Oliver Labs, Frank–Olaf Schreyer

Mathematik für Informatiker
Teil 1,2 und 3

Grundlagen, Analysis in einer Veränderlichen, Lineare Algebra, Analysis in mehreren Veränderlichen, Wahrscheinlichkeitstheorie und Statistik, Numerik

Version vom 6. Mai 2020, 9:51 Uhr

Inhaltsverzeichnis

Teil IV Mehrdimensionale Analysis

Teil V Wahrscheinlichkeitstheorie und Statistik

Abbildungsverzeichnis

Vorwort

Teil I

Grundlagen

Einführung

1

Logik und Beweismethoden

Die Regeln der Logik bilden die Grundlagen der mathematischen Argumentation. Zu den zentralen Anwendungen in der Informatik gehören:

- Schaltkreisentwurf,
- Entwicklung von Programmiersprachen,
- Verifikation von Hard- und Software,
- Suchen in Datenbanken,
- Automatisches Beweisen.

1.1 Logische Aussagen

Definition 1.1. *Eine **logische Aussage** ist ein Satz, dem genau ein Wahrheitswert wahr (w) oder falsch (f) zugeordnet ist.*

Beispiel 1.2.

1. Saarbrücken ist die Hauptstadt des Saarlandes. (w)
2. $2 + 2 = 7$. (f)
3. Im Saarland lebt es sich besser als in Rheinland–Pfalz. (subjektiv!)
4. $x + 1 = 3$. (keine logische Aussage, da x nicht spezifiziert ist)

1.2 Verknüpfungen von Aussagen

Durch **logische Operatoren** lassen sich aus logischen Aussagen, etwa A, B, neue formulieren, die dann **logische Formeln** heißen:

- $A \wedge B$ (A und B sind wahr), **Konjunktion**
- $A \vee B$ (A oder B oder beide sind wahr), **Disjunktion**
- $\neg A$ (A ist nicht wahr), **Negation**

Weitere Verknüpfungen sind:

- $A \Rightarrow B$(aus A folgt B), **Implikation**,
- $A \iff B$(A ist genau dann wahr, wenn B wahr ist), **Äquivalenz**.

Der Wahrheitswert dieser Aussagen hängt vom Wahrheitswert von A und B ab und ist über die **Wahrheitstafel** festgelegt:

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

Die Reihenfolge bei Ausführung von logischen Operationen legen wir durch Klammern fest.

Beispiel 1.3.

$$(A \Rightarrow B) \iff ((\neg A) \vee B) \tag{1.1}$$

Die zugehörige Wahrheitstabelle ist:

A	B	$A \Rightarrow B$	$\neg A$	$(\neg A) \vee B$	(1.1)
w	w	w	f	w	w
w	f	f	f	f	w
f	w	w	w	w	w
f	f	w	w	w	w

Um Klammern zu vermeiden, legen wir fest, dass $\neg$ die höchste Bindungs–Priorität, $\wedge$ und $\vee$ die mittlere Priorität sowie $\Rightarrow$ und $\iff$ die niedrigste Priorität haben. Die Formel (1.1) schreibt sich damit kürzer:

$$(A \Rightarrow B) \iff \neg A \vee B.$$

1.2.1 Erfüllbarkeit logischer Formeln

Im Folgenden fassen wir $A, B, C, D \ldots$ als **logische Variablen** auf, also Größen, die entweder wahr (w) oder falsch (f) sind. Aus diesen können wir dann mit den Operationen neue Aussagen formulieren.

Definition 1.4. *Eine logische Formel in den Variablen $A, B, C, \ldots$ ist **erfüllbar**, wenn es eine Belegung von $A, B, C, \ldots$ mit Wahrheitswerten gibt, so dass die Gesamtaussage wahr wird.*

Die Aufgabe, algorithmisch zu entscheiden, ob eine logische Formel erfüllbar ist, ist von zentraler Bedeutung in der Informatik.

Beispiel 1.5. Die Formel

$$(A \vee B) \wedge (A \vee \neg B) \wedge (\neg A \vee B) \wedge (\neg A \vee \neg B)$$

ist nicht erfüllbar.

Ob Erfüllbarkeit zu entscheiden schnell geht, hängt von der Struktur der logischen Formel ab. Wir unterscheiden die **disjunktive Normalform**

$$(x_{11} \wedge x_{12} \wedge \cdots \wedge x_{1n_1}) \vee (x_{21} \wedge x_{22} \wedge \cdots \wedge x_{2n_2}) \vee \cdots \vee (x_{k1} \wedge x_{k2} \wedge \cdots \wedge x_{kn_k}),$$

wobei $x_{ij} \in \{A, \neg A, B, \neg B, \ldots, w, f\}$ und die **konjunktive Normalform**:

$$(x_{11} \vee x_{12} \vee \cdots \vee x_{1n_1}) \wedge (x_{21} \vee x_{22} \vee \cdots \vee x_{2n_2}) \wedge \cdots \wedge (x_{k1} \vee x_{k2} \vee \cdots \vee x_{kn_k}).$$

In der disjunktiven Normalform ist dies einfach, für die konjunktive Normalform schwer, selbst im Fall $n_j = 3$. Diese Aussage ist der Grundpfeiler der Komplexitätstheorie.

1.2.2 Tautologien

Definition 1.6. *Eine logische Formel ist eine **logische Tautologie**, wenn sie unabhängig von der Belegung von logischen Variablen mit Wahrheitswerten wahr ist.*

Beispiel 1.7.

1. $(A \Rightarrow B) \iff ((\neg A) \vee B)$.
2. $(A \iff B) \iff (A \wedge B) \vee (\neg A \wedge \neg B)$. Dies zeigt die Wahrheitstafel:

A	B	$A \wedge B$	$\neg A \wedge \neg B$	$(A \wedge B) \vee (\neg A \wedge \neg B)$	$A \iff B$
w	w	w	f	w	w
w	f	f	f	f	f
f	w	f	f	f	f
f	f	f	w	w	w

Bemerkung 1.8. Zu entscheiden, ob eine logische Formel eine Tautologie ist, ist wenigstens so schwer wie die Entscheidung der Erfüllbarkeit. Ist X eine logische Formel, so gilt: X ist nicht erfüllbar genau dann, wenn $(X \Rightarrow f)$ eine Tautologie ist.

Beispiel 1.9. Weitere Tautologien:

3. $(A \Rightarrow B) \wedge (B \Rightarrow C) \Rightarrow (A \Rightarrow C)$ (**Transitivität der Implikation**)
4. Die **Gesetze von de Morgan**:

$$\neg(A \wedge B) \iff \neg A \vee \neg B,$$
$$\neg(A \vee B) \iff \neg A \wedge \neg B.$$

Dies folgt aus der Wahrheitstabelle:

A	B	$\neg(A \wedge B)$	$\neg A \vee \neg B$	$\neg(A \vee B)$	$(\neg A \wedge \neg B)$	Gesetze von de Morgan
w	w	f	f	f	f	w
w	f	w	w	f	f	w
f	w	w	w	f	f	w
f	f	w	w	w	w	w

Wir fassen die wichtigsten elementaren Tautologien zusammen:

Satz 1.10. *Die folgenden Formeln sind Tautologien:*

1. $A \vee B \iff B \vee A$,
   $A \wedge B \iff B \wedge A$ *(**Kommutativgesetz**).*
2. $(A \vee B) \vee C \iff A \vee (B \vee C)$,
   $(A \wedge B) \wedge C \iff A \wedge (B \wedge C)$ *(**Assoziativgesetze**).*
   Also macht $A \vee B \vee C$ und $A \wedge B \wedge C$ Sinn.
3. $A \wedge (B \vee C) \iff (A \wedge B) \vee (A \wedge C)$,
   $A \vee (B \wedge C) \iff (A \vee B) \wedge (A \vee C)$ *(**Distributivgesetze**)*
4. $A \vee f \iff A$,
   $A \wedge w \iff A$ *(**Identitätsgesetze**)*
5. $A \vee (\neg A) \iff w$ *(**Satz vom ausgeschlossenen Dritten**)*,
   $A \wedge (\neg A) \iff f$ *(**Satz vom Widerspruch**)*
6. $\neg(A \wedge B) \iff \neg A \vee \neg B$,
   $\neg(A \vee B) \iff \neg A \wedge \neg B$ *(**de Morgansches Gesetz**)*
7. $\neg(\neg A) \iff A$ *(**Doppelte Verneinung**)*
8. $A \vee A \iff A$,
   $A \wedge A \iff A$ *(**Idempotenzgesetze**)*
9. $A \Rightarrow B \iff (\neg B \Rightarrow \neg A)$ *(**Kontraposition**)*
10. $(A \Rightarrow B) \wedge (B \Rightarrow C) \Rightarrow (A \Rightarrow C)$ *(**Transitivität der Implikation**)*
11. $(\neg A \Rightarrow f) \iff A$ *(**Widerspruchsbeweis**)*

Beweis. Durch Aufstellen der Wahrheitstafeln. □

Korollar 1.11. *Jede logische Formel lässt sich mit Hilfe der Tautologien 1. – 11. aus dem Satz in konjunktive oder disjunktive Normalform bringen.*

Beweis. $(A \iff B) \iff (A \wedge B) \vee (\neg A \wedge \neg B)$ können wir verwenden, um $\iff$ – Zeichen zu beseitigen. $(A \Rightarrow B) \iff \neg A \vee B$ beseitigt $\Rightarrow$ – Zeichen. Die Regeln von de Morgan $\neg(A \vee B) \iff \neg A \wedge \neg B$, $\neg(A \wedge B) \iff \neg A \vee \neg B$ erlauben es uns, Negationszeichen nach Innen zu ziehen. Schließlich erlauben es die Distributivgesetze (3), die Formel in Richtung konjunktiver (disjunktiver) Normalform zu vereinfachen. □

1.3 Beweismethoden

Beweise werden in der Mathematik verwendet, um nachzuweisen, dass gewisse Sätze wahr sind. Dabei haben Tautologien eine wichtige Rolle.

Wir können z.B. $(A \Rightarrow B) \wedge (B \Rightarrow C) \Rightarrow (A \Rightarrow C)$ benutzen, um aus einem bekannten Satz A den Satz C in zwei Schritten zu beweisen. In der Informatik werden Beweise beispielsweise verwendet, um:

- die Korrektheit von Programmen nachzuweisen,
- zu zeigen, dass Programme terminieren,
- die Laufzeit eines Algorithmus in Abhängigkeit von der Eingabegröße der Daten zu analysieren,
- Zertifizierung des Outputs eines Programmes zu erreichen.

Zwei spezielle Beweismethoden heben wir heraus:

- Beweis durch Widerspruch,
- Beweis mit vollständiger Induktion.

1.3.1 Beweis durch Widerspruch

Der vielleicht älteste Beweis durch Widerspruch findet sich in Euklids Elementen.[1] Er handelt von **Primzahlen**, also natürlichen Zahlen $p \in \mathbb{N}, p > 1$, die nur durch 1 und sich selbst ohne Rest teilbar sind.

Satz 1.12 (Euklid). *Es gibt unendlich viele Primzahlen.*

Beweis. Angenommen, es gäbe nur endlich viele Primzahlen, etwa $p_1, \ldots, p_n$. Betrachten wir $q = p_1 \cdots p_n + 1$, so ist q durch keine der Zahlen p_j teilbar, da Rest 1 bei der Division bleibt. Also ist q selbst oder Primteiler von q eine Primzahl, die in der Liste $p_1, \ldots, p_n$ nicht vorkommt. Ein Widerspruch. □

[1] Geschrieben von Euklid um 325 v. Chr. war das Buch mit dem Titel *Die Elemente* über mehr als 2000 Jahre das wichtigste Mathematik–Buch.

1.3.2 Vollständige Induktion

Gegeben sei eine Aussage $A(n)$ für jede **natürliche Zahl** $n \in \mathbb{N} = \{1, 2, \dots\}$. Um die Aussage $A(n)$ für alle n zu zeigen, gehen wir wie folgt vor. Wir zeigen:

1. $A(1)$ gilt. (**Induktionsanfang**),
2. Für beliebiges n folgt unter der Voraussetzung, dass $A(n)$ gilt (genannt **Induktionsvoraussetzung** oder kurz I.-V.), dass auch $A(n+1)$ zutrifft (**Induktionsschritt**). Dies wird häufig auch kurz $n \to n+1$geschrieben.

Ist dies getan, so wissen wir:

$$A(1) \text{ ist wahr } \Rightarrow A(2) \text{ ist wahr } \Rightarrow A(3) \text{ ist wahr } \Rightarrow \cdots$$

Also ist $A(n)$ wahr für alle n. Diese Beweistechnik heißt **vollständige Induktion**.

Beispiel 1.13. Wir zeigen:

$$A(n)\colon 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Beweis. $A(1)\colon 1 = \frac{1\cdot(1+1)}{2}$, d.h. $A(1)$ ist wahr.

Für den Induktionsschritt $n \to n+1$ dürfen wir also annehmen, dass die Induktionsvoraussetzung $A(n)$ für ein $n \in \mathbb{N}$ wahr ist. Damit folgt:

$$\begin{aligned} 1 + 2 + \cdots + n + n + 1 &= (1 + 2 + \cdots + n) + (n+1) \\ &\overset{I.-V.}{=} \frac{n(n+1)}{2} + (n+1) \\ &= (n+1) \cdot \frac{n+2}{2}. \end{aligned}$$

Dies zeigt: $A(n+1)$.

Ein alternativer Beweis ist folgender:

$$\begin{array}{rccccc} & 1 & +2 & + & \cdots & +n \\ & n & +(n-1) & + & \cdots & +1 \\ \hline = & (n+1) & +(n+1) & + & \cdots & +(n+1) \end{array}$$

Dies ist aber gerade: $n(n+1) = 2(1 + \cdots + n)$. □

Bemerkung 1.14. Das Induktionsprinzip ist eine Aussage, die unsere Vorstellung von natürlichen Zahlen präzisiert: Ist $M \subset \mathbb{N}$,so dass gilt[2]:

[2] $\subset$ bezeichnet eine **Teilmenge**, $\subsetneq$ bezeichnet eine **echte Teilmenge**, d.h. eine Teilmenge, die nicht die ganze Menge ist.

$$(1 \in M) \text{ und } (n \in M \Rightarrow (n+1) \in M),$$

so folgt: $M = \mathbb{N}$.

Eine dazu äquivalente Aussage ist: Jede nicht leere Teilmenge $N \subset \mathbb{N}$ hat ein kleinstes Element. Betrachte $N = \mathbb{N} \backslash M$.

Definition 1.15. *Sei M eine Menge. Dann bezeichnet*

$$2^M := \{N \mid N \subset M\}$$

*die Menge aller Teilmengen von M, die sogenannte **Potenzmenge** von M. Manchmal wird statt 2^M auch $\mathcal{P}(M)$geschrieben.*

*Ist M eine endliche Menge, dann bezeichnet $|M|$ die **Anzahl der Elemente** von M.*

Satz 1.16. *Sei M eine endliche Menge. Dann gilt:*

$$|2^M| = 2^{|M|}.$$

Beispiel 1.17.

- $M = \emptyset$ (die **leere Menge**): $2^\emptyset = \{\emptyset\}$, also $|2^\emptyset| = 1 = 2^0$.
- $M = \{1\}$: $2^{\{1\}} = \{\emptyset, \{1\}\}$, also $|2^{\{1\}}| = 2 = 2^1$.
- $M = \{1,2\}$: $2^{\{1,2\}} = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$, also $|2^{\{1,2\}}| = 4 = 2^2$.

Beweis (von Satz 1.16). Ohne Einschränkung der Allgemeinheit können wir annehmen, dass $M = \{1, 2, \ldots, n\}$.

Induktionsanfang: ist bereits erbracht für $n = 0$ oder $n = 1$.

Induktionsschritt $n \to n+1$:

$$\begin{aligned} 2^{\{1,\ldots,n+1\}} &= \{N \subset \{1, \ldots, n+1\}\} \\ &= \{N \subset \{1, \ldots, n\}\} \cup \underbrace{\{N \subset \{1, \ldots, n+1\} \mid n+1 \in N\}}_{=\{N \mid N = N' \cup \{n+1\}, \text{ wobei } N' \subset \{1,\ldots,n\}\}}. \end{aligned}$$

Dabei bezeichnet die Notation $\cup$ die Vereinigung zweier Mengen, die disjunkt sind (also kein Element gemeinsam haben, siehe auch Abschnitt 2.4). Es folgt:

$$\begin{aligned} |2^{\{1,\ldots,n+1\}}| &= \left|2^{\{1,\ldots,n\}}\right| + \left|\left\{N \mid N = N' \cup \{n+1\}, N' \in 2^{\{1,\ldots,n\}}\right\}\right| \\ &= |2^{\{1,\ldots,n\}}| + |2^{\{1,\ldots,n\}}| \\ &\overset{I.-V.}{=} 2^n + 2^n = 2 \cdot 2^n = 2^{n+1} \\ &= 2^{|\{1,\ldots,n+1\}|}. \end{aligned}$$

□

1.3.3 Summen– und Produktzeichen

Induktion taucht auch in rekursiven Definitionen auf:

Definition 1.18 (Summen– und Produktzeichen). *Gegeben sind $n \in \mathbb{N}$* ***reelle Zahlen*** *$a_1, \ldots, a_n \in \mathbb{R}$. Wir setzen:*

$$\sum_{k=1}^{n} a_k := a_1 + \cdots + a_n.$$

Präzise: $\sum_{k=1}^{0} a_k := 0$ *(****leere Summe****) und* ***rekursiv:***

$$\sum_{k=1}^{n} a_k := \Big(\sum_{k=1}^{n-1} a_k\Big) + a_n.$$

Analog definieren wir

$$\prod_{k=1}^{n} a_k = a_1 \cdots a_n$$

exakter durch: $\prod_{k=1}^{0} a_k = 1$ *und*

$$\prod_{k=1}^{n} a_k = \Big(\prod_{k=1}^{n-1} a_k\Big) \cdot a_n.$$

Beispiel/Definition 1.19. Die Zahl

$$n! := \prod_{k=1}^{n} k = 1 \cdot 2 \cdot 3 \cdots n$$

heißt **Fakultät** von n (gelesen: n Fakultät). Insbesondere gilt: $0! = 1$.

Beispiel 1.20. Für jedes $n \in \mathbb{N} \cup \{0\}$ gilt:

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

Beweis mit vollständiger Induktion:
Induktionsanfang: $n = 0$ oder $n = 1$:

$$\sum_{k=1}^{1} k^2 = n^2 \stackrel{!}{=} \frac{1 \cdot 2 \cdot 3}{6},$$

was richtig ist.

Induktionsschritt $n \to n+1$:

$$\begin{aligned}
\sum_{k=1}^{n+1} k^2 &= \Big(\sum_{k=1}^{n} k^2\Big) + (n+1)^2 \\
&\overset{I.-V.}{=} \frac{n(n+1)(2n+1)}{6} + (n+1)^2 \\
&= (n+1)\cdot\Big(\frac{n(2n+1)}{6} + n + 1\Big) = \frac{n+1}{6}\cdot\Big(2n^2 + n + 6n + 6\Big) \\
&= \frac{n+1}{6}\cdot\Big(2n^2 + 7n + 6\Big) = \frac{n+1}{6}\cdot\Big((n+2)(2n+3)\Big) \\
&= \frac{(n+1)\cdot(n+2)\cdot(2(n+1)+1)}{6},
\end{aligned}$$

was die Aussage beweist.

Beispiel/Definition 1.21. Wir betrachten nochmals das Beispiel

$$h\colon \mathbb{Z} \to \mathbb{Z},\ n \mapsto h(n) := \sum_{k=1}^{n} k^2$$

von eben, wobei $\mathbb{Z} = \{0, 1, -1, 2, -2, \dots\}$die Menge der **ganzen Zahlen** bezeichnet. Wir fragen uns nun, wie wir selbst auf die Formel hätten kommen können. Es erscheint klar, dass für $n \geq 0$ gilt:

$$h(n) = \sum_{k=1}^{n} k^2 \approx \int_0^n t^2\, dt = [\frac{1}{3}t^3]_0^n = \frac{1}{3}n^3.$$

Daraus leiten wir die Hypothese ab, dass auch die Summe durch ein sogenanntes **Polynom** vom **Grad** 3 in n beschrieben wird:

$$h(n) = \sum_{k=1}^{n} k^2 = a_3n^3 + a_2n^2 + a_1n + a_0, \text{ für gewisse } a_i \in \mathbb{Q},$$

wobei $\mathbb{Q}$ die Menge der **rationalen Zahlen** bezeichnet, die wir erst in Beispiel 3.12 sauber einführen werden. Natürlich kann dies nicht für alle ganzen Zahlen $n \in \mathbb{Z}$ korrekt sein, da $h(n) = 0$ für alle $n < 0$ und da ein Polynom p, das für unendlich viele Werte den Funktionswert 0 ergibt, schon das **Nullpolynom** (d.h. $p(n) = 0$ für alle n) sein muss. Wir können also nur hoffen, eine solche Formel für $n \geq 0$ zu finden. Offenbar ist $h(0) = 0$ und $g(0) = a_0$, so dass sofort $a_0 = 0$ folgt.

Eine Strategie, die weiteren a_i zu bestimmen, ist folgende: Ist $f\colon \mathbb{Z} \to \mathbb{Z}$ eine Abbildung, so definieren wir die erste **Differenzfunktion**

$$\Delta f\colon \mathbb{Z} \to \mathbb{Z},\ (\Delta f)(n) = f(n) - f(n-1).$$

Bezeichnen wir in unserem Beispiel $g(n) = a_3n^3 + a_2n^2 + a_1n + a_0$, so ergeben sich als erste und weitere Differenzenfunktionen:

$$\begin{aligned}
g(n) &= a_3n^3 + a_2n^2 + a_1n + a_0,\\
(\Delta g)(n) &= a_3(n^3 - (n-1)^3) + a_2(n^2 - (n-1)^2) + a_1(n - (n-1))\\
&= a_3(3n^2 - 3n + 1) + a_2(2n - 1) + a_1,\\
(\Delta^2(g))(n) &= 3a_3(n^2 - (n-1)^2) + \cdots = 6a_3n - 6a_3 + 2a_2,\\
(\Delta^3(g))(n) &= 6a_3.
\end{aligned}$$

Hierbei ist $(\Delta^k(g))(n) := (\Delta(\cdots(\Delta(g))))(n)$ die k-fache Anwendung der Funktion Δ auf g. Wir sehen damit, dass $\Delta^3(g)$ nicht mehr von n abhängt, dass wir also a_3 direkt ablesen können, wenn wir nur Werte $(\Delta^3(g))(n)$ für n genügend groß berechnet haben. Dazu betrachten wir folgende Tabelle für unser $h(n) = \sum_{k=1}^{n} k^2$:

n	$h(n)$	$\Delta h(n)$	$\Delta^2 h(n)$	$\Delta^3 h(n)$
0	0	0	0	0
1	1	1	1	1
2	5	4	3	2
3	14	9	5	2
4	30	16	7	2

Ist also wirklich $h(n) = g(n)$ für alle $n = 0, 1, 2, \ldots$, so muss gelten (bei $\Delta^3 h(0)$, $\Delta^3 h(1)$, $\Delta^3 h(2)$ geht $h(-1)$ ein):

$$2 = (\Delta^3(h))(3) = (\Delta^3(g))(3) = 6a_3, \text{ also } a_3 = \frac{1}{3}.$$

Wir betrachten nun die neue Funktion $i(n) := h(n) - \frac{1}{3}n^3$, von der wir annehmen, dass sie für $n \geq 0$ durch ein quadratisches Polynom beschrieben wird. Wir können a_2 also wieder aus einem einzigen Wert aus einer Tabelle ablesen, da $\Delta^2(g)$ nicht von n abhängt, falls $a_3 = 0$ ist, und genauer den Wert $2a_2$ annimmt, wie wir weiter oben berechnet haben:

n	$i(n) = h(n) - \frac{1}{3}n^3$	$\Delta i(n)$	$\Delta^2 i(n)$
0	$0 - 0 = 0$	0	0
1	$1 - \frac{1}{3} = \frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$
2	$5 - \frac{8}{3} = \frac{7}{3}$	$\frac{5}{3}$	$\frac{3}{3} = 1$
3	$14 - 9 = 5$	$\frac{8}{3}$	1
4	$30 - \frac{64}{3} = \frac{26}{3}$	$\frac{11}{3}$	1

Dies liefert:

$$1 = 2a_2, \text{ also } a_2 = \frac{1}{2}.$$

Wir fahren analog fort, definieren also $j(n) := h(n) - \frac{1}{3}n^3 - \frac{1}{2}n^2$, von dem wir annehmen, dass es ein Polynom vom Grad 1 in n ist für $n \geq 0$. Wir wir oben berechnet haben, ergibt sich für $(\Delta g)(n)$ mit $a_3 = 0$ und $a_2 = 0$ aber der Wert a_1, der unabhängig von n ist. Wir können demnach a_1 aus folgender Tabelle ablesen:

n	$j(n) = i(n) - \frac{1}{2}n^2 = h(n) - \frac{1}{3}n^3 - \frac{1}{2}n^2$	$(\Delta j)(n)$
0	0	0
1	$\frac{2}{3} - \frac{1}{2} = \frac{1}{6}$	$\frac{1}{6}$
2	$\frac{7}{3} - \frac{1}{2} \cdot 4 = \frac{2}{6}$	$\frac{1}{6}$
3	$5 - \frac{1}{2} \cdot 9 = \frac{1}{2} = \frac{3}{6}$	$\frac{1}{6}$
4	$\frac{26}{3} - \frac{1}{2} \cdot 16 = \frac{2}{3} = \frac{4}{6}$	$\frac{1}{6}$

Insgesamt haben wir also das Polynom

$$g(n) = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n = \frac{n(n+1)(2n+1)}{6}$$

gefunden. Wie wir im vorigen Beispiel 1.20 schon bewiesen haben, ist dies auch tatsächlich das gesuchte und es gilt für jedes $n \in \{0, 1, 2, \dots\}$:

$$\sum_{k=1}^{n} k^2 = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n = \frac{n(n+1)(2n+1)}{6}.$$

1.3.4 Die Fibonacci–Zahlen

Wir definieren die **Fibonacci–Zahlen** $f_n$rekursiv:

$$f_0 := 0, f_1 := 1, \ f_{n+1} := f_n + f_{n-1} \text{ für } n = 1, 2, 3, \dots$$

Die ersten Werte sind:

$$0, 1, 1, 2, 3, 5, 8, 13, \dots$$

Satz 1.22. *Die n-te Fibonacci–Zahl ist*[3]

$$f_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right).$$

[3]Für eine exakte Einführung der Quadratwurzel siehe Abschnitt 5.6. Vorläufig werden wir auf das aus der Schule bekannte Wissen zurückgreifen. Demnach ist $\sqrt{a}$ für $0 < a \in \mathbb{R}$ eine positive Zahl, so dass $\sqrt{a}^2 = a$ ergibt.

Beweis. Es gilt:

$$f_0 = \frac{1}{\sqrt{5}}(1-1) = 0,$$
$$f_1 = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2} - \frac{1-\sqrt{5}}{2}\right) = \frac{1}{\sqrt{5}} \cdot \left(\frac{2\sqrt{5}}{2}\right) = 1.$$

Wir beweisen die Aussage

$$A(n)\colon f_k = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k\right) \text{ für } k = 0, \ldots, n$$

mit vollständiger Induktion. Den Induktionsanfang $A(1)$ haben wir oben bereits erledigt.

Für den Induktionsschritt $A(n) \Rightarrow A(n+1)$ betrachten wir:

$$\begin{aligned} f_{n+1} &= f_n + f_{n-1} \quad \text{(nach Definition)} \\ &\overset{\text{I.–V.}}{=} \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right) + \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n-1}\right) \\ &= \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1} \cdot \left(\frac{1+\sqrt{5}}{2} + 1\right) - \left(\frac{1-\sqrt{5}}{2}\right)^{n-1} \cdot \left(\frac{1-\sqrt{5}}{2} + 1\right)\right). \end{aligned}$$

Nun gilt: $\frac{1+\sqrt{5}}{2} + 1 = \frac{3+\sqrt{5}}{2}$ und $\left(\frac{1+\sqrt{5}}{2}\right)^2 = \cdots = \frac{3+\sqrt{5}}{2}$. Analog: $\frac{1-\sqrt{5}}{2} + 1 = \frac{3-\sqrt{5}}{2} = \left(\frac{1-\sqrt{5}}{2}\right)^2$ und damit:

$$\begin{aligned} f_{n+1} &= \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1+2} - \left(\frac{1-\sqrt{5}}{2}\right)^{n-1+2}\right) \\ &= \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right). \end{aligned}$$

□

Wie wir auf diese Formel kommen konnten, werden wir im nächsten Semester (Abschnitt 24.4.2) lernen.

Aufgaben

Aufgabe 1.1 (Wahrheitstafel). Geben Sie die Wahrheitstafel der folgenden logischen Formel an:

$$A \wedge \neg B \Rightarrow (C \vee A \Leftrightarrow (B \Rightarrow C \wedge A)).$$

Ist die Formel eine Tautologie, erfüllbar oder unerfüllbar?

Aufgabe 1.2 (Vier Zeugen). Ein Kommissar hat zu einem Verbrechen 4 Zeugen vernommen. Aus den Vernehmungen hat er folgende Schlussfolgerungen gezogen:

- Wenn der Butler die Wahrheit sagt, dann auch der Koch.
- Koch und Gärtner können nicht beide die Wahrheit sagen.
- Gärtner und Hausmeister lügen nicht beide.
- Wenn der Hausmeister die Wahrheit sagt, dann lügt der Koch.

1. Modellieren Sie die Informationen des Kommissar als logische Formeln. Verwenden Sie dazu die Variablen B, K, G und H.
2. Bei welchen Zeugen kann der Kommissar sicher sein, dass sie lügen? Bei welchen kann er sicher sein, dass sie die Wahrheit sagen? Erklären Sie, wie Sie auf Ihr Ergebnis kommen!

Aufgabe 1.3 (Zwei Investmentbänker). Ein Mann ist bei einer Kurz–Beratung mit zwei Investmentbänkern, A und B genannt, in der er herausfinden möchte, ob er seine Erbschaft lieber in die Anlagemöglichkeit 1 oder in die Anlagemöglichkeit 2 investieren soll. Leider lässt die kostenlose Beratung der Bank nur eine einzige Ja/Nein–Frage an nur einen der beiden Berater zu. Ein Freund hatte ihn zuvor davon informiert, dass einer der beiden immer die Wahrheit sagt und dass der andere stets lügt. Der Freund wusste aber unglücklicherweise nicht mehr, welcher der beiden welcher ist. Mit welcher Frage kann der Mann herausfinden, welche die gute und welche die schlechte Anlagemöglichkeit ist?

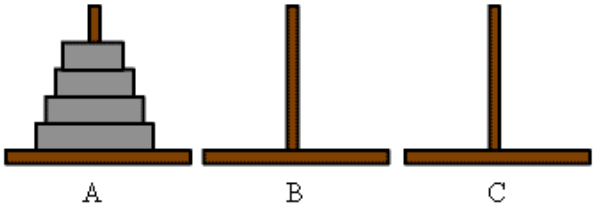
Aufgabe 1.4 (Logische Verknüpfungen). Sei $\overline{\wedge}$ das Zeichen für *nicht und*, d.h. für zwei logische Variablen A, B ist $A \mathbin{\overline{\wedge}} B = \neg(A \wedge B)$.

1. Stellen Sie die drei logischen Verknüpfungen $\neg$, $\wedge$ und $\vee$ jeweils ausschließlich durch $\overline{\wedge}$ dar.
2. Seien $X_1, ..., X_n$ logische Variablen und $f(X_1, ..., X_n)$ eine beliebige logische Funktion mit in X_1 bis X_n mit gegebener Wahrheitstafel. Zeigen Sie: f lässt sich durch $\overline{\wedge}$ darstellen.

Aufgabe 1.5 (Induktion). Finden Sie eine geschlossene Formel, die nur von $n \in \mathbb{N}$ abhängt, für

$$\sum_{k=1}^{n} k^3$$

(beispielsweise mit der in der Vorlesung erläuterten Methode, oder auch anders) und beweisen Sie die Formel per Induktion.



Aufgabe 1.6 (Die Türme von Hanoi). Das Spiel *Die Türme von Hanoi* besteht aus 3 Spielfeldern, auf denen $n \in \mathbb{N}$ Scheiben paarweise verschiedener Größe gestapelt werden können. Zu Beginn des Spiels sind alle Scheiben auf einem der Spielfelder der Größe nach gestapelt (die unten liegende Scheibe ist die größte, wie im Bild zu sehen). Ziel des Spiels ist es, den Anfangsstapel auf ein anderes Feld zu versetzen, so dass er dort wieder in der gleichen Stapel–Reihenfolge liegt. Dazu darf in jedem Spielzug die oberste Scheibe eines beliebigen Turms auf einen anderen Turm, der keine kleinere Scheibe enthält, gelegt werden.

Geben Sie einen Algorithmus an (Papierform genügt), der dieses Problem löst, und beweisen Sie die Korrektheit Ihres Algorithmus. Stellen Sie eine Formel für die Anzahl der notwendigen Züge auf und beweisen Sie diese mit vollständiger Induktion.

Aufgabe 1.7 (Erfüllbarkeit, konjunktive Normalform). Finden Sie für die folgenden Aussagen jeweils heraus, ob sie erfüllbar oder sogar eine Tautologie sind?

1. $(X \Rightarrow (Y \Rightarrow Z)) \iff ((X \wedge Y) \Rightarrow Z)$,
2. $(A \wedge B) \vee (A \Rightarrow B)$.

Geben Sie für die zweite Aussage auch die konjunktive Normalform an.

2

Mengen und Abbildungen

Die **Mengenlehre** ist das fundamentale Hilfsmittel zur Spezifizierung mathematischer Objekte. In der Informatik wird sie beispielsweise überall dort verwendet, wo Alphabete, Halbgruppen, Algebren, Verbände eine Rolle spielen. Zu den unmittelbaren Anwendungen gehören Datenbanken.

2.1 Mengentheoretische Sprechweisen

Eine **Menge** M ist eine Kollektion wohlbestimmter Objekte, die wir **Elemente** von M nennen. Mengen lassen sich auf zwei Weisen spezifizieren:

1. Aufzählen der Elemente,
2. durch eine charakteristische Eigenschaft.

Beispiel 2.1.

1. $\{a, b, c, \ldots, z\}$ ist die Menge der Buchstaben des Alphabets.
2. $\{a, b, a\} = \{a, b\}$: mehrfaches Aufzählen von Elementen ändert die Menge nicht.
3. $\{b, a\} = \{a, b\}$: auf die Reihenfolge kommt es beim Aufzählen der Elemente einer Menge nicht an.
4. Elemente von Mengen können auch Städte sein:

$$\begin{aligned} H &= \{\text{Hauptstädte der Bundesländer}\} \\ &= \{\text{Berlin, Bremen, Hamburg, Saarbrücken, Hannover, Kiel,} \\ &\quad \text{Schwerin, Magdeburg, Potsdam, Düsseldorf, Dresden, Erfurt,} \\ &\quad \text{Mainz, Wiesbaden, Stuttgart, München}\} \end{aligned}$$

5. $\emptyset = \{\}$, die **leere Menge**.
6. Alle reellen Zahlen x **mit der Eigenschaft** (kurz | geschrieben) $x^2 - x - 1$:

$$\{ x \in \mathbb{R} \mid x^2 - x - 1 = 0 \} = \left\{ \frac{1 + \sqrt{5}}{2}, \frac{1 - \sqrt{5}}{2} \right\}.$$

   Dies kann man beispielsweise mit der aus der Schule bekannten p, q–Formel berechnen.

Wichtige Mengen von Zahlen haben spezielle Notationen; einige davon haben wir bereits kennen gelernt:

- $\mathbb{N} = \{1, 2, 3, \dots\}$, Menge der natürlichen Zahlen,
- $\mathbb{Z} = \{0, 1, -1, 2, -2, \dots\}$, Menge der ganzen Zahlen.
- $\mathbb{Q} = \{\frac{a}{b} \mid a, b \in \mathbb{Z}, b \neq 0\}$, Menge der rationalen Zahlen (siehe Beispiel 3.12).
- $\mathbb{R} = \{$ unendliche Dezimalzahlen $\}$, Menge der reellen Zahlen (siehe dazu auch Kapitel 4).
- $\mathbb{C}$ = Menge der komplexen Zahlen (siehe Abschnitt 7.1).

Ist M eine Menge und a ein Element, so schreiben wir $a \in M$. $a \notin M$ steht für: a ist kein Element von M. $M \ni a$ steht für: M enthält das Element a.

2.2 Teilmengen und Venn-Diagramme

Eine **Teilmenge** N einer Menge M ist eine Menge, für die gilt:

$$a \in N \Rightarrow a \in M.$$

Schreibweisen:

- $N \subset M$: N ist eine Teilmenge von M.
- $N \not\subset M$: N ist keine Teilmenge von M.
- $N \subsetneq M$: N ist eine **echte Teilmenge** von M, d.h. $N \subset M$, aber $N \neq M$.

Für Teilmengen A, B einer Menge M bezeichnet

$$A \cap B = \{x \in M \mid x \in A \text{ und } x \in B\}$$

den **Durchschnitt**. Mit **Venn-Diagrammen** kann man Beziehungen von Teilmengen oft besonders anschaulich darstellen; siehe Abb. 2.1 für den Durchscnitt. Die Menge

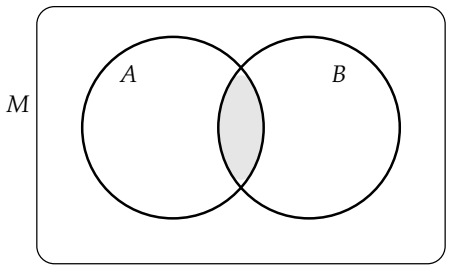


Abbildung 2.1. Der Schnitt $A \cap B$ zweier Mengen $A, B \subset M$, hervorgehoben durch die graue Einfärbung.

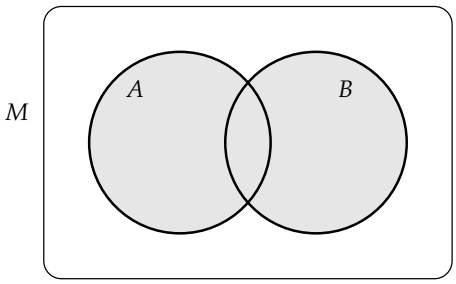


Abbildung 2.2. Die Vereinigung $A \cup B$ zweier Mengen $A, B \subset M$, hervorgehoben durch die graue Einfärbung.

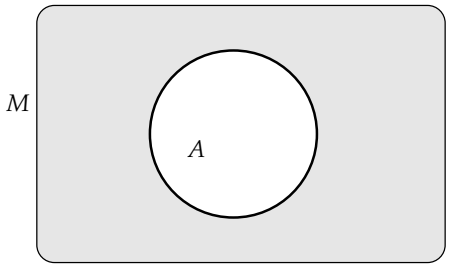


Abbildung 2.3. Das Komplement $\overline{A} = M \backslash A$ einer Menge $A \subset M$, hervorgehoben durch die graue Einfärbung.

$$A \cup B = \{x \in M \mid x \in A \text{ oder } x \in B\}$$

heißt **Vereinigung** von A und B; siehe Abb. 2.2. Die Menge

$$\bar{A} = M \backslash A = \{x \in M \mid x \notin A\}$$

heißt **Komplement** von A in M, siehe Abb. 2.3. Die Menge

$$A \backslash B = \{x \in M \mid x \in A \text{ und } x \notin B\}$$

heißt **Differenzmenge** von A und B, siehe Abb. 2.4. Manchmal wird statt $A \backslash B$ auch $A - B$ geschrieben.

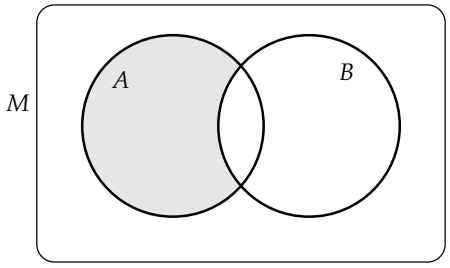


Abbildung 2.4. Die Differenz $A \backslash B$ zweier Mengen $A, B \subset M$, hervorgehoben durch die graue Einfärbung.

2.3 Rechenregeln für Mengen

Satz 2.2 (Rechenregeln für Mengen). *Es seien $A, B, C \subset M$. Dann gilt:*

1. $A \cap B = B \cap A$.
   $A \cup B = B \cup A$, ***Kommutativgesetze***
2. $(A \cap B) \cap C = A \cap (B \cap C)$.
   $(A \cup B) \cup C = A \cup (B \cup C)$, ***Assoziativgesetze***,
3. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, ***Distributivgesetze***,
4. $A \cup \emptyset = A$, $A \cap M = A$, ***Identitätsgesetze***.
5. $A \cup (M \backslash A) = A \cup \bar{A} = M$,
   $A \cap (M \backslash A) = A \cap \bar{A} = \emptyset$, ***Mengen und ihr Komplement***.
6. $\overline{(A \cap B)} = \bar{A} \cup \bar{B}$,
   $\overline{(A \cup B)} = \bar{A} \cap \bar{B}$, ***Gesetze von de Morgan***.

7. $\overline{(\bar{A})} = A$, ***Gesetz vom doppelten Komplement***.

Beweis. Die Aussagen folgen unmittelbar aus den analogen Aussagen der Logik. Alternativ mit Venn–Diagrammen (siehe Abb. 2.5 und 2.6): □

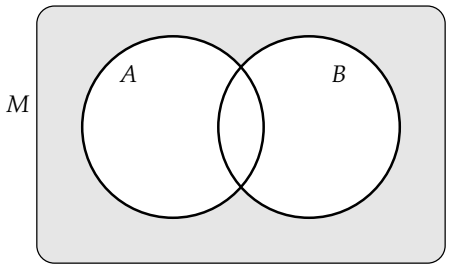


Abbildung 2.5. $\overline{A} \cap \overline{B} = \overline{A \cup B}$, hervorgehoben durch die graue Einfärbung.

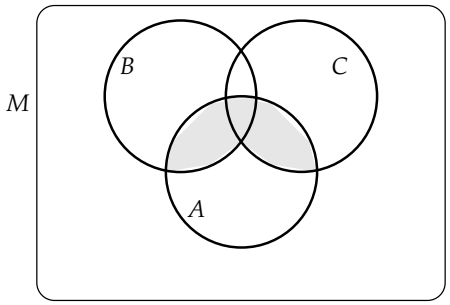


Abbildung 2.6. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, hervorgehoben durch die graue Einfärbung.

2.4 Disjunkte Mengen

Sind A und B endlich, dann gilt:

$$|A \cup B| = |A| + |B| - |A \cap B|,$$

da in der Summe $|A| + |B|$ die Elemente von $A \cap B$ doppelt gezählt würden. Besser:

$$|A \cup B| + |A \cap B| = |A| + |B|,$$

denn diese Formel macht auch für unendliche Mengen Sinn: $|A| = \infty$. A und B heißen **disjunkt**, wenn $A \cap B = \emptyset$.

2.5 Kartesische Produkte

Es seien A und B Mengen. Dann ist das **kartesische Produkt**

$$A \times B = \{(a, b) \mid a \in A, b \in B\}$$

die Menge der **geordneten Paare** von Elementen aus A und Elementen aus B.

Beispiel 2.3.

1. $\{a, b, \ldots, h\} \times \{1, 2, \ldots, 8\}$ ist beim Schach gebräuchlich.
2. Mit $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(a, b) \mid a, b \in \mathbb{R}\}$ lassen sich Punkte in der Ebene spezifizieren.

Für die Anzahl der Elemente eines kartesischen Produktes gilt:

$$|A \times B| = |A| \cdot |B|.$$

2.6 Definition des Binomialkoeffizienten

Die Potenzmenge 2^M einer Menge M, die aus den Teilmengen von M besteht, hatten wir schon eingeführt.

Definition 2.4. *Die Anzahl der k-elementigen Teilmengen einer n-elementigen Menge bezeichnen wir mit* $\binom{n}{k}$ *(gelesen: n über k, englisch n choose k).* $\binom{n}{k}$ *heißt auch* ***Binomialkoeffizient****.*

2.7 Eine Formel für den Binomialkoeffizienten

Der Binomialkoeffizient kann durch Fakultäten (siehe Definition 1.19) ausgedrückt werden:

Satz 2.5. *Für* $0 \leq k \leq n$ *gilt:*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Beweis. Induktion nach n. Für $n = 0$ ist die Aussage richtig. Die leere Menge $\emptyset$ hat genau eine 0-elementige Teilmenge, nämlich $\emptyset$. Also: $\binom{0}{0} = 1 = \frac{0!}{0!\cdot 0!}$ gilt.

Num zum Induktionsschritt $n \to n+1$: Für $k = 0$ ist die Aussage klar: Auch $\{1, \ldots, n+1\}$ hat genau eine 0-elementige Teilmenge, nämlich $\emptyset$, also gilt:

$$\binom{n+1}{0} = 1 = \frac{(n+1)!}{0!(n+1)!},$$

wie behauptet. Sei also $k \geq 1$. Wir betrachten

$$\binom{n+1}{k} = \left|\left\{A \subset \{1, \ldots, n+1\} \mid |A| = k\right\}\right|.$$

Die Menge auf der rechten Seite zerlegt sich disjunkt in

$$\left\{A \subset \{1, \ldots, n\} \mid |A| = k\right\} \cup \left\{A' \cup \{n+1\} \mid A' \subset \{1, \ldots, n\}, |A'| = k-1\right\}.$$

Also:

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

Die Induktionsvoraussetzung gibt nun:

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

□

Der Beweis des vorigen Satzes zeigt insbesondere:

Lemma 2.6. *Für $n, k \in \mathbb{N}$ gilt:*

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

Beweis. □

Der Name Binomialkoeffizient kommt von folgendem Satz:

n							
0				1			
1			1		1		
2		1		2		1	
3	1		3		3		1
4		1		4		6	
5	1		5		10		10
6		1		6		15	

n						
0			1			
1		1		1		
2	1		2		1	
3		3		3		1
4	1		4		6	
5		5		10		10
6	1		6		15	

n						
0			1			
1		1		1		
2	1		2		1	
3		3		3		1
4	1		4		6	
5		5		10		10
6	1		6		15	

Abbildung 2.7. Das Pascalsche Dreieck mit den Einträgen $\binom{n}{k}$ für $k = 0, \ldots, n$. Diese Darstellung suggeriert (siehe auch Lemma 2.6): $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ für $k \geq 1$.

Satz 2.7 (Binomische Formel). *Es seien $a, b \in \mathbb{R}$ und $n \in \mathbb{N}$. Dann gilt (siehe auch Abb. 2.7):*

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

Beweis (von Satz 2.7). Induktion nach n. Induktionsanfang: $n = 1$.

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

Induktionsschritt $n \to n+1$:

$$\begin{aligned}
(a+b)^{n+1} &= (a+b)(a+b)^n \\
&\overset{\text{I.-V.}}{=} (a+b) \cdot \sum_{k=0}^{n} \binom{n}{k} a^{n-k}\, b^k \\
&= \sum_{k=0}^{n} \binom{n}{k} a^{n-k+1}\, b^k + \sum_{k=0}^{n} \binom{n}{k} a^{n-k}\, b^{k+1} \\
&= \binom{n}{0} a^{n+1} + \sum_{k=1}^{n} \binom{n}{k} a^{n-k+1}\, b^k + \sum_{l=1}^{n+1} \binom{n}{l-1} a^{n-(l-1)}\, b^l \\
&= \binom{n+1}{0} a^{n+1} + \sum_{k=1}^{n} \left(\binom{n}{k} + \binom{n}{k-1} \right) a^{n-k+1}\, b^k + \binom{n}{n} b^{n+1} \\
&\overset{\text{Lemma 2.6}}{=} \binom{n+1}{0} a^{n+1} + \sum_{k=1}^{n} \binom{n+1}{k} a^{n+1-k}\, b^k + \binom{n+1}{n+1} b^{n+1} \\
&= \sum_{k=0}^{n+1} \binom{n+1}{k} a^{n+1-k}\, b^k.
\end{aligned}$$

□

2.8 Abbildungen

Um Beziehungen zwischen Mengen zu studieren, benötigen wir sogenannte Abbildungen.

2.8.1 Definition und erste Beispiele

Definition 2.8. *Eine **Abbildung** $f\colon M \to N$ zwischen zwei Mengen M und N ist (gegeben durch) eine Vorschrift, die jedem Elemente $a \in M$ ein Element $f(a) \in N$ zuordnet. M heißt dabei **Definitionsmenge** und N **Zielmenge** der Abbildung.*

Beispiel/Definition 2.9.

1. $\{$ Studierende der UdS $\} \to \mathbb{N}$, $x \mapsto$ Matrikelnummer .
2. $f\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto f(x) = x^2$ (siehe Abb. 2.8)

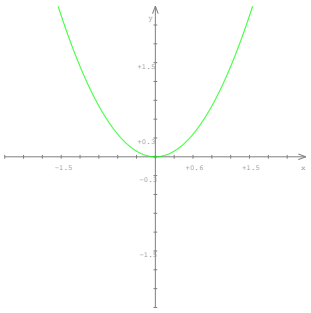


Abbildung 2.8. Graph einer Parabel mit Gleichung $f(x) = x^2$.

Zu einer Abbildung $f\colon M \to N$ heißt die Teilmenge

$$G_f = \{(x, y) \in M \times N \mid y = f(x)\}$$

der **Graph der Abbildung** f. Aus dem Graphen lässt sich die Abbildung zurückgewinnen:

$$G_f \cap (\{a\} \times N) = \{(a, f(a))\}.$$

3. Der **ganzzahlige Anteil** einer reellen Zahl ist durch folgende Abbildung gegeben (siehe auch Abb. 2.9):

$$\text{entier}\colon \mathbb{R} \to \mathbb{Z} \subset \mathbb{R},\ x \mapsto y = \text{entier}(x) = \lfloor x \rfloor = \max\left\{ n \in \mathbb{Z} \mid n \leq x \right\}.$$

$\lceil x \rceil = \min\{ n \in \mathbb{Z} \mid n \geq x \}$.

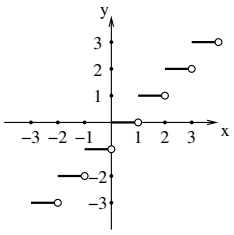


Abbildung 2.9. Graph der entier Funktion. Ein kleiner, leerer Kreis zeigt dabei an, dass der umkreiste Punkt nicht zum Graphen gehört.

4. Sei $A \subset \mathbb{R}$. Dann heißt $\chi_A \colon \mathbb{R} \to \{0,1\} \subset \mathbb{R}$,

$$\chi_A = \begin{cases} 0, & \text{falls } a \notin A, \\ 1, & \text{falls } a \in A. \end{cases}$$

   die **charakteristische Funktion** für A.

5. Der Graph der Funktion $\chi_{\mathbb{Q}}$ lässt sich schlecht zeichnen.

2.8.2 Injektivität, Surjektivität und Bijektivität

Definition 2.10. *Sei $f \colon M \to N$ eine Abbildung und $A \subset M$. Dann heißt zu $A \subset M$ die Menge*

$$f(A) = \{f(a) \mid a \in M\} \subset N$$

*das **Bild** von A unter f. Zu $B \subset N$ heißt die Menge*

$$f^{-1}(B) = \{a \in M \mid f(a) \in B\}$$

*das **Urbild** von B. f^{-1} ist eine (neue) Abbildung $f^{-1} \colon 2^N \to 2^M$ zwischen den Potenzmengen von N und M. Für ein Element $b \in N$ schreiben wir kürzer: $f^{-1}(b) = f^{-1}(\{b\})$.*

Im Verlaufe der Vorlesung werden wir sehen, dass die folgenden Eigenschaften von Abbildungen immer wieder eine zentrale Rolle spielen werden:

Definition 2.11. *Eine Abbildung $f \colon M \to N$ heißt **injektiv** (lat. iniacere: hineinwerfen), wenn $x_1, x_2 \in M, x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$ gilt. f heißt **surjektiv** (lat. suriectere: überwerfen), wenn $f(M) = N$. Eine Abbildung, die injektiv und surjektiv ist, heißt **bijektiv**.*

Beispiel 2.12. f in Abb. 2.10 ist weder injektiv noch surjektiv: $f(a) = f(b)$, $3 \notin f(M)$.

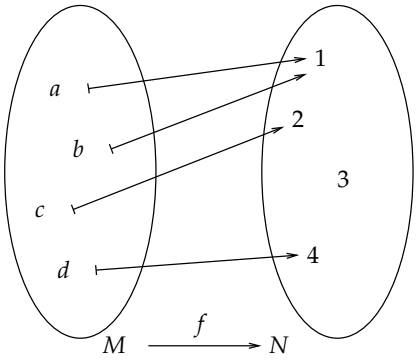


Abbildung 2.10. Injektivität und Surjektivität: Die gezeigte Abbildung f ist weder injektiv noch surjektiv.

Satz 2.13. *Sei $f\colon M \to N$ eine Abbildung zwischen endlichen Mengen mit $|M| = |N|$. Dann sind äquivalent:*

1. *f ist injektiv,*
2. *f ist surjektiv,*
3. *f ist bijektiv.*

Beweis. 2. $\Rightarrow$ 3. Sei f surjektiv. Dann gilt:

$$|N| = \sum_{b\in N} |\{b\}| \leq \sum_{b\in N} |f^{-1}(b)| = |M|.$$

Wegen $|N| = |M|$ muss Gleichheit gelten. Also:

$$|\{b\}| = |f^{-1}(b)| = 1 \text{ für alle } b \in N,$$

d.h. f ist bijektiv.

Wir zeigen nun: 1. $\Rightarrow$ 3. Sei dazu f injektiv. Dann gilt:

$$|N| = \sum_{b\in N} |\{b\}| \geq \sum_{b\in N} |f^{-1}(b)| = |M|$$

wegen $|N| = |M|$ gilt wieder:

$$|\{b\}| = |f^{-1}(b)| = 1 \text{ für alle } b,$$

d.h. f ist bijektiv.

Die verbleibenden Implikationen 3. $\Rightarrow$ 1. und 3. $\Rightarrow$ 2. sind nach Definition der Bijektivität richtig. $\square$

Mehrere wichtige Resultate folgen direkt daraus:

Korollar 2.14 (aus dem Beweis). *Seien M, N endliche Mengen. Ist $f\colon M \to N$ injektiv, so gilt: $|M| \leq |N|$. Ist $f\colon M \to N$ surjektiv, so gilt: $|M| \geq |N|$.*

Korollar 2.15 (Schubfachprinzip). *Seien M, N endliche Mengen. Eine Abbildung $f\colon M \to N$ mit $|M| > |N|$ ist nicht injektiv.*

Eine nette Anwendung davon ist folgende:

Proposition 2.16. *Unter beliebigen $n^2 + 1$ vielen Punkten $P_1, \ldots, P_{n^2+1}$ in einem Quadrat der Kantenlänge n gibt es zwei Punkte mit Abstand $\leq \sqrt{2}$.*

Beweis. Zwei Punkte in einem Quadrat der Kantenlänge 1 haben Abstand $\leq \sqrt{2}$ nach dem **Satz von Pythagoras**: $1^2 + 1^2 = (\sqrt{2})^2$. Wir zerlegen

$$Q = \{(x, y) \mid 0 \leq x < n, 0 \leq y < n\}$$

disjunkt in n^2 Quadrate (siehe Abb. 2.11):

$$Q_{ij} = \{(x, y) \mid i - 1 \leq x < i, j - 1 \leq y < j\}$$

und definieren eine Abbildung

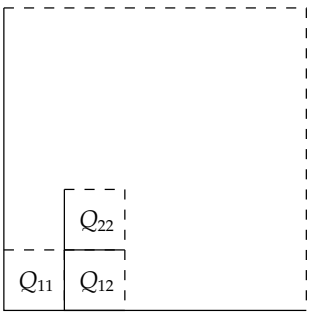


Abbildung 2.11. Zerlegung eines Quadrates.

$$f\colon \{1, \ldots, n^2 + 1\} \to \{1, \ldots, n\} \times \{1, \ldots, n\} = \{(i, j) \in \mathbb{N}^2 \mid i \leq n, j \leq n\},$$

durch $f(k) := (i, j)$, falls der Punkt $P_k \in Q_{ij}$. Da wir eine disjunkte Vereinigung haben, ist dies eine Abbildung. Nach dem Schubfachprinzip ist sie nicht injektiv und die Behauptung folgt. □

2.8.3 Weitere Notationen zu Abbildungen

Einige weitere Notationen zu Abbildungen, die häufig verwendet werden, sind die folgenden:

Definition 2.17. *Sind M und N Mengen, so bezeichnet $N^M = \{f : M \to N\}$ die Menge aller Abbildungen von M nach N.*

Beispiel 2.18. $\{0,1\}^M = \{f : M \to \{0,1\}\}$.

Bemerkung 2.19. Es gilt für endliche Mengen M, N:

$$|N^M| = |N|^{|M|},$$

daher die Notation. In der Tat müssen wir, um f festzulegen, für jedes $a \in M$ ein Bild $f(a)$ auswählen. Hierfür haben wir $|N|$ Wahlmöglichkeiten, insgesamt also $|N|^{|M|}$ Wahlmöglichkeiten.

2.8.4 Komposition von Abbildungen

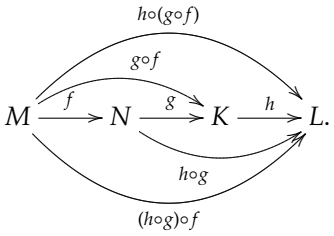
Definition 2.20. *Sind $f : M \to N$ und $g : N \to K$ Abbildungen, so ist die **Komposition** (oder **Hintereinanderausführung**) $g \circ f : M \to K$ (gelesen: g **verknüpft mit** f oder g **nach** f) durch $(g \circ f)(a) := g(f(a))$ definiert.*

Komposition von Abbildungen ist assoziativ:

Satz 2.21. *Für Abbildungen $f : M \to N$, $g : N \to K$ und $h : K \to L$ gilt:*

$$h \circ (g \circ f) = (h \circ g) \circ f,$$

kurz zusammengefasst:



Beweis. Es gilt:

$$\begin{aligned}(h \circ (g \circ f))(a) &= h((g \circ f)(a)) \\ &= h(g(f(a))) \\ &= (h \circ g)(f(a)) \\ &= ((h \circ g) \circ f)(a)\end{aligned}$$

für alle a aus M. Also:

$$h \circ (g \circ f) = (h \circ g) \circ f.$$

□

2.9 Existenz– und All–Quantor

Die Phrasen *für alle* und *es existiert* tauchen in der Mathematik und in der Informatik häufig auf. Wir verwenden daher die Notation $\forall$ für *für alle* und $\exists$ für *es existiert ein*.

Beispiel 2.22.

$$f\colon M \to N \text{ ist surjektiv} \iff \forall b \in N\ \exists\, a \in M \text{ mit } b = f(a).$$

Bei der Negation von Aussagen mit All– und Existenzquantoren verwandelt sich $\forall$ in $\exists$ und $\exists$ in $\forall$ ähnlich wie bei den Gesetzen von de Morgan.

Beispiel 2.23. $f\colon M \to N$ ist nicht surjektiv

$$\begin{aligned}&\iff \neg(\forall\, b \in N\ \exists\, a \in M : f(a) = b) \\ &\iff \exists\, b \in N\ \neg(\exists\, a \in M : f(a) = b) \\ &\iff \exists\, b \in N\ \forall\, a \in M : \neg(f(a) = b) \\ &\iff \exists\, b \in N\ \forall\, a \in M : f(a) \neq b.\end{aligned}$$

2.10 Indizes

Häufig werden Notationen wie $a_1, \ldots, a_{101} \in \mathbb{Z}$ verwendet. Was ist dies formal? $i \mapsto a_i$ ist eine Abbildung wie $f\colon \{1, \ldots, 101\} \to \mathbb{Z}$, $f(i) = a_i$.

Beispiel/Definition 2.24. Sei I eine beliebige Menge und $(A_i)_{i \in I}$ eine **Familie von Teilmengen** $A_i \subset M$ einer weiteren Menge M, d.h. eine Abbildung $I \to 2^M, i \mapsto A_i$. Dann ist der **Durchschnitt**

$$\bigcap_{i \in I} A_i = \{x \in M \mid x \in A_i\ \forall\, i \in I\}$$

und die **Vereinigung**:

$$\bigcup_{i \in I} A_i = \{x \in M \mid \exists\, i \in I \text{ mit } x \in A_i\}.$$

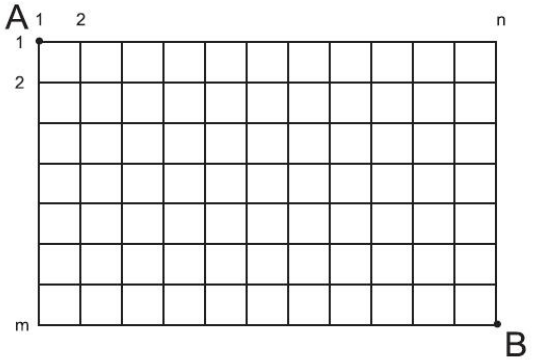
Aufgaben

Aufgabe 2.1 (Schubfachprinzip und Kartesische Produkte). Gegeben seien 101 paarweise verschiedene ganze Zahlen $a_1, \ldots, a_{101} \in \mathbb{Z}$. Zeigen Sie: Es gibt eine Teilfolge $a_{i_1}, a_{i_2}, \ldots, a_{i_{11}}$, $i_1 < \cdots < i_{11}$, von 11 Zahlen, so dass die Folge entweder monoton fallend ($a_{i_1} > \cdots > a_{i_{11}}$) oder monoton steigend ($a_{i_1} < \cdots < a_{i_{11}}$) ist.

Aufgabe 2.2 (Injektivität und Surjektivität). Seien M und N endliche Mengen. Wieviele injektive Abbildungen gibt es von M nach N? Wieviele surjektive Abbildungen gibt es von M nach N, wenn N zwei, drei oder vier Elemente enthält? Haben Sie eine Idee für den allgemeinen Fall $|N| = n \in \mathbb{N}$?

Aufgabe 2.3 (Potenzmengen). Sei M eine beliebige Menge und 2^M ihre Potenzmenge. Zeigen Sie: Es existiert keine bijektive Abbildung zwischen M und 2^M.

Aufgabe 2.4 (Wege durch eine Stadt). In einem amerikanischen Stadtplan mit n Avenues und m Streets, die ein Gitter aus gleich großen Quadraten bilden (siehe Abbildung unten), wollen Sie von einem Eckpunkt A aus zum gegenüberliegendem Eckpunkt B gehen. Wieviele kürzeste Wege gibt es?



Aufgabe 2.5 (Induktion). Zeigen Sie mit vollständiger Induktion, dass für $n \in \mathbb{N}$ mit $n \geq 2$ gilt:

$$\sum_{k=2}^{n} \binom{k}{2} = \binom{n+1}{3}.$$

Aufgabe 2.6 (Binomialkoeffizienten). Zeigen Sie: Für alle $n, k, s, t \in \mathbb{N}$ gelten die folgenden drei Gleichungen:

$$\binom{n}{k+1} = \binom{n}{k}\frac{n-k}{k+1}, \tag{2.1}$$

$$\sum_{i=0}^{n} i \cdot \binom{n}{i} = n \cdot 2^{n-1}, \tag{2.2}$$

$$\binom{s+t}{n} = \sum_{i=0}^{n} \binom{s}{i}\binom{t}{n-i}. \tag{2.3}$$

Geben Sie eine Interpretation von Gleichung (2.1) über die Definition des Binomialkoeffizienten.

3

Äquivalenzrelationen und Kongruenzen

In der Mathematik und Informatik hat man es oft mit Relationen zu tun.

3.1 Äquivalenzrelationen

Beispiel 3.1. $\geq$ (größer gleich) ist eine Relation auf $\mathbb{R}$. Für je zwei Zahlen $x, y \in \mathbb{R}$ ist die Relation $x \geq y$ entweder wahr oder falsch.

Definition 3.2. *Sei M eine Menge. Eine Teilmenge $R \subset M \times M$ nennen wir eine **Relation**. Für $x, y \in M$ ist die Relation erfüllt, wenn $(x, y) \in R$. Manchmal schreibt man dann auch xRy.*

Beispiel 3.3.

1. Für $\geq$ ist $R_{\geq} \subset \mathbb{R} \times \mathbb{R}$ die Winkelhalbierende des rechten oberen und linken unteren Quadranten des Koordinatensystems gemeinsam mit allen Punkten darunter (Abb. 3.1).

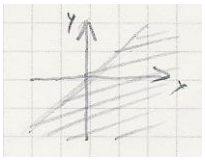


Abbildung 3.1. Die Relation $\geq$ auf $\mathbb{R}^2$.

2. Für Gleichheit $=$ ist $R_{=} \subset \mathbb{R} \times \mathbb{R}$ genau die Winkelhalbierende des rechten oberen und linken unteren Quadranten (Abb. 3.2).

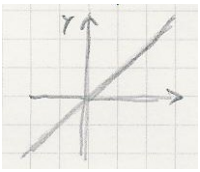


Abbildung 3.2. Die Relation = auf $\mathbb{R}^2$.

In diesem Kapitel möchten wir Äquivalenzrelationen studieren. Unser Ziel ist es, den Begriff *gleich* zu *äquivalent* bzw. *ähnlich* abzuschwächen. Zunächst ein paar Beispiele dazu:

Beispiel 3.4.

1. Sei $f: M \to N$ eine Abbildung. Wir sagen $a, b \in M$ sind **äquivalent**, in Zeichen $a \sim b$, wenn $f(a) = f(b)$ gilt.
2. Seien $M = \mathbb{Z}$ und $d \in \mathbb{Z}_{>1}$. Zwei Zahlen $a, b \in \mathbb{Z}$ heißen **kongruent modulo** d, in Zeichen $a \equiv b \mod d$, wenn $a - b$ durch d teilbar ist.

Welche Eigenschaften sollen Äquivalenzrelationen haben?

Definition 3.5. *Eine Teilmenge $R \subset M \times M$ heißt **Äquivalenzrelation** (wir schreiben $a \sim b$ statt $(a, b) \in R$), wenn folgende Eigenschaften erfüllt sind:*

1. ***Reflexivität**: $a \sim a \; \forall \; a \in M$.*
2. ***Symmetrie**: $a \sim b \Rightarrow b \sim a \; \forall a, b \in M$.*
3. ***Transitivität**: $a \sim b$ und $b \sim c \Rightarrow a \sim c \; \forall a, b, c \in M$.*

Beispiel 3.6.

0. Die Relation = ist eine Äquivalenzrelation
1. Für $f: M \to N$, definert $a \sim b$, falls $f(a) = f(b)$ eine Äquivalenzrelation auf M. Dies ist eine Äquivalenzrelation, da = auf N eine Äquivalenzrelation ist.
2. $a \equiv b \mod d$ ist eine Äquivalenzrelation auf $\mathbb{Z}$.
   a) $a - a = 0 \cdot d$.
   b) $a - b = k \cdot d \Rightarrow b - a = (-k) \cdot d$.
   c) $a - b = k \cdot d$ und $b - c = l \cdot d \Rightarrow a - c = a - b + b - c = k \cdot d + l \cdot d = (k + l) \cdot d$.
3. $\geq$ auf $M = \mathbb{R}$ ist keine Äquivalenzrelation. Zwar gilt die Reflexivität, da $x \geq x \; \forall x \in \mathbb{R}$, und $x \geq y, y \geq z \Rightarrow x \geq z$ (Transitivität), aber $x \geq y \Rightarrow y \geq x$ ist im Allgemeinen falsch.

Satz/Definition 3.7. *Sei $\sim$ eine Äquivalenzrelation auf M. Zu $a \in M$ heißt*

$$[a] := \{b \in M \mid b \sim a\}$$

*die **Äquivalenzklasse** von M und jedes Element $b \in [a]$ heißt ein **Repräsentant** von $[a]$. Je zwei Äquivalenzklassen $[a]$ und $[b]$ sind entweder gleich oder disjunkt (d.h. sie haben leeren Durchschnitt).*

Beispiel 3.8. Für ($\equiv \mod 3$) sind die Äquivalenzklassen

$$\begin{aligned}[0] &= \{\dots, -6, -3, 0, 3, 6, \dots\},\\ [1] &= \{\dots, -5, -2, 1, 4, 7, \dots\},\\ [2] &= \{\dots, -4, -1, 2, 5, 8, \dots\}.\end{aligned}$$

Es gilt:

$$[0] \cup [1] \cup [2] = \mathbb{Z}.$$

Beweis (des Satzes 3.7). Zu zeigen ist, dass aus $[a] \cap [b] \neq \emptyset$ folgt, dass $[a] = [b]$. Sei dazu etwa $c \in [a] \cap [b]$ und $d \in [a]$. Dann gilt: $d \sim a \sim c \sim b$, also wegen der Transitivität $d \sim b$ und daher $d \in [b]$. Dies zeigt: $[a] \subset [b]$. Die umgekehrte Inklusion $[b] \subset [a]$ folgt analog. Also insgesamt: $[a] = [b]$. □

Definition 3.9. *Sei $\sim$ eine Äquivalenzrelation auf M. Mit*

$$M/\!\sim \; = \{[a] \mid a \in M\} \subset 2^M$$

*(Sprechweise: M modulo $\sim$) bezeichnen wir die **Menge der Äquivalenzklassen**. Die Abbildung*

$$\pi \colon M \to M/\!\sim,\ a \mapsto [a]$$

*heißt **kanonische Äquivalenzklassenabbildung**.*

Beispiel 3.10. Für ($\equiv \mod 3$) auf $\mathbb{Z}$ ist $\pi \colon \mathbb{Z} \to \{[0], [1], [2]\}$ die Abbildung $n \mapsto$ [Rest von n bei Division durch 3].

Bemerkung 3.11. Offenbar gilt:

$$\pi(a) = \pi(b) \iff [a] = [b].$$

Jede Äquivalenzrelation ist also im Prinzip vom Typ 1 in Beispiel 3.6 mit $f = \pi \colon M \to N = M/\!\sim$. Das Urbild des Elements $[a] \in M/\!\sim$ ist $\pi^{-1}([a]) = [a] \subset M$.

Einer der wesentlichen Anwendungen von Äquivalenzrelationen ist es, dass man mit ihrer Hilfe aus bekannten Mengen M neue interessante Mengen $M/\!\sim$ konstruieren kann.

Beispiel 3.12. Die Konstruktion der Menge der **rationalen Zahlen** aus den ganzen Zahlen $\mathbb{Z}$. Dazu betrachten wir $M = \mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$ und definieren eine Äquivalenzrelation auf M durch:

$$(p_1, q_1) \sim (p_2, q_2), \text{ falls } p_1 \cdot q_2 = p_2 \cdot q_1.$$

Die Äquivalenzklasse $[(p, q)]$ wird üblicherweise mit $\frac{p}{q}$ bezeichnet, wobei p dann **Zähler** und q **Nenner** heißen. Also:

$$\mathbb{Q} := \Big(\mathbb{Z} \times (\mathbb{Z}\backslash\{0\})\Big) / \sim .$$

Wir müssen einsehen, dass $\sim$ wirklich eine Äquivalenzrelation ist. Reflexivität und Symmetrie sind klar. Transitivität: Seien $(p_1, q_1) \sim (p_2, q_2)$ und $(p_2, q_2) \sim (p_3, q_3)$, also $p_1q_2 = p_2q_1$ und $p_2q_3 = p_3q_2$. Es folgt: $p_1q_2q_3 = p_2q_1q_3 = p_2q_3q_1 = p_3q_2q_1$, also $q_2(p_1q_3 - p_3q_1) = 0$. Nun gilt: $q_2 \neq 0$. In $\mathbb{Z}$ folgt daher: $p_1q_3 - p_3q_1 = 0$, d.h. $(p_1, q_1) \sim (p_3, q_3)$.

Addition und Multiplikation erklären wir repräsentantenweise:

$$\frac{p}{q} + \frac{r}{s} := \frac{ps + qr}{qs} \text{ und } \frac{p}{q} \cdot \frac{r}{s} := \frac{p \cdot r}{q \cdot s}.$$

Diese Definition ist nicht unproblematisch, da die rechten Seiten von der Auswahl der Repräsentanten $(p, q) \in \frac{p}{q}$ und $(r, s) \in \frac{r}{s}$ abhängen. Das Beispiel

$$\frac{1}{2} + \frac{1}{3} = \frac{1 \cdot 3 + 2 \cdot 1}{2 \cdot 3} = \frac{5}{6}$$
$$\frac{2}{4} + \frac{-1}{-3} = \frac{2 \cdot (-3) + 4 \cdot (-1)}{4 \cdot (-3)} = \frac{-10}{-12}$$

suggeriert aber, dass dies vielleicht doch kein Problem ist. Um allgemein einzusehen, dass

$$+ : M/\sim \times M/\sim \to M/\sim \text{ und } \cdot : M/\sim \times M/\sim \to M/\sim$$

wohldefinierte Abbildungen sind, müssen wir zeigen, dass das Ergebnis nicht von der Wahl der Repräsentanten abhängt. Zum Beispiel ist für $(p_1, q_1) \sim (p_2, q_2)$ zu zeigen, dass $(p_1s + q_1r, q_1s) \sim (p_2s + q_2r, q_2s)$ gilt. Also ist

$$(p_1s + q_1r) \cdot q_2s \stackrel{!}{=} (p_2s + q_2r) \cdot q_1s$$

zu zeigen, was äquivalent zu $p_1s \cdot q_2s = p_2s \cdot q_1s$ ist. Dies folgt aber aus $p_1q_2 = p_2q_1$ durch Multiplikation mit s^2. Die Unabhängigkeit von der Auswahl $(r_2, s_2) \in \frac{r}{s}$ zeigt man genauso. Die Wohldefiniertheit der Multiplikation ist ähnlich, aber einfacher.

Schließlich lässt sich $\mathbb{Z}$ als Teilmenge von $\mathbb{Q}$ auffassen mit Hilfe der Abbildung:

$$\mathbb{Z} \hookrightarrow \mathbb{Q},\ n \mapsto \frac{n}{1}.$$

Dabei bezeichnet der Pfeil $\hookrightarrow$ eine injektive Abbildung. Gelegentlich verwenden wir auch den Pfeil $\twoheadrightarrow$; dieser steht für eine surjektive Abbildung. Jedes Element $\frac{p}{q} \in \mathbb{Q}$ hat einen ausgezeichneten Repräsentanten $(a, b) \in \frac{p}{q} = \frac{a}{b}$ mit a, b teilerfremd, $b > 0$.

Bemerkung 3.13. Im Allgemeinen gibt es bei Äquivalenzrelationen keine ausgezeichneten Repräsentanten. Dies sieht man beispielsweise an ähnlichen Dreiecken: Zwei Dreiecke heißen **ähnlich**, wenn die drei Winkel (α, β, γ) und $(\alpha', \beta', \gamma')$ bis auf die Reihenfolge übereinstimmen (Abb. 3.3). Einen ausgezeichneten Repräsentanten gibt es nicht (Abb. 3.4).

Abbildung 3.3. Zwei ähnliche Dreiecke.

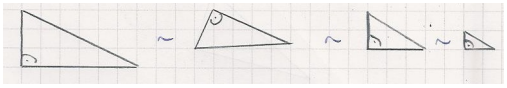


Abbildung 3.4. Ähnliche Dreiecke.

3.2 Kongruenzen

Im Folgenden möchten wir die Relation $(\equiv \mod n)$ auf $\mathbb{Z}$ näher studieren. Für $\mathbb{Z}/(\equiv \mod n)$ schreiben wir kürzer $\mathbb{Z}/n$. Jedes Element $[a]$ von $\mathbb{Z}/n$ hat einen ausgezeichneten Repräsentanten $i \in \{0, 1, \ldots, n-1\}$, nämlich den Rest i bei Division von a durch n. Die Restklasse von i ist

$$[i] = \{i + kn \mid k \in \mathbb{Z}\}.$$

Häufig wird auch die Notation

$$\bar{i} = [i]$$

verwendet. Die Menge

$$\mathbb{Z}/n = \{\bar{0}, \bar{1}, \ldots, \overline{n-1}\}$$

hat also n Elemente. Elemente von $\mathbb{Z}/n$ lassen sich addieren und multiplizieren:

$$\bar{i} + \bar{j} = \overline{i+j},\ \bar{i} \cdot \bar{j} = \overline{i \cdot j}.$$

Beispiel 3.14. $n = 6$.

1. Es gilt: $(\bar{2} + \bar{2}) + \bar{5} = \bar{4} + \bar{5} = \bar{9} = \bar{3}$.

   Außerdem ist: $\bar{2} + (\bar{2} + \bar{5}) = \bar{2} + \bar{7} = \bar{2} + \bar{1} = \bar{3}$.
2. Es gilt: $(\bar{2} \cdot \bar{2}) \cdot \bar{5} = \bar{4} \cdot \bar{5} = \overline{20} = \bar{2}$.

   Außerdem ist: $\bar{2} \cdot (\bar{2} \cdot \bar{5}) = \bar{2} \cdot \overline{10} = \bar{2} \cdot \bar{4} = \bar{8} = \bar{2}$.

Diese Addition und Multiplikation genügt den üblichen Gesetzen von Addition und Multiplikation, z.B. den Assoziativgesetzen:

$$(\bar{i} + \bar{j}) + \bar{k} = \bar{i} + (\bar{j} + \bar{k})$$
$$(\bar{i} \cdot \bar{j}) \cdot \bar{k} = \bar{i} \cdot (\bar{j} \cdot \bar{k}).$$

Distributivgesetze:

$$(\bar{i} + \bar{j}) \cdot \bar{k} = \bar{i} \cdot \bar{k} + \bar{j} \cdot \bar{k}.$$

Kommutativgesetze:

$$\bar{i} + \bar{j} = \bar{j} + \bar{i}$$
$$\bar{i} \cdot \bar{j} = \bar{j} \cdot \bar{i}.$$

Um dies einzusehen, zeigt man am Besten, dass in der Definition

$$\bar{i} + \bar{j} = \overline{i+j}, \quad \bar{i} \cdot \bar{j} = \overline{i \cdot j}$$

das Ergebnis nicht von der Auswahl $i \in \bar{i}$ und $j \in \bar{j}$ abhängt: Sind also zwei verschiedene Repräsentanten der gleichen Äquivalenzklasse i_1 und i_2 bzw. j_1 und j_2 gegeben, d.h.

$$i_1 \equiv i_2 \mod n, \quad j_1 \equiv j_2 \mod n.$$

so folgt tatsächlich:

$$i_1 + j_1 \equiv i_2 + j_2 \mod n, \quad i_1 \cdot j_1 \equiv i_2 \cdot j_2 \mod n.$$

Dann vererben sich nämlich die Rechengesetze direkt aus denen für + und · in $\mathbb{Z}$. Sei also $i_1 \equiv i_2 \mod n$, d.h. $i_1 - i_2 = k \cdot n$ für ein gewisses $k \in \mathbb{Z}$. Daraus folgt:

$$i_1 + j - (i_2 + j) = k \cdot n, \text{ also } i_1 + j \equiv i_2 + j \mod n.$$

Analog:

$$i_1 j - i_2 j = jkn \Rightarrow i_1 j \equiv i_2 j \mod n,$$

d.h. bei der Verknüpfung hängt das Ergebnis nicht von der Wahl eines Repräsentanten der Klasse $\overline{i}$ ab. Eine analoge Rechnung zeigt Entsprechendes für die Klasse $\overline{j}$.

Bemerkung 3.15. In $\mathbb{Z}/6$ gilt:

$$\overline{2} \cdot \overline{3} = \overline{6} = \overline{0} \in \mathbb{Z}/6,$$

obwohl $\overline{2}, \overline{3} \neq \overline{0}$. Aus $a \cdot b = a \cdot c$ kann man also $b = c$ in $\mathbb{Z}/n$ nicht schließen.

Eine Ausnahme bildet der Fall, dass $n = p$ eine Primzahl ist, denn aus

$$a \cdot b \equiv 0 \mod p$$

folgt: $p \mid a \cdot b$ und daher $p \mid a$ oder $p \mid b$ (dies werden wir in Korollar 3.32 beweisen).

Es folgt:

Satz/Definition 3.16. *Sei p eine Primzahl und $\overline{a} \in \mathbb{Z}/p$, $\overline{a} \neq \overline{0}$. Dann gibt die Multiplikation mit $\overline{a}$ eine bijektive Abbildung*

$$\mathbb{Z}/p \to \mathbb{Z}/p, \ \overline{b} \mapsto \overline{a} \cdot \overline{b}.$$

*Das Urbild von $\overline{1}$ bezeichnen wir mit $\overline{a}^{-1}$ und heißt **Inverses** von a. Das Inverse $\overline{a}^{-1}$ wird also durch ein Element $u \in \mathbb{Z}$ repräsentiert, so dass*

$$u \cdot a \equiv 1 \mod p$$

*gilt. u heißt **Inverses** von $a \mod p$.*

Beweis. Ist $\overline{a} \cdot \overline{b_1} = \overline{a} \cdot \overline{b_2}$, so folgt: $\overline{a} \cdot (\overline{b_1} - \overline{b_2}) = \overline{0}$. Somit gilt $p \mid a(b_1 - b_2)$, aber nach Voraussetzung $p \nmid a$ folgt: $p \mid b_1 - b_2$ und damit: $\overline{b_1} = \overline{b_2}$.

Also: $\mathbb{Z}/p \to \mathbb{Z}/p$, $\overline{b} \mapsto \overline{a} \cdot \overline{b}$ ist injektiv und somit bijektiv, da $\mathbb{Z}/p$ endlich ist. D.h. $\exists\, \overline{u}$ mit $\overline{u} \cdot \overline{a} = \overline{1} \iff \exists\, u \in \mathbb{Z}$: $u \cdot a \equiv 1 \mod p$. □

Beispiel 3.17. In $\mathbb{Z}/5$ gilt: $\overline{2} \in \mathbb{Z}/5$ und $\overline{2}^{-1} = \overline{3}$, da $2 \cdot 3 = 6 \equiv 1 \mod 5$.

3.3 Simultanes Lösen von Kongruenzen

Gegeben zwei natürliche Zahlen $n, m \in \mathbb{Z}_{>1}$. Dann haben wir eine Abbildung

$$\mathbb{Z} \to \mathbb{Z}/n \times \mathbb{Z}/m,\ i \mapsto (i \mod n, i \mod m).$$

Wir fragen: Ist die Abbildung surjektiv? Mit anderen Worten: Gibt es für gegebene $a, b \in \mathbb{Z}$ ein $x \in \mathbb{Z}$ mit

$$\begin{aligned} x &\equiv a \mod n, \\ x &\equiv b \mod m\,? \end{aligned}$$

Diese Fragestellung taucht auch bei der Kalenderrechnung auf:

Beispiel 3.18. In wie vielen Tagen fällt der nächste Vollmond auf einen Sonntag?

Vollmond ist alle 30 Tage (in etwa), Sonntag alle 7 Tage. Der nächste Vollmond ist Donnerstag, der 13.11., also in 6 Tagen. Ist x die gesuchte Anzahl (heute ist Freitag, der 8.11.), so gilt:

$$\begin{aligned} x &\equiv 6 \mod 30, \\ x &\equiv 2 \mod 7. \end{aligned}$$

Beispiel 3.19. Zahnräder, die ineinander greifen:

$$\begin{aligned} x &\equiv 2 \mod 10, \\ x &\equiv 5 \mod 12. \end{aligned}$$

Es gibt keine Lösung, denn aus den beiden Kongruenzen folgt:

$$\begin{aligned} x &\equiv 0 \mod 2, \\ x &\equiv 1 \mod 2, \end{aligned}$$

was ein Widerspruch ist.

Wie wir eben gesehen haben, sind simultane Kongruenzen nicht immer lösbar und zwar ist

$$\begin{aligned} x &\equiv a \mod n, \\ x &\equiv b \mod m, \end{aligned}$$

höchstens lösbar, wenn $a \equiv b \mod \mathrm{ggT}(n, m)$ = größter gemeinsamer Teiler von n und m (englisch: greatest common divisor gcd). Dass diese Bedingung hinreichend ist, ist Gegenstand des sogenannten chinesischen Restsatzes.

Wie berechnet man den ggT? Die Schulmethode ist folgende: Seien $n, m \in \mathbb{N}$. Dann schreibt man:

$$n = p_1^{e_1} \cdot \dots \cdot p_r^{e_r} = \prod_{i=1}^{r} p_i^{e_i},$$

und

$$m = p_1^{f_1} \cdot \dots \cdot p_r^{f_r} = \prod_{i=1}^{r} p_i^{f_i},$$

d.h. man ermittelt eine Primfaktorzerlegung von n und m in p_i paarweise verschiedene Primzahlen mit Vielfachheiten $e_i, f_i \in \mathbb{Z}_{\geq 0}$ (einige der e_i bzw. f_i dürfen 0 sein). Dann ist

$$\mathrm{ggT}(n, m) = p_1^{\min(e_1, f_1)} \cdots p_r^{\min(e_r, f_r)} = \prod_{i=1}^{r} p_i^{\min(e_1, f_1)}.$$

Aber: Faktorisieren ist schwierig. Nur in einigen Spezialfällen kann man Primfaktoren leicht erkennen. Jeder kennt aus der Schule die Regel, dass eine Zahl genau dann durch 2 teilbar ist, wenn ihre letzte Ziffer gerade ist. Außerdem ist bekannt, dass eine Zahl genau dann durch 3 teilbar ist, wenn ihre Quersumme es ist. Einige weitere Beispiele:

Beispiel 3.20.

Neuner–Regel: Ist $n \in \mathbb{N}$ gegeben in der Form

$$n = \sum_{i=0}^{r} a_i \cdot 10^i, \; a_i \in \{0, \dots, 9\},$$

also im in der Schule üblichen Zehnersystem, so gilt:

$$n \equiv Q(n) := \sum_{i=0}^{r} a_i \mod 9,$$

da $10 \equiv 1 \mod 9$. $Q(n)$ heißt **Quersumme** von n.

Elfer–Regel: Ist n wie oben gegeben, so ist die **alternierende Quersumme** die Zahl $Q'(n)$ mit:

$$Q'(n) := \sum_{i=0}^{r} (-1)^i a_i.$$

Dann gilt: $n \equiv Q'(n) \mod 11$, da $-1 \equiv 10 \mod 11$. Beispielsweise ist

$$Q'(143) = 3 + (-4) + 1 = 0 \Rightarrow 11 \text{ ist ein Teiler von } 143,$$

genauer gilt: $143 = 11 \cdot 13$, wie wir dann leicht berechnen können.

3.4 Das RSA–Verfahren

Faktorisieren ist so schwierig, dass dieses als Grundlage eines der weitverbreitesten **öffentlichen Verschlüsselungsverfahren** herangezogen wird, nämlich **RSA** (Rivest-Shamir-Adleman, 1978).

3.4.1 Öffentliche Kryptosysteme

Diffie und Hellmann haben das Prinzip sogenannter **öffentlicher Kryptosysteme** beschrieben:

Alice möchte Bob eine Nachricht x schicken. Da es sein könnte, dass Eve mithört, verschlüsselt Alice die Nachricht x in eine codierte Nachricht c. Bob möchte dann wieder aus der codierten Nachricht c in die ursprüngliche Nachricht x zurückgewinnen, ohne dass Eve dies kann. Dies funktioniert nach Diffie–Hellman folgendermaßen:

Die Verschlüsselungsmethode von Bob soll öffentlich zugänglich sein. Alice kann diese anwenden, um die Nachricht zu verschlüsseln. Eve soll auch dann keine Chance zum Entschlüsseln haben, wenn sie das Verschlüsselungsverfahren kennt.

RSA ist eine Umsetzung dieser Idee, die den sogenannten kleinen Satz von Fermat benutzt.

3.4.2 Der kleine Satz von Fermat

Um den kleinen Satz von Fermat, den wir erst im nächsten Semester beweisen, zu formulieren, benötigen wir folgende Definition.

Definition 3.21. *Die **Eulersche φ-Funktion** $\varphi\colon \mathbb{N}_{>1} \to \mathbb{N}$ ist definiert durch:*

$$\varphi(n) = \left|\left\{a \mid 1 \leq a < n \text{ mit } \operatorname{ggT}(a,n) = 1\right\}\right|.$$

Beispielsweise ist $\varphi(6) = 2$, da nur 1 und 5 mit 6 keinen gemeinsamen Teiler besitzen. Mit dieser Notation können wir nun den Satz formulieren:

Satz 3.22 (kleiner Satz von Fermat). *Sei $x \in \mathbb{Z}$ mit $\operatorname{ggT}(x,n) = 1$. Dann gilt:*

$$x^{\varphi(n)} \equiv 1 \mod n.$$

Beweis. später □

Beispiel 3.23. $n = 6$, $\varphi(n) = 2$, $x \in \{1,5\}$. Es gilt: $5^2 = 25 \equiv 1 \mod 6$, wie vom Satz vorausgesagt.

Es gibt auch einen sogenannten (großen) Satz von Fermat, auch **Fermats letzter Satz** genannt, obwohl Fermat ihn wohl nur vermutet hatte und erst Wiles ihn beweisen konnte:

Satz 3.24 (Satz von Wiles (auch: Fermats letzter Satz), Vermutung: Fermat (17. Jhdt.), Beweis: Wiles (1997)). *Für $n \geq 3$ hat die **diophantische Gleichung** (d.h. eine Gleichung, für deren ganzzahlige Lösungen man sich interessiert)*

$$a^n + b^n = c^n$$

keine Lösung $(a, b, c) \in \mathbb{Z}^3$ mit $a \cdot b \cdot c \neq 0$.

3.4.3 Das RSA–Verfahren

Der kleine Satz von Fermat erlaubt es uns nun, das RSA–Verfahren zu erläutern. Bevor Alice eine Nachricht schicken kann, muss Bob seinen öffentlichen und seinen geheimen Schlüssel produzieren:

- Bob wählt zwei große Primzahlen p_B, q_B mit ungefähr 100 Dezimalstellen und berechnet:

  $$n_B = p_B \cdot q_B \text{ und } \varphi(n_B) = (p_B - 1) \cdot (q_B - 1).$$

  Dass $\varphi(n_B) = (p_B - 1) \cdot (q_B - 1)$ gilt, ist nicht besonders schwierig zu zeigen; wir werden auf solche Fragen im nächsten Semester näher eingehen.
- Anschließend wählt Bob eine Zahl d_B mit

  $$\text{ggT}(d_B, \varphi(n_B)) = 1$$

  und berechnet eine Zahl e_B, so dass:

  $$d_B \cdot e_B \equiv 1 \mod \varphi(n_B).$$

- Veröffentlicht werden die Zahlen n_B und d_B, geheim bleiben die Zahlen p_B, q_B, $\varphi(n_B)$ und e_B.

Damit kann Alice nun die Nachricht verschlüsseln und Bob sie wieder entschlüsseln:

- Alice codiert die Nachricht in eine Zahl $x < n_B$ und berechnet daraus die verschlüsselte Zahl c:

  $$c \equiv x^{d_B} \mod n_B.$$

- c wird über den Nachrichtenkanal an Bob gesendet.
- Bob berechnet seinerseits $y \equiv c^{e_B} \mod n_B$.

Dieses y liefert tatsächlich die ursprüngliche Nachricht x, denn es gilt:

$$y \equiv c^{e_B} \equiv (x^{d_B})^{e_B} \equiv x^{d_B \cdot e_B}.$$

Nun ist:

$$d_B \cdot e_B = 1 + k \cdot \varphi(n_B) \text{ und } y \equiv x \cdot (x^{\varphi(n_B)})^k.$$

x und n_B sind mit nahezu 100%-iger Wahrscheinlichkeit teilerfremd:

$$\frac{\varphi(n_B)}{n_B} = \frac{n_B - p_B - q_B + 1}{n_B} \approx 1,$$

da $n_B \gg (p_B + q_B)$ (d.h. n_B ist **wesentlich größer** als $p_B + q_B$, ohne dass wir das Wort wesentlich hier genauer spezifizieren, entsprechend benutzt man $\ll$ für **wesentlich kleiner**). Mit an Sicherheit grenzender Wahrscheinlichkeit lässt sich der kleine Satz von Fermat anwenden, also:

$$y \equiv x \cdot (x^{\varphi(n_B)})^k \equiv x \cdot 1^k \equiv x \mod n_B.$$

Da $x < n_B$ gilt, kennt Bob also die Nachricht x, weil ja $x \equiv y \mod n_B$.

Eve müsste e_B kennen, um die Nachricht zu entschlüsseln, wozu im Wesentlichen die Kenntnis von $\varphi(n_B) = (p_B - 1) \cdot (q_B - 1)$ benötigt wird, was wiederum heißt, dass n_B in $p_B \cdot q_B$ faktorisiert werden muss. Das ist aber, wie schon erwähnt, sehr schwierig.

3.5 Der euklidische Algorithmus

Faktorisieren von Zahlen ist schwierig, wie wir im Abschnitt 3.4 über das RSA–Verfahren erfahren haben. Gemeinsame Teiler zweier gegebener Zahlen sind im Gegensatz dazu aber recht einfach zu finden, wie wir hier sehen werden.

3.5.1 Der Algorithmus

Algorithmus 3.25 (Der erweiterte euklidische Algorithmus).

Input: $a, b \in \mathbb{Z}$.

Output: $d = \text{ggT}(a, b)$ *und* $u, v \in \mathbb{Z}$ *mit* $d = ua + vb$.

u, v *nennt man* ***Bézoutkoeffizienten*** *für den größten gemeinsamen Teiler.*

1. Wir setzen $x_1 = a$, $x_2 = b$.
2. Für $i \geq 2$ mit $x_i \neq 0$ berechnen wir $x_{i+1} = x_{i-1} - c_i x_i$ mit $c_i \in \mathbb{Z}$ und $0 \leq x_{i+1} < |x_i|$, bis schließlich $x_{n+1} = 0$.

3. Dann ist $d = x_n = \text{ggT}(a, b)$.

Der erweiterte Teil des Algorithmus ist folgender:

4. Wir setzen $u_1 = 1, u_2 = 0$ und $v_1 = 0, v_2 = 1$.
5. Für $i = 2, \ldots, n-1$ sei dann

$$\begin{aligned} u_{i+1} &= u_{i-1} - c_i u_i \\ v_{i+1} &= v_{i-1} - c_i v_i. \end{aligned}$$

6. Dann gilt für $u = u_n$ und $v = v_n$:

$$d = ua + vb.$$

Zunächst ein Beispiel dazu:

Beispiel 3.26. $a = 1517$, $b = 1221$. Gemäß des Algorithmus setzen wir: $x_1 = a = 1517$, $x_2 = b = 1221$. Wir erhalten durch Division mit Rest:

$$\begin{aligned} 1517 : 1221 = 1 = c_2, &\text{ Rest } x_3 = 296, \\ 1221 : 296 = 4 = c_3, &\text{ Rest } x_4 = 37, \\ 296 : 37 = 8 = c_4, &\text{ Rest } x_5 = 0. \end{aligned}$$

Damit ist $x_4 = d = \text{ggT}(a, b)$. Für den erweiterten Teil setzen wir $u_1 = 1, u_2 = 0$, $v_1 = 0, v_2 = 1$. Weiter erhalten wir:

$$\begin{aligned} u_3 &= u_1 - c_2 \cdot u_2 = 1 - 1 \cdot 0 = 1, \\ v_3 &= v_1 - c_2 \cdot v_2 = 0 - 1 \cdot 1 = -1, \\ u_4 &= u_2 - c_3 \cdot u_3 = 0 - 4 \cdot 1 = -4, \\ v_4 &= v_2 - c_3 \cdot v_3 = 1 - 4 \cdot (-1) = 5. \end{aligned}$$

Zur Überprüfung der Aussage des Algorithmus berechnen wir:

$$(-4) \cdot 1517 + 5 \cdot 1221 = -6068 + 6105 = 37,$$

wie behauptet.

Bevor wir die Korrektheit des Algorithmus beweisen, führen wir noch eine Notation ein:

Definition 3.27. *Seien $x, y \in \mathbb{Z}$. Dann schreiben wir $x \mid y$, falls x die Zahl y **teilt** (in $\mathbb{Z}$), d.h. falls es ein $k \in \mathbb{Z}$ gibt mit $x \cdot k = y$, und $x \nmid y$, falls nicht.*

Beweis (der Korrektheit von Algorithmus 3.25).

1. Der Algorithmus terminiert, da alle $x_k \in \mathbb{Z}$ und
$$|x_2| > x_3 > \cdots > x_n \geq 0.$$

2. Wir zeigen $x_n \mid x_k$ mit absteigender Induktion nach k. Die Fälle $k = n$, d.h. $x_n \mid x_n$, sowie $k = n+1$, d.h. $x_n \mid x_{n+1}$, sind klar, da $x_{n+1} = 0$. Für den Induktionsschritt seien $x_n \mid x_{k+1}$ und $x_n \mid x_k$ schon bekannt. Es folgt:
$$x_n \mid (x_{k+1} + c_k x_k) \overset{\text{Def.}}{=} x_{k-1}.$$
Also:
$$x_n \mid x_1 = a \text{ und } x_n \mid x_2 = b,$$
d.h. x_n ist ein gemeinsamer Faktor von a und b.

3. Sei e ein Teiler von a und b. Wir zeigen sogar: $e \mid x_k$ für alle k. Wieder ist der Induktionsanfang klar: $e \mid x_1$ und $e \mid x_2$. Für den Induktionsschritt seien $e \mid x_{i-1}$ und $e \mid x_i$ bekannt, woraus folgt:
$$e \mid (x_{i-1} - c_i x_i) = x_{i+1}.$$
Insbesondere gilt also: $e \mid x_n$, d.h. $d = x_n$ ist der größte gemeinsame Teiler von a und b.

4. Um die erweiterte Aussage zu zeigen, beweisen wir $x_i = u_i a + v_i b$ für alle i. Für $i = 1, 2$ ist dies klar wegen $u_1 = 1, v_1 = 0, x_1 = a$ und $u_2 = 0, v_2 = 1, x_2 = b$. Für den Induktionsschritt betrachten wir:
$$\begin{aligned} u_{i+1}a + v_{i+1}b &\overset{\text{Def.}}{=} (u_{i-1} - c_i u_i)a + (v_{i-1} - c_i v_i)b \\ &= (u_{i-1}a + v_{i-1}b) - c_i(u_i a + v_i b) \\ &\overset{\text{I.-V.}}{=} x_{i-1} - c_i x_i \\ &\overset{\text{Def.}}{=} x_{i+1}. \end{aligned}$$
Der Fall $i = n$:
$$d = x_n = u_n a + v_n b = ua + vb,$$
was die letzte Behauptung beweist.

□

Korollar 3.28. *Seien $a, n \in \mathbb{Z}$. Es existiert ein $\overline{u} \in \mathbb{Z}/n$ mit $\overline{u} \cdot \overline{a} = \overline{1} \in \mathbb{Z}/n$ genau dann, wenn* $\mathrm{ggT}(a, n) = 1$. *Ist dies der Fall und sind u, v die Bézoutkoeffizienten in $ua + vn = 1$, dann gilt: $\overline{u} = [u]$.*

Beweis. $\overline{u} \cdot \overline{a} = \overline{1} \iff ua + vn = 1$ für ein gewisses $v \in \mathbb{Z}$. $\Rightarrow$ Jeder Teiler von a und n teilt 1. $\Rightarrow \mathrm{ggT}(a, b) = 1$ und $ua + vn = 1 \Rightarrow \overline{u} \cdot \overline{a} = \overline{1} \in \mathbb{Z}/n$. □

3.5.2 Der chinesische Restsatz

Satz 3.29 (Chinesischer Restsatz). *Es seien $n, m \in \mathbb{Z}_{>1}$ und $a, b \in \mathbb{Z}$. Die simultanen Kongruenzen*

$$\begin{cases} x \equiv a & \mod n, \\ x \equiv b & \mod m \end{cases}$$

haben eine Lösung $x \in \mathbb{Z}$ genau dann, wenn $a \equiv b \mod \mathrm{ggT}(n, m)$. In diesem Fall ist $x \in \mathbb{Z}$ bis auf Vielfache des ***kleinsten gemeinsamen Vielfaches***

$$\mathrm{kgV}(n, m) := \frac{n \cdot m}{\mathrm{ggT}(n, m)}$$

eindeutig bestimmt, d.h. ist $x_0 \in \mathbb{Z}$ eine Lösung, so ist

$$\{x_0 + l \cdot \mathrm{kgV}(n, m) \mid l \in \mathbb{Z}\}$$

die Menge aller Lösungen.

Beweis. Die Notwendigkeit hatten wir bereits gesehen (vor dem RSA–Algorithmus). Um, falls

$$a \equiv b \mod \mathrm{ggT}(n, m) = d$$

gilt, eine Lösung zu konstruieren, berechnen wir d mit dem erweiterten euklidischen Algorithmus:

$$d = un + vm.$$

Ist nun $a = b + kd$ für ein $k \in \mathbb{Z}$, so ist $x = a - kun$ eine Lösung, denn $x \equiv a \mod n$ ist klar und

$$\begin{aligned} x = a - kun &= a - k(d - vm) \\ &= b + kvm \equiv b \mod m. \end{aligned}$$

Ist y eine weitere Lösung, so gilt:

$$x - y \equiv 0 \mod n \text{ und } x - y \equiv 0 \mod m.$$

Aus $n, m \mid x - y$ folgt $\mathrm{kgV}(n, m) \mid x - y$, also:

$$y = x + l \cdot \mathrm{kgV}(n, m) \text{ für ein gewisses } l \in \mathbb{Z},$$

d.h. y ist tatsächlich in der angegebenen Menge aller Lösungen. □

Beispiel 3.30. Ein Himmelskörper sei alle 30 Tage gut zu beobachten, das letzte Mal vor 5 Tagen. Ich habe leider nur sonntags Zeit, heute ist Mittwoch. Wann kann ich das nächste Mal den Himmelskörper sehen?

Wir modellieren dies zunächst mit Hilfe von Kongruenzen:

$$x \equiv 5 = a \mod 30 = n, x \equiv 3 = b \mod 7 = m.$$

Dann berechnen wir mit dem euklidischen Algorithmus gemäß des Beweises des chinesischen Restsatzes u und v mit

$$1 = \text{ggT}(30, 7) = d = un + vm = u \cdot 30 + v \cdot 7.$$

Wir wissen nach dem Satz, dass dies geht, da $a \equiv b \mod 1$ für alle ganzen Zahlen gilt. Es ergibt sich: $u = -3$, $v = 13$. Nun schreiben wir:

$$5 = a = b + kd = 3 + 2 \cdot 1,$$

woraus sich die Lösung

$$x = a - kun = 5 - 2 \cdot (-3) \cdot 30 = 185 \equiv -25 \mod 210$$

ergibt. Also: in 25 Tagen kann ich den Himmelskörper beobachten (das hätten wir übrigens auch zu Fuß recht einfach berechnen können).

Wäre heute aber ein Montag, d.h. $b = 1$, so würden wir folgendes erhalten:

$$5 = a = b + kd = 1 + 4 \cdot 1,$$

also: $x = 5 - 4 \cdot (-3) \cdot 30 = 365 \equiv 155 \equiv -55 \mod 210$. Ich müsste also noch 55 Tage warten.

Beispiel 3.31. Oft kann man mit Hilfe des chinesischen Restsatzes ganze Zahlen schon an wenigen Kongruenzen erkennen. Wir berechnen hier die ganzen Zahlen, die folgende Kongruenzen erfüllen:

$$\begin{aligned} x &\equiv 1 \mod 2, \\ x &\equiv 2 \mod 3, \\ x &\equiv 3 \mod 5, \\ x &\equiv 2 \mod 7. \end{aligned}$$

$2, 3, 5, 7$ haben offenbar paarweise keine gemeinsamen Teiler außer 1. Der chinesische Restsatz sagt uns also, dass eine simultane Lösung für alle vier Gleichungen existiert. Der Beweis des Satzes erklärt auch, wie wir eine solche Lösung finden, nämlich mit Hilfe des erweiterten euklidischen Algorithmus.

Es gilt zunächst einmal:

$$\text{ggT}(2, 3) = 1 = u \cdot 2 + v \cdot 3 \text{ mit } u = -1 \text{ und } v = 1,$$

d.h. wegen $1 = 2 + (-1) \cdot 1$ ist

$$x_{12} = 1 - (-1) \cdot (-1) \cdot 2 = -1$$

eine simultane Lösung der ersten beiden Kongruenzen, also eine Zahl, für die gilt: $x_{12} \equiv -1 \mod 2 \cdot 3 = 6$.

Weiter gilt: $\mathrm{ggT}(6,5) = 1 = u \cdot 6 + v \cdot 5$ mit $u = 1$ und $v = -1$, d.h. wegen $-1 = 3 + (-4) \cdot 1$ ist $x_{123} = -1 - (-4) \cdot 1 \cdot 6 = 23$ eine Lösung der ersten drei Kongruenzen, also $x \equiv 23 \mod 2 \cdot 3 \cdot 5 = 30$.

Zuletzt betrachten wir noch: $\mathrm{ggT}(30,7) = 1 = u \cdot 30 + v \cdot 7$ mit $u = -3$ und $v = 13$, d.h. wegen $23 = 2 + 21 \cdot 1$ ist

$$x_{1234} = 23 - 21 \cdot (-3) \cdot 30 = 1913 \equiv 23 \mod 210$$

eine simultane Lösung aller vier Kongruenzen.

3.5.3 Weitere Folgerungen aus dem euklidischen Algorithmus

Mit Hilfe des euklidischen Algorithmus können wir nun Primzahlen auf andere Weise charakterisieren:

Korollar 3.32. *Sei $p \in \mathbb{Z}_{>1}$. Folgende Aussagen sind äquivalent:*

1. *p ist eine Primzahl.*
2. *$d \mid p, d \in \mathbb{Z}_{>0} \Rightarrow d \in \{1, p\}$.*
3. *$\forall a, b \in \mathbb{Z}$ gilt: $p \mid a \cdot b \Rightarrow p \mid a$ oder $p \mid b$.*

Beweis. Die Äquivalenz der ersten beiden Aussagen ist die Definition einer Primzahl; wir müssen also nur noch die Äquivalenz der letzten beiden Aussagen zeigen:

3. $\Rightarrow$ 2.: Angenommen, 2. ist nicht erfüllt. Dann existiert ein Teiler $a \mid p$ mit $1 < a < p$. Sei also $a \cdot b = p$. Dann gilt $p \mid a \cdot b$, aber $p \nmid a$ und $p \nmid b$, da $1 < a, b < p$. Also: 3. ist nicht erfüllt.

2. $\Rightarrow$ 3.: Sei 2. erfüllt und $p \mid a \cdot b$. Angenommen, $p \nmid a$. Wir müssen dann $p \mid b$ zeigen. Wegen $p \nmid a$ gilt $\mathrm{ggT}(a,p) < p$ und wegen 2. folgt: $\mathrm{ggT}(a,p) = 1$. Nach dem erweiterten euklidischen Algorithmus existieren $u, v \in \mathbb{Z}$ mit $1 = ua + vp$. Damit folgt:

$$p \mid ab \Rightarrow p \mid uab \Rightarrow p \mid (uab + vpb) = (ua + vp) \cdot b = b,$$

also: $p \mid b$.

□

Wir können jetzt zeigen, dass jede ganze Zahl in Primfaktoren zerlegbar ist:

Satz 3.33 (Fundamentalsatz der Arithmetik). *Sei $0 \neq n \in \mathbb{Z}$. Dann existieren $\varepsilon \in \{1, -1\}$, $r \in \mathbb{N}$ und nicht notwendig verschiedene Primzahlen $p_1, \ldots, p_r$, so dass*

$$n = \varepsilon p_1 \cdots p_r = \varepsilon \prod_{i=1}^{r} p_i.$$

Die Darstellung ist eindeutig bis auf die Reihenfolge der Faktoren.

Beweis. Existenz: Induktion nach $|n|$. Ohne Einschränkung (oft *o.E.* abgekürzt) sei $n > 0$, also $\varepsilon = 1$. Ist $n = 1$, so ist $\varepsilon = 1$ und $r = 0$. Ist n eine Primzahl, dann können wir $r = 1$, $p_1 = n$ wählen.

Andernfalls existieren Faktoren $a, b \in \mathbb{Z}_{>1}$ mit $a \cdot b = n$. Wegen $|a|, |b| < |n|$ existieren für a und b Primfaktorzerlegungen nach der Induktionsvoraussetzung, etwa

$$a = p_1 \cdots p_{r_a}, \quad b = q_1 \cdots q_{r_b}.$$

Dann ist $r = r_a + r_b$ und

$$n = a \cdot b = p_1 \cdots p_{r_a} \cdot q_1 \cdots q_{r_b}.$$

Eindeutigkeit: Ohne Einschränkung sei $n > 0$. Angenommen,

$$n = p_1 \cdots p_r = q_1 \cdots q_s$$

mit p_i, q_j Primzahlen. Wir müssen zeigen, dass dies bis auf Reihenfolge die gleiche Faktorisierung ist. Insbesondere müssen wir $r = s$ zeigen. Wieder verwenden wir Induktion, und zwar nach r. Wegen

$$p_r \mid n = q_1 \cdots q_s$$

gilt $p_r \mid q_k$ für ein gewisses k nach Eigenschaft 3. von Korollar 3.32 über die Charakterisierung von Primzahlen. Da q_k eine Primzahl ist, folgt $p_r = q_k$ und nach Umnummerierung der q_j dürfen wir $k = s$ annehmen. Wir haben also

$$p_1 \cdots p_{r-1} \cdot p_r = q_1 \cdots q_{s-1} \cdot p_r.$$

Es folgt:

$$p_1 \cdots p_{r-1} = q_1 \cdots q_{s-1}.$$

Dies ist ein Produkt aus weniger Faktoren, so dass nach der Induktionsvoraussetzung folgt:

$$r - 1 = s - 1, \quad \text{d.h. } r = s$$

und

$$p_1 \cdots p_{r-1} = q_1 \cdots q_{r-1},$$

bis auf Reihenfolge der Faktoren. □

Wir haben weiter oben das kleinste gemeinsame Vielfache schon kennen gelernt; mit dem obigen Satz können wir es nun formal einführen:

Definition 3.34. *Seien $a, b \in \mathbb{Z}_{>0}$. Dann bezeichnet* $\mathrm{kgV}(a, b)$ *das* ***kleinste gemeinsame Vielfache*** *(englisch: lowest common multiple* $\mathrm{lcm}(a, b)$*) von a und b. Das* kgV *existiert und es gilt:*

$$\mathrm{kgV}(a, b) = \frac{a \cdot b}{\mathrm{ggT}(a, b)}.$$

Aufgaben

Aufgabe 3.1 (Äquivalenzrelationen). Auf $M = \mathbb{N} \times \mathbb{N}$ definieren wir eine Relation $\sim$ durch

$$(a, b) \sim (c, d) \Longleftrightarrow a + d = b + c$$

1. Zeigen Sie, dass $\sim$ eine Äquivalenzrelation auf M ist.
2. Beschreiben Sie die Äquivalenzklassen [(1,1)] und [(3,1)].
3. Wir definieren eine Addition auf $M/\sim$ durch komponentenweise Addition, d.h.:
$$[(a, b)] + [(c, d)] = [(a + c, b + d)].$$
Zeigen Sie die Wohldefiniertheit, d.h. zeigen Sie, dass für $(a, b) \sim (a', b')$ und $(c, d) \sim (c', d')$ auch $(a + c, b + d) \sim (a' + c', b' + d')$ gilt.
4. Die Menge $M/\sim$ mit der so definierten Addition ist eine in der Mathematik wohlbekannte Menge. Welchen Namen hat diese Menge?

Aufgabe 3.2 (Der chinesische Schäfer). Ein chinesischer Schäfer hat ein Herde von höchstens 200 Tieren. Um sie exakt zu zählen, lässt er sie des Abends immer zu zweit durch ein Gatter laufen und stellt fest, dass ein Tier übrig bleibt. Am nächsten Abend lässt er die Tiere immer zu dritt durchs Gatter laufen und stellt ebenfalls fest, dass eins übrigbleibt. Am dritten Tage macht er dasselbe mit 5 Schafen und stellt wieder fest, dass eines übrig bleibt. Am vierten Abend schließlich lässt er 7 Schafe auf einmal durchs Gatter, und es bleibt kein Schaf übrig. Wie groß ist die Herde?

Aufgabe 3.3 (Größter gemeinsame Teiler).

1. Sei $a = 2387$ und $b = 2079$. Bestimmen Sie ohne Computer den größten gemeinsamen Teiler $d = \mathrm{ggT}(a, b)$ und die Bézoutkoeffizienten u und v, d.h. finden Sie u und v mit $au + bv = d$.
2. Sei $a = 139651$ und $b = 111649$. Bestimmen Sie ohne Computer das kleinste gemeinsame Vielfache von a und b.

Aufgabe 3.4 (Gemeinsame Teiler). Verwenden Sie das Computeralgebrasystem MAPLE, um folgendes Experiment durchzuführen: Wählen Sie zufällig 10.000 Paare von Zahlen zwischen 0 und 10^6 und zählen Sie, wie viele Paare einen gemeinsamen Teiler ungleich eins haben. Wiederholen Sie das Experiment für jeweils 10.000 Paare zwischen 0 und 10^9 bzw. 0 und 10^{12}. Geben Sie einen Ausdruck Ihres Maple-Programmes mit ab.

Aufgabe 3.5 (Inverse in $\mathbb{Z}/n$).

1. Zeigen Sie: $\overline{a} \in \mathbb{Z}/n$ hat genau dann ein multiplikatives Inverses $\overline{u} \in \mathbb{Z}/n$, wenn a und n keinen gemeinsamen Teiler haben, d.h. wenn $\mathrm{ggT}(a, n) = 1$ ist.
2. Zeigen Sie: Dieses Inverse $\overline{u}$ von $\overline{a}$ ist dann in $\mathbb{Z}/n$ eindeutig bestimmt.

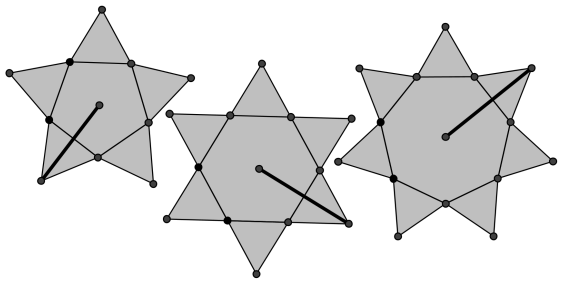
Aufgabe 3.6 (Zur Qualität von Primzahltests mit Hilfe des kleinen Satzes von Fermat). Nach dem kleinen Satz von Fermat gilt: p ist Primzahl $\Longrightarrow$ $a^{p-1} \equiv 1 \mod p \ \forall a$ mit $\mathrm{ggT}(a, p) = 1$.

Verwenden Sie MAPLE, um zu zeigen, dass die Umkehrung nicht gilt, d.h. finden Sie eine zusammengesetzte Zahl n, für die $a^{n-1} \equiv 1 \mod n$ für alle a mit $\mathrm{ggT}(a, n) = 1$ gilt.

Aufgabe 3.7 (Kongruenzen). Sei p eine Primzahl und seien $a, b \in \mathbb{N}$. Zeigen Sie:

$$(a + b)^p = a^p + b^p \mod p.$$

Aufgabe 3.8 (Zahnräder). In der unten stehenden Skizze sehen Sie drei Zahnräder, die ineinander greifen und die sich um ihren jeweiligen Mittelpunkt drehen lassen. Gibt es eine Einstellung der Zahnräder, so dass alle drei Zeiger (d.h. die dicken Striche) nach oben zeigen? Falls ja, geben Sie an, um wieviele Zacken man das linke Rad in welche Richtung drehen muss, damit alle drei Zeiger nach oben zeigen?



Teil II

Analysis in einer Veränderlichen

Einführung

4

Die reellen Zahlen

Mit $\mathbb{R}$ bezeichnen wir die Menge der (unendlichen) Dezimalzahlen, der reellen Zahlen. Wir fassen die Eigenschaften von $\mathbb{R}$ in einigen Axiomen zusammen, auf die wir alle weiteren Sätze aufbauen werden.

4.1 Die Körperaxiome

Auf $\mathbb{R}$ sind zwei Verknüpfungen $+, \cdot$ erklärt:

$$\begin{aligned} +\colon\ & \mathbb{R} \times \mathbb{R} \to \mathbb{R},\ (a,b) \mapsto a+b, \\ \cdot\colon\ & \mathbb{R} \times \mathbb{R} \to \mathbb{R},\ (a,b) \mapsto a \cdot b. \end{aligned}$$

$\mathbb{R}$ ist ein Körper im Sinne der folgenden Definition:

Definition 4.1. *Eine Menge K zusammen mit zwei Abbildungen*

$$\begin{aligned} +\colon\ & K \times K \to K,\ (a,b) \mapsto a+b, \\ \cdot\colon\ & K \times K \to K,\ (a,b) \mapsto a \cdot b. \end{aligned}$$

*heißt ein **Körper** (Achtung! englisch: **field**), wenn folgende Axiome erfüllt sind:*

K1 (Axiome der Addition):

K1.1: Die Addition ist assoziativ:

$$(a+b)+c = a+(b+c) \quad \forall\, a,b,c \in K.$$

K1.2: Die Addition ist kommutativ:

$$a+b = b+a \quad \forall\, a,b \in K.$$

K1.3: Existenz der 0 *(**neutrales Element der Addition**):*

$$\exists\, 0 \in K : a + 0 = a \quad \forall\, a \in K.$$

K1.4: Existenz des Negativen:

$$\forall a \in K\ \exists\, a' \in K \text{ mit } a' + a = 0.$$

*a' heißt **Negatives** zu a und wird üblicherweise mit $-a$ bezeichnet.*

K2 (Axiome der Multiplikation):

K2.1: Die Multiplikation ist assoziativ:

$$(a \cdot b) \cdot c = a \cdot (b \cdot c)\ \forall a, b, c \in K.$$

K2.2: Die Multiplikation ist kommutativ:

$$a \cdot b = b \cdot a \quad \forall\, a, b \in K.$$

K2.3: Existenz der 1 *(**neutrales Element der Multiplikation**):*

$$\exists\, 1 \in K : 1 \cdot a = a \quad \forall\, a \in K.$$

K2.4: Existenz des Inversen:

$$\forall a \in K \backslash \{0\}\ \exists\, a' \in K \text{ mit } a' \cdot a = 1.$$

*a' heißt **Inverses** zu a und wird üblicherweise mit a^{-1} oder $\frac{1}{a}$ bezeichnet.*

K3 (Distributivgesetze): Man kann ausmultiplizieren:

$$a \cdot (b + c) = a \cdot b + a \cdot c \quad \forall a, b, c \in K,$$

und

$$(a + b) \cdot c = a \cdot c + b \cdot c \quad \forall a, b, c \in K.$$

Außer den reellen Zahlen kennen wir schon einige andere Körper:

Beispiel 4.2.

1. $\mathbb{Q}$ ist ein Körper.
2. Ist p eine Primzahl, so ist $\mathbb{F}_p := (\mathbb{Z}/p, +, \cdot)$ ein Körper.
3. Insbesondere ist $\mathbb{F}_2$ ein Körper. Addition und Multiplikation kann man durch folgende **Verknüpfungstafeln** angeben:

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

4. $(\mathbb{Z}, +, \cdot)$ ist kein Körper, da das Axiome K2.4 nicht erfüllt ist, d.h. es existiert nicht für alle ganzen Zahlen ein Inverses in $\mathbb{Z}$, beispielsweise für $2 \in \mathbb{Z}$. Genauer existiert hier nur für $1, -1$ ein Inverses.

4.2 Ringe

Definition 4.3. *Eine Menge $(R, +, \cdot)$ mit zwei Verknüpfungen, für die alle Körperaxiome bis auf K2.4 gefordert werden, heißt ein kommutativer **Ring** mit 1. Wir werden meißt nur Ring dazu sagen, da bei uns allgemeine Ringe selten eine Rolle spielen werden. Ein allgemeiner Ring ist eine Menge, bei der außer K2.4 auch K2.2 und K2.3 nicht gefordert werden.*

Beispiel 4.4.

1. Jeder Körper ist ein Ring.
2. $(\mathbb{Z}, +, \cdot)$ ist ein Ring.
3. $(\mathbb{Z}/n, +, \cdot)$ ist ein Ring für jedes $n \in \mathbb{Z}_{>1}$. Dieses Beispiel zeigt, dass in Ringen aus $a \cdot b = 0$ nicht unbedingt folgt, dass $a = 0$ oder $b = 0$ gilt.

4.3 Folgerungen aus den Körperaxiomen

Aus den Körperaxiomen können wir recht schnell einige Folgerungen ziehen, die uns im Fall der reellen Zahlen geläufig sind:

Proposition 4.5 (Eigenschaften von Körpern). *Sei $(K, +, \cdot)$ ein Körper. Dann gilt:*

1. *Die neutralen Elemente $0, 1 \in K$ sind eindeutig bestimmt.*
2. *Das Negative $-a$ zu $a \in K$ und das Inverse a^{-1} zu $a \in K \backslash \{0\}$ sind eindeutig bestimmt. Wir schreiben $a - b$ für $a + (-b)$ und $\frac{a}{b}$ für $a \cdot b^{-1}$.*
3. $(-1) \cdot (-1) = 1$, $0 \cdot a = 0$ $\forall a \in K$.
4. *In Summen und Produkten kommt es nicht auf Klammern und Reihenfolge an, denn $a_1 + \cdots + a_n$ und $a_1 \cdots a_n$ haben stets den gleichen Wert, unabhängig davon, wie wir Klammern setzen und die Reihenfolge wählen.*

Beweis.

1. Sei $0' \in K$ ein weiteres Nullelement. Dann gilt: $0' + a = a$ $\forall a \in K$, insbesondere:
$$0 = 0' + 0 \overset{\text{K1.2}}{=} 0 + 0' \overset{\text{K1.3}}{=} 0'.$$
Genauso:
$$1 = 1' \cdot 1 = 1 \cdot 1' = 1'.$$

2. Sei a' ein weiteres Negatives zu a. Dann gilt also: $a' + a = 0$ und daher:

$$\begin{aligned} -a &= 0 + (-a) = (a' + a) + (-a) \\ &= a' + (a + (-a)) = a' + ((-a) + a) \\ &= a' + 0 = 0 + a' = a'. \end{aligned}$$

   Für Inverse a^{-1} argumentieren wir analog.

3. Es gilt:

$$0 \cdot a = (0 + 0) \cdot a = 0 \cdot a + 0 \cdot a.$$

   Daraus folgt:

$$\begin{aligned} 0 &= -(0 \cdot a) + 0 \cdot a = -0 \cdot a + (0 \cdot a + 0 \cdot a) \\ &= (-(0 \cdot a) + 0 \cdot a) + 0 \cdot a = 0 + 0 \cdot a \\ &= 0 \cdot a. \end{aligned}$$

   Für die andere Aussage gehen wir ähnlich vor:

$$\begin{aligned} 0 &= 0 \cdot (-1) = ((-1) + 1) \cdot (-1) \\ &= (-1) \cdot (-1) + 1 \cdot (-1) \\ &= (-1) \cdot (-1) + (-1). \end{aligned}$$

   Dies zeigt: $(-1) \cdot (-1) = 1$, wegen der Eindeutigkeit des Negativen.

4. Wir betrachten die Addition und führen Induktion nach n durch. Für $n = 2$ ist nichts zu zeigen. Der Fall $n = 3$, d.h.

$$(a_1 + a_2) + a_3 = a_1 + (a_2 + a_3)$$

   ist das Assoziativgesetz K1.1.

   Für den Induktionsschritt $n - 1 \to n$ nehmen wir nun an, dass $n \geq 4$ und dass die Behauptung für kleinere n schon gezeigt ist. Sei

$$(a_1 + \cdots + a_k) + (a_{k+1} + \cdots + a_n)$$

   die äußerste Klammerung. Wir zeigen:

$$(a_1 + \cdots + a_k) + (a_{k+1} + \cdots + a_n) = a_1 + (a_2 + (\cdots + (a_{n-1} + a_n) \cdots)).$$

   Im Fall $k = 1$ ist dies klar mit der Induktionsvoraussetzung. Sei also $k \geq 2$. Dann gilt mit der Induktionsvoraussetzung:

$$a_1 + \cdots + a_k = a_1 + (a_2 + \cdots + a_k),$$

   also:

$$\begin{aligned}(a_1 + \cdots + a_k) + (a_{k+1} + \cdots + a_n) &= (a_1 + (a_2 + \cdots + a_k)) + (a_{k+1} + \cdots + a_n) \\ &\overset{\text{K1.1 (Ass.)}}{=} a_1 + ((a_2 + \cdots + a_k) + (a_{k+1} + \cdots + a_n)) \\ &\overset{\text{I.–V.}}{=} a_1 + (a_2 + (a_3 + \cdots + a_n)).\end{aligned}$$

Für die Multiplikation argumentiert man genauso.

Schließlich noch zur Reihenfolge: Wir wissen aus dem schon Bewiesenen und dem Kommutativgesetz, dass

$$a_1 + a_2 + \cdots + a_n = a_2 + a_1 + a_3 + \cdots a_n.$$

Da wir jede Permutation der Summanden von $a_1 + \cdots + a_n$ durch wiederholtes Vertauschen benachbarter Summanden erhalten können, folgt die Behauptung.

□

Da endliche Körper in der Informatik von besonderer Bedeutung sind, geben wir noch ein Beispiel dazu:

Beispiel 4.6. Der Körper $\mathbb{F}_3$ mit genau drei Elementen, oft bezeichnet mit $0, 1, -1$, d.h. $\mathbb{F}_3 = \{0, 1, -1\}$, hat folgende Verknüpfungstafeln:

$+$	0	1	-1
0	0	1	-1
1	1	-1	0
-1	-1	0	1

$\cdot$	0	1	-1
0	0	0	0
1	0	1	-1
-1	0	-1	1

Die Verknüpfungstabellen von $\mathbb{F}_2$ und $\mathbb{Z}/2$ sind identisch und jene von $\mathbb{F}_3$ und $\mathbb{Z}/3$ gehen durch $-1 \mapsto 2$ ineinander über, d.h. die beiden Körper haben die gleiche Struktur. Allgemein haben wir schon gesehen, dass $\mathbb{Z}/p$ genau dann ein Körper ist, wenn p eine Primzahl ist.

Gibt es einen Körper mit genau 4 Elementen? Wie wir eben bemerkt haben, kann $\mathbb{Z}/4$ kein Körper sein, weil 4 keine Primzahl ist. Die Frage wird in den Übungsaufgaben beantwortet werden.

4.4 Die Anordnungsaxiome

Auf $\mathbb{R}$ sind gewisse Elemente mit $x > 0$ ausgezeichnet.

Definition 4.7. *Ein **angeordneter Körper** ist ein Körper K zusammen mit Teilmengen positiver Elemente*

$$\{x \in K \mid x > 0\},$$

so dass folgende Axiome erfüllt sind:

A1: Für jedes Element $x \in K$ ist genau eine der Eigenschaften $x > 0$, $x = 0$ oder $-x > 0$ erfüllt.

A2: Sind $x > 0$ und $y > 0$, so auch $x + y > 0$.

A3: Sind $x > 0$ und $y > 0$, so auch $x \cdot y > 0$.

Bemerkung 4.8.

1. Ist K ein angeordneter Körper, dann ist $\mathbb{Q} \hookrightarrow K$ auf natürliche Weise, wir können die rationalen Zahlen also als Teilmenge jedes angeordneten Körpers ansehen.

   Dies können wir wie folgt beweisen. Wir betrachten zunächst die Abbildung:

   $$\mathbb{N} \to K,\ n \mapsto \sum_{i=1}^{n} 1_K = n \cdot 1_K,$$

   wobei 1_K das 1–Element im Körper K bezeichnet. Alle Bilder dieser Abbildung sind positiv, denn nach A1–A3 ist:

   $$1_K = 1_K \cdot 1_K = (-1_K) \cdot (-1_K) > 0, \text{ da } 1_K \neq 0 \text{ und } 1_K > 0 \text{ oder } -1_K > 0.$$

   Die Abbildung ist injektiv, da, falls $n \cdot 1_K = m \cdot 1_K, n \geq m$, folgt: $(n - m) \cdot 1_K = 0$ und dies ist nur für $n = m$ möglich. Endliche Körper lassen sich daher nach dem Schubfachprinzip nicht anordnen ($\mathbb{N} \to \mathbb{F}$ ist für einen endlichen Körper $\mathbb{F}$ (z.B. für $\mathbb{F} = \mathbb{F}_p$) nicht injektiv). Also: $\mathbb{N} \hookrightarrow K$.

   $\mathbb{Z} \hookrightarrow K$ ist dann durch $-n \mapsto n \cdot (-1_K)$ für $n > 0$ definiert. Schließlich definieren wir

   $$\mathbb{Q} \hookrightarrow K,\ \frac{a}{b} \mapsto \frac{a \cdot 1_K}{b \cdot 1_K}.$$

2. Für endliche Körper $\mathbb{F}_q = \mathbb{F}$ mit q Elementen ist das kleinste Element $p \in \mathbb{N}$ mit $p \cdot 1_{\mathbb{F}} = 0$ eine Primzahl.

   Wäre nämlich $a \cdot b \cdot 1_K = 0$ mit $a \cdot 1_K \neq 0$ und $b \cdot 1_K \neq 0$, so ergäbe $a \cdot b \cdot 1_K = (a \cdot 1_K) \cdot (b \cdot 1_K) = 0$ einen Widerspruch: Produkte von Elementen ungleich Null sind in einem Körper stets ungleich Null, denn:

   $$a \cdot b = 0 \in K, a \neq 0 \Rightarrow b = 1 \cdot b = (a^{-1} \cdot a) \cdot b = a^{-1} \cdot (a \cdot b) = a^{-1} \cdot 0 = 0,$$

   d.h. $b = 0$.

   Diese Primzahl heißt **Charakteristik** von $\mathbb{F}_q$, notiert: $\text{char}(\mathbb{F}_q) = p$. Ist $\mathbb{Q} \subset K$, so schreiben wir $\text{char}(K) = 0$.

Bemerkung/Definition 4.9. Ist K ein angeordneter Körper, so definieren wir eine Relation $>$ auf K durch

$$x > y :\iff x - y > 0.$$

Die Relation $<$ ist definiert durch: $x < y :\iff y > x$ und die Relation $\geq$ durch:

$$x \geq y :\iff x = y \text{ oder } x > y.$$

$\geq$ ist eine reflexive und transitive Relation: Offenbar ist nämlich $x \geq x \; \forall x \in K$ und ferner folgt aus $x \geq y$ und $y \geq z$, dass $x \geq z$, da:

$$x - y \geq 0, y - z \geq 0 \Rightarrow x - y + y - z = x - z \geq 0.$$

Durch

$$|x| := \begin{cases} x, & \text{falls } x \geq 0, \\ -x, & \text{falls } -x \geq 0, \end{cases}$$

ist der **Betrag** von x definiert. Es gilt:

1. $|x| \geq 0$.
2. $|x| = 0$ genau dann, wenn $x = 0$.
3. $|x \cdot y| = |x| \cdot |y| \; \forall x, y \in K$.
4. Δ–Ungleichung:
$$|x + y| \leq |x| + |y|.$$

Die ersten Aussagen sind klar oder einfach zu zeigen, die letzte zeigt man, indem man alle Fälle von Vorzeichen durchspielt: $x > 0, x = 0, -x > 0$, $y > 0, y = 0, -y > 0$.

Definition 4.10. *Ein **archimedisch angeordneter Körper** ist ein angeordneter Körper K, der zusätzlich das Axiom*

A4: $\forall x \in K \; \exists \, n \in \mathbb{N}$ *mit* $n = n \cdot 1_K > x$.

erfüllt.

$\mathbb{R}$ und $\mathbb{Q}$ sind Beispiele für archimedisch angeordnete Körper. Ein Beispiel für einen nicht archimedisch angeordneten Körper können wir hier noch nicht geben. In einem nicht archimedisch angeordneten Körper gibt es Elemente $x \in K$, die größer als jedes $n \in \mathbb{N}$, also in gewissem Sinne *unendlich groß* sind.

4.5 Irrationale Zahlen

Das letzte Axiom, das wir für die Charakterisierung von $\mathbb{R}$ benötigen, ist das sogenannte **Vollständigkeitsaxiom**. Bevor wir dieses besprechen, sei daran erinnert, warum wir mit $\mathbb{Q}$ nicht zufrieden waren.

Beispiel 4.11. 1. Nach dem Satz des Pythagoras ist die Länge der Diagonalen c in einem Quadrat der Kantenlänge $a = b = 1$ wegen $a^2 + b^2 = c^2$ gerade

$$c = \sqrt{2}.$$

Es gilt:

$$\sqrt{2} \notin \mathbb{Q}.$$

Man sagt auch, dass $\sqrt{2}$ **irrational** ist. Nehmen wir nämlich an, dass $\sqrt{2} = \frac{p}{q}$ mit $p, q \in \mathbb{Z}$ und ggT$(p, q) = 1$, dann folgt:

$$2 = \frac{p^2}{q^2} \iff 2q^2 = p^2 \Rightarrow p \text{ ist gerade },$$

da p^2 von 2 geteilt wird. Also:

$$p = 2k \Rightarrow 2q^2 = (2k)^2 = 4k^2 \Rightarrow q^2 = 2k^2 \Rightarrow q \text{ ist gerade }.$$

2 ist daher ein gemeinsamer Teiler von p und q, was aber ein Widerspruch zur Voraussetzung ggT$(p, q) = 1$ ist. Daher muss die Annahme falsch gewesen sein.

2. Wir wollen allgemein zwei Strecken a, b vergleichen. Wir sagen, dass a und b **kommensurabel** (lat. zusammen messbar) sind, falls es ganze Zahlen $k, l \in \mathbb{Z}$ gibt, mit:

$$a = k \cdot d, \quad b = l \cdot d,$$

wobei d eine gemeinsame Teilstrecke ist (Abb. 4.1). Zwei Strecken heißen

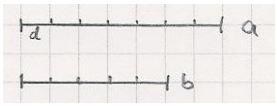


Abbildung 4.1. Kommensurabilität am Beispiel zweier Strecken a und b, die beide von einer Teilstrecke d geteilt werden. Hier: $a = 7 \cdot d$, $b = 5 \cdot d$, d.h. $\frac{a}{b} = \frac{7}{5}$.

inkommensurabel, wenn sie kein solches gemeinsames Teilstück haben. Wenn es aber eines gibt, dann können wir es finden, indem wir analog zum euklidischen Algorithmus die kürzere Strecke von der Größeren abtragen und den Rest nehmen und mit der kleineren und dem Rest fortfahren.

Die Pythagoräer haben die Existenz von inkommensurablen Strecken entdeckt: sie konnten beweisen, dass die Seite und die Diagonalen in einem regelmäßigen Fünfeck inkommensurabel sind. Das eben beschriebene Verfahren des fortwährenden Wegnehmens des Restes bricht nämlich nicht ab.

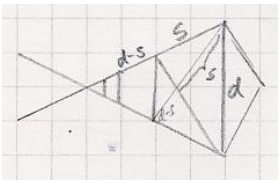


Abbildung 4.2. Die Inkommensurabilität am regelmäßigen Fünfeck: das Verfahren des fortwährenden Wegnehmens des Restes bricht nicht ab.

Aufgaben

Aufgabe 4.1 (Endliche Körper).

1. Gibt es einen Körper mit genau 4 Elementen? Falls ja, so geben Sie die Verknüpfungstafeln an. Beweisen Sie Ihre Antwort.
2. Gibt es einen Körper mit genau 6 Elementen? Falls ja, so geben Sie die Verknüpfungstafeln an. Beweisen Sie Ihre Antwort.

Aufgabe 4.2 (Karatsubas Algorithmus). Der klassische Multiplikationsalgorithmus für zwei Zahlen a und b mit höchstens $n = 2^k$ Bits benötigt $O(n^2)$ viele Schritte. Karatsubas Algorithmus funktioniert wie folgt:

- **Input:** Zwei Zahlen a und b mit höchstens $n = 2^k$ Bits.
- Zerlege $a = a_1 + a_2 2^{\frac{n}{2}}$ und $b = b_1 + b_2 2^{\frac{n}{2}}$.
- Berechne rekursiv $c_1 = a_1 b_1$, $c_3 = a_2 b_2$, $c_2 = (a_1 + a_2)(b_1 + b_2) - c_1 - c_3$.
- Gebe als Ergebnis $c = c_1 + c_2 2^{\frac{n}{2}} + c_3 2^n$ zurück.

Zeigen Sie: Die Laufzeit des Algorithmus ist in $O(n^{\log_2 3}) \subset O(n^{1,59})$.

Aufgabe 4.3 (Irrationale Zahlen). Zeigen Sie:

1. $\sqrt{3}$, $\sqrt{15}$, $\sqrt{45}$ sind irrational,
2. $\sqrt[3]{2}$ ist irrational,
3. $\sqrt{p}$ ist irrational für jede Primzahl p.

5

Konvergenz

Konvergenz ist die zentrale Idee der Analysis.

5.1 Folgen

Definition 5.1. *Eine **Folge** (a_n) $(= (a_n)_{n\in\mathbb{N}})$ reeller Zahlen ist eine Abbildung*

$$\mathbb{N} \to \mathbb{R},\ n \mapsto a_n.$$

*Üblicherweise bekommt die Abbildung keinen Namen, sondern man verwendet Indizes: a_n heißt n-tes **Folgenglied**.*

Beispiel 5.2.

1. $(a_n) = (\frac{1}{n})$, d.h. $a_n = \frac{1}{n}$:
$$\Big(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\Big).$$
2. $(a_n) = (2^n)$:
$$2, 4, 8, 16, \dots.$$

Beispiele von Folgen kennen wir aus Intelligenztests. Dort besteht die Aufgabe oft darin, das nächste Glied einer begonnenen Folge anzugeben, z.B.:

$$2, 4, 3, 6, 5, 10, 9, 18, \dots$$

Hier ist das Gesetz dazu:

$$a_{n+1} = \begin{cases} 2 \cdot a_n, & \text{für } n \text{ ungerade,} \\ a_n - 1, & \text{für } n \text{ gerade.} \end{cases}$$

Weitere Beispiele:

Beispiel 5.3.

1. Die Folge $(f_n) = (0, 1, 1, 2, 3, 5, 8, 13, \dots)$, also
$$f_{n+1} = f_n + f_{n-1} \text{ für } n \geq 2 \text{ und } f_1 = 0, f_2 = 1,$$
heißt Folge der **Fibonacci–Zahlen** (siehe auch Abschnitt 1.3.4).
2. Die Folge
$$(a_n) = (1, 2, 4, 6, 10, 12, 16, 18, 22, \dots)$$
gehorcht dem Gesetz: $a_n = n$–te Primzahl -1.

In der Analysis werden Folgen verwendet, um eine gesuchte Zahl besser und besser zu approximieren:

$$(3, 3.1, 3.14, 3.141, 3.1415, \dots)$$

approximiert die Kreiszahl

$$\pi = 3.141592\dots$$

immer besser, je größer n ist. Wir geben dieser Idee einen präzisen Sinn:

Definition 5.4. *Sei (a_n) eine Folge reeller Zahlen und $a \in \mathbb{R}$ eine weitere Zahl. (a_n) heißt **konvergent** gegen a, in Zeichen*

$$\lim_{n\to\infty} a_n = a,$$

wenn $\forall \varepsilon > 0\ \exists\, n_0 = n_0(\varepsilon) \in \mathbb{N}$, so dass $|a_n - a| < \varepsilon\ \forall\ n \geq n_0$.

*a heißt **Limes** oder **Grenzwert** der Folge (a_n).*

Beispiel 5.5.

1. $(a_n) = (\frac{1}{n})$. Es gilt: $\lim_{n\to\infty} \frac{1}{n} = 0$, denn ist $\varepsilon > 0$ vorgegeben, so existiert nach dem Archimedischen Axiom ein n_0 mit $n_0 > \frac{1}{\varepsilon}$ und daher:
$$\varepsilon > \frac{1}{n_0} \geq \frac{1}{n}\ \forall n \geq n_0.$$
Also gilt:
$$\left|\frac{1}{n} - 0\right| = \left|\frac{1}{n}\right| = \frac{1}{n} < \varepsilon\ \forall n \geq n_0.$$
2. Es gilt:
$$\lim_{n\to\infty} \frac{n+1}{n} = 1.$$
In der Tat gilt:
$$\left|\frac{n+1}{n} - 1\right| = \frac{1}{n} \leq \frac{1}{n_0} < \varepsilon \text{ für } n_0 < \varepsilon\ ,$$
wie im vorigen Beispiel. Also: $n_0 = \lceil \frac{1}{\varepsilon} \rceil$.

3. Die **konstante Folge** (a_n) mit $a_n = a$ für ein gewisses festes a konvergiert gegen a.
4. Die Folge $((-1)^n)$ konvergiert nicht, denn: Ist $\lim_{n\to\infty}(-1)^n = a$ für ein a, so existiert zu $\varepsilon = 1$ ein n_0, so dass

$$\left|(-1)^n - a\right| < 1 \ \forall n \geq n_0.$$

Insbesondere gilt:

$$\begin{aligned} 2 &= \left|(-1)^{n_0} - a + a + (-1)^{n_0+1}\right| \\ &\overset{\Delta\text{-Ungl.}}{\leq} \left|(-1)^{n_0} - a\right| + \left|a - (-1)^{n_0+1}\right| \\ &< 1 + 1 = 2. \end{aligned}$$

Dies ist ein Widerspruch.

Bemerkung 5.6. Für jedes $\varepsilon > 0$ liegen bis auf endlich viele Folgenglieder a_n alle Folgenglieder einer gegen a konvergenten Folge in dem Intervall

$$]a - \varepsilon, a + \varepsilon[,$$

wobei dies wie folgt definiert ist.

Definition 5.7. *Wir definieren für $a, b \in \mathbb{R}$ mit $a \leq b$ die folgenden **Intervalle**:*

$$\begin{aligned} [a,b] &:= \{x \in \mathbb{R} \mid a \leq x \leq b\} \ \textit{(**geschlossenes Intervall**),} \\ ]a,b[&:= \{x \in \mathbb{R} \mid a < x < b\} \ \textit{(**offenes Intervall**),} \\ [a,b[&:= \{x \in \mathbb{R} \mid a \leq x < b\} \ \textit{(**halboffenes Intervall**),} \\ ]a,b] &:= \{x \in \mathbb{R} \mid a < x \leq b\} \ \textit{(**halboffenes Intervall**).} \end{aligned}$$

Manchmal schreibt man für $]a,b[$ auch (a,b) etc.

Bemerkung 5.8. Der Grenzwert $a = \lim a_n$ einer konvergenten Folge ist eindeutig bestimmt.

Beweis. Sei $a' \in \mathbb{R}$ ein weiterer Grenzwert. Dann gibt es zu $\varepsilon > 0$ Zahlen n_1, n_2, so dass

$$|a_n - a| < \varepsilon \ \forall \ n \geq n_1, \ |a_n - a'| < \varepsilon \ \forall \ n \geq n_2.$$

Dann gilt für $n \geq \max(n_1, n_2)$:

$$|a - a'| = |a - a_n + a_n - a'| \leq |a - a_n| + |a_n - a'| < \varepsilon + \varepsilon = 2\varepsilon.$$

Also: $|a - a'| < 2\varepsilon$ für jedes beliebige $\varepsilon > 0$ und daher: $a = a'$. □

Satz 5.9 (Rechenregeln für Grenzwerte). *Es seien $(a_n), (b_n)$ zwei konvergente Folgen mit Grenzwerten $a = \lim a_n$, $b = \lim b_n$. Dann gilt:*

1. *Auch die Folge (a_n+b_n) ist konvergent mit Grenzwert $a+b$. Mit anderen Worten:*

$$\lim_{n\to\infty}(a_n+b_n)=\lim_{n\to\infty}a_n+\lim_{n\to\infty}b_n,$$

   falls die rechte Seite existiert.

2. *Die Folge der Produkte $(a_n\cdot b_n)$ konvergiert mit Grenzwert $a\cdot b$ bzw.:*

$$\lim_{n\to\infty}(a_n\cdot b_n)=\lim_{n\to\infty}a_n\cdot\lim_{n\to\infty}b_n,$$

   falls die rechte Seite existiert.

3. *Ist $b\neq 0$ und $b_n\neq 0$ für alle n, so konvergiert auch die Folge $\left(\frac{a_n}{b_n}\right)$ und es gilt:*

$$\lim_{n\to\infty}\frac{a_n}{b_n}=\frac{\lim a_n}{\lim b_n}=\frac{a}{b}.$$

Beweis.

1. Sei $\varepsilon>0$ vorgegeben. Nach Voraussetzung $\exists n_1,n_2\in\mathbb{N}$, so dass

$$|a_n-a|<\frac{\varepsilon}{2}\ \forall\ n\geq n_1 \text{ und } |b_n-b|<\frac{\varepsilon}{2}\ \forall\ n\geq n_2.$$

   Dann gilt für $n_0=\max(n_1,n_2)$:

$$|a_n+b_n-(a+b)|=|a_n-a+b_n-b|\overset{\Delta\text{–Ungl.}}{\leq}|a_n-a|+|b_n-b|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon\ \forall n\geq n_0.$$

2. Wir verwenden die Δ–Ungleichung in der Form:

$$\begin{aligned}|a_nb_n-ab|&=|a_nb_n-a_nb+a_nb-ab|\\&\leq|a_nb_n-a_nb|+|a_nb-ab|=|a_n|\cdot|b_n-b|+|b|\cdot|a_n-a|.\end{aligned}$$

   Nach Voraussetzung existiert zu $\varepsilon=1$ ein n_1, so dass

$$|a_n-a|<1\ \forall n\geq n_1\Rightarrow|a_n|\leq|a|+1\ \forall n\geq n_1,\ |b_n|\leq|b|+1.$$

   Für $\frac{\varepsilon}{2(|a|+1)}>0$ existiert ein n_3, so dass

$$|a_n-a|<\frac{\varepsilon}{2(|b|+1)}\ \forall n\geq n_3.$$

   Sei dann $n_0=\max(n_1,n_2,n_3)$. Dann gilt:

$$\begin{aligned}|a_nb_n-ab|&\leq|a_n|\cdot|b-b_n|+|b|\cdot|a_n-a|\\&\leq(|a|+1)\cdot|b-b_n|+(|b|+1)\cdot|a_n-a|\\&<(|a|+1)\cdot\frac{\varepsilon}{2(|a|+1)}+(|b|+1)\cdot\frac{\varepsilon}{2(|b|+1)}\\&=\frac{\varepsilon}{2}+\frac{\varepsilon}{2}\\&=\varepsilon.\end{aligned}$$

   für alle $n\geq n_0=\max(n_1,n_2,n_3)$.

3. Die wesentliche Abschätzung ist dieses Mal:

$$\begin{aligned}
|\frac{a_n}{b_n} - \frac{a}{b}| &= |\frac{a_n b - a b_n}{b_n b}| \\
&= \frac{|a_n b - ab + ab - a b_n|}{|b_n b|} \\
&\leq \frac{|a_n b - ab| + |ab - a b_n|}{|b_n| \cdot |b|} \\
&\leq \frac{|a_n - a| \cdot |b|}{|b_n| \cdot |b|} + \frac{|a| \cdot |b - b_n|}{|b_n| \cdot |b|}.
\end{aligned}$$

Zu $\varepsilon = \frac{|b|}{2} > 0$ $\exists n_1$, so dass $|b_n - b| < \frac{|b|}{2}$ $\forall n \geq n_1$, d.h.

$$|b_n| > |b| - \frac{|b|}{2} = \frac{|b|}{2} \; \forall n \geq n_1$$

wegen der Konvergenz von (b_n) gegen $b \neq 0$. Also gilt für $n \geq n_1$:

$$\left|\frac{a_n}{b_n} - \frac{a}{b}\right| \leq \frac{|a_n - a|}{|b_n|} + |a| \cdot \frac{|b_n - b|}{|b_n| \cdot |b|} \leq \frac{|a_n - a|}{|b|/2} + \frac{|a| \cdot |b_n - b|}{|b|/2}.$$

Sei $\varepsilon > 0$. dann existiert ein n_2, so dass $|a_n - a| < \frac{\varepsilon \cdot |b|}{4}$ $\forall n \geq n_2$ und es existiert ein n_3, so dass $|b_n - b| < \frac{1}{(|a|+1)} \cdot \frac{|b|^2}{4} \cdot \varepsilon$. Damit gilt:

$$\begin{aligned}
|\frac{a_n}{b_n} - \frac{a}{b}| &< \frac{\varepsilon \cdot |b|}{4} \cdot \frac{2}{|b|} + (|a| + 1) \cdot \frac{|b|^2}{4 \cdot (|a| + 1)} \cdot \varepsilon \cdot \frac{2}{|b|^2} \\
&= \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

für alle $n \geq n_0 = \max(n_1, n_2, n_3)$.

□

Beispiel 5.10. Wir betrachten die Folge (a_n) mit

$$a_n = \frac{n^2 + 2n - 1}{2n^2 + 3n + 1} = \frac{1 + \frac{2}{n} - \frac{1}{n^2}}{2 + \frac{3}{n} + \frac{1}{n^2}}.$$

Nun gilt:

$$\lim_{n\to\infty} \frac{1}{n} = 0 \;\Rightarrow\; 0 = \lim_{n\to\infty} \frac{1}{n} \cdot \lim_{n\to\infty} \frac{1}{n} = \lim_{n\to\infty} \frac{1}{n^2}.$$

Es folgt:

$$\lim_{n\to\infty}(1 + \frac{2}{n} - \frac{1}{n^2}) = \lim_{n\to\infty} 1 + \lim_{n\to\infty} 2\frac{1}{n} - \lim_{n\to\infty} \frac{1}{n^2} = 1 + 2 \cdot 0 - 0 = 1.$$

Ähnlich können wir $\lim_{n\to\infty}(2 + \frac{3}{n} + \frac{1}{n^2}) = 2$ zeigen, so dass sich insgesamt ergibt:

$$\lim_{n\to\infty} a_n = \frac{1 + \frac{2}{n} - \frac{1}{n^2}}{2 + \frac{3}{n} + \frac{1}{n^2}} = \frac{1}{2}.$$

5.2 Beispiele für Folgen in der Informatik

Sei A ein Algorithmus, den wir mit Eingabe unterschiedlicher Längen n aufrufen können.

Beispiel 5.11. Addition zweier Zahlen mit n Ziffern. Sei $a_n :=$ maximale Laufzeit dieses Algorithmus für eine Eingabe der Länge n. (a_n) ist eine Folge reeller Zahlen. Schriftliche Addition ist bekannt:

$$\begin{array}{cccccc} a_{n-1} & \cdots & a_2 & a_1 & a_0 \\ b_{n-1} & \cdots & b_2 & b_1 & b_0 \\ c_{n-1} & \cdots & c_2 & c_1 & \\ \hline d_{n-1} & \cdots & d_2 & d_1 & d_0. \end{array}$$

Die Addition von zwei bzw. drei einstelligen Zahlen kann man in einer Tabelle nachschlagen. Brauchen wir hierfür t Takte, so benötigen wir insgesamt $n \cdot t$ Takte. Ist ein Takt s Sekunden lang, so erhalten wir:

$$a_n = n \cdot t \cdot s.$$

5.3 Landau–Symbole (O– und o–Notation)

Sei (a_n) eine Folge positiver reeller Zahlen und (b_n) eine weitere solche Folge. Dann sagen wir:

$$b_n \in O(a_n) \quad (b_n) \text{ wächst höchstens wie } (a_n),$$

falls eine Konstante $c > 0$ und ein $n_0 \in \mathbb{N}$ existieren, so dass

$$|b_n| \leq c \cdot a_n \ \forall n \geq n_0.$$

Beispiel 5.12. $n \cdot t \cdot s \in O(n)$. Die Aussage, dass sich zwei Zahlen mit n Stellen in der Laufzeit $O(n)$ addieren lassen, ist in vielerlei Hinsicht viel besser und informativer als die genaue Formel, da sie über die Laufzeit für große n eine gut vergleichbare Aussage macht.

Wir schreiben

$$(b_n) \in o(a_n),$$

falls $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$, also:

$$o(n) = \left\{(b_n) \;\middle|\; \lim_{n\to\infty} \frac{b_n}{a_n} = 0\right\}.$$

$O(n)$ und $o(n)$ heißen auch **Landau–Symbole** (es gibt noch weitere davon, wir beschränken uns hier aber auf diese beiden).

5.4 Aufwandsanalyse der Multiplikation

Bei der schriftlichen Multiplikation zweier Zahlen a, b mit n bzw. m Ziffern müssen wir $n{\cdot}m$ Gedächtnisleistungen für das kleine Einmaleins durchführen; der Aufwand ist also insgesamt in $O(n{\cdot}m)$ bzw. $O(n^2)$, falls $n = m$. Es geht auch schneller:

Algorithmus 5.13 (Karatsuba, 1962).

Input: *Zwei natürliche Zahlen a, b mit $n = 2^k$ Binärstellen.*

Output: *Das Produkt $a \cdot b$.*

1. *Schreibe $a = a_0 + a_1 2^{n/2}$, $b = b_0 + b_1 2^{n/2}$ mit a_0, a_1, b_0, b_1 Zahlen mit 2^{k-1} Binärziffern.*
2. *Berechne $a_0 b_0$, $(a_0+a_1)(b_0+b_1)$, $a_1 b_1$ durch rekursives Aufrufen des Algorithmus.*
3. *Gebe das Ergebnis*

$$a_0 b_0 + [(a_0 + a_1)(b_0 + b_1) - a_0 b_0 - a_1 b_1] \cdot 2^{n/2} + a_1 b_1 2^n$$

aus.

Der dritte Schritt involviert die Addition von mehreren Zahlen mit $2^{n/2}$ Binärstellen, ist also von der Ordnung $O(\frac{n}{2}) = O(n)$. Entscheidend ist, dass im zweiten Schritt nur drei statt vier Multiplikationen notwendig sind. Daher ist nämlich der Aufwand für die Multiplikation von zwei n–stelligen Binärzahlen von der Ordnung

$$O(n^{\log_2 3}) \subset O(n^{1.59}),$$

wie wir in einer Aufgabe zeigen werden. Dies ist viel besser als $O(n^2)$.

Multiplikation geht noch wesentlich schneller; es gilt nämlich sogar:

Satz 5.14 (Schönhage–Strassen, 1971). *Zwei n–stellige Zahlen lassen sich mit Aufwand $O(n \log n \log\log n)$ multiplizieren.*

Näheres dazu in Veranstaltungen zu Datenstrukturen und Algorithmen oder Computeralgebra.

5.5 Das Vollständigkeitsaxiom

Wir möchten das sogenannte **Vollständigkeitsaxiom** von $\mathbb{R}$ formulieren. Grob gesehen sagt das Axiom, dass jede Folge, die so aussieht wie eine konvergente Folge, tatsächlich konvergiert.

Definition 5.15. *Sei (a_n) eine Folge reeller Zahlen. (a_n) heißt* ***nach oben beschränkt****,* ***nach unten beschränkt*** *bzw.* ***beschränkt****, wenn es eine Konstante $k \in \mathbb{R}$, genannt* ***Schranke****, gibt, so dass*

$$\begin{aligned} a_n &\leq k \ \forall n \in \mathbb{N}, \\ a_n &\geq k \ \forall n \in \mathbb{N} \text{ bzw.} \\ |a_n| &\leq k \ \forall n \in \mathbb{N}. \end{aligned}$$

(a_n) heißt ***monoton fallend*** *bzw.* ***monoton steigend*** *(auch* ***monoton wachsend*** *genannt), wenn $a_n \geq a_{n+1}$ $\forall n \in \mathbb{N}$ bzw. $a_n \leq a_{n+1}$ $\forall n \in \mathbb{N}$. Eine* ***monotone*** *Folge ist eine Folge, die monoton steigend oder monoton fallend ist.*

(a_n) heißt ***streng monoton fallend****,* ***streng monoton wachsend*** *bzw.* ***streng monoton****, falls die entsprechenden Ungleichungen strikt sind.*

Bemerkung 5.16. Jede konvergente Folge ist beschränkt.

Beweis. Sei (a_n) eine konvergente Folge mit $\lim a_n = a$. Dann gibt es zu $\varepsilon = 1$ ein n_0, so dass $|a_n - a| \leq 1$ $\forall n \geq n_0$. Es folgt: $|a_n| \leq |a| + 1$ $\forall n \geq n_0$. Also: $M = \max\{|a_1|, \ldots, |a_{n-1}|, |a| + 1\}$ ist die Schranke für $|a_n|$. □

Definition 5.17. *Eine Folge (a_n) heißt* ***Cauchy–Folge****, wenn*

$$\forall \varepsilon > 0 \ \exists n_0 \in \mathbb{N}, \text{ so dass } |a_n - a_m| < \varepsilon \ \forall n, m \geq n_0.$$

Bemerkung 5.18. Jede konvergente Folge ist eine Cauchy–Folge.

Beweis. Sei (a_n) eine konvergente Folge mit $\lim a_n = a$ und sei $\varepsilon > 0$ vorgegeben. Da (a_n) gegen a konvergiert, $\exists n_0$, so dass $|a_n - a| < \frac{\varepsilon}{2}$ $\forall n \geq n_0$. Also:

$$|a_n - a_m| = |a_n - a + a - a_m| \overset{\Delta\text{–Ungl.}}{\leq} |a_n - a| + |a - a_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \ \forall n, m \geq n_0.$$

□

Das Vollständigkeitsaxiom von $\mathbb{R}$ können wir nun so formulieren:

Satz 5.19 (Vollständigkeitsaxiom (Cauchy–Kriterium)). *Jede Cauchy–Folge reeller Zahlen konvergiert in $\mathbb{R}$, d.h. ist (a_n) eine Cauchy–Folge, dann $\exists$ $a \in \mathbb{R}$, so dass (a_n) gegen a konvergiert.*

Das Vollständigkeitsaxiom lässt sich auch anders formulieren:

Satz 5.20 (Vollständigkeitsaxiom, 2. Version). *Jede monotone beschränkte Folge konvergiert.*

Beweis (nur eine Plausibilitätsüberlegung). Diese Version machen wir uns für $\mathbb{R}$, definiert als Menge der unendlichen Dezimalzahlen, plausibel: Sei $\mathbb{R} = \{$ Dezimalzahlen $\}$. Es genügt, die Aussage für monoton fallende Folgen positiver Zahlen zu zeigen. Sei (a_n) eine Folge reeller Zahlen mit $a_n \geq a_{n+1}\ \forall n$ und $a_n \geq 0\ \forall n$. Wir bestimmen sukzessive die Dezimalzahlentwicklung des Grenzwerts a. Sei dazu $\lfloor a_1 \rfloor = k$ der ganze Anteil von a_1. Dann liegen alle Glieder der Folge in dem Intervall $[0, k+1[$, was wir in $k+1$ disjunkte Teilintervalle

$$[0,1[\cup[1,2[\cup\cdots\cup[k,k+1[$$

zerlegen. Da (a_n) monoton fallend ist, gibt es genau ein Teilintervall $[i, i+1[$, das unendlich viele Glieder der Folge enthält. Dann ist i der ganze Anteil des Grenzwerts. Anschließend zerlegen wir $[i, i+1[$ wiederum in 10 Teilintervalle:

$$[i, i+1[= [i.0, i.1[\cup[i.1, i.2[\cup\cdots\cup[i.9, i+1[.$$

Wieder gibt es genau ein Teilintervall $[i.i_1, i.(i_1+1)[$, welches unendlich viele Elemente der Folge enthält. Dann ist i_1 die erste Nachkommastelle (bzw. Nachpunktstelle in der obigen Notation) des Grenzwerts. Als nächstes zerlegen wir $[i.i_1, i.i_1 + 0.1[$ in 10 Teilintervalle, um die zweite Nachkommastelle i_2 zu bestimmen usw. Auf diese Weise können wir sämtliche Dezimalstellen i_k des Grenzwertes

$$a = i.i_1 i_2 i_3 \ldots$$

bestimmen. □

Warum ist das Vorige kein Beweis? Der Grund ist, dass in der Definition der unendlichen Dezimalzahlen schon der Begriff Konvergenz eingeht. Dass Folgen, die durch Dezimalbruchentwicklung definiert werden, wie

$$3.14, 3.141, 3.1415, \ldots$$

konvergieren, können wir so nicht zeigen.

Definition 5.21. *Sei (a_n) eine Folge und*

$$n_1 < n_2 < \cdots < n_k < \cdots$$

*eine streng monoton wachsende Folge natürlicher Zahlen. Dann nennen wir die Folge $(a_{n_k})_{k\in\mathbb{N}}$ eine **Teilfolge** von (a_n).*

Satz 5.22. *Jede Folge besitzt eine monotone Teilfolge.*

Beweis. Sei (a_n) eine Folge. Wir nennen das Glied a_m einen **Hochpunkt** der Folge, wenn

$$a_m > a_n\ \forall n \geq m.$$

Hat die Folge (a_n) nur endlich viele Hochpunkte, dann besitzt (a_n) eine monoton wachsende Teilfolge. Ist nämlich a_m der letzte Hochpunkt, so setzen wir $n_1 = m+1$ und zu a_{n_1} gibt es nach Voraussetzung ein $n_2 > n_1$ mit $a_{n_1} \leq a_{n_2}$, zu a_{n_2} ein $n_3 > n_2$ mit $a_{n_2} \leq a_{n_3}$ und rekursiv ein a_{n_k} ein $n_{k+1} > n_k$ mit $a_{n_k} \leq a_{n_{k+1}}$. (a_{n_k}) ist dann die gesuchte monoton wachsende Teilfolge.

Hat andererseits (a_n) unendlich viele Hochpunkte, so bildet die Teilfolge der Hochpunkte, $n_1 < n_2 < \cdots < n_k < \cdots$ mit n_k definiert durch a_{n_k} ist der k–te Hochpunkt eine monoton fallende Teilfolge. □

Satz 5.23. *Die zweite Version (Satz 5.20) des Vollständigkeitsaxioms impliziert die erste (Satz 5.19).*

Um diesen Satz beweisen zu können, benötigen wir folgende Aussage:

Satz 5.24. *Jede Cauchy–Folge ist beschränkt.*

Beweis. Sei (a_n) eine Cauchy–Folge. Dann gibt es zu $a = 1$ ein n_0, so dass

$$|a_n - a_m| < 1 \ \forall n, m \geq n_0.$$

Insbesondere gilt also:

$$|a_n - a_{n_0}| < 1 \ \forall n \geq n_0$$

und damit:

$$|a_n| \leq |a_{n_0}| + 1 \ \forall n \geq n_0.$$

Mit

$$M = \max\Big\{ |a_1, \ldots, |a_{n_0-1}|, |a_{n_0}| + 1 \Big\}$$

haben wir dann eine Schranke für die Folge (a_n) gefunden. □

Beweis (von Satz 5.23). Jede Teilfolge einer Cauchy–Folge (a_n) ist ebenfalls eine. Insbesondere gibt es eine monotone beschränkte Teilfolge (a_{n_k}) von (a_n). Nach der zweiten Version des Vollständigkeitsaxioms hat (a_{n_k}) einen Grenzwert

$$\lim_{k\to\infty} a_{n_k} = a.$$

Wir zeigen, dass auch $\lim_{n\to\infty} a_n = a$ gilt. Sei dazu $\varepsilon > 0$ vorgegeben. Dann existiert ein k_0 mit

$$|a_{n_k} - a| < \frac{\varepsilon}{2} \ \forall\, k \geq k_0$$

und es gibt ein n_0 mit

$$|a_m - a_n| < \frac{\varepsilon}{2} \ \forall n \geq n_0.$$

Sei nun $N = \max\{k_0, n_0\}$. Für $n \geq N$ gilt dann:

$$|a_n - a| = |a_n - a_{n_{k_0}}| + |a_{n_{k_0}} - a| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2},$$

was zu zeigen war. □

Satz 5.25. *Das Vollständigkeitsaxiom (Satz 5.19) impliziert die zweite Version des Vollständigkeitsaxioms (Satz 5.20).*

Beweis. Wir verwenden das Argument der Plausibilitätsüberlegung für die Richtigkeit der zweiten Version. Sei (a_n) eine monotone, beschränkte Folge. Sei M eine Schranke, also $|a_n| < M \ \forall n \in \mathbb{N}$. Wir zeigen zunächst, dass (a_n) eine Cauchy–Folge ist. Sei $\varepsilon > 0$ vorgegeben. Wir wählen $N \in \mathbb{N}$ mit $N > \frac{M}{2\varepsilon}$ (N existiert nach dem archimedischen Axiom, siehe Def. 4.10). Dann ist

$$[-M, M[\subset [-M, -M + 2\varepsilon N[= \bigcup_{k=0}^{2N-1} [-M + k\varepsilon, -M + (k+1)\varepsilon[$$

eine disjunkte Zerlegung. Wegen der Monotonie der Folge (a_n) gibt es genau ein Teilintervall

$$[\, -M + k\varepsilon,\ -M + (k+1)\varepsilon\, [,$$

das alle bis auf endlich viele a_n enthält. Gilt etwa

$$a_n \in [\, -M + k\varepsilon,\ -M + (k+1)\varepsilon\, [\quad \forall n \geq n_0,$$

so folgt:

$$|a_n - a_m| < \varepsilon \ \forall n, m \geq n_0.$$

Nach dem Cauchy–Kriterium hat die Folge (a_n) einen Grenzwert $a \in \mathbb{R}$. Somit ist gezeigt, dass jede monotone, beschränkte Folge konvergiert, wenn das Cauchy–Kriterium erfüllt ist. □

Weitere Axiome zur Charakterisierung der rellen Zahlen benötigen wir nicht:

Satz 5.26. *Seien R und R' zwei archimedisch angeordnete Körper, die beide das Vollständigkeitsaxiom erfüllen. Dann gibt es genau eine bijektive Abbildung*

$$\varphi \colon R \to R',$$

die alle Strukturen erhält. Etwa: $\varphi(a+b) = \varphi(a) + \varphi(b)$, $a > b \Rightarrow \varphi(a) > \varphi(b)$ usw.

Beweis. Da R und R' angeordnet sind und die Charakteristik 0 haben, lässt sich $\mathbb{Q}$ als Teilmenge von R und R' auffassen:

$$\begin{array}{ccc} R & \xrightarrow{\exists \varphi} & R' \\ \uparrow & & \uparrow \\ \mathbb{Q} & \xrightarrow{\mathrm{id}_{\mathbb{Q}}} & \mathbb{Q} \end{array}$$

Wir werden φ so konstruieren, dass

$$\varphi\big|_{\mathbb{Q}} : \mathbb{Q} \to R', \quad \varphi\big|_{\mathbb{Q}}(x) = \varphi(x)$$

(man sagt φ **eingeschränkt auf** $\mathbb{Q}$) die Identität $\mathrm{id}_{\mathbb{Q}}$ auf $\mathbb{Q}$ ist, d.h. $\varphi\big|_{\mathbb{Q}} = \mathrm{id}_{\mathbb{Q}}$. Sei dazu $a \in R$ gegeben. Zu $\varepsilon = \frac{1}{n}$ wählen wir eine rationale Zahl $a_n \in]a-\varepsilon, a+\varepsilon[$. Zum Beispiel können wir a_n in der Form $\frac{m}{n}$ wählen mit einem geeigneten $m \in \mathbb{Z}$, denn da der Durchmesser des Intervalls $(a+\varepsilon)-(a-\varepsilon) = 2\varepsilon = \frac{2}{n}$ ist, enthält es wenigstens eine der Zahlen $\frac{m}{n}$.

Die Folge (a_n) konvergiert dann gegen $a \in R$: Zu $\varepsilon > 0$ existiert ein $n_0 > \frac{1}{\varepsilon}$ nach dem archimedischen Axiom und $|a_n - a| < \frac{1}{n} \leq \frac{1}{n_0} < \varepsilon \ \forall n \geq n_0$. Die Bilder $\varphi(a_n) \in \mathbb{Q} \subset R'$ sind schon definiert, da $a_n \in \mathbb{Q}$. Die Folge $(\varphi(a_n))$ ist eine Cauchy–Folge in R', da das Cauchy–Kriterium für ε der Gestalt $\varepsilon = \frac{1}{m} \in \mathbb{Q}$ mit $m \in \mathbb{N}$ erfüllt ist. Nach dem archimedischen Axiom genügt es, solche ε zu betrachten.

Sei $a' = \lim_{n\to\infty} \varphi(a_n) \in R'$, der Grenzwert a' existiert nach dem Vollständigkeitsaxiom in R'. Wir definieren dann $\varphi(a) := a' \in R'$. Man kann sich ohne allzu große Mühe folgendes überlegen:

1. Die so definierte Abbildung $\varphi\colon R \to R'$ ist wohldefiniert, d.h. unabhängig von der Wahl der Cauchy–Folge (a_n) in $\mathbb{Q}$ mit $\lim a_n = a$.
2. φ ist bijektiv.
3. φ respektiert sämtliche Strukturen. Also: $\varphi(a+b) = \varphi(a) + \varphi(b)$, $\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b)$, $a < b \Rightarrow \varphi(a) < \varphi(b)$.

□

5.6 Quadratwurzeln

Die Einführung der reellen Zahlen haben wir beispielsweise motiviert durch $\sqrt{2} \notin \mathbb{Q}$. Für jede reelle Zahl $b > 0$ existiert $a = \sqrt{b} \in \mathbb{R}$, d.h. eine reelle Zahl $a \geq 0$, genannt die **Quadratwurzel** von b, mit $a^2 = b$. Wir können dies aus den Axiomen folgern:

Satz 5.27. *Sei $b > 0$ eine reelle Zahl. Die durch*

$$a_{n+1} = \frac{1}{2}\Big(a_n + \frac{b}{a_n}\Big)$$

rekursiv definierte Folge konvergiert für jeden beliebigen Startwert $a_0 > 0$ und der Grenzwert $a = \lim_{n\to\infty} a_n$ (> 0) erfüllt $a^2 = b$, d.h. $a = \sqrt{b}$.

Beweis. Wir gehen schrittweise vor:

1. Zunächst einmal ist $a_n \neq 0$ zu beweisen, damit die Folge überhaupt definiert ist. Wir zeigen: $a_n > 0\ \forall n$ mit Induktion. $a_0 > 0$ ist nach Voraussetzung richtig. Induktionsschritt: $a_n > 0, b > 0 \Rightarrow \frac{b}{a_n} > 0 \Rightarrow a_n + \frac{b}{a_n} > 0$ $\Rightarrow a_{n+1} = \frac{1}{2}(a_n + \frac{b}{a_n}) > 0$.

2. Es gilt $a_n^2 \geq b\ \forall b \geq 1$, denn:

$$\begin{aligned} a_n^2 - b &= \frac{1}{4}\Big(a_{n-1} + \frac{b}{a_{n-1}}\Big)^2 - b \\ &= \frac{1}{4}\Big(a_{n-1}^2 + 2b + \frac{b^2}{a_{n-1}^2}\Big) - b \\ &= \frac{1}{4}\Big(a_{n-1}^2 - 2b + \frac{b^2}{a_{n-1}^2}\Big) \\ &= \frac{1}{4}\Big(a_{n-1} - \frac{b}{a_{n-1}}\Big)^2 \\ &\geq 0. \end{aligned}$$

3. Es gilt: $a_{n+1} \leq a_n$ für alle $n \geq 1$, denn:

$$\begin{aligned} a_n - a_{n+1} &= a_n - \frac{1}{2}\Big(a_n + \frac{b}{a_n}\Big) \\ &= \frac{1}{2}\Big(a_n - \frac{b}{a_n}\Big) = \frac{1}{2a_n}\Big(a_n^2 - b\Big) \\ &\geq 0. \end{aligned}$$

   nach dem vorigen Schritt.

4. $(a_n)_{n\geq 1}$ ist also eine monoton fallende Folge positiver Zahlen. Nach dem Vollständigkeitsaxiom (zweite Version) existiert daher der Grenzwert $a = \lim_{n\to\infty} a_n$.

5. Wir zeigen: $a^2 = b$. Wegen dem vorigen Schritt konvergiert die Folge $(a_{n+1} \cdot a_n)$ ebenfalls und es gilt:

$$\lim(a_n \cdot a_{n+1}) = \lim a_n \cdot \lim a_{n+1} = a^2.$$

   Andererseits ist

$$a_n \cdot a_{n+1} = \frac{1}{2}\Big(a_n^2 + b\Big),$$

   also:

$$\begin{aligned} a^2 &= \lim\Big(a_n \cdot a_{n+1}\Big) = \lim \frac{1}{2}\Big(a_n^2 + b\Big) \\ &= \frac{1}{2}\Big((\lim a_n)^2 + b\Big) = \frac{1}{2}\Big(a^2 + b\Big) \end{aligned}$$

   nach den Rechenregeln für Grenzwerte. Schließlich folgt damit: $\frac{1}{2}a^2 = \frac{1}{2}b$ bzw. $a^2 = b$.

□

Beispiel 5.28. Wir wenden den im Beweis von Satz 5.27 angegebenen Algorithmus zur Berechnung der Quadratwurzel an:

1. $b = 4$, $a_0 = 1$. Der korrekte Grenzwert ist also $\sqrt{b} = 2$. Es ergibt sich folgende Tabelle:

n	a_n	$\frac{b}{a_n}$
0	1	4
1	2.5	1.6
2	2.05	$1.\overline{95121}$
3	2.0006097	...

2. Für $b = 2$ und $a_0 = 1$ erhalten wir als Grenzwert $\sqrt{2}$:

n	a_n	$\frac{b}{a_n}$
0	1	2
1	1.5	1.3333...
2	1.41666...	1.41176
3	1.4142	1.414211
4	1.414213	...

Bemerkung 5.29. Sei $b > 0$. Die Konvergenz der Folge

$$a_{n+1} = \frac{1}{2}\Big(a_n + \frac{b}{a_n}\Big)$$

gegen $a = \sqrt{b}$ ist bemerkenswert schnell: Wir definieren den **relativen Fehler** f_n von a_n durch die Formel

$$a_n = a \cdot (1 + f_n).$$

Dann ist $f_n \geq 0$ für $n \geq 1$. Einsetzen in die Gleichung $a_{n+1} = \frac{1}{2}(a_n + \frac{b}{a_n})$ ergibt:

$$a(1 + f_{n+1}) = \frac{1}{2}(a(1 + f_n) + \frac{a^2}{a(1 + f_n)}$$

bzw.

$$1 + f_{n+1} = \frac{1}{2}\Big((1 + f_n) + \frac{1}{1 + f_n}\Big) = \frac{1}{2} \cdot \frac{2 + 2f_n + f_n^2}{1 + f_n}.$$

Es folgt:

$$f_{n+1} = \frac{1}{2} \cdot \frac{f_n^2}{1 + f_n} \leq \frac{1}{2} \cdot \min\{f_n, f_n^2\}.$$

Ist also der relative Fehler $f_n \geq 1$, so halbiert er sich wenigstens in jedem Schritt. Ist nun $f_n < 1$, dann ist $f_{n+1} = \frac{1}{2} \cdot f_n^2$. In diesem Fall verdoppeln sich die relevanten Stellen mit jedem Iterationsschritt. Man spricht daher von **quadratischer Konvergenz**.

Bemerkung 5.30 (Monotonie der Quadratwurzel). Seien $x, y \in \mathbb{R}_{>0}$. Dann gilt:

$$\sqrt{x} > \sqrt{y} \quad \Longleftrightarrow \quad x > y,$$

denn:

$$(\sqrt{x} - \sqrt{y}) \cdot \underbrace{(\sqrt{x} + \sqrt{y})}_{>0} = x - y \quad \Longrightarrow \quad \left(\sqrt{x} - \sqrt{y} > 0 \iff x - y > 0\right).$$

5.7 Zur Existenz der reellen Zahlen

Nach Satz 5.26 ist es egal, wie wir uns von der Existenz von $\mathbb{R}$ überzeugen. Zwei Konstruktionen von $\mathbb{R}$ aus $\mathbb{Q}$ sind gebräuchlich:

- Cauchy–Folgen modulo Nullfolgen,
- Dedekindsche[1] Schnitte.

Beide werden wir kurz erläutern.

5.7.1 Cauchy–Folgen modulo Nullfolgen

Definition 5.31. *Eine **Nullfolge** (a_n) ist eine Folge, die gegen 0 konvergiert. Wir betrachten jetzt*

$$\begin{aligned} M &= \{(a_n) \mid (a_n) \textit{ ist eine Cauchy–Folge rationaler Zahlen }\} \\ &= \{(a_n) \mid \forall N > 0 \exists n_0 : |a_n - a_m| < \frac{1}{N} \forall n, m \geq n_0\}. \end{aligned}$$

Offenbar ist

$$M \subset \mathbb{Q}^{\mathbb{N}} = \{\mathbb{N} \to \mathbb{Q}\}.$$

Wir definieren auf M eine Äquivalenzrelation durch

$$(a_n) \sim (b_n) \iff (a_n - b_n) \textit{ ist eine Nullfolge }.$$

Dann können wir

$$\mathbb{R} := M/\sim$$

[1] Dedekind: Deutscher Mathematiker (1831–1916)

als Definition von $\mathbb{R}$ verwenden:

Addition und Multiplikation definieren wir repräsentantenweise auf $M/\sim$:

$$[(a_n)] + [(b_n)] := [(a_n + b_n)].$$

Dies ist wohldefiniert, da die Summe zweier Nullfolgen eine Nullfolge ist. Die Multiplikation

$$[(a_n)] \cdot [(b_n)] := [(a_n \cdot b_n)]$$

ist wohldefiniert, da Cauchy–Folgen beschränkt sind und da das Produkt einer beschränkten Folge mit einer Nullfolge eine Nullfolge ergibt.

Mit diesen beiden Verknüpfungen wird $M/\sim$ ein Körper:

- *Die 0 ist die Äquivalenzklasse, die aus allen Nullfolgen besteht.*
- *Die 1 wir von der Konstanten Folge $(1)_{n\in\mathbb{N}}$ repräsentiert.*
- *Ist $[(a_n)] \neq 0$, also (a_n) ist eine Cauchy–Folge, die keine Nullfolge ist, so existiert ein $a = \frac{1}{N} > 0$ und ein n_0, so dass*

  $$|a_n| \geq \frac{1}{N} \; \forall n \geq n_0.$$

  Die Folge

  $$(b_n) \text{ mit } b_n = \begin{cases} 1, & \text{falls } n < n_0, \\ \frac{1}{a_n}, & \text{falls } n \geq n_0, \end{cases}$$

  repräsentiert das Inverse.
- *Schließlich zum Vollständigkeitsaxiom für $M/\sim$: Ist (a_k) eine Cauchy–Folge in $M/\sim$ mit a_k repräsentiert durch die Folge $a_k = [(a_{k_n})_{n\in\mathbb{N}}]$, so ist die Diagonalfolge (a_{n_n}) ebenfalls eine Cauchy–Folge in $\mathbb{Q}$ und man kann recht einfach zeigen, dass tatsächlich gilt:*

  $$\lim_{n\to\infty} a_k = [(a_{nn})].$$

5.7.2 Dedekindsche Schnitte

Definition 5.32. *Eine disjunkte Zerlegung*

$$\mathbb{Q} = U \dot\cup V \text{ mit } U, V \neq 0 \text{ und } u < v \; \forall u \in U \text{ und } v \in V$$

heißt ***Dedekindscher Schnitt****.*

Ein **gut gewählter** rationaler Dedekindscher Schnitt ist ein Dedekindscher Schnitt der Gestalt

$$U_r = \{x \in \mathbb{Q} \mid x \leq r\}, \; V_r = \{x \in \mathbb{Q} \mid x > r\},$$

wobei $r \in \mathbb{Q}$.

Den Schnitt

$$U'_r = \{x \in \mathbb{Q} \mid x < r\},\ V'_r = \{x \in \mathbb{Q} \mid x \geq r\}$$

nennen wir **schlecht gewählt**.

Alle anderen Dedekindschen Schnitte nennen wir **irrational**.

Gut gewählte Schnitte sind entweder irrationale Schnitte oder gut gewählte rationale Schnitte. Dann können wir

$$\mathbb{R} := \{ \text{ gut gewählter Dedekindscher Schnitt } \} \subset 2^{\mathbb{Q}} \times 2^{\mathbb{Q}}$$

zur Definition von $\mathbb{R}$ nehmen. Der Nachweis sämtlicher Axiome ist länglich, aber nicht schwierig.

5.8 Der Satz von Bolzano–Weierstrass

Nun forumulieren wir noch eine sehr nützliche Aussage:

Satz 5.33 (Bolzano–Weierstrass). *Jede beschränkte Folge reeller Zahlen (a_n) hat eine konvergente Teilfolge.*

Beweis. Sei M eine Schranke für (a_n), etwa

$$-M \leq a_n \leq M.$$

Wir zerteilen sukzessive das Intervall

$$[-M, M] = [-M, 0] \cup [0, M]$$

in jeweils halb so große Intervalle. Wir setzen $N_1 := -M, M_1 := M$ und $n_1 = 1$. Sukzessive wählen wir

$$N_k \leq a_{n_k} \leq M_k,$$

so dass unendlich viele Glieder der Folge (a_n) im Intervall $[N_k, M_k]$ liegen. Ist dies für k getan, dann zerlegen wir

$$[N_k, M_k] = [N_k, \frac{N_k + M_k}{2}] \cup [\frac{N_k + M_k}{2}, M_k]$$

und wählen

$$N_{k+1} = \begin{cases} \frac{N_k+M_k}{2}, & \text{wenn } [\frac{N_k+M_k}{2}, M_k] \text{ } \infty \text{ viele Glieder der Folge enthält,} \\ N_k, & \text{sonst,} \end{cases}$$

$$M_{k+1} = \begin{cases} M_k, & \text{wenn } [\frac{N_k+M_k}{2}, M_k] \text{ } \infty \text{ viele Glieder der Folge enthält,} \\ \frac{N_k+M_k}{2}, & \text{sonst.} \end{cases}$$

Schließlich wählen wir

$$n_{k+1} > n_k, \text{ so dass } a_{n_{k+1}} \in [N_{k+1}, M_{k+1}].$$

Dann ist (a_{n_k}) eine Cauchy–Folge und damit die gesuchte konvergente Teilfolge. □

Definition 5.34. *Sei $M \subset \mathbb{R}$ eine Teilmenge. Eine reelle Zahl A heißt **obere Schranke** von M, wenn $a \leq A$ $\forall a \in M$ gilt. M heißt **nach oben beschränkt**, wenn es eine obere Schranke gibt.*

Satz/Definition 5.35. *Jede nicht leere nach oben beschränkte Teilmenge $M \subset \mathbb{R}$ besitzt eine kleinste obere Schranke, d.h. $\exists$ obere Schranke $A \in \mathbb{R}$ von M, so dass $A \leq A'$ für jede andere obere Schranke A' von M gilt. Wir nennen A das **Supremum** von M, geschrieben:*

$$\sup M := A = \textit{ kleinste obere Schranke von } M.$$

*Analog definieren wir für **nach unten beschränkte** nicht leere Teilmengen $M \subset \mathbb{R}$:*

$$\inf M := \textit{ größte untere Schranke von } M,$$

*das **Infimum** von M.*

Beweis (der Existenz des Supremums). Sei A_0 eine obere Schranke von M und $a_0 \in M$, also $a_0 \leq A_0$. Wir definieren zwei Folgen

$$(a_n) \text{ mit } a_n \in M$$

und

$$(A_n) \text{ obere Schranke von } M,$$

so dass (a_n) monoton wächst, (A_n) monoton fällt und

$$0 \leq A_n - a_n \leq 2^{-n}(A_0 - a_0)$$

gilt. Dann konvergieren beide Folgen und es gilt:

$$\lim a_n = \lim A_n.$$

Dieser Grenzwert ist das Supremum von M.

Seien a_n, A_n schon gewählt. Dann wählen wir

$$A_{n+1} = \begin{cases} \frac{A_n + a_n}{2}, & \text{falls} \frac{A_n + a_n}{2} \text{ eine obere Schranke ist,} \\ A_n, & \textit{sonst.} \end{cases}$$

$$a_{n+1} = \begin{cases} \text{ein Element von } M > \frac{A_n + a_n}{2}, & \text{falls} \frac{A_n + a_n}{2} \text{ keine obere Schranke ist,} \\ a_n, & \textit{sonst.} \end{cases}$$

Die Folgen (a_n) und (A_n) haben dann die gewünschte Eigenschaft, wie man leicht nachprüfen kann. Das Infimum ergibt sich analog. □

Beispiel 5.36. Sei

$$M = \{(-1)^n \cdot (1 - \frac{1}{n}) \mid n \in \mathbb{N}\}.$$

Dann gilt: $\sup(M) = 1$, $\inf(M) = -1$.

5.9 Mächtigkeit

Wieviele reelle Zahlen gibt es?

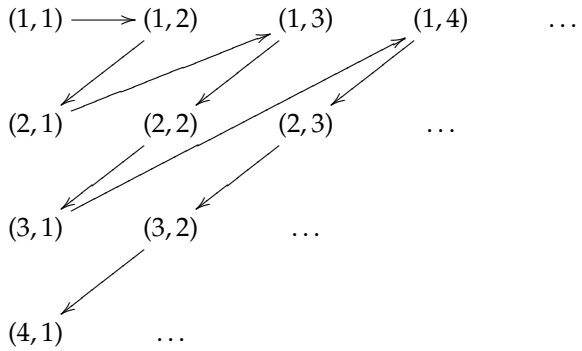
Definition 5.37. *Sei M eine Menge. M heißt **abzählbar**, wenn es eine surjektive Abbildung $\varphi\colon \mathbb{N} \to M$ gibt.*

Beispiel 5.38.

1. Jede endliche Menge ist abzählbar.
2. $\mathbb{Z}$ ist abzählbar: $\mathbb{Z} = \{0, 1, -1, 2, -2, \dots\}$, genauer:

$$\varphi\colon \mathbb{N} \to \mathbb{Z},\ \varphi(n) = \begin{cases} 0, & n = 1, \\ \frac{1}{2}n, & n \text{ gerade}, \\ -\frac{1}{2}(n-1), & n \text{ ungerade } \geq 3. \end{cases}$$

3. $\mathbb{N} \times \mathbb{N}$ ist abzählbar. Wir definieren $\varphi\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ durch



Bemerkung 5.39. Ist M abzählbar unendlich, dann gibt es auch eine bijektive Abbildung $\varphi\colon \mathbb{N} \to M$.

Beweis. Sei $\varphi\colon \mathbb{N} \to M$ surjektiv. Wir definieren $n_1 = 1$,

$$\psi(1) = \varphi(n_1)$$

und rekursiv, falls $\psi(1), \dots, \psi(k)$ schon definiert sind,

$$n_{k+1} = \min\Big\{n \mid \varphi(n) \notin \{\psi(1), \dots, \psi(k)\}\Big\}$$

und $\psi(k+1) = \varphi(n_{k+1})$. □

Satz 5.40. *Es sei $M = \bigcup_{k=1}^{\infty} M_k$ eine abzählbare Vereinigung von abzählbaren Mengen M_k. Dann ist auch M abzählbar.*

Beweis. Sei

$$\psi\colon \mathbb{N} \to \mathbb{N}\times\mathbb{N},\ n \mapsto \psi(n) = (\psi_1(n), \psi_2(n))$$

die Abzählung von $\mathbb{N}\times\mathbb{N}$ und $\varphi_k\colon \mathbb{N} \to M_k$ eine Abzählung von M_k. Dann ist die Abbildung

$$\Phi\colon \mathbb{N} \to \bigcup_{k=1}^{\infty} M_k,\ \Phi(n) = \varphi_{\psi_1(n)}(\psi_2(n))$$

eine Abzählung von M. □

Korollar 5.41. $\mathbb{Q}$ *ist abzählbar.*

Beweis. Sei $M_k = \{\frac{n}{k} \mid n \in \mathbb{Z}\} \subset \mathbb{Q}$. Dann gilt: $M_k \overset{\cong}{\leftarrow} \mathbb{Z} \overset{\cong}{\leftarrow} \mathbb{N}$ ist abzählbar, also auch $\mathbb{Q} = \bigcup_{k=1}^{\infty} M_k$. □

Definition 5.42. *Eine Menge, die nicht abzählbar ist, heißt* ***überabzählbar****.*

Satz 5.43 (Cantors zweites Diagonalargument, 1877[2]). $\mathbb{R}$ *ist überabzählbar.*

Beweis. Es genügt zu zeigen, dass $[0,1[\subset \mathbb{R}$ überabzählbar ist.

Angenommen,

$$\mathbb{N} \to [0,1[,\ n \mapsto a_n$$

ist eine Abzählung. Wir betrachten die Dezimalbruchentwicklung der a_n:

$$a_n = 0.a_{n1}a_{n2}\ldots a_{nk}\ldots$$

mit $a_{nk} \in \{0,\ldots,9\}$ die k–te Nachkommastelle von a_n. Dann sei die Zahl $c = 0.c_1c_2\ldots c_k\ldots$ mit Ziffern c_k durch

$$c_k = \begin{cases} 1, & a_{nn} \neq 1, \\ 2, & a_{kk} = 1. \end{cases}$$

definiert. Offenbar gilt dann $c \neq a_n$, da $c_n \neq a_{nn}$. Also ist

$$\mathbb{N} \to [0,1[,\ n \mapsto a_n$$

nicht surjektiv. Dies ist ein Widerspruch zur Annahme. □

Die Mengenlehre von Cantor beschäftigt sich mit beliebig großen Mengen.

[2] 1874 gab er schon einen anderen Beweis. Sein erstes Diagonalargument zeigt, dass die rationalen Zahlen abzählbar sind.

Definition 5.44 (Cantor). *Zwei Mengen M, N heißen **gleichmächtig**, wenn es eine bijektive Abbildung $\varphi\colon M \to N$ gibt. M heißt **wenigstens so mächtig wie** N, wenn es eine surjektive Abbildung $\psi\colon M \twoheadrightarrow N$ gibt. M heißt **echt mächtiger als** N, wenn es eine surjektive Abbildung $M \twoheadrightarrow N$, aber keine bijektive Abbildung $M \to N$ gibt.*

Man kann unter Verwendung des sogenannten **Auswahlpostulats** (auch **Auswahlaxiom**) zeigen, dass M wenigstens so mächtig wie N und N wenigstens so mächtig wie M impliziert, dass M und N gleichmächtig sind. Das Auswahlpostulat ist dabei folgendes Axiom der Mengenlehre:

Definition 5.45 (Auswahlaxiom). *Sei $(M_i)_{i\in I}$ eine Familie von nichtleeren Mengen. Dann existiert eine Abbildung*

$$a\colon I \to \bigcup_{i\in I} M_i,$$

so dass $a(i) \in M_i$ für alle $i \in I$ gilt. Mit anderen Worten kann man aus den nicht leeren Mengen M_i gleichzeitig je ein Element auswählen.

Man kann zeigen, dass die Potenzmenge 2^M einer Menge M stets echt mächtiger als M ist.

Aufgaben

Aufgabe 5.1 (Die Landau–Symbole). Welche der folgenden Aussagen gilt?

1. $\frac{n^3+n+1}{2n^2-5} \in O(n)$.
2. $\frac{n^3+n+1}{2n^2-5} \in o(n)$.
3. $\frac{2n-5}{20n\sqrt{n}+1000} \in O\left(\frac{1}{n}\right)$.
4. $\frac{20n\sqrt{n}+1000}{2n-5} \in O(n)$.

Aufgabe 5.2 (Quantoren und ε). Sei (a_n) eine Folge in $\mathbb{R}$ und $a \in \mathbb{R}$. Welche Implikationen bestehen zwischen den folgenden sechs Aussagen?

1. $\forall \varepsilon > 0\ \exists n_0 :\ \forall n \geq n_0\ |a_n - a| < \varepsilon$,
2. $\exists \varepsilon > 0\ \exists n_0 :\ \forall n \geq n_0\ |a_n - a| < \varepsilon$,
3. $\forall \varepsilon > 0\ \forall n_0 :\ \forall n \geq n_0\ |a_n - a| < \varepsilon$,
4. $\exists \varepsilon > 0\ \forall n_0 :\ \forall n \geq n_0\ |a_n - a| < \varepsilon$,
5. $\exists n_0\ \forall \varepsilon > 0 :\ \forall n \geq n_0\ |a_n - a| < \varepsilon$,
6. $\exists n_0\ \forall \varepsilon > 0 :\ \exists n \geq n_0\ |a_n - a| < \varepsilon$.

Geben Sie Beispiele von Folgen an, die zeigen, dass weitere Implikationen nicht bestehen.

Aufgabe 5.3 (Quantoren und ε). Für $n \in \mathbb{N}$ definieren wir die Folgen:

$$\begin{aligned} a_n &= \sqrt{n+1000} - \sqrt{n}, \\ b_n &= \sqrt{n+\sqrt{n}} - \sqrt{n}, \\ c_n &= \sqrt{n+\frac{n}{1000}} - \sqrt{n}. \end{aligned}$$

Zeigen Sie: Für $1 \leq n < 1.000.000$ gilt $a_n > b_n > c_n$, aber

$$\lim_{n\to\infty} a_n = 0, \text{ und } \lim_{n\to\infty} b_n = \frac{1}{2}.$$

und die Folge $(c_n)_{n\in\mathbb{N}}$ ist unbeschränkt.

Aufgabe 5.4 (Nullfolgen). Eine Folge (a_n) heißt **Nullfolge**, falls $\lim_{n\to\infty} a_n = 0$. Sei (a_n) eine Nullfolge, (b_n) eine Cauchy–Folge und (c_n) eine beschränkte Folge. Zeigen Sie:

1. $(a_n + b_n)$ ist eine Cauchy–Folge.
2. $(a_n + c_n)$ und $(b_n + c_n)$ sind beschränkte Folgen.

Aufgabe 5.5 (Konvergenz von Folgen). Wir definieren die Folge $(a_n)_{n\in\mathbb{N}_0}$ durch $a_0 = 1$ und

$$a_n = \sqrt{1 + a_{n-1}} \quad \forall n \geq 1.$$

Zeigen Sie, dass die Folge konvergiert, und bestimmen Sie den Grenzwert.

Aufgabe 5.6 (Infimum und Supremum). Seien

$$M_1 = \{x \in \mathbb{Q} \mid x^2 < 2\} \subset \mathbb{R} \text{ und } M_2 = \{x \in \mathbb{Q} \mid x^2 > 2\} \subset \mathbb{R}.$$

Welche dieser Mengen besitzt ein Supremum, welche ein Infimum? Geben Sie es jeweils an, wenn es existiert.

Aufgabe 5.7 (Abzählbarkeit). Zeigen Sie:

1. Die Menge aller endlichen Teilmengen von $\mathbb{N}$ ist abzählbar.
2. Die Menge aller Teilmengen von $\mathbb{N}$ ist überabzählbar.

Aufgabe 5.8 (Grenzwerte von Folgen / Teilfolgen). Zeigen Sie, dass die Folge

$$a_n = \sin\left(\pi \sqrt{n}\right) + \cos\left(\pi \log_2(n)\right)$$

eine konvergente Teilfolge hat und geben Sie eine solche an.

Aufgabe 5.9 (Mächtigkeit). Zeigen Sie, dass die Potenzmenge $2^{\mathbb{R}}$ von $\mathbb{R}$ echt mächtiger ist als die Menge $\mathbb{R}$ selbst.

6

Reihen

...

6.1 Definition und erste Eigenschaften

Definition 6.1. *Sei* $(a_k)_{k\in\mathbb{N}_0}$ *eine Folge. Die Folge* (s_n) *der* ***Partialsummen***

$$s_n = \sum_{k=0}^{n} a_k$$

nennen wir eine ***Reihe****, die wir mit*

$$\sum_{k=0}^{\infty} a_k$$

bezeichnen. Ist die Folge (s_n) *konvergent, so verwenden wir für den Grenzwert* $s = \lim_{n\to\infty} s_n$ *die Notation*

$$s = \sum_{k=0}^{\infty} a_k$$

und nennen die Reihe ***konvergent****.*

Also: $\sum_{k=0}^{\infty} a_k$ *hat zwei Bedeutungen:*

1. *die Folge der Partialsummen* $s_n = \sum_{k=0}^{n} a_k$*,*
2. *der Grenzwert* $\lim_{n\to\infty} s_n$*, falls er existiert.*

Häufig treten Folgen in der Form von Reihen auf:

Beispiel 6.2.

1. $\sum_{n=1}^{\infty} \frac{1}{n}$,
2. $\sum_{n=0}^{\infty} \frac{1}{2^n}$,
3. **Dezimalbruchentwicklung**: $d_k \in \{0, \dots, 9\}$:

$$\sum_{k=1}^{\infty} d_k \cdot 10^{-k} = 0.d_1 d_2 d_3 \dots$$

   Wir werden sehen, dass solche Reihen stets konvergieren.
4. Ist $(c_n)_{n\in\mathbb{N}}$ eine beliebige Folge, dann ist mit $a_1 = c_1$ und $a_k = c_k - c_{k-1}$ für $k \geq 2$ eine Reihe $\sum_{n=1}^{\infty} a_k$ gegeben, deren Partialsummen gerade die Folge (c_k) bilden.

Eigentlich sind Reihen also gar nichts Neues. Gelegentlich lässt sich dies verwenden, um Grenzwerte auszurechnen:

Beispiel 6.3 (Teleskopreihen). Wir zeigen:

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1.$$

Idee:

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}.$$

Also:

$$\begin{aligned} s_k = \sum_{n=1}^{k} \frac{1}{n(n+1)} &= \sum_{n=1}^{k} \Big(\frac{1}{n} - \frac{1}{n+1}\Big) \\ &= 1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \dots + \frac{1}{k} - \frac{1}{k+1} \\ &= 1 - \frac{1}{k+1}. \end{aligned}$$

Tatsächlich ist demnach $\lim_{k\to\infty} s_k = 1$.

Satz 6.4 (Cauchy–Kriterium für Reihen). *Eine Reihe $\sum_{k=0}^{\infty} a_k$ konvergiert genau dann, wenn $\forall \varepsilon > 0$ ein n_0 existiert, so dass:*

$$\Big|\sum_{k=n}^{m} a_k\Big| < \varepsilon \ \forall\ m, n \geq n_0.$$

Insbesondere ist also bei einer konvergenten Reihe $\sum a_k$ die Folge (a_k) eine Nullfolge.

Beweis. klar. □

6.2 Konvergenzkriterien für Reihen

Die Konvergenz von Reihen einzusehen ist manchmal sehr einfach, wie wir gleich an Beispielen sehen werden. Einige Kriterien für ihre Konvergenz sind besonders leicht anzuwenden. Wir gehen hier auf die wichtigsten ein.

Definition 6.5. *Eine **alternierende Reihe** ist eine Reihe der Gestalt*

$$\sum_{k=0}^{\infty}(-1)^k a_k,$$

wobei $a_k \geq 0$ $\forall k$.

Satz 6.6 (Leibnizkriterium). *Ist (a_n) eine monoton fallende Nullfolge, so ist die alternierende Reihe*

$$\sum_{k=0}^{\infty}(-1)^k a_k$$

konvergent.

Beweis. Wir betrachten die Teilfolgen (s_{2n}) und (s_{2n+1}) der geraden und ungeraden Partialsummen. Dann gilt:

$$\begin{aligned} s_{2n+2} &= s_{2n} - a_{2n+1} + a_{2n+2} \leq s_{2n}, \text{ da } a_{2n+1} \geq a_{2n+2}, \text{ und} \\ s_{2n+1} &= s_{2n-1} + \underbrace{a_{2n} - a_{2n+1}}_{\geq 0} \geq s_{2n-1}. \end{aligned}$$

Ferner ist

$$s_{2n+1} = s_{2n} - a_{2n+1} \leq s_{2n}.$$

Es folgt:

$$s_0 \geq s_2 \geq \cdots \geq s_{2n} \geq s_{2n+2} \cdots \text{ und } \cdots s_{2n+1} \geq \cdots \geq s_3 \geq s_1.$$

Daher ist (s_{2n}) eine monoton fallende, nach unten durch s_1 beschränkte Folge und (s_{2n+1}) ist monoton wachsend, nach oben beschränkt durch s_0. Daher existieren die Grenzwerte $\lim s_{2n}$ und $\lim s_{2n+1}$ und

$$\lim s_{2n} - \lim s_{2n+1} = \lim(s_{2n} - s_{2n+1}) = \lim a_{2n+1} = 0.$$

Daher sind beide Grenzwerte identisch, d.h. $\lim s_{2n} = \lim s_{2n+1} = s$ für ein gewisses $s \in \mathbb{R}$, also:

$$\lim s_k = s, \text{ also } \sum_{n=0}^{\infty}(-1)^n a_n = s.$$

□

Beispiel 6.7. Die Reihen

$$\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

und

$$\sum_{n=0}^{\infty}(-1)^{n}\frac{1}{2n+1} = 1 - \frac{1}{3} + \frac{1}{5} - \cdots$$

konvergieren nach dem Leibnizkriterium.

Schwieriger ist es, ihre Grenzwerte zu bestimmen. Es gilt:

$$\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n} = \ln 2$$

und

$$\sum_{n=0}^{\infty}(-1)^{n}\frac{1}{2n+1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \frac{\pi}{4}.$$

Wir werden dies erst zu einem späteren Zeitpunkt beweisen können.

Satz 6.8 (Geometrische Reihe). *Sei $q \in \mathbb{R}$. Die Reihe $\sum_{n=0}^{\infty} q^n$ konvergiert genau dann, wenn $|q| < 1$. In dem Fall ist*

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

Beispiel 6.9. Es gilt:

$$\sum_{n=0}^{\infty}\frac{1}{2^n} = \sum_{n=0}^{\infty}\left(\frac{1}{2}\right)^n = \frac{1}{1-\frac{1}{2}} = 2.$$

Die ersten Glieder dieser Reihe sind: $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots$.

Beweis (des Satzes 6.8 über die geometrische Reihe). $|q| < 1$ ist notwendig, da sonst (q^n) keine Nullfolge ist. Wir zeigen zunächst ($q \neq 1$ vorausgesetzt):

$$s_n = \sum_{k=0}^{n} q^n = \frac{1-q^{n+1}}{1-q}$$

mit Induktion nach n. Es gilt:

$$s_0 = \sum_{k=0}^{0} q^k = q^0 = 1 = \frac{1-q}{1-q}.$$

Induktionsschritt $n \to n+1$:

$$\begin{aligned} s_{n+1} &= s_n + q^{n+1} \overset{\text{I.-V.}}{=} \frac{1-q^{n+1}}{1-q} + q^{n+1} \\ &= \frac{1-q^{n+1}+(1-q)q^{n+1}}{1-q} \\ &= \frac{1-q^{n+2}}{1-q}. \end{aligned}$$

Wegen $|q| < 1$ gilt $\lim_{n\to\infty} q^{n+1} = 0$ und daher:

$$\sum_{k=0}^{n} q^k = \frac{1-q^{n+1}}{1-q} \overset{n\to\infty}{\longrightarrow} \frac{1}{1-q}.$$

□

Beispiel 6.10. Wie man aus der Schule weiß, gilt tatsächlich:

$$0.99999\ldots =: 0.\overline{9} = \sum_{n=1}^{\infty} 9 \cdot \frac{1}{10^n} = \frac{9}{10} \cdot \sum_{k=0}^{\infty} \Big(\frac{1}{10}\Big)^k = \frac{9}{10} \cdot \frac{1}{1-\frac{1}{10}} = 1.$$

Beispiel 6.11. Die Reihe

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots$$

heißt **harmonische Reihe**. Sie konvergiert nicht, denn:

$$1 + \frac{1}{2} + \underbrace{(\frac{1}{3} + \frac{1}{4})}_{\geq \frac{1}{4}+\frac{1}{4}=\frac{1}{2}} + \underbrace{(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8})}_{\geq \frac{4}{8}=\frac{1}{2}} + \underbrace{(\frac{1}{9} + \cdots + \frac{1}{16})}_{\geq \frac{8}{16}=\frac{1}{2}} + \cdots,$$

also:

$$s_{2^k} = \sum_{k=1}^{2^k} \frac{1}{n} \geq 1 + \underbrace{\frac{1}{2} + \cdots + \frac{1}{2}}_{k \text{ Summanden}} \geq 1 + \frac{k}{2} = \frac{k+2}{2}.$$

Die Folge der Partialsummen ist daher unbeschränkt und damit die Reihe **divergent** (d.h. nicht konvergent). Beispielsweise ist die Reihe $\sum_{k=0}^{n}(-1)^k$ auch divergent.

Die Idee, eine Folge mit einer einfacheren zu vergleichen, wollen wir genau ausformulieren:

Definition 6.12. *Es seien $\sum_{n=1}^{\infty} b_n$, $\sum_{n=1}^{\infty} a_n$ zwei Reihen. Dann heißt $\sum a_n$ eine **Majorante** von $\sum b_n$, falls*

$$|b_n| \leq a_n.$$

*$\sum a_n$ heißt **Minorante** von $\sum b_n$, wenn*

$$a_n \leq b_n.$$

Satz 6.13 (Majorantenkriterium). *Sei $\sum_{n=1}^{\infty} a_n$ eine konvergente Majorante der Reihe $\sum_{n=1}^{\infty} b_n$. Dann konvergiert die Reihe $\sum_{n=1}^{\infty} b_n$.*

Beweis. Das Cauchy–Kriterium für Reihen liefert: Für jedes $\varepsilon > 0$ existiert ein n_0 mit

$$\Big|\sum_{k=n}^{m} b_k\Big| \leq \sum_{k=n}^{m} |b_k| \leq \sum_{k=n}^{m} a_k < \varepsilon,$$

falls $n, m \geq n_0$, da $\sum a_n$ konvergiert. Nun zeigt das Kriterium auch, dass die Reihe $\sum b_n$ konvergiert. □

In logischer Negation heißt dies:

Korollar 6.14. *Eine Reihe $\sum_{n=1}^{\infty} a_n$ mit einer divergenten Minorante divergiert.*

Beispiel 6.15. Wir zeigen, dass

$$\sum_{n=1}^{\infty} \frac{1}{n^2}$$

konvergiert:

Es reicht, zu zeigen dass $\sum_{n=1}^{\infty} \frac{1}{(n+1)^2}$ konvergiert, da es auf den ersten Summanden nicht ankommt. Nun gilt: $\frac{1}{(n+1)^2} \leq \frac{1}{n(n+1)}$, also ist die Teleskopreihe $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ eine konvergente Majorante.

Schwieriger ist es, den Grenzwert zu bestimmen. Es gilt:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Der Beweis verwendet sogenannte **Fourierreihen** (siehe dazu Kapitel 27 bzw. genauer Korollar 27.8).

Weitere Kriterien lassen sich aus dem Majorantenkriterium herleiten:

Satz 6.16 (Quotientenkriterium). *Sei $\sum_{n=1}^{\infty} a_n$ eine Reihe mit $a_n \neq 0\ \forall n \geq n_0$ für ein gewisses $n_0 \in \mathbb{N}$. Existiert ein q mit $0 < q < 1$, so dass*

$$\left|\frac{a_{n+1}}{a_n}\right| \leq q\ \forall n \geq n_0,$$

so konvergiert die Reihe. Insbesondere gilt:

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = q < 1 \Rightarrow \sum_{n=1}^{\infty} a_n \text{ konvergiert.}$$

Bemerkung 6.17.

1. $\left|\frac{a_{n+1}}{a_n}\right| < 1\ \forall n \geq n_0$ reicht nicht: Beispielsweise divergiert die harmonische Reihe $\sum_{n=1}^{\infty} \frac{1}{n}$, aber
$$\left|\frac{\frac{1}{n+1}}{\frac{1}{n}}\right| = \frac{n}{n+1} < 1\ \forall n.$$
2. Die Reihe $\sum \frac{1}{n^2}$ konvergiert, obwohl
$$\frac{(\frac{1}{n+1})^2}{\frac{1}{n^2}} = \frac{n^2}{(n+1)^2} \to 1 \text{ für } n \to \infty.$$
3. Offensichtlicherweise divergiert eine Reihe $\sum_{n=1}^{\infty} a_n$, falls gilt:
$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = q > 1.$$

Beweis (des Quotientenkriteriums, Satz 6.16). Ohne Einschränkung sei $|\frac{a_{n+1}}{a_n}| \leq q < 1\ \forall n \geq 0$. Mit Induktion folgt dann: $|a_n| \leq |a_0| \cdot q^n$. Der Induktionsanfang ist klar, wir zeigen also den Induktionsschritt $n \to n+1$:

$$\begin{aligned} |a_{n+1}| &\leq |a_n| \cdot q \text{ (nach Voraussetzung)} \\ &\leq |a_0| \cdot q \cdot q^n \text{ nach I.-V.} \\ &= |a_0| \cdot q^{n+1}. \end{aligned}$$

Also ist $|a_0| \cdot \sum_{n=0}^{\infty} q^n$ eine konvergente Majorante (geometrische Reihe). □

Wir definieren für $x \in \mathbb{R}_{\geq 0}$ und $n \in \mathbb{N}$ die n–te Wurzel $\sqrt[n]{x} =: x^{1/n}$ als die Umkehrfunktion der Funktion $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R},\ f(x) = x^n$, siehe Abb. 6.1. Damit können wir folgendes nützliches Kriterium herleiten:

Satz 6.18 (Wurzelkriterium). *Sei $\sum_{n=0}^{\infty} a_n$ eine Reihe, für die es ein q mit $0 < q < 1$ gibt mit*

$$\sqrt[n]{|a_n|} \leq q\ \forall n,$$

so ist $\sum_{n=0}^{\infty} a_n$ konvergent.

Beweis. Nach Voraussetzung ist $|a_n| \leq q^n\ \forall n$. Also ist $\sum_{n=0}^{\infty} q^n$ eine konvergente Majorante (geometrische Reihe). □

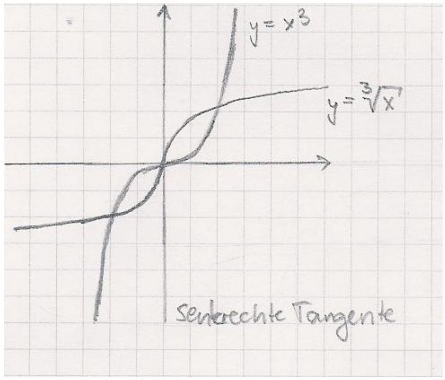


Abbildung 6.1. Die dritte Wurzel als Umkehrfunktion.

6.3 Umordnung von Reihen

Bei endlichen Summen spielt die Reihenfolge des Addierens keine Rolle. Bei Reihen ist dies anders. In manchen Fällen, darf man dennoch umordnen, wie wir sehen werden.

Beispiel 6.19. Die **alternierende harmonische Reihe** ist:

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

Nach dem Leibnizkriterium konvergiert diese gegen einen Wert $s \in \mathbb{R}$. Es gilt: $s \geq 1 - \frac{1}{2} = \frac{1}{2}$. Wir ändern nun die Reihenfolge der Summation

$$1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - \frac{1}{10} \cdots$$

Also:

$$\cdots = \sum_{k=1}^{\infty}\left(\frac{1}{2k-1} - \frac{1}{4k-2} - \frac{1}{4k}\right).$$

In dieser Reihe taucht jeder Stammbruch $\frac{1}{n}$ genau einmal mit dem richtigen Vorzeichen auf. Nun gilt $\frac{1}{2k-1} - \frac{1}{4k-2} = \frac{1}{4k-2}$, also:

$$\begin{aligned}\sum_{k=1}^{\infty}\left(\frac{1}{2k-1} - \frac{1}{4k-2} - \frac{1}{4k}\right) &= \sum_{k=1}^{\infty}\left(\frac{1}{4k-2} - \frac{1}{4k}\right) \\ &= \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{1}{2k-1} - \frac{1}{2k}\right) = \frac{1}{2}s \neq s,\end{aligned}$$

da $s > 0$. Bei der alternierenden harmonischen Reihe kommt es daher auf die Reihenfolge der Summanden an.

Bei gewissen Reihen darf man aber doch umsortieren:

Definition 6.20. *Eine Reihe $\sum_{n=0}^{\infty} a_n$ heißt **absolut konvergent**, wenn*

$$\sum_{n=0}^{\infty} |a_n|$$

konvergiert.

Bemerkung 6.21. Aus dem Cauchy–Kriterium folgt, dass aus absolut konvergent schon konvergent folgt, denn:

$$\Big|\sum_{k=n}^{m} a_k\Big| \overset{\Delta\text{–Ungl.}}{\leq} \sum_{k=n}^{m} |a_k|.$$

Satz 6.22 (Kleiner Umordnungssatz). *Sei $\sum_{n=1}^{\infty} a_n$ eine absolut konvergente Reihe und $\tau\colon \mathbb{N} \to \mathbb{N}$ eine Bijektion. Dann ist auch die Reihe $\sum_{n=1}^{\infty} a_{\tau(n)}$ absolut konvergent und es gilt:*

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} a_{\tau(n)}.$$

Satz 6.23 (Großer Umordnungssatz). *Sei $\sum_{n=1}^{\infty} a_n$ eine absolut konvergente Reihe und $(I_k)_{k\in\mathbb{N}}$ eine Familie von disjunkten Teilmengen $I_k \subset \mathbb{N}$ mit $\bigcup_{k=1}^{\infty} I_k = \mathbb{N}$, wobei I_k sowohl endlich als auch abzählbar sein darf. Dann ist jede der Reihen $\sum_{j\in I_k} a_j$ absolut konvergent und für die Grenzwerte $s_k = \sum_{j\in I_k} a_j$ ist die Reihe $\sum_{k=1}^{\infty} s_k$ absolut konvergent mit Grenzwert ebenfalls*

$$\sum_{k=1}^{\infty} s_k = \sum_{n=1}^{\infty} a_n.$$

Satz 6.24 (Cauchy–Produkt von Reihen). *Es seien $\sum_{i=0}^{\infty} a_i$ und $\sum_{j=0}^{\infty} b_j$ zwei absolut konvergente Reihen und die Folge (d_k) durch die Formel*

$$d_k = \sum_{i=0}^{k} a_i b_{k-i}$$

definiert. Dann ist auch die Reihe $\sum_{k=0}^{\infty} d_k$ absolut konvergent mit Grenzwert

$$\sum_{k=0}^{\infty} d_k = \Big(\sum_{i=0}^{\infty} a_i\Big) \cdot \Big(\sum_{j=0}^{\infty} b_j\Big).$$

Beweis. Wir betrachten die bijektive Abzählung

$$\varphi\colon \mathbb{N}_0 \to \mathbb{N}_0 \times \mathbb{N}_0,\ n \mapsto \varphi(n) = (\alpha(n), \beta(n))$$

und die Reihe $\sum_{n=0}^{\infty} a_{\alpha(n)} b_{\beta(n)}$. Wir zeigen zunächst, dass auch diese Reihe absolut konvergiert. Hierfür reicht es,

$$\sum_{n=0}^{N} |a_{\alpha(n)} b_{\beta(n)}| \leq \Big(\sum_{i=0}^{\infty} |a_i|\Big) \cdot \Big(\sum_{j=0}^{\infty} |b_j|\Big) < \infty$$

für beliebiege N zu zeigen (beschränkte monotone Folgen konvergieren). Dazu sei N gewählt; wir betrachten:

$$\begin{aligned} i_0 &= \max\{\alpha(0), \ldots, \alpha(N)\}, \\ j_0 &= \max\{\beta(0), \ldots, \beta(N)\}. \end{aligned}$$

Dann gilt:

$$\begin{aligned} \sum_{n=0}^{N} |a_{\alpha(n)} b_{\beta(n)}| &\leq \sum_{i=0}^{i_0} |a_i| \cdot \sum_{j=0}^{j_0} |b_j| \\ &\leq \Big(\sum_{i=0}^{\infty} |a_i|\Big) \cdot \Big(\sum_{j=0}^{\infty} |b_j|\Big) < \infty \end{aligned}$$

nach Voraussetzung. Die Reihe $\sum_{k=0}^{\infty} d_k = \sum_{k=0}^{\infty} \sum_{i=0}^{k} a_i b_{k-i}$ ist eine Umordnung der absolut konvergenten Reihe $\sum_{n=1}^{\infty} a_{\alpha(n)} b_{\beta(n)}$, das Produkt $\sum_{i=0}^{\infty} a_i \cdot \sum_{j=0}^{\infty} b_j = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i b_j$ ebenfalls. Nach dem großen Umordnungssatz sind alle diese Reihen absolut konvergent und haben den gleichen Grenzwert $\sum_{k=0}^{\infty} d_k = (\sum_{i=0}^{\infty} a_i)(\sum_{j=0}^{\infty} b_j)$. □

Definition 6.25. *Sei (q_n) eine Folge reeller Zahlen. Das **unendliche Produkt** $\prod_{k=1}^{\infty} q_k$ heißt **konvergent**, wenn der Grenzwert $q = \lim_{n\to\infty} \prod_{k=1}^{n} q_k$ der Partialprodukte existiert. Wir bezeichnen den Grenzwert mit $q = \prod_{k=1}^{\infty} q_k$.*

Satz 6.26 (Euler). *Sei s eine natürliche Zahl ≥ 2 und p_k die k–te Primzahl. Das Produkt $\prod_{k=1}^{\infty} \frac{1}{1-p_k^{-s}}$ konvergiert und hat den gleichen Grenzwert wie $\sum_{n=1}^{\infty} \frac{1}{n^s}$. Für $s = 1$ divergiert das Produkt und die Reihe. Insbesondere gibt es unendlich viele Primzahlen.*

Beweis. Wir zeigen dies in mehreren Schritten:

1. Für $s \geq 2$ ist $\frac{1}{n^s} \leq \frac{1}{n^2}$. Also konvergieren alle Reihen $\sum_{n=1}^{\infty} \frac{1}{n^s}$ absolut nach dem Majorantenkriterium, da $\sum_{n=1}^{\infty} \frac{1}{n^2}$ konvergiert.

2. Die Reihe
$$\frac{1}{1-\frac{1}{p^s}} = \sum_{l=0}^{\infty} (\frac{1}{p^s})^l$$
konvergiert absolut für jede Primzahl p, da $(\frac{1}{p})^s < \frac{1}{p} < 1$ gilt und die Reihe daher eine geometrische Reihe darstellt.

3. Das endliche Produkt
$$\prod_{k=1}^{r} \sum_{e=0}^{N} (\frac{1}{p^e})^s = \sum_{\substack{n \in \mathbb{N},\ n \text{ hat die} \\ \text{Primfaktoren } p_1, \dots, p_r \\ \text{mit Multiplizität } \leq N}} \frac{1}{n^s}$$
nach dem Satz über die eindeutige Primfaktorzerlegung. Es folgt, indem wir $N \to \infty$ betrachten:
$$\prod_{k=1}^{r} \frac{1}{1-p^{-s}} = \sum_{\substack{n \in \mathbb{N},\ n \text{ hat die} \\ \text{Primfaktoren } p_1, \dots, p_r}} \frac{1}{n^s}$$
nach dem großen Umordnungssatz. Schließlich ergibt sich:
$$\prod_{k=1}^{\infty} \frac{1}{1-p_k^{-s}} = \sum_{n=1}^{\infty} \frac{1}{n^s},$$
nochmals wegen des Großen Umordnungssatzes.

□

Korollar 6.27. *Sei N eine große Zahl und ω_N die Wahrscheinlichkeit, dass zwei zufällig gewählte Zahlen $a, b \in \mathbb{N}$ mit $0 < a \leq N$, $0 < b \leq N$ keinen gemeinsamen Faktor haben. Dann gilt:*
$$\lim_{N\to\infty} \omega_N = \frac{6}{\pi^2} = 0.60792\cdots \approx 60\%.$$

Beweis (nur Beweisidee). Als gemeinsame Primfaktoren kommen nur Primzahlen $p \leq N$ in Frage. Für $p \ll N$ ist die Wahrscheinlichkeit, dass p ein Teiler eines zufällig gewählten $a \in \{1, \dots, N\}$ $\gtrapprox \frac{p-1}{p} = 1 - \frac{1}{p}$. Die Wahrscheinlichkeit, dass a und b beide p als Teiler haben, ist $\gtrapprox \frac{p^2-1}{p^2} = 1 - \frac{1}{p^2}$. Es folgt: $\omega_N \gtrapprox \prod_{p\leq N}(1 - \frac{1}{p^2})$ bzw. $\prod_{p\leq N} \frac{1}{1-\frac{1}{p^2}} \gtrapprox \omega_N^{-1}$. Daher:
$$\lim_{N\to\infty} \frac{1}{\omega_N} = \prod_{p \text{ Primzahl}} \frac{1}{1-\frac{1}{p^2}} \overset{\text{Euler}}{=} \sum_{n=1}^{\infty} \frac{1}{n^2} \overset{\text{Fourierreihen}}{=} \frac{\pi^2}{6}.$$

Es folgt:

$$\lim_{N\to\infty} \omega_N = \frac{6}{\pi^2} = 0.60792\ldots$$

Leider können wir den Teil, der Fourierreihen verwendet, hier noch nicht erklären, siehe dazu wiederum Kapitel 27 bzw. genauer Korollar 27.8. □

Aufgaben

Aufgabe 6.1 (Grenzwerte von Reihen). Untersuchen Sie folgende Reihen auf Konvergenz und geben Sie, falls er existiert, den Grenzwert an:

1. $\sum_{n=0}^{\infty} \frac{3^n}{4^{n+1}}$
2. $\sum_{n=2}^{\infty} \frac{2}{n^2-1}$

Aufgabe 6.2 (Konvergenz von Reihen). Untersuchen Sie, ob folgende Reihen konvergieren:

1. $\sum_{n=1}^{\infty} \frac{n+4}{n^2-3n+1}$
2. $\sum_{n=1}^{\infty} \frac{n!}{n^n}$

Aufgabe 6.3 (Konvergenz von Reihen). Sei (a_n) eine monoton fallende Folge in $\mathbb{R}_{>0}$.

Zeigen Sie: $\sum_{n=0}^{\infty} a_n$ konvergiert genau dann, wenn $\sum_{k=0}^{\infty} 2^k a_{2^k}$ konvergiert.

Aufgabe 6.4 (Umordnung). Sei $\sum_{n=0}^{\infty} a_n$ eine konvergente, aber nicht absolut konvergente Reihe. Zeigen Sie:

1. Die Teilreihe der positiven Glieder wächst unbeschränkt, die Teilreihe der negativen Glieder fällt unbeschränkt.
2. Für jede reelle Zahl $a \in \mathbb{R}$ gibt es eine Umordnung $\tau\colon \mathbb{N}_0 \to \mathbb{N}_0$, so dass die Reihe $\sum_{n=0}^{\infty} a_{\tau(n)}$ den Grenzwert a hat.

Aufgabe 6.5 (Konvergenz von Reihen). Untersuchen Sie folgende Reihen auf Konvergenz:

1. $\sum_{n=0}^{\infty} \frac{(-1)^n}{3-2n}$
2. $\sum_{n=1}^{\infty} \frac{(1+(-1)^n \cdot \frac{1}{2})^n}{n^2}$
3. $\sum_{n=1}^{\infty} \frac{n^4}{3^n}$

Aufgabe 6.6 (Reihen: Konvergenz / Grenzwerte). Für welche $\alpha > 0$ konvergiert die Reihe

$$\sum_{n=2}^{\infty} \frac{1}{n(\ln n)^{\alpha}} \quad ?$$

Hinweis: Integralkriterium.

7

Potenzreihen

Viele aus der Schule bekannte Funktionen wie exp, sin, cos werden am Besten über Potenzreihen definiert. Hier werden wir die wichtigsten Eigenschaften dieser Reihen kennen lernen. Insbesondere zählen dazu Resultate über deren Konvergenz und Umordnungsmöglichkeiten. Da dies über den reellen Zahlen nicht gut zu behandeln ist, beginnen wir nach ersten Beispielen mit einer Einführung in die sogenannten komplexen Zahlen.

Definition 7.1. *Sei $(a_n)_{n\in\mathbb{N}_0}$ eine Folge reeller Zahlen und $x \in \mathbb{R}$. Eine Reihe der Gestalt $\sum_{n=0}^{\infty} a_n x^n$ heißt **Potenzreihe**.*

Beispiel 7.2. $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$, falls $|x| < 1$.

Potenzreihen werden häufig herangezogen, um Funktionen zu definieren:

Beispiel 7.3. Man definiert die **Exponentialfunktion**

$$\exp\colon \mathbb{R} \to \mathbb{R}$$

durch die Potenzreihe

$$\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Wir müssen uns noch überlegen, dass diese Reihe für jedes $x \in \mathbb{R}$ konvergiert. Mit dem Quotientenkriterium

$$\left| \frac{\frac{x^{n+1}}{(n+1)!}}{\frac{x^n}{n!}} \right| = \frac{|x|}{n+1} \underset{n\to\infty}{\longrightarrow} 0$$

folgt dies leicht.

Sinus und **Cosinus** (Abb. 7.7) definiert man durch die Formeln

$$\sin(x) := \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$
$$\cos(x) := \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

Der Bereich, in dem Konvergenzreihen konvergieren, hat eine einfache geometrische Beschreibung. Am prägnantesten wird diese, wenn wir auch komplexe Zahlen betrachten.

7.1 Komplexe Zahlen

Definition 7.4. *Die **komplexen Zahlen** $\mathbb{C}$ sind als Menge definiert durch $\mathbb{C} = \mathbb{R}^2$. Addition und Multiplikation sind auf $\mathbb{C}$ wie folgt erklärt (siehe auch Abb. 7.1):*

$$(a, b) + (c, d) := (a + c, b + d),$$
$$(a, b) \cdot (c, d) := (ac - bd, ad + bc).$$

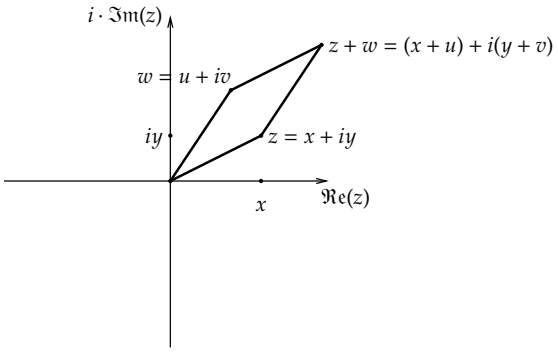


Abbildung 7.1. Die Addition komplexer Zahlen.

Das Nullelement ist damit $0 = (0, 0)$ und das Einselement der Multiplikation ist

$$1 = (1, 0) \in \mathbb{C}.$$

Das Element

$$i := (0, 1) \in \mathbb{C}$$

ist dann ein Element mit

$$i^2 = (-1, 0) = -1$$

und heißt ***imaginäre Einheit****. Jede komplexe Zahl hat damit die eindeutige Darstellung*

$$z = x + iy = (x, y) \text{ mit } x, y \in \mathbb{R}.$$

x heißt ***Realteil*** *von z und y* ***Imaginärteil****. Notation:*

$$x = \mathfrak{Re}(z), \quad y = \mathfrak{Im}(z).$$

Für das Rechnen mit komplexen Zahlen muss man sich nur $i^2 = -1$ merken und distributiv ausmultiplizieren:

$$\begin{aligned}(a + ib)(c + id) &= ac + ibc + aid + ibid \\ &= ac + i(bc + ad) + i^2bd \\ &= (ac - bd) + i(bc + ad).\end{aligned}$$

Satz 7.5. $(\mathbb{C}, +, \cdot)$ *ist ein Körper.*

Beweis. Nur die Existenz der multiplikativen Inversen ist nicht völlig trivial nachzurechnen. Sei also $z = x + iy \in \mathbb{C}$, $z \neq 0$, d.h. $(x, y) \neq (0, 0)$. Was ist $z^{-1} = \frac{1}{z}$?

$$\frac{1}{z} = \frac{1}{x + iy} \cdot \frac{x - iy}{x - iy} = \frac{x - iy}{x^2 + y^2},$$

also:

$$\mathfrak{Re}(\frac{1}{z}) = \frac{x}{x^2 + y^2}, \quad \mathfrak{Im}(\frac{1}{z}) = \frac{-y}{x^2 + y^2}.$$

In der Tat gilt:

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

□

Definition 7.6. *Für $z = x + iy$ heißt*

$$\bar{z} = x - iy$$

die ***konjugiert komplexe*** *Zahl (die Abb. $z \mapsto \bar{z}$ heißt entsprechend* ***komplexe Konjugation****) und*

$$|z| = \sqrt{x^2 + y^2} \in \mathbb{R}_{\geq 0}$$

der ***Betrag*** *von z (s. Abb. 7.2).*

Proposition 7.7 (Rechenregeln für komplexe Zahlen). *Seien $z = x + iy$, $w = u + iv \in \mathbb{C}$. Dann gilt:*

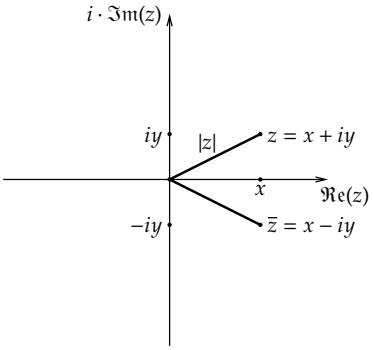


Abbildung 7.2. Die konjugiert komplexe Zahl.

1. $\Re\mathfrak{e}(z) = \frac{1}{2}(z + \overline{z})$, $\Im\mathfrak{m}(z) = \frac{1}{2i}(z - \overline{z})$.
2. $|z|^2 = z \cdot \overline{z}$.
3. *Eigenschaften des Betrags. Für $z, w \in \mathbb{C}$ gilt:*
   a) $|z| \geq 0$, *außerdem:* $|z| = 0 \iff z = 0$,
   b) $|z \cdot w| = |z| \cdot |w|$,
   c) $|z + w| \leq |z| + |w|$. *(Δ–Ungleichung, s. Abb. 7.3)*

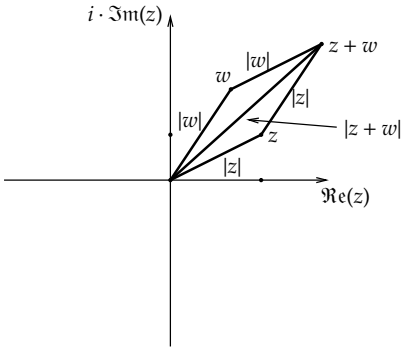


Abbildung 7.3. Eigenschaften des Betrags komplexer Zahlen. Das Bild veranschaulicht die Dreiecksungleichung $|z + w| \leq |z| + |w|$.

Beweis. Wir zeigen nur die Δ–Ungleichung, die anderen Regeln sind einfach nachzuweisen:

$$\begin{aligned}
|z+w|^2 &= (z+w)(\overline{z}+\overline{w}) \\
&= z\overline{z} + \underbrace{w\overline{z} + z\overline{w}}_{2\cdot\Re\mathrm{e}(w\overline{z})} + w\overline{w} \\
&= z\overline{z} + 2\Re\mathrm{e}(w\overline{z}) + w\overline{w} \\
&\leq |z|^2 + 2|z||w| + |w|^2 \\
&= (|z|+|w|)^2.
\end{aligned}$$

Da die Wurzelfunktion monoton ist, folgt: $|z+w| \leq |z| + |w|$. □

Definition 7.8. *Eine Folge (z_n) komplexer Zahlen **konvergiert** gegen $z \in \mathbb{C}$, falls*

$$\forall \varepsilon > 0 \; \exists \, n_0 : \; |z_n - z| < \varepsilon \; \forall n \geq n_0.$$

Bemerkung 7.9. Äquivalent dazu, dass die Folge (z_n) komplexer Zahlen den Grenzwert $z \in \mathbb{C}$ hat, ist:

$$\lim_{n\to\infty} \Re\mathrm{e}(z_n) = \Re\mathrm{e}(z) \quad \text{und} \quad \lim_{n\to\infty} \Im\mathrm{m}(z_n) = \Im\mathrm{m}(z).$$

Beweis. Für eine beliebige Zahl $w \in \mathbb{C}$ gilt, wegen der Dreiecks–Ungleichung und da die Diagonale in einem Quadrat mit Seitenlänge a gerade $\sqrt{2}$–mal so lang ist wie die Seite:

$$\sqrt{2}\max\Big\{|\Re\mathrm{e}(w)|, |\Im\mathrm{m}(w)|\Big\} \geq |w| \geq \max\Big\{|\Re\mathrm{e}(w)|, |\Im\mathrm{m}(w)|\Big\},$$

wie die Abbildung 7.4 verdeutlicht. Daraus folgt die Behauptung. □

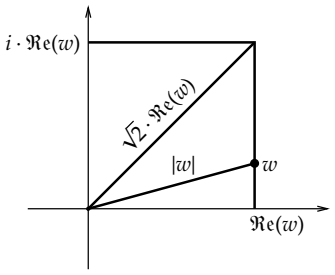


Abbildung 7.4. Obere und untere Schranke für den Betrag einer komplexen Zahl: $\sqrt{2}\max\{|\Re\mathrm{e}(w)|, |\Im\mathrm{m}(w)|\} \geq |w| \geq \max\{|\Re\mathrm{e}(w)|, |\Im\mathrm{m}(w)|\}$. Im Bild ist der Fall $|\Re\mathrm{e}(w)| \geq \Im\mathrm{m}(w)$ veranschaulicht.

Satz/Definition 7.10. *$\mathbb{C}$ ist **vollständig**, d.h. jede Cauchy–Folge komplexer Zahlen konvergiert gegen ein $z \in \mathbb{C}$.*

Formal: Ist (z_n) eine Folge komplexer Zahlen, so gilt:

$$\forall \varepsilon > 0 \; \exists \, n_0 : |z_n - z_m| < \varepsilon \; \forall n, m \geq n_0 \; \Rightarrow \; \exists z \in \mathbb{C} : \lim_{n \to \infty} z_n = z \in \mathbb{C}.$$

Beweis. (z_n) ist eine Cauchy–Folge $\iff$ $(\mathfrak{Re}(z_n))$ und $(\mathfrak{Im}(z_n))$ bilden Cauchy–Folgen reeller Zahlen. Sie konvergieren jeweils gegen x bzw. y. Der Grenzwert der Folge (z_n) ist daher $z = x + y$. □

Bemerkung 7.11. $\mathbb{C}$ lässt sich nicht anordnen.

Beweis. Angenommen, $>$ sei eine Anordnung auf $\mathbb{C}$. Da Quadrate immer ≥ 0 sind, folgt: $i^2 = -1 > 0$ und $1 = 1^2 > 0 \; \Rightarrow \; -1 + 1 = 0 > 0$, was nicht sein kann. □

Ein ganz wesentlicher Grund, aus dem man komplexe Zahlen in der Mathematik betrachtet, ist das folgende Resultat:

Satz/Definition 7.12 (Fundamentalsatz der Algebra, ohne Beweis). *Sei $p(z) = a_n z^n + \cdots + a_1 z = a_0$ ein **Polynom** mit $a_i \in \mathbb{C}$ vom **Grad** n, d.h. $a_n \neq 0$. Dann hat p eine **Nullstelle**, d.h. $\exists z_1 \in \mathbb{C} : \; p(z_1) = 0$.*

Für den Grad von p aus dem Satz schreiben wir auch $\deg(p) := n$. Die Menge aller Polynome in einer Variablen z und Koeffizienten in einem Körper K bezeichnen wir mit $K[z]$, also hier $p \in \mathbb{C}[z]$. Die Teilmenge der Polynome in einer Variablen z mit Koeffizienten in K vom Grad $\leq n$ schreiben wir $K[z]_{\leq n}$.

Wir geben für diesen Satz in dieser Vorlesung keinen Beweis. Man kann aber leicht einsehen, dass aus der Existenz einer Nullstelle induktiv schon folgt:

Korollar 7.13. *Jedes Polynom $p(z) = a_n z^n + \cdots + a_1 z = a_0$ vom Grad $n > 0$ mit $a_i \in \mathbb{C}$ faktorisiert in **Linearfaktoren**:*

$$p(z) = a_n \cdot (z - z_1) \cdot (z - z_n),$$

für gewisse $z_i \in \mathbb{C}$, wobei die z_i nicht unbedingt paarweise verschieden sein müssen.

In der Physik ist ein Hauptgrund, komplexe Zahlen zu verwenden, dass sich die Quantenmechanik ohne komplexe Zahlen nicht beschreiben lässt.

7.2 Der Konvergenzradius

Satz 7.14. *Sei $\sum_{n=0}^{\infty} a_n z^n$ eine Potenzreihe. Wenn diese Reihe für ein $z_0 \in \mathbb{C} \backslash \{0\}$ konvergiert, dann konvergiert die Reihe für alle $z \in \{z \in \mathbb{C} \mid |z| < |z_0|\}$ absolut.*

Beweis. Wir verwenden das Majorantenkriterium. Da die Reihe $\sum_{n=0}^{\infty} a_n z_0^n$ konvergiert, bildet die Folge $(a_n z_0^n)$ eine Nullfolge. Sie ist daher beschränkt, etwa $|a_n z_0^n| \leq M \; \forall n \geq 0$. Für $z \in \mathbb{C}$ mit $|z| < |z_0|$ ergibt sich:

$$|a_n z^n| = |a_n| \cdot |z_0^n| \cdot \left|\frac{z}{z_0}\right|^n \leq M \cdot \left|\frac{z}{z_0}\right|^n.$$

Da $|\frac{z}{z_0}| < 1$ gilt, ist die geometrische Reihe $\sum_{n=0}^{\infty} M \cdot |\frac{z}{z_0}|^n$ eine konvergente Majorante. □

Definition 7.15. *Sei $\sum_{n=0}^{\infty} a_n z^n$ eine Potenzreihe. Dann heißt*

$$R := \sup\left\{ |z_0| \;\middle|\; \sum_{n=0}^{\infty} a_n z_0^n \text{ konvergiert} \right\} \in [0, \infty]$$

($\sup A = \infty$, falls A nach oben nicht beschränkt ist) der ***Konvergenzradius*** *der Potenzreihe. Es gilt: Die Reihe konvergiert für alle $z \in \{z \in \mathbb{C} \mid |z| < R\}$ und divergiert für alle $z \in \{z \in \mathbb{C} \mid |z| > R\}$ wegen Satz 7.14.*

Auf dem Kreisrand $\{z \in \mathbb{C} \mid |z| = R\}$ kann Konvergenz vorliegen, muss aber nicht (Abb. 7.5).

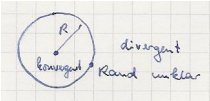


Abbildung 7.5. Der Konvergenzradius einer Potenzreihe.

Beispiel 7.16.

1. $\sum_{n=0}^{\infty} z^n$ hat Konvergenzradius $R = 1$.
2. Die Reihe $\sum_{n=1}^{\infty} \frac{z^n}{n}$ konvergiert im Punkt $z_0 = -1$ und divergiert für $z_1 = 1$ (alternierende) harmonische Reihe, daher folgt: $R = 1$.
3. $\sum_{n=0}^{\infty} \frac{z^n}{n!}$ hat Konvergenzradius $R = \infty$, da sie für beliebig große $x \in \mathbb{R}$ konvergiert.

Satz 7.17. *Sei $(a_n)_{n \in \mathbb{N}_0}$ eine Folge komplexer Zahlen mit $a_n \neq 0 \; \forall n$. Existiert der Grenzwert $q = \lim_{n \to \infty} \left|\frac{a_{n+1}}{a_n}\right|$, so hat die Potenzreihe $\sum_{n=0}^{\infty} a_n z^n$ den Konvergenzradius*

$$R = \begin{cases} \frac{1}{q}, & \text{falls } q > 0, \\ \infty, & \text{falls } q = 0. \end{cases}$$

Beweis. Quotientenkriterium. □

Die obige Formel ist nicht immer anwendbar (z.B. beim Sinus). Eine Formel, die dagegen immer funktioniert, ist folgende:

Definition 7.18. *Sei (a_n) eine Folge reeller Zahlen. Dann heißt*

$$\limsup_{n\to\infty}(b_n) := \lim_{n\to\infty} \sup\left\{ b_k \;\middle|\; k \geq n \right\}$$

*der **Limes Superior** von (b_n). Ist (b_n) nach oben nicht beschränkt, so setzen wir:*

$$\limsup b_n = +\infty.$$

Analog ist

$$\liminf_{n\to\infty} b_n := \lim_{n\to\infty} \inf\{b_k \mid k \geq n\},$$

*der **Limes Inferior**, erklärt.*

Satz 7.19 (Formel von Cauchy–Hadamard). *Seien $\sum_{n=0}^{\infty} a_n z^n$ eine Potenzreihe und $q = \limsup_{n\to\infty}(\sqrt[n]{|a_n|})$. Dann hat die Potenzreihe $\sum_{n=0}^{\infty} a_n z^n$ den Konvergenzradius*

$$R = \begin{cases} 0, & \text{falls } q = \infty, \\ \frac{1}{q}, & \text{falls } 0 < q < \infty, \\ \infty, & \text{falls } q = 0. \end{cases}$$

Beweis. Wurzelkriterium. □

7.3 Der Umordnungssatz

Satz 7.20. *Sei $\sum_{n=1}^{\infty} a_n$ eine Reihe komplexer Zahlen.*

1. *Ist $\sum_{n=1}^{\infty} a_n$ absolut konvergent und ist $\tau\colon \mathbb{N} \to \mathbb{N}$ eine bijektive Abbildung, dann ist auch die Reihe*

$$\sum_{n=0}^{\infty} a_{\tau(n)}$$

absolut konvergent und es gilt:

$$\sum_{n=0}^{\infty} a_{\tau(n)} = \sum_{n=0}^{\infty} a_n.$$

2. ***(Großer Umordnungssatz)*** *Sei $(I_k)_{k\in\mathbb{N}}$ eine Familie von endlichen oder unendlich disjunkten Teilmengen $I_k \subset \mathbb{N}$ mit $\dot{\bigcup}_{k=1}^{\infty} I_k = \mathbb{N}$. Dann ist für jedes k die Reihe*

$$s_k := \sum_{j\in I_k} a_j$$

absolut konvergent und die Reihe der Grenzwerte $\sum_{k=1}^{n} s_k$ ebenfalls und zwar mit Grenzwert

$$\sum_{k=1}^{\infty} s_k = \sum_{n=1}^{\infty} a_n.$$

Beweis. Die Aussage von 1. ist ein Spezialfall von 2. mit $I_k = \{\tau(k)\}$; wir müssen also nur 2. beweisen. Zunächst zur absoluten Konvergenz von

$$\sum_{j\in I_k} a_j := \lim_{N\to\infty}\Big(\sum_{j\in I_k, j\le N} a_j\Big),$$

wobei wir in einer beliebigen Reihenfolge summieren. Im Fall von I_k endlich und bei der 1. Teilaussage ist dies klar. Für den anderen Fall verwenden wir, dass 1. schon gezeigt ist.

Da die Partialsummen

$$\sum_{j=I'} |a_j| \le \sum_{j\in I''} |a_j|$$

monoton steigen für endliche Teilmengen $I' \subset I'' \subset I_k$, genügt es zu zeigen, dass sie beschränkt bleiben. Dies ist klar, da

$$\sum_{j\in I'} |a_j| \le \sum_{n=0}^{N} |a_n| \le \sum_{n=0}^{\infty} |a_n| < \infty,$$

wobei $N = \max\{j \mid j \in I'\}$. Für die absolute Konvergenz von $\sum_{k=1}^{\infty} s_k$ gehen wir genauso vor:

$$\begin{aligned}
\sum_{k=1}^{l} |s_k| &= \sum_{k=1}^{l} \lim_{N\to\infty} \Big|\sum_{j\in I_k, j\le N} a_j\Big| \\
&\le \lim_{N\to\infty} \sum_{k=1}^{l} \sum_{j\in I_k, j\le N} |a_j| \\
&= \lim_{N\to\infty} \sum_{j\in\bigcup_{k=1}^{l} I_k} |a_j| \\
&\le \lim_{N\to\infty} \sum_{j=1}^{N} |a_j| = \sum_{n=1}^{\infty} |a_n| < \infty.
\end{aligned}$$

Für die Gleichheit der Grenzwerte betrachten wir $s = \sum_{n=1}^{\infty} a_n$ und zu $\varepsilon > 0$ ein n_0, so dass

$$\sum_{n=n_0}^{\infty} |a_n| < \varepsilon \text{ und } \Big| s - \sum_{n=1}^{n_0-1} a_n \Big| < \varepsilon$$

gilt. Sei nun

$$k_0 := \max\Big\{ k \;\Big|\; \{1, \dots, n_0 - 1\} \cap I_k \neq \emptyset \Big\}.$$

Dann gilt für $k_1 \geq k_0$:

$$\begin{aligned} \Big| s - \sum_{k=1}^{k_1} s_k \Big| &\leq \Big| s - \sum_{n=1}^{n_0-1} a_n \Big| + \sum_{k=1}^{k_1} \sum_{n \in I_k, n \geq n_0} |a_n| \\ &\leq \Big| s - \sum_{n=1}^{n_0-1} a_n \Big| + \sum_{n \geq n_0} |a_n| < 2\varepsilon. \end{aligned}$$

□

7.4 Die komplexe Exponentialfunktion

Eine der wichtigsten Potenzreihen ist jene, die die sogenannte Exponentialfunktion definiert. Beispielsweise existiert ein interessanter Zusammenhang zu Sinus und Cosinus.

Definition 7.21. *Die Abbildung*

$$\exp\colon \mathbb{C} \to \mathbb{C},\ z \mapsto \exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

heißt ***komplexe Exponentialfunktion****.*

Satz 7.22 (Funktionalgleichung der Exponentialfunktion). *Für $z, w \in \mathbb{C}$ gilt:*

$$\exp(z + w) = \exp(z) \cdot \exp(w).$$

Beweis. Wir betrachten das Cauchy–Produkt der absolut konvergenten Reihen

$$\sum_{k=0}^{\infty} \frac{z^k}{k!} \text{ und } \sum_{n=0}^{\infty} \frac{w^n}{n!}.$$

Es ergibt sich mit der binomischen Formel:

$$d_n = \sum_{k=0}^{n} \frac{z^k}{k!} \cdot \frac{w^{n-k}}{(n-k)!} = \sum_{k=0}^{n} \binom{n}{k} \frac{z^k \cdot w^{n-k}}{n!} = \frac{(z+w)^n}{n!}.$$

Nach dem großen Umordnungssatz gilt nun:

$$\begin{aligned}\exp(z+w) &= \sum_{n=0}^{\infty} \frac{(z+w)^n}{n!} \\ &= \Big(\sum_{k=0}^{\infty} \frac{z^k}{k!}\Big) \cdot \Big(\sum_{l=0}^{\infty} \frac{w^l}{l!}\Big) \\ &= \exp(z) \cdot \exp(w).\end{aligned}$$

□

Korollar 7.23. *Es gilt:*

1. $\exp(0) = 1$,
2. $\exp(-z) = \frac{1}{\exp(z)}$ *und daher insbesondere* $\exp(z) \in \mathbb{C}^* \ \forall z \in \mathbb{C}$, *wobei* $\mathbb{C}^* := \mathbb{C}\backslash\{0\}$. *Die komplexe Exponentialfunktion ist also eine Abbildung* $\mathbb{C} \to \mathbb{C}^*$ *(siehe auch Abb. 7.6).*

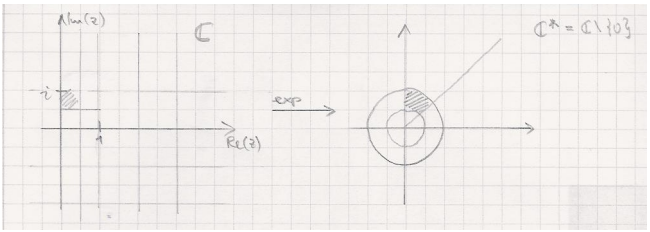


Abbildung 7.6. Die Wirkung von exp auf $\mathbb{C}$.

Beweis. $1 = \exp(0) = \exp(z + (-z)) = \exp(z) \cdot \exp(-z)$. □

Setzen wir für z einen **rein imaginären** Wert ein, d.h. $z = iy$, $y \in \mathbb{R}$, so erhalten wir:

$$\begin{aligned}\exp(iy) &= \sum_{n=0}^{\infty} \frac{(iy)^n}{n!} \\ &= \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k}}{2k!} \ + \ i \cdot \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k+1}}{(2k+1)!} \\ &= \cos(y) + i \sin(y),\end{aligned}$$

also einen Zusammenhang zwischen der komplexen Exponentialfunktion und Sinus und Cosinus. Man schreibt häufig auch

$$e^z := \exp(z)$$

bzw. für reelle $x \in \mathbb{R}$ entsprechend e^x. Die Zahl $e = e^1 = \exp(1)$ wird auch **Eulersche Zahl** genannt. Mit dieser Notation gilt:

Satz 7.24 (Zusammenhang zwischen der komplexen Exponentialfunktion und Sinus und Cosinus). *Für $z = x + iy \in \mathbb{C}$ gilt:*

$$\exp(x + iy) = e^x \cdot (\cos y + i \sin y).$$

Die Additionstheoreme für Sinus und Cosinus folgen mit Hilfe der obigen Formel aus denen der Exponentialfunktion:

Satz 7.25 (Additionstheoreme für Sinus und Cosinus). *Seien $\alpha, \beta \in \mathbb{R}$. Dann:*

$$\begin{aligned}\cos(\alpha + \beta) &= \cos\alpha \cdot \cos\beta - \sin\alpha \cdot \sin\beta,\\ \sin(\alpha + \beta) &= \sin\alpha \cdot \cos\beta + \cos\alpha \cdot \sin\beta.\end{aligned}$$

Insbesondere gilt (mit der Notation $\sin^k \alpha := (\sin\alpha)^k$ und entsprechend für den $\cos$):

$$1 = \sin^2\alpha + \cos^2\alpha.$$

Beweis. Es gilt:

$$\begin{aligned}\cos(\alpha + \beta) + i\sin(\alpha + \beta) &= \exp(i(\alpha + \beta))\\ &= \exp(i\alpha) \cdot \exp(i\beta)\\ &= (\cos\alpha + i\sin\alpha) \cdot (\cos\beta + i\sin\beta)\\ &= (\cos\alpha\cos\beta - \sin\alpha\sin\beta)\\ &\qquad + i(\cos\alpha\sin\beta + \sin\alpha\cos\beta),\end{aligned}$$

wobei sich die letzte Gleichheit gemäß der Definition der Multiplikation in $\mathbb{C}$ ergibt. Realteil und Imaginärteil dieser Formel ergeben die Behauptung.

Für den Zusatz betrachten wir

$$1 = \cos(0) = \cos\Big(\alpha + (-\alpha)\Big) = \cos(\alpha)\cos(-\alpha) - \sin(\alpha)\sin(-\alpha).$$

Da $\cos(-\alpha) = \cos(\alpha)$ und $\sin(-\alpha) = -\sin(\alpha)$ gilt, weil die Potenzreihe des Cosinus nur gerade Terme und jene des Sinus nur ungerade Terme hat, folgt $1 = \cos^2\alpha + \sin^2\alpha$, wie behauptet. □

Mit Hilfe des Satzes von Pythagoras kann man Sinus und Cosinus nun auf dem Einheitskreis einzeichnen (Abb. 7.7).

Die Multiplikation der komplexen Zahlen ergibt sich direkt. Seien dazu $z, w \in \mathbb{C}$. Dann existiert ein Winkel φ, genannt **Argument** von z und entsprechend ψ, so dass

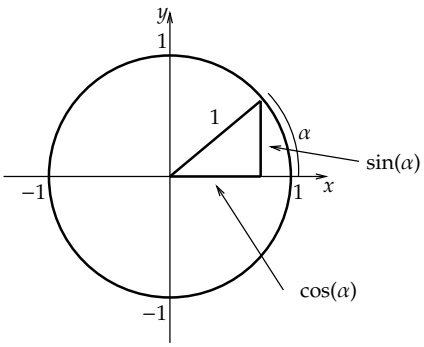


Abbildung 7.7. Sinus und Cosinus am Einheitskreis; α ist im Bogenmaß eingezeichnet.

$$z = |z| \cdot (\cos\varphi + i \sin\varphi) \quad \text{und} \quad w = |w| \cdot (\cos\psi + i \sin\psi).$$

Damit erhalten wir mit Hilfe der Additionstheoreme für Sinus und Cosinus für das Produkt von z und w:

$$\begin{aligned} z{\cdot}w &= |z{\cdot}w| \cdot (\cos\varphi + i \sin\varphi) \cdot (\cos\psi + i \sin\psi) \\ &= |z{\cdot}w| \cdot \Big((\cos\varphi \cdot \cos\psi - \sin\varphi \cdot \sin\psi) + i(\sin\varphi \cdot \cos\psi + \cos\varphi \cdot \sin\psi)\Big) \\ &= |z| \cdot |w| \cdot \Big(\cos(\varphi + \psi) + i \sin(\varphi + \psi)\Big). \end{aligned}$$

Mit anderen Worten:

Bemerkung 7.26. Bei der Multiplikation zweier komplexer Zahlen multipliziert sich der Betrag und es addieren sich die Argumente (Abb. 7.8).

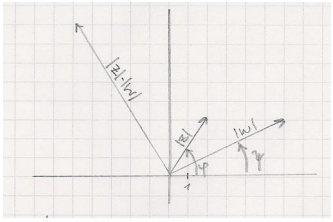


Abbildung 7.8. Multiplikation zweier komplexer Zahlen.

Aufgaben

Aufgabe 7.1 (Komplexe Zahlen). Bestimmen und zeichnen Sie für $r = \frac{1}{2}$, $r = 1$ und $r = 2$ jeweils die Menge:

$$\left\{z \in \mathbb{C} : \left|\frac{z-1}{z+1}\right| < r\right\}.$$

Aufgabe 7.2 (Grenzwertvertauschung, $\limsup$ und $\liminf$).

1. Sei
$$a_{j,k} = \begin{cases} \frac{k}{j} & j \geq k, \\ \frac{k-j}{k} & j < k. \end{cases}$$
Bestimmen Sie:
$$\lim_{k\to\infty} \lim_{j\to\infty} a_{j,k} \quad \text{und} \quad \lim_{j\to\infty} \lim_{k\to\infty} a_{j,k}.$$

2. Sei (a_n) die Folge mit $a_n := (-1)^n + \frac{1}{n}$. Berechnen Sie
$$\limsup_{n\to\infty} a_n \quad \text{und} \quad \liminf_{n\to\infty} a_n.$$

Aufgabe 7.3 (Kombinatorik).

1. Wir definieren a_k durch:
$$(1 + x^2 + x^3)^{10} = \sum_{k=0}^{30} a_k x^k.$$
Zeigen Sie: a_k ist die Anzahl der Möglichkeiten, k identische Kugeln in 10 Urnen so zu verteilen, dass in jeder Urne anschließend 2, 3 oder keine Kugel liegt.
2. Bestimmen Sie a_{20} mit Hilfe von Maple.

Aufgabe 7.4 (Konvergenzradien). Bestimmen Sie die Konvergenzradien der folgenden Reihen:

1. $\sum_{n=1}^{\infty} \frac{x^n}{n^2}$
2. $\sum_{n=1}^{\infty} n^2 x^n$
3. $\sum_{n=1}^{\infty} 2^n x^n$
4. $\sum_{n=1}^{\infty} \frac{x^n}{2^n}$

Können Sie den Grenzwert im Falle der Konvergenz bestimmen?

Aufgabe 7.5 (Cauchy–Produkt von Reihen). Für $n \in \mathbb{N}$ sei $a_n := b_n := (-1)^n \cdot \frac{1}{\sqrt{n+1}}$ und $c_n := \sum_{k=0}^{n} a_{n-k} b_k$. Zeigen Sie, dass die Reihen $\sum_{n=0}^{\infty} a_n$ und $\sum_{n=0}^{\infty} b_n$ konvergieren, ihr Cauchy–Produkt $\sum_{n=0}^{\infty} c_n$ aber nicht.

Aufgabe 7.6 (Additionstheoreme für Sinus und Cosinus).

1. Zeigen Sie, dass für jedes $n \in \mathbb{N}$ Polynome $p_n(x, y)$ und $q_n(x, y)$ in zwei Variablen x, y mit reellen Koeffizienten existieren, so dass

$$\sin(nt) = p_n(\sin(t), \cos(t)) \quad \text{und} \quad \cos(nt) = q_n(\sin(t), \cos(t))$$

   für alle $t \in \mathbb{R}$ gilt.

2. Berechnen Sie $p_n(x, y)$ und $q_n(x, y)$ für $n = 2, 3, 4$.

Aufgabe 7.7 (Konvergenzradius von Potenzreihen). Bestimmen Sie den Konvergenzradius der folgenden Potenzreihen:

1. $\sum_{n=1}^{\infty} (n^4 - 3n^3)x^n$,
2. $\sum_{k=1}^{\infty} \frac{3^k + (-2)^k}{k}(x + 1)^k$.

8

Stetigkeit

...

8.1 Definition und Folgenkriterium

Definition 8.1. *Sei $D \subseteq \mathbb{R}$. Eine **reellwertige Funktion** auf D ist eine Abbildung*

$$f : D \to \mathbb{R}.$$

*D heißt **Definitionsbereich** von f. Typischerweise ist D ein Intervall oder eine Vereinigung von Intervallen.*

Die Menge

$$G_f := \left\{ (x, y) \;\middle|\; x \in D, y = f(x) \right\} \subset \mathbb{R}^2$$

*heißt **Graph** der Funktion.*

Beispiel 8.2.

1. $y = f(x) = x^2$, siehe Abb. 8.1.
2. $y = \lfloor x \rfloor = \text{entier}(x)$, siehe Abb. 8.2.

Im zweiten Beispiel hat der Graph „Sprünge"; stetige Funktionen sind im Wesentlichen solche, für die das nicht der Fall ist. Präzise definieren wir dies wie folgt:

Definition 8.3. *Sei $f : D \to \mathbb{R}$ eine Funktion und $x_0 \in D$ ein Punkt. f heißt **stetig in** x_0, wenn*

$$\forall \varepsilon > 0 \; \exists \delta > 0 : \; |f(x) - f(x_0)| < \varepsilon \; \forall x \in D \text{ mit } |x - x_0| < \delta$$

*gilt. f heißt **stetig auf** D, wenn f in allen Punkten $x_0 \in D$ stetig ist.*

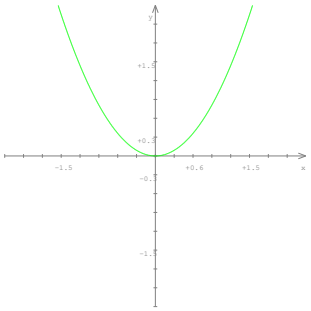


Abbildung 8.1. Graph einer Parabel mit Gleichung $f(x) = x^2$.

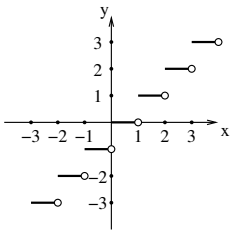


Abbildung 8.2. Graph der entier Funktion. Ein kleiner, leerer Kreis zeigt dabei an, dass der umkreiste Punkt nicht zum Graphen gehört.

Im Englischen heißt stetig *continuous*; auch im Deutschen werden wir gelegentlich den Begriff **kontinuierlich** statt stetig verwenden.

Beispiel 8.4.

1. $f(x) = x$ ist stetig. Zu ε können wir $\delta = \varepsilon$ wählen.
2. $f(x) = x^2$ ist stetig in allen Punkten. Die wesentliche Abschätzung ist

$$\begin{aligned} |x^2 - x_0^2| &= |x + x_0| \cdot |x - x_0| \\ &\leq |2x_0 + 1| \cdot |x - x_0| \quad \forall x \text{ mit } |x - x_0| < 1. \end{aligned}$$

   Entsprechend ergibt sich $|x^2 - x_0^2| < \varepsilon \ \forall x$ mit $|x - x_0| < \frac{\varepsilon}{2|x_0|+1}$. Also können wir

$$\delta = \min\left\{1, \frac{\varepsilon}{2|x_0| + 1}\right\}$$

   wählen. δ hängt sowohl von ε also auch von x_0 ab.
3. entier: $\mathbb{R} \to \mathbb{R}$ ist nicht stetig in $x_0 = 0$: Zu $\varepsilon = 1$ und $\delta > 0$ beliebig klein existiert ein Punkt x mit $|x - x_0| < \delta$ und $-1 < x < 0$. Für diese gilt:

$$|\operatorname{entier}(x) - \operatorname{entier}(0)| = |-1 - 0| = 1 \geq \varepsilon.$$

Allgemein gilt: entier ist in allen Punkten $x_0 \in \mathbb{R}\backslash\mathbb{Z}$ stetig und in allen Punkte $x_0 \in \mathbb{Z}$ unstetig.

4. Die konstanten Funktionen $f\colon \mathbb{R} \to \mathbb{R}$ mit $f(x) = c$ sind stetig.

Satz 8.5 (Folgenkriterium für Stetigkeit). *Sei $f\colon D \to \mathbb{R}$ eine Funktion und $x_0 \in D$ ein Punkt. f ist stetig in x_0 genau dann, wenn für alle Folgen (x_n) mit $x_n \in D$ und $\lim_{n\to\infty} x_n = x_0$ gilt:*

$$f(x_0) = \lim_{n\to\infty} f(x_n).$$

Das Folgenkriterium ist oft gut geeignet, um Unstetigkeit zu zeigen, wie wir am folgenden Beispiel sehen werden. Der Nachweis der Stetigkeit ist in der Regel einfacher mit der ε-δ-Definition.

Beispiel 8.6. Wir betrachten die Funktion (Abb. 8.3):

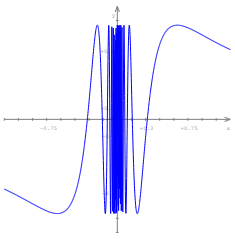


Abbildung 8.3. Die Funktion $\sin(\frac{1}{x})$ in der Nähe von 0.

$$f(x) = \begin{cases} 0, & \text{falls } x = 0, \\ \sin(\frac{1}{x}), & \text{falls } x \neq 0. \end{cases}$$

Bekanntlich gilt (wir werden in 11.14 und 11.15 die Zahl $\pi \in \mathbb{R}$ definieren und die Aussage beweisen):

$$1 = \sin\frac{\pi}{2} = \sin\Big(\frac{\pi}{2} + 2k\pi\Big)$$

für jedes $k \in \mathbb{Z}$. Also gilt für $x_k = \frac{1}{\frac{\pi}{2}+2k\pi}$ zwar

$$\lim_{k\to\infty} x_k = 0, \quad \text{aber} \quad \lim_{k\to\infty} f(x_k) = 1 \neq 0.$$

Da $\sin(\frac{3}{2}\pi + 2k\pi) = -1$ ist, gilt für $x'_k = \frac{1}{\frac{3}{2}\pi+2k\pi}$ zwar $\lim_{k\to\infty} x'_k = 0$, aber $\lim f(x'_k) = -1$.

Es folgt, dass man keinen Wert für $f(0)$ finden kann, so dass f im Nullpunkt stetig ergänzt wird.

Beweis (für das Folgenkriterium für Stetigkeit, Satz 8.5). f sei stetig in x_0 und (x_n) eine Folge in D mit $\lim x_n = x_0$. Da f in x_0 stetig ist, existiert zu $\varepsilon > 0$ ein $\delta > 0$, so dass:

$$|f(x) - f(x_0)| < \varepsilon \ \forall x \in D \text{ mit } |x - x_0| < \delta.$$

Wegen $\lim x_n = x_0$ gibt es zu $\delta > 0$ ein n_0, so dass $|x_n - x_0| < \delta \ \forall n \geq n_0$. Also:

$$|f(x_n) - f(x_0)| < \varepsilon \ \forall n \geq n_0,$$

d.h. $\lim_{n\to\infty} f(x_n) = f(x_0)$.

Umgekehrt nehmen wir nun an, f sei nicht stetig. Dann existiert ein $\varepsilon > 0$, so dass für jedes $\delta > 0$ ein $x \in D$ existiert mit $|x - x_0| < \delta$ mit $|f(x) - f(x_0)| > \varepsilon$. Wir wenden diese Aussage für alle $\delta = \frac{1}{n}$ an und erhalten eine Folge (x_n) in D mit $\lim x_n = x_0$, aber

$$|f(x_n) - f(x_0)| > \varepsilon \ \forall n.$$

Also konvergiert $(f(x_n))_{n\in\mathbb{N}}$ nicht gegen $f(x_0)$. □

Wir geben nun noch einige einfache Sätze, mit denen wir aus stetigen Funktionen weitere bilden können:

Satz 8.7 (Rechenregeln für stetige Funktionen). *Es seien $f, g\colon D \to \mathbb{R}$ Funktionen.*

1. *Sind f, g in x_0 stetig, so sind auch $f + g$ und $f \cdot g$ in x_0 stetig.*
2. *Sind f, g in x_0 stetig und ist $g(x_0) \neq 0$, dann ist auch*

$$\frac{f}{g}\colon D' \to \mathbb{R}$$

mit $D' = \{x \in D \mid g(x) \neq 0\} \subset D$ stetig in $x_0 \in D'$.

Beweis. Analog zu der entsprechenden Aussage für Grenzwerte von Folgen. □

Daraus folgt sofort:

Korollar 8.8.

1. *Polynome*
$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$
mit Konstanten $a_0, \ldots, a_n \in \mathbb{R}$ sind stetige Funktionen $f\colon \mathbb{R} \to \mathbb{R}$.
2. ***Rationale Funktionen**, d.h. Abbildungen der Form $\frac{f}{g}\colon D \to \mathbb{R}$ mit Polynomen f, g, sind stetig im Definitionsbereich $D = \{x \in \mathbb{R} \mid g(x) \neq 0\}$.*

8.2 Der Zwischenwertsatz und Anwendungen

Einer der ganz zentralen Sätze über stetige Funktionen ist folgender:

Satz 8.9 (Zwischenwertsatz). *Sei $f\colon [a,b] \to \mathbb{R}$ eine stetige Funktion und c ein Wert **zwischen** $f(a)$ und $f(b)$, d.h. $f(a) \le c \le f(b)$, falls $f(a) \le f(b)$ und $f(b) \le c \le f(a)$, falls $f(a) \ge f(b)$. Dann existert ein $\xi \in [a,b]$ (siehe auch Abb. 8.4), so dass*

$$f(\xi) = c.$$

Insbesondere folgt, dass jede stetige Funktion mit $f(a) < 0$ und $f(b) > 0$ eine Nullstelle in $[a,b]$ hat.

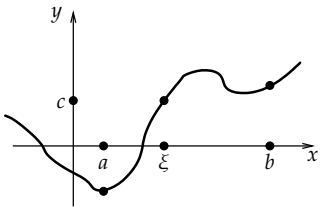


Abbildung 8.4. Offenbar ist im Bild $f(a) \le c \le f(b)$, so dass nach dem Zwischenwertsatz ein ξ mit $f(\xi) = c$ existiert.

Beweis. Indem wir zu $\pm(f(x) \pm c)$ übergehen, genügt es, die zweite Aussage zu zeigen. Sei also $f(a) < 0$ und $f(b) > 0$. Wir konstruieren induktiv monotone Folgen, die gegen die Nullstelle konvergieren mit dem sogenannten **Intervallhalbierungsalgorithmus**: Wir setzen zunächst $x_0 = a$ und $y_0 = b$. Sind x_n und y_n schon konstruiert, so betrachten wir $\overline{x} = \frac{x_n + y_n}{2}$ und $f(\overline{x})$. Dann seien

$$x_{n+1} = \begin{cases} \overline{x}, & \text{falls } f(\overline{x}) < 0, \\ x_n, & \text{sonst,} \end{cases}$$

$$y_{n+1} = \begin{cases} y_n, & \text{falls } f(\overline{x}) < 0, \\ \overline{x}, & \text{sonst.} \end{cases}$$

Dann gilt offenbar:

1. $f(x_n) < 0 \ \forall n$ und $f(y_n) \ge 0 \ \forall n$.
2. $|y_n - x_n| = 2^{-n}(b-a)$.
3. (x_n) ist monoton steigend und (y_n) monoton fallend.

Beide Folgen konvergieren also und wegen

$$\lim_{n\to\infty}(y_n - x_n) = \lim_{n\to\infty} 2^{-n}(b-a) = 0$$

ist $\xi = \lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n$. Wegen der Stetigkeit von f gilt:

$$f(\xi) = \lim_{n\to\infty} f(x_n) \leq 0,$$

da $f(x_n) < 0\ \forall n$ und

$$f(\xi) = \lim_{n\to\infty} f(y_n) \geq 0,$$

da $f(y_n) \geq 0\ \forall n$. Also folgt: $f(\xi) = 0$. □

Satz/Definition 8.10 (Existenz von Maximum und Minimum stetiger Funktionen). *Es sei $f\colon [a,b] \to \mathbb{R}$ eine stetige Funktion auf einem abgeschlossenen, beschränkten Intervall. Dann existieren $x_{\max}, x_{\min} \in [a,b]$ mit*

$$f(x_{\max}) = \sup\big\{f(x) \mid x \in [a,b]\big\},$$
$$f(x_{\min}) = \inf\big\{f(x) \mid x \in [a,b]\big\}.$$

Insbesondere ist f beschränkt.

*Wir sagen: f nimmt in $x_{\max}$ das **Maximum** an, geschrieben:*

$$\max_{x\in[a,b]} f(x) := f(x_{\max}).$$

*Analog für das **Minimum**:* $\min_{x\in[a,b]} f(x) := f(x_{\min})$.

Bemerkung 8.11. Dass $[a,b]$ abgeschlossen ist, ist wesentlich: Die Funktion $f(x) = \frac{1}{x}$ nimmt auf $]0,\infty[$ kein Maximum und Minimum an.

Beweis (der Existenz von Maximum und Minimum, Satz 8.10). Sei

$$M = \sup\big\{f(x) \mid x \in [a,b]\big\} \in \mathbb{R} \cup \{\infty\}.$$

Wir wählen eine Folge (x_n) mit $x_n \in [a,b]$, so dass $\lim_{n\to\infty} f(x_n) = M$ (bzw. $f(x_n)$ unbeschränkt wächst, falls $M = \infty$). Nach dem Satz von Bolzano–Weierstrass 5.33 hat (x_n) eine konvergente Teilfolge (x_{n_k}). Sei $x_{\max} := \lim_{k\to\infty} x_{n_k}$. Dann gilt:

$$f(x_{\max}) = \lim_{k\to\infty} f(x_{n_k}) = M$$

wegen der Stetigkeit von f. Insbesondere ist $M < \infty$. □

Bemerkung 8.12. Im vorigen Beweis kann man den Übergang zu einer Teilfolge im Allgemeinen nicht vermeiden. Dies zeigt das Beispiel: $f(x) = 1 - (x^2-1)^2$ auf $[-2,2]$ (Abb. 8.5). Die Ausgangsfolge (x_n) könnte zwischen den zwei Maxima $x_{\max} = \pm 1$ hin und her springen.

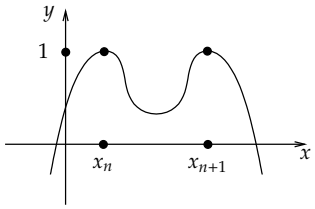


Abbildung 8.5. Eine Funktion mit zwei Maxima auf dem selben Niveau.

Definition 8.13. *Eine Funktion $f\colon I \to \mathbb{R}$ auf einem Intervall heißt* ***monoton wachsend*** *(oder* ***monoton steigend****), wenn $f(x_1) \leq f(x_2)$ für $x_1 < x_2$;* ***streng monoton wachsend*** *(oder* ***streng monoton steigend****), wenn $f(x_1) < f(x_2)$ für $x_1 < x_2$. Analog sind* ***monoton fallend*** *und* ***streng monoton fallend*** *definiert.*

f heißt ***streng monoton****, falls f entweder streng monoton wachsend oder streng monoton fallend ist.*

Satz 8.14.

1. *Sei $f\colon I \to \mathbb{R}$ eine stetige Funktion auf einem Intervall. Dann ist $J = f(I) \subset \mathbb{R}$ ebenfalls ein Intervall.*
2. *Ist f außerdem streng monoton, dann ist die Abbildung $f\colon I \to J$ bijektiv. Mit $f^{-1}\colon J \to I \subset \mathbb{R}$ bezeichnen wir dann die* ***Umkehrfunktion****, d.h. die Abbildung mit $f^{-1}(f(x)) = x\ \forall x \in I$.*

Beweis. 1. J ist ein Intervall, wenn mit $y_1, y_2 \in J$ auch alle Punkte zwischen y_1 und y_2 in J liegen. Dies ist der Fall nach dem Zwischenwertsatz 8.9.

2. Ist f streng monoton, so ist $f\colon I \to \mathbb{R}$ injektiv. Also ist $f\colon I \to J$ injektiv und surjektiv, d.h. insbesondere bijektiv, so dass die Umkehrfunktion $f^{-1}\colon J \to I$ erklärt ist.

□

Definition 8.15. *Sei $f\colon D \to \mathbb{R}$ eine Funktion und $x_0 \in \mathbb{R}\backslash D$ ein Punkt, für den es eine Folge (x_n) mit $x_n \in D$ und $\lim_{n\to\infty} x_n = a$ gibt. Existiert für jede Folge (x_n) auf D mit $\lim_{n\to\infty} x_n = a$ der Grenzwert $\lim_{n\to\infty} f(x_n)$, so dass alle diese Grenzwerte gleich sind, so bezeichnen wir mit*

$$\lim_{x\to a} f(x) := \lim_{n\to\infty} f(x_n)$$

den gemeinsamen ***Grenzwert****.*

Beispiel 8.16. Die Funktion

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

ist zunächst nur auf $D = \mathbb{R}\backslash\{1\}$ definiert. Da aber

$$\lim_{x\to 1} f(x) = \lim_{x\to 1} \frac{(x-1)(x+1)}{x-1} = \lim_{x\to 1}(x+1) = 2$$

ist, lässt sich die Funktion $f\colon D \to \mathbb{R}$ zu einer stetigen Funktion $\tilde{f}\colon \mathbb{R} \to \mathbb{R}$ fortsetzen: $\tilde{f}(x) = x + 1$.

Definition 8.17. *Die Notation* $\lim_{x\nearrow a} f(x)$ *verwenden wir, wenn wir nur Folgen* (x_n) *mit* $x_n < a$ *betrachten. Analog:* $\lim_{x\searrow a} f(x)$.

Aufgaben

Aufgabe 8.1 (Stetigkeit). Bestimmen Sie, in welchen Punkten die folgende Funktion stetig ist:

$$f(x) = \begin{cases} -x + 1, & x \le -1, \\ x^2 + 5x + 7, & -1 < x \le 0, \\ x + 7, & x > 0. \end{cases}$$

Aufgabe 8.2 (Stetigkeit). Die drei Funktionen $f, g, h\colon \mathbb{R} \to \mathbb{R}$ seien folgendenmaßen definiert:

$$f(x) = \begin{cases} x, & x \in \mathbb{Q}, \\ 1 - x, & x \notin \mathbb{Q}, \end{cases}$$

$$g(x) = \begin{cases} 1, & x \in \mathbb{Q}, \\ 0, & x \notin \mathbb{Q}, \end{cases}$$

$$h(x) = \begin{cases} \frac{1}{q}, & x = \frac{p}{q} \in \mathbb{Q} \text{ mit } p, q \in \mathbb{Z} \text{ teilerfremd}, q > 0, \\ 0, & x \notin \mathbb{Q}. \end{cases}$$

Zeigen Sie: f ist nur in $\frac{1}{2}$ stetig, g ist nirgendwo stetig und h ist genau in allen irrationalen x stetig.

Aufgabe 8.3 (Leinenwurf). In einem Raum ist eine Leine von der Fensterwand zur gegenüberliegenden Wand gespannt. Jetzt wird die Leine an beiden Seiten gelöst und irgendwie in die Mitte des Raumes geworfen.

Zeigen Sie: Es gibt einen Punkt auf der Leine, der genauso weit von der Fensterwand entfernt ist wie zuvor.

Aufgabe 8.4 (Stetige Funktionen).

1. Gibt es eine stetige Funktion $f\colon \mathbb{R} \to \mathbb{R}$, die jeden ihrer Werte genau zweimal annimmt?
2. Gibt es eine stetige Funktion $f\colon \mathbb{R} \to \mathbb{R}$, die jeden ihrer Werte genau dreimal annimmt?

9

Differentiation

...

9.1 Differenzierbarkeit

Definition 9.1. *Sei $f : I \to \mathbb{R}$ eine Funktion auf einem Intervall und $x_0 \in I$. f heißt in x_0 **differenzierbar** (kurz auch: **diffbar**), falls der Grenzwert*

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

existiert.

Geometrisch lässt sich der **Differenzenquotient**

$$\frac{f(x) - f(x_0)}{x - x_0}$$

als Steigung der **Sekante** durch die Punkte $(x_0, f(x_0))$ und $(x, f(x))$ des Graphen G_f interpretieren (Abb. 9.1). Der Grenzwert lässt sich also als Steigung der Tangente an G_f im Punkt $(x_0, f(x_0))$ interpretieren und f ist in x_0 differenzierbar, wenn G_f in $(x_0, f(x_0))$ vernünftig eine nicht senkrechte Tangente zugeordnet werden kann.

Definition 9.2. *$f : I \to \mathbb{R}$ ist auf I differenzierbar, wenn f in jedem Punkt $x_0 \in I$ differenzierbar ist. Die Funktion*

$$f' : I \to \mathbb{R},\ f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

*nennen wir dann die **Ableitung** von f auf I.*

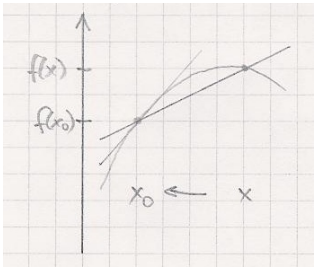


Abbildung 9.1. Differenzenquotient als Sekantensteigung. Gezeigt ist die Sekante durch $(x_0, f(x_0))$ und $(x, f(x))$ mit Steigung $\frac{f(x)-f(x_0)}{x-x_0}$.

Bemerkung 9.3. 1. Differenzierbarkeit ist fundamental, um Begriffe wie Geschwindigkeit in der Physik überhaupt definieren zu können. Beschreibt $f\colon I \to \mathbb{R},\ t \mapsto f(t)$ die Bewegung eines Punktes $f(t)$ in $\mathbb{R}$, so ist $f'(t)$ die Geschwindigkeit zum Zeitpunkt t.

2. Von Newton stammt die Notation $f'(x)$ bzw. $\dot{f(t)}$ bei Ableitungen nach der Zeit. Leibniz hat die in gewisser Weise bessere Notation

$$\frac{df}{dx}(x_0) = f'(x_0)$$

verwendet.

Satz 9.4. *Sei $f\colon I \to \mathbb{R}$ eine Funktion und $x_0 \in I$. Dann gilt: f ist differenzierbar in $x_0 \Rightarrow f$ ist stetig in x_0.*

Beweis. Existiert $\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$, dann auch der Grenzwert

$$\lim_{x\to x_0}\Big(\frac{f(x)-f(x_0)}{x-x_0}\cdot(x-x_0)\Big) = \lim_{x\to x_0}(f(x)-f(x_0))$$

und ist

$$\lim_{x\to x_0}(f(x)-f(x_0)) = \lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}\cdot\lim_{x\to x_0}x-x_0 = 0.$$

Also: $\lim_{x\to x_0} f(x) = f(x_0)$, d.h. f ist stetig in x_0 nach dem Folgenkriterium. □

Beispiel 9.5.

1. $f(x) = x^2$ ist in jedem Punkt $x_0 \in \mathbb{R}$ differenzierbar:

$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0} = \lim_{x\to x_0}\frac{x^2-x_0^2}{x-x_0} = \lim_{x\to x_0}(x+x_0) = 2x_0.$$

Also: $f'(x) = 2x$.

2. Konstante Funktionen $f(x) = c$ sind differenzierbar mit $f'(x) = 0$.

In den Übungsaufgaben werden wir Funktionen kennen lernen, die zwar stetig, aber an einigen Stellen nicht differenzierbar sind. In vielen einfachen Fällen kann man solche Stellen dadurch erkennen, dass der Graph einen Knick hat, wie z.B. die Betragsfunktion im Ursprung (siehe Abb. 9.2).

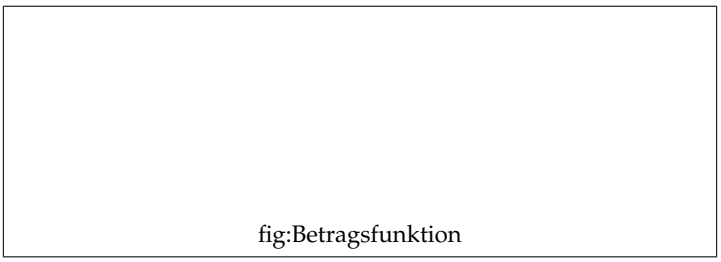


Abbildung 9.2. SKIZZE FEHLT!

Es gibt aber auch Funktionen, wie beispielsweise die sogenannte **Koch–Kurve**, die zwar überall stetig, aber nirgends differenzierbar sind.

Definition 9.6. *Sei $f\colon I \to \mathbb{R}$ eine differenzierbare Funktion auf einem Intervall und $x_0 \in I$. f heißt in x_0 **stetig differenzierbar** (kurz auch: **stetig diffbar**), falls die Ableitung f' stetig ist.*

In den Übungsaufgaben werden wir eine Funktion kennen lernen, die zwar differenzierbar, aber nicht stetig differenzierbar ist.

9.2 Rechenregeln für Ableitungen

Satz 9.7 (Rechenregeln für Ableitungen). *Seien $f, g\colon I \to \mathbb{R}$ in $x_0 \in I$ differenzierbare Funktionen. Dann sind*

$$f + g\colon I \to \mathbb{R} \text{ und } f \cdot g\colon I \to \mathbb{R}$$

in x_0 ebenfalls differenzierbar mit Ableitungen

$$\begin{aligned}(f + g)'(x_0) &= f'(x_0) + g'(x_0)\\ (f \cdot g)'(x_0) &= f'(x_0) \cdot g(x_0) + g'(x_0) \cdot f(x_0).\end{aligned}$$

*Die zweite Regel heißt auch **Leibnizregel** oder **Produktregel**.*

Ist $g(x_0) \neq 0$, so ist auch

$$\frac{f}{g} \colon I \setminus \{\, x \mid g(x) = 0 \,\} \to \mathbb{R}$$

*in x_0 differenzierbar und es gilt die **Quotientenregel**:*

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'g - fg'}{g^2}(x_0).$$

Beweis. Die Aussage zu $f + g$ folgt direkt aus der Definition. Zum Nachweis der Produktregel betrachten wir

$$\begin{aligned}\frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0} &= \frac{f(x) - f(x_0)}{x - x_0} g(x) + f(x_0)\frac{g(x) - g(x_0)}{x - x_0} \\ &\xrightarrow[x \to x_0]{} f'(x_0)g(x) + f(x_0)g'(x_0).\end{aligned}$$

Für die Aussage über den Quotienten untersuchen wir zunächst den Spezialfall $\left(\frac{1}{g}\right)' = \frac{-g'}{g^2}$:

$$\frac{\frac{1}{g(x)} - \frac{1}{g(x_0)}}{x - x_0} = \frac{1}{g(x)g(x_0)} \frac{g(x_0) - g(x)}{x - x_0} \xrightarrow[x \to x_0]{} \frac{1}{g^2(x_0)}(-g'(x_0)).$$

Die allgemeine Regel folgt aus dem Spezialfall mit der Produktregel:

$$\left(f \cdot \frac{1}{g}\right)' = f' \cdot \frac{1}{g} + f \cdot \frac{-g'}{g^2} = \frac{f'g - fg'}{g^2}.$$

□

Beispiel 9.8.

1. Die Funktion $f(x) = x^n$ ist für beliebige $n \in \mathbb{Z}$ auf ihrem Definitionsbereich diffbar mit
$$f'(x) = nx^{n-1}.$$
Für $n \geq 0$ haben wir dies schon gesehen. Sei also $n = -k$, $k > 0$, d.h. $f(x) = \frac{1}{x^k}$. Die Quotientenregel ergibt:
$$f'(x) = \frac{-kx^{k-1}}{(x^k)^2} = -kx^{-k+1} = nx^{n-1}.$$

2. Rationale Funktionen $r = \frac{f}{g}$ sind auf ihrem Definitionsbereich diffbar. Ist der Bruch gekürzt, so heißen die Nullstellen von g auch **Polstellen** der rationalen Funktion.

Satz 9.9 (Kettenregel). *Es seien $f\colon I \to \mathbb{R}$, $g\colon J \to \mathbb{R}$ diffbar mit $f(I) \subset J$. Dann ist auch die Komposition*

$$g \circ f\colon I \to \mathbb{R},\ (g \circ f)(x) = g(f(x)),$$

diffbar mit

$$(g \circ f)'(x) = g'(f(x)) \cdot f'(x).$$

*Den Faktor $f'(x)$ nennt man hierbei **innere Ableitung**.*

Beweis. Es gilt:

$$\frac{(g \circ f)(x+h) - (g \circ f)(x)}{h} = \frac{g(f(x+h)) - g(f(x))}{f(x+h) - f(x)} \cdot \frac{f(x+h) - f(x)}{h}.$$

Da mit $h \to 0$ auch $f(x+h) \to f(x)$ gilt, da f stetig ist, folgt:

$$\frac{g(f(x+h)) - g(f(x))}{f(x+h) - f(x)} \xrightarrow[h\to 0]{} g'(f(x))$$

und dann:

$$\frac{f(x+h) - f(x)}{h} \to f'(x).$$

Dieses Argument ist gültig, sofern $f(x+h) - f(x) \neq 0$. Ist aber $f(x+h_n) - f(x) = 0$ für eine Nullfolge (h_n), so folgt $f'(x) = 0$ und $\frac{g(f(x+h_n)) - g(f(x))}{h_n} = 0$, also gilt auch in diesem Fall

$$\lim_{n\to\infty} \frac{g(f(x+h_n)) - g(f(x))}{h_n} = g'(f(x)) \cdot f'(x) = 0.$$

□

Beispiel 9.10. Wir betrachten für $f(x) = x^2 + 1$ und $g(x) = x^3$ die Hintereinanderausführung $(g \circ f)(x) = (x^2+1)^3$. Die Kettenregel ergibt:

$$\left((x^2+1)^3\right)' = 3(x^2+1)^2 \cdot 2x.$$

Wenn wir zunächst ausmultiplizieren, erhalten wir:

$$(x^6 + 3x^4 + 3x^2 + 1)' = 6x^5 + 12x^3 + 6x.$$

Beide Ergebnisse stimmen überein.

Satz 9.11 (Ableitung der Umkehrfunktion). *Sei $f\colon I \to \mathbb{R}$ eine streng monotone diffbare Funktion, $J = f(I)$ und $x_0 \in I$ ein Punkt mit $f'(x_0) \neq 0$. Dann ist die Umkehrfunktion $f^{-1}\colon J \to I \subset \mathbb{R}$ in $y_0 = f(x_0)$ diffbar mit*

$$(f^{-1})'(y_0) = \frac{1}{f'\left(f^{-1}(y_0)\right)} = \frac{1}{f'(x_0)}.$$

Zunächst eine Merkregel für die Formel: Da $f \circ f^{-1} = \mathrm{id}_I$, gilt: $(f \circ f^{-1})' = 1$. Andererseits liefert die Kettenregel $1 = f'(f^{-1}(y_0)) \cdot (f^{-1})'(y_0)$ und die Formel folgt. Dies benutzt allerdings schon die Differenzierbarkeit der Umkehrfunktion, so dass hierfür noch ein Beweis nötig ist.

Beweis (des Satzes 9.11 über die Ableitung der Umkehrfunktion). Nach Voraussetzung ist $f'(x_0) \neq 0$, also $\frac{f(x)-f(x_0)}{x-x_0} \neq 0$ für x nahe x_0. Da mit $x \to x_0$ auch $y = f(x) \to y_0 = f(x_0)$ folgt, erhalten wir:

$$\frac{f^{-1}(y) - f^{-1}(y_0)}{y - y_0} = \frac{x - x_0}{f(x) - f(x_0)} \xrightarrow[y \to y_0]{} \frac{1}{f'(x_0)}.$$

□

Beispiel/Definition 9.12 (k–te Wurzel). Sei $g(x) = \sqrt[k]{x} = x^{1/k}$, $k \in \mathbb{N}$, die Umkehrfunktion von der streng monotonen Funktion $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, $f(x) = x^k$. Dann erhalten wir: $f'(x) = kx^{k-1} \neq 0$ für $x \neq 0$. Es folgt: $g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ ist auf $\mathbb{R}_{>0}$ diffbar mit

$$g'(x) = \frac{1}{k(\sqrt[k]{x})^{k-1}} = \frac{1}{k} x^{\frac{1-k}{k}} = \frac{1}{k} x^{\frac{1}{k}-1}.$$

Erneut ist die Exponentenregel gültig.

Aufgaben

Aufgabe 9.1 (Produktregel). Seien $D \subset \mathbb{R}$ und $f, g : D \to \mathbb{R}$ zwei n Mal differenzierbare Funktionen. Zeigen Sie:

$$(f \cdot g)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(n-k)} g^{(k)}.$$

Aufgabe 9.2 (Approximierung). Schreiben Sie zwei Programme (z.B. mit Maple), um $\sqrt[k]{a}$ zu berechnen; verwenden Sie dabei einmal das Intervall-Halbierungsverfahren und einmal das Newtonverfahren. Zählen Sie, wieviele Iterationen Ihre beiden Verfahren benötigen, um $\sqrt[4]{3}$ mit einer Genauigkeit von mindestens 10^{-5} zu berechnen. Verwenden Sie $[1, 2]$ als Startintervall für das Intervall-Verfahren und 1 als Startwert für das Newtonverfahren.

Aufgabe 9.3 (Optimierung). Eine Konservendose von 320 ml Inhalt soll so dimensioniert werden, dass der Blechverbrauch minimal ist. Wir nehmen dabei an, die Konservendose sei ein perfekter Zylinder. Welche Höhe und welchen Durchmesser hat die Dose?
Hinweis: Dabei dürfen Sie die aus der Schule bekannten Formeln für Volumen und Mantel eines Zylinders verwenden.

Aufgabe 9.4 (Mehrfache Nullstellen). Für welche $a, b \in \mathbb{R}$ hat $f(x) = x^3 - ax + b$ eine doppelte Nullstelle (d.h. eine Stelle x_0 mit $f(x_0) = f'(x_0) = 0$)? Für welche a, b hat die Funktion genau eine, zwei bzw. drei reelle Nullstellen?

Aufgabe 9.5 (Differenzierbarkeit). Zeigen Sie:

1. Die Betragsfunktion $|.|\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto |x|$, ist in $x \neq 0$ differenzierbar, in $x = 0$ aber nicht.
2. Die Funktion
$$x \mapsto \begin{cases} x^2 \cdot \sin \frac{1}{x}, & \text{falls } x \neq 0, \\ 0, & \text{falls } x = 0, \end{cases}$$
ist differenzierbar, aber nicht zweimal differenzierbar.

10

Mittelwertsatz und lokale Extrema

Ableitungen werden häufig eingesetzt, um lokale Extrema, d.h. Minima oder Maxima, zu bestimmen. Eines der bekanntesten Verfahren zur Bestimmung einer Nullstelle einer differenzierbaren Funktion, das Newtonverfahren, benutzt ebenfalls Ableitungen.

10.1 Die erste Ableitung

Definition 10.1. *Seien $f: I \to \mathbb{R}$ eine Funktion und $x_0 \in I$. f hat in x_0 ein* ***lokales Maximum*** *bzw.* ***lokales Minimum****, wenn $\exists\, h > 0$, so dass $]x_0 - h, x_0 + h[\subset I$ und*

$$f(x_0) \geq f(x) \;\; \text{bzw.} \;\; f(x_0) \leq f(x) \; \forall x \in]x_0 - h, x_0 + h[.$$

Ein ***lokales Extremum*** *ist ein lokales Maximum oder Minimum. Gilt*

$$f(x_0) > f(x) \text{ bzw. } f(x_0) < f(x) \; \forall x \in]x_0 - h, x_0 + h[, x \neq x_0,$$

so spricht man von einem ***isolierten Extremum****.*

Absolute Maxima *(auch* ***globale Maxima****) sind Stellen x_0, für die $f(x_0) \geq f(x)$ $\forall x \in I$ gilt.* ***Absolute Minima*** *und* ***absolute Extrema*** *(auch* ***globale Minima*** *und* ***globale Extrema****) sind analog definiert.*

Satz 10.2. *Hat $f:]a, b[\to \mathbb{R}$ in $x_0 \in]a, b[$ ein lokales Extremum und ist f in x_0 diffbar, so gilt $f'(x_0) = 0$.*

Beweis. Wir betrachten den Fall eines lokalen Maximums. Es gilt:

$$\frac{f(x) - f(x_0)}{x - x_0} \leq 0 \text{ für } x > x_0 \;\; \text{und} \;\; \frac{f(x) - f(x_0)}{x - x_0} \geq 0 \text{ für } x < x_0.$$

Es folgt

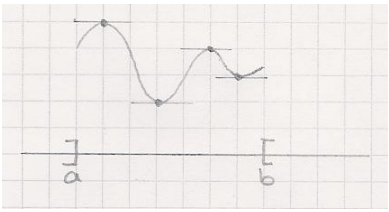


Abbildung 10.1. Die Ableitung in einem lokalen Extremum verschwindet.

$$0 \geq \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) \geq 0$$

und damit die Behauptung. □

Bemerkung 10.3. $f'(x) = 0$ ist notwendig, aber nicht hinreichend für lokale Extrema einer diffbaren Funktion, wie das folgende Beispiel (Abb. 10.2) zeigt: $f(x) = x^3$ erfüllt $f'(0) = 0$, aber $x_0 = 0$ ist kein lokales Extremum.

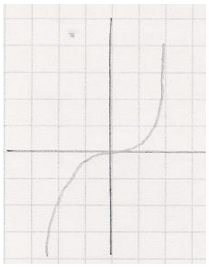


Abbildung 10.2. Eine verschwindende Ableitung ist kein hinreichendes Kriterium für die Existenz eines lokalen Extremums, wie die Abbildung zeigt. Hier ist $f(x) = x^3$.

Satz 10.4 (Satz von Rolle). *Sei* $f\colon [a,b] \to \mathbb{R}$ *eine stetige und auf* $]a,b[$ *diffbare Funktion mit* $f(a) = f(b)$. *Dann existiert ein* $\xi \in]a,b[$ *mit* $f'(\xi) = 0$ *(Abb. 10.3).*

Beweis. Ist f konstant, dann hat jedes $\xi \in]a,b[$ diese Eigenschaft. Anderenfalls verwenden wir, dass f auf $[a,b]$ sowohl Maximum als auch Minimum annimmt. Da $f(a) = f(b)$ und f nicht konstant ist, können Maximum und Minimum nicht beide die Randpunkte sein. Es folgt, dass f auf $]a,b[$ ein Extremum an der Stelle ξ annimmt, das also $f'(\xi) = 0$ erfüllt. □

Satz 10.5 (Mittelwertsatz (MWS)). *Sei* $f\colon [a,b] \to \mathbb{R}$ *stetig und auf* $]a,b[$ *diffbar. Dann existiert ein* $\xi \in]a,b[$ *mit (siehe auch Abb. 10.4):*

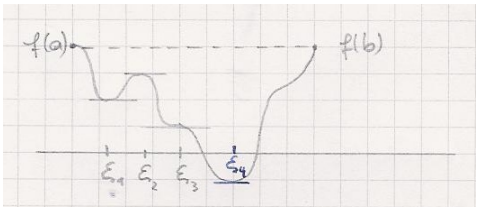


Abbildung 10.3. Der Satz von Rolle.

$$\frac{f(b)-f(a)}{b-a} = f'(\xi).$$

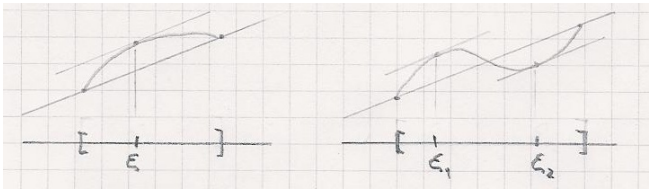


Abbildung 10.4. Der Mittelwertsatz.

Beweis. Wir betrachten

$$F(x) = f(x) - \frac{f(b)-f(a)}{b-a}(x-a).$$

Dann gilt:

$$F(a) = f(a) = F(b).$$

F ist diffbar mit $F'(x) = f'(x) - \frac{f(b)-f(a)}{b-a}$. Nach dem Satz von Rolle existiert ein $\xi \in]a,b[$ mit $F'(\xi) = 0$. □

Korollar 10.6. *Sei $f\colon [a,b] \to \mathbb{R}$ stetig und in $]a,b[$ diffbar. Außerdem nehmen wir an, dass $m, M \in \mathbb{R}$ existieren mit $m \le f'(x) \le M \quad \forall x \in]a,b[$. Dann gilt für $x_1 < x_2$ mit $a \le x_1 < x_2 \le b$ (Abb. 10.5):*

$$m\cdot(x_2-x_1) \le f(x_2) - f(x_1) \le M\cdot(x_2-x_1).$$

Beweis. Nach dem Mittelwertsatz gilt für $x_1 < x_2$: $m \le \frac{f(x_2)-f(x_1)}{x_2-x_1} \le M$. □

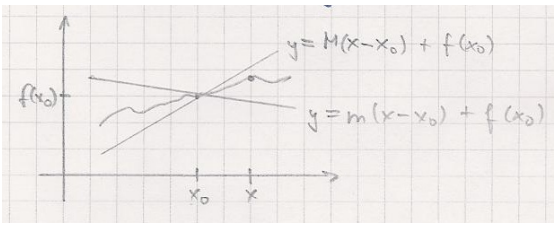


Abbildung 10.5. Schranken für die Differenz zweier Funktionswerte.

Korollar 10.7. *Sei $f\colon [a,b] \to \mathbb{R}$ stetig und in $]a,b[$ diffbar. Gilt $f'(x) = 0\ \forall x \in]a,b[$, so ist f konstant.*

Beweis. Wäre f nicht konstant, so gäbe es x_1, x_2 mit $f(x_1) \neq f(x_2)$ und dann mit dem Mittelwertsatz ein $\xi \in]a,b[$ mit $f'(\xi) \neq 0$, im Widerspruch zur Voraussetzung. □

Satz 10.8. *Es sei $f\colon [a,b] \to \mathbb{R}$ stetig und in $]a,b[$ diffbar. Gilt $f'(x) > 0$ (bzw. ≥ 0, < 0, ≤ 0) $\forall x \in]a,b[$, dann ist f streng monoton wachsend (bzw. monoton wachsend, streng monoton fallend, monoton fallend).*

Beweis. Wir verwenden den Mittelwertsatz: Angenommen, es existieren x_1, x_2 mit $x_1 < x_2$, aber $f(x_1) \geq f(x_2)$, so existiert ξ mit $f'(\xi) \leq 0$ im Widerspruch zur Voraussetzung. □

10.2 Höhere Ableitungen

Definition 10.9 (höhere Ableitungen). *Sei $f\colon I \to \mathbb{R}$ diffbar. f heißt 2–mal diffbar, wenn $f'\colon I \to \mathbb{R}$ ebenfalls diffbar ist. $f^{(2)} := f'' := (f')'$ bezeichnet dann die 2–te Ableitung. Allgemeiner ist f n–mal diffbar, wenn*

$$f^{(n)} := (f^{(n-1)})'$$

existiert.

Satz 10.10 (hinreichendes Kriterium für Extrema). *Sei $f\colon]a,b[\to \mathbb{R}$ zweimal diffbar. Ist $f'(x_0) = 0$ und $f''(x_0) \neq 0$, so hat f in x_0 ein isoliertes lokales Extremum. Dieses ist ein Maximum, wenn $f''(x_0) < 0$ und ein Minimum, wenn $f''(x_0) > 0$.*

Beweis. Wir betrachten den Fall, dass $f'(x_0) = 0$ und $f''(x_0) < 0$. Dann gilt:

$$\lim_{x \to x_0} \frac{f'(x) - f'(x_0)}{x - x_0} < 0.$$

Es folgt, dass ein $\varepsilon > 0$ existiert, so dass

$$f'(x) > 0 \text{ für } x \in]x_0 - \varepsilon, x_0[\text{ und } f'(x) < 0 \text{ für } x \in]x_0, x_0 + \varepsilon[.$$

Es folgt, dass f in $]x_0 - a, x_0]$ streng monoton wachsend und in $[x_0, x_0 + \varepsilon[$ streng monoton fallend ist. x_0 ist also ein isoliertes Maximum. □

Beispiel 10.11. Sei $f(x) = x^2$. Dann ist $f'(x) = 2x$, $f''(x) = 2 > 0$ $\forall x$, also $f'(0) = 0, f''(0) > 0$. Daher ist 0 ein Minimum von f. Analog ist 0 ein Maximum von $g(x) = -x^2$, da $g'(0) = 0$ und $g'(0) < 0$. Siehe auch Abb. 10.6.

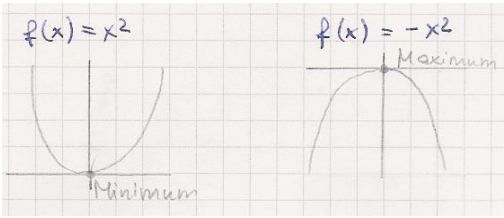


Abbildung 10.6. Parabeln mit Maximum bzw. Minimum.

Definition 10.12. *Sei $f\colon]a, b[\to \mathbb{R}$ dreimal diffbar. Ein Punkt $x_0 \in]a, b[$ mit $f''(x_0) = 0$, $f'''(x_0) \neq 0$ heißt **Wendepunkt** von f. Ist $f'(x_0) = 0$, so heißt x_0 **Sattelpunkt** von f.*

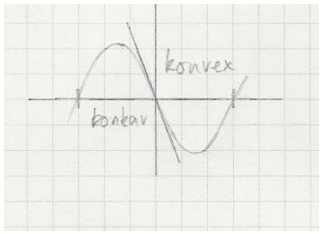


Abbildung 10.7. Die Umgebung eines Wendepunktes.

Definition 10.13. *Sei $I \subset \mathbb{R}$ ein Intervall $f\colon I \to \mathbb{R}$ heißt **konvex** (Abb. 10.8), wenn $\forall x_1, x_2 \in I$ und alle λ mit $0 \leq \lambda \leq 1$:*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

*f heißt **konkav**, wenn $-f$ konvex ist.*

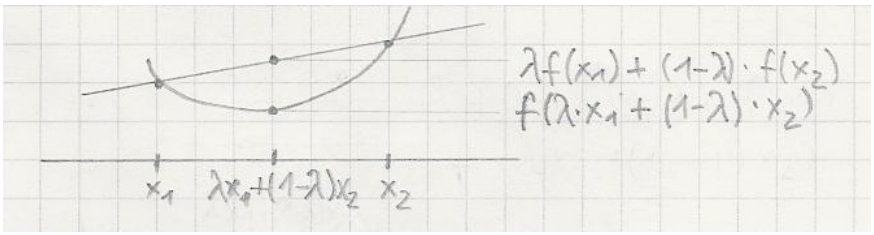


Abbildung 10.8. Definition von konvex.

Beispiel 10.14. Sei $f(x) = x^3 - x$. Es ist $f''(x) = 6x$, $f'''(x) = 6 \neq 0$. Diese Funktion ändert sich im Wendepunkt $x_0 = 0$ von konkav zu konvex.

Satz 10.15. *Sei $f\colon I \to \mathbb{R}$ eine zwei Mal diffbare Funktion auf einem Intervall I. Dann gilt:*

$$f \text{ ist konvex} \iff f''(x) \geq 0 \ \forall x \in I.$$

10.3 Das Newtonverfahren zur Berechnung von Nullstellen

Mit den bisherigen Resultaten in diesem Kapitel können wir nun also die aus der Schule bekannte **Kurvendiskussion** zum Studium des Aussehens reeller Funktionen in einer Variablen durchführen, sofernd diese ausreichend oft differenzierbar sind. Ein Problem haben wir allerdings noch vernachlässigt, nämlich die Berechnung der Nullstellen solcher Funktionen. Von einigen speziellen Funktionen, wie beispielsweise Polynomen vom Grad ≤ 2 können wir sie bestimmen, doch wie sieht es im Allgemeinen aus? Hierzu liefert ein **Iterationsverfahren**, nämlich das sogenannte **Newtonverfahren**, näherungsweise eine Möglichkeit, zumindest, wenn man sich schon *nahe genug* an einer der Nullstellen befindet.

Sei also $f\colon [a,b] \to \mathbb{R}$ eine zweimal diffbare Funktion mit $f(a) < 0$, $f(b) > 0$. Die Idee ist folgende: Ist x_0 ein Startwert, so setzen wir:

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)},$$

d.h. x_{n+1} ist die Nullstelle der Tangente in $(x_n, f(x_n))$ an den Graphen von f (Abb. 10.9).

Satz/Definition 10.16. *Sei $f\colon [a,b] \to \mathbb{R}$ zweimal diffbar, $f(a) < 0$, $f(b) > 0$ und konvex. Dann gilt:*

1. *Es gibt genau eine Nullstelle $\xi \in [a,b]$.*

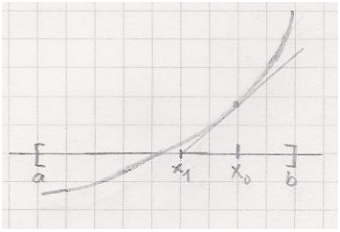


Abbildung 10.9. Die Idee des Newtonverfahrens.

2. *Ist* $x_0 \in [a,b]$ *ein beliebiger Startwert mit* $f(x_0) \geq 0$, *so konvergiert die Folge* (x_n) *mit* $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ *monoton fallend gegen* ξ.
3. *Ist* $f'(x) \geq c > 0$ *und* $f''(x) < k \; \forall x \in [\xi, b]$, *so gilt die Abschätzung:*

$$|x_{n+1} - x_n| \leq |\xi - x_n| \leq \frac{k}{2c}|x_n - x_{n-1}|^2.$$

Man sagt deshalb, dass das Newtonverfahren ***quadratisch konvergiert****.*

Das Newtonverfahren ist wegen der quadratischen Konvergenz meist wesentlich schneller als das Intervallhalbierungsverfahren aus dem Beweis des Zwischenwertsatzes 8.9.

Beispiel 10.17. Wir betrachten $f(x) = x^2 - a$. Dann ist

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$$

unser Verfahren zur Berechnung der Quadratwurzel $\sqrt{a}$ aus Satz 5.27.

Beweis (des Satzes 10.16 zum Newtonverfahren).

1. f hat eine Nullstelle nach dem Zwischenwertsatz und genau eine, da f konvex ist (siehe Abb. 10.10).
2. f ist konvex. Daher: $f(x_n) > 0 \Rightarrow f'(x_n) > 0$ und $\xi \leq x_{n+1} < x_n$ (s. Abb. 10.11). Die Folge (x_n) ist wohldefiniert, monoton fallend, beschränkt und daher konvergent. Der Grenzwert erfüllt: $f(\overline{x}) = 0$, also: $\overline{x} = \xi$.
3. Da f' monoton wächst, gilt $f'(x) \geq c > 0 \; \forall x \in [\xi, b]$. Mit dem Mittelwertsatz folgt: $|\xi - x_n| \leq \frac{f(x_n)}{c}$. Um $f(x_n)$ abzuschätzen, betrachten wir die Hilfsfunktion

$$\psi(x) = f(x) - f(x_{n-1}) - f'(x_{n-1})(x - x_{n-1}) - \frac{k}{2}(x - x_{n-1})^2.$$

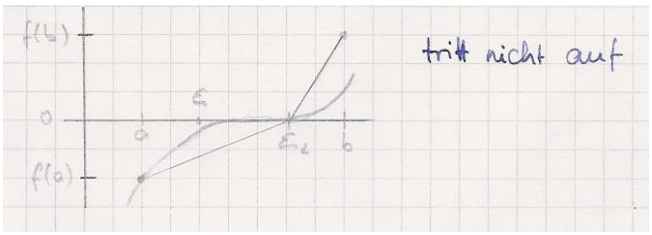


Abbildung 10.10. Konvexität erzwingt: höchstens eine Nullstelle.

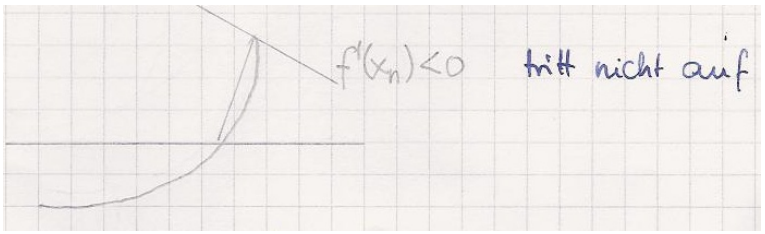


Abbildung 10.11. Konvexität erzwingt: Steigung positiv.

Dafür gilt:

$$\psi'(x) = f'(x) - f'(x_{n-1}) - k(x - x_{n-1})$$
$$\psi''(x) = f''(x) - k \leq 0 \; \forall x \in]\xi, b[.$$

ψ fällt also monoton. Da $\psi'(x_{n-1}) = 0$ ist, folgt: $\psi'(x) \geq 0 \forall x \in]\xi, x_{n-1}[$. Da außerdem $\psi(x_{n-1}) = 0$ ist, gilt auch: $\psi(x) \leq 0 \; \forall x \in]\xi, x_{n-1}[$ und insbesondere $\psi(x_n) \leq 0$, d.h. $f(x_n) \leq \frac{k}{2}(x_n - x_{n-1})^2$, da $f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0$. Also: $|x_{n+1} - x_n| \leq |\xi - x_n| \leq \frac{k}{2c}(x_n - x_{n-1})^2$.

□

Aufgaben

Aufgabe 10.1 (Kurvendiskussion). Diskutieren Sie die folgenden Funktionen, d.h. bestimmen Sie alle Nullstellen, lokale Minima und Maxima, Wendepunkte, Polstellen, den Definitionsbereich und das asymptotische Verhalten. Fertigen Sie jeweils eine Skizze an.

$$f_1(x) = \frac{x^3 - 3x}{x^2 - 4}$$
$$f_2(x) = xe^{-\frac{1}{x}}$$
$$f_3(x) = 2\cos x - x^2$$

Aufgabe 10.2 (Extrema). Sei $a \in \mathbb{R}$. Bestimmen Sie alle Minima und Maxima der Funktion

$$f(x) = \sin(x + a)\sin(x - a).$$

11

Spezielle Funktionen

Wir besprechen nun einige wichtige Beispiele differenzierbarer Funktionen, wie die Exponentialfunktion, den Logarithmus, sowie einige trigonometrische Funktionen, z.B.: Sinus, Cosinus, Tangens, Arcussinus, Arcustangens. Dabei geben wir auch eine exakte Definition der Kreiszahl π.

11.1 Die Exponentialfunktion

In Beispiel 7.3 haben wir die Exponentialfunktion $\exp\colon \mathbb{R} \to \mathbb{R}$ bereits durch $e^x = \exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ mit Konvergenzradius $R = \infty$ definiert und anschließend erste Eigenschaften, wie $e^{(x_1+x_2)} = e^{x_1} \cdot e^{x_2}$ und $e^{-x} = \frac{1}{e^x}$ mit $e = \exp(1)$, hergeleitet.

Unser erstes Ziel ist es nun, die Differenzierbarkeit der Exponentialfunktion zu zeigen. Zunächst im Nullpunkt:

$$\frac{\exp(x) - \exp(0)}{x - 0} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{n!} \xrightarrow[x \to 0]{} ?$$

Lemma 11.1 (Restgliedabschätzung der Exponentialreihe). *Wir definieren für $N \in \mathbb{N}$ die Funktion r_{N+1} durch*

$$\exp(x) = \sum_{n=0}^{N} \frac{x^n}{n!} + r_{N+1}(x).$$

Dann gilt:

$$|r_{N+1}(x)| \leq 2\frac{|x|^{N+1}}{(N+1)!} \text{ für } |x| \leq 1 + \frac{N}{2}.$$

Beweis. Es ist $r_{N+1}(x) = \sum_{n=N+1}^{\infty} \frac{x^n}{n!}$, also:

$$\begin{aligned}
|r_{N+1}(x)| &\leq \sum_{n=N+1}^{\infty} \frac{|x^n|}{n!} \\
&= \frac{|x|^{N+1}}{(N+1)!}\Big(1 + \frac{|x|}{N+2} + \frac{|x|^2}{(N+2)(N+3)} + \cdots\Big) \\
&\leq \frac{|x|^{N+1}}{(N+1)!} \cdot \sum_{k=0}^{\infty}\Big(\frac{|x|}{N+2}\Big)^k \\
&\leq \frac{|x|^{N+1}}{(N+1)!} \cdot \frac{1}{1-\frac{1}{2}} \\
&= 2 \cdot \frac{|x|^{N+1}}{(N+1)!},
\end{aligned}$$

wie behauptet. □

Satz 11.2. *Die Funktion* $\exp\colon \mathbb{R} \to \mathbb{R}$ *ist diffbar mit* $\exp'(x) = \exp(x)$.

Beweis. Wir zeigen zunächst, dass $\exp'(0) = 1$ gilt. Dazu bemerken wir, dass $\frac{\exp(x)-1}{x} = \sum_{n=1}^{N} \frac{x^{n-1}}{n!} + \frac{r_{N+1}(x)}{x}$. Wegen

$$0 \leq |\frac{r_{N+1}(x)}{x}| \leq \frac{1}{2}\frac{|x|^N}{(N+1)!} \xrightarrow[x\to 0]{} 0$$

folgt:

$$\exp'(0) = \lim_{x\to 0} \frac{\exp(x) - \exp(0)}{x} = \lim_{x\to 0}\Big(\sum_{n=1}^{N} \frac{x^{n-1}}{n!}\Big) + 0 = 1.$$

Im allgemeinen Fall $x_0 \in \mathbb{R}$ verwenden wir das Additionstheorem:

$$\frac{\exp(x_0+h) - \exp(x_0)}{h} = \exp(x_0) \cdot \frac{\exp(h) - 1}{h} \xrightarrow[n\to\infty]{} \exp(x_0) \cdot 1 = \exp(x_0).$$

Tatsächlich folgt also: $\exp' = \exp$. □

Direkt folgt:

Korollar 11.3. exp *ist streng monoton steigend und konvex (Abb. 11.1).*

Das häufige Auftreten der Exponentialfunktion bei der Beschreibung von Naturvorgängen liegt daran, dass $y = e^{cx}$ eine Lösung der Differentialgleichung $y' = cy$ ist. Genauer gilt:

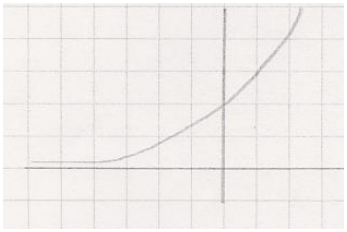


Abbildung 11.1. Die Exponentialfunktion ist streng monoton steigend und konvex.

Satz 11.4. *Sei $f\colon I \to \mathbb{R}$ eine Funktion auf einem Intervall, die $f' = cf$ erfüllt. Dann gilt:*

$$f(x) = f(x_0) \cdot e^{c(x-x_0)},$$

wobei $x_0 \in I$ ein beliebiger fester Punkt ist.

Beweis. Wir betrachten die Funktion $h(x) = f(x) \cdot e^{-cx}$. Dann gilt:

$$h'(x) = f'(x)e^{-cx} + f(x)(-c)e^{-cx} = cf(x)e^{-cx} - cf(x)e^{-cx} = 0$$

für jedes $x \in I$. Es folgt, dass h konstant ist. Setzen wir x_0 ein, so erhalten wir:

$$h(x) = f(x)e^{-cx} = h(x_0) = f(x_0)e^{-cx_0},$$

also: $f(x) = f(x_0)e^{c(x-x_0)}$. □

Die Abbildung $\exp\colon \mathbb{R} \to \mathbb{R}_{>0}$ ist bijektiv und wegen des Additionstheorems ein sogenannter **Isomorphismus von Gruppen** $(\mathbb{R}, +) \to (\mathbb{R}_{>0}, \cdot)$, d.h. in diesem Spezialfall: $\exp(x + y) = \exp(x) \cdot \exp(y)$ für alle $x, y \in \mathbb{R}$. Genauer werden wir Gruppen im zweiten Semester kennen lernen.

11.2 Der Logarithmus

Definition 11.5. *Die Umkehrfunktion*

$$\ln\colon \mathbb{R}_{>0} \to \mathbb{R}$$

von exp *heißt der* ***natürliche Logarithmus****.*

Satz 11.6 (Eigenschaften des Logarithmus). *Es gilt:*

1. $\ln(x_1 \cdot x_2) = \ln x_1 + \ln x_2$.
2. ln *ist diffbar mit* $(\ln x)' = \frac{1}{x}$.

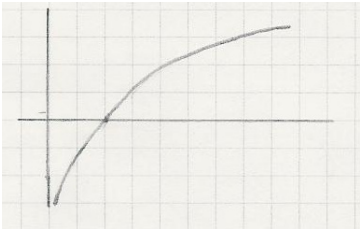


Abbildung 11.2. ln ist konkav und monoton wachsend.

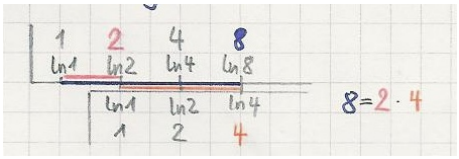


Abbildung 11.3. Der Rechenschieber basiert auf dem Logarithmus.

3. ln *ist konkav und monoton wachsend (Abb. 11.2).*

Beweis. 1. Dies folgt aus dem Additionstheorem der Exponentialfunktion. Hierauf basiert der **Rechenschieber**, siehe Abb. 11.3.

2. Nach dem Satz über die Ableitung der Umkehrfunktion ist

$$ln'(x) = \frac{1}{\exp(\ln x)} = \frac{1}{x}.$$

3. Dies folgt aus den Eigenschaften von exp.

□

Definition 11.7. *Sei $a \in \mathbb{R}_{>0}$. Dann definieren wir die **Exponentiation zu einer beliebigen Basis** durch*

$$a^x := e^{x \cdot \ln a}.$$

Satz 11.8. *Es gilt:*

1. $a^{x_1+x_2} = a^{x_1} \cdot a^{x_2}$.
2. *Für $x = \frac{p}{q} \in \mathbb{Q}$ gilt: $a^x = \sqrt[q]{a^p}$. Dies stimmt mit der alten Definition 9.12 überein.*
3. *Die Funktion $x \mapsto a^x$ ist diffbar mit $(a^x)' = \ln a \cdot a^x$.*

Beweis. Die erste und letzte Aussage sind klar. Für die verbleibende betrachten wir zunächst ganzzahlige positive Exponenten $n \in \mathbb{Z}_{>0}$. Es gilt:

$$a^n (= \underbrace{a \cdots a}_{n \text{ Mal}}) = (e^{\ln a})^n = e^{n \ln a}.$$

Ist $n \in \mathbb{Z}_{<0}$, also $n = -k < 0$, so folgt:

$$a^n = \frac{1}{a^k} = \frac{1}{e^{k \ln a}} = e^{-k \ln a} = e^{n \ln a}.$$

Im Allgemeinen Fall müssen wir zeigen, dass $\sqrt[q]{a^p} = e^{\frac{p}{q} \ln a}$, was äquivalent ist zu $a^p = (e^{\frac{p}{q} \ln a})^q$ und Letzteres ist tatsächlich gleich $e^{p \ln a} = (e^{\ln a})^p = a^p$. □

Definition 11.9. *Für $a \in \mathbb{R}_{>0}$ bezeichnet*

$$\log_a : \mathbb{R}_{>0} \to \mathbb{R}$$

die Umkehrfunktion von $x \mapsto a^x$.

Bemerkung 11.10. $\log_a x$ ist diffbar mit

$$(\log_a)'(x) = \frac{1}{\ln a \cdot a^{\log_a x}} = \frac{1}{x \ln a}.$$

Besonders wichtig für die Informatik ist $\log_2 n$, die Anzahl der **Binärstellen** einer natürlichen Zahl n, d.h. die Anzahl der Bits Information.

Beispiel 11.11. Wir betrachten die Funktion

$$f : \mathbb{R}_{>0} \to \mathbb{R},\ f(x) = x^x.$$

f ist diffbar nach der Kettenregel: $x^x = e^{x \ln x}$ und

$$f'(x) = e^{x \ln x}\Big(\ln x + x \cdot \frac{1}{x}\Big) = (1 + \ln x) \cdot e^{x \ln x} = (1 + \ln x) \cdot x^x.$$

Funktionen wie $x \mapsto x^x$ tauchen in der Komplexitätstheorie auf: Einer der besten bekannten Algorithmen, um eine Zahl n mit $x = \log_2 n$ Binärstellen zu faktorisieren, hat die Laufzeit $O(e^{\frac{1}{2} x \log_2 x})$.

11.3 Trigonometrische Funktionen

Wir hatten Sinus und Cosinus in Beispiel 7.3 bereits durch Potenzreihen definiert:

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

Außerdem haben wir in Satz 7.25 bereits die Additionstheoreme und insbesondere $\sin^2 x + \cos^2 x = 1 \; \forall x \in \mathbb{R}$ gezeigt.

Satz 11.12. *Die Funktionen* $\sin, \cos \colon \mathbb{R} \to \mathbb{R}$ *sind diffbar mit*

$$\sin' = \cos, \quad \cos' = -\sin.$$

Beweis. Zunächst zeigen wir: $\sin'(0) = 1 = \cos(0)$ und $\cos'(0) = 0 = \sin 0$:

$$\sin'(0) = \lim_{h \to 0} \frac{\sin h - 0}{h} = \lim_{h \to 0} \Big(\sum_{k=0}^{\infty} (-1)^k \frac{h^{2k}}{(2k+1)!} \Big) = 1,$$

$$\cos'(0) = \lim_{h \to 0} \frac{\cos h - 1}{h} = \lim_{h \to 0} \Big(\sum_{k=1}^{\infty} (-1)^k \frac{h^{2k-1}}{(2k)!} \Big) = 0.$$

Allgemein erhalten wir mit den Additionstheoremen:

$$\frac{\sin(x_0 + h) - \sin(x_0)}{h} = \sin x_0 \cdot \frac{\cos h - 1}{h} + \cos x_0 \cdot \frac{\sin h - 0}{h} \xrightarrow[h \to 0]{} \cos x_0.$$

Entsprechend ergibt sich für den Cosinus:

$$\frac{\cos(x_0 + h) - \cos(x_0)}{h} = \cos x_0 \cdot \frac{\cos h - 1}{h} - \sin x_0 \cdot \frac{\sin h - 1}{h} \xrightarrow[h \to 0]{} -\sin x_0.$$

□

Als nächstes werden wir die Zahl π definieren. Dazu zunächst ein Hilfssatz:

Lemma 11.13. *Es gilt:*

1. $\cos(0) = 1, \cos(2) < 0.$
2. $\sin(x) > 0$ *für* $x \in]0, 2[$.

Beweis. 1. Da wir $\cos(0) = 1$ schon im Beweis von Satz 11.12 gesehen haben, beginnen wir mit $\cos(2)$. Dies ist eine alternierende Reihe monoton fallender Glieder, denn:

$$\frac{2^{2k}}{2k!} > \frac{2^{2k+2}}{(2k+2)!} \iff (2k+1)(2k+2) > 2^2 \iff k \geq 1.$$

Es folgt:

$$-1 = 1 - \frac{2^2}{2} \leq \cos 2 \leq 1 - \frac{2^2}{2} + \frac{2^4}{4!} = 1 - 2 + \frac{16}{24} = -1 + \frac{2}{3} = -\frac{1}{3}.$$

2. Es gilt:
$$\frac{x^{2k+1}}{(2k+1)!} > \frac{x^{2k+3}}{(2k+3)!} \iff (2k+2)(2k+3) > x^2.$$
Für $k \geq 0$ und $0 \leq x \leq 2$ ist aber tatsächlich: $2k+3 \geq 2k+2 \geq x \geq 0$, d.h. $(2k+2)(2k+3) > x^2$. Für $x \in]0,2]$ folgt:
$$x \geq \sin x \geq x - \frac{x^3}{6} = x \cdot \Big(1 - \frac{x^2}{6}\Big) \geq x \cdot \Big(1 - \frac{2^2}{6}\Big) > 0,$$
was zu zeigen war.

□

Damit können wir nun π definieren:

Korollar/Definition 11.14. cos *ist in* $[0,2]$ *monoton fallend und wegen* $\cos(0) = 1$, $\cos(2) < 0$ *hat* cos *genau eine Nullstelle in* $[0,2]$. *Wir definieren die (auch* ***Kreiszahl*** *genannte) Zahl* π *durch:* $\frac{\pi}{2}$ *ist die Nullstelle von* cos *in* $[0,2]$. *Also:*
$$\cos\Big(\frac{\pi}{2}\Big) = 0, \quad \sin\Big(\frac{\pi}{2}\Big) = 1.$$

Leicht ergeben sich nun die folgenden Formeln:

Satz 11.15 (Verschiebungen von Sinus und Cosinus). *Es gilt:*

1. $\sin(x + \frac{\pi}{2}) = \cos x, \ \cos(x + \frac{\pi}{2}) = -\sin x.$
2. $\sin(x + \pi) = -\sin x, \ \cos(x + \pi) = -\cos x.$
3. $\sin(x + 2\pi) = \sin x, \ \cos(x + 2\pi) = \cos x.$

Man sagt, dass Sinus und Cosinus **periodische Funktionen** mit **Periode** 2π sind (Abb. 11.4). Der Wert von π ist
$$\pi = 3.1415\ldots$$

Bemerkung 11.16 (zur Bedeutung von Sinus und Cosinus).

1. $[0, 2\pi[\to \mathbb{R}^2, \ t \mapsto (\cos t, \sin t)$ parametrisiert den Einheitskreis.
2. Ist f eine Lösung der sogenannten **Differentialgleichung** (d.h. einer Gleichung, in der eine gesuchte Funktion y sowie eine oder mehrere ihrer Ableitungen auftreten)
$$y'' = -w^2 y,$$
so ist $f(x) = a\cos(wx) + b\sin(wx)$. Sinus und Cosinus tauchen bei der Beschreibung von Schwingungsvorgängen auf.

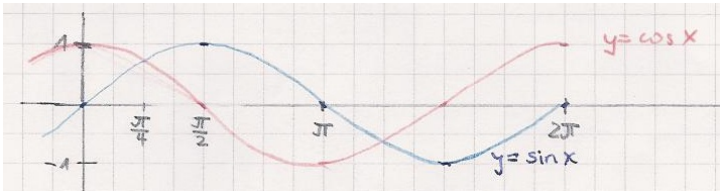


Abbildung 11.4. Funktionsgraphen von Sinus und Cosinus.

Definition 11.17. *Die Abbildung*

$$\tan\colon \mathbb{R} \setminus \left\{\frac{\pi}{2} + \pi k \mid k \in \mathbb{Z}\right\} \to \mathbb{R}, \quad x \mapsto \tan x = \frac{\sin x}{\cos x}$$

heißt ***Tangens****, sein Kehrwert*

$$\cot\colon \mathbb{R} \setminus \{k\pi \mid k \in \mathbb{Z}\} \to \mathbb{R}, \quad x \mapsto \cot x = \frac{1}{\tan x} = \frac{\cos x}{\sin x}$$

Cotangens*. Siehe auch Abb. 11.5.*

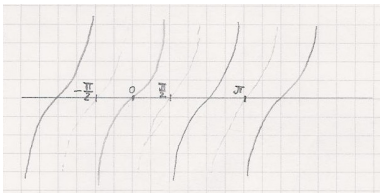


Abbildung 11.5. Funktionsgraph des Tangens.

Nach der Quotientenregel gilt:

$$\tan'(x) = \frac{\cos x \cos x - (-\sin x)\sin x}{\cos^2 x} = \frac{1}{\cos^2 x}.$$

Satz/Definition 11.18.

1. $\tan\colon \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[\to \mathbb{R}$ *ist streng monoton steigend. Die Umkehrfunktion*

$$\arctan\colon \mathbb{R} \to \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[\subset \mathbb{R}$$

heißt ***Arcustangens****.*

2. *Die Abbildung*

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

ist streng monoton steigend. Die Umkehrfunktion

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

heißt ***Arcussinus****. Siehe dazu auch Abb. 11.6.*

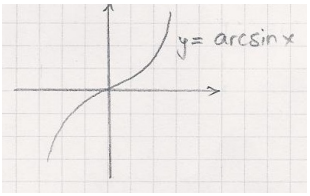


Abbildung 11.6. Funktionsgraph von Arcussinus.

Satz 11.19.

1. arcsin *ist auf* $]-\frac{\pi}{2}, \frac{\pi}{2}[$ *diffbar mit*

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

2. arctan *ist diffbar mit*

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

Beweis.

1. Wir haben schon gesehen, dass $\sin' = \cos$. Damit erhalten wir:

$$\begin{aligned}\arcsin'(x) &= \frac{1}{\cos(\arcsin(x))} \\ &= \frac{1}{\sqrt{1-\sin^2(\arcsin(x))}} \\ &= \frac{1}{\sqrt{1-x^2}}.\end{aligned}$$

2. Wir wissen bereits, dass $\tan'(x) = \frac{1}{\cos^2(x)}$. Damit folgt:

$$\begin{aligned}\arctan'(x) &= \cos^2(\arctan(x)) \\ &= \frac{\cos^2(\arctan(x))}{\sin^2(\arctan(x)) + \cos^2(\arctan(x))} \\ &= \frac{1}{1 + \tan^2(\arctan(x))} \\ &= \frac{1}{1 + x^2}.\end{aligned}$$

□

Analog kann man auch einen **Arcuscosinus** definieren und dessen Ableitung ausrechnen, nämlich

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

als Umkehrfunktion von $\cos\colon [0, \pi] \to [-1, 1]$. Mit den Additionstheoremen (Satz 7.25) sieht man recht leicht, dass man die Arcusfunktionen ineinander umrechnen kann:

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

Analog zur Ableitung des Arcussinus erhält man jene des Arcuscosinus:

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

Aufgaben

Aufgabe 11.1 (Eine Abschätzung für den Logarithmus). Seien $x, y \in \mathbb{R}$, $x, y > 0$, positive Zahlen. Zeigen Sie:

$$\frac{\ln x + \ln y}{2} \leq \ln \frac{x + y}{2}.$$

12

Asymptotisches Verhalten und Regel von L'Hospital

Grenzwerte rationaler Funktionen sind mit den bisherigen Mitteln oft nicht einfach zu berechnen. Die Regel von L'Hospital[1] ist in solchen Situationen oft hilfreich. Insbesondere werden wir damit das asymptotische Verhalten rationaler Funktionen recht einfach untersuchen können.

12.1 Die Regel von L'Hospital

Satz 12.1 (Regel von L'Hospital). *Seien* $f, g\colon [a,b] \to \mathbb{R}$ *stetige Funktionen, auf* $]a,b[$ *differenzierbar mit* $g'(x) \neq 0\ \forall x \in]a,b[$ *und* $f(a) = g(a) = 0$. *Existiert* $\lim_{x \searrow a} \frac{f'(x)}{g'(x)}$, *dann existiert auch* $\lim_{x \searrow a} \frac{f(x)}{g(x)}$ *und es gilt:*

$$\lim_{x \searrow a} \frac{f(x)}{g(x)} = \lim_{x \searrow a} \frac{f'(x)}{g'(x)}.$$

Bevor wir dies beweisen, zunächst ein Beispiel:

Beispiel 12.2. $f(x) = \sin(x)$, $g(x) = e^x - 1$, $[a,b] = [0,1]$. Der Quotient $\frac{f(0)}{g(0)} = \frac{0}{0}$ macht keinen Sinn, aber

$$\frac{f'(x)}{g'(x)} = \frac{\cos x}{e^x}$$

ist stetig in $x = 0$ mit

[1] Wikipedia sagt dazu: Die Regel ist nach Guillaume Francois Antoine, Marquis de L'Hospital (1661–1704) benannt. L'Hospital veröffentlichte sie 1696 in seinem Buch *Analyse des infiniment petits pour l'intelligence des lignes courbes*, dem ersten Lehrbuch der Differentialrechnung. Er hatte sie aber nicht selbst entdeckt, sondern von Johann Bernoulli übernommen.

$$\lim_{x \searrow 0} \frac{f'(x)}{g'(x)} = \frac{\cos 0}{e^0} = \frac{1}{1} = 1.$$

Also existiert

$$\lim_{x \searrow 0} \frac{\sin x}{e^x - 1} = \lim_{x \searrow 0} \frac{\cos x}{e^x} = \frac{1}{1} = 1.$$

Die Idee des Beweises der Regel von L'Hospital ist es, ein Analogon des Mittelwertsatzes anzuwenden, nämlich folgendes:

Lemma 12.3. *Seien $f, g\colon [a,b] \to \mathbb{R}$ stetig, auf $]a,b[$ diffbar mit $g'(x) \neq 0\ \forall x \in]a,b[$ und $g(a) \neq g(b)$. Dann existiert ein $\xi \in]a,b[$, so dass:*

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$

Beweis. Wir betrachten die Funktion

$$h(x) = f(x) - \frac{f(b) - f(a)}{g(b) - g(a)} \cdot \big(g(x) - g(a)\big).$$

Es gilt offenbar: $h(a) = f(a) = h(b)$. Mit dem Satz von Rolle existiert daher ein $\xi \in]a,b[$, so dass:

$$0 = h'(\xi) = f'(\xi) - \frac{f(b) - f(a)}{g(b) - g(a)} g'(\xi).$$

Da $g'(\xi) \neq 0$ nach Voraussetzung, folgt:

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

□

Beweis (von L'Hospitals Regel, Satz 12.1). Da $g'(x) \neq 0\ \forall x \in]a,b[$ und $g(a) = 0$, ist $g(x) \neq 0\ \forall x \in]a,b[$ nach dem Satz von Rolle. Ferner gilt nach dem Lemma:

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f'(\xi)}{g'(\xi)}$$

für ein $\xi \in]a,x[$. Mit $x \searrow a$ strebt auch $\xi \searrow a$. Also:

$$\lim_{x \searrow a} \frac{f(x)}{g(x)} = \lim_{\xi \searrow a} \frac{f'(\xi)}{g'(\xi)} = \lim_{x \searrow a} \frac{f'(\xi)}{g'(\xi)}.$$

□

Von der sehr nützlichen Regel von L'Hospital gibt es viele Varianten. Um einige wichtige davon formulieren zu können, benötigen wir folgende Grenzwertbegriffe:

Definition 12.4 (Konvergenz für $x \to \infty$). *Sei $f\colon [a,\infty[\to \mathbb{R}$ eine Funktion. Wir sagen, $f(x)$ **strebt gegen** $c \in \mathbb{R}$ für x gegen ∞, in Zeichen*

$$\lim_{x\to\infty} f(x) = c,$$

falls

$$\forall \varepsilon > 0\ \exists N:\ |f(x) - c| < \varepsilon\ \forall x \geq N.$$

Wir sagen: $\lim_{x\to\infty} f(x) = \infty$, *falls*

$$\forall M > 0\ \exists N > 0:\ f(x) > M\ \forall x > N.$$

Für $f\colon]a,b] \to \mathbb{R}$ schreiben wir $\lim_{x\searrow a} f(x) = \infty$, falls

$$\forall M > 0\ \exists \varepsilon > 0:\ f(x) > M\ \forall x > a \text{ mit } |x - a| < \varepsilon.$$

Analog lassen sich $\lim_{x\to-\infty} f(x) = c$ oder etwa $\lim_{x\nearrow b} f(x) = -\infty$ definieren.

Die Varianten folgen direkt aus der ursprünglichen Regel:

Korollar 12.5 (Varianten der Regel von L'Hospital).

1. *Seien $f, g\colon]a,b[\to \mathbb{R}$ diffbare Funktionen mit $g'(x) \neq 0\ \forall x \in]a,b[$ und*

   $$\lim_{x\searrow a} f(x) = \infty = \lim_{x\searrow a} g(x).$$

   Existiert $\lim_{x\searrow a} \frac{f'(x)}{g'(x)}$, dann existiert auch $\lim_{x\searrow a} \frac{f(x)}{g(x)}$ und es gilt:

   $$\lim_{x\searrow a} \frac{f(x)}{g(x)} = \lim_{x\searrow a} \frac{f'(x)}{g'(x)}.$$

2. *Seien $f, g\colon]a,\infty[\to \mathbb{R}$ diffbare Funktionen mit $g'(x) \neq 0\ \forall x \in]a,\infty[$ und*

   $$\lim_{x\to\infty} f(x) = 0 = \lim_{x\to\infty} g(x)$$

   oder

   $$\lim_{x\to\infty} f(x) = \infty = \lim_{x\to\infty} g(x).$$

   Existiert $\lim_{x\to\infty} \frac{f'(x)}{g'(x)}$, dann existiert auch $\lim_{x\to\infty} \frac{f(x)}{g(x)}$ und es gilt:

   $$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}.$$

Beispiel 12.6. Wir zeigen: Für jedes $n \in \mathbb{N}$ ist

$$\lim_{x\to\infty} \frac{x^n}{e^x} = 0.$$

Es gilt: $\lim_{x\to\infty} x^n = \lim_{x\to\infty} e^x = \infty$. Die Regel von L'Hospital liefert dann:

$$\lim_{x\to\infty} \frac{x^n}{e^x} = \lim_{x\to\infty} \frac{nx^{n-1}}{e^x} = \cdots = n! \lim_{x\to\infty} \frac{1}{e^x} = 0.$$

Man sagt: e^x wächst schneller als jedes Polynom.

12.2 Asymptotisches Verhalten rationaler Funktionen

Für Folgen haben wir in Abschnitt 5.3 die O– und o–Notation eingeführt. Analog nun für Funktionen, um Aussagen wie die obige, dass e^x schneller als jedes Polynom wächst, präzise formulieren zu können.

Definition 12.7 (O– und o– Notation für Funktionen). *Seien $f, g\colon [a, \infty[\to \mathbb{R}$ Funktionen. Wir schreiben*

$$f \in O(g) \text{ für } x \to \infty,$$

*falls $\exists\, c > 0\ \exists\, M$, so dass $|f(x)| \leq c \cdot g(x)\ \forall x \geq M$, und sagen f **liegt in groß** O **von** g. Wir sagen $f \in o(g)$, f **liegt in klein** o **von** g, falls $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$.*

Beispiel 12.8.

1. $x^n \in o(e^x)$ für $x \to \infty$ für jedes $n \in \mathbb{N}$, wie wir gerade gesehen haben.
2. Sei $f(x) = a_n x^n + \cdots + a_0 \in \mathbb{R}[x]$ ein Polynom. Dann gilt: $f(x) \in O(x^n)$ für $x \to \infty$. Genauer gilt: Für jedes $C = |a_n| + \varepsilon, \varepsilon > 0,\ \exists\, M > 0$, so dass:

$$|f(x)| \leq C \cdot x^n\ \forall x \geq M.$$

Sei $h(x) = \frac{f(x)}{g(x)}$ eine rationale Funktion mit Polynomen

$$f(x) = a_n x^n + \cdots + a_0 \in \mathbb{R}[x],$$
$$g(x) = b_m x^m + \cdots + b_0 \in \mathbb{R}[x],$$

vom Grad n bzw. m, d.h. $a_n, b_m \neq 0$. Dann gilt, beispielsweise mit der Regel von L'Hospital:

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0, \quad \text{falls } n < m,$$
$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \frac{a_n}{b_m}, \quad \text{falls } n = m,$$
$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \begin{cases} +\infty, & \text{falls } n > m \text{ und } \frac{a_n}{b_m} > 0, \\ -\infty, & \text{falls } n > m \text{ und } \frac{a_n}{b_m} < 0. \end{cases}$$

Im letzten Fall lässt sich eine wesentlich präzisere Aussage machen:

Satz 12.9 (Division mit Rest). *Seien $f, g \in \mathbb{R}[x]$ Polynome in einer Variablen x mit reellen Koeffizienten. Dann existieren eindeutig bestimmte Polynome $q(x), r(x) \in \mathbb{R}[x]$, so dass:*

$$f(x) = q(x) \cdot g(x) + r(x) \quad \textit{und} \quad \deg r < \deg g.$$

Beweis. Zunächst zur Existenz, mit Induktion nach $\deg f$. Für $\deg f < \deg g$ können wir $q = 0$ und $r = f$ wählen. Ist $n = \deg f \geq \deg g = m$, etwa

$$f = a_n x^n + \cdots + a_0, \quad g = b_m x^m + \cdots + b_0$$

mit $a_n \neq 0 \neq b_m$, so betrachten wir

$$q_0 := \frac{a_n}{b_m} x^{n-m} \text{ und } f_1 := f - q_0 \cdot g.$$

Es gilt: $\deg f_1 < \deg f$. Wir können daher induktiv voraussetzen, dass eine Darstellung $f_1 = q_1 g + r_1$ existiert, also:

$$f = f_1 + q_0 \cdot g = (q_0 + q_1) \cdot g + r_1.$$

Nun zur Eindeutigkeit: Angenommen, $f = q \cdot g + r$ und $f = \tilde{q} \cdot g + \tilde{r}$ mit $\deg g > \deg r$, $\deg g > \deg \tilde{r}$, sind zwei verschiedene Darstellungen, d.h. insbesondere $q \neq \tilde{q}$. Dann ist

$$0 = \underbrace{(q - \tilde{q})}_{=:\bar{q}} \cdot g + \underbrace{(r - \tilde{r})}_{=:\bar{r}} \text{ mit } \bar{q} \neq 0, \text{ d.h. } \bar{q} \cdot g + \bar{r} = 0 \text{ mit } \deg(\bar{q}) \geq 0.$$

Es folgt:

$$\deg(\bar{q} \cdot g) \geq 0 + \deg g > \deg(\bar{r}),$$

also $\bar{q} \cdot g + \bar{r} \neq 0$, ein Widerspruch. Also: $q = \tilde{q}$ und $r = \tilde{r}$. □

Beispiel 12.10. Wir betrachten die rationale Funktion

$$h(x) = \frac{f(x)}{g(x)} = \frac{x^3}{x^2 - 1}.$$

Der Beweis des Satzes zur Division mit Rest gibt einen Algorithmus an, um diese durchzuführen. Hier ergibt sich:

$$\begin{array}{l} x^3 : (x^2 - 1) = x + \frac{x}{x^2-1}, \\ \underline{x^3 - x} \\ \quad x \end{array}$$

d.h. $q(x) = x$, $r(x) = x$, also $f(x) = x^3 = x \cdot (x^2 - 1) + x = q(x) \cdot g(x) + r(x)$.

Es folgt: $h(x) \in x + o(1)$, d.h. asymptotisch verhält sich $h(x)$ in etwa wie x. In der Nähe von 0 unterscheidet sich h allerdings sehr von der Funktion $x \mapsto x$. Beispielsweise hat h die Polstellen $x = \pm 1$ und einen Sattelpunkt in 0 (siehe auch Abb. 12.1).

Wie wir im Beispiel gesehen haben, liefert Division mit Rest sofort auch eine Aussage über das asymptotische Verhalten rationaler Funktionen:

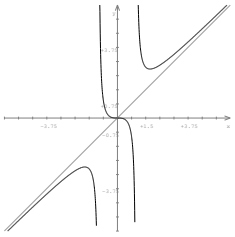


Abbildung 12.1. Graph einer rationalen Funktion mit Polstellen bei $x = \pm 1$, einem Sattelpunkt in $x = 0$ und asymptotischen Verhalten wie die Gerade $x \mapsto x$.

Korollar 12.11. *Sei* $h(x) = \frac{f(x)}{g(x)}$ *eine rationale Funktion und*

$$f(x) = q(x) \cdot g(x) + r(x)$$

mit $\deg r < \deg g$. *Dann verhält sich* h *für* $x \to \pm\infty$ *wie* $q(x)$, *genauer:*

$$h(x) \in q(x) + o(1).$$

Um zu sehen, dass Asymptoten nicht immer Geraden sein müssen, betrachten wir zum Abschluss noch ein etwas komplizierteres Beispiel:

Beispiel 12.12. Wir betrachten die rationale Funktion $h(x) = \frac{x^4+1}{x^2+1}$. Division mit Rest liefert:

$$\begin{array}{l} (x^4+1) : (x^2+1) = x^2 - 1 + \frac{2}{x^2+1}, \\ \underline{x^4 + x^2} \\ \quad -x^2 + 1 \\ \quad \underline{-x^2 - 1} \\ \qquad\quad 2 \end{array}$$

d.h. $q = x^2 - 1$ und $r = 2$. Offenbar hat $h(x)$ keine Polstelle. Wir bestimmen die Extrema, um eine Skizze des Graphen von $h(x)$ zeichnen zu können. Für die Ableitung ergibt sich:

$$h'(x) = \frac{(x^2+1) \cdot 4 \cdot x^3 - 2 \cdot x \cdot (x^4+1)}{(x^2+1)^2} = \frac{2x^5 + 4x^3 - 2x}{(x^2+1)^2}.$$

Eine Extremstelle muss also erfüllen $2x^5 + 4x^3 - 2x = 0$, d.h. $x = 0$ oder $x^4 + 2x^2 - 1 = 0$. Mit $z = x^2$ erhalten wir die **quadratische Gleichung** $z^2 + z - 1 = 0$, für die die p, q–Formel folgende beiden Lösungen ergibt:

$$z_{1,2} = -1 \pm \sqrt{1+1}.$$

Da $-1-\sqrt{2}<0$ keine reelle Lösung für x liefert, verbleiben die beiden Stellen

$$x_{1,2} = \pm\sqrt{-1+\sqrt{2}} \approx \pm 0.64.$$

Man kann nachrechnen, dass $x = 0$ ein lokales Maximum und $x_{1,2}$ lokale Minima sind. Da sich $h(x)$ für $x \to \pm\infty$ wie $q = x^2 - 1$ verhält, erhalten wir in etwa das in Abb. 12.2 gezeigte Schaubild.

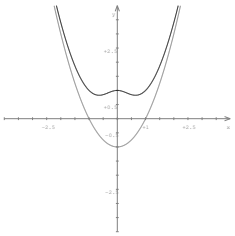


Abbildung 12.2. Eine rationale Funktion mit der Parabel $x \mapsto x^2 - 1$ als Asymptote.

Aufgaben

Aufgabe 12.1 (Wachstumsverhalten gegen Unendlich). Sortieren Sie die Funktionen

$$\begin{array}{ll} f_1(x) = x^{\ln x} & f_4(x) = 3^x \\ f_2(x) = e^{x \ln x} & f_5(x) = x^3 \\ f_3(x) = x^x & f_6(x) = e^x \ln x \end{array}$$

nach dem Wachstum für $x \to \infty$ (Begründung!).

Aufgabe 12.2 (Grenzwerte). Zeigen Sie:

$$\lim_{x\to 0} \frac{2\cos x + e^x + e^{-x} - 4}{x^4} = \frac{1}{6},$$

$$\lim_{x\to 0} \frac{\sqrt{\cos ax} - \sqrt{\cos bx}}{x^2} = \frac{b^2 - a^2}{4} \quad \text{für } a, b \in \mathbb{R}.$$

Aufgabe 12.3 (Grenzwerte). Prüfen Sie, ob folgende Grenzwerte existieren, und bestimmen Sie diese gegebenenfalls:

1. $\lim_{x\searrow 0} \frac{\ln x}{\cot x}$,

2. $\lim_{x\to\frac{\pi}{2}} \frac{\tan(3x)}{\tan(x)}$,
3. $\lim_{x\searrow 1}\Big(\ln(x)\cdot\ln(1-x)\Big)$.

Aufgabe 12.4 (Die Eulersche Zahl). Zeigen Sie $\lim_{n\to\infty}\Big(n\ln(1+\frac{1}{n})\Big) = 1$ und folgern Sie daraus:

$$\lim_{n\to\infty}\Big(1+\frac{1}{n}\Big)^n = e.$$

13

Integration

Sei $f\colon [a,b] \to \mathbb{R}$ eine Funktion auf einem abgeschlossenen Intervall. Wir wollen die Fläche unter dem Graphen von f bestimmen.

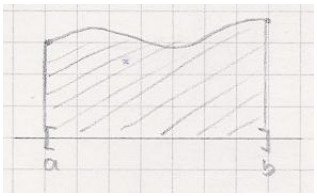


Abbildung 13.1. Die Fläche unter einem Graphen.

Grundidee ist es, ein Approximationsverfahren zu verwenden. Schließen wir f durch zwei sogenannte Treppenfunktionen ein, $\varphi \leq f \leq \psi$, so ist klar, dass die Fläche unter f größer als die unter φ und kleiner als die unter ψ ist (Abb. 13.2).

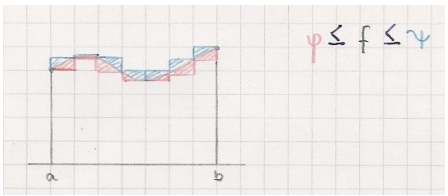


Abbildung 13.2. Approximation durch Treppenfunktionen.

13.1 (Riemann–)Integrierbarkeit

Definition 13.1. *Eine* ***Treppenfunktion*** $\varphi\colon [a,b] \to \mathbb{R}$ *ist eine Funktion, zu der es eine Unterteilung* $a = t_0 < t_1 < \cdots < t_n = b$ *des Intervalls* $[a,b]$ *gibt, so dass* φ *auf den offenen Intervallen* $]t_{i-1}, t_i[$ *konstant ist, d.h. für jedes* $i \in \{1,\dots,n\}$ *gibt es ein* $c_i \in \mathbb{R}$*, so dass für die Einschränkung* $\varphi\big|_{]t_{i-1},t_i[}$ *gilt:*

$$\varphi\big|_{]t_{i-1},t_i[}\colon\ ]t_{i-1},t_i[\to \mathbb{R}, \quad \varphi\big|_{]t_{i-1},t_i[}(x) = \varphi(x) = c_i.$$

Für $\varphi(t_i)$ *ist nichts vorausgesetzt. Das* ***Integral einer Treppenfunktion*** *ist*

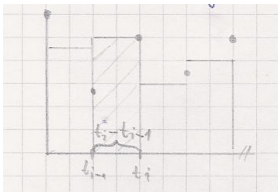


Abbildung 13.3. Treppenfunktionen auf Teilintervallen.

$$\int_a^b \varphi(x)\,dx := \sum_{i=1}^{n} c_i(t_i - t_{i-1}).$$

Eine solche Summe heißt ***Riemmannsche Summe****. Mit deren Hilfe können wir Integrale komplizierterer Funktionen definieren: Sei dazu* $f\colon [a,b] \to \mathbb{R}$ *eine beliebige beschränkte Funktion. Das* ***Oberintegral*** *von* f *ist*

$$\int_a^{*b} f := \inf\Big\{ \int_a^b \psi\,dx \;\Big|\; \psi \geq f,\ \psi \text{ Treppenfunktion} \Big\},$$

das ***Unterintegral***

$$\int_{*a}^b f := \sup\Big\{ \int_a^b \varphi\,dx \;\Big|\; \varphi \leq f,\ \varphi \text{ Treppenfunktion} \Big\}.$$

f *heißt* ***integrierbar*** *(genauer:* ***Riemann–integrierbar****), falls*

$$\int_a^{*b} f = \int_{*a}^b f$$

gilt. In diesem Fall definieren wir das ***Integral der beschränkten Funktion*** $f\colon [a,b] \to \mathbb{R}$ *durch:*

$$\int_a^b f(x)\,dx := \int_a^{*b} f = \int_{*a}^b f.$$

Bemerkung 13.2. 1. f beschränkt ist notwendig, damit es Treppenfunktionen φ, ψ mit $\varphi \le f \le \psi$ gibt.

2. f ist genau dann integrierbar, wenn es zu jedem $\varepsilon > 0$ Treppenfunktionen φ, ψ mit $\varphi \le f \le \psi$ gibt, so dass

$$(\,0 \le\,) \int_a^b (\psi - \varphi)\, dx < \varepsilon.$$

Beispiel 13.3. Treppenfunktionen sind integrierbar.

Satz 13.4. *Monotone Funktionen* $[a,b] \to \mathbb{R}$ *sind integrierbar.*

Beweis. Sei $f\colon [a,b] \to \mathbb{R}$ monoton steigend und $\varepsilon > 0$ vorgegeben. Wir wählen n so groß, dass

$$\varepsilon > \frac{1}{n} \cdot (b-a) \cdot \Big(f(b) - f(a)\Big),$$

setzen $h = \frac{b-a}{n}$ und betrachten die Zerlegung

$$t_i = a + i{\cdot}h,\ i = 0, \dots, n.$$

Für die Treppenfunktionen φ, ψ mit

$$\varphi\Big|_{[t_{i-1},t_i[} = f(t_{i-1}), \quad \psi\Big|_{[t_{i-1},t_i[} = f(t_i)$$

und $\varphi(b) = \psi(b) = f(b)$ gilt dann $\varphi \le f \le \psi$ wegen der Monotonie. Andererseits:

$$\begin{aligned}\int \psi\, dx - \int \varphi\, dx &= \sum_{i=1}^{n} \Big(f(t_i) - f(t_{i-1})\Big){\cdot}h \\ &= h{\cdot}\Big(f(b) - f(a)\Big) = \frac{b-a}{n}{\cdot}\Big(f(b) - f(a)\Big) < \varepsilon.\end{aligned}$$

□

Beispiel 13.5. Seien $0 \le b$. Was ist $\int_0^b x^2\, dx$? $f(x) = x^2$ ist monoton auf $[0,b]$, also existiert das Integral. Zur sogenannten **äquidistanten Unterteilung**

$$t_i = i \cdot \frac{b}{n}, \quad i = 0, \dots, n,$$

des Intervalls $[0,b]$ und der Treppenfunktion ψ mit $\psi\Big|_{]t_{i-1},t_i[} = f(t_i)$ und $h = \frac{b}{n}$ gilt:

$$\int_0^b \psi\, dx = \sum_{i=1}^{n} i^2{\cdot}h^2{\cdot}h = \frac{n(n+1)(2n+1)}{6} \cdot \frac{b^3}{n^3} \underset{n\to\infty}{\longrightarrow} \frac{b^3}{3}.$$

Also: $\int_0^b x^2\, dx = \frac{b^3}{3}$.

Beispiel 13.6. Die Funktion

$$f\colon [0,1] \to \mathbb{R},\ f(x) = \begin{cases} 1, & \text{für } x \in \mathbb{Q}, \\ 0, & \text{für } x \notin \mathbb{Q}, \end{cases}$$

ist nicht integrierbar, denn für jedes Paar φ, ψ von Treppenfunktionen mit $\varphi \leq f \leq \psi$ gilt:

$$\int_0^1 \varphi(x)\,dx \leq 0, \quad \int_0^1 \psi(x)\,dx \geq 1,$$

da wegen $]t_{i-1}, t_i[\cap \mathbb{Q} \neq \emptyset$ für φ gilt: $\varphi\big|_{]t_{i-1},t_i[} \leq 0$ und analog $\psi\big|_{]t_{i-1},t_i[} \geq 1$ wegen $]t_{i-1}, t_i[\cap (\mathbb{R}\backslash\mathbb{Q}) \neq \emptyset$.

Satz 13.7 (Integrierbarkeit stetiger Funktionen). *Sei $f\colon [a,b] \to \mathbb{R}$ stetig. Dann ist f über $[a,b]$ integrierbar.*

Für den Beweis benötigen wir mehr als nur die punktweise Stetigkeit:

Definition 13.8. *Eine Funktion $f\colon I \to \mathbb{R}$ heißt **gleichmäßig stetig**, wenn $\forall \varepsilon > 0$ $\exists\, \delta$, so dass*

$$|f(x_1) - f(x_0)| < \varepsilon \ \forall\ x_0, x_1 \in I \text{ mit } |x_1 - x_0| < \delta.$$

Der entscheidende Unterschied zur Stetigkeit in allen Punkten ist, dass hier $\delta = \delta(\varepsilon)$ nicht von x_0 abhängt, sondern nur von ε.

Beispiel 13.9. Die Funktion

$$f\colon \mathbb{R}_{>0} \to \mathbb{R},\ f(x) = \frac{1}{x}$$

ist stetig, aber nicht gleichmäßig stetig. Zu $\varepsilon > 0$ und $x_0 \to 0$ muss $\delta = \delta(\varepsilon, x_0)$ immer kleiner gewählt werden, wie man leicht nachrechnen kann.

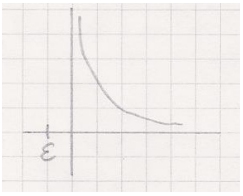


Abbildung 13.4. $1/x$ ist nicht gleichmäßig stetig.

Satz 13.10. *Sei $f\colon [a,b] \to \mathbb{R}$ eine auf einem abgeschlossenen, beschränkten Intervall stetige Funktion. Dann ist f gleichmäßig stetig.*

Beweis. Angenommen, f ist nicht gleichmäßig stetig, d.h. $\exists\ \varepsilon > 0$, so dass eine Folge (δ_n) mit $\delta_n \xrightarrow[n\to\infty]{} 0$ gibt und Folgen (x_n), (y_n), so dass

$$|f(x_n) - f(y_n)| \geq \varepsilon, \text{ obwohl } |x_n - y_n| < \delta_n.$$

Nach Bolzano–Weierstrass hat die Folge (x_n) eine konvergente Teilfolge (x_{n_k}); sei

$$x_0 = \lim_{k\to\infty}(x_{n_k}) \in [a,b]$$

deren Grenzwert. Da f in x_0 stetig ist, existiert zu $\frac{\varepsilon}{2}$ ein δ, so dass

$$|f(x) - f(x_0)| < \frac{\varepsilon}{2} \text{ für } x \in [a,b] \text{ mit } |x - x_0| < \delta.$$

Wir wählen jetzt k so groß, dass $|x_{n_k} - x_0| < \frac{\delta}{2}$ und $\delta_{n_k} < \frac{\delta}{2}$. Dann gilt auch:

$$|y_{n_k} - x_0| \leq |y_{n_k} - x_{n_k}| + |x_{n_k} - y_0| \leq \delta_{n_k} + \frac{\delta}{2} < \delta$$

und

$$|f(x_{n_k}) - f(y_{n_k})| \leq |f(x_{n_k}) - f(x_0)| + |f(y_{n_k}) - f(x_0)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

im Widerspruch zu $|f(x_n) - f(y_n)| \geq \varepsilon\ \forall n$. □

Beweis (von Satz 13.7 über die Integrierbarkeit stetiger Funktionen). Es sei $f\colon [a,b] \to \mathbb{R}$ stetig. Nach Satz 13.10 ist f sogar gleichmäßig stetig. Sei $\varepsilon > 0$ nun vorgegeben. Wir konstruieren Treppenfunktionen

$$\varphi \leq f \leq \psi \text{ mit } \int_a^b (\psi - \varphi)\,dx < \varepsilon.$$

Zu $\frac{\varepsilon}{b-a} > 0$ wählen wir ein $\delta > 0$, so dass

$$|f(x) - f(y)| < \frac{\varepsilon}{b-a}\ \forall\, x, y \in [a,b] \text{ mit } |x - y| < \delta.$$

Wir wählen dann n so groß, dass $h = \frac{b-a}{n} < \delta$ und $t_i = a + i{\cdot}h$. Dann sei $\varphi\big|_{[t_{i-1},t_i[} = \min\{f(x) \mid x \in [t_{i-1},t_i]\}$ und $\psi\big|_{[t_{i-1},t_i[} = \max\{f(x) \mid x \in [t_{i-1},t_i]\}$. Da Minimum und Maximum angenommen werden und $h < \delta$ ist, gilt

$$0 \leq \psi(x) - \varphi(x) < \frac{\varepsilon}{b-a} \Rightarrow 0 \leq \int_a^b (\psi(x) - \varphi(x))\,dx < \frac{\varepsilon}{b-a}{\cdot}(b-a) < \varepsilon.$$

□

Satz 13.11 (Eigenschaften des Integrals). *Es seien $f, g\colon [a,b] \to \mathbb{R}$ integrierbare Funktionen, $c \in \mathbb{R}$ eine Konstante. Es gilt:*

1. (***Linearität des Integrals***) *Auch $c\cdot f$ und $f+g$ sind integrierbar mit*
$$\int_a^b c\cdot f(x)\,dx = c\cdot\int_a^b f(x)\,dx,$$
$$\int_a^b \big(f(x)+g(x)\big)\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.$$
2. (***Monotonie des Integrals***) $f\le g \Rightarrow \int_a^b f(x)\,dx \le \int_a^b g(x)\,dx$.
3. *Mit f sind auch die Funktionen $f_+ := \max(f,0)$, $f_- := \max(-f,0)$ und $|f| = f_+ + f_-$ integrierbar.*
4. *Zu $p\in\mathbb{R}_{>0}$ ist auch $|f|^p$ integrierbar. Insbesondere ist auch*
$$f\cdot g = \frac{1}{4}\big(|f+g|^2 - |f-g|^2\big)$$
integrierbar.

Beweis. 1. Mit Treppenfunktionen $\varphi\le f, \psi\le g$ ist auch $\varphi+\psi$ eine Treppenfunktion und es gilt: $\varphi+\psi\le f+g$. Es folgt:
$$\int_{a*}^b (f+g)\,dx \ge \int_{a*}^b f\,dx + \int_{a*}^b g\,dx,$$
da wir das Supremum nehmen. Analog:
$$\int^{*} (f+g)\,dx \le \int^{*} f\,dx + \int^{*} g\,dx$$
und die Kette
$$\int_{*} f\,dx + \int_{*} g\,dx \le \int_{*}(f+g)\,dx \le \int^{*}(f+g)\,dx \le \int^{*} f\,dx + \int^{*} g\,dx$$
impliziert Gleichheit.

2. Aus $f\ge g$ folgt ($\varphi\le f \Rightarrow \varphi\le g$) und daher
$$\int_{*} f\,dx \le \int_{*} g\,dx.$$
Analog: $\int^{*} f\,dx \le \int^{*} g\,dx$.

3. Mit φ ist auch $\varphi_+ = \max(\varphi,0)$ eine Treppenfunktion und $\varphi\le f\le\psi \Rightarrow \varphi_+\le f_+\le\psi_+$. Außerdem ist
$$0\le \int_a^b (\psi_+ - \varphi_+)\,dx \le \int_a^b (\psi-\varphi)\,dx < \varepsilon$$
für geeignete φ,ψ. Also ist f_+ integrierbar und dann auch $f_- = -(f-f_+)$ und $|f| = f_+ + f_-$ wegen der Linearität des Integrals.

4. Da f beschränkt ist, können wir wegen der Linearität des Integrals $0 \leq |f| \leq 1$ annehmen. Für Treppenfunktionen $0 \leq \varphi \leq |f| \leq \psi \leq 1$ gilt dann

$$\varphi^p \leq |f|^p \leq \psi^p \text{ und } 0 \leq (\psi^p - \varphi^p) \leq p(\psi - \varphi)$$

nach dem Mittelwertsatz, angewendet auf die Funktion $x \mapsto x^p$ auf dem Intervall $[0,1]$:

$$\frac{b^p - a^p}{b-a} = p\xi^{p-1} \leq p \text{ für } [a,b] \in [0,1].$$

Also:

$$\int_a^b (\psi^p - \varphi^p) \leq p \int_a^b (\psi - \varphi)\, dx < \varepsilon$$

für φ, ψ geeignet. Schließlich ist $f \cdot g$ integrierbar wegen der Formel

$$f \cdot g = \frac{1}{4}\Big((f+g)^2 - (f-g)^2\Big),$$

dem Bewiesenen für $p = 2$ und der Linearität des Integrals.

□

Satz 13.12 (Mittelwertsatz der Integralrechnung). *Seien $f, g\colon [a,b] \to \mathbb{R}$ Funktionen, f stetig, g integrierbar und $g(x) \geq 0\ \forall x$. Dann existiert ein $\xi \in [a,b]$, so dass*

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

Insbesondere: $\exists\ \xi \in [a,b]$, so dass

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

Beweis. Seien

$$M = \max\{f(x) \mid x \in [a,b]\},\ \ m = \min\{f(x) \mid x \in [a,b]\}.$$

Dann gilt:

$$mg(x) \leq f(x)g(x) \leq Mg(x),$$

da $g \geq 0$. Die Monotonie des Integrals ergibt:

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

Ist $\int_a^b g(x)\, dx = 0$, dann ist nichts mehr zu zeigen. Andernfalls existiert nach dem Zwischenwertsatz ein ξ, so dass

$$m \leq \frac{\int_a^b f(x)g(x)\, dx}{\int_a^b g(x)\, dx} = f(\xi) \leq M.$$

Die Behauptung folgt. Der Spezialfall ist der Fall $g(x) = 1\ \forall x \in [a,b]$. □

Die Berechnung von Integralen mittels Ober– und Untersumme und Grenzwertbildung ist mühselig. Die Hauptmethode, Integrale zu bestimmen, ist es, sämtliche Integrale

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

für eine variable Obergrenze t gleichzeitig zu bestimmen.

Satz 13.13. *Sei $f\colon [a,b] \to \mathbb{R}$ eine integrierbare Funktion. Dann ist f auch über jedem abgeschlossenen Teilintervall von $[a,b]$ integrierbar und es gilt (s. Abb. 13.5):*

$$\int_a^t f(x)\,dx + \int_t^b f(x)\,dx = \int_a^b f(x)\,dx \;\forall t \in]a,b[.$$

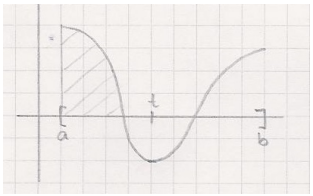


Abbildung 13.5. Integrierbarkeit auf Teilintervallen.

Beweis. Schränke jede Approximation durch Treppenfunktionen auf das Teilintervall ein. □

Definition 13.14. *Sei $f\colon [a,b] \to \mathbb{R}$ integrierbar. Wir setzen*

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

für vertauschte Ober– und Untergrenze.

13.2 Stammfunktionen

Bemerkung/Definition 13.15. Sei $f\colon I \to \mathbb{R}$ eine stetige Funktion. Eine **Stammfunktion** $F\colon I \to \mathbb{R}$ ist eine diffbare Funktion mit $F' = f$. F ist durch f bis auf eine Konstante eindeutig bestimmt.

Beweis. Ist $G\colon I \to \mathbb{R}$ eine weitere Stammfunktion, dann gilt:

$$(G - F)' = f - f = 0,$$

also $G - F = c$ eine konstante Funktion bzw. $G = F + c$. □

Satz 13.16 (Hauptsatz der Differential– und Integralrechnung). *Sei $f\colon I \to \mathbb{R}$ eine stetige Funktion und $a \in I$. Dann gilt:*

1. *$F\colon I \to \mathbb{R}$ mit $F(x) = \int_a^x f(t)\,dt$ ist eine Stammfunktion von f.*
2. *Ist G eine Stammfunktion von f, so gilt:*

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

Es gibt zwei übliche Kurzschreibweisen hierfür:

$$G(x)\Big|_a^b := \Big[G(x)\Big]_a^b := G(b) - G(a).$$

Beweis. 1. Nach dem Mittelwertsatz der Integralrechnung gilt:

$$F(x) - F(x_0) = \int_{x_0}^{x} f(t)\,dt = f(\xi)(x - x_0)$$

für einen Wert $\xi \in]x, x_0[$ (Abb. 13.6), also:

$$\frac{F(x) - F(x_0)}{x - x_0} = f(\xi).$$

Da mit $x \to x_0$ auch $\xi \to x_0$ und f stetig ist, folgt:

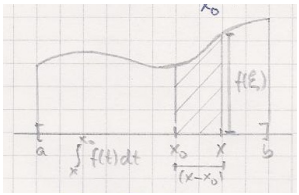


Abbildung 13.6. Anwendung des MWS der Integralrechnung.

$$F'(x_0) = \lim_{x \to x_0} \frac{F(x) - F(x_0)}{x - x_0} = \lim_{\xi \to x_0} f(\xi) = f(x_0).$$

2. Nach der Bemerkung 13.15 von eben gilt:
$$G(x) = F(x) + c$$
für ein gewisses $c \in \mathbb{R}$. Es folgt:
$$G(b) - G(a) = F(b) - F(a) = \int_a^b f(x)\,dx$$
nach der Definition von F.

□

Definition 13.17. *Das **unbestimmte Integral** $\int f(x)\,dx$ bezeichnet eine Stammfunktion von f.*

Beispiel 13.18. Wir haben bereits gesehen, dass die folgenden Stammfunktionen tatsächlich die behaupteten Ableitungen haben:

1. $\int x^\alpha\,dx = \frac{x^{\alpha+1}}{\alpha+1}, \quad \alpha \neq -1.$
2. $\int \frac{1}{x}\,dx = \ln|x|.$
3. $\int e^x\,dx = e^x.$
4. $\int \sin x\,dx = -\cos x.$
5. $\int \cos x\,dx = \sin x.$
6. $\int \frac{1}{1+x^2}\,dx = \arctan x.$
7. $\int \frac{1}{\sqrt{1+x^2}}\,dx = \operatorname{arcsin} x.$

Jede Ableitungsregel liefert eine Regel für die Berechnung von Stammfunktionen. Die Kettenregel ergibt:

Satz 13.19 (Substitutionsregel). *Sei $f\colon I \to \mathbb{R}$ stetig, $\varphi\colon [a,b] \to I$ stetig differenzierbar und $\alpha = \varphi(a), \beta = \varphi(b)$. Dann gilt:*
$$\int_\alpha^\beta f(x)\,dx = \int_a^b f(\varphi(t)) \cdot \varphi'(t)\,dt.$$

Beweis. Sei $F(x) = \int f(x)\,dx$. Dann ist $F \circ \varphi$ differenzierbar nach dem Hauptsatz der Differential– und Integralrechnung und die Ableitung ist
$$(F \circ \varphi)'(t) = F'(\varphi(t)) \cdot \varphi'(t) = f(\varphi(t)) \cdot \varphi'(t)$$
nach der Kettenregel. Also ist $F \circ \varphi$ eine Stammfunktion von $(f \circ \varphi) \cdot \varphi'$ und:
$$\int_a^b f(\varphi(t))\varphi'(t)\,dt = F(\varphi(b)) - F(\varphi(a)) = F(\beta) - F(\alpha) = \int_\alpha^\beta f(x)\,dx.$$

□

Beispiel 13.20. Recht häufig kann man folgenden Spezialfall der Substitutionsregel anwenden:

1. Sei $g\colon [a,b] \to \mathbb{R}$ stetig diffbar mit $g(t) \neq 0\ \forall t$. Dann gilt:

$$\int \frac{g'(t)}{g(t)}\,dt = \ln|g(t)|.$$

   In der Notation der Substitutionsregel ist hier also $f(x) = \frac{1}{x}$ und $\varphi = g$.

   Wir rechnen dies explizit nach: Ohne Einschränkung ist, wegen des Zwischenwertsatzes, $g > 0$, d.h. $g(t) > 0\ \forall t$ (sonst betrachten wir $-g$ mit $(-g)' = -g'$). Die Kettenregel liefert:

$$(\ln(g(t))' = \frac{1}{g(t)} \cdot g'(t).$$

2. Das eben Gezeigte können wir beispielsweise bei folgender Rechnung benutzen:

$$\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx = -\int \frac{-\sin x}{\cos x}\,dx = -\ln|\cos x|.$$

Bemerkung 13.21. Häufig merkt man sich die Substitutionsregel in der Form

$$x = \varphi(t), \quad \frac{dx}{dt} = \varphi'(t),$$

also „$dx = \varphi'(t)\,dt$". Wir haben zwar nicht formal nachgewiesen, dass eine solche Schreibweise sinnvoll ist, es liefert aber das richtige Ergebnis:

$$F(x) = \int f(x)\,dx \Rightarrow F(\varphi(t)) = \int f(\varphi(t))\cdot\varphi'(t)\,dt,$$

indem wir x durch $\varphi(t)$ und dx durch $\varphi'(t)\,dt$ ersetzen.

Die Produktregel $(fg)' = f'g + fg'$ impliziert, analog zur Folgerung der Substitutionsregel aus der Kettenregel:

Satz 13.22 (Partielle Integration). *Es seien $f, g\colon I \to \mathbb{R}$ stetig diffbare Funktionen. Dann gilt:*

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

bzw.

$$\int_a^b f'(x)g(x)\,dx = f(x)g(x)\Big|_a^b - \int_a^b f(x)g'(x)\,dx$$

Beweis. Produktregel. □

Beispiel 13.23.

1. Es gilt:
$$\int_0^\pi x \sin x \, dx = (-x \cos x)\Big|_0^\pi + \sin x\Big|_0^\pi = -\pi \cdot (-1) = \pi,$$
da wir bei der partiellen Integration $g(x) = x$, d.h. $g'(x) = 1$ und $f'(x) = \sin x$, d.h. $f(x) = \cos x$ wählen können.

2. Wir berechnen $\int e^{-x} \sin x \, dx$. Dazu setzen wir $f'(x) = e^{-x}$, d.h. $f(x) = -e^{-x}$, $g(x) = \sin x$, d.h. $g'(x) = \cos x$. Partielle Integration liefert:
$$\int e^{-x} \sin x \, dx = -e^{-x} \sin x + \int e^{-x} + \cos x \, dx.$$
Auf das letzte Integral wenden wir wieder partielle Integration an und erhalten:
$$\int e^{-x} \cos x \, dx = -e^{-x} \cos x - \int (-e^{-x})(-\sin x) \, dx.$$
Insgesamt folgt
$$2 \cdot \int e^{-x} \sin x \, dx = -e^{-x}(\sin x + \cos x),$$
also:
$$\int e^{-x} \sin x \, dx = -\frac{1}{2} e^{-x}(\sin x + \cos x).$$
Zur Sicherheit machen wir die Probe:
$$\left(-\frac{1}{2} e^{-x}(\sin x + \cos x)\right)' = \frac{1}{2} e^{-x}(\sin x + \cos x) - \frac{1}{2} e^{-e}(\cos x + \sin x),$$
was tatsächlich $e^{-x} \sin x$ ergibt.

3. Wir zeigen, dass $\int_0^{2\pi} \sin^2 x \, dx = \pi$. Dazu setzen wir $f(x) = \sin x$, $g' = \sin x$, so dass $g = -\cos x$, $f' = -\cos x$ und erhalten:
$$\int \sin^2 x \, dx = -\sin x \cos x + \int \cos^2 x \, dx.$$
Da $\cos^2 x = 1 - \sin^2 x$ ist, folgt, wie eben:
$$\int \sin^2 x \, dx = \frac{1}{2}(x - \sin x \cos x).$$
Einsetzen der Grenzen 0 und 2π liefert die Behauptung.

4. Ähnlich folgt: $\int_0^{2\pi} \sin x \cos x \, dx = \frac{1}{2} \sin^2 x\Big|_0^{2\pi} = 0$, denn mit $f(x) = \sin x$ und $g'(x) = \cos x$ gilt:
$$\int \sin x \cos x \, dx = \sin^2 x - \int \sin x \cos x.$$

13.3 Elementare Funktionen

Definition 13.24. *Die Menge der **elementaren Funktionen** ist die kleinste Menge von Funktionen, die folgendes erfüllt:*

1. x^n, $\sin x$, $\tan x$, e^x *und deren Umkehrfunktionen sind elementar.*
2. *Summen, Produkte und Quotienten von elementaren Funktionen sind elementar.*
3. *Kompositionen von elementaren Funktionen sind elementar.*

Satz 13.25 (ein Satz von Liouville). *Nicht jede elementare Funktion hat eine elementare Stammfunktion.*

Beweis. Die Funktion e^{-x^2} besitzt keine elementare Stammfunktion, wie bereits Liouville[1] zeigte. □

Leider können wir den Satz hier nicht nachweisen. Allerdings können wir das folgende positive Resultat, zumindest in groben Zügen, herleiten:

Satz 13.26. *Rationale Funktionen sind elementar integrierbar.*

Wir betrachten zunächst folgendes Beispiel:

Beispiel 13.27. Wir suchen die Stammfunktion

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

Die Idee dazu ist es, die **Partialbruchzerlegung** von $y = \frac{1}{1-x^2}$ zu betrachten. y hat Polstellen bei ± 1. Deren Partialbruchzerlegung ist dann eine Zerlegung der Form:

$$\frac{1}{1-x^2} = \frac{A}{1-x} + \frac{B}{1+x}.$$

Dieser **Ansatz** liefert $A(1+x) + B(1-x) = 1$, d.h. $A = B = \frac{1}{2}$, also:

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

Auf unser ursprüngliches Problem angewendet erhalten wir:

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

Beweis (von Satz 13.26, nur Beweisidee). Wir gehen in drei Schritten vor:

[1] Joseph Liouville (24. März 1809 in Saint-Omer – 8. September 1882 in Paris), französischer Mathematiker.

1. Zunächst bemerken wir, dass wir folgende Stammfunktionen kennen:

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

$$\int \frac{1}{x^n}\,dx = \begin{cases} \ln|x|, & \text{falls } n = 1, \\ \frac{1}{1-n}\cdot\frac{1}{x^{n-1}}, & \text{falls } n > 1, \end{cases}$$

$$\int \frac{x}{(1+x^2)^n}\,dx = \begin{cases} \frac{1}{2}\ln(1+x^2), & \text{falls } n = 1, \\ \frac{1}{2(1-n)}\cdot\frac{1}{(1+x^2)^{n-1}}, & \text{falls } n > 1. \end{cases}$$

Im Fall $n > 1$ ergibt sich die letzte Gleichung folgendermaßen:

$$\begin{aligned} \int \frac{1}{(1+x^2)^n}\,dx &= \int \frac{dx}{(1+x^2)^{n-1}} + \int x\cdot\frac{x}{(1+x^2)^{n-1}}\,dx \\ &= \int \frac{dx}{(1+x^2)^{n-1}} \\ &\quad + x\cdot\frac{1}{2(2-n)}\cdot\frac{1}{(1+x^2)^{n-2}} + \frac{1}{2(n-2)}\cdot\int \frac{dx}{(1+x^2)^{n-2}}, \end{aligned}$$

weil die Integrale auf der rechten Seite induktiv bekannt sind, falls $n \geq 3$. Der Fall $n = 2$ ist dem Leser überlassen.

2. Wie im Beispiel berechnen wir eine Partialbruchzerlegung. Wir starten mit einer rationalen Funktion $f(x) = \frac{g(x)}{h(x)}$. Division mit Rest liefert $q(x)$ und $r(x)$, so dass:

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

Den Nenner $h(x)$ faktorisieren wir in k lineare Faktoren $l_i(x) \in \mathbb{R}[x]$ (d.h. $\deg l_i(x) = 1$) und l quadratische Faktoren $q_j(x) \in \mathbb{R}[x]$ (d.h. $\deg q_j(x) = 2$), wobei q_j nicht mehr in Linearfaktoren zerlegbar sind:

$$h(x) = \prod_{i=1}^{k} l_i(x)^{e_i} \cdot \prod_{j=1}^{l} q_j(x)^{f_j}.$$

Dass eine solche Zerlegung über $\mathbb{R}$ immer existiert, können wir hier leider nicht beweisen.

Man kann nun zeigen, dass dann Konstanten $a_{im}, b_{jn}, c_{jn} \in \mathbb{R}$ existieren, so dass wir folgende Partialbruchzerlegung erhalten:

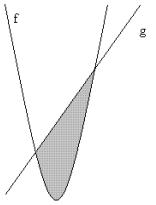
$$\frac{r(x)}{h(x)} = \sum_{i=1}^{k}\sum_{m=1}^{e_i} \frac{a_{im}}{l_i^m} + \sum_{j=1}^{l}\sum_{n=1}^{f_i} \frac{b_{jn}x + c_{jn}}{q_i^n}.$$

3. Nun transformieren wir $\frac{1}{l_i}$ in $\frac{1}{x}$ und $\frac{1}{q_j}$ in $\frac{1}{1+x^2}$ vermöge eines **affinen Koordinatenwechsels**, d.h. einer Abbildung $x \mapsto \alpha x + \beta$, und wenden die obigen Spezialfälle an. Auch hier können wir nicht erklären, warum ein solcher Koordinatenwechsel immer existiert.

□

Aufgaben

Aufgabe 13.1 (Flächeninhalt). Sei $f = 3x^2$ und $g = 3x + 6$. Bestimmen Sie den Flächeninhalt zwischen den Graphen von f und g, d.h. die graue Fläche in der Zeichnung:



Aufgabe 13.2 (Integrale). Berechnen Sie folgende Integrale:

1. $\int_1^2 \left(x^2 + \frac{1}{x}\right)\, dx$,
2. $\int_0^{\sqrt{\pi}} \left(5x \sin(x^2)\right)\, dx$,
3. $\int_0^{2\pi} \left(x^2 \sin(2x) + 3x \sin(2x)\right)\, dx$,
4. $\int_{-1}^1 \sqrt{1 - x^2}\, dx$ (verwenden Sie hier die Substitution $x = \sin t$).

Aufgabe 13.3 (Stammfunktion einer rationalen Funktion). Sei $f : D \to \mathbb{R}$ definiert durch

$$f(x) = \frac{1}{x^2 - 3x + 2},$$

wobei $D \subset \mathbb{R}$ ihr maximal möglicher Definitionsbereich sei.

Leiten Sie die Stammfunktion von f mit Hilfe von Partialbruchzerlegung her.

Aufgabe 13.4 (Maximierung von Integralen). Bestimmen Sie den Wert bzw. die Werte, an denen

$$H(x) = \int_0^{x^2} (9 - t^2) \cdot e^{-t}\, dt$$

maximal wird.

Aufgabe 13.5 (Stammfunktionen). Finden Sie folgende Stammfunktionen:

1. $\int \frac{x^6}{x^4+3x^2+2}\,dx$ (Hinweis: Division mit Rest, dann Partialbruchzerlegung),
2. $\int e^x \sin x\,dx$,
3. $\int \frac{(\ln x)^n}{x}\,dx$ für $n \in \mathbb{N}$.

14

Uneigentliche Integrale

Bisher haben wir Integrale nur dann ausgerechnet, wenn die Grenzen beide endlich waren. Wir werden sehen, dass wir in einigen Fällen auch ∞ als Grenze zulassen können und dass dies viele interessante Anwendungen hat, beispielsweise auf die Konvergenz von Reihen.

Definition 14.1. *Sei $f\colon [a,\infty[\to\mathbb{R}$ eine stetige Funktion. Dann setzen wir*

$$\int_a^\infty f(x)\,dx := \lim_{b\to\infty}\int_a^b f(x)\,dx,$$

*falls der Grenzwert existiert und nennen in diesem Fall f über $[a,\infty[$ **uneigentlich integrierbar** und $\int_a^\infty f(x)\,dx$ **konvergent**.*

Beispiel 14.2. Wir betrachten die Funktion $f(x)=x^{-s}$ für $s\in\mathbb{R}$. Der Grenzwert

$$\int_1^\infty x^{-s}\,dx = \lim_{b\to\infty}\Big(\frac{1}{1-s}x^{1-s}\Big|_1^b\Big) = \lim_{b\to\infty}\frac{1}{1-s}(1-b^{1-s})$$

existiert, falls $s>1$. Also gilt in diesem Fall:

$$\int_1^\infty x^{-s}\,dx = \frac{1}{1-s}.$$

Satz 14.3 (Integralkriterium für Reihen). *Sei $f\colon [0,\infty[\to\mathbb{R}$ eine monoton fallende positive Funktion. Die Reihe $\sum_{n=0}^\infty f(n)$ konvergiert genau dann, wenn das Integral $\int_0^\infty f(x)\,dx$ existiert.*

Beweis. Wegen der Monotonie von f erhalten wir für jedes $k\in\mathbb{N}$ offenbar Schranken von oben und unten für $\int_0^k f(x)\,dx$ (siehe Abb. 14.1):

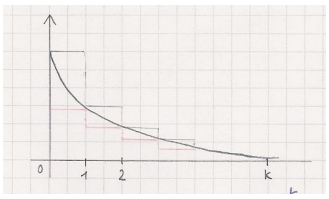


Abbildung 14.1. Ober– und Untersumme.

$$\sum_{n=0}^{k-1} f(n) \geq \int_0^k f(x)\,dx \geq \sum_{n=1}^{k} f(n).$$

Die Behauptung folgt. □

Korollar/Definition 14.4. *Der Grenzwert*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

existiert für $s > 1$. $\zeta(s)$ heißt ***Riemannsche Zetafunktion****.*

Beweis. Im Beispiel 14.2 haben wir gesehen, dass $\int_1^\infty x^{-s}dx$ für $s > 1$ konvergiert. Das Integralkriterium für Reihen liefert nun die Behauptung. □

Definition 14.5. *Sei $f\colon]a,b]$ eine stetige Funktion auf einem halboffenen Intervall. Dann schreiben wir*

$$\int_a^b f(x)\,dx := \lim_{t \searrow a} \int_t^b f(x)\,dx,$$

falls der Grenzwert existiert.

Beispiel 14.6. Das Integral $\int_0^1 x^\alpha\,dx$ existiert, falls $\alpha < -1$. Im Gegensatz dazu existiert das Integral $\int_0^1 \frac{1}{x}\,dx$ nicht, da: $\ln x \underset{x\to 0}{\longrightarrow} -\infty$.

Definition 14.7. *Eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ heißt* ***uneigentlich integrierbar****, falls die Grenzwerte*

$$\lim_{b\to\infty} \int_0^b f(x)\,dx \quad \textit{und} \quad \lim_{a\to-\infty} \int_a^0 f(x)\,dx$$

existieren. In diesem Fall schreiben wir:

$$\int_{-\infty}^{\infty} f(x)\,dx := \lim_{a\to\infty} \int_a^0 f(x)\,dx + \lim_{b\to\infty} \int_0^b f(x)\,dx.$$

Beispiel 14.8.

1. Es gilt:
$$\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = \lim_{a\to-\infty,\, b\to\infty} \Big(\arctan x\,\Big|_a^b\Big) = \frac{\pi}{2}.$$
2. Der Grenzwert $\int_{-\infty}^{\infty} e^{-x^2}\,dx$ existiert, denn $e^{-x^2} \in O\Big(\frac{1}{1+x^2}\Big)$. Man kann zeigen, dass gilt:
$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

Aufgaben

Aufgabe 14.1 (Uneigentliche Integrale). Überprüfen Sie, ob folgende uneigentliche Integrale existieren und berechnen Sie ggf. ihre Werte:

1. $\int_0^{\infty} xe^{-x^2}\,dx$,
2. $\int_0^1 \frac{e^x}{x}\,dx$,
3. $\int_1^{\infty} \frac{1}{1+\ln x}\,dx$,
4. $\int_{-\infty}^{\infty} \frac{1}{x^2-3x+2}\,dx$.

15

Taylorpolynom und Taylorreihe

Sei $f\colon I \to \mathbb{R}$ eine n–mal stetig diffbare Funktion, $x_0 \in I$. Wir wollen f in der Nähe von x_0 durch ein Polynom approximieren. Die „beste" Approximation durch ein lineares Polynom ist die Tangente (Abb. 15.1)

$$L(x) = f(x_0) + f'(x_0) \cdot (x - x_0).$$

Wir werden nun erfahren, wie man mit Polynomen von höhrem Grad noch bessere Approximationen erhalten kann.

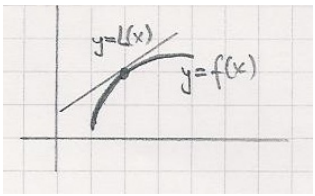


Abbildung 15.1. Approximation durch die Tangente.

Definition 15.1. *Seien $f\colon I \to \mathbb{R}$ eine n–mal stetig diffbare Funktion und $x_0 \in I$. Dann heißt*

$$T^n_{x_0} f := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

*das n–te **Taylorpolynom** von f in x_0.*

$T^n_{x_0} f$ hat offenbar den gleichen Wert und die gleichen ersten n Ableitungen in x_0 wie f.

Satz 15.2 (Taylorsche Formel). *Seien $f\colon I \to \mathbb{R}$ eine $(n+1)$–mal stetig diffbare Funktion und $x_0 \in I$. Dann gilt:*

$$f(x) = (T^n_{x_0} f)(x) + R_{n+1}(x) \left(= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)(x-x_0)^k}{k!} + R_{n+1}(x)\right)$$

mit dem Restglied

$$R_{n+1}(x) = \int_{x_0}^{x} f^{(n+1)}(t) \cdot \frac{(x-t)^n}{n!}\, dt.$$

Ausführlich:

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f^{(2)}(x_0)}{2}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \int_{x_0}^{x} f^{(n+1)}(t)\frac{(x-t)^n}{n!}dt.$$

Beweis. Wir verwenden Induktion nach n und partielle Integration. Der Induktionsanfang $n = 0$ gilt nach dem Hauptsatz der Differential– und Integral—Rechnung:

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)\, dt.$$

Für den Induktionsschritt $n-1 \to n$ betrachten wir das Restglied:

$$\begin{aligned} R_n(x) &= \int_{x_0}^{x} \underbrace{f^{(n)}(t)}_{f} \cdot \underbrace{\frac{(x-t)^{n-1}}{(n-1)!}}_{g'}\, dt \\ &\overset{g=-\frac{(x-t)^n}{n!}}{=} -f^{(n)}(t) \cdot \frac{(x-t)^n}{n!}\Big|_{x_0}^{x} + \int_{x_0}^{x} f^{(n+1)}(t)\frac{(x-t)^n}{n!}\, dt \\ &= f^{(n)}(x_0)\frac{(x-x_0)^n}{n!} + R_{n+1}(x). \end{aligned}$$

Da

$$T^n_{x_0} f - T^{n-1}_{x_0} f = f^{(n)}(x_0)\frac{(x-x_0)^n}{n!},$$

folgt die Behauptung. □

Wegen der aufwändigen Berechnungen ist für die praktische Anwendung dieser Formel ein Computeralgebra–System sehr hilfreich. In der Vorlesung wurden mit Maple einige Beispiele vorgeführt.

Satz 15.3 (Lagrangeform des Restglieds). *Sei $f\colon I \to \mathbb{R}$ eine $(n+1)$–mal stetig diffbare Funktion und $x_0 \in I$. Dann $\exists\, \xi \in [x_0, x]$ (bzw. $\xi \in [x, x_0]$, wenn $x < x_0$), so dass*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}.$$

Beweis. Wir wenden den Mittelwertsatz der Integralrechnung auf

$$R_{n+1}(x) = \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n dt$$

an und erhalten:

$$R_{n+1}(x) = f^{(n+1)}(\xi) \cdot \int_{x_0}^{n} \frac{(x-t)^n}{n!} dt = f^{(n+1)}(\xi)\frac{(x-x_0)^{n+1}}{(n+1)!}.$$

□

Beispiel 15.4. $f(x) = \sin x$, $x_0 = 0$.

$$(T_0^{2n+1}\sin)(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!},$$

da

$$\sin^{(k)}(0) = \begin{cases} 0, & k \text{ gerade}, \\ (-1)^{\frac{k-1}{2}}, & k \text{ ungerade}. \end{cases}$$

Fehlerabschätzung für Sinus:

$$|R_{n+1}(x)| = |f^{(n+1)}(\xi)| \cdot \frac{|x|^{n+1}}{(n+1)!} \leq \frac{|x|^{n+1}}{(n+1)!} \leq \frac{R^{n+1}}{(n+1)!}$$

für $|x| \leq R$. Es gilt: $\frac{R^{n+1}}{(n+1)!} \leq \varepsilon < 1 \iff$

$$(n+1)\ln R \leq \ln \varepsilon + \sum_{k=1}^{n+1} \ln k \leq \ln \varepsilon + \int_1^{n+1} \ln x \, dx.$$

Es folgt:

$$\begin{aligned} &(n+1)\ln R \leq \ln \varepsilon + (x \ln x - x)\Big|_1^{n+1} \\ \iff &\ln R \leq \frac{\ln \varepsilon}{n+1} + \ln(n+1) - 1 + \frac{1}{n+1} \\ \iff &R \leq (e\varepsilon)^{\frac{1}{n+1}} \cdot \frac{n+1}{e}. \end{aligned}$$

Abbildung 15.2 zeigt die Taylorpolynome des Sinus vom Grad $1, 3, 5, 7$.

Beispiel 15.5. Wir betrachten die Funktion $f(x) = (1+x)^\alpha$, etwa $\alpha = \frac{1}{2}$. Es gilt:

$$f^{(k)}(x) = \alpha \cdot (\alpha - 1) \cdots (\alpha - k + 1) \cdot (1+x)^{\alpha - k}.$$

Daher folgt:

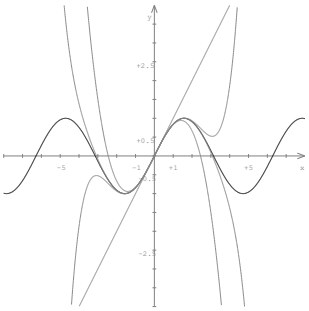


Abbildung 15.2. Die Taylorpolynome des Sinus vom Grad $1, 3, 5, 7$.

$$T_0^n f(x) = \sum_{k=1}^{n} \binom{\alpha}{k} x^k,$$

wobei für $\alpha \in \mathbb{R}$ der **Binomialkoeffizient** definiert ist durch

$$\binom{\alpha}{k} := \frac{\alpha}{1} \cdot \frac{\alpha - 1}{2} \cdots \frac{\alpha - k + 1}{k}.$$

Definition 15.6. *Sei $f : I \to \mathbb{R}$ eine ∞–oft diffbare Funktion und $x_0 \in I$. Dann heißt*

$$T_{x_0} f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

*die **Taylorreihe** von f mit **Entwicklungspunkt** x_0. Die Taylorreihe $T_{x_0} f$ ist eine Potenzreihe in $(x - x_0)$.*

Beispiel 15.7.

1. $(T_0 \exp)(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ ist die definierende Potenzreihe von exp.
2. Die Funktion $f(x) = (1 + x)^\alpha$ hat die Taylorreihe

$$\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k.$$

Frage 15.8. *Konvergiert die Taylorreihe von f gegen f?*

Antwort.

1. Eine Taylorreihe hat nicht notwendig positive Konvergenzradien.
2. Selbst, wenn $R > c > 0$ gilt, konvergiert die Taylorreihe nicht notwendig gegen f auf $]-R + x_0, x_0 + R[$.

Beispiel 15.9. Wir betrachten die Funktion

$$f\colon \mathbb{R} \to \mathbb{R},\ x \mapsto f(x) = \begin{cases} e^{-\frac{1}{x^2}}, & x \neq c, \\ 0, & \text{sonst.} \end{cases}$$

Wir zeigen: f ist ∞–oft diffbar und $f^{(n)}(0) = 0\ \forall n$. Insbesondere ist die Taylorreihe $= 0$ und konvergiert nicht gegen f.

Wir beweisen dazu mit Induktion nach n, dass

$$f^{(n)}(x) = \begin{cases} p_n\left(\frac{1}{x}\right) \cdot e^{-\frac{1}{x^2}}, & \text{falls } x \neq 0, \\ 0, & \text{falls } x = 0, \end{cases}$$

wobei p_n ein Polynom ist. Der Induktionsanfang ist klar. Zunächst betrachten wir den Fall $x \neq 0$:

$$\begin{aligned} \left(p_n\left(\frac{1}{x}\right) \cdot e^{-\frac{1}{x^2}}\right)' &= \left(p_n\left(\frac{1}{x}\right)\right)' \cdot e^{-\frac{1}{x^2}} + p_n\left(\frac{1}{x}\right) \cdot \frac{2}{x^3} \cdot e^{-\frac{1}{x^2}}. \\ &= \left(p_n'\left(\frac{1}{x}\right) \cdot \frac{-1}{x^2} + p_n\left(\frac{1}{x}\right) \cdot \frac{2}{x^3}\right) \cdot e^{-\frac{1}{x^2}}. \end{aligned}$$

Dann ist

$$p_{n+1}(t) = -p_n'(t) \cdot t^2 - 2p_n(t) \cdot t^3.$$

An der Stelle $x = 0$ gilt dann:

$$f^{(n+1)}(0) = \lim_{x \to 0}\left(\frac{1}{x} p_n\left(\frac{1}{x}\right) \cdot e^{-\frac{1}{x^2}}\right) = 0,$$

da exp schneller wächst als jedes Polynom, wie wir in Beispiel 12.6 gesehen haben.

Die folgende Aussage ist wegen Frage und Antwort 15.8 weniger trivial, als es erscheinen könnte:

Satz 15.10 (Binomische Reihe). *Für $|x| < 1$ gilt:*

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k.$$

Beweis. Der Konvergenzradius der Reihe ist $R = 1$. Wir möchten das Quotientenkriterium anwenden:

$$\left|\frac{\binom{\alpha}{k+1} \cdot x^{k+1}}{\binom{\alpha}{k} \cdot x^k}\right| = \left|\frac{\alpha - k}{k+1}\right| \cdot |x| \underset{k \to \infty}{\longrightarrow} |x|.$$

Mit der Bemerkung 6.17 zum Quotientenkriterium zeigt dies, dass die Reihe konvergiert, wenn $|x| < 1$ und diviergiert, wenn $|x| > 1$ ist.

Wir zeigen:

$$R_{n+1}(x) = \frac{1}{n!}\int_0^x (x-t)^n f^{(n+1)}(t)\,dt = (n+1)\binom{\alpha}{n+1}\int_0^x (x-t)^n(1+t)^{\alpha-n-1}\,dt$$

konvergiert für $|x| < 1$ gegen 0. Nach dem Vorzeichen von x unterscheiden wir zwei Fälle.

Erster Fall $0 \le x < 1$: Sei $C = \max(1, (1+x)^\alpha)$. Dann gilt für $0 \le t \le x$

$$0 \le (1+t)^{\alpha-n+1} \le (1+t)^\alpha \le C.$$

Also

$$\begin{aligned}|R_{n+1}(x)| &= (n+1)\cdot\left|\binom{\alpha}{n+1}\right|\int_0^{|x|} |(x+t)^n(1-t)^{\alpha-n-1}|dt\\ &\le (n+1)\cdot\left|\binom{\alpha}{n+1}\right|\cdot C\int_0^x (x-t)^n dt = C\cdot\left|\binom{\alpha}{n+1}\right||x^{n+1}.\end{aligned}$$

Da $\sum_{n=0}^\infty \binom{\alpha}{n}x^n$ konvergiert, bekommen wir $|\binom{\alpha}{n+1}|x^{n+1} \longrightarrow 0$ für $n \to \infty$.

Der zweite Fall, $-1 < x < 0$, ist schwieriger, weil die Funktion $(1+x)^\alpha$ oder eine ihrer Ableitungen eine Polstelle bei $x = -1$ haben kann.

$$\begin{aligned}|R_{n+1}(x)| &= (n+1)\cdot\left|\binom{\alpha}{n+1}\right|\int_0^x |(x-t)^n(1-t)^{\alpha-n-1}|dt\\ &\le (n+1)\cdot\left|\binom{\alpha}{n+1}\right|\int_0^{|x|} \Big(\frac{|x|-t}{1-t}\Big)^n(1-t)^{\alpha-1}dt.\end{aligned}$$

Die Funktion $t \mapsto \frac{|x|-t}{1-t}$ ist monoton fallend in $[0, |x|]$, da die Ableitung negative ist

$$\frac{(1-t)(-1)-(-1)(|x|-t)}{(1-t)^2} = \frac{|x|-1}{(1-t)^2} < 0.$$

Also gilt

$$\Big(\frac{|x|-t}{1-t}\Big)^n \le |x|^n \text{ für } 0 \le t \le |x|.$$

Mit $C = |\alpha|\int_0^{|x|}(1-t)^{\alpha-1}dt$ erhalten wir

$$|R_{n+1}(x)| \le |\alpha|\left|\binom{\alpha-1}{n}\right|\cdot|x|^n\int_0^{|x|}(1-t)^{\alpha-1}dt \le C\cdot\left|\binom{\alpha-1}{n}\right|\cdot|x|^n,$$

und der Ausdruck konvergiert gegen 0 für $n \to \infty$, da auch die Reihe $\sum_{n=0}^\infty \binom{\alpha-1}{n}x^n$ konvergiert. □

Aufgaben

Aufgabe 15.1 (Taylorpolynom). Bestimmen Sie ohne Computer das Taylorpolynom 2. Grades von

1. $f\colon \mathbb{R} \to \mathbb{R},\ x \mapsto f(x) = \ln(\sin(x))$ im Punkt $x_0 = \frac{\pi}{2}$,
2. $g\colon \mathbb{R} \to \mathbb{R},\ x \mapsto g(x) = e^{\sqrt{x}} - e$ im Punkt $x_0 = 1$.

Aufgabe 15.2 (Taylorreihe). Berechnen Sie die Taylorreihe von

$$f(x) = \frac{x}{(x-1)(x+1)}$$

im Entwicklungspunkt $x_0 = 0$.

Zeigen Sie mit Hilfe von Partialbruchzerlegung, dass die Taylorreihe auf dem offenen Intervall $(-1, 1)$ gegen f konvergiert.

Aufgabe 15.3 (Taylorpolynome mit Computeralgebra). Berechnen Sie mit Hilfe von Maple die Taylorpolynome $T_0^k f$ der Ordnungen $k = 1, \ldots, 6$ im Entwicklungspunkt $x_0 = 0$ für

1. $f(x) = \tan x$
2. $f(x) = \sqrt{1+x}$

und plotten Sie jeweils die Graphen von f und der Taylorpolynome.

16

Konvergenz von Funktionenfolgen

Sei

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

eine Potenzreihe mit Konvergenzradius $R > 0$. Wir wollen zeigen, dass f in $]-R, R[$ ∞–mal diffbar ist und dass die Potenzreihe mit der Taylorreihe von f in $x_0 = 0$ übereinstimmt. Allein die Stetigkeit ist hierbei nicht offensichtlich.

16.1 Gleichmäßige Konvergenz

Definition 16.1. *Sei $f_n \colon I \to \mathbb{R}$ eine Folge von Funktionen auf einem Intervall. Die Folge (f_n) heißt **konvergent** (genauer: **punktweise konvergent**), wenn für jedes x die Folge $(f_n(x))$ konvergiert und dann ist die **Grenzfunktion***

$$f \colon I \to \mathbb{R},\ f(x) = \lim_{n\to\infty} f_n(x).$$

Wir schreiben: $\lim_{n\to\infty} f_n = f$.

Frage 16.2. *Wenn alle f_n stetig sind, ist dann auch* $\lim f_n$ *stetig?*

Antwort. Nein, nicht unbedingt. Dazu betrachten wir das Beispiel $f_n \colon [0,1] \to \mathbb{R},\ f_n(x) = x^n$. Dann existiert $f = \lim f_n$, aber

$$f(x) = \begin{cases} 0, & x \in [0,1[, \\ 1, & x = 1. \end{cases}$$

f ist also nicht stetig.

Wir müssen daher eine stärkere Forderung an die Konvergenz stellen:

Definition 16.3. *Sei* $(f_n : I \to \mathbb{R})_{n\in\mathbb{N}}$ *eine Folge von Funktionen auf einem Intervall* I. (f_n) ***konvergiert gleichmäßig*** *gegen eine* ***Grenzfunktion*** $f : I \to \mathbb{R}$, *wenn:*

$$\forall \varepsilon > 0 \; \exists\, n_0 : \; |f_n(x) - f(x)| < \varepsilon \; \forall n \geq n_0 \; \forall \; x \in I.$$

Satz 16.4 (Gleichmäßiger Limes stetiger Funktionen). *Ist* $(f_n : I \to \mathbb{R})$ *eine Folge stetiger Funktionen, die gleichmäßig gegen* f *konvergiert, so ist auch* f *stetig.*

Beweis. Seien $x_0 \in I$ und $\varepsilon > 0$ vorgegeben. Zu $\frac{\varepsilon}{3}$ existiert n_0 mit

$$|f_n(x) - f(x)| < \frac{\varepsilon}{3} \; \forall n \geq n_0 \; \forall \; x \in I.$$

Da f_{n_0} stetig ist, $\exists\, \delta > 0$, so dass

$$|f_{n_0}(x) - f_{n_0}(x_0)| < \frac{\varepsilon}{3} \; \forall x \in I \text{ mit } |x - x_0| < \delta.$$

Es folgt:

$$\begin{aligned} |f(x) - f(x_0)| &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(x_0)| + |f_n(x_0) - f(x_0)| \\ &< \varepsilon \; \forall x \text{ mit } |x - x_0| < \delta. \end{aligned}$$

□

Frage 16.5. *Lässt sich Integration mit Grenzwertbildung vertauschen?*

Antwort. Nein, nicht unbedingt. Wir betrachten dazu die **Zackenfunktion**

$$f_n(x) = \begin{cases} n^2 x & \text{falls } 0 \leq x \leq \frac{1}{n} \\ n^2(\frac{2}{n} - x) & \text{falls } \frac{1}{n} < x \leq \frac{2}{n} \\ 0 & \text{sonst ,} \end{cases}$$

siehe Abbildung 16.1. Es gilt: $\int_0^1 f_n(x)\, dx = 1 \; \forall \; n$ und $\lim_{n\to\infty} f_n = 0$ (punkt-

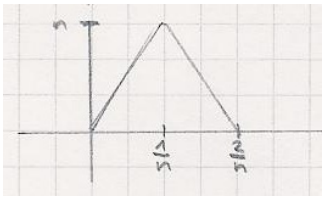


Abbildung 16.1. Die Zackenfunktion.

weise). Aber:

$$\lim_{n\to\infty} \int_0^1 f_n(x)\, dx = 1 \neq 0 = \int_0^1 \lim_{n\to\infty} f_n(x)\, dx.$$

Satz 16.6. *Sei $f_n\colon [a,b] \to \mathbb{R}$ eine Folge stetiger Funktionen auf einem abgeschlossenen Intervall, die gleichmäßig gegen $f\colon [a,b] \to \mathbb{R}$ konvergiert. Dann gilt:*

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \int_a^b f_n(x)\,dx.$$

Beweis. Zunächst einmal ist f ebenfalls stetig und deshalb integrierbar. Ferner:

$$\left|\int_a^b f(x)\,dx - \int_a^b f_n(x)\,dx\right| \le \int_a^b \left|f(x) - f_n(x)\right| dx \le \varepsilon(b-a),$$

falls n so groß ist, dass $|f(x) - f_n(x)| < \varepsilon\ \forall x \in [a,b]$. Die Behauptung folgt. □

Bemerkung 16.7. Für uneigentliche Integrale braucht man zusätzliche Voraussetzungen. Dies zeigt das Beispiel der **Zackenfunktion** in Abbildung 16.2. Offenbar ist $\lim f_n = 0$, sogar gleichmäßig, aber:

$$\int_0^\infty f_n(x)\,dx = 1 \neq 0 = \int_0^\infty 0\,dx.$$

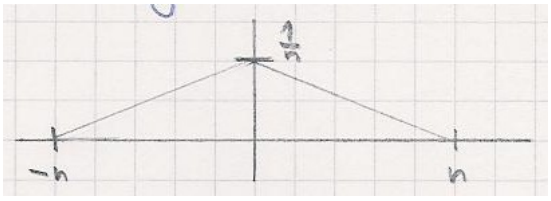


Abbildung 16.2. Eine Zackenfunktion.

Korollar 16.8. *Sei $f_n\colon [a,b] \to \mathbb{R}$ eine Folge von stetig diffbaren Funktionen, die punktweise gegen $f\colon [a,b] \to \mathbb{R}$ konvergiert. Konvergiert die Folge der Ableitungen (f_n') gleichmäßig, dann ist f diffbar und es gilt:*

$$f' = \lim_{n\to\infty} f_n'.$$

Beweis. Sei $f^* = \lim f_n'$. Nach dem Satz 16.4 ist f^* auf $[a,b]$ stetig. Ferner gilt:

$$f_n(x) = f_n(a) + \int_a^x f_n'(t)\,dt$$

für $x \in [a,b]$. Mit Satz 16.6 folgt:

$$f(x) = f(a) + \int_a^x f^*(t)\,dt.$$

Der Hauptsatz der Differential– und Integralrechnung liefert nun, dass f diffbar ist und dass $f' = f^*$. □

16.2 Anwendung auf Potenzreihen

Satz 16.9. *Sei $f(x) = \sum_{n=0}^{\infty} a_n x^n$ eine Potenzreihe mit Konvergenzradius $R > 0$, also $f :]-R, R[\to \mathbb{R}$. Dann haben die Potenzreihen*

1. $\sum_{n=1}^{\infty} n a_n x^{n-1}$,
2. $\sum_{n=0}^{\infty} a_n \frac{x^{n+1}}{n+1}$,

die wir durch gliedweise Differentiation bzw. Integration erhalten, den gleichen Konvergenzradius und konvergieren auf $]-R, R[$ gegen

1. $f'(x)$,
2. $\int_0^x f(x)\, dx$.

Insbesondere ist f unendlich oft diffbar und es gilt:

$$f^{(n)}(0) = a_n \cdot n!.$$

Beweis. Nach der Formel von Cauchy–Hadamard 7.19 ist $\sum a_n x^n$ für $|x| < R$ genau dann konvergent, wenn

$$\limsup_{n\to\infty} \sqrt[n]{|a_n x^n|} = \left(\limsup_{n\to\infty} \sqrt[n]{|a_n|}\right) \cdot R \leq 1,$$

also:

$$R = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|a_n|}}.$$

Da

$$\lim_{n\to\infty} \sqrt[n]{n} = \lim_{n\to\infty} n^{\frac{1}{n}} = \lim_{n\to\infty} e^{\frac{\ln n}{n}} = e^0 = 1,$$

haben $\sum_{n=1}^{\infty} n a_n x^{n-1}$ und $\sum_{n=0}^{\infty} a_n \frac{x^{n+1}}{n+1}$ den gleichen Radius. Die Folge der Partialsummen von f und f' konvergieren auf jedem echten Teilintervall $[-r, r]$ für $r < R$ gleichmäßig. Die Behauptung folgt daher auf $[-r, r]$ aus Satz 16.6 und Korollar 16.8. □

Beispiel 16.10.

1. Die **logarithmische Reihe** $\ln(1 + x)$ ist Stammfunktion von

$$f(x) = \frac{1}{1+x} = \sum_{k=0}^{\infty} (-1)^n x^n.$$

Gliedweise Integration und $\ln(1) = 0$ liefert:

$$\ln(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$$

für $|x| < 1$.

2. arctan ist Stammfunktion von

$$\frac{1}{1+x^2} = f(x) = \sum_{n=0}^{\infty} (-1)^n x^{2n}.$$

Integration und $\arctan(0) = 0$ liefert:

$$\arctan(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} \quad \text{für } |x| < 1.$$

3. Es gilt:

$$\sum_{n=0}^{\infty} n x^n = x \cdot \sum_{n=1}^{\infty} n x^{n-1}.$$

Es folgt:

$$\sum_{n=0}^{\infty} n x^n = x \cdot \Big(\frac{1}{1-x}\Big)' = \frac{x}{(1-x)^2} \quad \text{für } |x| < 1,$$

also zum Beispiel:

$$\sum_{n=1}^{\infty} n \Big(\frac{1}{2}\Big)^n = \frac{\frac{1}{2}}{(1-\frac{1}{2})^2} = 2.$$

In den Beispielen 1. und 2. konvergiert die Reihe auch für $x = 1$. Dies legt die Formeln

$$\sum_{n=0}^{\infty} (-1)^n \frac{1}{n+1} = \ln(1+1) = \ln 2,$$

$$\sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} = \arctan(1) = \frac{\pi}{4}$$

(da $\tan \frac{\pi}{4} = 1$) zumindest nahe. Dass dies wirklich der Fall ist, zeigt das folgende Resultat:

Satz 16.11 (Abelscher Grenzwertsatz). *Sei $\sum_{n=0}^{\infty} a_n$ eine konvergent Reihe reeller Zahlen. Dann hat die Potenzreihe $f(x) = \sum_{n=0}^{\infty} a_n x^n$ den Konvergenzradius $R \geq 1$ und für die Grenzfunktion $f :]-1, 1[\to \mathbb{R}$ gilt:*

$$\lim_{x \to 1} f(x) = \sum_{n=0}^{\infty} a_n.$$

Beweis. Drei Seiten. Forster. □

Beispiel 16.12. Wir betrachten zwei Reihen, von denen wir bereits wissen, dass sie konvergieren:

1. Die Reihe $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ konvergiert. Nach dem abelschen Grenzwertsatz ist daher durch $f(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \cdot x^n$ eine auf $]-1,1[$ stetige Funktion erklärt. Da $f(x) = \ln(1+x)$ für $x \in]-1,1[$ gilt, folgt: $f(1) = \ln(2)$. Also:

$$\ln(2) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots .$$

2. Wir betrachten $\sum_{n=0}^{\infty} (-1)^n \cdot \frac{x^{2n+1}}{2n+1} = \arctan x$ für $x \in]-1,1[$. Für $x = 1$ liegt ebenfalls Konvergenz vor, also:

$$\sum_{n=0}^{\infty} (-1)^n \cdot \frac{1}{2n+1} = \arctan 1 = \frac{\pi}{4}.$$

Aufgaben

Aufgabe 16.1 (. . .). . . .

Teil III

Lineare Algebra

Einführung

In der linearen Algebra werden Probleme und Phänomene untersucht, die mit Hilfe linearer Gleichungssysteme ausdrückbar sind. Insbesondere werden dabei auch Verfahren studiert, um explizit Lösungen für solche Probleme auszurechnen. Oft sind sowohl die Probleme und Lösungen als auch die Phänomene sehr anschaulich zu verstehen, wenn man deren geometrische Seite betont. Wir werden versuchen, dies in dieser Vorlesung möglichst oft zu realisiseren.

Anwendungen der linearen Algebra finden sich neben der Geometrie in vielen Bereichen der Mathematik und Informatik, aber auch des alltäglichen Lebens. Wir werden solche Anwendungen so oft wie möglich ansprechen, insbesondere, wenn sie die Informatik betreffen.

Um den Anschauungs– und Anwendungsbezug möglichst naheliegend zu halten, beginnen wir mit der linearen Algebra über den reellen Zahlen und deren Geometrie. Es wird sich im Verlauf der Vorlesung allerdings herausstellen, dass es oft sinnvoll ist, davon abzuweichen und lineare Algebra auch über anderen Zahlensystemen zu betreiben. Beispielsweise wäre das ganze Gebiet der Kodierungstheorie kaum denkbar ohne die lineare Algebra über endlichen Körpern. Andererseits wäre die Darstellung der Theorie viel zu kompliziert, ohne die Ausweitung der Zahlen auf die sogenannten komplexen Zahlen zu betrachten.

17

Der $\mathbb{R}^3$ und der $\mathbb{R}^n$

Die zum Teil aus der Schule bekannte Geometrie im Anschauungsraum $\mathbb{R}^3$ stellen wir vor und erfahren dabei, dass die Erweiterung auf den $\mathbb{R}^n$ in den meisten Fällen abstrakt gesehen gar keine Änderung benötigt.

17.1 Punkte im $\mathbb{R}^n$

Einen **Punkt** im Anschauungsraum $\mathbb{R}^3$ können wir nach Einführung eines **Koordinatensystems** durch ein **Tupel** (a_1, a_2, a_3) (ein Tupel ist eine durch Komma getrennte Menge von Objekten, bei denen es auf die Reihenfolge ankommt) von **Koordinaten** spezifizieren (Abb. 17.1). Meist werden die Koordinatenachsen zueinander senkrecht gewählt; in diesem Fall nennt man das Koordinatensystem **kartesisch** nach René Descartes (1596-1650), der Koordinatensysteme maßgeblich bekannt machte.

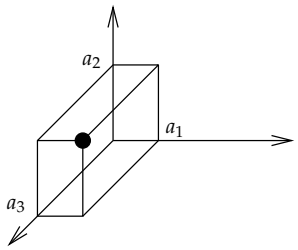


Abbildung 17.1. Der Punkt $(a_1, a_2, a_3) \in \mathbb{R}^3$.

Punkte im $\mathbb{R}^3$ und allgemein im $\mathbb{R}^n$ stellen wir mit **Zeilen-** oder **Spaltenvektoren** dar. In der linearen Algebra sind Spaltenvektoren üblich, bei denen dann das Komma weggelassen wird:

$$a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \in \mathbb{R}^3 \quad \text{bzw.} \quad a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^n.$$

Platzsparend werden Spaltenvektoren aber häufig $(a_1, a_2, \ldots, a_n)^t$ notiert, wobei ^{t} für **transponieren** steht. Der **Nullvektor** oder **Ursprung des Koordinatensystems** ist der Vektor $(0, \ldots, 0)^t \in \mathbb{R}^n$; er wird oft kurz 0 geschrieben. Aus dem Zusammenhang wird immer klar sein, ob 0 die Zahl oder den Vektor bezeichnet.

$\mathbb{R}^{3k}$ taucht zum Beispiel auf, wenn wir k Punkte im Raum gleichzeitig betrachten. R^n mit $n \approx 1.000.000$ wird in der Computertomographie oder bei der Diskretisierung von partiellen Differentialgleichungen verwendet. $n \approx 10^9$ taucht bei einer *Supportvektormaschine* auf, wie sie Google verwendet und $n \approx 10^6$ Dreiecke werden zur Approximation einer geschwungenen Oberfläche durch winzige Dreiecke verwendet.

Definition 17.1. *Zu zwei Vektoren*

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \; y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

*und einer Zahl $\lambda \in \mathbb{R}$ (genannt **Skalar**) setzen wir (siehe Abb. 17.2):*

$$x + y := \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}, \quad \lambda \cdot x := \begin{pmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

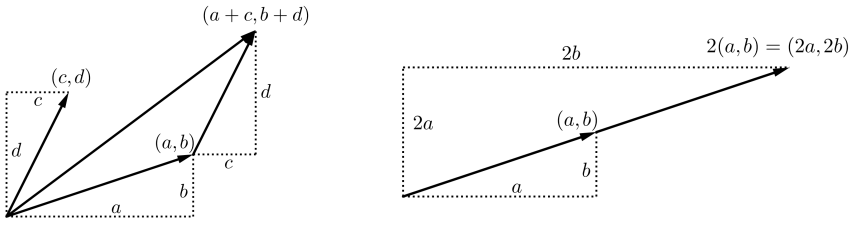


Abbildung 17.2. Summe zweier Vektoren $(a, b), (c, d) \in \mathbb{R}^2$ sowie Multiplikation eines Vektors $(a, b) \in \mathbb{R}^2$ mit dem Skalar $2 \in \mathbb{R}$.

17.2 Skalarprodukt, Euklidische Norm

Mit Hilfe des sogenannten Skalarprodukts werden wir nun Abstände und Winkel einführen und erste Eigenschaften der neuen Begriffe herleiten.

Definition 17.2. *Für $x, y \in \mathbb{R}^n$ ist*

$$\langle x, y\rangle := x_1 \cdot y_1 + \cdots + x_n \cdot y_n \in \mathbb{R}$$

(oder $x \cdot y = \langle x, y\rangle$) das ***Skalarprodukt*** *(genauer* ***Standard–Skalarprodukt*** *oder* ***euklidisches Skalarprodukt****) von x und y (engl. auch* ***dot–product*** *oder* ***inner product****).*

$$\|x\| := \|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} := \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = (\langle x, x\rangle)^{\frac{1}{2}} \in \mathbb{R}$$

heißt ***Betrag*** *oder* ***euklidische Norm*** *von x. Hierbei bezeichnet $\sqrt{a}$ die Quadratwurzel aus einer nicht-negativen reellen Zahl, d.h. es ist diejenige nicht-negative Zahl, für die $(\sqrt{a})^2 = a$ (s. Abschnitt 5.6 für Details).*

Zumindest im $\mathbb{R}^2$ und $\mathbb{R}^3$ lässt sich $\|x\|$ als Länge des Vektors $x \in \mathbb{R}^3$ interpretieren (wegen des Satzes von Pythagoras, siehe Proposition 17.3 und Bemerkung 17.7). Wir nennen $\|x\|$ daher auch die ***Länge des Vektors*** *$x \in \mathbb{R}^n$. Zu $x, y \in \mathbb{R}^n$ ist*

$$d(x, y) := \|x - y\|$$

der ***Abstand der Punkte*** *x und y.*

Die folgenden Eigenschaften folgen recht leicht aus diesen Defintionen:

Proposition 17.3 (Eigenschaften des Skalarprodukts). *Es gilt:*

1. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ *für alle* $x, y, z \in \mathbb{R}^n$ ***(Linearität)***,
2. $\langle \lambda x, y\rangle = \lambda\langle x, y\rangle$ *für alle* $x, y \in \mathbb{R}^n, \lambda \in \mathbb{R}$ ***(Linearität)***,
3. $\langle x, y\rangle = \langle y, x\rangle$ *für alle* $x, y \in \mathbb{R}^n$ ***(Symmetrie)***,
4. $\langle x, x\rangle \geq 0$ *und* $\langle x, x\rangle = 0$ *genau dann, wenn* $x = 0 \in \mathbb{R}^n$,
5. $\|x + y\|^2 = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2$ ***(Satz des Pythagoras)***
6. $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$ ***(Parallelogrammgleichung)***.

Beweis. Wir zeigen nur drei der Behauptungen:

Zu 1. $\langle x + y, z\rangle = \sum(x_i + y_i)z_i = \sum x_i z_i + \sum y_i z_i = \langle x, z\rangle + \langle y, z\rangle$.

Zu 5. Es gilt: $\|x+y\|^2 = \langle x+y, x+y\rangle = \langle x, x\rangle + 2\langle x, y\rangle + \langle y, y\rangle$. Zur Bezeichnung *Satz des Pythagoras* s. Bemerkung 17.7.

Zu 6. $\|x+y\|^2 + \|x-y\|^2 \overset{(5)}{=} \|x\|^2 + 2\langle x,y\rangle + \|y\|^2 + \|x\|^2 + 2\langle x,-y\rangle + \|-y\|^2$
$\overset{(2\&3)}{=} \|x\|^2 + 2\langle x,y\rangle + \|y\|^2 + \|x\|^2 - 2\langle x,y\rangle + \|y\|^2$, wie behauptet.

□

Nun kommen wir zu einer weiteren, sehr häufig verwendeten, aber nicht ganz so leicht nachzuweisenden Eigenschaft:

Satz 17.4 (Cauchy–Schwarz'sche Ungleichung). *Für $x, y \in \mathbb{R}^n$ gilt:*

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

Ferner gilt für $x \neq 0$ die Gleichheit $|\langle x,y\rangle| = \|x\| \cdot \|y\|$ genau dann, wenn $y = \lambda x$ für ein $\lambda \in \mathbb{R}$.

Beweis. Für $x = 0$ ist alles klar. Sei also $x \neq 0 \in \mathbb{R}^n$. Dann ist

$$\mu := \langle x,x\rangle = \sum_{i=1}^{n} x_i^2 > 0.$$

Ferner setzen wir: $\varphi := -\langle x,y\rangle$. Damit gilt:

$$\begin{aligned} 0 &\leq \langle \varphi x + \mu y, \varphi x + \mu y\rangle \\ &= \varphi^2\langle x,x\rangle + 2\varphi\mu\langle x,y\rangle + \mu^2\langle y,y\rangle \\ &= \mu\cdot\langle x,y\rangle^2 - 2\mu\cdot\langle x,y\rangle^2 + \mu\cdot\langle x,x\rangle\langle y,y\rangle \\ &= \mu\cdot\Big(-\langle x,y\rangle^2 + \langle x,x\rangle\langle y,y\rangle\Big). \end{aligned}$$

Da $\mu > 0$, folgt:

$$0 \leq -(\langle x,y\rangle)^2 + \|x\|^2 \cdot \|y\|^2 \quad \text{bzw.} \quad |\langle x,y\rangle|^2 \leq \|x\|^2 \cdot \|y\|^2.$$

Die Monotonie der Quadratwurzel $\sqrt{}$ (siehe Bemerkung 5.30) liefert die erste Behauptung.

Gilt Gleichheit, dann folgt:

$$0 = \langle \varphi x + \mu y, \varphi x + \mu y\rangle \;\Rightarrow\; \varphi x + \mu y = 0.$$

Also: $y = \lambda x$ mit $\lambda = -\frac{\varphi}{\mu}$. □

Die folgenden Eigenschaften der oben eingeführten Norm sind wieder leicht einzusehen:

Proposition 17.5 (Eigenschaften der Norm). *Für $x, y \in \mathbb{R}^n, \lambda \in \mathbb{R}$ gilt:*

1. *$\|x\| \geq 0$ und $\|x\| = 0$ genau dann, wenn $x = 0 \in \mathbb{R}^n$,*

2. $\|\lambda x\| = |\lambda| \cdot \|x\|$,
3. $\|x + y\| \leq \|x\| + \|y\|$ *(Δ-Ungleichung).*

Beweis. Nur 3. ist zu zeigen. Nach Definition gilt:

$$\begin{aligned}\|x + y\|^2 &= \langle x + y, x + y\rangle \\ &= \|x\|^2 + 2\langle x, y\rangle + \|y\|^2.\end{aligned}$$

Cauchy–Schwartz liefert nun:

$$\|x + y\|^2 \leq \|x\|^2 + 2\|x\|\cdot\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$

Die Behauptung folgt mit der Monotonie der Quadratwurzel (Bem. 5.30). □

Wir kommen nun zu der geometrischen Bedeutung des Skalarprodukts:

Definition 17.6. *Zwei Vektoren $x, y \in \mathbb{R}^n$ heißen **senkrecht** (auch **orthogonal**) zueinander (in Zeichen $x \perp y$), wenn $\langle x, y\rangle = 0$.*

Bemerkung 17.7.

1. Im $\mathbb{R}^2$ bzw. $\mathbb{R}^n$ stimmt dieser Begriff mit dem anschaulichen Begriff überein. Dies folgt aus der Formel (s. 17.3)

   $$\|x + y\|^2 = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2$$

   zusammen mit dem geometrischen Satz des Pythagoras $a^2+b^2 = c^2$ für die Seitenlängen a, b, c in einem rechtwinkligen Dreieck, wobei c die längste ist.

   Für die Seitenlängen der beiden in Abb. 17.3 erkennbaren rechtwinkligen Dreiecke gilt dann nämlich:

   $$a^2 + d^2 = \|x\|^2, \qquad a^2 + (d + \|y\|)^2 = \|x + y\|^2.$$

   Die Differenz dieser beiden Gleichungen ist:

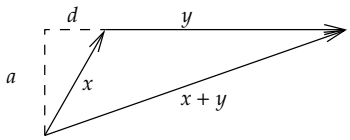


Abbildung 17.3. Anwendung des geometrischen Satzes des Pythagoras auf zwei rechtwinklige Dreiecke.

$$2d\|y\| + \|x\|^2 + \|y\|^2 = \|x+y\|^2 = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2$$

und somit folgt: $d\|y\| = \langle x, y\rangle$.

Also: $d = 0 \Leftrightarrow \langle x, y\rangle = 0 \overset{\text{Def}}{\Leftrightarrow} x \perp y \Leftrightarrow \|x\|^2 + \|y\|^2 = \|x+y\|^2$.

2. Den Beweis des geometrischen Satzes des Pythagoras liefert Abbildung 17.4. Dass die vier Dreiecke in Abb. 17.4 gleich groß sind, beruht dabei

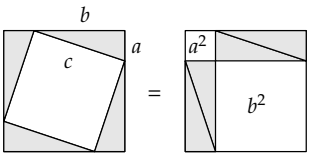


Abbildung 17.4. Dies zeigt: $c^2 = a^2 + b^2$, da die vier Dreiecke gleich groß sind.

auf der Tatsache, dass in einem Dreieck die Winkelsumme 180° ist, was wiederum aus dem **Parallelenaxiom** folgt (siehe Abb. 17.5). Dieses besagt: In einer Ebene α gibt es zu jeder Geraden g und jedem Punkt S außerhalb von g genau eine Gerade, die zu g parallel ist und durch den Punkt S geht. Ob das Parallelenaxiom in der Wirklichkeit gilt (nicht im $\mathbb{R}^3$), ist offen und Gegenstand der Astronomie. An dieser Stelle möchten wir auf zwei Bücher von Roger Penrose hinweisen: [Pena] und [Penb].

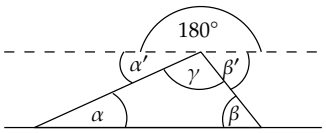


Abbildung 17.5. Das Parallelenaxiom liefert: $\alpha' = \alpha,\ \beta' = \beta$, also: $\alpha + \beta + \gamma = 180°$.

Definition 17.8. *Für zwei Vektoren $x, y \in \mathbb{R}^n$ definieren wir den **Winkel** θ zwischen x und y durch die Formel:*

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}.$$

Wegen der Cauchy–Schwarz'schen Ungleichung gilt:

$$-\|x\| \cdot \|y\| \le \langle x, y\rangle \le \|x\| \cdot \|y\|,$$

also $\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|} \in [-1, 1]$. Somit hat diese Definition einen Sinn (s. Abb. 17.6). Selbstverständlich wird der Winkel θ nur eindeutig durch seinen Cosinus

festgelegt, wenn man sich auf ein geeignetes Intervall für θ beschränkt, hier etwa auf das Intervall $[0, 2\pi[$. 2π ist die Länge des Vollkreises mit Radius 1; viele Eigenschaften von Cosinus und Sinus sowie Zusammenhänge zwischen diesen können recht leicht am Einheitskreis erklärt werden (s. Abb. 17.6).

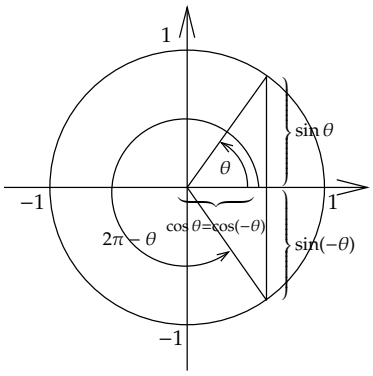


Abbildung 17.6. Cosinus und Sinus eines Winkels θ, visualisiert am Einheitskreis.

17.3 Geometrische Objekte im $\mathbb{R}^n$

Wir werden in diesem Abschnitt die eben eingeführten Begriffe verwenden, um Geraden, Ebenen und deren Verallgemeinerungen zu definieren und deren gegenseitige Lage, also insbesondere Abstände zwischen ihnen, zu studieren.

17.3.1 Geraden und Hyperebenen

Definition 17.9. *Eine **Gerade** $L \subseteq \mathbb{R}^n$ ist eine Teilmenge der Gestalt*

$$L = \{p + \lambda v \mid \lambda \in \mathbb{R}\} =: p + \mathbb{R} \cdot v,$$

*wobei $p \in L$ ein beliebiger **Aufpunkt** und $v \in \mathbb{R}^n \setminus \{0\}$ ein **Richtungsvektor** ist (s. Abb. 17.7).*

*Eine **Hyperebene** $H \subseteq \mathbb{R}^n$ ist eine Teilmenge der Gestalt*

$$H = \{x \in \mathbb{R}^n \mid a_1x_1 + \cdots + a_nx_n = b\} = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\},$$

*wobei $a = (a_1, \ldots, a_n)^t \in \mathbb{R}^n \setminus \{0\}$ und $b \in \mathbb{R}$. Der Vektor a heißt **Normalenvektor** von H (s. Abb. 17.8). Im $\mathbb{R}^3$ heißt eine Hyperebene einfach **Ebene**.*

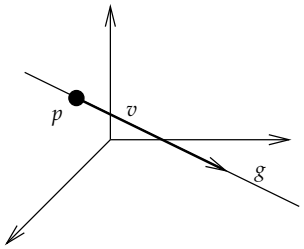


Abbildung 17.7. Eine Gerade g im $\mathbb{R}^3$, mit Aufpunkt p und Richtungsvektor v.

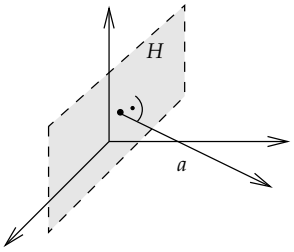


Abbildung 17.8. Eine (Hyper-)ebene in $\mathbb{R}^3$ und ein Normalenvektor a von H.

Für zwei Punkte $p, q \in H$ gilt für den Differenzvektor $v = p - q$

$$\langle a, v\rangle = \langle a, p - q\rangle = \langle a, p\rangle - \langle a, q\rangle = b - b = 0.$$

Also $a \perp p - q$. Daher der Name Normalenvektor.

17.3.2 Schnittpunkte

Sei $L \subseteq \mathbb{R}^n$ eine Gerade und $H \subseteq \mathbb{R}^n$ eine Hyperebene. Für die Schnittmenge $L \cap H$ gibt es drei Möglichkeiten:

1. $L \cap H = \{q\}$, besteht aus genau einem Punkt $q \in \mathbb{R}^n$,
2. $L \cap H = \emptyset$,
3. $L \subseteq H$.

Proposition 17.10. *2. oder 3. liegt genau dann vor, wenn der Richtungsvektor von L senkrecht zum Normalenvektor a von H ist.*

Beweis. Setzen wir die Parametrisierung $L = \{p + \lambda v | \lambda \in \mathbb{R}\}$ der Geraden in Gleichung $\langle a, x\rangle = b$ von H ein, so erhalten wir mit

$$\langle a, p + \lambda v\rangle = b \Leftrightarrow \lambda\langle a, v\rangle = b - \langle a, p\rangle \qquad (*)$$

eine Gleichung für λ.

Ist a nicht senkrecht zu v, d.h. $\langle a, v\rangle \neq 0$, dann ist die einzige Lösung

$$\lambda = \frac{b - \langle a, p\rangle}{\langle a, v\rangle}, \text{ also } L \cap H \ni q = p + \frac{b - \langle a, p\rangle}{\langle a, v\rangle} v.$$

Ist $\langle a, v\rangle = 0$, also $a \perp v$, dann hat (*) nur dann eine Lösung, wenn $b - \langle a, p\rangle = 0 \Leftrightarrow p \in H \Rightarrow L \subseteq H$, da dann $\lambda \in \mathbb{R}$ beliebig gewählt werden kann. Dies entspricht dem 3. Fall.
Ist $\langle a, v\rangle = 0$ und $b \neq \langle a, p\rangle$ dann $L \cap H = \emptyset$, 2. Fall. □

Definition 17.11. *In den Fällen 2. und 3., d.h. wenn $a \perp v$, so sagen wir: L ist eine zu H **parallele** Gerade.*

17.3.3 Abstände

Abstand zwischen Gerade und Punkt

Sei $L = \{p + \lambda v \mid \lambda \in \mathbb{R}\} \subseteq \mathbb{R}^n$ eine Gerade und $q \in \mathbb{R}^n$ ein weiterer Punkt. Dann ist für jeden Punkt $u \in L$ der Abstand $d(u, q) = \|u - q\|$.

Definition 17.12. *Wir definieren den **Abstand von** L **zu** q durch*

$$d(L, q) = \min_{u \in L} d(u, q).$$

Proposition/Definition 17.13. *Das Minimum $d(L, q)$ wird in genau einem Punkt u_q angenommen. Der Punkt u_q ist eindeutig durch die Eigenschaft, dass $u_q - q$ senkrecht zu dem Richtungsvektor v steht, bestimmt. u_q heißt **Fußpunkt** des **Lots** von q auf L (Abb. 17.9).*

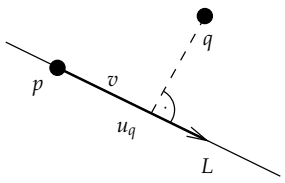


Abbildung 17.9. Das Lot des Punktes q auf die Gerade L.

Beweis. Die Gleichung $\langle v, p + \lambda v - q\rangle = 0$ hat genau eine Lösung nämlich

$$\lambda = \frac{\langle q - p, v\rangle}{\langle v, v\rangle},$$

da $\|v\|^2 \neq 0$. Also:

$$u_q = p + \frac{\langle q - p, v\rangle}{\langle v, v\rangle} \cdot v = p + \left\langle q - p, \frac{v}{\|v\|}\right\rangle \cdot \frac{v}{\|v\|}.$$

Jeder andere Punkt $u \in L$ hat einen größeren Abstand (s. Abb. 17.10), da $\|u - q\|^2 \overset{\text{Pythagoras}}{=} \|u_q - q\|^2 + \|u - u_q\|^2 \geq \|u_q - q\|^2$, also:

$$d(L, q) = \|u_q - q\|.$$

□

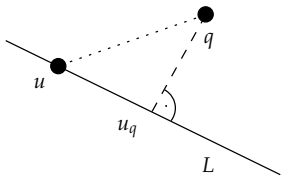


Abbildung 17.10. Der Abstand des Punktes q von der Geraden L ist $d(L, q) = \|u_q - q\|$.

Abstand zwischen Hyperebene und Punkt

Proposition/Definition 17.14. *Sei $H = \{x \mid \langle a, x\rangle = b\} \subseteq \mathbb{R}^n$ eine Hyperebene und $q \in \mathbb{R}^n$ ein Punkt. Dann definieren wir*

$$d(H, q) := \min_{u \in H} d(u, q).$$

Das Minimum $d(H, q)$ wird in genau einem Punkt $u_q \in H$ angenommen. u_q ist durch die Bedingung, dass die Differenz $u_q - q$ ein skalares Vielfaches des Normalenvektors a ist, eindeutig bestimmt. Die Abbildung

$$\mathbb{R}^n \to H, \quad q \to u_q$$

heißt ***orthogonale Projektion auf die Hyperebene*** *H (s. Abb. 17.11).*

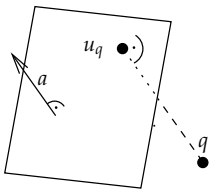


Abbildung 17.11. Die Orthogonale Projektion des Punktes q auf die Hyperebene H.

Beweis. Jeder andere Punkt $u \in H$ hat größeren Abstand zu q, nach Pythagoras. Um u_q auszurechnen, betrachten wir die Gerade $L = \{q + \lambda a \mid \lambda \in \mathbb{R}\} \subseteq \mathbb{R}^n$ und bestimmen $L \cap H$: $\langle q + \lambda a, a\rangle = b$ liefert $\lambda = \frac{b-\langle q,a\rangle}{\langle a,a\rangle}$, also

$$u_q = q + \frac{b - \langle a, q\rangle}{\langle a, a\rangle} \cdot a.$$

Der Abstand ist somit:

$$d(H, q) = \left\| \frac{b - \langle a, q\rangle}{\langle a, a\rangle} \cdot a \right\| = \frac{|b - \langle a, q\rangle|}{\langle a, a\rangle} \cdot \|a\| = \left| \frac{b}{\|a\|} - \left\langle \frac{a}{\|a\|}, q \right\rangle \right|.$$

Wählen wir den Normalenvektor **normiert**, d.h. von der Länge 1, dann gilt:

$$d(H, q) = |b - \langle a, q\rangle|.$$

In diesem Fall lässt sich $|b|$ als Abstand $d(H, 0)$ von H zum Nullpunkt interpretieren. Auch das Vorzeichen von b hat eine Interpretation:

$b > 0 \Leftrightarrow 0$ liegt in dem Halbraum $\{x \mid \langle a, x\rangle < b\}$

$\Leftrightarrow$ Der Normalenvektor, auf H angetragen, zeigt in den Halbraum, der 0 nicht enthält.

□

Abstand zwischen zwei Geraden

Definition 17.15. *Seien $L_1 = \{p_1 + \lambda v_1 \mid \lambda \in \mathbb{R}\}$ und $L_2 = \{p_2 + \lambda v_2 \mid \lambda \in \mathbb{R}\}$ zwei Geraden im $\mathbb{R}^n$. L_1 und L_2 heißen **parallel**, wenn $v_1 = \lambda v_2$ für ein $\lambda \in \mathbb{R}$, das heißt, wenn die Richtungsvektoren bis auf den Skalarfaktor übereinstimmen. L_1 und L_2 heißen **windschief**, wenn gilt:*

1. *L_1 und L_2 sind nicht parallel,*
2. *$L_1 \cap L_2 = \emptyset$.*

*$d(L_1, L_2) := \min_{x \in L_1, y \in L_2} d(x, y)$ nennen wir den **Abstand von** L_1 **zu** L_2.*

Proposition 17.16 (Abstand windschiefer Geraden). *Es seien L_1 und L_2 zwei windschiefe Geraden mit Richtungsvektoren v_1 bzw. v_2. Dann wird das Minimum $d(L_1, L_2)$ in genau einem Paar von Punkten $(\tilde{x}, \tilde{y}) \in L_1 \times L_2$ angenommen. $(\tilde{x}, \tilde{y})$ ist durch die Bedingung, dass $\tilde{x} - \tilde{y}$ senkrecht zu v_1 und v_2 steht, eindeutig bestimmt (Abb. 17.12).*

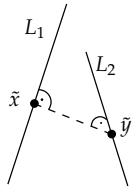


Abbildung 17.12. Abstand windschiefer Geraden.

Beweis. $(\tilde{x}, \tilde{y})$ erfülle die Bedingung. Für jedes andere Paar $(x, y) \in L_1 \times L_2$ gilt:

$$(x, y) = (\tilde{x} + \lambda_1 v_1,\ \tilde{y} + \lambda_2 v_2) \text{ für gewisse } \lambda_1, \lambda_2 \in \mathbb{R}.$$

Mit dieser Notation gilt

$$\begin{aligned} \|x - y\|^2 &= \|\tilde{x} + \lambda_1 v_1 - \tilde{y} - \lambda_2 v_2\|^2 \\ &= \|\tilde{x} - \tilde{y} + \lambda_1 v_1 - \lambda_2 v_2\|^2 \\ &= \|\tilde{x} - \tilde{y}\|^2 + \|\lambda_1 v_1 - \lambda_2 v_2\|^2 \end{aligned}$$

nach Pythagoras, da $\tilde{x} - \tilde{y}$ nach Voraussetzung zu jeder Linearkombination $\lambda_1 v_1 - \lambda_2 v_2$ senkrecht steht:

$$\langle \tilde{x} - \tilde{y},\ \lambda_1 v_1 - \lambda_2 v_2 \rangle = \lambda_1 \langle \tilde{x} - \tilde{y},\ v_1 \rangle - \lambda_2 \langle \tilde{x} - \tilde{y},\ v_2 \rangle = 0.$$

Insgesamt folgt: $\|x - y\|^2 \geq \|\tilde{x} - \tilde{y}\|^2$ (also auch $d(x, y) \geq d(\tilde{x}, \tilde{y})$) und Gleichheit gilt genau dann, wenn

$$\lambda_1 v_1 - \lambda_2 v_2 = 0 \Leftrightarrow \lambda_1 = \lambda_2 = 0,$$

da v_1 und v_2 keine skalaren Vielfachen voneinander sind. Wir haben somit:

$$d(L_1, L_2) = \|\tilde{x} - \tilde{y}\|.$$

Es bleibt zu zeigen, dass die angegebene Bedingung $\tilde{x}$ und $\tilde{y}$ eindeutig bestimmt. Wir schreiben: $\tilde{x} = p_1 + \lambda_1 v_1,\ \tilde{y} = p_2 + \lambda_2 v_2$ für gewisse $\lambda_1, \lambda_2 \in \mathbb{R}$ und $p_1, p_2, v_1, v_2 \in \mathbb{R}^n$. Wegen der Bedingung gilt nun:

$$\langle \tilde{x} - \tilde{y}, v_1 \rangle = 0, \quad \langle \tilde{x} - \tilde{y}, v_2 \rangle = 0,$$

also:

$$\langle p_1 - p_2 + \lambda_1 v_1 - \lambda_2 v_2, v_1 \rangle = 0, \quad \langle p_1 - p_2 + \lambda_1 v_1 - \lambda_2 v_2, v_2 \rangle = 0.$$

Dies liefert ein **lineares Gleichungssystem** (d.h. eine Menge von Bedingungen an λ_1, λ_2, die jeweils ein Polynom vom Grad 1 in den λ_i darstellen) für λ_1 und λ_2:

$$\lambda_1 \cdot \|v_1\|^2 - \lambda_2 \langle v_2, v_1 \rangle = \langle p_2 - p_1, v_1 \rangle, \quad \lambda_1 \cdot \langle v_1, v_2 \rangle - \lambda_2 \|v_2\|^2 = \langle p_2 - p_1, v_2 \rangle.$$

Wir könnten dies nun explizit lösen. Wir machen das hier aber nicht, weil wir in Kürze eine Maschinerie zur Lösung solcher Probleme kennen lernen werden, mit Hilfe der sogenannten **Matrixschreibweise**:

$$\begin{pmatrix} \|v_1\|^2 & -\langle v_2, v_1 \rangle \\ \langle v_1, v_2 \rangle & -\|v_2\|^2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \langle p_2 - p_1, v_1 \rangle \\ \langle p_2 - p_1, v_2 \rangle \end{pmatrix}.$$

Wir werden sehen, dass diese Gleichung genau dann eine eindeutig bestimmte Lösung $(\lambda_1, \lambda_2)^t \in \mathbb{R}^2$ hat, wenn

$$0 \neq \det \begin{pmatrix} \|v_1\|^2 & -\langle v_2, v_1 \rangle \\ \langle v_1, v_2 \rangle & -\|v_2\|^2 \end{pmatrix} := -\|v_1\|^2 \cdot \|v_2\|^2 + \langle v_1, v_2 \rangle^2 \leq 0,$$

wobei det() die sogenannte **Determinante** beschreibt, die wir gleich allgemein einführen werden. Nach der Cauchy–Schwarz'schen Ungleichung ist aber $|\langle v_1, v_2 \rangle| \leq \|v_1\| \cdot \|v_2\|$ und es gilt $<$, da v_1 und v_2 nicht skalare Vielfache voneinander sind. □

Aufgaben

Aufgabe 17.1 (Abstand im $\mathbb{R}^3$). Berechnen Sie den Abstand zwischen den folgenden beiden Geraden im $\mathbb{R}^3$:

$$g : \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \qquad h : \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} + \mu \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}.$$

Aufgabe 17.2 (Spiegel). Im Punkt $A = \begin{pmatrix} -3 \\ -3 \\ 5 \end{pmatrix}$ befinde sich ein Auge, mit Blickrichtung $v = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$.

Ein Objekt habe sein Zentrum im Punkt $O = \begin{pmatrix} -3 \\ 3 \\ -1 \end{pmatrix}$.

Ferner genüge ein (unendlich großer) Spiegel der Gleichung $x = 0$.

1. In welchem Punkt P des Spiegels sieht man das Objekt?
2. Wie groß ist der Winkel OPA?

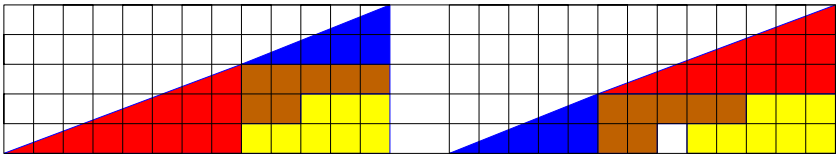
Aufgabe 17.3 (Winkel im $\mathbb{R}^4$).

1. Definieren Sie den Winkel zwischen zwei Hyperebenen im $\mathbb{R}^n$ in sinnvoller Weise.
2. Berechnen Sie den Winkel zwischen den folgenden beiden Hyperebenen im $\mathbb{R}^4$:

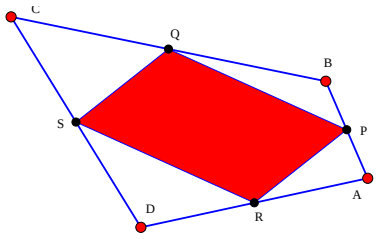
$$H_1 = \left\{ x \in \mathbb{R}^4 \mid \left\langle \begin{pmatrix} 1 \\ 0 \\ 2 \\ 1 \end{pmatrix}, x \right\rangle = 0 \right\}, \qquad H_2 = \left\{ x \in \mathbb{R}^4 \mid \left\langle \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, x - \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \right\rangle = 0 \right\}.$$

Aufgabe 17.4 (Geometrie in der Ebene).

1. Woher kommt die Lücke?



2. Seien vier beliebige Punkte $A, B, C, D \in \mathbb{R}^2$ gegeben. Diese bilden ein Viereck $ABCD$. Die Mittelpunkte der Seiten AB, BC, CD, DA bezeichen wir mit P, Q, R, S (in dieser Reihenfolge). Zeigen Sie: Das Viereck $PQRS$ ist ein Parallelogramm.



18

Abstrakte Vektorräume

$\mathbb{R}^3$ und $\mathbb{R}^n$ sind Beispiele von Vektorräumen. Wir wollen Vektorräume auch in abstrakterer Form einführen, da deren Verwendung sehr häufig notwendig ist, wie wir noch im weiteren Verlauf der Vorlesung sehen werden. Zunächst einmal wollen wir für Skalare auch andere Zahlbereiche zulassen. Zum Beispiel kommen $\mathbb{R}, \mathbb{Q}, \mathbb{C}$ und endliche Körper in Frage.

18.1 Definitionen

Da wir diesen Vorlesungsteil möglichst unabhängig vom ersten Teil machen möchten, besprechen wir hier kurz den Begriff des Körpers. Weitere Details sind im ersten Vorlesungsteil nachzulesen.

Darauf aufbauend führen wir dann den Begriff des Vektorraumes ein.

Definition 18.1. *Ein **Körper** ist ein Tupel $(K, +, \cdot)$ aus einer Menge K und zwei Abbildungen*

$$+ : \ K \times K \to K, \ \ (a, b) \to a + b$$
$$\cdot : \ K \times K \to K, \ \ (a, b) \to a \cdot b,$$

die folgenden Axiomen genügen:

K1: Axiome der Addition
K1.1 (Assoziativgesetz)

$$(a + b) + c = a + (b + c) \quad \forall\, a, b, c \in K.$$

K1.2 (Existenz der Null)

$$\exists\, 0 \in K, \text{ so dass } 0 + a = a \quad \forall\, a \in K.$$

K1.3 (Existenz des Negativen)

$$\forall\, a \in K \;\; \exists - a \in K, \text{ so dass } \; -a + a = 0 \quad (a - b := a + (-b)).$$

K1.4 (Kommutativgesetz)

$$a + b = b + a \quad \forall\, a, b \in K.$$

(Eine Menge K mit Verknüpfung +, die obige Axiome erfüllt, wird auch als ***abelsche Gruppe*** *bezeichnet.)*

K2: Axiome der Multiplikation
K2.1 (Assoziativgesetz)

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c \quad \forall\, a, b, c \in K.$$

K2.2 (Existenz der Eins)

$$\exists\, 1 \in K^* := K \setminus \{0\}, \text{ so dass } a \cdot 1 = a \quad \forall\, a \in K.$$

K2.3 (Existenz des Inversen)

$$\forall\, a \in K^* \quad \exists\, a^{-1} \in K, \text{ so dass } a \cdot a^{-1} = 1.$$

K2.4 (Kommutativgesetz der Multiplikation)

$$a \cdot b = b \cdot a \quad \forall\, a, b \in K.$$

K3: Distributivgesetze *(Punktrechung geht vor Strichrechnung)*

$$\begin{aligned} a \cdot (b + c) &= a \cdot b + a \cdot c \\ (a + b) \cdot c &= a \cdot c + b \cdot c \quad \forall\, a, b, c \in K. \end{aligned}$$

Insbesondere ist also $(K^*, \cdot)$ eine abelsche Gruppe.

Beispiel 18.2. $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ sind Körper.

$(\mathbb{Z}, +, \cdot)$ ist kein Körper, da $n \in \mathbb{Z}$ mit $|n| \geq 2$ kein Inverses hat.

Definition 18.3. *Lassen wir in der Definition des Körpers das Axiom K2.3 weg, so erhalten wir die Axiome eines* ***kommutativen Rings mit*** 1.

Beispiel/Definition 18.4. Für $a \in \mathbb{Z}$ bezeichnen wir mit $\overline{a}$ den Rest bei der Division durch $p \in \mathbb{Z}$. Sei p eine Primzahl. Dann ist

$$\begin{aligned} \mathbb{F}_p &:= \{\overline{0}, \overline{1}, \ldots, \overline{p-1}\} \\ &= \text{die Menge der Reste bei der Division durch } p \text{ in } \mathbb{Z} \end{aligned}$$

ein Körper (mit p Elementen) vermöge der folgenden Verknüpfungen:

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

Häufig läßt man $\bar{}$ weg und schreibt die Elemente als $0, 1, \ldots$. Für Details s. Kapitel 3.

Beispiel 18.5. • $\mathbb{F}_2 = \{\bar{0}, \bar{1}\}$. Verknüpfungstafeln:

$+$	$\bar{0}$	$\bar{1}$
$\bar{0}$	$\bar{0}$	$\bar{1}$
$\bar{1}$	$\bar{1}$	$\bar{0}$

$\cdot$	$\bar{0}$	$\bar{1}$
$\bar{0}$	$\bar{0}$	$\bar{0}$
1	$\bar{0}$	1

also $\bar{1} + \bar{1} = \bar{0} \in \mathbb{F}_2$ (da $1 + 1 \equiv 0 \mod 2$). Häufig schreibt man der Einfachheit halber auch 0 statt $\bar{0}$ und 1 statt $\bar{1}$. Der Körper $\mathbb{F}_2$ ist in vielerlei Hinsicht bemerkenswert; beispielswiese gilt dort: $-1 = +1$.

- $\mathbb{F}_3 = \{\bar{0}, \bar{1}, \bar{2}\}$. Verknüpfungstafeln:

$+$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{0}$
$\bar{2}$	$\bar{2}$	$\bar{0}$	$\bar{1}$

$\cdot$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{2}$	$\bar{0}$	$\bar{2}$	$\bar{1}$

da beispielsweise $2+1 \equiv 0 \mod 3$ und $2+2 \equiv 1 \mod 3$. Auch der Körper $\mathbb{F}_3$ ist außergewöhnlich: Wie man an den Verknüpfungstafeln sehen kann, ist $\bar{2} = -\bar{1}$; daher schreibt man die drei Elemente von $\mathbb{F}_3$ auch häufig der Einfachheit halber $0, 1, -1$ statt $\bar{0}, \bar{1}, \bar{2}$.

Bemerkung 18.6. Ist n eine zusammengesetzte Zahl, etwa eine echte Primzahlpotenz, dann ist $\mathbb{Z}/n := \{0, 1, \ldots, n-1\}$ kein Körper, sondern lediglich ein kommutativer Ring mit 1.

Beispielsweise hat 2 in $\mathbb{Z}/6$ kein Inverses bzgl. der Multiplikation: Es gibt kein $x \in \mathbb{F}_6$, so dass $x \cdot 2 \equiv 1 \mod 6$, da:

$$1 \cdot 2 = 2 \equiv 2 \mod 6, \quad 2 \cdot 2 = 4 \equiv 4 \mod 6, \quad 3 \cdot 2 = 6 \equiv 0 \mod 6,$$
$$4 \cdot 2 = 8 \equiv 2 \mod 6, \quad 5 \cdot 2 = 10 \equiv 4 \mod 6, \quad 0 \cdot 2 = 0 \equiv 0 \mod 6.$$

Bemerkung 18.7. $0 \cdot a = 0 \quad \forall\, a \in K$ und $(-1) \cdot (-1) = 1$ gilt in allen Körpern.

Definition 18.8. *Sei K (genauer $(K, +, \cdot)$) ein Körper. Ein K-**Vektorraum** (kurz K–**VR**) ist ein Tupel $(V, +, \cdot)$, wobei V eine Menge ist, zusammen mit zwei Abbildungen*

$$+ : V \times V \to V, \ (v, w) \to v + w$$
$$\cdot : K \times V \to V, \ (\lambda, v) \to \lambda \cdot v,$$

die den folgenden Axiomen genügen:

VR 1: Axiome der Vektoraddition

VR 1.1 Assoziativität: $u + (v + w) = (u + v) + w \quad \forall\, u, v, w \in V$.

VR 1.2 Existenz der Null[1]*:* $\exists 0 \in V$*, so dass* $0 + v = v \quad \forall v \in V$.

VR 1.3 Existenz des Negativen: $\forall v \in V \exists -v \in V$ *so dass* $-v + v = 0$.

VR 1.4 $v + w = w + v \quad \forall v, w \in V$.

Mit anderen Worten $(V, +)$ *ist eine abelsche Gruppe.*

VR 2: Axiome der Skalarmultiplikation

VR 2.1 $(\lambda \cdot \mu) \cdot v = \lambda \cdot (\mu \cdot v) \quad \forall v \in V \quad \forall \lambda \in K$.

VR 2.2 $1 \cdot v = v \quad \forall v \in V$ *gilt für das Einselement* $1 \in K$.

VR 3: Distributivgesetze

$$(\lambda + \mu) \cdot v = \lambda \cdot v + \mu \cdot v \quad \forall \lambda, \mu \in K, \quad \forall v \in V,$$
$$\lambda \cdot (v + w) = \lambda \cdot v + \lambda \cdot w \quad \forall \lambda \in K \quad \forall v, w \in V.$$

Die Elemente $\lambda \in K$ *heißen* ***Skalare****, die Elemente* $v \in V$ *heißen* ***Vektoren****.*

Bemerkung/Definition 18.9. Verlangen wir nicht mehr, dass K ein Körper ist, sondern nur, dass $R = K$ ein (kommutativer) Ring mit 1 ist, so erhalten die Definition eines **(Links-)Moduls** über R.

Die Theorie der Module ist deutlich verschieden von der Theorie der Vektorräume.

18.2 Beispiele von Vektorräumen

1. $\mathbb{R}^n$ ist ein $\mathbb{R}$-Vektorraum, $\mathbb{Q}^n$ ein $\mathbb{Q}$-VR und allgemein

$$K^n = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mid x_i \in K \right\}$$

   ein K-Vektorraum, wenn K ein Körper ist.

2. Die Polynome

$$\mathbb{R}[x] := \{p = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \mid n \in \mathbb{N}, a_i \in \mathbb{R}\}$$

   bilden einen $\mathbb{R}$-Vektorraum: Seien $p = 5x^2 - 3x$, $q = x^3 + x - 1 \in \mathbb{R}[x]$. Dann ist:

[1] Achtung: Mit 0 bezeichnen wir sowohl die Zahl $0 \in K$, also auch den Null-Vektor $0 = (0, \ldots, 0)^t$, als auch den Nullvektorraum $\{(0, \ldots, 0)^t\}$. Es wird immer aus dem Kontext verständlich sein, welche 0 gemeint ist.

$$p + q = x^3 + 5x^2 - 2x - 1 \in \mathbb{R}[x],$$
$$\frac{1}{2} \cdot p = \frac{5}{2}x^2 - \frac{3}{2}x \in \mathbb{R}[x].$$

Außerdem ist beispielsweise $(x-1)^2 + 2(x+1) \in \mathbb{R}[x]$.

3. Die Mengen (siehe für die Definitionen Kapitel 9)

$$C^0[a,b] := \{f\colon [a,b] \to \mathbb{R} \mid f \text{ ist stetig }\},$$
$$C^1[a,b] := \{f\colon [a,b] \to \mathbb{R} \mid f \text{ ist stetig und differenzierbar }\},$$
$$C^\infty[a,b] = \{f\colon [a,b] \to \mathbb{R} \mid f \text{ ist unendlich oft differenzierbar }\}$$

sind $\mathbb{R}$-Vektorräume; hierbei ist $[a,b] = \{x \mid a \leq x \leq b\}$ das sogenannte abgeschlossene Intervall der reellen Zahlen zwischen a und b (s.5.7 für Details). Vektorräume von Funktionen spielen beispielsweise in der Bildbearbeitung eine Rolle.

$$\mathbb{R}^{[a,b]} := \{f\colon [a,b] \to \mathbb{R} \mid f \text{ ist Abbildung }\}$$

ist ebenfalls ein $\mathbb{R}$-Vektorraum.

4. Sei K ein Körper und M eine Menge. Dann ist

$$K^M := \{f\colon M \to K\}$$

ein K-Vektorraum.

5. In der Kodierungstheorie verwendet man häufig Vektorräume über endlichen Körpern, etwa $K = \mathbb{F}_2$:

$$V = \mathbb{F}_2^n = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \;\middle|\; x_i \in \{0,1\} \right\},$$

die Menge der n–Tupel von Elementen aus $\mathbb{F}_2$. Allgemein definiert man für einen endlichen Körper $K = \mathbb{F}_p$ und zwei Vektoren

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},\ y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{F}_p^n$$

die **Hammingdistanz**:

$$d(x,y) := |\{i \mid x_i \neq y_i\}|.$$

Beispielsweise ist

$$d\left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right) = 2.$$

In der Kodierungstheorie ist ein **Code** eine Teilmenge $C \subseteq \mathbb{F}_p^n$. Ein Element $x \in C$ heißt ein **Codewort**. Die **Minimaldistanz** von C ist

$$D = \min_{x,y \in C,\, x \neq y} d(x, y).$$

Rauscht der Kanal so wenig, dass man weniger als $\frac{D}{2}$ Fehler erwarten darf, dann können wir das Wort x aus dem übertragenen Wort y zurückbekommen, indem wir

$$z \in C \text{ bestimmen mit } d(z, y) = \min_{c \in C} d(c, y).$$

Bei weniger als $\frac{D}{2}$ Fehlern in der Übertragung gilt $z = x$.

Besonders häufig werden für Codes Teilmengen verwendet, die selbst wieder Vektorräume sind, nämlich sogenannte Untervektorräume. Dazu kommen wir im nächsten Abschnitt.

18.3 Untervektorräume

Definition 18.10. *Sei V ein K-Vektorraum. Eine nicht leere Teilmenge $U \subseteq V$ heißt **Untervektorraum** (kurz **UVR**), wenn*

1. $u_1, u_2 \in U \Rightarrow u_1 + u_2 \in U$,
2. $u \in U, \lambda \in K \Rightarrow \lambda \cdot u \in U$.

Bemerkung 18.11. Insbesondere ist dann mit $u \in U$ auch $(-1) \cdot u = -u \in U$. Mit anderen Worten

1. $+ \colon U \times U \to V,\ (u_1, u_2) \mapsto u_1 + u_2 \in U$,
2. $\cdot \colon K \times U \to V,\ (k, u) \mapsto k \cdot u \in U$.

$(U, +, \cdot)$ ist dann ein Vektorraum.

Bemerkung 18.12. Ist $U \subseteq V$ ein Untervektorraum, dann gilt $0 \in U$. Denn $U \neq \emptyset$ und somit: $\exists\, u \in U \Rightarrow -u \in U \Rightarrow 0 = -u + u \in U$.

Frage 18.13. *Welche der folgenden Teilmengen des $\mathbb{R}^2$ sind Untervektorräume (s. Abb. 18.1 und 18.2)?*

$$U_1 = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; x_1 + 2x_2 = 0 \right\},$$
$$U_2 = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_2 \geq 0 \right\},$$
$$U_3 = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_1 + x_2 \leq 1 \right\},$$
$$U_4 = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_1 + x_2 = 1 \right\}.$$

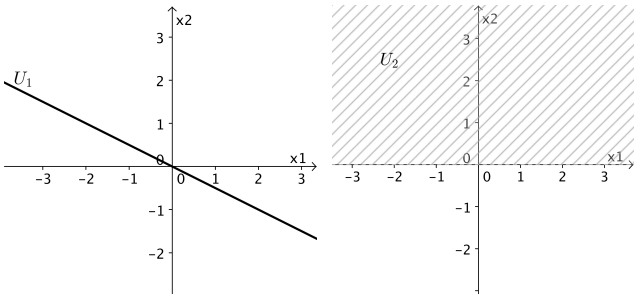


Abbildung 18.1. Die Teilmengen U_1 und U_2 des $\mathbb{R}^2$.

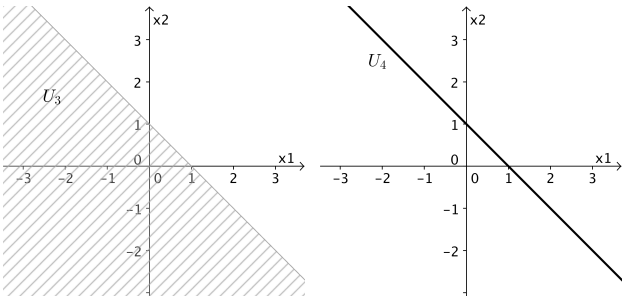


Abbildung 18.2. Die Teilmengen U_3 und U_4 des $\mathbb{R}^2$.

Antwort. Nur U_1 ist ein Untervektorraum.

U_2 ist kein Untervektorraum, weil $(0,1)^t \in U_2$, aber $-(0,1)^t = (0,-1)^t \notin U_2$.

U_3 ist kein Untervektorraum, weil große skalare Vielfache von $u \in U_3 \setminus \{0\}$ nicht in U_3 liegen.

U_4 ist kein Untervektorraum, weil z.B. $(\frac{1}{2}, \frac{1}{2})^t \in U_4$, aber $2 \cdot (\frac{1}{2}, \frac{1}{2})^t = (1,1)^t \notin U_4$ und weil $0 \notin U_4$. □

Bemerkung 18.14. 1. Eine Gerade L (Hyperebene H) $\subseteq \mathbb{R}^n$ ist ein Untervektorraum genau dann, wenn $0 \in L$, (bzw. $0 \in H$).

2. Sind $U, W \subseteq V$ Untervektorräume, dann ist auch $U \cap W \subseteq V$ ein Untervektorraum.
3. Der kleinste Untervektorraum von V ist der Nullraum $0 = \{0\}$.
4. Sind $U, W \subseteq V$ Untervektorräume, dann ist im Allgemeinen $U \cup W \subseteq V$ kein Untervektorraum.

   Beweis. Seien $V = \mathbb{R}^2$ und

   $$U = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_1 = 0 \right\}, \; W = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_2 = 0 \right\}.$$

   Dann ist die Menge

   $$U \cup W = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_1 \cdot x_2 = 0 \right\}$$

   kein Untervektorraum von V, denn:

   $$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in U \cup W, \text{ aber } \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \notin U \cup W.$$

   □

18.4 Lineare Unabhängigkeit und Basen

Unser Ziel ist es, einen Dimensionsbegriff für abstrakte K-Vektorräume V zu entwickeln. Wir wollen $\dim V \in \mathbb{N} \cup \{\infty\}$ definieren.

Natürlich soll $\dim \mathbb{R}^n = n$, und für $L, H \subseteq \mathbb{R}^n$ Gerade bzw. Hyperebene mit $0 \in L$ ($0 \in H$) soll $\dim L = 1$ und $\dim H = n - 1$ gelten. Anschaulich ist die Dimension die minimale Anzahl von Vektoren, die wir benötigen, um V aufzuspannen.

Definition 18.15. *1. Seien V ein K-Vektorraum und $v_1, \ldots, v_n \in V$ Vektoren. Eine **Linearkombination** von $v_1, \ldots, v_n$ ist ein Ausdruck*

$$v = \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_n v_n \in V,$$

wobei $\lambda_1, \ldots, \lambda_n \in K$. Mit

$$\langle v_1, \ldots, v_n \rangle := \text{Spann}(v_1, \ldots, v_n) := \{\lambda_1 v_1 + \cdots + \lambda_n v_n \mid \lambda_i \in K\} \subseteq V$$

bezeichnen wir den ***Spann*** *von $v_1, \ldots, v_n$ (oder ausführlicher: den von $v_1, \ldots, v_n$* ***aufgespannten Untervektorraum****).*

Wir setzen: $\langle \emptyset \rangle := \{0\} =: 0$.

$\langle v_1, \ldots, v_n \rangle \subseteq V$ ist der kleinste Untervektorraum von V, der $v_1, \ldots, v_n$ enthält.

2. *$v_1, \ldots v_n$* ***erzeugen*** *V, wenn $\langle v_1, \ldots, v_n \rangle = V$. Mit anderen Worten: $\forall\, v \in V \quad \exists\, \lambda_1, \ldots, \lambda_n \in K : \; v = \lambda_1 v_1 + \cdots + \lambda_n v_n$. Wir sagen auch, die Familie $\{v_i\}_{i=1,\ldots,n}$ bildet ein* ***Erzeugendensystem*** *von V.*
3. *$v_1, \ldots, v_n$ heißen* ***linear unabhängig****, wenn $\forall\, \lambda_1, \ldots, \lambda_n \in K$ gilt:*

$$0 = \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_n v_n \quad \Rightarrow \quad \lambda_1 = \cdots = \lambda_n = 0.$$

Andernfalls heißen $v_1, \ldots, v_n$ ***linear abhängig****.*

Beispiel 18.16. $V = \mathbb{R}^3$. Wir betrachten die vier Vektoren:

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},\; v_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},\; v_3 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix},\; v_4 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^3.$$

1. v_1, v_2 sind linear unabhängig:

$$\lambda_1 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 + \lambda_2 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = 0 \;\Rightarrow\; \lambda_1 = \lambda_2 = 0.$$

2. v_1, v_2, v_3, v_4 sind linear abhängig, zum Beispiel: $1 \cdot v_1 + 1 \cdot v_2 - 1 \cdot v_3 + 0 \cdot v_4 = 0$, also sogar v_1, v_2, v_3 sind schon linear abhängig (Abb. 18.3). Wir werden noch sehen, dass die lineare Abhängigkeit klar ist, weil lineare Unabhängigkeit schon aus Dimensionsgründen nicht sein kann.
3. v_1, v_2, v_4 sind linear unabhängig, weil

$$0 = \lambda_1 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \lambda_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 + \lambda_2 + \lambda_3 \\ \lambda_1 + \lambda_3 \\ \lambda_2 + \lambda_3 \end{pmatrix}$$

$\Rightarrow \lambda_3 = -\lambda_1 = -\lambda_2$ wegen der 2. und 3. Komponenten.

Eingesetzt in die erste Komponente ergibt sich:

$$\lambda_1 + \lambda_1 - \lambda_1 = \lambda_1 = 0 \;\Rightarrow \lambda_3 = 0 \;\Rightarrow \lambda_2 = 0.$$

4. v_1, v_2, v_4 bilden ein Erzeugersystem, weil sich jeder Vektor v, etwa $v = (b_1, b_2, b_3)^t$, als Linearkombination dieser Vektoren darstellen lässt.

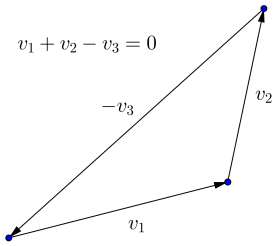


Abbildung 18.3. Ist die Summe dreier Vektoren der Nullvektor, so bedeutet dies, dass sie aneinander gehängt einen geschlossenen Vektorzug ergeben.

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \lambda_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 + \lambda_2 + \lambda_3 \\ \lambda_1 + \lambda_3 \\ \lambda_2 + \lambda_3 \end{pmatrix}$$

liefert nämlich ein Gleichungssystem,

$$\begin{aligned} (I) \quad & b_1 = \lambda_1 + \lambda_2 + \lambda_3 \\ (II) \quad & b_2 = \lambda_1 + \lambda_3 \\ (III) \quad & b_3 = \lambda_2 + \lambda_3, \end{aligned}$$

welches eine Lösung hat:

I - II - III ergibt: $b_1 - b_2 - b_3 = -\lambda_3 \Rightarrow \lambda_3 = b_2 + b_3 - b_1$.
Dies in II eingesetzt liefert: $b_2 = \lambda_1 + b_2 + b_3 - b_1 \Rightarrow \lambda_1 = b_1 - b_3$.
Dies wiederum in III eingesetzt: $b_3 = \lambda_2 + b_2 + b_3 - b_1 \Rightarrow \lambda_2 = b_1 - b_2$.
Also:

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} b_1 - b_3 \\ b_1 - b_2 \\ -b_1 + b_2 + b_3 \end{pmatrix}.$$

Probe: ✓

Fazit: Lineare Abhängigkeit und Erzeugung zu entscheiden läuft darauf hinaus, lineare Gleichungssysteme zu lösen.

Beispiel/Definition 18.17. $V = \mathbb{R}[x] = \{$ aller Polynome $\}$ ist ein $\mathbb{R}$-Vektorraum. Für ein Polynom

$$p = a_d x^d + a_{d-1} x^{d-1} + \cdots + a_1 x + a_0, \quad a_i \in \mathbb{R},$$

heißen die a_i **Koeffizienten** von p. Ist $a_d \neq 0$, dann hat p den **Grad** $\deg p := d$. Wir setzen $\deg 0 := -\infty$. Es gilt[2]:

[2]Das Zeichen $\cong$ bedeutet, dass die beiden Vektorräume im Wesentlichen gleich sind; genauer werden wir dies erst später definieren.

$$\begin{aligned}\mathbb{R}[x]_{\leq d} &:= \{p \in \mathbb{R}[x] \mid \deg p \leq d\} \\ &= \{a_d x^d + \cdots + a_1 x + a_0 \mid a_i \in \mathbb{R}\} \\ &\cong \mathbb{R}^{d+1}.\end{aligned}$$

Z.B.: $\mathbb{R}[x]_{\leq 3} \ni 1, x, x^2, x^3$ sind linear unabhängig und erzeugen $\mathbb{R}[x]_{\leq 3}$. Dagegen bilden

$$p_1 = x^2 + 1,\ p_2 = x^2 - 1,\ p_3 = x^3 + 1,\ p_4 = x^3 - 1 \in \mathbb{R}[x]_{\leq 3}$$

kein Erzeugendensystem für $\mathbb{R}[x]_{\leq 3}$, da:

$$\langle x^2 + 1, x^2 - 1, x^3 + 1, x^3 - 1\rangle \subseteq \{p = a_3 x^3 + a_2 x^2 + a_1 x + a_0 \mid a_1 = 0\}.$$

p_1, p_2, p_3, p_4 sind vielmehr linear abhängig, da die Relation $\lambda_1 p_1 + \lambda_2 p_2 + \lambda_3 p_3 + \lambda_4 p_4 = 0$ für $\lambda_1 = 1, \lambda_2 = -1, \lambda_3 = -1, \lambda_4 = 1$ eine **nicht–triviale lineare Relation** (d.h., nicht alle $\lambda_i = 0$) ist.

Beispiel/Definition 18.18. Seien $k, d \in \mathbb{N}$ und $a, b \in \mathbb{R}$ mit $a < b$.

$$\mathbb{R}[x]_{\leq d} \subseteq \mathbb{R}[x]_{\leq d+1} \subseteq \mathbb{R}[x] \subseteq C^\infty[a,b] \subseteq C^k[a,b] \subseteq C^0[a,b] \subseteq \mathbb{R}^{[a,b]}$$

ist eine **aufsteigende Kette** von Untervektorräumen des Vektorraumes $\mathbb{R}^{[a,b]} = \{f\colon [a,b] \to \mathbb{R}\}$.

Definition 18.19. *Sei V ein Vektorraum und seien $v_1, \ldots, v_n \in V$ Vektoren. $\beta = \{v_i\}_{i=1,\ldots,n}$ ist eine **Basis** von V, wenn*

1. *$\{v_1, \ldots, v_n\}$ den Vektorraum V erzeugen und*
2. *$\{v_1, \ldots, v_n\}$ linear unabhängig sind.*

Beispiel 18.20. 1. $\mathbb{R}^n$. Die Vektoren $e_1, \ldots, e_n$,

$$e_1 = \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix},\ \ldots,\ e_i = \begin{pmatrix}0\\\vdots\\1\\\vdots\\0\end{pmatrix},\ \ldots,\ e_n = \begin{pmatrix}0\\\vdots\\0\\1\end{pmatrix}$$

mit nur einer 1 an der i-ten Position und sonst 0-en, bilden eine Basis des $\mathbb{R}^n$, die sogenannte Standardbasis.

2. $\mathbb{R}[x]_{\leq 3}$ hat die Basis a, x, x^2, x^3, $a \neq 0$. Es gibt aber auch andere Basen. Beispielsweise ist $1, x-a, (x-a)^2, (x-a)^3$ mit $a \neq 0$ auch eine Basis von $\mathbb{R}[x]_{\leq 3}$. Dies ist die Basis, welche für das dritte Taylorpolynom $T_a^3 f$ verwendet wird (siehe Kapitel 15).

Bemerkung 18.21. Ist $\{v_1, \ldots, v_n\} \subseteq V$ eine Basis, dann hat jeder Vektor $w \in V$ eine eindeutig bestimme Darstellung

$$w = \lambda_1 v_1 + \cdots + \lambda_n v_n \text{ mit } \lambda_i \in K$$

als Linearkombination.

Beweis. Existenz ist die erste Bedingung, Eindeutigkeit die zweite: die Differenz zweier Darstellungen ist nämlich eine Relation, die nach der zweiten Bedingung trivial ist.

Ausführlicher: Sei $w \in V$. Da eine Basis ein Erzeugendensystem ist, gibt es $\lambda_i \in K$, so dass

$$w = \lambda_1 v_1 + \cdots + \lambda_n v_n.$$

Ist $w = \lambda_1' v_1 + \cdots + \lambda_n' v_n$ eine weitere Darstellung, so gilt für die Differenz:

$$0 = (\lambda_1 - \lambda_1') v_1 + \cdots + (\lambda_n - \lambda_n') v_n.$$

Wegen der Definition einer Basis, folgt: $\lambda_1 - \lambda_1' = 0, \ldots, \lambda_n - \lambda_n' = 0$. Damit gilt aber schon: $\lambda_1 = \lambda_1', \ldots, \lambda_n = \lambda_n'$. Dies zeigt die Eindeutigkeit. □

Satz 18.22. *Sei V ein K–VR und seien $v_1, \ldots, v_n \in V$ Vektoren. Äquivalent sind:*

1. *$\{v_1, \ldots, v_n\}$ ist eine Basis von V.*
2. *$\{v_1, \ldots, v_n\}$ bilden ein unverlängerbares System von linear unabhängigen Vektoren.*
3. *$\{v_1, \ldots, v_n\}$ bilden ein unverkürzbares Erzeugendensystem von V.*
4. *Jeder Vektor $w \in V$ hat genau eine Darstellung $w = \lambda_1 v_1 + \cdots + \lambda_n v_n$ mit $\lambda_i \in K$ als Linearkombination von $v_1, \ldots, v_n$.*

Beweis. 1. ⇒ 4.: Das ist Bemerkung 18.21.

4. ⇒ 2. und 4. ⇒ 3. sind jeweils klar.

2., 3. ⇒ 1. nach der Definition einer Basis.
Es bleibt also 2. ⇔ 3. zu zeigen.

2. ⇒ 3.: Sei $v_1, \ldots, v_n$ ein unverlängerbares System von linear unabhängigen Vektoren. Mit jedem weiteren Vektor $0 \neq w \in V$ erhalten wir also ein System von linear abhängigen Vektoren, also:

$$\exists\, \lambda_1, \ldots, \lambda_n, \lambda_{n+1} \in K: \quad \lambda_1 v_1 + \cdots + \lambda_n v_n + \lambda_{n+1} w = 0,$$

wobei wenigstens ein $\lambda_i \neq 0$. Es ist $\lambda_{n+1} \neq 0$, da $v_1, \ldots, v_n$ linear unabhängig sind. Da K ein Körper ist, gilt $\frac{1}{\lambda_{n+1}} \in K$.
Es folgt:

$$w = (-\frac{\lambda_1}{\lambda_{n+1}})v_1 + \cdots + (-\frac{-\lambda_n}{\lambda_{n+1}}) \cdot v_n.$$

Dies gilt für beliebige $w \in V$, d.h. $v_1, \ldots, v_n$ erzeugen V. $v_1, \ldots, v_n$ ist unverkürzbar, das heißt nach Weglassen eines der Vektoren haben wir kein Erzeugendensystem mehr. Wenn wir zum Beispiel v_n weglassen und $v_1, \ldots, v_{n-1}$ noch ein Erzeugendensystem wäre, gäbe es eine Darstellung

$$v_n = \mu_1 v_1 + \ldots mu_{n+1} v_{n+1},$$

d.h. $v_1, \ldots, v_n$ wäre linear abhängig. Dies widerspricht der Voraussetzung.

3. $\Rightarrow$ 2.: Sei $v_1, \ldots, v_n$ ein unverkürzbares Erzeugendensystem. Dann sind $v_1, \ldots, v_n$ linear unabhängig. In der Tat: Wäre

$$\lambda_1 v_1 + \ldots \lambda_n v_n = 0$$

ein nicht triviale Relation, etwa mit $\lambda_n \neq 0$, dann:

$$v_n = \left(\frac{-\lambda_1}{\lambda_n}\right) v_1 + \cdots + \left(\frac{-\lambda_{n-1}}{\lambda_n}\right) v_{n-1}.$$

Doch dann wären schon $v_1, \ldots, v_{n-1}$ erzeugend, im Widerspruch zur Voraussetzung wäre also $v_1, \ldots, v_n$ zu einem linear unabhängigem System $v_1, \ldots, v_{n-1}$ zu verkürzen. □

Beispiel 18.23. Das Erzeugendensystem $v_1 = (1,0), v_2 = (0,1), v_3 = (1,1)$ von Vektoren des $\mathbb{R}^2$ ist verkürzbar. Wir können sogar jeden beliebigen der drei Vektoren weglassen und erhalten immer noch ein System, das ganz $\mathbb{R}^2$ erzeugt: $\langle v_1, v_2 \rangle = \langle v_1, v_3 \rangle = \langle v_2, v_3 \rangle = \mathbb{R}^2$.

Bei $w_1 = (1,0), w_2 = (0,1), w_3 = (0,2)$ können wir allerdings w_1 nicht weglassen, da w_2 und w_3 nur die y-Achse erzeugen:

$$\langle w_2, w_3 \rangle = \{(a,b) \in \mathbb{R}^2 \mid a = 0\} \subsetneq \mathbb{R}^2.$$

18.5 Dimension

Da wir nun wissen, was eine Basis eines K–Vektorraumes ist, können wir nun endlich dessen Dimension definieren:

Definition 18.24. *Sei V ein K-Vektorraum. Dann definieren wir die **Dimension** von V durch*

$$\dim V := \dim_k V = \begin{cases} n, & \text{falls } V \text{ eine Basis } v_1, \ldots, v_n \text{ aus } n \text{ Vektoren hat}, \\ \infty, & \text{sonst}. \end{cases}$$

Der zweite Fall $\dim V = \infty$ tritt genau dann ein, wenn V kein endliches Erzeugendensystem hat.

Beispiel 18.25. 1. $\dim K^n = n$, da $e_1, \dots, e_n$ eine Basis ist.

2. $\dim \mathbb{R}[x] = \infty$, da wir aus endlich vielen Polynomen nur Polynome von einem beschränkten Grad linear kombinieren können.
3. $\dim \mathbb{R}[x]_{\leq d} = d + 1$, da $1, x, x^2, \dots, x^d$ eine Basis ist.

Bemerkung 18.26. Es ist nicht klar, dass die obige Definition der Dimension eine vernünftige ist. Unklar ist bislang, ob je zwei Basen von V gleich viele Elemente haben. Also müssen wir zeigen, dass die Dimension wohldefiniert ist, d.h. unabhängig von der Wahl der Basis. Dies zu zeigen ist unser nächstes Ziel.

Lemma 18.27 (Austauschlemma). *Sei $v_1, \dots, v_n$ eine Basis von V und $w \in V$ ein weiterer Vektor mit $w \neq 0$. Dann existiert ein $i \in \{1, \dots, n\}$, so dass wir nach Austausch von v_i mit w nach wie vor eine Basis haben. Ist etwa $i = 1$, was man durch Umnummerierung erreichen kann, dann ist also $w, v_2, \dots, v_n$ eine Basis von V.*

Beweis. w ist eine Linearkombination

$$w = \lambda_1 v_1 + \cdots + \lambda_n v_n$$

für gewisse $\lambda_i \in K$, da $v_1, \dots, v_n$ ein Erzeugendensystem bilden. Wenigstens ein $\lambda_i \in K$ ist $\neq 0$, da w nicht der Nullvektor ist. Nach Umnummerieren können wir $\lambda_1 \neq 0$ annehmen.

Wir zeigen:

1. $w, v_1, \dots, v_n$ ist ein Erzeugendensystem.
2. $w, v_1, \dots, v_n$ sind linear unabhängig.

Zunächst einmal gilt:

$$v_1 = \frac{1}{\lambda_1} w + \Big(\frac{-\lambda_2}{\lambda_1}\Big) v_2 + \cdots + \Big(\frac{-\lambda_n}{\lambda_1}\Big) v_n,$$

da $\frac{1}{\lambda_1} \in K$ existiert.

Sei $u \in V$ ein beliebiger Vektor. Dann existieren $\mu_1, \dots, \mu_n \in k$, so dass

$$u = \mu_1 v_1 + \cdots + \mu_n v_n,$$

da $v_1, \dots, v_n$ ganz V erzeugen. Also:

$$u = \mu_1\Big(\frac{1}{\lambda_1}w + \Big(\frac{-\lambda_2}{\lambda_1}\Big)v_2 + \cdots + \Big(\frac{-\lambda_n}{\lambda_1}\Big)v_n\Big) + \lambda_2 v_2 + \cdots + \mu_n v_n$$
$$= \frac{\mu_1}{\lambda_1}w + \Big(\mu - \lambda_2\frac{\mu_1}{\lambda_1}\Big)v_2 + \cdots + \Big(\mu_n - \lambda_n\frac{\mu_1}{\lambda_1}\Big)v_n.$$

D.h., $w, v_2, \ldots, v_n$ erzeugen V.

Zur linearen Unabhängigkeit: Angenommen,

$$0 = \mu_1 w + \mu_2 v_2 + \cdots + \mu_n v_n, \quad \mu_i \in K.$$

Einsetzen der Ausgangsgleichung für w liefert

$$0 = \mu_1\lambda_1 v_1 + (\mu_2 + \mu_1\lambda_2)v_2 + \cdots + (\mu_n + \mu_1\lambda_n)v_n.$$

Da $v_1, \ldots, v_n$ linear unabhängig sind, folgt

$$\mu_1\lambda_1 = 0, \quad \mu_2 + \mu_1\lambda_2 = 0, \quad \ldots, \quad \mu_n + \mu_1\lambda_n = 0 \in K.$$

Nach Voraussetzung gilt:

$$\lambda_1 \neq 0 \Rightarrow \mu_1 = \frac{1}{\lambda_1}\mu_1\lambda_1 = 0.$$

Einsetzen liefert: $\mu_2 = \cdots = \mu_n = 0$. □

Satz 18.28 (Austauschsatz von Steinitz). *Sei V ein K-Vektorraum und $v_1, \ldots, v_r$ eine Basis von V und $w_1, \ldots, w_n$ eine Familie von linear unabhängigen Vektoren. Dann gilt $n \leq r$ und es existieren Indices $i_1, \ldots, i_n \in 1, \ldots, r$ so dass wir nach Austausch von v_{i_k} mit w_k nach wie vor eine Basis haben. Gilt etwa $i_1 = 1, \ldots, i_n = n$, was durch Umnummerierung von $v_1, \ldots, v_r$ erreicht werden kann, dann ist also $w_1, \ldots, w_n, v_{n+1}, \ldots, v_r$ eine Basis von V.* Achtung*: $n \leq r$ wird bewiesen und nicht vorausgesetzt.*

Beweis. Induktion nach n.
Für $n = 0$ ist nichts zu zeigen. Sei also $n \geq 1$ und der Satz für $n-1$ schon gezeigt. Dann gilt $r \geq n - 1$ und wir müssen nur noch den Fall $r = n - 1$ ausschließen. Nach der Induktionsvoraussetzung können wir nach Umnummerierung von $v_1, \ldots, v_r$ annehmen, dass

$$w_1, \ldots, w_{n-1}, v_n, \ldots, v_r$$

eine Basis ist, denn auch die Familie $w_1, \ldots, w_{n-1}$ ist linear unabhängig. w_n hat eine Darstellung

$$w_n = \lambda_1 w_1 + \cdots + \lambda_{n-1}w_{n-1} + \lambda_n v_n + \cdots + \lambda_r v_r \text{ mit } \lambda_i \in K.$$

Nicht alle Koeffizienten $\lambda_n, \ldots, \lambda_r$ können 0 sein. Insbesondere $r \geq n$ denn sonst wären $w_1, \ldots, w_n$ linear abhängig, im Widerspruch zur Voraussetzung.

Also ist einer der Koeffizienten $\lambda_n, \ldots, \lambda_r$ nicht 0; nach Umnummerieren von $v_n, \ldots, v_r$ können wir annehmen, dass $\lambda_n \neq 0$. Nach dem Austauschlemma ist dann auch

$$w_1, \ldots, w_n, v_{n+1}, \ldots, v_r$$

eine Basis von V. □

Korollar 18.29. *Je zwei Basen eines endlich–dimensionalen K-Vektorraums V haben gleich viele Elemente. Insbesondere ist*

$$\dim V := \begin{cases} n, & \text{falls } \exists \text{ Basis } v_1, \ldots, v_n, \\ \infty, & \text{sonst} \end{cases}$$

wohldefiniert.

Beweis. Es seien $w_1, \ldots, w_n$ und $v_1, \ldots, v_r$ Basen von V. Dann sind $w_1, \ldots, w_n$ linear unabhängig und nach dem Austauschsatz ist deshalb $n \leq r$. Die Ungleichung $r \leq n$ folgt durch Vertauschen der Rolle der w's und der v's. Also $r = n$. Gibt es keine endliche Basis, dann ist V nicht endlich erzeugt, da man aus jedem endlichen Erzeugendensystem durch eventuelles Weglassen eine Basis erhält. □

Korollar 18.30 (Basisergänzungssatz). *Sei $v_1, \ldots, v_n$ eine Familie linear unabhängiger Vektoren in einem endlich–dimensionalen Vektorraum V und sei $r = \dim V < \infty$. Dann kann man diese Familie zu einer Basis*

$$v_1, \ldots, v_n, v_{n+1}, \ldots, v_r$$

von V ergänzen.

Beweis. Nach dem vorigen Korollar ist $n \leq r$. Ist $n < r$, so gibt es, ebenfalls wegen des Korollars, einen Vektor $w \in V \setminus \langle v_1, \ldots, v_n \rangle$. Induktiv können wir dies fortführen, bis wir schließlich eine Basis erhalten. □

Tragen wir alle bisherigen Resultate zusammen, erhalten wir:

Korollar 18.31. *1. Jeder endlich–dimensionale Vektorraum besitzt eine Basis.*

2. *Ist V ein Vektorraum der Dimension $n = \dim V < \infty$, dann ist jede Familie von mehr als n Vektoren in V linear abhängig.*
3. *Sei $U \subseteq V$ ein Untervektorraum. Dann gilt:* $\dim U \leq \dim V$.
   Ist V endlich–dimensional und $\dim U = \dim V$, *so folgt $U = V$.*

Beweis. □

Bemerkung 18.32. Für unendlich–dimensionale Vektorräume

$$U \subseteq V \text{ mit } \dim U = \dim V = \infty$$

kann man auf $U = V$ nicht schließen. Zum Beispiel: $\mathbb{R}[t] \subset C^0[a, b]$, da nicht jede stetige Funktion ein Polynom ist.

Aufgaben

Aufgabe 18.1 (Lineare Unabhängigkeit).

1. Prüfen Sie, ob die folgenden Vektoren linear unabhängig sind. Bestimmen Sie in jedem Fall die Dimension des aufgespannten Raumes und geben Sie eine Basis an.
   a) $(1,1,0)^t$, $(1,0,1)^t$, $(0,1,1)^t \in (\mathbb{F}_2)^3$.
   b) $(1,2,3)^t$, $(2,3,4)^t$, $(3,4,5)^t \in \mathbb{R}^3$.
   c) $(5,0,5,-4)^t$, $(0,5,-5,-3)^t$, $(5,-5,10,-1)^t$, $(-4,-3,-1,5)^t \in \mathbb{R}^4$.
2. Für welche $\lambda \in \mathbb{R}$ sind die Vektoren $(2,\lambda,3)^t$, $(1,-1,2)^t$, $(-\lambda,4,-3)^t \in \mathbb{R}^3$ linear abhängig? Stellen Sie für diese λ den letzten Vektor als Linearkombination der ersten beiden dar.

Aufgabe 18.2 (Untervektorräume). Welche der folgenden Mengen U_i sind Untervektorräume der Vektorräume V_i? Berechnen Sie in diesen Fällen auch deren Dimension.

1. $V_1 := \mathbb{R}^5,\ U_1 := \{p \in \mathbb{R}^5 \mid \|p\| = 1\}$.
2. $V_2 := \mathbb{R}^4,\ U_2 := \{(x,y,z,w) \in \mathbb{R}^4 \mid x+y+z+w = 0,\ w = 0\}$.
3. $V_3 := \mathbb{R}^3,\ U_3 := \left\{p \in \mathbb{R}^3 \mid \langle p, (1,2,3)^t\rangle = 0\right\}$.
4. $V_4 := \mathbb{R}^4,\ U_4 := \{(x,y,z,w) \in \mathbb{R}^4 \mid (x+y)\cdot(x-y) = 0\}$.
5. $V_5 := \mathbb{R}[x]_{\leq 3} = \{ax^3+bx^2+cx+d \mid a,b,c,d \in \mathbb{R}\},\ U_5 := \{p \in \mathbb{R}[x]_{\leq 3} \mid b+d = 0,\ a+c = 0\}$.

Aufgabe 18.3 (Basen). Sei $\mathbb{F} := \mathbb{F}_5$ der Körper mit fünf Elementen.

1. Wie viele Elemente hat $\mathbb{F}^3$?
2. Wie viele verschiedene Basen hat $\mathbb{F}^3$?

Aufgabe 18.4 (Kodierungstheorie).

Parity Check Ist ein Daten-Wort $w = (w_1, w_2, \ldots, w_{19}) \in (\mathbb{F}_2)^{19}$ gegeben, so setzen wir:
$(v_1, v_2, \ldots, v_{19}, v_{20}) := (w_1, w_2, \ldots, w_{19}, p) \in (\mathbb{F}_2)^{20}$, wobei p die Parität des Wortes w ist, d.h.:

$$p = \begin{cases} 0, \text{ falls } w_1 + w_2 + \cdots + w_{19} \equiv 0 \mod 2, \\ 1, \text{ falls } w_1 + w_2 + \cdots + w_{19} \equiv 1 \mod 2. \end{cases}$$

Wir nehmen an, dass bei der Übermittlung eines Wortes $v \in (\mathbb{F}_2)^{20}$ höchstens ein Buchstabe fehlerhaft beim Empfänger ankommt. Zeigen Sie, dass

der Empfänger unter dieser Annahme erkennen kann, welche Wörter nicht korrekt übertragen wurden und welche er daher nochmals anfragen muss.

Hamming Code Für ein Daten-Wort $w = (w_1, w_2, w_3) \in (\mathbb{F}_2)^4$ werden beim Hamming-Code drei Parity-Check-Bits p_1, p_2, p_3 hinzugefügt, um einen Ein-Bit-Übertragungsfehler auch korrigieren zu können. Das übertragene Wort ist dann $v = (v_1, \ldots, v_7) = (p_1, p_2, w_1, p_3, w_2, w_3, w_4) \in (\mathbb{F}_2)^7$. Hierbei sind p_i, $i = 1, 2, 3$, Paritäten gewisser Teil-Wörter von v. Das Teil-Wort t_i enthält 2^{i-1} Bits von v ab dem 2^{i-1}-ten Bit, enthält die nächsten 2^{i-1} Bits nicht, enthält die nächsten 2^{i-1}-ten Bits aber wieder, usw. t_1 ist also das Teil-Wort $(v_1, v_3, v_5, v_7) = (p_1, w_1, w_2, w_4)$, $t_2 = (v_2, v_3, v_6, v_7)$, $t_3 = (v_4, v_5, v_6, v_7)$. Lassen wir den ersten Buchstaben von t_i weg, so erhalten wir ein neues Wort, das wir s_i nennen. p_i, $i = 1, 2, 3$, ist nun definiert als die Parität des Wortes s_i.
Wie lauten die Daten, die als $a = (0, 0, 1, 1, 0, 1, 0)$, $b = (1, 0, 1, 0, 1, 0, 1)$, $c = (1, 1, 1, 1, 1, 1, 0)$ empfangen wurden, unter der Annahme, dass maximal ein Bit falsch übertragen wurde?

Aufgabe 18.5 (Austauschbarkeit von Basiselementen).
Seien $v_1 = (1, 3, -2, 2)^t$, $v_2 = (-3, 2, -1, 1)^t$, $v_3 = (1, 3, -2, 3)^t$.

$$V := \langle v_1, v_2, v_3 \rangle \subset \mathbb{R}^4.$$

1. Ist es möglich, einen der Vektoren v_1, v_2, v_3 durch $v = (-5, -4, 3, -5)^t$ auszutauschen? Wenn ja, welchen?
2. Ist es möglich, einen der Vektoren v_1, v_2, v_3 durch $w = (-1, 2, -3, 4)^t$ auszutauschen? Wenn ja, welchen?
3. Finden Sie einen Vektor $v_4 \in \mathbb{R}^4$, der v_1, v_2, v_3 zu einer Basis des $\mathbb{R}^4$ ergänzt.

Aufgabe 18.6 (Basen von Untervektorräumen). Seien

$$U := \left\langle \begin{pmatrix} 2 \\ 5 \\ 9 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ -3 \end{pmatrix} \right\rangle \quad \text{und} \quad W := \left\langle \begin{pmatrix} -3 \\ 1 \\ 6 \end{pmatrix}, \begin{pmatrix} 5 \\ 3 \\ 0 \end{pmatrix} \right\rangle$$

Unterräume des $\mathbb{R}^3$. Bestimmen Sie eine Basis des Unterraums $U \cap W$.

19

Matrizen und Lineare Gleichungssysteme

19.1 Definition und Beispiele

Beispiel 19.1. Wir wollen das Gleichungssystem

$$\begin{aligned} x_1 + 2x_2 - 5x_3 &= 1 \\ 2x_1 + 3x_2 - 7x_3 &= 3 \\ 3x_1 + 4x_2 - 8x_3 &= 13 \end{aligned}$$

(systematisch) lösen.

Idee: Da der Koeffizient von x_1 in der ersten Gleichung $\neq 0$, können wir diese Gleichung verwenden, um x_1 aus den beiden anderen Gleichungen zu entfernen.

$$\begin{aligned} x_1 + 2x_2 - 5x_3 &= 1 \\ -x_2 + 3x_3 &= 1 \\ -2x_2 + 7x_3 &= 10 \end{aligned}$$

ist ein äquivalentes Gleichungssystem.

Anschließend lösen wir das kleinere System (d.h. die unteren beiden Gleichungen) mit der selben Idee. Zunächst:

$$\begin{aligned} x_1 + 2x_2 - 5x_3 &= 1, \\ -x_2 + 3x_3 &= 1, \\ x_3 &= 8, \end{aligned}$$

also $x_3 = 8$. Dies in die vorletzte Gleichung eingesetzt ergibt:

$$-x_2 + 3 \cdot 8 = 1, \quad \text{also } x_2 = 23.$$

Schließlich finden wir:

$$x_1 + 2 \cdot 23 - 5 \cdot 8 = 1 \Rightarrow x_1 = -5.$$

$\Rightarrow x = (-5, 23, 8)^t$ ist der eindeutig bestimmte Lösungsvektor.

Bemerkung 19.2. 1. Auch geometrisch ist es einzusehen, dass es genau eine Lösung gibt: Jede der drei Gleichungen definiert eine Ebene = Hyperebene in $\mathbb{R}^3$ und drei Ebenen schneiden sich in der Regel in einem Punkt.

2. Eigentlich ist es überflüssig, die Variablen x_1, x_2, x_3 jeweils hinzuschreiben, denn allein aus der Position des Koeffizienten können wir schon schließen, welche Variable dazugehört.

Definition 19.3. *1. Sei K ein Körper. Eine $m \times n$ **Matrix** mit Einträgen in K ist eine Tabelle*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \in K^{m \times n}$$

von Körperelementen $a_{ij} \in K$. m ist die Anzahl der Zeilen und n die Anzahl der Spalten von A. Wir schreiben auch

$$A = (a_{ij})_{i=1,\dots,m;j=1,\dots,n} = (a_{ij}) \in K^{m \times n}.$$

2. Es seien $A = (a_{ij}) \in K^{m \times n}$ und $B = (b_{jk}) \in K^{n \times r}$ zwei Matrizen, so dass die Spaltenzahl von A mit der Zeilenzahl von B übereinstimmt. Dann ist das Produkt

$$C = A \cdot B = (c_{ik}) \in K^{m \times r}$$

durch die Formel $c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}$ erklärt.

Beispiel 19.4. $A = \begin{pmatrix} 2 & 3 & 5 \\ 1 & 4 & 6 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 1 \\ -1 & 2 \\ 1 & 3 \end{pmatrix}$. Also, $A \cdot B = \begin{pmatrix} 4 & 23 \\ 3 & 27 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$ $(1 \cdot 1 + 2 \cdot 4 + 3 \cdot 6 = 27)$. In diesem speziellen Fall ist auch das Produkt $B \cdot A$ erklärt, da (zufälligerweise) $m = 2 = r$:

$$B \cdot A = \begin{pmatrix} 1 & 1 \\ -1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 2 & 3 & 5 \\ 1 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 3 & 7 & 11 \\ 0 & 5 & 7 \\ 5 & 15 & 23 \end{pmatrix} \in \mathbb{R}^{3 \times 3}.$$

Insbesondere ist $A \cdot B \neq B \cdot A$.

Spaltenvektoren wie

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in K^{n\times 1}$$

können wir mit $n \times 1$ Matrizen identifizieren.

Das allgemeine Gleichungsystem

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\qquad &\vdots\,\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

können wir knapper schreiben:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

oder noch kürzer

$$A \cdot x = b,$$

wobei

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \in K^{m\times n}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in K^m = K^{m\times 1}$$

und $x = (x_1, \dots, x_n)^t \in K^{n\times 1}$ der Vektor der Unbestimmten ist.

19.2 Der Gaußalgorithmus zum Lösen linearer Gleichungssysteme

Um ein Gleichungssystem explizit zu lösen, dürfen wir es äquivalent umformen, z.B. einer Gleichung eine andere dazu addieren. Dies in Termen von Matrizen führt auf den Begriff der Zeilenoperation. Der Gaußalgorithmus gibt ein Verfahren an, mit Hilfe solcher Zeilenoperationen Gleichungssysteme zu lösen.

Definition 19.5. *Sei $A = (a_{ij}) \in K^{m\times n}$ eine $m \times n$-Matrix und seien $a_i \in K^n$, $i = 1, \dots, m$, die Zeilenvektoren von A. Eine **elementare Zeilenoperation** (oder **elementare Zeilenumformung**) ist eine der folgenden Operationen:*

I) Multiplikation einer Zeile mit einem Skalar $\lambda \neq 0 \in K$:

$$A = \begin{pmatrix} \vdots \\ a_i \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} \vdots \\ \lambda a_i \\ \vdots \end{pmatrix} =: A_I.$$

II) Addieren der i-ten Zeile zur j-ten Zeile:

$$A = \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_i + a_j \\ \vdots \end{pmatrix} =: A_{II}.$$

III) Addieren des λ-fachen ($\lambda \in K^*$) der i-ten Zeile zur j-ten Zeile:

$$A = \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ \lambda a_i + a_j \\ \vdots \end{pmatrix} =: A_{III}.$$

IV) Vertauschen der i-ten Zeile mit der j-ten Zeile:

$$A = \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ a_i \\ \vdots \end{pmatrix} =: A_{IV}.$$

Bemerkung 19.6. Die Operation vom Typ III und VI kann man auch durch wiederholtes Anwenden von I und II erhalten.

Beweis. III)

$$A = \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} \overset{\text{I}}{\mapsto} \begin{pmatrix} \vdots \\ \lambda a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} \overset{\text{II}}{\mapsto} \begin{pmatrix} \vdots \\ \lambda a_i \\ \vdots \\ \lambda a_i + a_j \\ \vdots \end{pmatrix} \overset{\text{I}}{\mapsto} \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ \lambda a_i + a_j \\ \vdots \end{pmatrix}.$$

IV)

$$A = \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} \overset{\text{III}}{\mapsto} \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j - a_i \\ \vdots \end{pmatrix} \overset{\text{II}}{\mapsto} \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ a_j - a_i \\ \vdots \end{pmatrix} \overset{\text{III}}{\mapsto} \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ -a_i \\ \vdots \end{pmatrix} \overset{\text{I}}{\mapsto} \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ a_i \\ \vdots \end{pmatrix}.$$

□

Definition 19.7. *Eine Matrix $A = (a_{ij}) \in K^{m\times n}$ hat **Zeilenstufenform**, wenn sie folgende Form hat*[1]*:*

$$\begin{pmatrix} \underline{|a_{1j_1}} & * & & & * \\ & \underline{|a_{2j_2}} & * & & \\ & & \underline{|a_{3j_3}} & * & \\ & & & \ddots & \\ & & & & \underline{|a_{rj_r}} \\ 0 & & & & \end{pmatrix}, \quad a_{ij_i} \neq 0.$$

Genauer: Falls $\exists r$ mit $0 \leq r \leq m$ und Indices $1 \leq j_1 < \cdots < j_r \leq n$, so dass:

1. *$a_{ij} = 0$, falls $1 \leq i \leq r$ und $j < j_i$ oder $i > r$.*
2. *$a_{ij_i} \neq 0$ für $i = 1, \ldots, r$.*

Satz 19.8 (Gaußalgorithmus). *Sei $A \in K^{m\times n}$. Dann lässt sich A durch eine Folge von elementaren Zeilenumformungen in eine Matrix $\widetilde{A}$, die in Zeilenstufenform ist, umformen.*

Beweis. Induktion nach m. Für $m = 1$ ist die Aussage trivial richtig. Sei also $m \geq 2$. Wir betrachten die erste Spalte von A:

$$a_1 = \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix}.$$

[1] Hierbei steht $*$ für Einträge, die nicht genauer spezifiziert sind (sie dürfen auch 0 oder gar nicht vorhanden sein). Die 0 in der linken unteren Ecke repräsentiert die Tatsache, dass alle Einträge, die links oder unterhalb der Linien sind, 0 sind. Solche Notationen werden oft bei Matrizen verwendet.

1. Fall: Ist $a_{11} \neq 0$, dann setzen wir $j_1 = 1$ und addieren jeweils das $(-\frac{a_{i1}}{a_{11}})$-fache der ersten Zeile zur i-ten Zeile. Anschließend hat A die Gestalt:

$$\left(\begin{array}{c|ccc} \underline{a_{11}} & a_{12} & & * \\ 0 & a'_{22} & \cdots & a'_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & a'_{m2} & \cdots & a'_{m2} \end{array}\right) = \left(\begin{array}{c|c} \underline{a_{11}} & * \\ 0 & A' \end{array}\right)$$

und wir fahren induktiv mit der Matrix A' fort, die nämlich weniger Spalten als A hat.

2. Fall: $a_{11} = 0$, aber $a_1 \neq 0$. Ist etwa $a_{i1} \neq 0$, so vertauschen wir die i-te und die 1-te Zeile

$$\begin{pmatrix} a_{11} & a_{12} & \cdots \\ \vdots & \vdots & \\ a_{i1} & a_{i2} & \cdots \\ \vdots & \vdots & \end{pmatrix} \mapsto \begin{pmatrix} a_{i1} & a_{i2} & \cdots \\ \vdots & \vdots & \\ a_{11} & a_{12} & \cdots \\ \vdots & \vdots & \end{pmatrix}$$

und fahren wie im 1. Fall fort.

3. Fall: $a_1 = 0$. In diesem Fall ist $j_1 > 1$ und wir fahren mit der Teilmatrix

$$\left(\begin{array}{c|c} 0 & \\ \vdots & A' \\ 0 & \end{array}\right), \quad A' \in K^{m\times(n-1)}$$

genauso fort, bis die erste Spalte keine Nullspalte ist. Dann trifft auf die neue Matrix Fall 1 oder 2 zu.

□

Beispiel 19.9.

$$A = \begin{pmatrix} 1 & 3 & 4 \\ 0 & 0 & 2 \\ 2 & 0 & 7 \\ 1 & 4 & 5 \end{pmatrix} \overset{\text{IIIs}}{\mapsto} \left(\begin{array}{c|cc} \underline{1} & 3 & 4 \\ 0 & 0 & 2 \\ 0 & -6 & -1 \\ 0 & 1 & 1 \end{array}\right) \overset{\text{IV}}{\mapsto} \left(\begin{array}{c|cc} \underline{1} & 3 & 4 \\ 0 & \underline{1} & 1 \\ 0 & -6 & -1 \\ 0 & 0 & 2 \end{array}\right) \overset{\text{III}}{\mapsto} \left(\begin{array}{c|cc} \underline{1} & 3 & 4 \\ 0 & \underline{1} & 1 \\ 0 & 0 & \underline{5} \\ 0 & 0 & 2 \end{array}\right) \overset{\text{III}}{\mapsto} \left(\begin{array}{c|cc} \underline{1} & 3 & 4 \\ 0 & \underline{1} & 1 \\ 0 & 0 & \underline{5} \\ 0 & 0 & 0 \end{array}\right) = \widetilde{A}.$$

$$
\begin{aligned}
B &= \left(\begin{array}{cccc} 4 & 2 & 7 & 5 \\ 3 & 0 & 6 & 4 \\ 1 & 0 & 2 & 4 \end{array}\right) \\
&\overset{\text{IVs}}{\mapsto} \left(\begin{array}{c|ccc} \underline{4} & 2 & 7 & 5 \\ 0 & -3/2 & 6-21/4 & 4-15/4 \\ 0 & -1/2 & 2-7/4 & 1-5/4 \end{array}\right) = \left(\begin{array}{c|ccc} \underline{4} & 2 & 7 & 5 \\ 0 & -3/2 & 3/4 & 1/4 \\ 0 & -1/2 & 1/4 & -1/4 \end{array}\right) \\
&\overset{(3)\mapsto(3)-1/3\cdot(2)}{\mapsto} \left(\begin{array}{c|cc|c} \underline{4} & 2 & 7 & 5 \\ 0 & \underline{-3/2} & 3/4 & 1/4 \\ 0 & 0 & 0 & -1/4-1/12 \end{array}\right) = \left(\begin{array}{c|cc|c} \underline{4} & 2 & 7 & 5 \\ 0 & \underline{-3/2} & 3/4 & 1/4 \\ 0 & 0 & 0 & -1/3 \end{array}\right) \\
&= \widetilde{B}.
\end{aligned}
$$

Es leuchtet unmittelbar ein, dass man, wenn man solche Rechnungen per Hand durchführen möchte, sinnvollerweise versuchen wird, Brüche und zu kleine Zahlen zu vermeiden. Dies ist manchmal möglich, indem man Zeilen– oder Spaltenvertauschungen vornimmt. Im nächsten Semester werden wir auf die damit zusammenhängende Thematik (genannt **Pivotierung** oder **Pivotwahl**) noch genauer eingehen, denn auch wenn man auf einem Computer mit Fließkommazahlen arbeitet, möchte man nummerische Fehler möglichst klein halten und beispielsweise vom Betrag her sehr kleine Zahlen vermeiden. Auch dies ist mit geeigneten Vertauschungen oft möglich.

Satz 19.10. *Sei $A \in K^{m\times n}$, $b \in K^m$. Das Gleichungssystem*

$$Ax = b$$

hat eine Lösung $x \in K^n$ genau dann, wenn die erweiterte Matrix

$$(A \,\vdots\, b) \in K^{m\times(n+1)}$$

eine Zeilenstufenform hat, bei der Spalte $n+1$ keine Stufe ist.

Beweis. Wir bringen die erweiterte Matrix $(A \,\vdots\, b)$ in Zeilenstufenform durch elemetare Zeilenumformungen:

$$
(\widetilde{A} \;\vdots\; \widetilde{b}) = \left(\begin{array}{cccc|c} \underline{*} & & & & * \\ & \underline{*} & & & \\ & & \underline{*} & & \\ & & & \ddots & \vdots \\ & & & & \underline{*} \\ 0 & & & & \end{array}\right) \begin{array}{l} \\ \\ \\ \\ \widetilde{b}_1 \\ \\ \\ \\ \widetilde{b}_r \end{array},
$$

mit r Stufen. Ist die letzte Stufe bei Spalte $n+1$, dann hat das Gleichungssystem keine Lösung, da

$$0x_1 + \cdots + 0x_n = 0 \neq \widetilde{b}_r.$$

Andernfalls ist die Gestalt

$$\left(\begin{array}{ccccccc|c}
\underline{\widetilde{a}_{1j_1}} & * & & & & & & * \\
0 & \underline{\widetilde{a}_{2j_2}} & * & & & & & \\
 & 0 & \underline{\widetilde{a}_{3j_3}} & * & & & & \\
 & & 0 & \ddots & & & & \\
 & & & & \underline{\widetilde{a}_{rj_r}} & * & * & \\
0 & & & & 0 & 0 & &
\end{array}\right)$$

mit $j_r < n+1$ und $\widetilde{b}_{r+1} = \cdots = \widetilde{b}_m = 0$. Man kann dann die x_j mit $j \notin \{j_1, \ldots, j_r\}$ beliebig wählen und dann $x_{j_r}, x_{j_{r-1}}, \ldots, x_{j_1}$ sukzessive aus den Gleichungen

$$\begin{aligned}
\widetilde{a}_{rj_r}x_{j_r} + \widetilde{a}_{rj_r+1}x_{j_r+1} + \cdots + \widetilde{a}_{rn}x_n &= \widetilde{b}_r \\
\vdots &= \vdots \\
\sum_{j=j_k}^{n} \widetilde{a}_{1j}x_j &= \widetilde{b}_k \\
\vdots &= \vdots \\
\sum_{j=j_1}^{n} \widetilde{a}_{1j}x_j &= \widetilde{b}_1
\end{aligned}$$

bestimmen:

$$\begin{aligned}
x_{j_r} &= \frac{1}{\widetilde{a}_{rj_r}}\left(\widetilde{b}_r - \sum_{j=j_r+1}^{n} \widetilde{a}_{rj}x_j\right) \\
&\vdots \\
x_{j_k} &= \frac{1}{\widetilde{a}_{kj_k}}\left(\widetilde{b}_k - \sum_{j=j_k+1}^{n} \widetilde{a}_{kj}x_j\right) \\
&\vdots
\end{aligned}$$

□

Beispiel 19.11. Wir suchen den Lösungsvektor $x \in \mathbb{R}^3$ des Gleichungssystems:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}.$$

Die erweiterte Matrix ist:

$$\left(\begin{array}{ccc:c} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{array}\right).$$

Der Gaußalgorithmus liefert:

$$\left(\begin{array}{c|ccc:c} 1 & 2 & 3 & 4 \\ \hline 0 & -1 & -2 & -3 \end{array}\right).$$

Wir können $x_3 = t \in \mathbb{R}$ beliebig wählen. Dann ist $-x_2 - 2t = -3$, also $x_2 = 3 - 2t$. Eingesetzt in die erste Zeile liefert das: $x_1 + 2 \cdot (3 - 2t) + 3t = 4$, also $x_1 = -2 + t$.

Wir finden also die Lösungsmenge:

$$L = \left\{ \begin{pmatrix} t-2 \\ 3-2t \\ t \end{pmatrix} \,\middle|\, t \in \mathbb{R} \right\} \subset \mathbb{R}^3.$$

Diese ist eine Gerade:

$$L = \left\{ \begin{pmatrix} -2 \\ 3 \\ 0 \end{pmatrix} + t \cdot \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

mit Aufpunkt $\begin{pmatrix} -2 \\ 3 \\ 0 \end{pmatrix}$ und Richtungsvektor $\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$.

19.3 Aufwand des Gaußalgorithmus (im Fall $n = m$)

Um den Aufwand des Gaußalgorithmus im Fall $n = m$ zu berechnen, beginnen wir mit der Matrix:

$$(A, b) = \left(\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & b_n \end{array}\right).$$

1. Schritt: Wir bringen (A, b) in Zeilenstufenform:

$$(\tilde{A}, \tilde{b}) = \begin{pmatrix} * & * & * & \dots & * & * \\ 0 & * & * & \dots & * & * \\ 0 & 0 & * & \dots & * & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & * & * \end{pmatrix},$$

wobei wir annehmen, dass die Matrix A genau n Stufen hat.

2. Schritt: Rückwärts einsetzen.

Wir betrachten die Anzahl der Multiplikationen und Additionen in K, die dabei durchzuführen sind:

Additionen im 1. Schritt: Die erste Spalte in Zeilenstufenform zu bringen benötigt

$$\underbrace{(n-1)}_{\text{Zeilen}} \cdot \underbrace{(n+1)}_{\text{Spalten}}$$

Additionen. Für alle Spalten sind dies insgesamt

$$\sum_{k=1}^{n} (k+1)(k-1) \approx \sum_{k=1}^{n} k^2 \approx \frac{n^3}{3} \in O(n^3)$$

Additionen, d.h. im Wesentlichen ein konstantes Vielfaches von n^3.

Multiplikationen im 1. Schritt: Ähnlich: $\sum_{k=1}^{n} k^2 \in O(n^3)$.

Operationen im 2. Schritt: Man zeigt analog, dass dies in $O(n^2)$ liegt.

Insgesamt sind also $O(n^3)$ Körperoperationen nötig.

Bemerkung 19.12. Es ist offen, was asymptotisch optimal ist. Arbeiten von Strassen zeigen: Es kommt auf den Aufwand $A \cdot B$ aus $A, B \in K^{n\times n}$ an. Die naive Betrachtungsweise anhand der Definition liefert: $O(n^3)$, nämlich n^3 Multiplikationen und $n^2(n-1)$ Additionen.

Satz 19.13 (1969, Strassen). *Seien $A, B \in K^{n\times n}$. Dann kann man $A \cdot B$ mit $O(n^{\log_2 7})$ (man bemerke: $O(n^{\log_2 8}) = O(n^3)$) Operationen ausrechnen.*

Es gibt neuere, asymptotisch noch bessere Algorithmen, siehe dazu Bemerkung 20.29. Zunehmend wichtig werden Algorithmen, bei denen mehrere Operationen parallel ablaufen können, weil moderne Rechner mehrere, teils sogar sehr viele, Operationen gleichzeitig ablaufen lassen können, u.a. durch Ausnutzung der Graphikkartenchips. Solche Herangehensweisen sind in der obigen Überlegung nicht berücksichtigt.

Aufgaben

Aufgabe 19.1 (Multiplikation von Matrizen). Seien

$$A = \begin{pmatrix} 1 & -2 & 3 & -4 \\ -3 & 2 & -1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & -2 \\ 1 & -1 \\ 2 & 0 \\ 3 & 1 \end{pmatrix}.$$

Berechnen Sie AB und BA. Können Sie $AB \neq BA$ auch ohne Rechnen einsehen?

Aufgabe 19.2 (Matrizen und Kommutativität).

1. Bestimmen Sie alle Matrizen der Form
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad a, b, c, d \in \mathbb{R},$$
für die gilt:
$$A \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} A.$$
2. Man sagt, eine $n \times n$ Matrix M **kommutiert** mit einer $n \times n$ Matrix N, wenn $MN = NM$. Bestimmen Sie alle reellen 2×2 Matrizen, die mit jeder reellen 2×2 Matrix kommutieren.

Aufgabe 19.3 (Gauß-Algorithmus). Lösen Sie das folgende lineare Gleichungssystem mit dem Gauß-Algorithmus:
$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 25 \\ 125 \end{pmatrix}.$$

Aufgabe 19.4 (Schnelles Berechnen des Matrixprodukts). Zeigen Sie (die Aussagen in () müssen dabei nicht bewiesen werden!):

1. Der folgende Algorithmus (von Strassen (1969)) berechnet das Matrixprodukt, er benötigt für $n = 2$ insgesamt 7 (< 8) Multiplikationen (und 15 Additionen, dabei Zwischenergebnisse benutzen!).
2. Für $n = 2^k, k \in \mathbb{N}$, benötigt er $O(n^{\log_2 7})$ ($< O(n^3)$) Multiplikationen (bzw. Mult. und Add.).

Eingabe: $A, B \in \mathbb{R}^{n \times n}$ und $n = 2^k$ für ein $k \in \mathbb{N}$.
Ausgabe: Das Produkt $AB \in \mathbb{R}^{n \times n}$.

1. Falls $n = 1$, dann schreibe $A = (a)$, $B = (b)$. **Ausgabe:** (ab).
2. Sonst schreiben wir: $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$, mit $A_{ij}, B_{ij} \in \mathbb{R}^{(n/2) \times (n/2)}$.
3. Wir setzen nun:
$$\begin{aligned} P_1 &= A_{11}B_{11}, & P_5 &= (A_{21} + A_{22}) \cdot (B_{12} - B_{11}), \\ P_2 &= A_{12}B_{21}, & P_6 &= ((A_{21} + A_{22}) - A_{11}) \cdot (B_{22} - (B_{12} - B_{11})), \\ P_3 &= (A_{12} - ((A_{21} + A_{22}) - A_{11})) \cdot B_{22}, & P_7 &= (A_{11} - A_{21}) \cdot (B_{22} - B_{12}). \\ P_4 &= A_{22} \cdot ((B_{22} - (B_{12} - B_{11})) - B_{21}), \end{aligned}$$
4. **Ausgabe:** $\begin{pmatrix} P_1 + P_2 & ((P_1 + P_6) + P_5) + P_3 \\ ((P_1 + P_6) + P_7) - P_4 & ((P_1 + P_6) + P_7) + P_5 \end{pmatrix}$.

20

Lineare Abbildungen

Matrizen, die wir im vorigen Abschnitt betrachtet haben, beschreiben, wie wir sehen werden, auf natürliche Weise sogenannte lineare Abbildungen. Um Lösungen von Gleichungssystemen besser strukturell verstehen zu können, betrachten wir daher nun lineare Abbildungen genauer. Mit deren Hilfe werden wir gewisse Räume, genannt Kern und Bild, einführen können, die essentiell für das Verständnis von linearen Abbildungen und damit auch für das Lösen linearer Gleichungssysteme sind.

20.1 Grundlegende Definitionen

Definition 20.1. *Es seien V, W zwei K-Vektorräume. Eine (K-) **lineare Abbildung** (oder **Vektorraumhomomorphismus**) von V nach W ist eine Abbildung*

$$f\colon\ V \to W,$$

die Folgendes erfüllt:

1. $f(v_1 + v_2) = f(v_1) + f(v_2) \quad \forall\, v_1, v_2 \in V \quad$ *und*
2. $f(\lambda v) = \lambda f(v) \quad \forall \lambda \in K, \forall\, v \in V.$

Die Menge aller Vektorraumhomomorphismen von V nach W bezeichnen wir mit $\mathrm{Hom}(V, W)$ *bzw. zur Verdeutlichung des Körpers K mit* $\mathrm{Hom}_K(V, W)$.

Beispiel 20.2.

1. Seien $V = K^n$, W ein weiterer K-VR und $\{w_1, \ldots, w_n\} = \mathcal{A}$ eine Familie von Vektoren von W. Dann ist

$$\varphi_{\mathcal{A}}\colon K^n \to W, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \sum_{l=1}^{n} x_i w_i$$

eine K-lineare Abbildung.

2. $V = K^n$, $W = K^m$, $A = (a_{ij}) \in K^{m\times n}$. Die Abbildung

$$\varphi_A\colon K^n \xrightarrow{A} K^m, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto Ax$$

ist K-linear.

3. Die Translation eines Vektors $b \in \mathbb{R}^n, b \neq 0$,

$$\mathbb{R}^n \mapsto \mathbb{R}^n, \quad x \mapsto x + b$$

ist nicht linear, wie die folgende Bemerkung zeigt.

Insbesondere ist also beispielsweise die Funktion $f\colon \mathbb{R} \to \mathbb{R},\ x \mapsto 3x + 1$ keine lineare Abbildung, aber $g\colon \mathbb{R} \to \mathbb{R},\ x \mapsto 7x$ schon (Abb. 20.1).

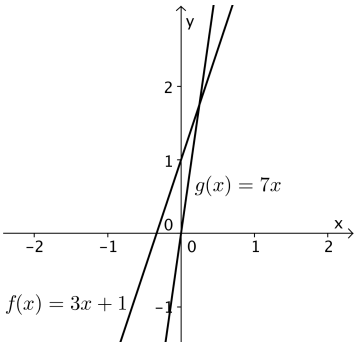


Abbildung 20.1. Die Funktionen $f(x) = 3x + 1$ und $g(x) = 7x$. Dabei ist f keine lineare Abbildung, z.B. weil $f(0) \neq 0$ ist, g aber schon.

Bemerkung 20.3. Ist $f\colon V \to W$ eine lineare Abbildung zwischen zwei K-Vektorräumen. Dann gilt: $f(0_V) = 0_W$ (kurz: $f(0) = 0$).

Beweis. Es gilt $0_V = 0_K \cdot 0_V$, also: $f(0_V) = f(0_K \cdot 0_V) = 0_K \cdot f(0_V) = 0_W$. □

Die in der folgenden Bemerkung verwendeten Begriffe injektiv, surjektiv und bijektiv werden ausführlich in Abschnitt 2.8.2 beschrieben. Knapp gesagt

ist eine Abbildung injektiv, wenn jedes Element der Bildmenge höchstens ein Urbild hat, surjektiv, wenn jedes solche mindestens ein Urbild hat, und bijektiv, wenn jedes solche genau ein Urbild hat.

Bemerkung/Definition 20.4. Einen injektiven Vektorraumhomomorphismus $f\colon V \to W$ nennen wir einfach **Monomorphismus**, einen surjektiven nennen wir **Epimorphismus**. Ein bijektiver Vektorraumhomomorphismus $f\colon V \to W$ heißt **Isomorphismus**; V und W heißen dann **isomorph** ($V \cong W$). Ist f ein Isomorphismus, dann ist die Umkehrabbildung $f^{-1} : W \to V$ ebenfalls ein Isomorphismus.

Beweis. Zu zeigen ist: f^{-1} ist K-linear. Sind $w_1, w_2 \in W$ und $v_j = f^{-1}(w_j)$, $j = 1, 2$, also $w_j = f(v_j)$, dann gilt:

$$w_1 + w_2 = f(v_1) + f(v_2) = f(v_1 + v_2),$$

also: $f^{-1}(w_1 + w_2) = f^{-1}(f(v_1 + v_2)) = v_1 + v_2 = f^{-1}(w_1) + f^{-1}(w_2)$.
Sind außerdem $w \in W$, $\lambda \in K$ und $v = f^{-1}(w)$, also $w = f(v)$, dann gilt:

$$\lambda \cdot w = \lambda \cdot f(v) = f(\lambda \cdot v),$$

also: $f^{-1}(\lambda \cdot w) = f^{-1}(f(\lambda \cdot v)) = \lambda \cdot v = \lambda \cdot f^{-1}(w)$. □

20.2 Kern und Bild

Satz/Definition 20.5. *Sei $f\colon V \to W$ eine lineare Abbildung zwischen zwei K-Vektorräumen. Dann ist der **Kern** von f*

$$\operatorname{Ker} f = \{v \in V \mid f(v) = 0\} \subseteq V$$

*ein Untervektorraum von V und das **Bild** von f*

$$\operatorname{Bild} f = f(V) = \{w \in W \mid \exists\, v \in V : f(v) = w\} \subseteq W$$

(auch manchmal $\operatorname{im}(f)$ *geschrieben, von engl.* image*) ein Untervektorraum von W.*

Beweis. Zur Abgeschlossenheit des Kerns: Seien $v_1, v_2 \in \operatorname{Ker} f \Rightarrow f(v_1) = 0 = f(v_2) \Rightarrow f(v_1 + v_2) = f(v_1) + f(v_2) = 0 + 0 = 0$, da f linear ist $\Rightarrow v_1 + v_2 \in \operatorname{Ker} f$. Seien nun $v \in \operatorname{Ker} f$, $\lambda \in K$, $f(\lambda v) = \lambda f(v) = \lambda \cdot 0 = 0 \Rightarrow \lambda v \in \operatorname{Ker} f$.

Zum Bild: Seien $w_1, w_2 \in \operatorname{Bild} f$, etwa $w_j = f(v_j)$. Dann ist $w_1 + w_2 = f(v_1) + f(v_2) = f(v_1 + v_2) \in \operatorname{Bild}(f)$ wegen der Linearität von f. Seien nun wieder $w = f(v) \in \operatorname{Bild} f$, $\lambda \in K$, dann: $\lambda w = \lambda f(v) = f(\lambda v) \in \operatorname{Bild}(f)$. □

Satz 20.6. *Sei $f\colon V \to W$ ein Vektorraumhomomorphismus.*
Die Abbildung f ist ein Monomorphismus genau dann, wenn

$$\operatorname{Ker} f = 0 = \{0\} \subseteq V.$$

Beweis. Angenommen $\operatorname{Ker} f = 0, v_1, v_2 \in V$ und $f(v_1) = f(v_2)$. Wir müssen zeigen, dass dann $v_1 = v_2$ gilt:

$$\begin{aligned}
&\Rightarrow f(v_1 - v_2) = f(v_1) - f(v_2) = 0\\
&\Rightarrow v_1 - v_2 \in \operatorname{Ker} f = \{0\}\\
&\Rightarrow v_1 - v_2 = 0\\
&\Rightarrow v_1 = v_2
\end{aligned}$$

Also f ist injektiv.

Die umgekehrte Richtung ist klar, da bei einer injektiven Abbildung nur die 0 auf 0 abgebildet wird. □

20.3 Vorgabe der Bilder einer Basis

Satz 20.7. *Sei V ein K-Vektorraum endlicher Dimension und $\mathcal{B} = \{v_1, \ldots, v_n\}$ eine Basis.*

1. *Dann ist die Abbildung*

$$\varphi_{\mathcal{B}}\colon K^n \to V, \quad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \sum_{i=1}^{n} x_i v_i$$

   ein Isomorphismus.

2. *Ist W ein weiterer K-Vektorraum und $A = \{w_1, \ldots, w_n\}$ eine beliebige Familie von Vektoren aus W, dann ist die Abbildung*

$$\varphi_{\mathcal{A}}^{\mathcal{B}}\colon V \to W,\ v = \sum_{i=1}^{n} \lambda_i v_i \mapsto \sum_{i=1}^{n} \lambda_i w_i$$

   linear.

Beweis. Surjektivität ist klar nach Definition einer Basis. Wegen der Eindeutigkeit der Darstellung in einer Basis ist $(0, \ldots, 0)^{\mathrm{t}}$ der einzige Vektor, der auf $0 \in V$ abgebildet wird. Die Linearität beider Abbildungen ist einfach nachzuweisen und wird daher hier nicht vorgeführt. □

Beispiel 20.8. Seien $V = \mathbb{R}[t]_{\leq d}, \mathcal{B} = \{1, t, t^2, \ldots, t^d\}$. Dann ist

$$\mathbb{R}^{d+1} \to \mathbb{R}[t]_{\leq d}, \quad \begin{pmatrix} a_0 \\ \vdots \\ a_d \end{pmatrix} \mapsto \sum_{i=0}^{n} a_i t^i$$

ein Isomorphismus. Ein anderer ist:

$$\mathbb{R}^{d+1} \to \mathbb{R}[t]_{\leq d}, \quad \begin{pmatrix} a_0 \\ \vdots \\ a_d \end{pmatrix} \mapsto \sum_{i=0}^{d} a_i (t-\alpha)^i, \quad \alpha \in \mathbb{R}.$$

Bemerkung 20.9.

1. Bezeichnet $\varepsilon = \varepsilon_n = \{e_1, \ldots, e_n\} \subseteq K^n$ die Standardbasis, dann ist offenbar:
$$\varphi_{\mathcal{B}} = \varphi_{\mathcal{B}}^{\varepsilon}.$$
2. *Vorgabe der Bilder einer Basis:* Die zweite Aussage von Satz 20.7 besagt, dass die Bilder einer Basis unter einer linearen Abbildung beliebig vorgeschrieben werden können, etwa $v_i \mapsto w_i$, $i = 1, \ldots, n$, und dass damit die lineare Abbildung festgelegt ist. Denn jeder Vektor $v \in V$ entsteht als Linearkombination der v_i.
3. *Isomorphie gleich-dimensionaler Vektorräume:* Es gilt nach der ersten Aussage des Satzes 20.7:
$$\dim_K V = n \Rightarrow V \cong K^n,$$
da V eine Basis besitzt. Alle n-dimensionalen K-Vektorräume sind also isomorph.

20.4 Matrixdarstellungen einer linearen Abbildung

Definition 20.10. *Seien V, W zwei endlich–dimensionale K-Vektorräume und $\mathcal{A} = \{v_1, \ldots, v_n\}$ bzw. $\mathcal{B} = \{w_1, \ldots, w_m\}$ Basen. Ist $f\colon V \to W$ eine lineare Abbildung, dann betrachten wir die Skalare $a_{ij} \in K$ definiert durch*

$$f(v_j) = a_{1j}w_1 + a_{2j}w_2 + \cdots + a_{mj}w_m = \sum_{i=1}^{m} a_{ij} w_i \in W.$$

(Dies basiert auf der Eindeutigkeit der Darstellung eines Vektors als Linearkombination einer Basis.) Dann heißt die Matrix

$$A = (a_{ij}) = M_{\mathcal{B}}^{\mathcal{A}}(f) \in K^{m \times n}$$

*die **Matrixdarstellung** von f bezüglich der Basen $\mathcal{A}$ und $\mathcal{B}$.*

Beispiel 20.11. Sei $V = \mathbb{R}[x]_{\leq d}$, und seien $\alpha_0, \ldots, \alpha_d \in \mathbb{R}$. Die Abbildung

$$\varphi\colon V \to \mathbb{R}^{d+1}, \quad p \mapsto \begin{pmatrix} p(\alpha_0) \\ \vdots \\ p(\alpha_d) \end{pmatrix}$$

ist $\mathbb{R}$-linear, weil $(p+q)(\alpha_i) = p(\alpha_i) + q(\alpha_i)$ und $(\lambda p)(\alpha_i) = \lambda p(\alpha_i)$. Um die Darstellung von φ bezüglich der Basen $\mathcal{A} = \{1, t, \ldots, t^d\}$ und der Standardbasis $\varepsilon = \{e_1, \ldots, e_{d+1}\} \subseteq \mathbb{R}^{d+1}$ zu berechnen, betrachten wir für $i = 0, 1, 2, \ldots, d$ die Bilder der Basisvektoren aus A und stellen diese als Linearkombination der Basis ε des Zielvektorraumes dar:

$$\varphi(t^i) = \begin{pmatrix} \alpha_0^i \\ \vdots \\ \alpha_d^i \end{pmatrix} = \alpha_0^i \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \alpha_1^i \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \alpha_d^i \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Daher hat φ die Matrixdarstellung

$$\begin{pmatrix} 1 & \alpha_0 & \alpha_0^2 & \ldots & \alpha_0^d \\ 1 & \alpha_1 & \ldots & \ldots & \alpha_1^d \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \alpha_d & \cdots & \cdots & \alpha_d^d \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Merkregel 20.12. *$A \in K^{m\times n}$ mit Spalten $A = (a_1, \ldots, a_n)$. Dann ist die j-te Spalte*

$$a_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} \in K^m$$

das Bild des j-ten Einheitsvektors von K^n.

Beispiel 20.13. Seien $V = \mathbb{R}[t]_{\leq d}$, $W = \mathbb{R}[t]_{\leq d-1}$ mit Basen $\mathcal{A} = \{1, t, \ldots, t^d\}$, $\mathcal{B} = \{1, t, \ldots, t^{d-1}\}$. Sei

$$\varphi\colon V \to W, \ \ p \mapsto p'$$

die Abbildung, die ein Polynom auf seine Ableitung abbildet. Dann gilt:

$$\varphi(t^k) = (t^k)' = k \cdot t^{k-1}.$$

Also:

$$M_{\mathcal{B}}^{A}(\varphi) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d \end{pmatrix} \in \mathbb{R}^{d\times(d+1)}.$$

Satz 20.14. *Sei $f\colon V \to W$ eine lineare Abbildung zwischen zwei Vektorräumen V und W. Ist $A = (a_{ij}) = M_{\mathcal{B}}^{\mathcal{A}}(f) \in K^{m\times n}$ die Matrixdarstellung bzgl. der Basen $\mathcal{A} = \{v_1, \ldots, v_n\}$ und $\mathcal{B} = \{w_1, \ldots, w_m\}$, so gilt:*

1. *Das Diagramm*

$$\begin{array}{ccc} V & \xrightarrow{f} & W \\ {\scriptstyle\varphi_{\mathcal{A}}}\uparrow & & \uparrow{\scriptstyle\varphi_{\mathcal{B}}} \\ K^n & \xrightarrow{A} & K^m \end{array}$$

   ***kommutiert**, das heißt*

$$f(\varphi_{\mathcal{A}}(x)) = \varphi_{\mathcal{B}}(Ax) \quad \forall x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in K^n.$$

2. *Jede Matrix $A \in K^{m\times n}$ liefert eine Abbildung f, so dass das Diagramm kommutiert.*

Beweis. Zu 1.: Der untere Pfeil ist

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \overset{A}{\mapsto} Ax = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{pmatrix} \in K^m.$$

Unter $\varphi_{\mathcal{B}}$ (rechter Pfeil) geht dies über in:

$$\varphi_{\mathcal{B}}(Ax) = \varphi_{\mathcal{B}} \begin{pmatrix} \sum_{j=1}^{n} a_{1j}x_i \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{pmatrix}.$$

Betrachten wir nun den anderen Weg: Der linke Pfeil ist

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \overset{\varphi_{\mathcal{A}}}{\mapsto} \sum_{j=1}^{n} x_j v_j.$$

Unter f geht dies über in (oberer Pfeil):

$$\begin{aligned}
\sum_{j=1}^{n} x_j v_j \qquad &\mapsto \qquad f\left(\sum_{j=1}^{n} x_j v_j\right) \\
&\overset{f \text{ linear}}{=} \qquad \sum_{j=1}^{n} x_j f(v_j) \\
&\overset{\text{Def. Matrix zu lin. Abb.}}{=} \qquad \sum_{j=1}^{n} x_j \sum_{i=1}^{m} a_{ij} w_i \\
&\overset{\text{neu klammern}}{=} \qquad \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} x_j\right) w_i.
\end{aligned}$$

Schauen wir uns nun die Ergebnisse, die wir auf den beiden Wegen in der rechten oberen Ecke erhalten haben, an, so sehen wir:

$$\varphi_{\mathcal{B}} \begin{pmatrix} \sum_{j=1}^{n} a_{1j} x_i \\ \vdots \\ \sum_{j=1}^{n} a_{mj} x_j \end{pmatrix} \overset{!}{=} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} x_j\right) w_i,$$

was nach Definition von $\varphi_{\mathcal{B}}$ gilt, so dass das Diagramm kommutiert.

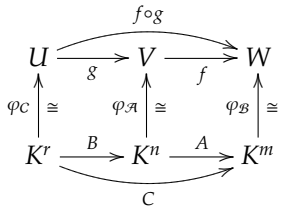
Zu 2.: Die Abbildung ist: $f = \varphi_{\mathcal{B}} \circ A \circ \varphi_{\mathcal{A}}^{-1}$. □

Wir können lineare Abbildungen zwischen endlich–dimensionalen Vektorräumen also vollständig auf Matrixebene verstehen und die Abbildungen $\varphi_{\mathcal{A}}$ usw. benutzen, um zwischen den ursprünglichen Vektorräumen und dem K^n hin– und herzuwechseln. Eine direkte Folgerung ist:

Korollar 20.15. *Es seien U, V, W drei K-Vektorräume mit Basen $\mathcal{C} = \{u_1, \ldots, u_r\}, \mathcal{A} = \{v_1, \ldots, v_n\}, \mathcal{B} = \{w_1, \ldots, w_m\}$, sowie $g\colon U \to V$, $f\colon V \to W$ zwei lineare Abbildungen. Dann gilt für die Matrixdarstellungen $A = M_{\mathcal{B}}^{\mathcal{A}}(f)$, $B = M_{\mathcal{A}}^{\mathcal{C}}(g)$ und $C = M_{\mathcal{B}}^{\mathcal{C}}(f \circ g)$, dass*

$$C = A \cdot B.$$

Mit anderen Worten: Das Diagramm



kommutiert; insbesondere heißt dies, dass das Matrixprodukt der Komposition von linearen Abbildungen entspricht.

Beweis. Wir betrachten die Hintereinanderausführung $f \circ g \colon U \to W$. Für einen Vektor u_k der Basis von U gilt:

$$\begin{aligned}(f \circ g)(u_k) &= f(g(u_k)) \\ &= f\Big(\sum_{j=1}^{n} b_{jk} v_j\Big) \\ &= \sum_{j=1}^{n} b_{jk} \cdot f(v_j) \\ &= \sum_{j=1}^{n} b_{jk} \cdot \sum_{i=1}^{m} a_{ij} w_i \\ &= \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} a_{ij} b_{jk}\Big) \cdot w_i.\end{aligned}$$

Also:

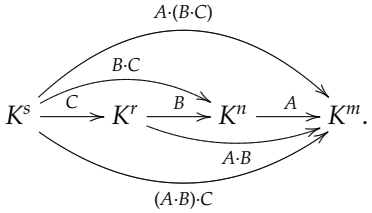
$$C = (c_{ik}) \in K^{m \times r} \text{ mit } c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}$$

ist die Matrixdarstellung von $f \circ g$. □

Korollar 20.16. *Das Matrixprodukt ist assoziativ, das heißt:*

$$(A \cdot B) \cdot C = A \cdot (B \cdot C) \quad \forall A \in K^{m \times n} \; \forall B \in K^{n \times r} \; \forall C \in K^{r \times s}.$$

Beweis. Die Komposition von linearen Abbildungen ist assoziativ:



□

20.5 Invertierbare Matrizen

Definition 20.17. *Eine quadratische Matrix $A \in K^{n \times n}$ heißt **invertierbar**, wenn die lineare Abbildung $f \colon K^n \to K^n$, $x \mapsto f(x) = Ax$, ein Isomorphismus ist.*

Die Matrixdarstellung der Umkehrabbildung $B = M_{\mathcal{E}}^{\mathcal{E}}(f^{-1}) \in K^{n\times n}$ *erfüllt:*

$$B \cdot A = E = (\delta_{kl}).$$

Hierbei bezeichnet $\delta_{kl} := \begin{cases} 1, & \text{falls } k = l \\ 0, & \text{sonst.} \end{cases}$ *das* ***Kroneckersymbol****; E ist also die Einheitsmatrix:*

$$E = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \in K^{n\times n}.$$

Wir definieren die ***Inverse****:*

$$A^{-1} := B.$$

Bemerkung 20.18. Es gilt: $f^{-1} \circ f = \mathrm{id}_{\mathbb{R}^n}$, also

$$M_{\mathcal{E}}^{\mathcal{E}}(f^{-1}) \cdot M_{\mathcal{E}}^{\mathcal{E}}(f) = M_{\mathcal{E}}^{\mathcal{E}}(\mathrm{id}_{\mathbb{R}^n}) \text{ d.h. } B \cdot A = E.$$

Da auch $f \circ f^{-1} = \mathrm{id}_{\mathbb{R}^n}$, gilt $A \cdot B = E$ ebenfalls.

A^{-1} ist durch A eindeutig bestimmt, denn es definiert die eindeutig bestimmte Umkehrabbildung

$$f\colon K^n \xrightarrow{A} K^n, \quad f^{-1} : K^n \xrightarrow{A^{-1}} K^n.$$

Satz/Definition 20.19. *Die Menge der quadratischen invertierbaren n×n–Matrizen über K bezeichnen wir mit*

$$\mathrm{GL}(n, K) := \{A \in K^{n\times n} \mid A \text{ ist invertierbar}\}$$

(GL steht für general linear*). Vermöge des Matrizenproduktes ist* $\mathrm{GL}(n, k)$ *eine Gruppe.*

Definition 20.20. *Eine* ***Gruppe*** $(G, \cdot)$ *ist eine Menge G, zusammen mit einer Verknüpfung* $\cdot$ *, das heißt einer Abbildung*

$$G \times G \to G, \quad (A, B) \mapsto A \cdot B,$$

die folgenden Axiomen genügt:

G1) Assoziativgesetz:

$$(A \cdot B) \cdot C = A \cdot (B \cdot C) \quad \forall A, B, C \in G.$$

G2) Existenz des neutralen Elements:

$$\exists E \in G \text{ mit } A \cdot E = A \quad \forall A \in G.$$

G3) Existenz von Inversen:

$$\forall A \in G, \exists A^{-1} \in G, \text{ so dass } A^{-1} \cdot A = E.$$

Eine Gruppe heißt ***abelsch*** *(nach N.H. Abel (1802-1829), oder* ***kommutativ****), falls für alle $g, h \in G$ gilt: $gh = hg$.*

Beweis (von Satz 20.19). Ist klar mit den vorigen Sätzen. □

Einige Beispiele von Gruppen sind:

Beispiel 20.21.

1. $(\mathbb{Z}, +)$ ist eine Gruppe:
   $a + b \in \mathbb{Z}$, $(a + b) + c = a + (b + c)$, $a + 0 = a \ \forall a \Rightarrow E = 0$, $a + (-a) = 0$
   (wird die Verknüpfung + verwendet, dann schreibt man für a^{-1} meist $-a$).
2. $(\mathbb{Z}^*, \cdot)$ ist keine Gruppe ($\mathbb{Z}^* := \mathbb{Z} \setminus \{0\}$):
   $1 \cdot a = a \quad \forall a \in \mathbb{Z}$, 1 ist also das neutrale Element ($E = 1$). Aber für $a \in \mathbb{Z}$ mit $|a| > 1$ existiert kein Inverses. Z.B: $\nexists\, b \in \mathbb{Z} : \ 2 \cdot b = 1$.
3. Sei K ein Körper. Dann sind $(K, +)$ und $(K^*, \cdot)$ abelsche Gruppen.

Bemerkung 20.22.

1. $\mathrm{GL}(n, K)$ ist nicht abelsch (siehe Übungsaufgaben).
2. Es gilt: $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$, denn $A \cdot B \cdot B^{-1} \cdot A^{-1} = A \cdot E \cdot A^{-1} = A \cdot A^{-1} = E$.

20.6 Berechnung der Inversen mit dem Gaußalgorithmus

Beispiel 20.23. Wir wollen die quadratische Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 3 & 4 & 6 \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

invertieren. Das heißt wir suchen ein $B \in \mathbb{R}^{3\times 3}$ mit $B \cdot A = E$ bzw. $A \cdot B = E$, wobei E die 3×3–Einheitsmatrix bezeichnet.

Die erste Spalte $(b_{11}, b_{21}, b_{31})^{\mathrm{t}}$ von B ist die Lösung des Gleichungssystems

$$\begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{21} \\ b_{31} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Analog für die zweite und dritte Spalte $b_i = (b_{1i}, b_{2i}, b_{3i})^t$. Dies sieht man besonders gut, wenn man das Matrizenprodukt $A \cdot B = E$ folgendermaßen notiert:

$$\begin{array}{cc} & \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \end{array}$$

Die drei Gleichungssysteme $A \cdot b_i = e_i$, $i = 1, 2, 3$, können wir simultan mit dem Gaußalgorithmus lösen:

1. Wir bilden die erweiterte Matrix

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 3 & 5 & 0 & 1 & 0 \\ 3 & 4 & 6 & 0 & 0 & 1 \end{array}\right).$$

2. Wir bringen diese mit dem Gaußalgorithmus auf Zeilenstufenform: Zuerst 2. Zeile - 2 × 1. Zeile, 3. Zeile - 3 × 1. Zeile,

$$\rightsquigarrow \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -1 & -1 & -2 & 1 & 0 \\ 0 & -2 & -3 & -3 & 0 & 1 \end{array}\right),$$

dann noch 3. Zeile - 2 × 2. Zeile:

$$\rightsquigarrow \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -1 & -1 & -2 & 1 & 0 \\ 0 & 0 & -1 & 1 & -2 & 1 \end{array}\right).$$

3. Wir machen die Diagonalelemente der linken Teilmatrix zu 1 (2. und 3. Zeile durchmultiplizieren mit −1):

$$\rightsquigarrow \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 & -1 & 0 \\ 0 & 0 & 1 & -1 & 2 & -1 \end{array}\right).$$

4. Wir räumen die Einträge oberhalb der Diagonalen des linken Blocks von unten nach oben aus. Zuerst 2. Zeile - 3. Zeile und 1. Zeile - 3 × 3. Zeile,

$$\rightsquigarrow \left(\begin{array}{ccc|ccc} 1 & 2 & 0 & 4 & -6 & 3 \\ 0 & 1 & 0 & 3 & -3 & 1 \\ 0 & 0 & 1 & -1 & 2 & -1 \end{array}\right),$$

dann noch 1. Zeile - 2 × 2. Zeile:

$$\rightsquigarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -2 & 0 & 1 \\ 0 & 1 & 0 & 3 & -3 & 1 \\ 0 & 0 & 1 & -1 & 2 & -1 \end{array}\right).$$

Die Inverse ist nun hinter dem Strich abzulesen:

$$A^{-1} = \begin{pmatrix} -2 & 0 & 1 \\ 3 & -3 & 1 \\ -1 & 2 & -1 \end{pmatrix}.$$

Dass dies auch allgemein so funktioniert, werden wir im nächsten Abschnitt sehen.

20.7 Der Gaußalgorithmus zur Berechnung der Inversen

Sei $A \in K^{n\times n}$.

1. A ist invertierbar genau dann, wenn die Zeilen–Stufenform genau n Stufen hat:

$$\tilde{A} = \begin{pmatrix} \underline{\tilde{a}_{11}} & * & * & * \\ 0 & \underline{\tilde{a}_{22}} & * & * \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \underline{\tilde{a}_{nn}} \end{pmatrix}$$

   Die Notwendigkeit ist klar, weil sonst die letzte Zeile eine Nullzeile ist. Die zugehörige lineare Abbildung ist nicht surjektiv, da dann nämlich $\tilde{A} \cdot x = (x_1, \ldots, x_{n-1}, 0)^t \ \forall x \in K^n$. Die andere Richtung der Behauptung zeigt der folgende Algorithmus.

2. Ist A invertierbar, so erhält man die inverse Matrix wie folgt:
   a) Wir bilden die um E erweiterte Matrix

$$(A \mid E) = \left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ \vdots & a_{22} & \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & 0 \\ a_{n1} & \cdots & a_{nn} & 0 & \cdots & 0 & 1 \end{array}\right)$$

   und bringen diese auf Zeilenstufenform (mit Zeilenoperationen)

$$(A \mid E) \rightsquigarrow (\tilde{A} \mid \tilde{B}) = \left(\begin{array}{ccc|c} \underline{\tilde{a}_{11}} & \cdots & \tilde{a}_{1n} & \\ & \ddots & \vdots & \tilde{b}_{ij} \\ 0 & & \underline{\tilde{a}_{nn}} & \end{array}\right)$$

   (möglicherweise Zeilenvertauschungen nötig).

b) Wir dividieren die k-te Zeile jeweils durch $\tilde{a}_{kk} \in K \setminus \{0\}$ (da die Zeilenstufenform genau n Stufen hat), und erhalten die Gestalt:

$$\rightsquigarrow (\tilde{\tilde{A}} \mid \tilde{\tilde{B}}) = \left(\begin{array}{ccc|c} 1 & \cdots & \tilde{\tilde{a}}_{1n} & \\ \vdots & \ddots & \vdots & \tilde{\tilde{b}}_{ij} \\ 0 & \cdots & 1 & \end{array}\right)$$

c) Wir räumen durch Zeilenoperationen die Einträge $\tilde{\tilde{a}}_{ij}$ sukzessive aus, etwa in der Reihenfolge:

$$\begin{array}{cccc} \tilde{\tilde{a}}_{n-1,n}, & \tilde{\tilde{a}}_{n-2,n}, & \ldots, & \tilde{\tilde{a}}_{1,n} \\ & \tilde{\tilde{a}}_{n-2,n-1}, & \ldots, & \vdots \\ & & \ddots & \vdots \\ & & & \tilde{\tilde{a}}_{1,2}. \end{array}$$

Dann haben wir eine Matrix:

$$\left(\begin{array}{ccc|c} 1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & B \\ 0 & \cdots & 1 & \end{array}\right).$$

Behauptung. Für die eben erhaltene Matrix gilt: $A^{-1} = B$.

Beweis. In der ersten Spalte von B steht $(b_{11}, \ldots, b_{n1})^{\mathrm{t}}$, die Lösung des Gleichungssystems:

$$A \begin{pmatrix} b_{11} \\ \vdots \\ b_{n1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Allgemein: Die k–te Spalte $(b_{\cdot k})$ von B löst:

$$A \begin{pmatrix} b_{1k} \\ \vdots \\ b_{nk} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = e_k.$$

Also insgesamt: $A \cdot B = E$. Es folgt: B ist invertierbar und $A = B^{-1}$, also auch $B \cdot B^{-1} = E$ und letztlich $B = A^{-1}$. $\square$

20.8 Klassifikationssatz/Struktursatz von Linearen Abbildungen

20.8.1 Die Resultate

Bzgl. unterschiedlicher Basen kann die Matrixdarstellung einer linearen Abbildung sehr verschieden aussehen. Die folgende Aussage erklärt, wie man aus einer Matrixdarstellung zu einer neuen kommt, wenn alle Basen (alte und neue) bekannt sind:

Satz/Definition 20.24 (Basiswechsel). *Es seien V, W zwei endlich-dimensionale K-Vektorräume mit Basen*

$$\mathcal{A} = \{v_1, \ldots, v_n\}, \mathcal{B} = \{w_1, \ldots, w_m\}$$

und $f\colon V \to W$ eine lineare Abbildung. $A = M_{\mathcal{B}}^{\mathcal{A}}(f)$ sei die Matrixdarstellung von f bezüglich dieser Basen.

Sind $\mathcal{A}' = \{v_1', \ldots, v_n'\}$, $\mathcal{B}' = \{w_1', \ldots, w_m'\}$, dann ergibt sich die Matrixdarstellung $B = M_{\mathcal{B}'}^{\mathcal{A}'}(f)$ in den neuen Basen wie folgt:

$$B = TAS^{-1},$$

*wobei $T = M_{\mathcal{B}'}^{\mathcal{B}}(\mathrm{id}_W)$, $S = M_{\mathcal{A}'}^{\mathcal{A}}(\mathrm{id}_V)$ die sogenannten **Basiswechselmatrizen** sind. Hierbei bezeichnen $\mathrm{id}_V : V \to V$ und $\mathrm{id}_W : W \to W$ jeweils die identischen Abbildungen.*

Mit anderen Worten: Das folgende Diagramm kommutiert:

$$\begin{array}{ccc} K^n & \xrightarrow{B} & K^m \\ \uparrow{\scriptstyle \varphi_{\mathcal{A}'}} & & \uparrow{\scriptstyle \varphi_{\mathcal{B}'}} \\ V & \xrightarrow{f} & W \\ \uparrow{\scriptstyle \varphi_{\mathcal{A}}} & & \uparrow{\scriptstyle \varphi_{\mathcal{B}}} \\ K^n & \xrightarrow{A} & K^m. \end{array} \qquad M_{\mathcal{A}'}^{\mathcal{A}}(\mathrm{id}_V)=S, \quad T=M_{\mathcal{B}'}^{\mathcal{B}}(\mathrm{id}_W)$$

Beweis. Klar nach Definition der Matrixdarstellung. Beispielsweise:

$$\begin{array}{ccc} W & \xrightarrow{\mathrm{id}_W} & W \\ \uparrow{\scriptstyle \varphi_{\mathcal{B}}} & & \uparrow{\scriptstyle \varphi_{\mathcal{B}'}} \\ K^m & \xrightarrow{T=M_{\mathcal{B}'}^{\mathcal{B}}(\mathrm{id}_W)} & K^m. \end{array}$$

□

Wählt man die Basen für eine gegebene lineare Abbildung geeignet, so ist es immer möglich, nahezu die Einheitsmatrix zu erhalten. Genauer:

Satz 20.25 (Klassifikationssatz/Struktursatz von linearen Abbildungen). *Sei $f\colon K^n \to K^m$ die durch die Matrix $A \in K^{m\times n}$ definierte lineare Abbildung. Dann existieren $S \in \mathrm{GL}(n,K), T \in \mathrm{GL}(m,K)$, so dass:*

$$TAS^{-1} = \left(\begin{array}{c|c} 1 & \\ \ddots & 0 \\ & 1 \\ \hline 0 & 0 \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c} 1 \\ \ddots \\ 1 \end{array}}\right\} r \\ \left.\vphantom{0}\right\} m-r \end{array}$$

$$\underbrace{\phantom{\begin{array}{c} 1 \\ \ddots \\ 1 \end{array}}}_{r} \quad \underbrace{}_{n-r}$$

für ein $r \le \min(n,m)$.

Beweis. Wir wählen Basen von K^n und K^m geschickt: Zunächst betrachten wir dazu den Kern

$$\operatorname{Ker} A = \{x \in K^n \mid Ax = 0\}.$$

Ist $d = \dim(\operatorname{Ker} A)$, so setzen wir $r = n - d$ (also $r \le n$). Zunächst wählen wir eine Basis von $\operatorname{Ker} A \subseteq K^n$, die wir mit $v_{r+1}, \dots, v_n$ durchnummerieren. Anschließend ergänzen wir diese durch Vektoren $v_1, \dots, v_r \in K^n$ zu einer Basis $\mathcal{A} = \{v_1, \dots, v_n\}$ von K^n.

Seien $w_i = f(v_i)$, $i = 1, \dots, r$, die Bilder der ersten r Vektoren. Dann sind $w_1, \dots, w_r \in K^m$ linear unabhängig: Wären sie nämlich abhängig, etwa

$$\lambda_1 w_1 + \cdots + \lambda_r w_r = 0,$$

so wäre

$$\lambda_1 v_1 + \cdots + \lambda_r v_r \in \operatorname{Ker} A = \langle v_{r+1}, \dots, v_n \rangle$$

das heißt $\{v_1, \dots, v_n\}$ wäre keine Basis, außer $\lambda_1 = \cdots = \lambda_r = 0$. Da also $w_1, \dots, w_r$ insgesamt r linear unabhängige Vektoren in K^m sind, folgt:

$$r \le m \ (= \dim K^m).$$

Wir ergänzen nun $w_1, \dots, w_r$ zu einer Basis $\mathcal{B} = \{w_1, \dots, w_r, w_{r+1}, \dots, w_m\}$ des K^m. Bezüglich der Basen $\mathcal{A}$ und $\mathcal{B}$ hat f die Gestalt:

$$M_{\mathcal{B}}^{\mathcal{A}}(f) = \left(\begin{array}{c|c} 1 & \\ \ddots & 0 \\ & 1 \\ \hline 0 & 0 \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c} 1 \\ \ddots \\ 1 \end{array}}\right\} r \\ \left.\vphantom{0}\right\} m-r \end{array}$$

$$\underbrace{\phantom{\begin{array}{c} 1 \\ \ddots \\ 1 \end{array}}}_{r} \quad \underbrace{}_{n-r}$$

Dies folgt sofort aus $f(v_i) = w_i$, $i = 1, \ldots, r$ und $f(v_j) = 0$, $i = r+1, \ldots, n$. Wenn also $S = M^{\mathcal{E}}_{\mathcal{A}}(\mathrm{id}_{K^n})$ und $T = M^{\mathcal{E}}_{\mathcal{B}}(\mathrm{id}_{K^m})$ die Basiswechselmatrizen sind, so folgt, dass das Diagramm (das aus dem äußeren Teil des Diagramms in vorigem Satz besteht)

$$\begin{array}{ccc} K^n & \xrightarrow{M^{\mathcal{A}}_{\mathcal{B}}(f)} & K^m \\ {\scriptstyle S}\uparrow & & \uparrow{\scriptstyle T} \\ K^n & \xrightarrow{A} & K^m \end{array}$$

kommutiert. D.h., $TAS^{-1} = M^{\mathcal{A}}_{\mathcal{B}}(f)$. □

Bezüglich geeigneter Basen kann jede lineare Abbildung zwischen endlich–dimensionalen Vektorräumen also durch eine sehr einfache Matrix beschrieben werden. Der Wechsel zur passenden Basis in Definitions– bzw. Ziel–Vektorraum wird jeweils von einer invertierbaren Matrix realisiert. Mit Hilfe des Beweises ist folgende oft hilfreiche Formel leicht einzusehen:

Korollar 20.26 (Dimensionsformel). *Sei $f\colon V \to W$ eine lineare Abbildung zwischen zwei K-Vektorräumen und* $\dim V < \infty$. *Dann gilt:*

$$\dim \mathrm{Bild}(f) + \dim \mathrm{Ker}(f) = \dim V.$$

Beweis. Ist $d = \dim \mathrm{Ker}\, f$ und $n = \dim V$, dann ist $\mathrm{Bild}(f) = \langle w_1, \ldots, w_r \rangle$ (siehe Beweis des Satzes), also $\dim \mathrm{Bild}(f) = r$, wobei $r = n - d$ (auch nach dem Beweis des Satzes). □

Bemerkung 20.27. Die Formel gilt auch, falls $\dim V = \infty$. Dann muss nämlich wenigstens einer der Vektorräume $\mathrm{Ker}\, f$ oder $\mathrm{Bild}\, f$ ebenfalls ∞-dimensional sein.

20.8.2 Geometrische Interpretation des Klassifikationssatzes

Sei $f\colon V \to W$ eine lineare Abbildung. Bezüglich geeigneter Basen bzw. Koordinaten ist f eine Parallelprojektion:

$$\begin{pmatrix} v_1 \\ \vdots \\ v_r \\ v_{r+1} \\ \vdots \\ v_n \end{pmatrix} \mapsto \begin{pmatrix} v_1 \\ \vdots \\ v_r \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Geometrisch sieht dies aus wie in Abbildung 20.2.

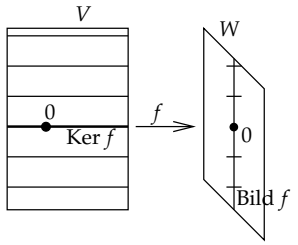


Abbildung 20.2. Geometrische Interpretation des Klassifikationssatzes linearer Abbildungen.

20.8.3 Anwendung für Gleichungssysteme

Sei $Ax = b$ mit $A \in K^{m\times n}$, $b \in K^m$, ein Gleichungssystem und

$$Ax = 0$$

das sogenannte **zugehörige homogene Gleichungssystem**.

Dann ist die Lösungsmenge des homogenen Gleichungssystems der Untervektorraum

$$\operatorname{Ker} A = \{x \in K^n \mid Ax = 0\}.$$

Ist $\tilde{x} \in K^n$ **eine** Lösung des i.A. **inhomogenen Gleichungssystems**

$$Ax = b,$$

dann ist dessen ganze Lösungsmenge

$$L_b = \{x \in K^n \mid Ax = b\} = \tilde{x} + \operatorname{Ker} A := \{\tilde{x} + x \in K^n \mid x \in \operatorname{Ker} A\}.$$

Es gilt nämlich für $x \in \operatorname{Ker} A$ und $\tilde{x}$ mit $A\tilde{x} = b$, dass $\tilde{x} + \operatorname{Ker} A \subseteq L_b$, da

$$\begin{aligned} A(\tilde{x} + x) &= A\tilde{x} + Ax \\ &= A\tilde{x} \\ &= b; \end{aligned}$$

umgekehrt gilt $L_b \subseteq \tilde{x} + \operatorname{Ker} A$:

$$\begin{aligned} x' \in L_b \Rightarrow A(x' - \tilde{x}) &= Ax' - A\tilde{x} \\ &= b - b \\ &= 0 \end{aligned}$$

$$\Rightarrow x' - \tilde{x} \in \operatorname{Ker} A,$$

also

$$x' \in \tilde{x} + \operatorname{Ker} A.$$

Insgesamt haben wir demnach $\tilde{x} + \operatorname{Ker} A \subseteq L_q \subseteq \tilde{x} + \operatorname{Ker} A$ eingesehen, so dass tatsächlich $L_b = \tilde{x} + \operatorname{Ker} A$, wenn eine Lösung $\tilde{x}$ von $Ax = b$ existiert.

Ist $b \notin \operatorname{Bild}(A)$, dann existiert kein $\tilde{x} \in K^n$ mit $A\tilde{x} = b$ und $L_b = \emptyset$.

Zusammenfassend können wir also das Problem, alle Lösungen von $Ax = b$ zu finden, lösen, indem wir die beiden einfacheren Probleme, alle Lösungen von $Ax = 0$ und nur eine Lösung von $Ax = b$ zu finden, lösen.

20.8.4 Spezialfall: Genauso viele Gleichungen wie Unbestimmte

Wir betrachten nun den wichtigen Spezialfall von Gleichungssystemen mit genauso vielen Gleichungen wie Unbestimmten, d.h. $A \in K^{n\times n}$.

Satz 20.28. *Sei $A \in K^{n\times n}$ und $b \in K^n$; mit f bezeichnen wir die zugehörige lineare Abbildung. Dann sind äquivalent:*

1. *A ist invertierbar, d.h. $A \in GL(n, K)$ bzw. f ist ein Isomorphismus.*
2. *$\operatorname{Ker} A = 0$, d.h. f ist ein Monomorphismus.*
3. *$\operatorname{Bild} A = K^n$, d.h. f ist ein Epimorphismus.*
4. *$Ax = b$ hat genau eine Lösung.*

Beweis. 1. $\iff$ 2. $\iff$ 3.: siehe Übungsaufgaben.

4. $\Rightarrow$ 2. & 3. & 1.: $Ax = b$ hat für **ein** b genau eine Lösung $\tilde{x}$. Die Menge aller Lösungen für dieses b ist dann aber $\tilde{x} + \operatorname{Ker} A$, d.h. es folgt $\operatorname{Ker} A = 0$, unabhängig von b. Da 1. bis 3. äquivalent sind, folgen auch die anderen Aussagen. □

Bemerkung 20.29. 1. Häufig will man das Gleichungssystem

$$Ax = b$$

für eine Matrix $A \in K^{n\times n}$ und viele verschiedene b ermitteln. Dann lohnt es sich, die Inverse A^{-1}, etwa mit Gauß, zu finden, weil dann für jedes b die Lösung einfach über $x = A^{-1}b$ zu berechnen ist.

2. Für $A \in GL(n, K)$ ist der Aufwand, A^{-1} zu berechnen, mit Hilfe des Gaußalgorithmus von der Größenordnung $O(n^3)$.
3. Eine Matrixmultiplikation

$$A \cdot B$$

auszurechnen für $A, B \in K^{n\times n}$ hat mit der Formel aus der Definition den Aufwand $O(n^3)$, denn es gibt n^2 Einträge von $A \cdot B$ und

$$c_{ik} = \sum_{j=1}^{n} a_{ik} b_{kj}$$

besteht aus n Termen.

Z.B.: $n = 2$. Der Aufwand ist 8 Multiplikationen. 7 Multiplikationen geht aber auch! Dies liefert für allgemeines n einen niedrigeren Aufwand (Strassen, 1969, wie schon in Satz 19.13 auf Seite 250 erwähnt): $O(n^{\log_2 7}) \approx O(n^{2,7})$. Inzwischen existieren asymptotisch bessere Algorithmen. Es scheint aber immer noch unklar, ob eine asymptotische Laufzeit von $O(n^2)$ möglich ist. Erstaunlicher Weise basieren neueste Ansätze auf Gruppentheorie.

Man kann außerdem zeigen, dass viele Matrixoperationen im Wesentlichen auf die Multiplikation von Matrizen zurückgeführt werden können. Beispielsweise ist das Invertieren von Matrizen genauso schnell wie das Multiplizieren, so dass ein Algorithmus in $O(n^2)$ für die Multiplikation auch einen Algorithmus für das Multiplizieren in $O(n^2)$ liefern würde.

20.9 Summen von Vektorräumen

Definition 20.30. *Seien V ein K-Vektorraum und $U, W \subseteq V$ zwei Untervektorraume. Dann bezeichnet*

$$U + W = \{v \in V \mid \exists\, u \in U, \exists\, w \in W : v = u + w\}$$

die Summe der Untervektorräume.

*Die **äußere** bzw. **direkte Summe** von U und W ist*

$$U \oplus W := U \times W = \{(u, w) \mid u \in U,\ w \in W\}.$$

Wir haben eine kanonische lineare Abbildung

$$f\colon U \oplus W \to V, \quad (u, w) \mapsto u + w.$$

Häufig wird $U \oplus W$ auch nur als Notation von $U + W$ verwendet, wenn $U \cap W = 0$.

Satz 20.31. *Mit obiger Notation gilt:*

$$\text{Bild}(f\colon U \oplus W \to V) = U + W$$

und

$$\text{Ker } f \cong U \cap W$$

vermöge

$$g\colon U \cap W \to U \oplus W, \quad x \mapsto (x, -x).$$

Beweis. Bild $g \cong U \cap W$ ist klar, weil $(x,-x) \mapsto x$ die Umkehrabbildung ist. Bild $f = U + W$ ist klar nach Definition von $U + W$.

Bild $g \subseteq \operatorname{Ker} f$ ebenso, da $f(x,-x) = x - x = 0$. Umgekehrt müssen wir noch zeigen, dass $\operatorname{Ker} f \subseteq \operatorname{Bild} g$ gilt, da dann $\operatorname{Ker} f \cong \operatorname{Bild} g$ folgt. Ist $(u,w) \in \operatorname{Ker} f \subseteq U \oplus W$, dann folgt aus $0 = f(u,w) = u + w$, dass $w = -u$, also $w, u \in U \cap W$, und $(u,w) = g(u) \in \operatorname{Bild} g$ und deshalb $\operatorname{Ker} f \subseteq \operatorname{Bild} g$. Insgesamt ergibt sich:

$$\operatorname{Bild} g = \operatorname{Ker} f.$$

Es folgt mit der Dimensionsformel:

$$\dim \operatorname{Bild} g = \dim \operatorname{Ker} f = \dim(U \oplus W) - \dim(U + W).$$

Da aber $\dim \operatorname{Ker} f = \dim(\operatorname{Ker}(U \oplus W \to U + W))$ ist, weil die v außerhalb von $U + W$ nicht im Bild sind und da $\operatorname{Ker}(U \oplus W \to U + W) = U \cap W$ ist, also $\operatorname{Ker} f \cong U \cap W$ ist, induziert g also einen Isomorphismus

$$U \cap W \xrightarrow{\cong} \operatorname{Ker}(U \oplus W \to U + W).$$

□

Korollar 20.32 (Dimensionsformeln). *$U, W \subseteq V$ seien Untervektorräume. Dann gilt:*

$$\begin{aligned} \dim U + \dim W &= \dim(U \cap W) + \dim(U + W) \\ \dim(U \oplus W) = \dim(U \times W) &= \dim U + \dim W \\ \dim(U \cap W) &= \dim \operatorname{Ker} f \\ \dim(U + W) &= \dim \operatorname{Bild} f. \end{aligned}$$

Beweis. Dies folgt mit dem vorigen Satz direkt aus der ersten Dimensionsformel (Korollar 20.26). □

Beispiel 20.33. $V = \mathbb{R}^3$. Wir betrachten die beiden Untervektorräume:

$$U_t = \left\langle \begin{pmatrix} \cos t \\ \sin t \\ 0 \end{pmatrix} \right\rangle, \; W = \left\langle \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \right\rangle.$$

Welche Dimensionen können für $U_t \cap W$ und $U_t + W$ auftreten?

Es gilt: $\dim U_t = 1, \; \dim W = 2 \; \forall t$. Außerdem ist

$$\dim(U_t \cap W) = 0, \; \dim(U_t + W) = 3,$$

falls

$$\left\{\begin{pmatrix}\cos t\\ \sin t\\ 0\end{pmatrix},\begin{pmatrix}1\\1\\1\end{pmatrix},\begin{pmatrix}1\\-1\\1\end{pmatrix}\right\}$$

eine Basis des $\mathbb{R}^3$ ist und dies passt zu den Formeln aus dem Korollar.
Ist dies nicht der Fall, dann:

$$\begin{pmatrix}\cos t\\ \sin t\\ 0\end{pmatrix}\in\left\langle\begin{pmatrix}1\\1\\1\end{pmatrix},\begin{pmatrix}1\\-1\\1\end{pmatrix}\right\rangle = W \quad (U_t \subseteq W),$$

also:

$$\dim U_t = 1,\ \dim W = 2,\ \dim(U_t \cap W) = 1,\ \dim(U_t + W) = 2.$$

Dies liegt vor, falls:

$$\begin{pmatrix}\cos t\\ \sin t\\ 0\end{pmatrix}\in\left\langle\begin{pmatrix}0\\2\\0\end{pmatrix}\right\rangle \implies t = \frac{\pi}{2} \text{ bzw. } t = \frac{\pi}{2} + k\pi,\ k \in \mathbb{Z}.$$

Aufgaben

Aufgabe 20.1 (Dimension).

1. Seien $U = \langle(1,0,1)^t,(1,-1,1)^t\rangle$ und $W = \langle(0,1,-1)^t,(1,1,0)^t\rangle$ zwei Untervektorräume von $V = (\mathbb{F}_3)^3$. Die Elemente des Körpers $\mathbb{F}_3$ bezeichnen wir hierbei wie üblich mit $-1, 0, +1$. Berechnen Sie Dimension und Basen für die Vektorräume: U, W, $U + W$, $U \cap W$.
2. Zeigen Sie: Für einen Vektorraum-Homomorphismus $\varphi\colon K^n \to K^n$, $n < \infty$, gilt:
   φ injektiv $\Leftrightarrow$ φ surjektiv $\Leftrightarrow$ φ bijektiv.
3. Seien $U_\lambda = \langle(1,1,1)^t,(\lambda,\lambda,-\lambda)^t\rangle$, $W_\lambda = \langle(\cos\lambda,\sin\lambda,0)^t,(\cos\lambda,\sin\lambda,1)^t\rangle$ Unterräume des $\mathbb{R}^3$. Für welche $\lambda \in \mathbb{R}$ ist $\dim(U_\lambda \cap W_\lambda)\ldots$ (a) $\ldots = 0$? (b) $\ldots = 1$? (c) $\ldots = 2$? (d) $\ldots = 3$?
   Fertigen Sie eine Skizze der Situation an.

Aufgabe 20.2 (Kern und Bild). Bestimmen Sie jeweils eine Basis von Kern und Bild derjenigen linearen Abbildungen, die durch folgende Matrizen definiert werden. Überprüfen Sie die Dimensionsformel für diese Beispiele.

$$A = \begin{pmatrix}1&2&3&4\\1&-2&1&3\\1&0&1&2\\1&0&3&5\end{pmatrix}\in\mathbb{R}^{4\times4},\quad B = \begin{pmatrix}1&2&3\\1&-2&1\\2&0&4\\-3&-2&-7\end{pmatrix}\in\mathbb{R}^{4\times3}.$$

Aufgabe 20.3 (Kern und Bild). Sei $n \in \mathbb{N}$. Welche der folgenden Aussagen sind richtig? Kurze Begründung:

(*a*) $\operatorname{Ker} A \subset \operatorname{Ker} A^2 \quad \forall A \in \mathbb{R}^{n\times n}$, (*b*) $\operatorname{Ker} A \supset \operatorname{Ker} A^2 \quad \forall A \in \mathbb{R}^{n\times n}$,

(*c*) $\operatorname{Bild} A \subset \operatorname{Bild} A^2 \quad \forall A \in \mathbb{R}^{n\times n}$, (*d*) $\operatorname{Bild} A \supset \operatorname{Bild} A^2 \quad \forall A \in \mathbb{R}^{n\times n}$.

Aufgabe 20.4 (Invertieren von Matrizen). Invertieren Sie die Matrix

$$A := \begin{pmatrix} 2 & -\frac{4}{3} & -\frac{4}{3} & -\frac{8}{3} \\ 0 & -\frac{4}{3} & -\frac{4}{3} & -\frac{2}{3} \\ 4 & \frac{2}{3} & -\frac{4}{3} & -\frac{8}{3} \\ 4 & 0 & -2 & -4 \end{pmatrix} \in \mathbb{Q}^{4\times 4}$$

mit Hilfe des Gaußalgorithmus.

Aufgabe 20.5 (Matrixdarstellung einer linearen Abbildung). Für eine Menge N bezeichne $\mathrm{id}_N \colon N \to N$ die identische Abbildung. Sei $V := \mathbb{R}[t]_{\leq d}$. Bestimmen Sie die Matrixdarstellung

$$A := M_{\mathcal{B}}^{\mathcal{A}}(\mathrm{id}_V)$$

von id_V bzgl. der Basen $\mathcal{A} := \{1, t, \ldots, t^d\}$ und $\mathcal{B} := \{1, t-\alpha, \ldots, (t-\alpha)^d\}$.

Aufgabe 20.6 (Invertierbarkeit von Matrizen). Zeigen Sie: Die **Vandermondsche Matrix**

$$A := \begin{pmatrix} 1 & \alpha_0 & \alpha_0^2 & \ldots & \alpha_0^d \\ 1 & \alpha_1 & \alpha_1^2 & \ldots & \alpha_1^d \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \alpha_d & \alpha_d^2 & \ldots & \alpha_d^d \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$$

ist genau dann invertierbar, wenn $\alpha_0, \ldots, \alpha_d \in \mathbb{R}$ paarweise verschieden sind.

21

Gruppen und Symmetrie

In diesem Abschnitt gehen wir etwas detaillierter auf den bereits erwähnten Begriff der Gruppe ein. Gruppen treten in sehr vielen Bereichen der Mathematik auf und sind daher von grundlegender Bedeutung.

21.1 Definition und erste Beispiele

Zwar haben wir auf Seite 262 schon im Zusammenhang mit der Gruppe $\mathrm{GL}(n, K)$ den Begriff der Gruppe erwähnt, trotzdem hier noch einmal die wesentlichen Eigenschaften: Eine Gruppe G ist eine Menge, auf der eine Verknüpfung existiert, die assoziativ ist (G1) und für die ein neutrales Element e existiert mit $ae = a\ \forall a \in G$ (G2) und für jedes Element a der Menge ein Inverses a' existiert mit $aa' = e$ (G3); oft schreibt man $a^{-1} := a'$. Ist in einer Gruppe G zusätzlich das Kommutativgesetz (G4) erfüllt,

$$\text{(G4)} \quad ab = ba \quad \forall a, b \in G,$$

so nennt man die Gruppe abelsch. Bei abelschen Gruppen verwendet man oft die additive Notation: $+$ für die Verknüpfung, 0 für das neutrale Element, $-a$ für das Inverse.

Beispiel 21.1. 1. Sei K ein Körper. $(\mathbb{Z}, +), (K, +), (K^*, \cdot) = (K \setminus \{0\}, \cdot)$ sind Gruppen.
$(\mathbb{Z} \setminus \{0\}, \cdot)$ ist keine Gruppe, da (G3) nicht erfüllt ist ($\frac{1}{2} \notin \mathbb{Z}$).

2. K Körper. $GL(n, K) = \{A \in K^{n \times n} \mid A \text{ ist invertierbar }\}$
ist eine Gruppe bezüglich des Matrizenprodukts.

$$e = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}.$$

3. Eine Abbildung

$$f : \mathbb{R}^n \to \mathbb{R}^n, \text{mit } f(0) = 0 \text{ und } \|f(x) - f(y)\| = \|x - y\| \quad \forall x, y \in \mathbb{R}^n$$

heißt **orthogonal**. Man kann zeigen: f ist linear und bijektiv, das heißt

$$f(x) = Ax \text{ für ein gewisses } A \in \mathrm{GL}(n, \mathbb{R}).$$

Die Menge der orthogonalen Abbildungen auf dem $\mathbb{R}^n$ bilden eine Gruppe, die sogenannte **Orthogonale Gruppe** $\mathrm{O}(n)$. Denn mit $x' = g(x)$ und $y' = g(y)$ ist

$$\|f(g(x)) - f(g(y))\| = \|f(x') - g(y')\| = \|x' - y'\| = \|x - y\|.$$

Man kann zeigen, dass die zugehörigen Matrizen (genannt **orthogonale Matrizen** genau jene sind mit der Eigenschaft $A^{\mathrm{t}} A = E$, wobei A^{t} die **transponierte Matrix** bezeichnet:

$$(A^{\mathrm{t}})_{ij} = A_{ji}.$$

Dies passt mit der entsprechenden Notation für Vektoren zusammen: aus einer $n \times 1$–Matrix wird eine $1 \times n$–Matrix.

Man kann zeigen, dass $\det A \in \{\pm 1\}$ für $A \in O(n)$, wobei $\det A$ die sogenannte Determinante von A bezeichnet, die wir noch studieren werden; man setzt $\mathrm{SO}(n) := \{A \in O(n) \mid \det A = 1\}$ (**Spezielle Orthogonale Gruppe im** $\mathbb{R}^n$). Die Menge der Drehungen im $\mathbb{R}^2$ um den Ursprung ist beispielsweise SO(2). Die Matrixdarstellung für eine Drehung um den Winkel α um den Ursprung erhält man aus $(1, 0)^{\mathrm{t}} \mapsto (\cos\alpha, \sin\alpha)^{\mathrm{t}}$ und $(0, 1)^{\mathrm{t}} \mapsto (-\sin\alpha, \cos\alpha)^{\mathrm{t}}$ (siehe auch Abb. 29.1):

$$\mathrm{O}(2) \supset \mathrm{SO}(2) = \left\{ \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \;\middle|\; \alpha \in [0, 2\pi[\right\}.$$

Beispiel 21.2. Seien G_1, G_2 Gruppen. $G_1 \times G_2 = \{(g_1, g_2) \mid g_1 \in G_1, g_2 \in G_2\}$ ist ebenfalls eine Gruppe vermöge:

$$(a_1, a_2) \circ (b_1, b_2) = (a_1 b_1,\ a_2 b_2).$$

Das neutrale Element ist: $e_{G_1 \times G_2} = (e_{G_1}, e_{G_2})$.

Bemerkung/Definition 21.3 (Elementare Eigenschaften von Gruppen). In jeder Gruppe G mit neutralem Element e gilt:

1. Das Neutrale e erfüllt auch: $e \cdot a = a \quad \forall a \in G$.
2. e ist eindeutig durch die Eigenschaft $a \cdot e = a \quad \forall a \in G$ charakterisiert.

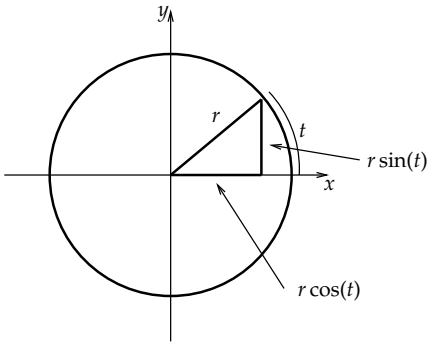


Abbildung 21.1. Sinus und Cosinus am Einheitskreis.

3. Das Inverse a' zu $a \in G$ erfüllt auch $a' \cdot a = e$.
4. Für festes a ist $a' \in G$ durch die Eigenschaft $a \cdot a' = e$ eindeutig bestimmt.

Meist schreibt man $a^{-1} := a'$ für das inverse Element und ab für $a \cdot b$.

Beweis. Zu 3.: Zu a' gibt es $a'' \in G$, so dass $a' \cdot a'' = e$ nach (G3). Es folgt:

$$\begin{aligned} a' \cdot a &\overset{G2}{=} (a' \cdot a) \cdot e = (a' \cdot a) \cdot (a' \cdot a'') \\ &\overset{G1}{=} ((a' \cdot a) \cdot a') \cdot a'' \\ &\overset{G1}{=} (a' \cdot (a \cdot a')) \cdot a'' \\ &\overset{G3}{=} (a' \cdot e) \cdot a'' \\ &\overset{G2}{=} a' \cdot a'' = e. \end{aligned}$$

Zu 1.: Wir können nun 3. verwenden:

$$ea \overset{G3}{=} (aa')a \overset{G1}{=} a(a'a) \overset{3.}{=} ae \overset{G2}{=} a.$$

Zu 2.: Sei $\tilde{e}$ ein weiteres neutrales Element. Dann gilt

$$\tilde{e} \overset{G2}{=} e \cdot \tilde{e} \overset{1.}{=} e.$$

e ist nämlich ebenfalls neutrales Element.

Zu 4.: Sei $\tilde{a}$ ein weiteres inverses Element zu a. Dann gilt:

$$e \overset{G3}{=} a \cdot \tilde{a} \overset{3.}{=} \tilde{a} \cdot a.$$

Es folgt:

$$\tilde{a} \overset{G2}{=} \tilde{a}e \overset{G3}{=} \tilde{a}(a \cdot a') \overset{G1}{=} (\tilde{a}a)a' \overset{3.}{=} ea' \overset{1.}{=} a'.$$

□

21.2 Permutationsgruppen

Sei M eine Menge. Wir setzen:

$$\mathrm{Bij}(M) := \{\sigma : M \to M \mid \sigma \text{ ist bijektiv } \}.$$

$\mathrm{Bij}(M)$ zusammen mit der Komposition von Abbildungen bilden eine Gruppe. Neutrales Element:

$$\mathrm{id}_M \colon M \to M, \quad x \mapsto x.$$

Inverses eines Elements σ: die Umkehrabbildung σ^{-1}. Außer für den Spezialfall zweielementiger Mengen, d.h. $|M| \leq 2$, ist $\mathrm{Bij}(M)$ keine abelsche Gruppe, wie wir in Beispiel 21.5 sehen werden.

21.2.1 Die Permutationsgruppen S_n

Definition 21.4. *Für $M = \{1, \ldots, n\}$ heißt*

$$S_n = \mathrm{Bij}(\{1, \ldots, n\})$$

*die **Gruppe der Permutationen** von $\{1, \ldots, n\}$. Ein Element $\sigma \in S_n$ nennt man eine **Permutation**.*

Häufig wird σ in Tabellenform angegeben:

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{pmatrix}.$$

Beispiel 21.5. Wir betrachten zwei Permutationen für den Fall $n = 3$:

$$\sigma = \begin{pmatrix} 1\,2\,3 \\ 2\,1\,3 \end{pmatrix} = \sigma \in S_3, \quad \tau = \begin{pmatrix} 1\,2\,3 \\ 2\,3\,1 \end{pmatrix} \in S_3.$$

Für diese gilt:

$$\sigma \circ \tau = \begin{pmatrix} 1\,2\,3 \\ 1\,3\,2 \end{pmatrix} \neq \tau \circ \sigma = \begin{pmatrix} 1\,2\,3 \\ 3\,2\,1 \end{pmatrix}.$$

Die Gruppe S_3 ist also nicht abelsch.

Allgemein gilt für beliebiges n, dass $|S_n| = n!$ ($= 1 \cdot 2 \cdots n$, siehe auch Beispiel 1.19). Denn um $\sigma(1)$ in

$$\sigma = \begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{pmatrix}$$

zu spezifizieren, haben wir n Wahlmöglichkeiten, anschließend für $\sigma(2)$ genau $n-1$ Wahlmöglichkeiten, …, für $\sigma(k)$ genau $n-(k-1)$ Wahlmöglichkeiten, da $\sigma(1), \ldots, \sigma(k-1)$ für $\sigma(k)$ nicht mehr in Frage kommen, also

$$|S_n| = n \cdot (n-1) \cdots 2 \cdot 1 = n!.$$

Beispiel 21.6. Die Symmetriegruppe des Quadrats hat 8 Elemente (s. Abb. 21.2). Es gibt 4 Spiegelungen und 4 Drehungen (um 90°, 180°, 270°, 360°; die letzte ist natürlich die identische Abbildung). Bezeichnet man die Ecken mit den Zahlen 1, 2, 3, 4, so kann man die Spiegelungen und Drehungen auch als Permutationen schreiben.

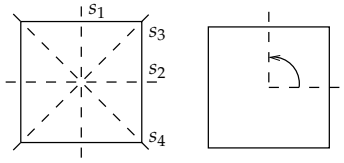


Abbildung 21.2. Die Symmetriegruppe des Quadrats: 4 Spiegelungen und 4 Drehungen.

Allgemein notieren wir mit D_{2n} die Symmetriegruppe des regulären n-Ecks, auch n–te **Diedergruppe** genannt. Sie hat $2n$ Elemente, nämlich n Drehungen und n Spiegelungen. Achtung: Manche Autoren schreiben für D_{2n} auch D_n; es muss also immer dazugesagt werden, welche der Notationen man verwendet.

21.2.2 Zykelschreibweise für Permutationen

Definition 21.7. *Seien $i_1, \ldots, i_k \in \{1, \ldots, n\}$ k paarweise verschiedene Elemente. Dann bezeichnet*

$$(i_1\ i_2\ \ldots\ i_k) \in S_n$$

die zyklische Vertauschung, die

$$i_j \text{ für } j = 1, \ldots, k-1 \text{ auf } i_{j+1} \text{ und } i_k \text{ auf } i_1$$

*abbildet und alle anderen Elemente von $\{1, \ldots, n\}$ festlässt. Eine solche Permutation heißt **Zykel**.*

Beispiel 21.8. Eine Permutation in Zykelschreibweise:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} = (1\ 2\ 3\ 4) \in S_4.$$

Eine weitere:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix} = (1\ 3) \in S_4.$$

Allgemein gilt:

Bemerkung 21.9. Jede Permutation $\sigma \in S_n$ ist die Komposition von **disjunkten Zyklen** (auch **elementfremden Zyklen**)

$$\sigma = (i_{11}\ i_{12}\ \dots\ i_{1k_1}) \cdot (i_{21}\ \dots\ i_{2k_2}) \cdots (i_{r1}\ \dots i_{rk_r}),$$

wobei die i_{jl} paarweise verschieden sind.

Beispiel 21.10. Einige Permutationen als Komposition elementfremder Zykel geschrieben:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = (1\ 2)(3\ 4) = (3\ 4)(1\ 2) \in S_4,$$
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 4 & 5 & 7 & 1 & 3 & 6 \end{pmatrix} = (1\ 2\ 4\ 7\ 6\ 3\ 5) \in S_7,$$
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 5 & 6 & 1 & 2 & 7 \end{pmatrix} = (1\ 3\ 5)(2\ 4\ 6)(7) = (1\ 3\ 5)(2\ 4\ 6) \in S_7.$$

21.2.3 Komposition von nicht disjunkten Zykeln

Es ist klar, dass Produkte disjunkter Zykel kommutativ sind. Für nicht disjunkte ist dies nicht unbedingt der Fall. Außerdem ist zunächst nicht klar, wie die Zykellänge eines Produktes von den Zykellängen der Elemente abhängt:

Beispiel 21.11. Es gilt:

$$(1\ 2\ 3)(3\ 4\ 5) = (1\ 2\ 3\ 4\ 5) \in S_5.$$

Wie allgemein für Abbildungen werden Kompositionen von Permutationen von rechts nach links berechnet ($(f \circ g)(x) = f(g(x))$):

$$(1\ 2)(3\ 4)(1\ 2\ 3\ 4) = (1)(2\ 4)(3) = (2\ 4)$$

und

$$(1\ 2\ 3\ 4)(3\ 4)(1\ 2) = (1\ 3)(2)(4) = (1\ 3).$$

Manche bevorzugen die Multiplikation von links.

Obwohl also nicht disjunkte Produkte nicht unbedingt kommutativ sind, sind sie oft sehr hilfreich, wie der folgende Satz zeigt.

Definition 21.12. *Eine **Transposition** in S_n ist eine Permutation τ der Gestalt $\tau = (kl)$.*

Satz 21.13. *Jede Permutation ist ein Produkt von Transpositionen.*

Beweis. Es reicht, dies für einen Zykel $(i_1 \dots i_k) \in S_n$ zu zeigen. Es gilt:

$$(i_1\ i_2\ \dots\ i_k) = (i_1\ i_2)(i_2\ i_3)\dots(i_{k-1}\ i_k),$$

denn

$$i_k \mapsto i_{k-1} \mapsto i_{k-2} \mapsto \dots \mapsto i_2 \mapsto i_1,$$
$$i_l \overset{(i_l\ i_{l+1})}{\mapsto} i_{l+1},\ \forall\, l < k.$$

□

Beispiel 21.14. Ein Dreierzykel ist Produkt von zwei Transpositionen:

$$(3\ 1\ 2) = (1\ 2\ 3) = (1\ 2) \cdot (2\ 3).$$

Bemerkung/Definition 21.15. Sei $\sigma \in S_n$ eine Permutation. Dann heißt

$$\operatorname{sign}(\sigma) := \prod_{i<j} \frac{\sigma(j) - \sigma(i)}{j - i} \overset{!}{\in} \{\pm 1\}$$

das **Signum** (oder Vorzeichen) von σ.

Beweis. Es gilt tatsächlich $\operatorname{sign}(\sigma) \in \{\pm 1\}$, da jeder Faktor $j - i$ des Nenners bis auf Vorzeichen auch im Zähler vorkommt: Schreiben wir nämlich

$$\tilde{j} = \sigma^{-1}(j), \quad \tilde{i} = \sigma^{-1}(i),$$

dann ist

$$\sigma(\tilde{j}) - \sigma(\tilde{i}) = j - i.$$

Ist $\tilde{j} > \tilde{i}$, so ist das Vorzeichen $+$, gilt $\tilde{j} < \tilde{i}$, dann ist es $-$. □

Satz 21.16. *Seien $\sigma, \tau \in S_n$. Dann gilt:*

1. $\operatorname{sign}(\sigma \circ \tau) = \operatorname{sign}\sigma \cdot \operatorname{sign}\tau$,
2. $\operatorname{sign}(\sigma) = (-1)^a$,

wobei a die Anzahl der Transpositionen in einer beliebigen Zerlegung von σ in ein Produkt von Transposition ist.

Beweis. 1. Es gilt:

$$\begin{aligned}
\operatorname{sign}(\sigma \circ \tau) &= \prod_{i<j} \frac{(\sigma \circ \tau)(j) - (\sigma \circ \tau)(i)}{j - i} \\
&= \prod_{i<j} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{\tau(j) - \tau(i)} \cdot \frac{\tau(j) - \tau(i)}{j - i} \\
&= \left[\prod_{i<j} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{\tau(j) - \tau(i)}\right]\left[\prod_{i<j} \frac{\tau(j) - \tau(i)}{j - i}\right] \\
&= \operatorname{sign}(\sigma) \cdot \operatorname{sign}(\tau),
\end{aligned}$$

da mit $\{i, j\}$ auch $\{\tau(i), \tau(j)\}$ alle 2-elementigen Teilmengen von $\{1, \dots, n\}$ durchläuft und

$$\frac{\sigma(\tau(j)) - \sigma(\tau(i))}{\tau(j) - \tau(i)} = \frac{\sigma(\tau(i)) - \sigma(\tau(j))}{\tau(i) - \tau(j)}.$$

2. Wegen 1. reicht es,

$$\operatorname{sign}(\tau) = -1$$

zu zeigen für eine Transposition $\tau = (k\ l),\ k < l$. Es gilt:

$$\begin{aligned}
\operatorname{sign}(\tau) &= \prod_{i<j} \frac{\tau(j) - \tau(i)}{j - i} \\
&= \prod_{i<j,\ i,j\neq k,l} \frac{j-i}{j-i} \prod_{i<j,\ i=k,j\neq l} \frac{j-l}{j-k} \prod_{i<j,\ j=k,i\neq l} \frac{l-i}{k-i} \\
&\quad \prod_{i<j,\ j=l,i\neq k} \frac{k-i}{l-i} \prod_{i<j,\ i=l,j\neq k} \frac{j-k}{j-l} \prod_{i<j,\ k=i,j=l} \frac{k-l}{l-k} \\
&= -1,
\end{aligned}$$

da das erste Produkt 1 ist, das zweite sich mit dem fünften wegkürzt, das dritte sich mit dem vierten wegkürzt und das letzte Produkt aus einem einzigen Faktor -1 besteht.

□

21.3 Gruppenhomomorphismen

Definition 21.17. *Ein **Gruppenhomomorphismus***

$$\varphi : G \to H$$

ist eine Abbildung zwischen Gruppen, die

$$\varphi(a \circ_G b) = \varphi(a) \circ_H \varphi(b)$$

$\forall a, b \in G$ *erfüllt.*

*Einen injektiven, surjektiven bzw. bijektiven Gruppenhomomorphismus nennt man auch (Gruppen–)**Epi-**, **Mono-** und bzw. **Isomorphismus**.*

Bemerkung 21.18. Sei $\varphi\colon G \to H$ ein Gruppenhomomorphismus. Dann gilt:

$$\varphi(e_G) = e_H.$$

Dies beweist man genauso wie für Homomorphismen von Vektorräumen, siehe Bem. 20.3.

Beispiel 21.19. 1. Jeder Vektorraumhomomorphismus $f\colon V \to W$ ist auch ein Gruppenhomomorphismus für die zugehörigen additiven Gruppen:

$$f\colon (V,+) \to (W,+).$$

2. $\varphi\colon S_4 \to S_3$, $\begin{pmatrix}1&2&3&4\\a&b&c&d\end{pmatrix} \mapsto \begin{pmatrix}1&2&3\\a&b&c\end{pmatrix}$ ist ein Gruppenhomomorphismus.
3. Das Exponenzieren

$$\exp\colon (\mathbb{R},+) \to (\mathbb{R}_{>0},\cdot),\ x \mapsto \exp(x) = e^x$$

ist ein Gruppenhomomorphismus, denn $e^{x+y} = e^x \cdot e^y$.

4. sign$\colon S_n \to \{\pm 1\}$ ist ein Gruppenhomomorphismus. Dies folgt direkt aus Satz 21.16.

Bemerkung/Definition 21.20. Ist $\varphi\colon G \to H$ ein Isomorphismus, dann ist auch $\varphi^{-1}\colon H \to G$ ein Gruppenhomomorphismus, also auch ein Isomorphismus. Schreibweise: $G \cong H$.

Beweis. Es gilt: $\forall\, h_1, h_2 \in H : \varphi^{-1}(h_1 \circ h_2) = \varphi^{-1}(h_1) \cdot \varphi^{-1}(h_2)$, denn:

$$\begin{aligned}\varphi(\varphi^{-1}(h_1)\cdot\varphi^{-1}(h_2)) &\overset{\varphi \text{ Homom.}}{=} \varphi(\varphi^{-1}(h_1))\cdot\varphi(\varphi^{-1}(h_2))\\ &= h_1 \cdot h_2\end{aligned}$$

$\Rightarrow \varphi^{-1}(h_1 \cdot h_2) = \varphi^{-1}(h_1)\cdot\varphi^{-1}(h_2)$, da φ bijektiv ist. □

Beispiel 21.21. Die Gruppe S_4 der Permutation auf 4 Buchstaben ≅ Symmetriegruppe des Tetraeders. Dies ist nicht sehr schwierig zu überprüfen; die Transpositionen entsprechen dabei den Spiegelungen an einer Symmetrieebene des Tetraeders durch zwei der Ecken.

Analog zu Vektorräumen können wir auch Teilmengen von Gruppen betrachten, die wieder Gruppen sind:

Definition 21.22. *Eine nichtleere Teilmenge $U \subseteq G$ heißt **Untergruppe**, wenn:*

UG1: $a, b \in U \Rightarrow a \circ b \in U$,

UG2: $a \in U \Rightarrow a^{-1} \in U$.

Bemerkung 21.23. Sei U Untergruppe. Dann ist U eine Gruppe mit neutralem Element $e_U = e_G \in U$.

Beweis. $U \neq \emptyset \Rightarrow \exists a \in U \overset{\text{UG2}}{\Rightarrow} a^{-1} \in U \overset{\text{UG1}}{\Rightarrow} a \cdot a^{-1} = e \in U$. □

Bemerkung 21.24. Die beiden Bedingungen UG1 und UG2 für eine Untergruppe sind äquivalent zu:

UG': $a, b \in U \Rightarrow ab^{-1} \in U$.

Beweis. UG1, UG2 $\Rightarrow$ UG' ist klar. Wir zeigen also:

UG' $\Rightarrow$ UG2: $U \neq \emptyset \Rightarrow \exists b \in U \overset{UG'}{\Rightarrow} e = bb^{-1} \in U \Rightarrow eb^{-1} = b^{-1} \in U$.

UG' $\Rightarrow$ UG1: Seien nun $a, b \in U \Rightarrow a, b^{-1} \in U \Rightarrow ab = a(b^{-1})^{-1} \in U$. □

Bemerkung 21.25. Sei $\varphi\colon G \to H$ ein Gruppenhomomorphismus. Dann ist sein Kern $\operatorname{Ker}\varphi := \{g \in G \mid \varphi(g) = e_H\} \subseteq G$ eine Untergruppe.

Beweis. Zunächst zeigen wir die Abgeschlossenheit: $a, b \in \operatorname{Ker}\varphi \Rightarrow \varphi(a) = e, \varphi(b) = e \Rightarrow \varphi(a \cdot b) = \varphi(a) \cdot \varphi(b) = e \cdot e = e \Rightarrow a \cdot b \in \operatorname{Ker}\varphi$.

Um nun noch zu beweisen, dass für jedes a auch $a^{-1} \in \operatorname{Ker}\varphi$, zeigen wir zunächst folgende Eigenschaft:

Lemma 21.26. *Sei $\varphi\colon G \to H$ ein Gruppenhomomorphismus. Dann gilt:*

$$H \ni (\varphi(a))^{-1} = \varphi(a^{-1}).$$

Beweis. Zu zeigen ist:

$$\varphi(a) \cdot \varphi(a^{-1}) \overset{!}{=} e,$$

da $(\varphi(a))^{-1} \in H$ eindeutig bestimmt ist. Es gilt:

$$\varphi(a) \cdot \varphi(a^{-1}) = \varphi(a \cdot a^{-1}) = \varphi(e_G) = e,$$

was zu zeigen war. □

Dies können wir nun benutzen:

$$a \in \operatorname{Ker}\varphi \Rightarrow \varphi(a) = e \Rightarrow \varphi(a^{-1}) = (\varphi(a))^{-1} = e^{-1} = e \Rightarrow a^{-1} \in \operatorname{Ker}\varphi$$

$\Rightarrow \operatorname{Ker}\varphi \subseteq G$ ist eine Untergruppe. □

Bemerkung 21.27. Sei $\varphi\colon G \to H$ ein Gruppenhomomorphismus. Dann ist $\operatorname{Bild}\varphi = \varphi(G) \subseteq H$ auch eine Untergruppe.

Beweis. $\varphi(a) \cdot \varphi(b) = \varphi(ab)$ liegt im Bild, $(\varphi(a))^{-1} = \varphi(a^{-1})$ ebenfalls. □

Beispiel 21.28. Die Gruppe

$$A_n := \operatorname{Ker}(\operatorname{sign}\colon S_n \to \{\pm 1\})$$

heißt **alternierende (Unter)gruppe** (von S_n).

Beispielsweise operiert die S_3 auf einem regelmäßigen Dreieck durch Vertauschung der Punkte (es gibt 6 solcher Vertauschungen). Die A_3 besteht nur aus den Vertauschungen, die zyklisch alle drei Ecken vertauschen (s. Abb. 21.3).

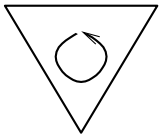


Abbildung 21.3. Die Gruppe A_3 operiert auf dem gleichseitigen Dreieck.

21.4 Gruppenoperationen

Definition 21.29. *Eine **Operation** einer Gruppe G auf einer Menge M ist eine Abbildung*

$$G \times M \to M, \quad (g, m) \mapsto g.m,$$

die den Regeln

O1: $g.(h.m) = (g \cdot h).m \quad \forall g, h \in G \quad \forall m \in M,$

O2: $e.m = m \quad \forall m \in M$

genügt.

Beispiel 21.30.

- S_n operiert auf $\{1, \ldots, n\}$:

$$S_n \times \{1, \ldots, n\} \to \{1, \ldots, n\}, \quad (\sigma, i) \mapsto \sigma(i).$$

Bemerkung 21.31. Sei $G \times M \to M$ eine Operation. Dann ist

$$G \to \mathrm{Bij}(M), \quad g \mapsto (g \colon M \to M, \quad m \mapsto g.m)$$

ein Gruppenhomomorphismus.

Definition 21.32. *Seien $G \times M \to M$ eine Gruppenoperation und $m \in M$. Dann heißt die Menge*

$$Gm := G.m := \{g.m \mid g \in G\}$$

*die **Bahn** von m (unter G).*

Beispiel 21.33. 1. Die Gruppe

$$\mathrm{SO}(2) = \left\{ \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \quad \alpha \in [0, 2\pi) \right\}$$

operiert auf $\mathbb{R}^2$. Ihre Bahnen sind konzentrische Kreislinien (s. Abb. 21.4).

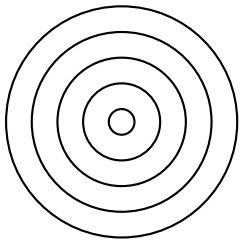


Abbildung 21.4. Die Bahnen der Operation von SO(2) auf $\mathbb{R}^2$.

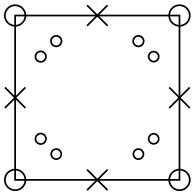


Abbildung 21.5. Einige Bahnen der Operation der D_8.

2. Die Symmetriegruppe D_8 des Quadrats operiert auf dem Quadrat. Für einige Bahnen s. Abb. 21.5; offenbar sind also nicht unbedingt alle Bahnen gleich lang, d.h. nicht alle haben gleich viele Elemente.

Definition 21.34. *Sei $G \times M \to M$ eine Gruppenoperation und $m \in M$. Dann heißt*

$$\mathrm{Stab}(m) = \{g \in G \mid g.m = m\}$$

*der **Stabilisator** von m.*

Beispiel 21.35. Sei T die Symmetriegruppe des Tetraeders mit Ecken $1, 2, 3, 4$ (s. Abb. 21.6). Dann hat Stab(1) sechs Elemente, nämlich die 3 Drehungen, die die Ecke 1 festlassen sowie die 3 Spiegelungen an Ebenen durch 1 und die Mitte einer der drei gegenüberliegenden Seiten.

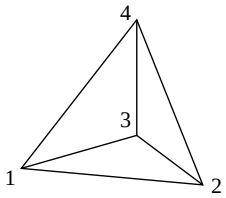


Abbildung 21.6. Die Symmetriegruppe des Tetraeders operiert auf dem Tetraeder.

Bemerkung 21.36. Stab(m) $\subseteq G$ ist eine Untergruppe.

Beweis. Zur Abgeschlossenheit: $a.m = m, b.m = m \Rightarrow (a \cdot b).m = a.(b.m) = a.m = m$. Die Existenz des Inversen ist klar. □

Definition 21.37. *Mit*

$$G\backslash M = \{G.m \mid m \in M\} \subseteq 2^M (= \mathcal{P}(M))$$

bezeichnet man den ***Bahnenraum*** *von G auf M.*

Bei ***Rechtsoperationen*** *(oder* ***Operation von rechts****) $(M \times G \to M, (m, g) \mapsto m.g)$ schreiben wir M/G für den Bahnenraum. Zur Verdeutlichung sagt man für Operation auch manchmal* ***Linksoperation*** *oder* ***Operation von links****.*

Bemerkung 21.38. Je zwei Bahnen Gm_1, Gm_2 sind entweder gleich oder disjunkt. Mit anderen Worten: *In der gleichen Bahn liegen* definiert eine Äquivalenzrelation auf M (siehe dazu Abschnitt 3.1).

Beweis. Angenommen, $Gm_1 \cap Gm_2 \neq \emptyset$. D.h.,

$$\exists\, g_1, g_2 \in G\colon\ g_1 m_1 = g_2 m_2.$$

$$\Rightarrow hm_2 = h(g_2^{-1} g_1 m_1) = (hg_2^{-1} g_1) m_1 \in Gm_1 \ \forall h \in G.$$

Somit gilt: $Gm_2 \subseteq Gm_1$. $Gm_1 \subseteq Gm_2$ folgt genauso. □

21.5 Index– und Bahnenformel

Ein wichtiges Beispiel von Operationen sind solche von Untergruppen auf einer gegebenen Gruppe. Mit ihrer Hilfe werden wir die sogenannte Indexformel und als Folgerung den Satz von Lagrange beweisen.

Beispiel 21.39. Sei $H \subseteq G$ eine Untergruppe. Dann operiert H auf G von links vermöge:

$$H \times G \to G \quad (h, g) \mapsto h.g = hg$$

und von rechts vermöge:

$$G \times H \to G, \quad (g, h) \mapsto g.h = gh.$$

Definition 21.40. *Sei $H \subseteq G$ eine Untergruppe.*

$$H\backslash G := \{Hg \mid g \in G\} \subseteq 2^G$$

heißt Menge der ***Links-Nebenklassen*** *von H in G.*

Entsprechend ist

$$G/H := \{gH \mid g \in G\} \subseteq 2^G$$

die Menge der ***Rechts-Nebenklassen*** *von $H \subset G$.*

Beispiel 21.41. H selbst ist die Nebenklasse: $H = eH = He$.

Definition 21.42. *Sei G eine Gruppe. Dann ist*

$$\operatorname{ord}(G) := |G| := \begin{cases} n \in \mathbb{N}, & \text{falls } G \text{ genau } n \text{ Elemente besitzt.} \\ \infty, & \text{sonst.} \end{cases}$$

*die **Ordnung der Gruppe** G. Sei $g \in G$, dann ist $\langle g \rangle := \{g^n \mid n \in \mathbb{Z}\} \subseteq G$ eine Untergruppe (für $n < 0$ ist $g^n = (g^{|n|})^{-1}$, wie üblich). Die **Ordnung des Elementes** g ist*

$$\operatorname{ord} g := \operatorname{ord}\langle g \rangle.$$

Offenbar gilt:

$$\operatorname{ord}(g) = \begin{cases} \infty, & \text{falls } g^n \neq e \quad \forall n \in \mathbb{N}, \\ \min\{n \in \mathbb{N}_\infty \mid g^n = e\}, & \text{sonst.} \end{cases}$$

Beispiel 21.43. 1. Die Ordnungen einiger Permutationen:

$$\begin{aligned} \operatorname{ord}(1\ 2\ 3) &= 3, \\ \operatorname{ord}(1\ 2\ 4)(3\ 5) &= 6, \\ \operatorname{ord}(1\ 2)(3\ 4) &= 2. \end{aligned}$$

2. Die Ordnung einer Spiegelung an einer Hyperebene ist 2.
3. In $\mathbb{Z}/10\mathbb{Z}$ gilt: $\operatorname{ord}(1) = \operatorname{ord}(3) = 10$, $\operatorname{ord}(2) = 5$, $\operatorname{ord}(5) = 2$.
4. Die Ordnung der 1 als Element von $(\mathbb{Z}, +)$ ist ∞. Hier sieht man auch, dass bei der Definition von $\langle g \rangle$ tatsächlich $n \in \mathbb{Z}$ und nicht $n \in \mathbb{N}$ verwendet werden sollte, da nämlich $\{n \cdot 1 \mid n \in \mathbb{N}\}$ keine Gruppe ist.

Satz/Definition 21.44. *Sei $H \subseteq G$ eine Untergruppe. Dann bezeichnet*

$$[G : H] := |G/H| = |H\backslash G|$$

*den **Index** von H und G.*

Beweis. Zu zeigen ist, dass es genauso viele Links– wie Rechtsnebenklassen gibt. Dies ist vollständig analog zum Beweis des folgenden Satzes. □

Satz 21.45 (Indexformel). *Sei $H \subseteq G$ eine Untergruppe. Dann gilt:*

$$|G| = [G : H] \cdot |H|.$$

Beweis. Je zwei Nebenklassen g_1H, g_2H haben gleich viele Elemente, denn

$$g_1H \to g_2H, \quad x \mapsto g_2g_1^{-1}x$$

ist eine Bijektion, denn die Umkehrabbildung ist die Multiplikation mit $g_1g_2^{-1}$. Also, da $G = \dot{\bigcup}_{gH \in G/H} gH$ die disjunkte Vereinigung der Bahnen ist, folgt:

$$\begin{aligned} |G| &= \sum_{gH \in G/H} |gH| \\ &= \sum_{gH \in G/H} |H| \\ &= |G : H| \cdot |H|, \end{aligned}$$

da $[G : H] = |G/H|$. □

Korollar 21.46. *Sei G eine Gruppe. Dann gilt:*

1. *Ist $H \subseteq G$ eine Untergruppe, $|G| < \infty$, dann gilt der **Satz von Lagrange**:*

   $$|H| \textit{ teilt } |G|.$$

   Dies wird auch $|H| \,\big|\, |G|$ geschrieben.
2. $|G| < \infty \Rightarrow \operatorname{ord}(g)$ *teilt* $\operatorname{ord}(G) = |G|$.

Beweis. Die erste Aussage folgt direkt aus dem vorigen Satz, die zweite ebenfalls, weil $\langle g \rangle = \{g^n \mid n \in \mathbb{Z}\} = \{g, g^2, g^3, \ldots, g^{\operatorname{ord} g} = e\}$ für jedes $g \in G$ eine Untergruppe von G ist. □

Beispiel 21.47.

1. A_4 ist eine Untergruppe von S_4 und es gilt:

   $$|S_4| = 24, \quad |A_4| = 12.$$

2. Für jede Primzahl p hat $\mathbb{Z}/p\mathbb{Z}$ nur die trivialen Untergruppen $\{0\}$ und $\mathbb{Z}/p\mathbb{Z}$ selbst.

Lemma 21.48. *Sei $G \times M \to M$ eine Operation und sei $m \in M$ fest gewählt. Dann ist die Abbildung*

$$G/\operatorname{Stab}(m) \to G.m, \quad gH \mapsto g.m$$

eine wohldefinierte Bijektion. D.h. für jedes $m \in M$ gilt $|G/\operatorname{Stab}(m)| = |G.m|$.

Beweis. Wohldefiniert: Wir schreiben $H = \operatorname{Stab}(m)$. Sei $g_1 \in gH$, z.B. $g_1 = gh$.
$\Rightarrow g_1m = ghm = gm$.

Surjektiv: Ist klar.

Injektiv: Angenommen, $g_1 m = g_2 m$. Dann gilt:

$$g_2^{-1} g_1 m = m \Rightarrow g_2^{-1} g_1 \in \mathrm{Stab}(m) = H.$$

Das liefert: $g_2 H = g_2((g_2^{-1} g_1)H) = g_1 H$, was zu zeigen war.

□

Hierbei möchten wir betonen, dass die Wohldefiniertheit einer Abbildung auf der Menge der Nebenklassen natürlich immer nachgewiesen werden muss, analog zu den Abbildungen auf Äquivalenzklassen im Abschnitt 3.1 z.B. die Konstruktion der rationalen Zahlen in Beispiel 3.12. Beispielsweise ist die Abbildung

$$z\colon \mathbb{Q} \to \mathbb{Q},\ \frac{p}{q} \mapsto p$$

erst wohldefiniert, wenn wir Zusätzliches verlangen, z.B. dass der Bruch gekürzt ist und dass $q > 0$ ist. Ansonsten ist beispielsweise nicht klar, was $z(\frac{2}{3})$ ist, weil $\frac{2}{3} = \frac{4}{6}$ und $2 \neq 4$ ist.

Das Lemma liefert direkt die folgende Formel:

Korollar 21.49 (Bahnenformel). *Sei $G \times M \to M$ eine Operation auf einer endlichen Menge. Dann gilt:*

$$|M| = \sum_{Gm \in G\backslash M} |Gm| = \sum_{Gm \in G\backslash M} [G : \mathrm{Stab}(m)].$$

Beweis. $M = \dot{\bigcup} Gm$ ist die disjunkte Vereinigung der Bahnen $Gm \in G\backslash M$ und mit dem Lemma 21.48 erhalten wir:

$$|Gm| = |G/\mathrm{Stab}(m)| = [G : \mathrm{Stab}(m)].$$

□

Beispiel 21.50. Wir betrachten S_4 als Symmetriegruppe des Tetraeders mit Ecken e_1, e_2, e_3, e_4 (s. auch Abb. 21.7):

- $\mathrm{Stab}(e_1) = S_{\mathrm{Dreieck}(e_2,e_3,e_4)} = S_3$,
- $S_4 e_1 = \{e_1, \ldots, e_4\}$,
- $|S_4|/|S_3| = \frac{4!}{3!} = 4 = |S_4 e_1|$,
- $|S_4 m_{12}| = 6$, denn $\mathrm{Stab}(m_{12}) = \{e, (1\ 2), (3\ 4), (1\ 2)(3\ 4)\}$, also:

$$|S_4 m_{12}| = \frac{4!}{|\mathrm{Stab}(m_{12})|} = \frac{24}{4} = 6.$$

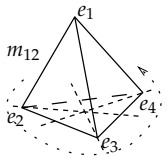


Abbildung 21.7. Die S_3 als Stabilisator einer Ecke des Tetraeders.

21.5.1 Anwendung: Klassifikation von Graphen

Definition 21.51. *Ein **ungerichteter schleifenfreier Graph** ist ein Tupel*

$$G = (V, E),$$

wobei V eine Menge (von Ecken *bzw. **Knoten**, engl. vertex) und $E \subseteq V \times V$ (**Kanten**, engl. edge) symmetrisch und disjunkt von der Diagonalen (d.h. **schleifenfrei**) ist.*

Schleifenfrei heißt ein Graph also, wenn kein Knoten mit sich selbst mit einer Kante verbunden ist.

Beispiel 21.52. Zwei Beispiele ungerichteter schleifenfreier Graphen mit $|V| = 4$ sind in Abb. 21.8 zu sehen. Diese sind **zusammenhängend**, d.h. von jeder Ecke eines Graphen gibt es einen Weg zu jeder anderen Ecke.

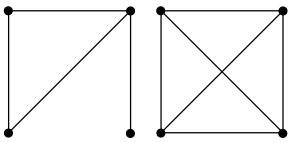


Abbildung 21.8. Zwei Beispiele zusammenhängender Graphen.

Definition 21.53. *Zwei Graphen $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$ heißen **isomorph** ($G_1 \cong G_2$), wenn es eine bijektive Abbildung*

$$\varphi : V_1 \to V_2$$

gibt, die Kanten / Nichtkanten in Kanten / Nichtkanten überführt. Das heißt:

$$(v, w) \in E_1 \Leftrightarrow (\varphi(v), \varphi(w)) \in E_2.$$

Beispiel 21.54. Abb. 21.9 zeigt zwei isomorphe Graphen. Ein Isomorphismus ist gegeben durch die Permutation (1432).

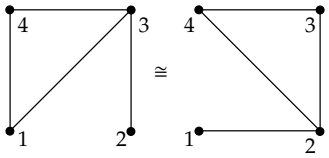


Abbildung 21.9. Zwei isomorphe Graphen.

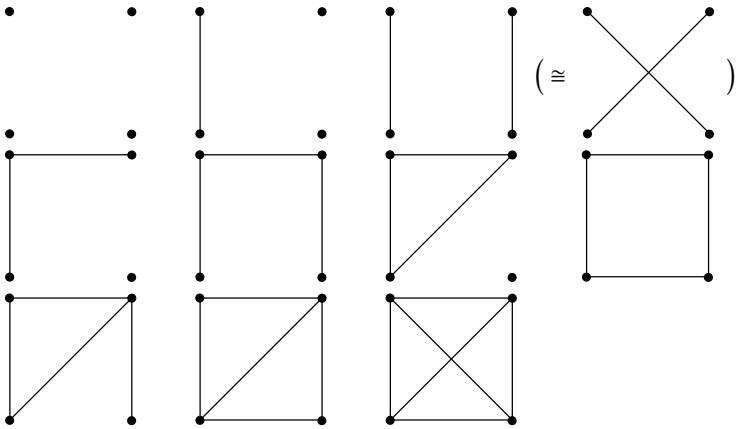


Abbildung 21.10. 10 Graphen mit 4 Knoten.

Beispiel 21.55. Wieviele Isomorphie-Klassen von Graphen mit 4 Knoten gibt es? Ist die Liste in Abb. 21.10 vollständig?

Sei M die Menge der Graphen mit genau 4 Ecken $\{1, \ldots, 4\}$. Es gilt:

$$|M| = 2^6,$$

da es 6 mögliche Kanten gibt: $\{1,2\}, \ldots, \{3,4\}$.

S_4 operiert auf M: $S_4 \times M \to M$. Wir fragen nach $|S_4 \backslash M|$, weil dies ja die Anzahl der verschiedenen Bahnen der Menge aller Graphen ist.

Um die Bahnengleichung überprüfen zu können, müssen wir die Stabilisatoren der obigen Graphen berechnen:

1. $\mathrm{Stab}(G_1) \cong S_4 \Rightarrow |\mathrm{Stab}(G_1) = 24|$ (davon gibt es nach Lemma 21.48 genau $\frac{|S_4|}{|\mathrm{Stab}(G_1)|} = 1$),
2. $\mathrm{Stab}(G_2) \cong \mathbb{Z}_2 \times \mathbb{Z}_2 \Rightarrow |\mathrm{Stab}(G_2)| = 4$ (davon gibt es also $\frac{|S_4|}{|\mathrm{Stab}(G_2)|} = \frac{24}{4} = 6$ Stück),
3. $\mathrm{Stab}(G_3) \cong \langle (12), (34), (13)(24) \rangle \cong D_8 \Rightarrow |\mathrm{Stab}(G_3)| = 8$ (3 Stück),

4. $\mathrm{Stab}(G_4) \cong \mathbb{Z}_2$ (mittlerer und einzelner fest!) (12 Stück),
5. $\mathrm{Stab}(G_5) \cong \mathbb{Z}_2$ (nur vertikale Spiegelung) (12 Stück),
6. $\mathrm{Stab}(G_6) \cong S_3$ (Dreiecke, einzelner fest) (4 Stück),
7. $\mathrm{Stab}(G_7) \cong D_8$ (Symmetriegruppe des Quadrats) (3 Stück),
8. $\mathrm{Stab}(G_8) \cong \mathbb{Z}_2$ (linke beiden vertauschbar) (12 Stück),
9. $\mathrm{Stab}(G_9) \cong \mathbb{Z}_2 \times \mathbb{Z}_2$ (jeweils diagonal gegenüber vertauschbar) (6 Stück),
10. $\mathrm{Stab}(G_{10}) \cong S_4$ (1).

Die Bahnengleichung liefert:

$$2^6 = 64 \stackrel{!}{=} 1 + 6 + 3 + 12 + 12 + 4 + 3 + 12 + 6 + 1 = 60.$$

Die Bahnengleichung ist also nicht erfüllt, d.h. es fehlt mindestens ein Graph. In der Tat, es fehlt der Graph G_{11} in Abb. 21.11. Es gilt: $\mathrm{Stab}(G_{11}) \cong S_3$; davon

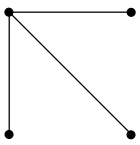


Abbildung 21.11. Der Graph, der in der Liste fehlt.

gibt es also $\frac{24}{6} = 4$ Stück. Damit ist die Bahnengleichung erfüllt.
Es gibt also genau 11 Typen!

Aufgaben

Aufgabe 21.1 (Bewegungen). Sei $f\colon \mathbb{R}^n \to \mathbb{R}^n$ eine Abbildung, für die gilt:

$$\|f(x) - f(y)\| = \|x - y\|.$$

Zeigen Sie: $\exists$ orthogonale Matrix $U \in \mathbb{R}^{n\times n}$ (d.h. $U^{\mathrm{t}} \cdot U = E$) und $\exists\, b \in \mathbb{R}^n$, s.d.:

$$f(x) = Ux + b.$$

Solche Abbildungen heißen **Bewegungen**.

Aufgabe 21.2 (Permutationen). Lässt sich bei dem bekannten Schiebespiel die linke der folgenden Konfigurationen in die Ausgangsstellung (rechts) überführen?

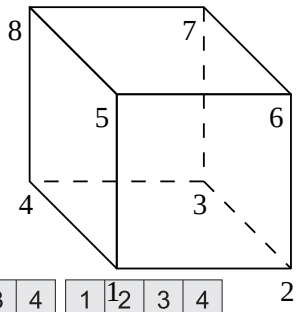
Aufgabe 21.3 (Zykelschreibweise für Permutationen). Geben Sie für die folgenden Permutationen deren Zykel-Schreibweise, Ordnung und Signum an:

$$\sigma_1 = \begin{pmatrix} 1&2&3&4&5&6&7&8 \\ 3&8&5&6&7&4&1&2 \end{pmatrix} \in S_8,$$

$$\sigma_2 = \begin{pmatrix} 1&2&3&4&5&6&7&8 \\ 2&7&4&5&3&6&8&1 \end{pmatrix} \circ \begin{pmatrix} 1&2&3&4&5&6&7&8 \\ 2&1&5&3&4&7&6&8 \end{pmatrix} \in S_8.$$

Aufgabe 21.4 (Permutationsgruppen). Geben Sie für jede der Permutations-Gruppen S_i, $i = 4, 5, 6, 7$, je ein Element $s_i \in S_i$ maximaler Ordnung an.

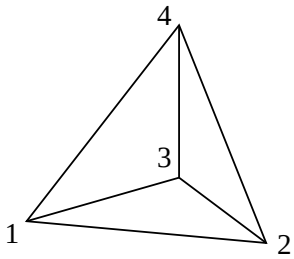
Aufgabe 21.5 (Symmetriegruppen). Welche Ordnung hat die Symmetriegruppe W des Würfels? Beschreiben Sie alle Elemente von W geometrisch.



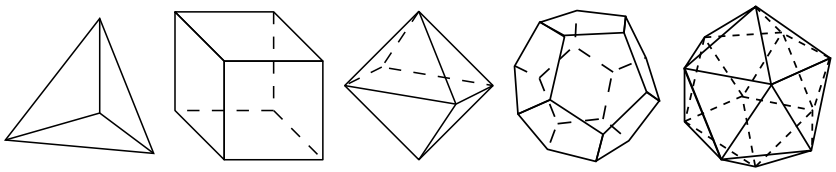
2	1	3	4
5	6	7	8
9	10	11	12
13	14	15	

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

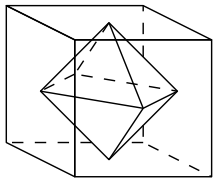
Aufgabe 21.6 (Symmetriegruppen). Bestimmen Sie sämtliche Untergruppen der S_4 mit Hilfe des Tetraeders (Typ der Untergruppe und Anzahl der Untergruppen der gleichen Art).



Aufgabe 21.7 (Symmetriegruppen der Platonischen Körper). Bestimmen Sie die Ordnungen der Symmetriegruppen sämtlicher **Platonischer Körper** (die nach der Anzahl Ihrer Flächen benannt sind):



Tetraeder Würfel Oktaeder Dodekaeder Ikosaeder

Tipp:  . Für Dodekaeder und Ikosaeder gibt es ein ähnliches Bild.

Aufgabe 21.8 (Operation durch Konjugation). Sei G eine Gruppe. Wir definieren durch Konjugation eine Gruppenoperation φ von G auf sich selbst:

$$\varphi\colon G \times G \to G,\ (g,x) \mapsto \varphi(g,x) := g.x := gxg^{-1}.$$

1. Zeigen Sie: Das ist tatsächlich eine Gruppen-Operation.
2. Eine **Konjugationsklasse** ist eine Bahn unter dieser Operation. Eine **Partition** von $n \in \mathbb{N}$ ist eine Darstellung der Form: $n = n_1 + n_2 + \cdots + n_k$ für gewisse $n_i \in \mathbb{N}$ und ein gewisses $k \in \mathbb{N}$ mit: $n_1 \geq n_2 \geq \cdots \geq n_k$.
   Zeigen Sie: Der Zykeltyp (was ist eine sinnvolle Definition?) von Elementen in S_n definiert eine Bijektion zwischen den Konjugations-Klassen von S_n und den Partitionen von n.

22

Determinanten

22.1 Existenz und Eindeutigkeit der Determinante

22.1.1 Motivation

Sei $A \in K^{n\times n}$ eine quadratische Matrix. Wir wollen A ein Element $\det A \in K$ zuordnen. Wir suchen:

$$\det\colon K^{n\times n} \to K$$

mit einigen *netten* Eigenschaften.

Im Fall $K = \mathbb{R}$ beispielsweise hat die Determinante einer Matrix

$$A = (a_{ij}) = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

mit Zeilen $a_1, \ldots, a_n$ eine elementargeometrische Interpretation:

$$\det A = \pm \operatorname{Vol}(\{\lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_i \in [0,1]\}),$$

das **Volumen des Parallelotops** (auch **Parallelepiped** genannt), das durch $a_1, \ldots, a_n$ aufgespannt wird.

Beispiel 22.1. Der allgemeine Begriff des Volumens des Parallelotops im $\mathbb{R}^n$ bedeutet für die uns geläufigen Spezialfälle folgendes: im Falle $n = 3$ das Volumen, im Falle $n = 2$ der Flächeninhalt (s. Abb. 22.1).

Auf diese Weise lässt sich die Determinante mathematisch nicht präzise definieren: Für $n \neq 4$ bräuchten wir zunächst einen Volumenbegriff und für einen beliebigen Körper K (etwa $K = \mathbb{F}_p$), ist es fraglich, ob es eine solche

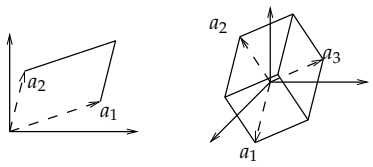


Abbildung 22.1. Parallelotope im $\mathbb{R}^n$, $n = 2, 3$.

Interpretation gibt. Wir können die Interpretation aber benutzen, um Regeln für die Abbildung det zu entdecken.

Für das Lösen linearer Gleichungssysteme wird die Eigenschaft, dass die Determinante genau dann verschieden von 0 ist, wenn die Matrix invertierbar ist, besonders interessant sein.

22.1.2 Definition

Definition 22.2. *Sei K ein Körper, $n \in \mathbb{Z}_{>0}$. Eine Abbildung*

$$\det\colon K^{n\times n} \to K, \quad A = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \mapsto \det A,$$

*häufig auch $|A| := \det A$ geschrieben, heißt **Determinante**, falls folgendes gilt:*

D1) det *ist **linear in jeder Zeile**. Genauer:*

a) Ist $a_i = a_i' + a_i''$, dann gilt

$$\det\begin{pmatrix} \vdots \\ a_i \\ \vdots \end{pmatrix} = \det\begin{pmatrix} \vdots \\ a_i' \\ \vdots \end{pmatrix} + \det\begin{pmatrix} \vdots \\ a_i'' \\ \vdots \end{pmatrix},$$

wobei die $\vdots$ andeuten, dass diese Zeilen bei allen drei Vektoren übereinstimmen.

b) Für jedes $\lambda \in K$ gilt:

$$\det\begin{pmatrix} \vdots \\ \lambda a_i \\ \vdots \end{pmatrix} = \lambda \det\begin{pmatrix} \vdots \\ a_i \\ \vdots \end{pmatrix}.$$

D2) det *ist* ***alternierend****, d.h.*

$$\det A = 0,$$

falls A zwei gleiche Zeilen hat.

D3) det *ist* ***normiert****, d.h. für die Einheitsmatrix $E_n \in K^{n\times n}$ gilt:*

$$\det E_n = 1.$$

Ziel dieses Paragraphen ist es zu zeigen, dass eine Determinantenabbildung

$$\det : K^{n\times n} \to K$$

existiert und dass diese außerdem eindeutig bestimmt ist.

Zunächst die Motivation für diese Forderungen, die in der Definition der Determinante an die Abbildung det gestellt werden:

Zu D1): Im Fall $n = 2$, $K = \mathbb{R}$, sieht das folgendermaßen aus:

a)
$$\det\begin{pmatrix} a_1 \\ a_2' + a_2'' \end{pmatrix} = \det\begin{pmatrix} a_1 \\ a_2' \end{pmatrix} + \det\begin{pmatrix} a_1 \\ a_2'' \end{pmatrix}.$$

In der anschaulichen Interpretation der Determinante als Volumen (d.h. hier für $n = 2$ als Flächeninhalt), lässt sich dies folgendermaßen umformulieren: Die Fläche des Parallelogramms, das von a_1 und $a_2' + a_2''$ aufgespannt wird, ist genauso groß wie die Summe der anderen beiden Flächeninhalte, weil das rechte eingefärbte Dreieck in Abb. 22.2 auf das linke verschoben werden kann. Alternativ könnte man die Tatsache benutzen, dass die Fläche zweier Parallelogramme mit gleicher Grundseite und gleicher Höhe die gleiche Fläche haben.

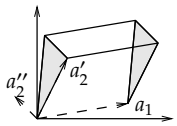


Abbildung 22.2. Illustration zur Determinanten-Eigenschaft D1a) für $n = 2$. Das Bild ist eine 2-dimensionale Veranschaulichung, auch wenn es 3-dimensional erscheinen mag.

b) Genauso lässt sich D1b) im $\mathbb{R}^2$ verstehen:

$$\det\begin{pmatrix} a_1 \\ \lambda \cdot a_2 \end{pmatrix} = \lambda \cdot \det\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

Anschaulich heißt dies: Streckt man einen der beiden Vektoren um das λ-fache, so vergrößert sich der Flächeninhalt ebenso um das λ-fache (s. Abb. 22.3).

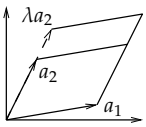


Abbildung 22.3. Illustration zur Determinanten-Eigenschaft D1b) für $n = 2$.

Zu D2): Wieder veranschaulichen wir uns dies im Fall $n = 2$, d.h. für die beiden Zeilen a_1 und a_2 der Matrix gilt $a_1 = a_2$, also:

$$\det\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \det\begin{pmatrix} a_1 \\ a_1 \end{pmatrix} = 0,$$

weil das Parallelogramm entartet ist und also gar keinen Flächeninhalt hat (s. Abb. 22.4).

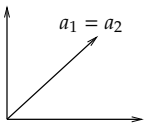


Abbildung 22.4. Ein entartetes Parallelogramm hat keinen Flächeninhalt.

Zu D3): Dies ist lediglich eine Frage der Konvention. Sicherlich ist es aber vernünftig, dem **Einheitsquadrat** ($n = 2$) bzw. dem **Einheitswürfel** ($n = 3$), jeweils mit Seitenlängen 1, das Volumen 1 zu geben.

22.1.3 Der Determinanten–Satz

Im folgenden Satz werden erste Eigenschaften der Determinante zusammengefasst. Im weiteren Verlauf dieses Kapitels werden wir zwar noch weitere kennen lernen, doch hier sind auch schon einige sehr wesentliche dabei, beispielsweise der bereits erwähnte Zusammenhang zwischen Determinanten und Invertierbarkeit von Matrizen.

Satz 22.3 (Determinanten–Satz). *Eine Determinante* $\det : K^{n\times n} \to K$ *hat folgende weitere Eigenschaften:*

D4) Für jedes $\lambda \in K$ *gilt:*

$$\det(\lambda \cdot A) = \lambda^n \cdot \det A.$$

Dies folgt sofort aus D1b), da wir n Zeilen mit einem Faktor λ *in* λA *haben.*

D5) Gibt es ein i mit $a_i = (0, \dots, 0)$*, dann ist (wegen D1b)):*

$$\det A = 0.$$

D6) Wenn B aus A durch Vertauschung von genau 2 *Zeilen entsteht, dann gilt:* $\det B = -\det A$*. Anders gesagt:*

$$\det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} = -\det \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ a_i \\ \vdots \end{pmatrix}.$$

D7) Ist $\lambda \in K$ *und entsteht B aus A durch Addition des* λ*-fachen der i–ten Zeile zur j–ten, dann ist* $\det B = \det A$*:*

$$\det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} = \det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j + \lambda a_i \\ \vdots \end{pmatrix}.$$

D8) Es sei $\sigma \in S_n$ *eine Permutation.* $e_1, \dots, e_n \in K^n$ *bezeichne die kanonischen Basisvektoren von* K^n *(als Zeilenvektoren). Dann gilt für die Determinante der sogenannten* ***Permutationsmatrizen****:*

$$\det \begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(n)} \end{pmatrix} = \operatorname{sign}(\sigma).$$

Mit der Notation $\mathrm{SO}(n) = \{A \in O(n) \mid \det(A) = 1\}$ *gilt daher:*

$$\begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(n)} \end{pmatrix} \in \mathrm{SO}(n) \;\Leftrightarrow\; \sigma \in A_n.$$

D9) Ist A eine ***obere Dreiecksmatrix****, also*

$$A = \begin{pmatrix} \lambda_1 & \cdots & \cdots & * \\ 0 & \lambda_2 & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix},$$

so ist $\det A = \lambda_1 \cdots \lambda_n$.

D10) Äquivalent sind:

1. $\det A \neq 0$,
2. $A \in GL(n, K)$,
3. *Die Zeilen* $a_1, \ldots, a_n \in K^n$ *von* A *sind linear unabhängig.*

Beweis.

D4)

$$\det \lambda A = \det \begin{pmatrix} \lambda \cdot a_1 \\ \vdots \\ \lambda \cdot a_n \end{pmatrix} \overset{D1b)}{=} \lambda^n \cdot \det \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \lambda^n \cdot \det A.$$

D5)

$$\det \begin{pmatrix} a_1 \\ \vdots \\ 0 \\ \vdots \\ a_n \end{pmatrix} = \det \begin{pmatrix} a_1 \\ \vdots \\ 0 \cdot a_k \\ \vdots \\ a_n \end{pmatrix} \overset{D1b)}{=} 0 \cdot \det \begin{pmatrix} a_1 \\ \vdots \\ a_k \\ \vdots \\ a_n \end{pmatrix} = 0 \in K.$$

D6)

$$0 \overset{D2)}{=} \det \begin{pmatrix} \vdots \\ a_i + a_j \\ \vdots \\ a_i + a_j \\ \vdots \end{pmatrix} \overset{D1a)}{=} \underbrace{\det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_i \\ \vdots \end{pmatrix}}_{=0 \ \text{(D2)}} + \det \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ a_i \\ \vdots \end{pmatrix} + \det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} + \underbrace{\det \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ a_j \\ \vdots \end{pmatrix}}_{=0 \ \text{(D2)}}$$

$$\Rightarrow \det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} + \det \begin{pmatrix} \vdots \\ a_j \\ \vdots \\ a_i \\ \vdots \end{pmatrix} = 0 \Rightarrow D6).$$

D7)

$$\det \begin{pmatrix} a_i \\ a_j + \lambda a_i \end{pmatrix} \overset{D1)}{=} \det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} + \lambda \cdot \det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_i \\ \vdots \end{pmatrix} \overset{D2)}{=} \det \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix}.$$

D8) Ist τ eine Transposition und σ eine Permutation, dann geht die Matrix $\begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(n)} \end{pmatrix}$ durch eine Vertauschung von genau zwei Zeilen in $\begin{pmatrix} e_{\tau\sigma(1)} \\ \vdots \\ e_{\tau\sigma(n)} \end{pmatrix}$ über.

Daraus folgt: $\det \begin{pmatrix} e_{\tau\sigma(1)} \\ \vdots \\ e_{\tau\sigma(n)} \end{pmatrix} = -\det \begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(n)} \end{pmatrix}$, nach D6).

$$\Rightarrow \det \begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(n)} \end{pmatrix} = (-1)^k = \operatorname{sign} \sigma,$$

falls $\sigma = \tau_1 \cdots \tau_k$, eine Komposition von k Transpositionen ist (siehe dazu auch Satz 21.16).

Bemerkung 22.4. Nach D6) und D7) können wir elementare Zeilenumformungen vom Typ III und IV in A vornehmen und erhalten eine Matrix $\tilde{A}$ mit

$$\det A = (-1)^k \det \tilde{A},$$

wobei k die Anzahl der Vertauschungen ist.

Die Aussage über SO(n) folgt, da die Orthogonalität der betrachteten Matrizen mit Zeilen e_i einfach einzusehen ist.

D9) Sei $\lambda_i = 0$ für ein $i \in \{1, 2, \ldots, n\}$. Durch elementare Zeilenumformungen vom Typ III und IV kann man A in eine Matrix $\tilde{A}$ überführen, die in Zeilenstufenform ist. Deren letzte Zeile ist eine Nullzeile, so dass $\det \tilde{A} = 0$ (nach D5). Andererseits ist nach D6) und D7):

$$\det A = \pm \det \tilde{A}.$$

Also ist $\det A = 0$ und die Behauptung ist im Fall, dass ein $\lambda_i = 0$ ist, bewiesen.

Wir müssen D9) nun noch für den Fall, dass $\lambda_i \neq 0 \; \forall \; i \in \{1, 2, \ldots, n\}$ gilt, zeigen. Hat A diese Eigenschaft, so gilt nach D1b):

$$\det A = \lambda_1 \cdots \lambda_n \cdot (\det B),$$

wobei B von der Form

$$B = \begin{pmatrix} 1 & & * \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

ist, also eine obere Dreiecksmatrix mit allen Diagonaleinträgen gleich 1. Da man B durch Zeilenumformungen vom Typ III in die Einheitsmatrix überführen kann, ist

$$\det B = \det E_n = 1.$$

Daraus folgt die Behauptung.

Da jede $n \times n$–Matrix in Zeilenstufenform auch eine obere Dreiecksmatrix ist, liefert D9) zusammen mit Bemerkung 22.4 (also D6) und D7)) einen Algorithmus, um die Determinante einer beliebigen $n \times n$–Matrix zu bestimmen.

Algorithmus 22.5 (Gauß-Algorithmus für Determinanten).

Input: $A \in K^{n \times n}$

Output: $\det(A)$

1) Bringe A durch Zeilenoperationen vom Typ III und IV auf Zeilenstufenform $\tilde{A}$. Sei k die Anzahl der Typ IV-Umformungen (d.h. die Anzahl der Zeilenvertauschungen).

2) Berechne d, das Produkt der Diagonaleinträge von $\tilde{A}$.

3) $\det A = (-1)^k \cdot d$.

Die Laufzeit dieses Algorithmus ist im Wesentlichen die des Gauß-Algorithmus selbst: $O(n^3)$ (Schritt 1 ist dominierend).

Beispiel 22.6. Wir berechnen zwei Determinanten auf diese Weise:

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} \overset{D7)}{=} \begin{vmatrix} 1 & 2 \\ 0 & -2 \end{vmatrix} \overset{D9)}{=} 1 \cdot (-2) = -2.$$

$$\begin{vmatrix} 0 & 1 & 2 \\ 3 & 2 & 1 \\ 1 & 2 & 0 \end{vmatrix} \overset{IV}{=} - \begin{vmatrix} 1 & 2 & 0 \\ 3 & 2 & 1 \\ 0 & 1 & 2 \end{vmatrix} \overset{III}{=} - \begin{vmatrix} 1 & 2 & 0 \\ 0 & -4 & 1 \\ 0 & 1 & 2 \end{vmatrix} \overset{IV}{=} \begin{vmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & -4 & 1 \end{vmatrix} \overset{III}{=} \begin{vmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 9 \end{vmatrix} = 9.$$

Achtung! Hierbei ist darauf zu achten, dass man natürlich nicht eine Zeile selbst vervielfachen darf, weil dies die Determinante ändert, z.B. bei einer Einheitsmatrix, bei der man eine Zeile vervielfacht.

D10) Wieder können wir durch elementare Zeilenumformungen vom Typ III und IV die $n \times n$–Matrix A in eine Matrix $\tilde{A}$ überführen, die Zeilenstufenform hat. Diese ist dann eine obere Dreiecksmatrix. Nach D9) ist deren Determinante genau dann $\neq 0$, wenn alle Diagonaleinträge $\neq 0$ sind, d.h. wenn es genau n Stufen gibt.

In Abschnitt 20.7 haben wir gesehen, dass eine quadratische $n \times n$–Matrix genau dann invertierbar ist, wenn sie genau n Stufen hat. Daher sind 1. und 2. also äquivalent.

Die Äquivalenz zur dritten Aussage folgt, da die Zeilen von $\tilde{A}$ genau dann linear unabhängig sind, wenn alle Diagonaleinträge $\neq 0$ sind und da die Zeilen von $\tilde{A}$ genau dann linear unabhängig sind, wenn es die Zeilen von A sind.

□

Korollar 22.7 (Eindeutigkeit der Determinante). *Durch die Bedingungen D1), D2) und D3) ist*

$$\det\colon K^{n\times n} \to K$$

eindeutig festgelegt.

Beweis. Wir können A durch elementare Zeilenumformungen in Zeilenstufenform bringen, was nur das Vorzeichen eventuell ändert. Mit D9) folgt die Behauptung. □

Satz 22.8 (Formel für die Determinante). *Ist K ein Körper und $n \in \mathbb{N}$, dann gibt es genau eine Abbildung*

$$\det\colon K^{n\times n} \to K,$$

die die Bedingungen D1-D3 erfüllt. Nämlich für $A = (a_{ij})$ gilt:

$$\det A = \sum_{\sigma\in S_n} \operatorname{sign}(\sigma)\cdot a_{1\sigma(1)}\cdots a_{n\sigma(n)}.$$

Beweis. Die Eindeutigkeit haben wir bereits eben in Korollar 22.7 gesehen. Wir zeigen nun zunächst, dass, falls eine solche Abbildung existiert, diese die angegebene Formel erfüllen muss. Auch dies folgt aus D1-D3: Schreiben wir nämlich für $A = (a_{ij}) \in K^{n\times n}$ die i–te Zeile als Linearkombination der Standard–Basis–Vektoren (als Zeilenvektoren!), nämlich $a_i = a_{i1}e_1 + \cdots + a_{in}e_n$, so ergibt die Regel D1 (Linearität in den Zeilen):

$$\begin{aligned}\det\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} &= \sum_{j_1=1}^{n} a_{1j_1}\cdot \det\begin{pmatrix} e_{j_1} \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \\ &= \cdots = \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\cdots\sum_{j_n=1}^{n} a_{1j_1}\cdots a_{nj_n}\cdot \det\begin{pmatrix} e_{j_1} \\ \vdots \\ e_{j_n} \end{pmatrix}.\end{aligned}$$

Wenn die Abbildung

$$j\colon \{1,\dots,n\} \to \{1,\dots,n\}, \quad i \mapsto j_i$$

nicht injektiv ist, dann gilt

$$\det\begin{pmatrix} e_{j_1} \\ \vdots \\ e_{j_n} \end{pmatrix} = 0,$$

weil zwei gleiche Zeilen vorkommen (D2). Von den urspünglich n^n Summanden sind also höchstens jene $n!$ von Null verschieden, die zu bijektiven Abbildungen j gehören, d.h. zu Permutationen aus S_n. Wir erhalten:

$$\det A = \sum_{\sigma \in S_n} a_{1\sigma(1)} \cdots a_{n\sigma(n)} \cdot \det \begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(n)} \end{pmatrix}.$$

D8) liefert nun:

$$\det A = \sum_{\sigma \in S_n} \operatorname{sign} \sigma \cdot a_{1\sigma(1)} \cdots a_{n\sigma(n)}.$$

Es bleibt noch die Existenz zu zeigen; dazu gleich (Seite 309). □

Die Formel ist meist nur für sehr kleine n zur tatsächlichen Berechnung einer Determinante nützlich, denn die Summe besteht aus $n!$ Summanden:

Beispiel 22.9.

n=1:

$$\det A = \det\left(a_{11}\right) = a_{11}.$$

n=2:

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11} \cdot a_{22} - a_{12} \cdot a_{21}.$$

n=3: Die **Regel von Sarrus** besagt:

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{array}{r} a_{11} \cdot a_{22} \cdot a_{33} \\ + a_{12} \cdot a_{23} \cdot a_{31} \\ + a_{13} \cdot a_{21} \cdot a_{32} \\ - a_{31} \cdot a_{22} \cdot a_{13} \\ - a_{32} \cdot a_{23} \cdot a_{11} \\ - a_{33} \cdot a_{21} \cdot a_{12}. \end{array}$$

Die Formel aus dem vorigen Satz hat $|S_n| = n!$ Summanden. Für $n = 4$ also 24. Man könnte auf die Idee kommen, für $n \neq 3$ die Regel von Sarrus auch anzuwenden; diese gilt aber nur für $n = 3$ (für $n = 4$ würde die falsch angewandte Regel von Sarrus nur 8 Summanden liefern, was natürlich vorteilhaft wäre).

Die Formel liefert einen Algorithmus in $O(n!)$ Körperoperationen Aufwand. Der Gaußalgorithmus braucht nur $O(n^3)$; man wird die Formel also nur in den seltensten Fällen zum konkreten Berechnen einer Determinante verwenden. Für theoretische Zwecke ist die Formel allerdings des öfteren hilfreich.

Beweis (Beweis der Existenz in Satz 22.8). Um die Existenz einzusehen, zeigen wir, dass die durch die Formel definierte Abbildung $\det\colon K^{n\times n} \to K$ tatsächlich die Bedingungen D1, D2 und D3 (s. Seite 300) erfüllt:

D1a) Sei $a_i = a'_i + a''_i$, d.h. $a_{ij} = a'_{ij} + a''_{ij}$. Es folgt:

$$\begin{aligned}\det\begin{pmatrix} a_1 \\ \vdots \\ a'_i + a''_i \\ \vdots \\ a_n \end{pmatrix} &= \sum_{\sigma\in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \cdots (a'_{i\sigma(i)} + a''_{i\sigma(i)}) \cdots a_{n\sigma(n)} \\ &= \sum_{\sigma\in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \cdots a'_{i\sigma(i)} \cdots a_{n\sigma(n)} + \sum \cdots a''_{i\sigma(i)} \cdots \\ &= \det\begin{pmatrix} \vdots \\ a'_i \\ \vdots \end{pmatrix} + \det\begin{pmatrix} \vdots \\ a''_i \\ \vdots \end{pmatrix}.\end{aligned}$$

D1b) λ zieht sich aus dem i–ten Faktor in jedem Summanden heraus.

D2) Angenommen die k-te und l-te Zeile von A sind gleich, $k \neq l$. Wir setzen: $\tau := (k\ l) \in S_n$. Dann ist:

$$S_n = A_n \dot\cup A_n\tau,$$

da $|A_n| = \frac{n!}{2} = |A_n\tau|$ und da die Vereinigung disjunkt ist. Wenn σ die Gruppe A_n durchläuft, so durchläuft $\sigma \circ \tau$ die Menge $A_n\tau$. Also gilt:

$$(**) \qquad \det A = \sum_{\sigma\in A_n} a_{1\sigma(1)} \cdots a_{n\sigma(n)} - \sum_{\sigma\in A_n} a_{1\sigma(\tau(1))} \cdots a_{n\sigma(\tau(n))}.$$

Da die k–te und l–te Zeile von A gleich sind und da außerdem die Multiplikation in K kommutativ ist, können wir die Summanden der rechten Seite nach Definition von τ umformen:

$$\begin{aligned} a_{1\sigma(\tau(1))} \cdots a_{k\sigma(\tau(k))} \cdots a_{l\sigma(\tau(l))} \cdots a_{n\sigma(\tau(n))} &= a_{1\sigma(1)} \cdots a_{k\sigma(l)} \cdots a_{l\sigma(k)} \cdots a_{n\sigma(n)} \\ &= a_{1\sigma(1)} \cdots a_{k\sigma(k)} \cdots a_{l\sigma(l)} \cdots a_{n\sigma(n)} \\ &= a_{1\sigma(1)} \cdots a_{n\sigma(n)}. \end{aligned}$$

Also heben sich in (∗∗) die beiden Summen gegeneinander auf.

D3) Es gilt für die Einheitsmatrix $E_n = (e_{ij})$:

$$\begin{aligned}\det A &= \sum_{\sigma\in S_n} \operatorname{sign}(\sigma) \cdot a_{1\sigma(1)} \cdots a_{n\sigma(n)} \\ &= \operatorname{sign}(\mathrm{id}) \cdot a_{11} \cdots a_{nn} = 1.\end{aligned}$$

Die Summe kollabiert hier auf einen einzigen Summanden, da in jedem anderen der $n!$ Summanden wenigstens ein Faktor 0 auftritt.

□

22.2 Weitere Eigenschaften der Determinante

Es wurden schon ganze Bücher über Eigenschaften und Formeln zu Determinanten geschrieben. Im Rahmen dieser Vorlesung müssen wir uns leider auf einige wesentliche beschränken. Dazu zählt sicher die erste, die wir vorstellen möchten, nämlich die Multiplikativität der Determinante. Dazu benötigen wir allerdings noch ein paar Notationen:

Lemma/Definition 22.10. *Für $A \in \mathrm{GL}(n, K)$ existiert eine Zerlegung*

$$A = C_1 \cdot C_2 \cdots C_s$$

*in sogenannte **Elementarmatrizen** C_k. Jede der C_k ist dabei von einem der Typen*

$$S_i(\lambda),\ Q_i^j,\ Q_i^j(\lambda) \text{ bzw. } P_i^j.$$

Dies sind die Matrizen, die durch Multiplikation von links Zeilenumformungen vom Typ I, II, III bzw. IV (siehe Definition 19.5) realisieren:

$$S_i(\lambda) = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & \vdots \\ \vdots & & \ddots & \lambda & \ddots & & \vdots \\ \vdots & & & \ddots & 1 & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \leftarrow i, \quad Q_i^j(\lambda) = \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \overset{j\text{–te Spalte}}{\downarrow} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \leftarrow i\text{–te Zeile},$$

sowie $Q_i^j := Q_i^j(1)$ und

$$P_i^j := \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \overset{i}{\downarrow} & & & & & \overset{j}{\downarrow} & \\ \end{matrix} \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & \vdots \\ 0 & 0 & 0 & 1 & 0 & 0 & \vdots \\ 0 & 0 & 0 & 0 & 1 & 0 & \vdots \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \\ \leftarrow i\text{–te Zeile} \\ \\ \\ \\ \leftarrow j\text{–te Zeile.} \\ \\ \end{matrix}$$

Beweis. Es ist sehr einfach nachzurechnen, dass die angegebenen Matrizen tatsächlich die entsprechenden Zeilenoperationen realisieren, so dass wir das hier nicht vorführen. Der Gaußalgorithmus zur Berechnung der inversen Matrix aus Abschnitt 20.7 besagt nun aber gerade, dass jede $n \times n$–Matrix mit solchen elementaren Zeilenumformungen in die Einheitsmatrix E_n zu bringen ist. Wir erhalten daher:

$$E_n = B_1 \cdot B_2 \cdots B_s \cdot A$$

für gewisse Elementarmatrizen B_i und ein $s \in \mathbb{N}$. Da außerdem die Inversen $C_{s-i+1} := (B_i)^{-1}$ der Elementarmatrizen wieder Elementarmatrizen sind, nämlich

$$(S_i(\lambda))^{-1} = S_i\Big(\frac{1}{\lambda}\Big), \quad (Q_i^j(\lambda))^{-1} = Q_i^j(-\lambda), \quad (P_i^j)^{-1} = P_i^j,$$

folgt, dass $C_1 \cdot C_2 \cdots C_s = A$ auch ein Produkt von Elementarmatrizen ist. □

Satz 22.11 (Determinanten–Multiplikationssatz / Multiplikativität der Determinante). *Für alle $A, B \in K^{n\times n}$ gilt:*

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

Beweis. Zunächst sei der **Rang** von A

$$\operatorname{rang}(A) := \dim \operatorname{Bild}(A) := \dim \operatorname{Bild}(K^n \to K^n,\ x \mapsto Ax) < n,$$

d.h. A ist nicht invertierbar. Wegen $\operatorname{Bild}(A\cdot B) \subseteq \operatorname{Bild}(A)$, also auch $\operatorname{rang}(A\cdot B) < n$, folgt mit D10):

$$\det A \cdot B = 0 = \det A \cdot \det B.$$

Nun sei $\operatorname{rang}(A) = n$, d.h. $A \in \mathrm{GL}(n, K)$. Nach dem Lemma existiert eine Zerlegung

$$A = C_1 \cdot C_2 \cdots C_s$$

in Elementarmatrizen.

Wir haben bereits in der Bemerkung 19.6 gesehen, dass wir die Zeilenoperationen vom Typ *III* und *IV* durch wiederholtes Anwenden der Typen *I* und *II* erhalten. Es reicht daher, Matrizen vom Typ $S_i(\lambda)$ und Q_i^j zu betrachten.

Wir zeigen, dass für eine Matrix C vom Typ $S_i(\lambda)$ oder Q_i^j gilt:

$$\det(C \cdot B) = \det(C) \cdot \det(B).$$

Für $C = S_i(\lambda)$ ist $\det(C) = \lambda$ und $\det(C \cdot B) = \lambda \cdot \det(B)$ (da Multiplikation mit $S_i(\lambda)$ lediglich eine Multiplikation der i–ten Zeile mit λ bewirkt), also $\det(C \cdot B) = \det(C) \cdot \det(B)$.

Für $C = Q_i^j$ ist $\det(C) = 1$ wegen D9 und $\det(C \cdot B) = \det(B)$, da C eine Addition einer Zeile zu einer anderen Zeile bewirkt.

Also ergibt sich letztendlich:

$$\begin{aligned}\det(A \cdot B) &= \det(C_1 \cdots C_s B) = \det(C_1) \cdot \det(C_2 \cdots C_s \cdot B)\\ &= \det(C_1) \cdots \det(C_s) \cdot \det(B)\\ &= \det(C_1 \cdots C_s) \cdot \det(B)\\ &= \det A \cdot \det B,\end{aligned}$$

was zu zeigen war. □

Bemerkung 22.12. Im Allgemeinen ist

$$\det(A + B) \neq \det(A) + \det(B),$$

es gilt nämlich zum Beispiel für $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ und $B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$:

$$1 = \det(A + B) \neq 0 = \det(A) + \det(B).$$

Satz 22.13. *Für jede Matrix $A \in K^{n\times n}$ gilt:*

$$\det A^{\mathrm{t}} = \det A.$$

Beweis. Ist $A = (a_{ij})$, so ist $A^{\mathrm{t}} = (a'_{ij})$ mit $a'_{ij} = a_{ji}$. Deshalb folgt mit der Formel für die Determinante aus Satz 22.8 :

$$\begin{aligned}\det A^{\mathrm{t}} &= \sum_{\sigma\in S_n} \operatorname{sign}\sigma\, a'_{1,\sigma(1)} \cdots a'_{n,\sigma(n)}\\ &= \sum_{\sigma\in S_n} \operatorname{sign}\sigma\, a_{\sigma(1),1} \cdots a_{\sigma(n),n}.\end{aligned}$$

Für jedes $\sigma \in \sigma_n$ gilt:

$$a_{\sigma(1),1} \cdots a_{\sigma(n),n} = a_{1,\sigma^{-1}(1)} \cdots a_{n,\sigma^{-1}(n)},$$

denn beide Produkte haben die gleichen Faktoren (möglicherweise in anderer Reihenfolge), denn ist $j = \sigma(i)$, so gilt $(\sigma(i), i) = (j, i) = (j, \sigma^{-1}(j))$.

Außerdem ist

$$\operatorname{sign}\sigma = \operatorname{sign}\sigma^{-1},$$

da $1 = \operatorname{sign}(\sigma \cdot \sigma^{-1}) = \operatorname{sign}(\sigma) \cdot \operatorname{sign}(\sigma^{-1})$. Damit folgt:

$$\begin{aligned}\det A^{\mathrm{t}} &= \sum_{\sigma\in S_n} \operatorname{sign}\sigma^{-1}\, a_{1,\sigma^{-1}(1)} \cdots a_{n,\sigma^{-1}(n)}\\ &\overset{(*)}{=} \sum_{\sigma\in S_n} \operatorname{sign}\sigma\, a_{1,\sigma(1)} \cdots a_{n,\sigma(n)}\\ &= \det A.\end{aligned}$$

(*) gilt, da mit σ auch σ^{-1} ganz S_n durchläuft, d.h. die Abbildung $S_n \to S_n$, $\sigma \mapsto \sigma^{-1}$ ist bijektiv. □

Bemerkung 22.14. Die Regeln D1, D2, D5, D6, D7, D10 gelten sinngemäß auch für Spalten, D9 auch für untere Dreiecksmatrizen.

22.3 Berechnung von Determinanten

Wir haben schon gesehen, dass es die Eigenschaften D6, D7 und D9 der Determinante erlauben, mit Hilfe des Gaußalgorithmus die Determinante einer beliebigen $n \times n$–Matrix prinzipiell auszurechnen. In dieserm Abschnitt stellen wir einige Formeln vor, die diese Berechnungen vereinfachen können.

Satz 22.15 (Kästchensatz). *Sei $n \geq 2$ und $A \in K^{n\times n}$ in der Form*

$$A = \begin{pmatrix} A_1 & C \\ 0 & A_2 \end{pmatrix} \qquad \textbf{\textit{(Blockmatrizen, Kästchenform).}}$$

mit $A_1 \in K^{n_1\times n_1}$, $A_2 \in K^{n_2\times n_2}$, $C \in K^{n_1\times n_2}$, dann gilt:

$$\det(A) = \det(A_1) \cdot \det(A_2).$$

Beweis. Siehe Übungsaufgabe 22.6. □

Per Induktion gilt die Aussage analog selbstverständlich auch für mehr als zwei Kästchen auf der Diagonalen.

Notation 22.16. *Sei $A = (a_{ij}) \in K^{n\times n}$. Mit A_{ij} bezeichnen wir die Matrix, die aus A entsteht, indem man a_{ij} durch 1 ersetzt, und alle anderen Einträge in Zeile i und Spalte j durch 0 ersetzt:*

$$A_{ij} = \begin{pmatrix} a_{1,1} & \dots & a_{1,j-1} & 0 & a_{1,j+1} & \dots & a_{1,n} \\ \vdots & & & \vdots & & & \vdots \\ a_{i-1,1} & & & 0 & & & a_{i-1,n} \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ a_{i+1,1} & & & 0 & & & a_{i+1,n} \\ \vdots & & & \vdots & & & \vdots \\ a_{n,1} & \dots & a_{n,j-1} & 0 & a_{n,j+1} & \dots & a_{n,n} \end{pmatrix}$$

*Die Matrix $\tilde{A} = (\tilde{a_{ij}}) \in K^{n\times n}$ mit $\tilde{a_{ij}} = \det A_{ji}$ heißt **komplementäre Matrix** zu A. Man beachte die umgekehrte Reihenfolge der Indices.*

Mit

$$A'_{ij} = \begin{pmatrix} a_{1,1} & \dots & \widehat{a_{1,j}} & \dots & a_{1,n} \\ \vdots & & \vdots & & \vdots \\ \widehat{a_{i,1}} & \dots & \widehat{a_{i,j}} & \dots & \widehat{a_{i,n}} \\ \vdots & & \vdots & & \vdots \\ a_{n,1} & \dots & \widehat{a_{n,j}} & \dots & a_{n,n} \end{pmatrix}$$

bezeichnen wir die Matrix, die durch Streichen der i–ten Zeile und der j–ten Spalte von A entsteht (nicht vorhandene Einträge werden hier mit einem $\widehat{\ }$ gekennzeichnet).

Bemerkung 22.17. Es gilt:

$$\det A_{ij} = (-1)^{i+j} \det A'_{ij}.$$

Beweis. Durch $(i-1)$ Vertauschungen benachbarter Zeilen und $(j-1)$ Vertauschungen benachbarter Spalten lässt sich A_{ij} überführen in:

$$\begin{pmatrix} 1 & 0 \cdots & 0 \\ 0 & & \\ \vdots & A'_{ij} & \\ 0 & & \end{pmatrix}.$$

Dies liefert mit Satz 22.15 über die Blockmatrizen:

$$\det A_{ij} = (-1)^{(i-1)+(j-1)} \det A'_{ij} = (-1)^{i+j} \det A'_{ij}.$$

□

Satz 22.18. *Sei $A \in K^{n\times n}$ und $\tilde{A}$ die komplementäre Matrix. Dann gilt:*

$$\tilde{A} \cdot A = A \cdot \tilde{A} = \det(A) \cdot E_n = \begin{pmatrix} \det A & & 0 \\ & \ddots & \\ 0 & & \det A \end{pmatrix}.$$

Beweis. Seien $a^1, \dots, a^n$ die Spaltenvektoren von A, und e^i der i-te Einheitsvektor. Sei $(a^1 \dots a^{j-1} e^i a^{j+1} \dots a^n)$ die Matrix, die aus A durch Ersetzen der j–ten Spalte durch e^i entsteht. Dann gilt:

$$(*) \quad \det(a^1 \dots a^{j-1} e^i a^{j+1} \dots a^n) = \det A_{ij},$$

denn man kann A_{ij} durch Typ *III*–Spaltenumformungen erhalten.

Sei $\tilde{A} \cdot A = (c_{ik})$, dann ist:

$$
\begin{aligned}
c_{ik} &= \sum_{j=1}^{n} \tilde{a_{ij}} \cdot a_{jk} = \sum_{j=1}^{n} a_{jk} \cdot \det A_{ji} \\
&\overset{(*)}{=} \sum_{j=1}^{n} a_{jk} \cdot \det(a^1 \dots a^{i-1} e^j a^{i+1} \dots a^n) \\
&\overset{D1}{=} \det(a^1 \dots a^{i-1} \sum_{j=1}^{n} a_{jk} e^j \, a^{i+1} \dots a^n) \\
&= \det(a^1 \dots a^{i-1} \, a^k \, a^{i+1} \dots a^n) \\
&\overset{D2}{=} \begin{cases} 0, & i \neq k \\ \det A, & i = k \end{cases} \\
&= \delta_{ik} \cdot \det A,
\end{aligned}
$$

wobei δ_{ik} das Kroneckersymbol bezeichnet. D.h.:

$$
\tilde{A} \cdot A = \begin{pmatrix} \det A & & 0 \\ & \ddots & \\ 0 & & \det A \end{pmatrix}.
$$

Die Gleichung $A \cdot \tilde{A} = \det A \cdot E_n$ beweist man analog. □

Eine sehr häufig eingesetzte Methode zur Berechnung der Determinante ist folgende:

Korollar 22.19 (Entwicklungssatz von Laplace). *Ist $n \geq 2$ und $A \in K^{n \times n}$, so gilt für jedes $i \in \{1, \dots, n\}$*

$$
\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A'_{ij} \quad \text{(\textbf{Entwicklung} nach der i–ten Zeile)}
$$

und für jedes $j \in \{1, \dots, n\}$

$$
\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A'_{ij} \quad \text{(Entwicklung nach der j–ten Spalte).}
$$

Beweis. Nach Satz 22.18 ist $\det A$ gleich dem i–ten Diagonaleintrag von $A \cdot \tilde{A}$:

$$
\begin{aligned}
\det A = \sum_{j=1}^{n} a_{ij} \cdot \tilde{a_{ji}} &= \sum_{j=1}^{n} a_{ij} \cdot \det A_{ij} \\
&= \sum_{j=1}^{n} a_{ij} \cdot (-1)^{i+j} \cdot \det A'_{ij},
\end{aligned}
$$

nach Bemerkung 22.17. □

Bemerkung 22.20. Genau genommen gibt Korollar 22.19 nur ein Verfahren an, um die Summanden von

$$\sum_{\sigma \in S_n} \operatorname{sign} \sigma \; a_{1\sigma(1)} \cdots a_{n\sigma(n)}$$

in einer speziellen Reihenfolge aufzuschreiben. Dies kann aber sehr nützlich sein, beispielsweise, wenn in einer Zeile oder Spalte viele Nullen stehen.

Beispiel 22.21. Nochmal das Beispiel von eben (Bsp. 22.6):

$$\begin{aligned}
\begin{vmatrix} 0 & 1 & 2 \\ 3 & 2 & 1 \\ 1 & 2 & 0 \end{vmatrix} &\overset{\text{Entw. nach 1. Spalte}}{=} 0 \cdot \begin{vmatrix} 2 & 1 \\ 2 & 0 \end{vmatrix} + (-1) \cdot 3 \cdot \begin{vmatrix} 1 & 2 \\ 2 & 0 \end{vmatrix} + 1 \cdot \begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} \\
&= 0 - 3 \cdot (-4) + 1 \cdot (-3) \\
&= 9.
\end{aligned}$$

Die durch $(-1)^{i+j}$ bewirkte Vorzeichenverteilung kann man sich als Schachbrettmuster vorstellen:

+	-	+	-
-	+	-	+
+	-	+	-
-	+	-	+

Satz 22.18 gibt auch eine Methode an, die Inverse einer Matrix A mit Hilfe der Determinante zu bestimmen:

Korollar 22.22 (Formel für die Inverse). *Sei $A \in \mathrm{GL}(n, K)$ eine invertierbare Matrix. Dann gilt:*

$$A^{-1} = \frac{1}{\det A} \tilde{A}.$$

Beweis. Dies folgt direkt aus Satz 22.18: $A \cdot \tilde{A} = \det A \cdot E_n$. □

Meist ist es praktischer, die Inverse mit dem Gaußalgorithmus zu berechnen, doch in manchen Fällen ist diese Formel doch hilfreich, etwa für sehr kleine Matrizen:

Beispiel 22.23. Für den Spezialfall $n = 2$ erhalten wir:

$$\begin{aligned}\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} &= \frac{1}{ad-bc}\begin{pmatrix} \det A_{11} & \det A_{21} \\ \det A_{12} & \det A_{22} \end{pmatrix} \\ &= \frac{1}{ad-bc}\begin{pmatrix} \det A'_{11} & -\det A'_{21} \\ -\det A'_{12} & \det A'_{22} \end{pmatrix} \\ &= \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \\ &= \begin{pmatrix} \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}.\end{aligned}$$

Bemerkung 22.24. Aus der Formel folgt insbesondere, dass sich jeder Eintrag von A^{-1} als rationaler Ausdruck mit Einträgen von A darstellen lässt. Mit Hilfe der mehrdimensionalen Analysis kann man folgern, dass die Abbildung

$$\mathrm{GL}(n,K) \to \mathrm{GL}(n,K),\ A \mapsto A^{-1}$$

differenzierbar (insbesondere stetig) ist.

Kommen wir nun zurück zur Lösung von Gleichungssystemen. Wie wir bereits wissen, ist das Gleichungssystem

$$Ax = b$$

für alle $A \in \mathrm{GL}(n,K)$ und alle $b = (b_1,\dots,b_n)^{\mathrm{t}} \in K^n$ eindeutig lösbar. Die Lösung x ist gegeben durch

$$x = A^{-1} \cdot b.$$

Man kann zunächst A^{-1} berechnen und mit b multiplizieren. Diese Schritte lassen sich wie folgt kombinieren:

Sind $a^1,\dots,a^n$ die Spalten von A, so hat A^{-1} in der i–ten Zeile und j–ten Spalte den Eintrag

$$\frac{\det A_{ji}}{\det A} = \frac{1}{\det A}\det(a^1 \dots a^{i-1}\ e^j\ a^{i+1} \dots a^n).$$

Daher folgt für die i-te Komponente von x:

$$\begin{aligned}x_i &= \sum_{j=1}^{n} \frac{1}{\det A} \cdot \det(a^1 \dots a^{i-1}\ e^j\ a^{i+1} \dots a^n) \cdot b_j \\ &= \frac{1}{\det A} \cdot \det(a^1 \dots a^{i-1}\ \sum_{j=1}^{n} b_j e^j\ a^{i+1} \dots a^n.\end{aligned}$$

Dies beweist:

Satz 22.25 (Cramersche Regel). *Seien $A \in \mathrm{GL}(n,K)$ und $b \in K^n$. Sei ferner $x = (x_1,\dots,x_n)^{\mathrm{t}} \in K^n$ die eindeutige Lösung von $Ax = b$. Dann gilt für jedes $i \in \{1,\dots,n\}$:*

$$x_i = \frac{\det(a^1 \dots a^{i-1}\ b\ a^{i+1} \dots a^n)}{\det A}.$$

Beispiel 22.26.

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + x_3 &= 1 \\ 3x_1 + 2x_2 + x_3 &= 0 \end{aligned}$$

$$\rightsquigarrow \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \cdot x = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Die Cramersche Regel liefert nun, da $\det A = -1$ ist:

$$x_1 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 2 & 1 \end{vmatrix}}{-1} = \frac{0}{-1} = 0,\ x_2 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 3 & 0 & 1 \end{vmatrix}}{-1} = \frac{1}{-1} = -1,\ x_3 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 3 & 2 & 0 \end{vmatrix}}{-1} = \frac{-2}{-1} = 2.$$

Auch die Cramersche Regel ist für die konkrete Lösung eines Gleichungssystems oft nicht die beste Wahl; mit Hilfe des Gaußalgorithmus geht dies meist schneller. Allerdings ist die Regel für theoretische Zwecke doch recht häufig einsetzbar.

Definition 22.27. *Für $A \in K^{m\times n}$ definieren wir **Rang**, **Spaltenrang** und **Zeilenrang** wie folgt:*

$$\begin{aligned} \operatorname{rang} A := \text{Spaltenrang } A &:= \dim \operatorname{Bild} A \\ \text{Zeilenrang } A &:= \text{Spaltenrang } A^{\mathrm{t}} = \dim \operatorname{Bild} A^{\mathrm{t}}. \end{aligned}$$

Proposition 22.28. *Für alle $A \in K^{m\times n}$ ist Zeilenrang A = Spaltenrang A.*

Beweis. Nach dem Struktursatz 20.25 für lineare Abbildungen gibt es invertierbare Matrizen $S \in \mathrm{GL}(m, K), T \in \mathrm{GL}(n, K)$, so dass:

$$S \cdot A \cdot T = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Offenbar gilt: Spaltenrang(SAT) = Zeilenrang(SAT). Da S und T Isomorphismen sind, gilt auch:

$$\text{Spaltenrang } SAT = \text{Spaltenrang } A \text{ bzw. Zeilenrang } SAT = \text{Zeilenrang } A.$$

Es folgt:

$$\text{Zeilenrang } A = \text{Zeilenrang } SAT = \text{Spaltenrang } SAT = \text{Spaltenrang } A,$$

wie behauptet. □

Definition 22.29. *Sei $A \in K^{m\times n}$ und $k \in \mathbb{N}$ mit $1 \leq k \leq \min\{n, m\}$. Eine $k \times k$-**Teilmatrix** von A ist eine Matrix $A' \in K^{k\times k}$, die aus A durch Streichen von $(m-k)$ Zeilen und $(n-k)$ Spalten entsteht.* $\det A'$ *nennt man einen $k \times k$-**Minor** . Offenbar hat A genau*

$$\binom{m}{k} \cdot \binom{n}{k}$$

verschiedene $k \times k$-Teilmatrizen/Minoren. Hierbei bezeichnet die Notation $\binom{a}{b}$ die Anzahl der b–elementigen Teilmengen in einer a–elementigen Menge, wie im Abschnitt 2.6 definiert und erläutert.

Satz 22.30 (Minorenkriterium für den Rang). *Sei $A \in K^{m\times n}$. Äquivalent sind:*

1. $\operatorname{rang} A = k$.
2. *Alle $(k+1) \times (k+1)$ Minoren von A sind 0, aber es gibt einen $k \times k$-Minor ungleich 0.*

Beweis. Wir zeigen für alle k die Äquivalenz von:

a) $\operatorname{rang} A \geq k$,

b) $\exists\, k \times k$–Teilmatrix A' mit $\det A' \neq 0$.

Hieraus folgt die Behauptung.

b) $\Rightarrow$ *a*): Sei A' solch eine Teilmatrix. Da $A' \in \mathrm{GL}(k, K)$, sind die k Spalten linear unabhängig. Damit sind auch die entsprechenden Spalten von A linear unabhängig. Also $\operatorname{rang} A \geq k$.

a) $\Rightarrow$ *b*): Ist $\operatorname{rang} A \geq k$, so hat A wenigstens k linear unabhängige Spalten. Sei $B \in K^{m\times k}$ die Teilmatrix dieser Spalten. Klar ist: $\operatorname{rang} B = k$. Wegen

$$\text{Zeilenrang } B = \text{Spaltenrang } B$$

(vorige Proposition) sind k Zeilen von B linear unabhängig. Wählen wir diese, erhalten wir eine $k \times k$-Teilmatrix A' von A mit $\operatorname{rang} A' = k$, d.h. $A' \in \mathrm{GL}(k, K)$, d.h. $\det A' \neq 0$.

□

Aufgaben

Aufgabe 22.1 (Determinanten). Berechnen Sie die Determinante der folgenden Matrix:

$$\begin{pmatrix} 1 & 2 & 1 & -1 \\ -1 & -1 & 3 & 1 \\ 2 & 5 & 7 & -1 \\ 1 & 1 & -11 & 2 \end{pmatrix} \in \mathbb{R}^{4\times 4}.$$

Aufgabe 22.2 (Determinante der Vandermondschen Matrix). Berechnen Sie mit vollständiger Induktion die Determinante der Vandermondschen Matrix

$$A := \begin{pmatrix} 1 & \alpha_0 & \alpha_0^2 & \dots & \alpha_0^d \\ 1 & \alpha_1 & \alpha_1^2 & \dots & \alpha_1^d \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \alpha_d & \alpha_d^2 & \dots & \alpha_d^d \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)},$$

wobei $d \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$. Vergleichen Sie das Ergebnis mit der entsprechenden Aufgabe 20.6.

Aufgabe 22.3 (Isomorphismen). Für welche $t \in \mathbb{R}$ ist die lineare Abbildung, die durch unten stehende Matrix $A \in \mathbb{R}^{6\times 6}$ definiert wird, ein Isomorphismus?

$$A = \begin{pmatrix} t & 1 & \frac{2}{3} & \frac{3}{7}t & \frac{5}{2}t & -3 \\ 1 & 2 & -3 & -7t & 8 & -2t \\ 0 & 0 & \frac{3}{7} & 2 & 0 & 0 \\ 0 & 0 & 1 & 7 & 0 & 0 \\ 0 & 0 & 3t & 2 & \frac{1}{3} & t \\ 0 & 0 & 3 & 2t & -2 & -3 \end{pmatrix}.$$

Aufgabe 22.4 (Determinanten und Geometrie).

1. Gegeben seien zwei verschiedene Punkte $P = (p_1, p_2), Q = (q_1, q_2) \in \mathbb{R}^2$. Geben Sie eine 3×3-Matrix A an, so dass
$$\left\{x_1, x_2 \in \mathbb{R}^2 \mid \det A = 0\right\}$$
genau die Gerade durch P und Q definiert.
Tipp: A sollte nicht nur reelle Zahlen, sondern auch die Variablen x_1 und x_2 als Einträge haben.
2. Wie lautet die analoge Beschreibung einer Hyperebene im $\mathbb{R}^n$ durch n Punkte?

Aufgabe 22.5 (Determinante). Sei $n \in \mathbb{N}$ und seien ferner $a_0, a_1, \dots, a_{n-1} \in \mathbb{R}$. Zeigen Sie:

$$\det \begin{pmatrix} x & -1 & & & 0 \\ 0 & x & -1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & x & -1 \\ a_0 & a_1 & \cdots & a_{n-1} & x + a_{n-1} \end{pmatrix} = x^n + a_{n-1}x^{n-1} + \cdots + a_0.$$

Aufgabe 22.6 (Kästchensatz). Beweisen Sie den Kästchensatz 22.15, der in der Vorlesung nur angegeben wurde.

Aufgabe 22.7 (Cramersche Regel). Für welche $\lambda \in \mathbb{R}$ ist die Matrix

$$A_\lambda = \begin{pmatrix} \lambda & 1 & 1 \\ 1 & \lambda & 1 \\ 1 & 1 & \lambda \end{pmatrix}$$

invertierbar? Invertieren Sie A_λ mit Hilfe der Cramerschen Regel in diesen Fällen.

23

Determinante eines Endomorphismus und Orientierung

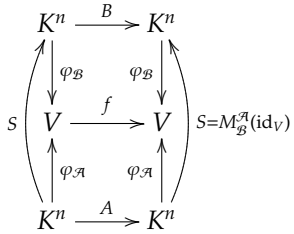
23.1 Definition der Determinante

Sei V ein K-Vektorraum, $\dim V < \infty$ und $f \in \mathrm{End}(V) := \mathrm{Hom}(V,V) := \mathrm{Hom}_K(V,V)$ ein **Endomorphismus**, wobei $\mathrm{Hom}(V,W)$ die Menge aller Vektorraumhomomorphismen von V nach W bezeichnet. Wir wollen $\det f$ definieren.

Dazu wählen wir eine Basis $\mathcal{A}$ von V, setzen $A = M^{\mathcal{A}}_{\mathcal{A}}(f)$ und definieren

$$\det(f) := \det A.$$

Um einzusehen, dass dies nicht von der Wahl einer Basis abhängt (d.h. dass $\det(f)$ wohldefiniert ist), überlegen wir uns, was passiert, wenn wir eine andere Basis $\mathcal{B}$ wählen:



Zu zeigen ist: $\det A = \det B$. Mit $S = M^{\mathcal{A}}_{\mathcal{B}}(\mathrm{id}_V) \in \mathrm{GL}(n,K)$ gilt: $B = S \cdot A \cdot S^{-1}$. Mit dem Determinantenmultiplikationssatz folgt:

$$\begin{aligned}\det B &= \det(S \cdot A \cdot S^{-1}) \\ &= \det S \cdot \det A \cdot \det S^{-1} \\ &= \det S \cdot (\det S)^{-1} \cdot \det A \\ &= \det A.\end{aligned}$$

Dies zeigt, dass die Determinante eines Endomorphismus tatsächlich unabhängig von der gewählten Basis ist. $\det(f)$ ist daher wohldefiniert.

23.2 Geometrie der Determinante eines Endomorphismus

Sei $K = \mathbb{R}$, $f\colon \mathbb{R}^n \to \mathbb{R}^n$, $A = M_{\mathcal{E}}^{\mathcal{E}}(f)$. Dann lässt sich $|\det f| = |\det A|$ als das Volumen des von $f(e_1), \ldots, f(e_n)$ aufgespannten Parallelotops interpretieren (s. Abb. 23.1), wie wir im Wesentlichen bereits in der einleitenden Motivation zu Determinanten von quadratischen Matrizen in Abschnitt 22.1.1 gesehen haben.

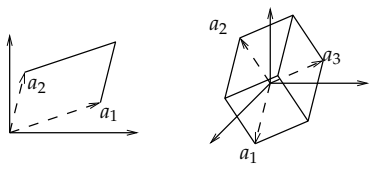


Abbildung 23.1. Parallelotope im $\mathbb{R}^n$, $n = 2, 3$.

23.3 Orientierung

Eben haben wir den Betrag der Determinante eines Endomorphismus als Volumen interpretiert. Was aber könnte das Vorzeichen der Determinante bedeuten?

Definition 23.1. *Sei V ein $\mathbb{R}$-Vektorraum,* $\dim V = n < \infty$. *Ein Endomorphismus $f\colon V \to V$ heißt* ***orientierungstreu****, wenn* $\det f > 0$. *Insbesondere ist f dann ein Isomorphismus.*

Beispiel 23.2. $V = \mathbb{R}^2$. Wir betrachten die Matrizen

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Die Auswirkung der beiden Matritzen auf den Buchstaben F sind in Abb. 23.2 dargestellt. Dies passt mit $\det A = +3$ und $\det B = -3$ zusammen.

Bemerkung/Definition 23.3. Was ist eine Orientierung von $V = \mathbb{R}^n$? Zwei Basen $\mathcal{A}, \mathcal{B}$ von V heißen **gleich orientiert**, falls $\det_{\mathcal{B}}^{\mathcal{A}}(\mathrm{id}_V) > 0$; andernfalls **entgegengesetzt orientiert**.

Nach dem Determinanten–Multiplikationssatz ist Gleichorientiertheit eine Äquivalenzrelation auf der Menge $\{\mathcal{A} \mid \mathcal{A} = \{v_1, \ldots, v_n\}$ Basis von $V\}$ (hierbei

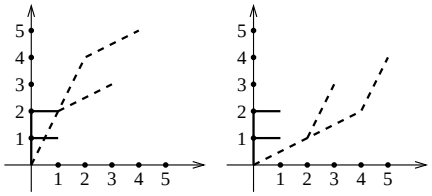


Abbildung 23.2. Orientierung am Buchstaben F: Das linke Bild zeigt F (durchgezogene Linien) und das Bild $A(F)$ unter der Matrix A aus Beispiel 23.2 (gestrichelt). Das rechte Bild zeigt das entsprechende Resultat für die Matrix B.

kommt es auf die Reihenfolge der Basiselemente an, trotz der Mengenschreibweise; siehe dazu auch eine Übungsaufgabe).

$$\{\mathcal{A} \mid \mathcal{A} = \{v_1, \ldots, v_n\} \text{ Basis von } V\} \,/\, \text{ Gleichorientiertheit}$$

besteht aus genau zwei Klassen: $\{e_1, \ldots, e_n\}$ und $\{e_{\sigma 1}, \ldots, e_{\sigma n}\}$, $\sigma \in S_n$, sind gleich orientiert genau dann, wenn $\operatorname{sign} \sigma = +1$.

Beispiel 23.4. Die Orientierung liefert im $\mathbb{R}^3$ eine Unterscheidung zwischen **rechtshändigen Koordinatensystemen** und **linkshändigen Koordinatensystemen**: $\{e_1, e_2, e_3\}$ ist ein rechtshändiges: zeigt der Daumen der rechten Hand in Richtung e_1, der Zeigefinger in Richtung e_2, so zeigt der Mittelfinger in Richtung e_3. Die linke Hand kann man entsprechend folgendermaßen in dieses Schema einpassen: Daumen nach e_3, Zeigefinger nach e_2, Mittelfinger nach e_1. Tatsächlich ist $\{e_3, e_2, e_1\} = \{e_{\sigma 1}, e_{\sigma 2}, e_{\sigma 3}\}$ für $\sigma = (13)$ und $\operatorname{sign}(13) = -1$.

Mit Hilfe des im Vorlesungsteil zur Analysis eingeführten Begriffes der Stetigkeit (Abschnitt 8) kann man folgendes schöne Kriterium für die positive Orientierung der Spalten einer invertierbaren Matrix geben. Da wir die Resultate der Analysis in diesem Abschnitt zur linearen Algebra aber nicht benutzen möchten, geben wir keinen kompletten Beweis, sondern nur eine kurze Bemerkung zur einen Richtung der Aussage:

Satz 23.5. *Seien $A \in \mathrm{GL}(n, \mathbb{R})$ und $a_1, \ldots, a_n$ die Spaltenvektoren von A. Dann sind die Basen $\{a_1, \ldots, a_n\}$ und $\{e_1, \ldots, e_n\}$ gleich orientiert genau dann, wenn es eine Abbildung*

$$\varphi \colon [0,1] \to \mathrm{GL}(n, \mathbb{R}),\ t \mapsto (\varphi_{ij}(t))$$

gibt mit $\varphi_{ij} \colon [0,1] \to \mathbb{R}$ stetig für alle $i, j \in \{1, 2, \ldots, n\}$ und $\varphi(0) = A$, $\varphi(1) = E$.

Beweis (nur Notwendigkeit). Die Notwendigkeit der Bedingung ergibt sich aus dem Zwischenwertsatz 8.9. Mit φ_{ij} ist auch $[0,1] \to \mathbb{R}$, $t \mapsto \det(\varphi(t))$ stetig, als Summe von Produkten von stetigen Funktionen. Da $\det(\varphi(t)) \neq 0\ \forall\ t$, folgt: $\det(\varphi(t))$ hat das gleiche Vorzeichen $\forall\ t \in [0,1]$. $\det \varphi(0) = \det A$ hat das gleiche Vorzeichen wie $\det \varphi(1) = \det E = 1 > 0$. □

Aufgaben

Aufgabe 23.1 (Determinante eines Endomorphismus). Sei $n \in \mathbb{N}$. Wir definieren:

$$f_n \colon \mathbb{R}[x]_{\leq n} \to \mathbb{R}[x]_{\leq n}, \quad p \mapsto (p \cdot x)',$$

wobei q' die Ableitung eines Polynoms $q \in \mathbb{R}[x]$ ist. Zeigen Sie, dass f_n ein Endomorphismus ist. Berechnen Sie die Determinante von f_5.

24

Eigenwerte und das charakteristische Polynom

24.1 Einleitung

Sei K ein Körper, $f\colon V \to W$ eine lineare Abbildung zwischen endlichdimensionalen K–Vektorräumen. Nach dem Struktursatz über lineare Abbildungen 20.25 existieren Basen $\mathcal{A}, \mathcal{B}$ von V bzw. W, so dass

$$M_{\mathcal{B}}^{\mathcal{A}}(f) = \left(\begin{array}{ccc|c} 1 & & & \\ & \ddots & & 0 \\ & & 1 & \\ \hline & & & \\ & 0 & & 0 \end{array}\right).$$

Bei einem Endomorphismus wollen wir $\mathcal{B} = \mathcal{A}$ wählen und fragen, ob $M_{\mathcal{A}}^{\mathcal{A}}(f)$ möglichst einfach ist. Etwas anders formuliert: Sei A zunächst beliebig. Gibt es dann Basiswechselmatrizen $S = M_{\mathcal{B}}^{\mathcal{A}}(\mathrm{id}_V)$, so dass $B = M_{\mathcal{A}}^{\mathcal{A}}(f) = S \cdot A \cdot S^{-1}$ (siehe dazu wieder das kommutative Diagramm auf Seite 323) möglichst einfach ist?

Mit anderen Worten: Wir betrachten die Operation

$$\mathrm{GL}(n,K) \times K^{n\times n} \to K^{n\times n},\ (S,A) \mapsto S \cdot A \cdot S^{-1}$$

von $GL(n,K)$ auf $K^{n\times n}$ durch **Konjugation** (Dies ist tatsächlich eine Operation, denn $EAE^{-1} = A$ für alle $A \in K^{n\times n}$ und $S(TAT^{-1})S^{-1} = (ST)A(ST)^{-1}$.) und fragen nach den Klassen bzgl. dieser Operation:

Definition 24.1. *Zwei quadratische Matrizen $A, B \in K^{n\times n}$ heißen **ähnlich** bzw. **konjugiert**, wenn sie in der gleichen Bahn (genannt **Konjugationsklasse**) bezüglich dieser Operation liegen, d.h. wenn ein $S \in \mathrm{GL}(n,K)$ existiert mit $B = S \cdot A \cdot S^{-1}$.*

24.2 Eigenwerte und Eigenvektoren

Der Schlüssel zur Lösung der Frage nach der möglichst einfachen Matrix bzgl. einer geeigneten Basis ist der Begriff des Eigenwerts:

Definition 24.2. *Sei V ein K–Vektorraum, $f\colon V \to V$ ein Endomorphismus, $\lambda \in K$. Der Skalar λ heißt **Eigenwert** (engl. **Eigenvalue**) von f, wenn es einen Vektor $0 \neq v \in V$ gibt, so dass*

$$f(v) = \lambda \cdot v.$$

*Solch ein v heißt dann **Eigenvektor** (engl. **Eigenvector**) von f zum Eigenwert λ.*

Achtung. Ein Eigenwert λ kann $0 \in K$ sein, ein Eigenvektor ist stets $\neq 0$.

Ein Eigenvektor ist also ein Vektor $v \neq 0$, dessen Bild $f(v)$ unter der linearen Abbildung auf der gleichen Ursprungsgeraden liegt wie vorher, für den also $f(v) \in \langle v \rangle$ gilt.

Beispiel 24.3. Im $\mathbb{R}^2$ hat demnach eine Drehung um den Ursprung um $\alpha \neq 0$ keinen Eigenvektor zu einem reellen Eigenwert, da keine Ursprungsgerade auf sich selbst abgebildet wird.

Im Gegensatz dazu hat die Spiegelung f an der x-Achse des $\mathbb{R}^2$ alle Vektoren $v \neq 0$ der x-Achse als Eigenvektoren zum Eigenwert 1, weil für diese $f(v) = 1 \cdot v$ gilt.

Abgesehen von der theoretischen Motivation, die wir in der Einleitung geliefert haben, gibt es unvorstellbar viele Anwendungen von Eigenwerten. Wir beginnen mit der Suchmaschine Google:

Beispiel 24.4. Wir betrachten zwei Anwendungen, in denen Eigenvektoren wesentlich zur Lösung eines Problems verwendet werden können:

1. City–Kunden und Outlet–Kunden,
2. Googles PageRank.

Beide sind sehr ausführlich im Internet beschrieben und wurden in der Vorlesung vorgestellt, hier aber nur verlinkt:

www.gm.fh-koeln.de/ konen/Mathe2-SS/Workshop-Google/PageRank-Workshop2-ext.pdf

Satz 24.5. *Es sei V ein K–Vektorraum, $n = \dim V < \infty$ und $f\colon V \to V$ ein Endomorphismus. Äquivalent sind:*

1. *V besitzt eine Basis aus Eigenvektoren von f.*

2. Es gibt eine Basis $\mathcal{B}$ von V, so dass

$$M_{\mathcal{B}}^{\mathcal{B}}(f) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \text{ mit } \lambda_i \in K.$$

Beweis. Ist $M_{\mathcal{B}}^{\mathcal{B}}(f)$ von der angegebenen Form, d.h. in **Diagonalgestalt** (eine solche Matrix heißt dann **Diagonalmatrix**) für $\mathcal{B} = \{v_1, \ldots, v_n\}$, dann gilt: $f(v_i) = \lambda_i v_i \; \forall i \;\Leftrightarrow\; v_1, \ldots, v_n$ ist eine Basis von Eigenvektoren. □

24.3 Das charakteristische Polynom

Bisher haben wir nur erfahren, warum Eigenwerte und –vektoren wichtig sind, aber nicht, wie wir sie berechnen können. Dieses Problem löst das charakteristische Polynom:

Definition 24.6. *Sei $A \in K^{n\times n}$ und $\lambda \in K$ beliebig. Dann heißt*

$$\mathrm{Eig}(A, \lambda) := \{v \in K^n \mid Av = \lambda v\}$$

*der **Eigenraum** von A zu λ.*

$$\chi_A(t) := \det(A - tE) \in K[t]$$

*heißt **charakteristisches Polynom** von A.*

Bemerkung 24.7. Für eine Matrix $A \in K^{n\times n}$ gilt also:

$$\lambda \in K \text{ ist ein Eigenwert von } A \Leftrightarrow \mathrm{Eig}(A, \lambda) \neq 0.$$

Die zentrale Eigenschaft ist nun:

Satz 24.8. *Seien $A \in K^{n\times n}$ und $\lambda \in K$. Dann gilt:*

$$\lambda \text{ ist ein Eigenwert von } A \Leftrightarrow \lambda \text{ ist eine Nullstelle von } \chi_A(t).$$

Beweis. Es gilt:

$$\begin{aligned} \lambda \text{ Eigenwert} &\Leftrightarrow Av = \lambda v \text{ für ein } v \neq 0 \\ &\Leftrightarrow (A - \lambda E) \cdot v = 0 \text{ hat eine nichttriviale Lösung } v \neq 0 \\ &\Leftrightarrow \mathrm{Eig}(A, \lambda) = \mathrm{Ker}(A - \lambda E) \neq 0 \\ &\Leftrightarrow \det(A - \lambda E) = 0 \\ &\Leftrightarrow \chi_A(\lambda) = 0. \end{aligned}$$

□

Im Zwei– und Drei–Dimensionalen ist die Geometrie aller eingeführten Begriffe sehr gut anschaulich verständlich und illustrierbar; dazu ein Beispiel:

Beispiel 24.9. Wir betrachten die Matrix

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

Die Operation der zugehörigen linearen Abbildung auf $\mathbb{R}^2$ verdeutlicht Abb. 24.1. Das charakteristische Polynom von A ist:

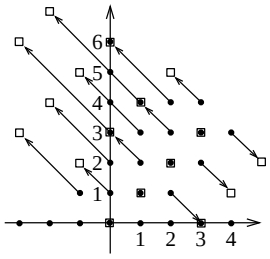


Abbildung 24.1. Die Operation der Matrix A aus Beispiel 24.9 auf $\mathbb{R}^2$.

$$\chi_A(t) = \det\begin{pmatrix} 2-t & -1 \\ -1 & 2-t \end{pmatrix} = (2-t)^2 - 1 = t^2 - 4t + 3 = (t-3)(t-1),$$

die Eigenwerte sind also $\lambda_1 = 3$, $\lambda_2 = 1$ wegen Satz 24.8.

Die Eigenräume zu diesen beiden Eigenwerten sind:

$$\operatorname{Eig}(A,3) = \operatorname{Ker}(A - 3E) = \operatorname{Ker}\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} = \left\langle \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\rangle,$$

$$\operatorname{Eig}(A,1) = \operatorname{Ker}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \left\langle \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\rangle.$$

Dies passt mit Abbildung 24.1 zusammen, dort gehen nämlich die beiden Winkelhalbierenden unter A in sich selbst über; wir sagen dann, dass diese Geraden **invariant** sind unter A. Allerdings wird auf der Winkelhalbierenden $\operatorname{Eig}(A,1)$ zwar jeder Punkt auf sich selbst abgebildet, doch auf $\operatorname{Eig}(A,3)$ wird jeder Vektor auf sein Dreifaches abgebildet. $\operatorname{Eig}(A,1)$ heißt daher **punktweise invariant** unter A.

Um nun A auf Diagonalgestalt zu bringen (dies sollte nach Satz 24.5 möglich sein, da die Eigenvektoren, die wir eben berechnet haben, eine Basis

des $\mathbb{R}^2$ bilden), definieren wir S^{-1} als Matrix, deren Spalten gerade aus den Basisvektoren der Eigenräume bestehen:

$$S^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad \Rightarrow \quad S = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Dann ist nämlich nach den Definitionen der Matrixmultiplikation und von Eigenwerten AS^{-1} eine Matrix, deren Spaltenvektoren jetzt einfach die Eigenvektoren (d.h. die Spalten von S^{-1}) multipliziert mit den zugehörigen Eigenwerten sind. Multiplikation dieser Matrix mit S von links ergibt daher eine Diagonalmatrix:

$$\begin{aligned} S \cdot A \cdot S^{-1} &= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

In diesen neuen **Koordinaten** (man kann den Wechsel der Basis auch als Wechsel des Koordinatensystems auffassen) ist die Abbildung also einfach zu verstehen; s. auch Abb. 24.2. Blicken wir nun nochmals zurück auf Abb. 24.1, so stellen wir fest, dass dies genau damit zusammen passt.

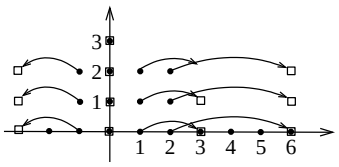


Abbildung 24.2. Die Operation der Matrix SAS^{-1} aus Beispiel 24.9, die Diagonalgestalt mit den Diagonaleinträgen 3 und 1 besitzt.

Bemerkung/Definition 24.10. Wir haben eben gesehen, dass die Nullstellen von $\chi_A(t)$ die Eigenwerte der durch $A \in K^{n\times n}$ definierten linearen Abbildung sind. Allgemeiner können wir auch ein **charakteristisches Polynom** $\chi_f(t)$ **eines Endomorphismus** $f\colon V \to V, n = \dim V < \infty$, definieren. Wir definieren:

$$\chi_f(t) := \det(f - t \cdot \mathrm{id}_V) = \det(A - t \cdot E),$$

wobei $A = M_{\mathcal{A}}^{\mathcal{A}}(f)$ und $\mathcal{A}$ irgendeine Basis von V ist. Wir müssen wieder zeigen, dass diese Definition wohldefiniert ist. Sei B also die Matrixdarstellung bzgl. einer anderen Basis $\mathcal{B}$:

$$B = M_{\mathcal{B}}^{\mathcal{B}}(f) = SAS^{-1}.$$

Dann gilt:

$$\begin{aligned}\chi_B(t) &= \det(B - t{\cdot}E) = \det(SAS^{-1} - t{\cdot}E)\\ &= \det(S(A - t{\cdot}E)S^{-1}), \qquad (\text{da } SES^{-1} = E)\\ &= \det S \cdot \det(A - t{\cdot}E) \cdot \det S^{-1}\\ &= \chi_A(t),\end{aligned}$$

d.h. $\chi_f(t)$ ist tatsächlich wohldefiniert.

Diese Rechnung zeigt insbesondere:

Satz 24.11. *Sind $A, B \in K^{n\times n}$ zueinander konjugierte Matrizen, dann gilt:*

$$\chi_A(t) = \chi_B(t).$$

Zueinander konjugierte Matrizen haben also die gleichen Eigenwerte.

Wir sehen uns das charakteristische Polynom einer quadratischen Matrix $A \in K^{n\times n}$ noch etwas genauer an. Es gilt, da die Determinante linear in jeder Zeile ist:

$$\begin{aligned}\chi_A(t) &= \det\begin{pmatrix} a_{11}-t & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22}-t & \dots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \dots & a_{nn}-t\end{pmatrix}\\ &= \det\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22}-t & \dots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \dots & a_{nn}-t\end{pmatrix} + \det\begin{pmatrix} -t & 0 & \dots & 0\\ a_{21} & a_{22}-t & \dots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \dots & a_{nn}-t\end{pmatrix}\\ &= \dots = \det A + \dots + \left(\sum_{i=1}^{n} a_{ii}\right)(-t)^{n-1} + (-t)^n\\ &= b_0 + b_1{\cdot}t + \dots + b_{n-1}{\cdot}t^{n-1} + b_n{\cdot}t^n \in K[t]\end{aligned}$$

für

$$b_0 = \det A, \quad b_{n-1} = (-1)^{n-1} \cdot \sum_{i=1}^{n} a_{ii}, \quad b_n = (-1)^n.$$

Übrigens folgt $b_0 = \det A$ auch direkt aus $\chi_A(0) = \det A$ wegen der Definition von $\chi_A(t)$. Auch die Behauptung über b_{n-1} und b_n kann man anders einsehen: Entwickeln von $\det(A - tE)$ nach der ersten Spalte liefert $(a_{11} - t) \cdot A_{11} + a_{21} \cdot A_{21} + \dots + a_{n1} \cdot A_{n1}$. Aber in A_{j1}, $j > 1$ kommen jeweils nur $n - 2$ Einträge mit t vor (weil immer zwei gestrichen werden), so dass die Determinante nach der Determinantenformel höchstens Grad t hat. Daher kommen t^{n-1} und t^n nur im Produkt $(a_{11} - t)\cdots(a_{nn} - t) = (-t)^n + \left(\sum_{i=1}^{n} a_{ii}\right)(-t)^{n-1} + \dots$ vor.

Als Koeffizienten des charakteristischen Polynoms kommen also sowohl $\det A$ als auch die vorzeichenbehaftete Summe der Diagonaleinträge der Matrix vor. Da letztere noch häufiger auftauchen wird, geben wir dieser Summe einen eigenen Namen:

Definition 24.12. *Sei $A \in K^{n\times n}$ eine Matrix.*

$$\operatorname{tr}(A) := \operatorname{Spur}(A) := \sum_{i=1}^{n} a_{ii}$$

*heißt die **Spur** (engl. **trace**) von A.*

Man kann zeigen, dass das charakteristische Polynom einer Matrix $A \in K^{n\times n}$ noch eine weitere sehr interessante Eigenschaft besitzt (dies ist der sogenannte **Satz von Cayley–Hamilton**): Es gilt:

$$\chi_A(A) = 0 \in K^{n\times n},$$

d.h. setzt man in $\chi_A(t)$ statt einer reellen Zahl die Matrix A ein, so erhält man die 0–Matrix. Wir können dies hier leider nicht beweisen, doch werden wir in einer Übungsaufgabe wenigstens ein etwas schwächeres Resultat kennen lernen.

24.4 Diagonalisierbarkeit

Kommen wir nun wieder zurück auf die Eingangsfrage danach, wie einfach eine zu einer gegebenen quadratischen Matrix ähnliche Matrix aussehen kann, und zwar insbesondere zu der Frage, unter welchen Bedingungen diese Diagonalgestalt besitzen kann.

Definition 24.13. *Sei $P(t) \in K[t]$ ein Polynom. Wir sagen, dass $P(t)$ über K in **Linearfaktoren zerfällt** genau dann, wenn es $\lambda_1, \dots, \lambda_n \in K, c \in K^*$, gibt, so dass*

$$P(t) = c \cdot (t-\lambda_1)\cdots(t-\lambda_n) = c \cdot \prod_{j=1}^{r} (t-\lambda'_j)^{m_j},$$

*wobei $m_j \in \mathbb{N}$ und die $\lambda'_1, \dots, \lambda'_r \in \{\lambda_1, \dots, \lambda_n\}$ paarweise verschieden sind. m_j heißt **Vielfachheit** der Nullstelle λ'_j. Es ist $\sum_{j=1}^{r} m_j = n$.*

Für ein beliebiges, nicht notwendig in Linearfaktoren zerfallendes, Polynom $P(t) \in K[t]$ und $\lambda \in K$ ist

$$\begin{aligned} m(P,\lambda) &:= \max\Big\{ m \;\Big|\; \exists\, Q(t) \in K[t], \text{ so dass } P(t) = (t-\lambda)^m Q(t) \Big\} \\ &= m, \text{ wobei } P(t) = (t-\lambda)^m \cdot Q(t) \text{ mit } Q(\lambda) \neq 0. \end{aligned}$$

*die **Vielfachheit** von λ als Nullstelle von P.*

Also gilt beispielsweise:

$$\begin{aligned} m(P,\lambda) = 0 &\quad\Leftrightarrow\quad \lambda \text{ ist keine Nullstelle von } P, \\ m(P,\lambda) = 1 &\quad\Leftrightarrow\quad \lambda \text{ ist eine einfache Nullstelle von } P, \\ m(P,\lambda) = 2 &\quad\Leftrightarrow\quad \lambda \text{ ist eine doppelte Nullstelle von } P. \end{aligned}$$

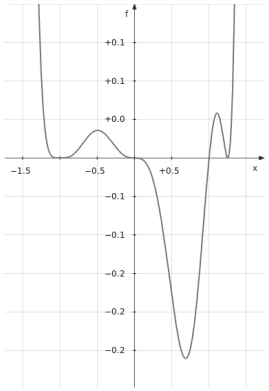


Abbildung 24.3. Das Polynom $p(x) = (x+1)^4 \cdot x^3 \cdot (x-1) \cdot (x-1.25)^2$ hat (von links) Nullstellen mit Vielfachheiten 4, 3, 1 und 2.

24.4.1 Ein Diagonalisierbarkeits–Kriterium

Wir haben eben die Vielfachheit einer Nullstelle eines Polynoms definiert. Ein erster Zusammenhang zu den Eigenräumen ist folgender:

Lemma 24.14. *Es gilt für jeden Eigenwert λ einer $n \times n$–Matrix $A \in K^{n\times n}$:*

$$\dim \mathrm{Eig}(A,\lambda) \leq m(\chi_A(t),\lambda).$$

Beweis. Wäre die Dimension des Eigenraums zu λ größer als $r = m(\chi_A(t),\lambda)$, so würde eine Basis $v_1,\ldots,v_k$ dieses Eigenraumes mit $k > r$ existieren. Ergänzen wir diese zu einer Basis $v_1,\ldots,v_n$ von K^n, so ist $S^{-1} = (v_1 \ldots v_n)$ invertierbar und es gilt:

$$AS^{-1} = (Av_1 \ \ldots \ Av_n) = (\lambda v_1 \ldots \lambda v_k \, Av_{k+1} \ldots Av_n).$$

Da $SS^{-1} = E$ ist, folgt:

$$\begin{aligned} SAS^{-1} &= S \cdot (\lambda v_1 \dots \lambda v_k \; Av_{k+1} \dots Av_n) \\ &= (\lambda Sv_1 \dots \lambda Sv_k \; SAv_{k+1} \dots SAv_n) \\ &= (\lambda \cdot e_1 \dots \lambda \cdot e_k \; SAv_{k+1} \dots SAv_n). \end{aligned}$$

Nach dem Kästchensatz ist dann $m(\chi_A(t), \lambda) \geq k > r$, ein Widerspruch. □

Gilt sogar Gleichheit und zerfällt $\chi_A(t)$ in Linearfaktoren, so ist A diagonalisierbar:

Definition 24.15. *$A \in K^{n\times n}$ heißt **diagonalisierbar** (über K), wenn $\exists S \in \mathrm{GL}(n, K)$: $SAS^{-1} = D$ für eine Diagonalmatrix D.*

Satz 24.16 (Diagonalisierbarkeits–Kriterium). *Sei $A \in K^{n\times n}$. A ist diagonalisierbar genau dann, wenn folgende beide Bedingungen erfüllt sind:*

1. *$\chi_A(t) \in K[t]$ (über K) in Linearfaktoren zerfällt,*
2. *für jede Nullstelle λ von $\chi_A(t)$ gilt: $m(\chi_a(t), \lambda) = \dim \mathrm{Eig}(A, \lambda)$.*

Manchmal nennt man $m(\chi_A(t), \lambda)$ auch **algebraische Vielfachheit** eines Eigenwertes λ und $\dim \mathrm{Eig}(A, \lambda)$ seine **geometrische Vielfachheit**. Mit dieser Terminologie heißt die zweite Bedingung, dass algebraische und geometrische Vielfachheit für alle Eigenwerte übereinstimmen sollen.

Bevor wir den Satz auf Seite 337 beweisen, zunächst ein paar Folgerungen und Bemerkungen:

Korollar 24.17. *Sei $A \in K^{n\times n}$. Hat das charakteristische Polynom $\chi_A(t)$ genau n verschiedene Nullstellen $\lambda_1, \dots, \lambda_n \in K$, dann ist A diagonalisierbar zu:*

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Beweis. Für jedes λ_i gilt

$$\begin{aligned} 1 &\leq \dim \mathrm{Eig}(A, \lambda_i) \\ &\leq m(\chi_A, \lambda_i) \\ &= 1, \end{aligned}$$

da alle Nullstellen einfach sind. □

Bemerkung 24.18. Ist A diagonalisierbar, etwa $SAS^{-1} = D$, und ist $k \in \mathbb{N}$, so gilt:

$$D^k = (SAS^{-1})^k = SA^kS^{-1},$$

also

$$A^k = S^{-1}D^kS.$$

Dies ist sehr hilfreich, um höhere Potenzen diagonalisierbarer Matrizen auszurechnen.

Bemerkung 24.19. 1. Für $K = \mathbb{C} = \{a + b\sqrt{-1} \mid a, b \in \mathbb{R}\}$ zerfällt jedes Polynom $p \in K[t]$ in einer Variablen t in Linearfaktoren, denn es gilt:

Satz 24.20 (Fundamentalsatz der Algebra, ohne Beweis). *Jedes Polynom $P(t) \in \mathbb{C}[t]$ vom Grad $d \geq 1$ hat eine Nullstelle (in $\mathbb{C}$).*

Die aus der Schule bekannte Polynomdivision (siehe dazu auch den euklidischen Algorithmus aus Abschnitt 3.5) liefert daher:

Korollar 24.21. *Jedes Polynom $P(t) \in \mathbb{C}[t]$ zerfällt in Linearfaktoren.*

Dies erste Bedingung des Satzes ist für $K = \mathbb{C}$ also immer erfüllt. Für $K = \mathbb{R}$ ist dies natürlich nicht richtig, wie das Beispiel $t^2 + 1$ zeigt.

2. Die zweite Bedingung ist nicht immer erfüllt, auch, wenn es die erste ist. Ist beispielsweise

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

so folgt $\chi_A(t) = (1-t)^2$, d.h. $m(\chi_A(t), 1) = 2$, aber

$$\dim \operatorname{Eig}(A, 1) = \dim \ker \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 1.$$

Lemma 24.22. *$A \in K^{n\times n}$. Seien $\lambda_1, \ldots, \lambda_r$ paarweise verschiedene Eigenwerte von A und $v_1, \ldots, v_r$ zugehörige Eigenvektoren. Dann sind $v_1, \ldots, v_r$ linear unabhängig. Genauer gilt: Die Summe*

$$\operatorname{Eig}(A, \lambda_1) \oplus \operatorname{Eig}(A, \lambda_2) \oplus \cdots \oplus \operatorname{Eig}(A, \lambda_r) \subseteq K^n$$

ist direkt, also nach Definition:

$$\operatorname{Eig}(A, \lambda_j) \cap \sum_{\substack{l=i \\ l\neq j}}^{r} \operatorname{Eig}(A, \lambda_l) = 0 \qquad \forall\, j \in \{1, \ldots, r\}.$$

Beweis. Induktion nach r. Der Fall $r = 1$ ist trivial.

Nehmen wir also an, dass:

$$v_r \in \operatorname{Eig}(A, \lambda_r) \cap \left(\sum_{j=1}^{r-1} \operatorname{Eig}(A, \lambda_j) \right),$$

etwa

$$v_r = w_1 + \cdots + w_{r-1} \quad \text{für gewisse } w_j \in \operatorname{Eig}(A, \lambda_j).$$

Es folgt, da λ_j, $j = 1, \ldots, r-1$, die Eigenwerte zu den Eigenvektoren w_j sind:

$$\lambda_r v_r = A v_r = A w_1 + \cdots + A w_{r-1} = \lambda_1 w_1 + \cdots + \lambda_{r-1} w_{r-1}.$$

Wieder durch Verwendung von $v_r = w_1 + \cdots + w_{r-1}$ erhalten wir:

$$0 = (\lambda_1 - \lambda_r) w_1 + \cdots + (\lambda_{r-1} - \lambda_r) w_{r-1}.$$

Da die Summe $\operatorname{Eig}(A, \lambda_1) \oplus \cdots \oplus \operatorname{Eig}(A, \lambda_{r-1}) \subseteq K^n$ direkt ist nach Induktionsvoraussetzung, folgt:

$$(\lambda_j - \lambda_r) w_j = 0 \in \operatorname{Eig}(A, \lambda_j).$$

Es gilt aber $\lambda_j \neq \lambda_r$ nach Voraussetzung und somit $w_j = 0$, $j = 1, \ldots, r-1$, also schließlich: $v_r = w_1 + \cdots + w_{r-1} = 0$. □

Damit können wir nun das Diagonalisierungskriterium nachweisen:

Beweis (Beweis des Satzes 24.16). Zur Notwendigkeit der Bedingungen: Ist A diagonalisierbar, dann ist wegen Satz 24.11

$$\chi_A(t) = \chi_D(t) = \prod_{j=1}^{n} (\lambda_j - t)$$

zerfallend. Ferner:

$$\begin{aligned} \dim \operatorname{Eig}(A, \lambda_j) &= \dim \operatorname{Eig}(D, \lambda_j) \\ &= \dim \ker \begin{pmatrix} \lambda_1 - \lambda_j & & 0 \\ & \ddots & \\ 0 & & \lambda_n - \lambda_j \end{pmatrix} \\ &= |\{i \in \{1, 2, \ldots, n\} \mid \lambda_i = \lambda_j\}| \\ &= m(\chi_A(t), \lambda_j). \end{aligned}$$

Wir haben also noch zu zeigen, dass die Bedingungen auch hinreichend für die Diagonalisierbarkeit sind: Es sei dazu

$$\chi_A(t) = \prod_{i=1}^{n} (\lambda_i - t) = \prod_{j=1}^{r} (\lambda_j - t)^{m_j},$$

wobei $\lambda_1, \ldots, \lambda_r$ die paarweise verschiedenen Eigenwerte bezeichnen, d.h. $\sum_{j=1}^{r} m_j = n$. Nach dem Lemma gilt:

$$\mathrm{Eig}(A,\lambda_1)\oplus\cdots\oplus\mathrm{Eig}(A,\lambda_r)\subseteq K^n,$$

d.h. insbesondere $\dim\Big(\mathrm{Eig}(A,\lambda_1)\oplus\cdots\oplus\mathrm{Eig}(A,\lambda_r)\Big)=\sum_{j=1}^{r}\dim\mathrm{Eig}(A,\lambda_j)$. Nach der 2. Bedingung ist aber

$$\sum_{j=1}^{r}\dim\mathrm{Eig}(A,\lambda_j)=\sum_{j=1}^{r}m_j=\deg\chi_A(t)=n=\dim K^n.$$

Insgesamt zeigt dies:

$$\mathrm{Eig}(A,\lambda_1)\oplus\cdots\oplus\mathrm{Eig}(A,\lambda_r)=K^n.$$

Fügen wir Basen der Eigenräume $\mathrm{Eig}(A,\lambda_j)$, $j=1,\dots,r$, zu einer Basis $\{v_1,\dots,v_n\}$ von K^n zusammen, dann hat bezüglich dieser Basis der Endomorphismus A Diagonalgestalt. Genauer: Ist nämlich wie im Beweis zum Lemma 24.14

$$S^{-1}=(v_1\ \dots\ v_n)$$

die Matrix, deren Spalten diese Basisvektoren sind, so ist

$$SAS^{-1}=\begin{pmatrix}A_1 & & 0\\ & \ddots & \\ 0 & & A_r\end{pmatrix},$$

wobei die $k_i\times k_i$–Kästchen $A_i\in K^{k_i\times k_i}$ mit $k_i=$ Vielfachheit des Eigenwertes λ_i gerade die Diagonalmatrizen

$$A_i=\begin{pmatrix}\lambda_i & & 0\\ & \ddots & \\ 0 & & \lambda_i\end{pmatrix}\in K^{k_i\times k_i}$$

sind. □

24.4.2 Anwendung: Lineare Rekursionen

Wie wir in einer Übungsaufgabe am Beispiel der Fibonacci–Zahlen sehen werden, können wir für eine **lineare Rekursion**, d.h. eine Formel der Form

$$r_n=ar_{n-1}+br_{n-2},\quad r_0=a_0,\ r_1=a_1,$$

eine **geschlossene Formel** für r_n (d.h. eine Formel, in der zwar n, nicht aber die r_i vorkommen) mit Hilfe von Eigenwerten und –vektoren herleiten.

Dies beruht darauf, dass offenbar:

$$\begin{pmatrix} r_n \\ r_{n-1} \end{pmatrix} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} r_{n-1} \\ r_{n-2} \end{pmatrix}.$$

Den nächsten Wert, r_{n+1}, können wir nun mit Hilfe linearer Algebra berechnen:

$$\begin{pmatrix} r_{n+1} \\ r_n \end{pmatrix} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} r_n \\ r_{n-1} \end{pmatrix} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}^2 \cdot \begin{pmatrix} r_{n-1} \\ r_{n-2} \end{pmatrix}.$$

Um r_{n+k} zu berechnen, benötigen wir also A^k. Ist aber A diagonalisierbar, etwa $SAS^{-1} = D$, so folgt $A^k = S^{-1}D^kS$, was einfach zu berechnen ist. Dies kann man benutzen, um eine geschlossene Formel für r_n anzugeben. In den Übungsaufgaben werden wir dies verwenden, um eine solche Formel für die **Fibonacci–Zahlen** (diese haben wir bereits im ersten Semester in Abschnitt 1.3.4 gesehen, konnten dort aber die Herkunft der Formel nicht erklären)

$$f_n := f_{n-1} + f_{n-2},\ n \geq 2, \quad f_0 = 0, f_1 = 1,$$

herzuleiten. Die Folge beginnt folgendermaßen:

$$(f_0, f_1, f_2, \dots) = (0, 1, 1, 2, 3, 5, 8, 13, 21, 34, \dots).$$

24.5 Die Jordansche Normalform

Über den komplexen Zahlen ist jede Matrix zu einer recht einfachen Matrix, ihrer sogenannten Jordanschen Normalform konjugiert. Betrachten wir also die Operation $\mathrm{GL}(n, \mathbb{C}) \times \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$, $(S, A) \mapsto SAS^{-1}$. Wir wissen bereits, dass über $\mathbb{C}$ aus

$$\dim \mathrm{Eig}\,(A, \lambda) = m(\chi_A(t), \lambda) \quad \forall \lambda \in W$$

folgt, dass die Matrix A diagonalisierbar ist. Wir haben in der Bahn von A also eine Diagonalmatrix als Repräsentanten:

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Als Grenzwert von diagonalisierbaren Matrizen tauchen nicht diagonalisierbare auf: Beispielsweise ist die Matrix $\begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{pmatrix}$ für $\lambda_1 \neq \lambda_2$ diagonalisierbar, aber

$$\lim_{\lambda_2 \to \lambda_1} \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix} =: A$$

hat nur λ_1 als Eigenwert und es gilt:

$$m(\chi_A(t), \lambda_1) = 2 > \dim \operatorname{Eig}(A, \lambda_1) = \dim \operatorname{Ker} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 1.$$

Allgemein haben wir stets folgende Repräsentanten:

Definition 24.23. *Ein **Jordankästchen** der Größe k zum Eigenwert λ ist die Matrix:*

$$J(\lambda, k) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \in \mathbb{C}^{k \times k}.$$

Mit der gleichen Begründung wie für die 2×2–Matrizen oben ist ein solches Kästchen für $k \geq 2$ nicht diagonalisierbar, weil $m(\chi_{J(\lambda,k)}(t), \lambda) = k$, aber $\dim \operatorname{Eig}(J(\lambda, k), \lambda) = 1$.

Satz 24.24 (Jordansche Normalform, ohne Beweis). *Sei $A \in \mathbb{C}^{n \times n}$ eine quadratische Matrix mit komplexen Einträgen. Dann existieren $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$, $k_1, \ldots, k_r \in \mathbb{N}$ und $S \in \mathrm{GL}(n, \mathbb{C})$, so dass:*

$$SAS^{-1} = J = \begin{pmatrix} J(\lambda_1, k_1) & 0 & \cdots & 0 \\ 0 & J(\lambda_2, k_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & J(\lambda_r, k_r) \end{pmatrix},$$

wobei die Eigenwerte $\lambda_1, \ldots, \lambda_r$ nicht notwendig paarweise verschieden sind.

Über den reellen Zahlen und über allgemeinen Körpern gibt es ein ähnliches Resultat, das aber etwas aufwändiger zu formulieren ist, so dass wir es hier nicht angeben.

Wir haben weder erklärt, wie man das Resultat über die Jordansche Normalform beweist, noch, wie man die k_i berechnet. Leider können wir das im Rahmen der Vorlesung auch nicht erledigen. Daher möchten wir an dieser Stelle darauf hinweisen, dass sehr viele Computeralgebra–Programme sowohl Eigenwerte und Eigenvektoren als auch die Jordansche Normalform berechnen können. Auch an der Universität des Saarlandes ist die Software Maple verfügbar und es kann durchaus hilfreich sein, sich einmal über die recht ausführliche und verständlichen Hilfeseiten soweit einzuarbeiten, dass man wenigstens einfache Berechnungen damit durchführen kann. In Maple existieren verschiedene Bibliotheken zur linearen Algebra, die entsprechende Berechnungen durchführen können. Beispielsweise liefert

```
with(linalg);
```

eine Liste aller in dieser Bibliothek zur Verfügung gestellten Prozeduren, die auch gleich zur Benutzung bereit stehen.

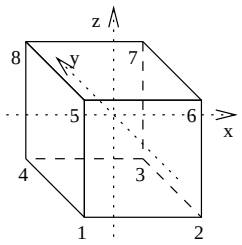
```
A := matrix(2,2,[1,2,3,4]);
eigenvalues(A);
eigenvectors(A);
jordan(A);
```

ermittelt die erfragten Ergebnisse ohne Wartezeit.

Alternativ kann man auch die Webseite www.WolframAlpha.com verwenden und dort in die Suchmaske `JordanDecomposition[{{1,2},{3,4}}]` eingeben.

Aufgaben

Aufgabe 24.1 (Eigenwerte und Eigenräume). Wir betrachten den Würfel W mit Ecken $(\pm 1, \pm 1, \pm 1) \in \mathbb{R}^3$, dessen Schwerpunkt also im Ursprung des Koordinatensystems liegt:



Berechnen Sie Matrixdarstellungen der folgenden linearen Abbildungen, die den Würfel auf sich selbst abbilden, und berechnen Sie Eigenwerte und Eigenräume dieser Matrizen:

1. D := Drehung um 180° um die Achse, die die Mittelpunkte der Strecken 15 und 37 verbindet.
2. S := Spiegelung an der Ebene, die durch die Punkte 2, 4, 6, 8 geht.
3. Die Abbildung $D \circ S$ (eine sogenannte Drehspiegelung).

Aufgabe 24.2 (). Die Vielfachheit eines Eigenwertes λ als Nullstelle des charakteristischen Polynoms nennen wir *algebraische Multiplizität* von λ, die Dimension des zu λ gehörenden Eigenraumes nennen wir *geometrische Multiplizität* von λ. Berechnen Sie algebraische und geometrische Multiplizität der folgenden Matrizen:

$$A = \begin{pmatrix} 2&0&0&0\\ 0&2&0&0\\ 0&0&2&0\\ 0&0&0&2 \end{pmatrix}, B = \begin{pmatrix} 2&1&0&0\\ 0&2&0&0\\ 0&0&2&0\\ 0&0&0&2 \end{pmatrix}, C = \begin{pmatrix} 2&1&0&0\\ 0&2&1&0\\ 0&0&2&0\\ 0&0&0&2 \end{pmatrix}, D = \begin{pmatrix} 2&1&0&0\\ 0&2&0&0\\ 0&0&2&1\\ 0&0&0&2 \end{pmatrix}, E = \begin{pmatrix} 2&1&0&0\\ 0&2&1&0\\ 0&0&2&1\\ 0&0&0&2 \end{pmatrix}.$$

Aufgabe 24.3 (Potenzen von Matrizen).

1. Berechnen Sie M^2, M^3, M^4 für die Matrix $M = \begin{pmatrix} 0&1&2&1\\ 0&0&1&1\\ 0&0&0&1\\ 0&0&0&0 \end{pmatrix}$.

   Was sind die Eigenwerte und Eigenräume von M?

2. Berechnen Sie Eigenwerte und Eigenräume von: $A = \begin{pmatrix} 19&-12\\ 30&-19 \end{pmatrix}$.

   Finden Sie eine Diagonalmatrix D und eine invertierbare Matrix S, so dass $A = SDS^{-1}$ und berechnen Sie A^{10000}.

Aufgabe 24.4 (Lineare Rekursion). Seien $a, b \in \mathbb{R}$. Es sei nun $x_0 = a$, $x_1 = b$ und $x_n = \frac{x_{n-1}+x_{n-2}}{2}$ für $n \geq 2$.

1. Schreiben Sie die Rekursion in der Form $y_n = A \cdot y_{n-1}$, wobei A eine 2×2 - Matrix ist und $y_i = \begin{pmatrix} x_i \\ x_{i-1} \end{pmatrix}$.
2. Finden Sie eine Diagonalmatrix D und eine invertierbare Matrix S, so dass $A = SDS^{-1}$.
3. Bestimmen Sie: $\lim_{n\to\infty} S^{-1}A^nS$.
4. Leiten Sie daraus $\lim_{n\to\infty} A^n$ und $\lim_{n\to\infty} x_n$ ab.

Aufgabe 24.5 (Relationen zwischen Matrizen).

1. Sei $A = \begin{pmatrix} 2&-1&0\\ -1&2&-1\\ 0&-1&2 \end{pmatrix}$.

   Zeigen Sie: $A^3 - 6A^2 + 10A - 4E_3 = 0$, wobei $E_3 \in K^{3\times3}$ die Einheitsmatrix ist.

2. Sei nun $A \in K^{n\times n}$ beliebig. Zeigen Sie: Es existiert ein Polynom $P(t) = b_rt^r + \cdots + b_1t + b_0 \in K[t]$, so dass: $b_rA^r + \cdots + b_1A + b_0E_n = 0$, wobei $E_n \in K^{n\times n}$ die Einheitsmatrix ist.

25

Hauptachsentransformation

Reelle symmetrische Matrizen spielen eine besondere Rolle, beispielsweise weil, wie wir sehen werden, alle ihre Eigenwerte reell sind. Die Symmetrie von Matrizen ist trotzdem keine Eigenschaft, die so speziell ist, dass sie nie vorkommt; im Gegenteil: jede Quadrik (also Kugel, Ellipsoide, Hyperboloide, etc., s. Abb. 25.1) lässt sich mit solch einer Matrix beschreiben. Dies wird es uns ermöglichen, mit Hilfe der linearen Algebra eine Klassifikation dieser geometrischen Objekte zu erreichen. Für wesentlich mehr Hintergrundinformationen zur Anwendungen der Linearen Algebra in der Geometrie ist das Buch von Gerd Fischer [Fis01] — ggf. in Kombination mit seinem Buch zur linearen Algebra [Fis08] — zu empfehlen.

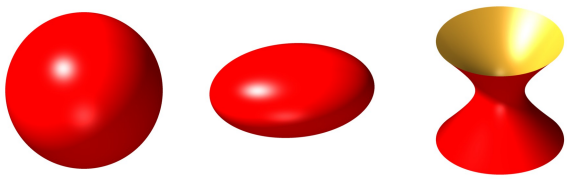


Abbildung 25.1. Einige Quadriken: Eine Kugel, ein Ellipsoid und ein einschaliger Hyperboloid.

Für die Informatik sind solche Flächen beispielsweise wichtig, weil die meisten Computer Aided Design Programme sie als Basis–Objekte zur Verfügung stellen, aus denen man kompliziertere mittels booleschen Operationen wie Vereinigung, Durchschnitt, etc. erzeugen kann. Außerdem kann man mit ihrer Hilfe geschwungene Objekte, wie Autokarosserien oder Flugzeuge, oft besser annähern als mit kleinen Dreiecken, weil von letzteren zu viele benötigt werden. Dies ist allerdings nicht trivial: Erstaunlicherweise stößt man

schon bei der exakten Berechnung der Schnittpunkte und –kurven von wenigen Quadriken auf große — von der aktuellen Forschung immer noch nicht zufriedenstellend gelöste — algorithmische Probleme, u.a. weil dabei die Koordinaten oft komplizierte Wurzelausdrücke beinhalten.

25.1 Symmetrische Matrizen

Im letzten Kapitel haben wir Kriterien dafür entwickelt, wann eine Matrix ähnlich zu einer recht einfachen Matrix, wie beispielsweise einer Diagonalmatrix oder einer Jordanmatrix ist. In allen Fällen ging das nur sehr gut, wenn alle Eigenwerte über dem Grundkörper existieren. Da aber für viele Polynome über den reellen Zahlen nicht alle Nullstellen über den reellen Zahlen existieren, erscheint die Frage sinnvoll, ob man einer Matrix unter gewissen Voraussetzungen ansehen kann, dass alle Eigenwerte reell sind. Betrachten wir die beiden Matrizen

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \Rightarrow \chi_A(t) = t^2 + 1, \text{ aber } \chi_B(t) = t^2 - 1 \text{ für } B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Das charakteristische Polynom zerfällt für die symmetrische Matrix B schon über den reellen Zahlen in Linearfaktoren; für die nicht symmetrische Matrix A müssen wir hierfür komplexe Zahlen zu Hilfe nehmen. Tatsächlich haben symmetrische reelle Matrizen immer nur reelle Eigenwerte; wir werden dies zwar erst im nächsten Kapitel beweisen, aber hier schon einige geometrische Folgerungen angeben.

Definition 25.1. *Eine Matrix $A = (a_{ij}) \in K^{n\times n}$ heißt **symmetrisch**, wenn*

$$A^t = A.$$

Bemerkung 25.2. Sei $A \in K^{n\times n}$. A symmetrisch ist äquivalent zu:

$$(x^t A^t)\, y = \langle Ax, y\rangle = \langle x, Ay\rangle = x^t A\, y \quad \forall\, x, y \in K^n.$$

Beweis. Für die Standard–Basis–Vektoren $x = e_i$ und $y = e_j$ ergibt sich:

$$e_i^t \cdot A^t \cdot e_j = (a_{1i}, \ldots, a_{ni}) \cdot e_j = a_{ji}$$

$$\text{und } e_i^t \cdot A \cdot e_j = e_i^t \cdot \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix} = a_{ij}.$$

Es muss also tatsächlich $a_{ij} = a_{ji}\ \forall i, j$ gelten. Die Umkehrung folgt, weil beliebige x und y sich als Linearkombinationen der Vektoren der Standardbasis schreiben lassen. □

Wie schon erwähnt, gilt Folgendes:

Satz 25.3. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. Dann hat A nur reelle Eigenwerte.*

Beweis. Später (Satz 26.5). □

Satz 25.4 (Hauptachsentransformation). *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Dann existiert eine orthogonale Matrix $S \in \mathrm{SO}(n)$, so dass*

$$S^{\mathrm{t}} A\, S = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

mit $\lambda_i \in \mathbb{R}$.

Beweis. Sei $\lambda \in \mathbb{R}$ ein Eigenwert und $v \in \mathbb{R}^n$ ein zugehöriger Eigenvektor mit Länge $\|v\| = 1$. Sei

$$W = v^{\perp} = \{w \in \mathbb{R}^n \mid \langle w, v\rangle = 0\}$$

der zu v orthogonale Untervektorraum (s. Abb. 25.2).

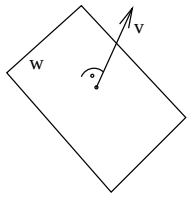


Abbildung 25.2. Das orthogonale Komplement $W = v^{\perp}$ eines Vektors v.

Wir zeigen: $Aw \in W\ \forall w \in W$:

$$\begin{aligned} \langle Aw, v\rangle &= \langle w, Av\rangle && \text{(weil } A \text{ symmetrisch ist)} \\ &= \langle w, \lambda v\rangle && \text{(weil } v \text{ ein Eigenvektor ist)} \\ &= \lambda\langle w, v\rangle \\ &= \lambda \cdot 0 \\ &= 0. \end{aligned}$$

Dies zeigt: $Aw \in v^{\perp} = W$.

Wir wählen nun eine Basis von W aus zueinander senkrecht stehenden normierten Vektoren $v_2, \ldots, v_n$ (die explizite Konstruktion solcher Vektoren liefert das Gram–Schmidt–Verfahren in Satz 26.13; in diesem Abschnitt

erläutern wir es noch nicht, weil wir zunächst den Schwerpunkt auf die geometrische Anwendung legen möchten). Wir setzen dann $v_1 := v$ und $S := (v_1, v_2, \ldots, v_n) \in \mathrm{GL}(n, \mathbb{R})$. Damit gilt: $\langle v_i, v_i \rangle = 1 \ \forall i$, $\langle v_i, v_j \rangle = 0 \ \forall i \neq j$. Insbesondere ist S nach Definition von $O(n)$ orthogonal, weil $S^{\mathrm{t}} S = E$. Wir setzen ferner: $w_j := Av_j \in W, \ j = 2, \ldots, n$. Es ergibt sich:

$$S^{\mathrm{t}} AS = S^{\mathrm{t}}(\lambda v_1, w_2, \ldots, w_n) = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & & & \\ 0 & & B & \\ 0 & & & \end{pmatrix}.$$

Die erste Zeile und Spalte folgen dabei aus $\langle v_i, v_j \rangle = \delta_{ij}$ und die Matrix $B \in \mathbb{R}^{(n-1)\times(n-1)}$ hat, da A symmetrisch ist und wir daher Bemerkung 25.2 angewenden können, die Einträge

$$v_i{}^{\mathrm{t}} w_j = v_i^t A v_j = \langle v_i, Av_j \rangle = \langle Av_i, v_j \rangle = v_j{}^{\mathrm{t}} w_i \ \text{ für } 2 \leq i, j \leq n,$$

ist also symmetrisch.

Per Induktion folgt, dass wir erreichen können, dass $S^{\mathrm{t}} A S$ die angegebene Form hat. Da $S \in \mathrm{O}(n)$, können wir durch Übergang von v zu $-v$ sogar $S \in \mathrm{SO}(n)$ erreichen. □

Beispiel 25.5. Sei $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$. Das charakteristische Polynom ist:

$$\chi_A(t) = (2-t)^2 - 1 = t^2 - 4t + 3 = (t-1)(t-3),$$

die Eigenwerte sind also: $\lambda_1 = 1, \lambda_2 = 3$. Der Eigenraum zum ersten dieser beiden ist:

$$\mathrm{Eig}(A, \lambda_1) = \mathrm{Ker} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \left\langle \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\rangle.$$

Es gilt: $\left\| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\| = \sqrt{2}$. Um einen normierten Vektor zu erhalten, setzen wir:

$$v_1 := v := \begin{pmatrix} \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{pmatrix} \left(\Rightarrow v_1^{\perp} = \left\langle \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\rangle \right).$$

Damit gilt: $A \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \in v^{\perp}$, wie behauptet.

Als Basis von $v^{\perp}$ wählen wir $v_2 = \begin{pmatrix} -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{pmatrix}$. Die gesuchte Matrix ist demnach:

$$S = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix}.$$

Nun können wir leicht nachrechnen, dass $\det S = 1$ und dass

$$S^{\mathrm{t}} A S = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} = \begin{pmatrix} & \cdots & \end{pmatrix} \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{3}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{3}{2}\sqrt{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

25.2 Klassifikation von Quadriken

Definition 25.6. *Eine Quadrik $Q \subseteq \mathbb{R}^n$ ist die **Lösungsmenge** einer quadratischen Gleichung*

$$q(x) = \sum_{i,j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} b_i x_i + c = 0,$$

*auch genannt **Nullstellenmenge** eines quadratischen Polynoms. In Matrixschreibweise:*

$$q(x) = x^{\mathrm{t}} A\, x + b^{\mathrm{t}} x + c,$$

wobei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ symmetrisch gewählt sei. Die Matrix

$$\tilde{A} = \begin{pmatrix} c & \frac{b_1}{2} & \dots & \frac{b_n}{2} \\ \frac{b_1}{2} & a_{11} & \dots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{b_n}{2} & a_{n1} & \dots & a_{nn} \end{pmatrix}$$

heißt erweiterte Matrix von q. Damit gilt dann:

$$q(x) = (1, x_1, \dots, x_n) \cdot \tilde{A} \cdot \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Beispiel 25.7. Einige Quadriken kennt man vermutlich schon aus der Schule. Beispielsweise liefert der Satz von Pythagoras unmittelbar, dass ein Kreis mit Radius r um den Ursprung im $\mathbb{R}^2$ beschrieben werden kann durch die Gleichung (s. Abb. 25.3):

$$x^2 + y^2 = r^2.$$

Denn drei positive reelle Zahlen a, b, c mit $a, b \leq c$ bilden genau dann ein rechtwinkliges Dreieck, wenn sie die Beziehung $a^2 + b^2 = c^2$ erfüllen. In drei Variablen liefert analog $x^2 + y^2 + z^2 = r^2$ eine Kugel.

Wie aber sehen allgemeinere Quadriken aus? Der folgende Satz liefert die Antwort:

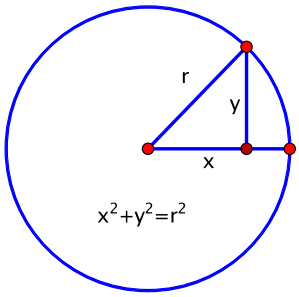


Abbildung 25.3. Der Kreis als Nullstellenmenge.

Satz/Definition 25.8 (Klassifikation von Quadriken in Dimensionen n). *Sei $q(x) = x^t A\, x + b^t x + c$ ein quadratisches Polynom mit reellen Koeffizienten und $\tilde{A}$ die erweiterte Matrix. Dann gibt es eine **Bewegung** (auch **euklidische Bewegung**), d.h. eine Abbildung*

$$f\colon \mathbb{R}^n \to \mathbb{R}^n, \quad f(y) = Sy + t, \ \text{ mit } S \in \mathrm{SO}(n), t \in \mathbb{R}^n,$$

*so dass $q(f(y)) = 0$ zu einer der folgenden Gleichungen (die auch **Normalformen** genannt werden) äquivalent ist. Dabei schreiben wir: $m = \operatorname{rang} A$, $\tilde{m} = \operatorname{rang} \tilde{A}$:*

(a) ($\tilde{m} = m$)

$$\frac{y_1^2}{\alpha_1^2} + \cdots + \frac{y_k^2}{\alpha_k^2} - \frac{y_{k+1}^2}{\alpha_{k+1}^2} - \cdots - \frac{y_m^2}{\alpha_m^2} = 0,$$

(b) ($\tilde{m} = m + 1$)

$$\frac{y_1^2}{\alpha_1^2} + \cdots + \frac{y_k^2}{\alpha_k^2} - \frac{y_{k+1}^2}{\alpha_{k+1}^2} - \cdots - \frac{y_m^2}{\alpha_m^2} = 1,$$

(c) ($\tilde{m} = m + 2$)

$$\frac{y_1^2}{\alpha_1^2} + \cdots + \frac{y_k^2}{\alpha_k^2} - \frac{y_{k+1}^2}{\alpha_{k+1}^2} - \cdots - \frac{y_m^2}{\alpha_m^2} = y_{m+1}.$$

Hierbei sind $\alpha_1, \ldots, \alpha_m \in \mathbb{R}_{>0}$ Konstanten und $0 \le k \le m$.

Bemerkung 25.9. 1) Da sich $\tilde{A}$ von A nur durch eine zusätzliche Zeile und Spalte unterscheidet, ist klar:

$$\tilde{m} \le m + 2.$$

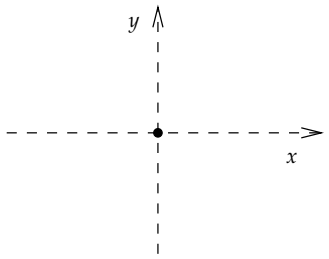
2) Ist $\det A \neq 0$, dann ist $m = n$ und $\tilde{m} \le n + 1$. Am *häufigsten* tritt Fall (b) mit $m = n$ ein ($\Leftrightarrow \det A \neq 0, \det \tilde{A} \neq 0$).

3) Genau wie $S \in \mathrm{SO}(n)$ behält eine Bewegung offenbar Abstände bei, da die Verschiebung um $t \in \mathbb{R}^n$ hierauf keinen Einfluss hat. Daher ist die Form einer Quadrik nach Anwendung der Bewegung identisch mit der Ausgangsform. Beispielsweise ist die Normalform eines Kreises nicht etwa eine belibiege Ellipse, sondern ein gleich großer Kreis mit Mittelpunkt im Ursprung.

Beispiel 25.10 ($n = 2$). Wir betrachten Quadriken in der Ebene $\mathbb{R}^2$. Es ergeben sich nach dem Klassifikationssatz folgende Fälle:

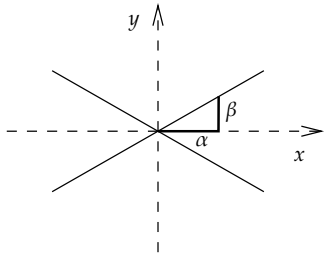
$m = \tilde{m} = 2, k = 2$: Ein Punkt:

$$\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} = 0$$



$m = \tilde{m} = 2, k = 1$: Zwei Geraden mit Steigungen $\frac{\beta}{\alpha}, -\frac{\beta}{\alpha}$:

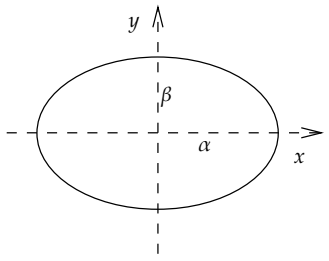
$$0 = \frac{x^2}{\alpha^2} - \frac{y^2}{\beta^2} = \left(\frac{x}{\alpha} + \frac{y}{\beta}\right)\left(\frac{x}{\alpha} - \frac{y}{\beta}\right)$$



$m = \tilde{m} = 2, k = 0$: Dies führt nach Multiplikation der Gleichung mit -1 wieder auf den schon betrachteten Fall mit $k = 2$ (ein Punkt).

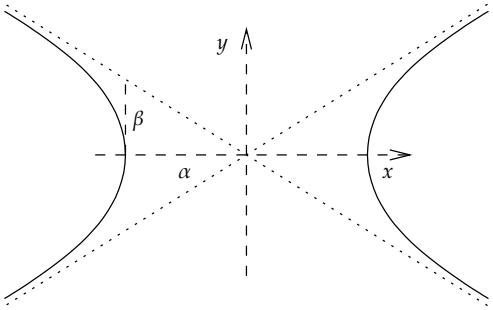
$m = 2, \tilde{m} = 3, k = 2$: Eine Ellipse mit Halbachsen der Längen α, β:

$$\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} = 1$$



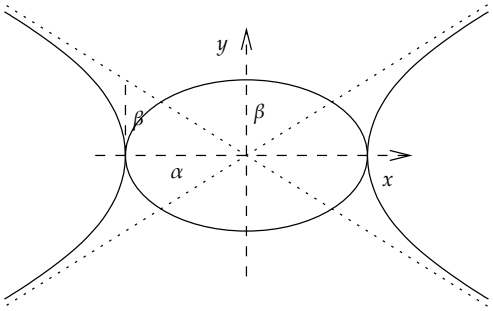
$m = 2, \tilde{m} = 3, k = 1$: Eine Hyperbel mit Halbachsen der Längen α, β:

$\frac{x^2}{\alpha^2} - \frac{y^2}{\beta^2} = 1$



Die folgende Abbildung zeigt Ellipse und Hyperbel mit den gleichen Halbachsen in einem Bild, um deren Zusammenhang deutlich zu machen:

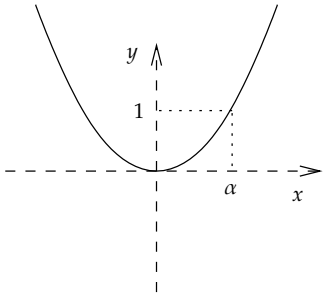
$\frac{x^2}{\alpha^2} \pm \frac{y^2}{\beta^2} = 1$



$m = 2, \tilde{m} = 3, k = 0$: $-\frac{x^2}{\alpha^2} - \frac{y^2}{\beta^2} = 1$ die leere Menge: $\emptyset$

$m = 1, \tilde{m} = 3, k = 1$: Eine Parabel:

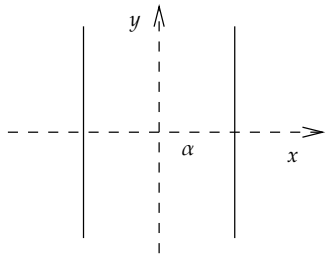
$\frac{x^2}{\alpha^2} = y$



$m = 1, \tilde{m} = 3, k = 0$: Dieser Fall ist nach Durchmultiplizieren mit -1 analog zum vorigen und liefert eine Parabel (allerdings nach unten offen).

$m = 1, \tilde{m} = 2, k = 1$: Zwei Geraden mit Abstand 2α:

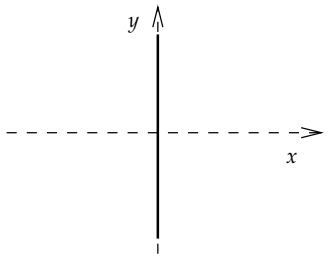
$\frac{x^2}{a^2} = 1 \ (\Leftrightarrow (\frac{x}{a}+1)(\frac{x}{a}-1) = 0)$



$m = 1, \tilde{m} = 2, k = 0$: $\frac{x^2}{a^2} = 1$: die leere Menge: $\emptyset$

$m = 1 = \tilde{m}, k = 1$: Eine doppelte Gerade:

$\frac{x^2}{a^2} = 0$



$m = 1 = \tilde{m}, k = 0$: analog zu eben nach Durchmultiplikation mit -1.

Beispiel 25.11. Wir möchten herausfinden, welchen Typ die folgende Quadrik hat:

$$q(x, y) = x^2 - xy + y^2 - x - y - 1 = 0.$$

Dazu schreiben wir sie zunächst mit Hilfe der erweiterten Matrix $\tilde{A}$:

$$q(x, y) = (1, x, y) \cdot \begin{pmatrix} -1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x \\ y \end{pmatrix}.$$

Es gilt $q(x, y) = 0 \Leftrightarrow 2q(x, y) = 0$; wir dürfen also statt unserer ursprünglichen Gleichung $q(x, y) = 0$ für die Quadrik auch die Gleichung $2q(x, y) = 0$ verwenden:

$$2q(x, y) = (1, x, y) \begin{pmatrix} -2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ x \\ y \end{pmatrix}.$$

Die rechte untere 2×2–Teilmatrix A haben wir im vorigen Beispiel untersucht und berechnet, dass mit

$$S = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} \in \mathrm{SO}(2) \quad \text{gilt} \quad S^{\mathrm{t}} A S = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

Diese Matrix benutzen wir, um neue Koordinaten x' und y' zu erhalten:

$$\begin{pmatrix} x \\ y \end{pmatrix} = S \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{2}x' - \frac{1}{2}\sqrt{2}y' \\ \frac{1}{2}\sqrt{2}x' + \frac{1}{2}\sqrt{2}y' \end{pmatrix}.$$

In diesen Koordinaten lautet die Gleichung der Quadrik $q(x, y) = 0$ nun:

$$\begin{aligned} 0 = q'(x', y') &= 3(x')^2 + (y')^2 - (\sqrt{2}x' + \sqrt{2}y') - (\sqrt{2}x' + \sqrt{2}y') - 2 \\ &= 3(x')^2 + (y')^2 - 2\sqrt{2}y' - 2 \\ &= 3(x')^2 + (y' - \sqrt{2})^2 - 4. \end{aligned}$$

Wir nehmen nun eine weitere Koordinatentransformation vor: $x'' = x', y'' = y' - \sqrt{2}$. In diesen Koordinaten lautet die Gleichung der Quadrik: $\frac{3}{4}(x'')^2 + \frac{1}{4}(y'')^2 = 1$ bzw.

$$\frac{(x'')^2}{(\frac{2}{3}\sqrt{3})^2} + \frac{(y'')^2}{2^2} = 1.$$

Sie hat also eine Normalform, die wir in der Liste aus Beispiel 25.10 finden: Die Quadrik $q(x, y) = 0$ ist demnach eine Ellipse. Um diese Ellipse auch in ihren ursprünglichen Koordinaten zeichnen zu können, drücken wir nun die neuen Koordinaten x'' und y'' in den alten Koordinaten x und y aus:

$$\begin{pmatrix} x \\ y \end{pmatrix} = S \begin{pmatrix} x' \\ y' \end{pmatrix} = S \begin{pmatrix} x'' \\ y'' + \sqrt{2} \end{pmatrix} \Rightarrow \begin{pmatrix} x'' \\ y'' + \sqrt{2} \end{pmatrix} = S^{\mathrm{t}} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} x'' \\ y'' \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix}.$$

Abbildung 25.4 zeigt die Ellipse sowohl in den neuen Koordinaten x'' und y'', als auch in den alten Koordinaten x und y.

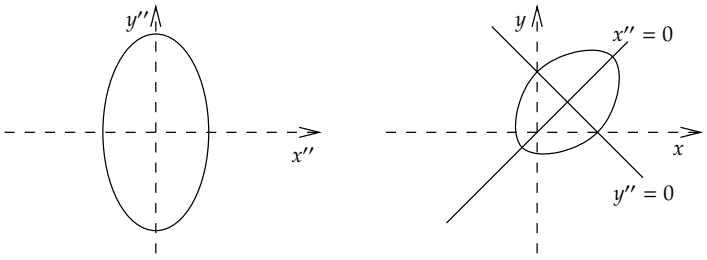


Abbildung 25.4. Eine Ellipse in neuen und in alten Koordinaten.

Bemerkung/Definition 25.12. Quadriken in der Ebene heißen auch **Kegelschnitte**, weil sie durch den Schnitt eines Kegels mit einer Ebene entstehen. Dies wusste bereits Apollonius von Perge (262 - 190 v.Chr.). Abbildung 25.5 zeigt den Kegel mit Gleichung $K: x^2 + y^2 = z^2$ und einige Schnitte. Einsetzen

Abbildung 25.5. Einige Schnitte eines Kegels.

von $z = r$ in K zeigt beispielsweise sofort, dass ein Schnitt mit der Ebene $z = r$ einen Kreis mit Radius r aus K ausschneidet. Eine Hyperbel erhält man durch Schnitt mit $y = r$. Der Leser kann sich leicht selbst überlegen, wie man die weiteren Kegelschnitte erhält.

Bemerkung 25.13 (Brennpunkte von Ellipsen). Kegelschnitte haben viele interessante Eigenschaften. Beispielsweise haben Ellipsen zwei sogenannte **Brennpunkte** (s. Abb. 25.6): Ein Lichtstrahl, der von einem der beiden Brennpunkte in eine beliebige Richtung ausgesendet wird, wird an der Ellipse so reflektiert, dass er durch den anderen Brennpunkt läuft. Die entsprechende

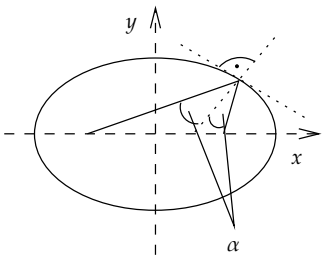


Abbildung 25.6. Die Brennpunktseigenschaft von Ellipsen.

Eigenschaft im Dreidimensionalen wurde beispielsweise in einigen Burgen benutzt, um Besucher, die eine heimliche Unterredung führen wollten, abzuhören: Bei einer Decke, die Ellipsoidenform hat, muss man nur die Besucher in den einen Brennpunkt stellen und im anderen Brennpunkt stehen, um der Unterhaltung zu lauschen, auch wenn sie flüsternd von statten geht.

Tatsächlich nutzt man die Brennpunktseigenschaft auch heute aus, insbesondere in der Variante für Parabeln. Diese haben nämlich nur einen Brennpunkt, siehe Abb. 25.7. Ein Parabolspiegel, der die Form einer um ihre Symmetrie-

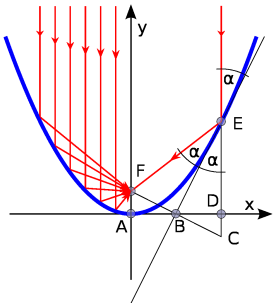


Abbildung 25.7. Der Brennpunkt einer Parabel.

achse rotierenden Parabel besitzt (genannt Paraboloid, siehe Abschnitt 25.3) hat dann auch nur einen Brennpunkt. Platziert man in diesem den Energieumwandler bzw. Empfänger, so ist es effizient möglich, Daten aus dem All zu empfangen oder Strom aus Sonnenenergie zu gewinnen.

Für viele weitere Eigenschaften von Quadriken und anderen geometrischen Objekten ist [HCV32] ein sehr lesenswertes Buch.

Beweis (des Satzes 25.8 zur Klassifikation der Quadriken im $\mathbb{R}^n$). Nach dem Satz über die Hauptachsentransformation $\exists S \in \mathrm{SO}(n)$, so dass

$$S^{\mathrm{t}} A S = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} = D$$

Ist k die Anzahl der positiven Eigenwerte, $m = \operatorname{rang} A$, dann können wir also ohne Beschränkung der Allgemeinheit (o.B.d.A.) annehmen, dass

$$A = \begin{pmatrix} \frac{1}{\alpha_1^2} & & & & & & 0 \\ & \ddots & & & & & \\ & & \frac{1}{\alpha_k^2} & & & & \\ & & & -\frac{1}{\alpha_{k+1}^2} & & & \\ & & & & \ddots & & \\ & & & & & -\frac{1}{\alpha_m^2} & \\ & & & & & & 0 \\ 0 & & & & & & 0 \end{pmatrix}$$

für gewisse $\alpha_i \in \mathbb{R}_{>0}$. Bezüglich der Koordinaten $y \in \mathbb{R}^n$ mit

$$x = Sy$$

schreibt sich $q(x) = x^t A\, x + b^t\, x + c$ nun:

$$q(Sy) = \sum_{i=1}^{k} \frac{y_i^2}{\alpha_i^2} - \sum_{j=k+1}^{m} \frac{y_j^2}{\alpha_j^2} + \sum_{l=1}^{n} \tilde{b}_l y_l + c$$

für gewisse $\tilde{b}_l$ (genauer: $\tilde{b}^t = b^t\, S$). Also ist die erweiterte Matrix bezüglich der y–Koordinaten von der Gestalt:

c	$\frac{\tilde{b}_1}{2}$ … … … … $\frac{\tilde{b}_m}{2}$	$\frac{\tilde{b}_{m+1}}{2}$ … $\frac{\tilde{b}_n}{2}$
$\frac{\tilde{b}_1}{2}$	$\frac{1}{\alpha_1^2}$	
⋮	⋱	
⋮	$\frac{1}{\alpha_k^2}$	
⋮	$-\frac{1}{\alpha_{k+1}^2}$	0
⋮	⋱	
$\frac{\tilde{b}_m}{2}$	$-\frac{1}{\alpha_m^2}$	
$\frac{\tilde{b}_{m+1}}{2}$		
⋮	0	0
$\frac{\tilde{b}_n}{2}$		

Wir möchten q auf noch schönere Form bringen. Dies erreichen wir durch die Translationen:

$$\tilde{y}_i = \begin{cases} y_i - \frac{\tilde{b}_i \alpha_i^2}{2}, & i \in \{1, 2, \ldots, k\} \\ y_i + \frac{\tilde{b}_i \alpha_i^2}{2}, & i \in \{k+1, \ldots, m\}. \end{cases}$$

q hat dann die Gestalt (d.h. die linearen Terme $\tilde{b}_l \tilde{y}_l$ sind für $l \leq m$ nicht vorhanden):

$$q(\tilde{y}) = \sum_{i=1}^{k} \frac{\tilde{y}_i^2}{\tilde{\alpha}_i^2} - \sum_{j=k+1}^{m} \frac{\tilde{y}_j^2}{\tilde{\alpha}_j^2} + \sum_{l=m+1}^{n} \tilde{b}_l \tilde{y}_l + \tilde{c}.$$

Die erweiterte Matrix sieht jetzt also folgendermaßen aus:

$\tilde{c}$	0	$\frac{\tilde{b}_{m+1}}{2} \cdots \frac{\tilde{b}_n}{2}$
0	$\ddots$	0
$\frac{\tilde{b}_{m+1}}{2}$ $\vdots$ $\frac{\tilde{b}_n}{2}$	0	0

Sind alle $\tilde{b}_{m+1} = \cdots = \tilde{b}_n = 0$ und $\tilde{c} = 0$ so sind wir in Fall (a).

Ist aber $\tilde{b}_{m+1} = \cdots = \tilde{b}_n = 0, \tilde{c} \neq 0$, so liefert Division durch $\tilde{c}$ den Fall (b).

Ist schließlich $(\tilde{b}_{m+1}, \ldots, \tilde{b}_n) \neq (0, \ldots, 0)$, so können wir neue Koordinaten $y'_{m+1}, \ldots, y'_n$ einführen, so dass

$$\sum_{l=m+1}^{n} \tilde{b}_l \tilde{y}_l = b'_{m+1} y'_{m+1}.$$

Translation (um $\tilde{c}$ auf null zu bringen) und Division durch b'_{m+1} liefert dann den Fall (c). □

25.3 Klassifikation von Quadriken im Fall $n = 3$

Im Fall $n = 3$, d.h. im $\mathbb{R}^3$, haben wir also eine Quadrik:

$$q(x, y, z) = (1\ x_1\ x_2\ x_3) \cdot \tilde{A} \cdot \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

wobei die erweiterte Matrix $\tilde{A}$ zu $A = (a_{ij})$ symmetrisch ist und die Form hat:

$$\tilde{A} = \begin{pmatrix} c & \frac{b_1}{2} & \frac{b_2}{2} & \frac{b_3}{2} \\ \frac{b_1}{2} & a_{11} & a_{12} & a_{13} \\ \frac{b_2}{2} & a_{21} & a_{22} & a_{23} \\ \frac{b_3}{2} & a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Wir schreiben: $m = \operatorname{rang} A$, $\tilde{m} = \operatorname{rang} \tilde{A}$. Damit gibt es die folgenden Fälle (bei den Graphiken ist das Koordinatensystem häufig etwas gedreht, damit man die Geometrie der Quadrik besser erkennen kann; unendlich große Oberflächen sind mit einer Kugel abgeschnitten):

1) $m = 3, \tilde{m} = 4, k = 3$, ein **Ellipsoid** (Abb. 25.8):

$$\frac{x_1^2}{\alpha_1^2} + \frac{x_2^2}{\alpha_2^2} + \frac{x_3^2}{\alpha_3^2} = 1$$

Abbildung 25.8. Ein Ellipsoid.

2) $m = 3 = \tilde{m}, k = 2$, ein **Kegel** (auch **Doppelkegel** genannt, s. Abb. 25.9):

$$\frac{x_1^2}{\alpha_1^2} + \frac{x_2^2}{\alpha_2^2} - \frac{x_3^2}{\alpha_3^2} = 0$$

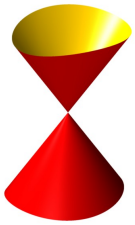


Abbildung 25.9. Ein Kegel.

3) $m = 3, \tilde{m} = 4, k = 1, 2$, Hyperboloiden (Abb. 25.10): Ein **einschaliger Hyperboloid** ist durch

$$\frac{x_1^2}{\alpha_1^2} + \frac{x_2^2}{\alpha_2^2} - \frac{x_3^2}{\alpha_3^2} = 1$$

gegeben ($k = 2$). Einen **zweischaligen Hyperboloiden** erhält man durch Änderung des Vorzeichens auf der rechten Seite (bzw. durch $k = 1$):

$$\frac{x_1^2}{\alpha_1^2} + \frac{x_2^2}{\alpha_2^2} - \frac{x_3^2}{\alpha_3^2} = -1 \quad \Leftrightarrow \quad \frac{x_1^2}{\alpha_1^2} - \frac{x_2^2}{\alpha_2^2} - \frac{x_3^2}{\alpha_3^2} = 1.$$

Die beiden Hyperboloiden entstehen auch als **Deformation** des Kegels, indem die α_i immer größer gewählt werden oder äquivalent auf der rechten Seite statt der 1 ein ε immer näher an 0 gewählt wird (Abb. 25.11):

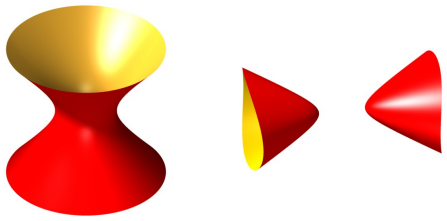


Abbildung 25.10. Ein– und zweischaliger Hyperboloid.

$$\frac{x_1^2}{\alpha_1^2} \pm \frac{x_2^2}{\alpha_2^2} - \frac{x_3^2}{\alpha_3^2} = \varepsilon.$$

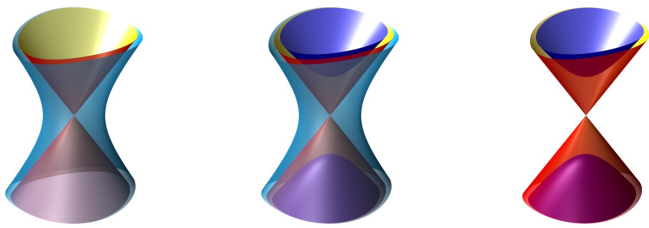


Abbildung 25.11. Hyperboloiden als Deformationen des Kegels: links der ein–, rechts der zweischalige und in der Mitte beide Deformationen gemeinsam.

4) $m = 2, \tilde{m} = 4$, Paraboloide (Abb. 25.12):

   Es gibt den **elliptischen Paraboloiden** ($k = 2$),

$$\frac{x_1^2}{\alpha_1^2} + \frac{x_2^2}{\alpha_2^2} = x_3,$$

   und den **hyperbolischen Paraboloiden** ($k = 1$):

$$\frac{x_1^2}{\alpha_1^2} - \frac{x_2^2}{\alpha_2^2} = x_3.$$

5) Allgemein nennt man jede Quadrik, deren Normalform nur von zwei Variablen abhängt, **Zylinder**.

   $m = 2, \tilde{m} = 3$: Hier (Abb. 25.13) erhalten wir einen **elliptischen Zylinder** ($k = 2$) und einen **hyperbolischen Zylinder** ($k = 1$):

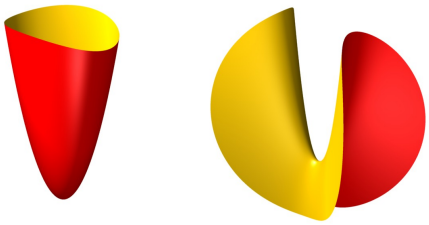


Abbildung 25.12. Ein elliptischer und ein hyperbolischer Paraboloid.

$$\frac{x_1^2}{\alpha_1^2} \pm \frac{x_2^2}{\alpha_2^2} = 1.$$

Ist beim elliptischen speziell $\alpha_1 = \alpha_2$, so nennt man diesen auch **Kreiszylinder**.

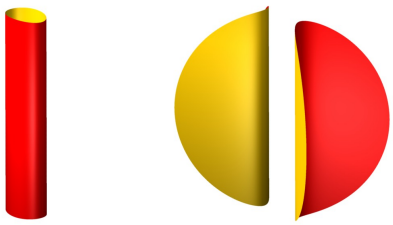


Abbildung 25.13. Elliptischer und hyperbolischer Zylinder.

$m = 1, \tilde{m} = 3$: Offenbar ist dies auch ein Zylinder, und zwar ein **parabolischer Zylinder** (Abb. 25.14):

$$\frac{x_1^2}{\alpha_1^2} = x_2.$$

$m = 2, \tilde{m} = 2$: Natürlich sind dies prinzipiell zwar auch Zylinder. Allerdings sind sie so speziell, dass man sie üblicherweise nicht als solche bezeichnet.

Für $k = 1$ erhalten wir nämlich

$$\frac{x_1^2}{\alpha_1^2} - \frac{x_2^2}{\alpha_2^2} = 0,$$

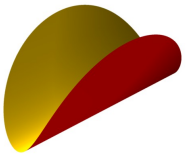


Abbildung 25.14. Ein parabolischer Zylinder.

Abbildung 25.15. Zwei Ebenen.

was (wegen einer binomischen Formel) in zwei lineare Faktoren, d.h. in zwei Ebenen, zerfällt.

Für $k = 2$ ergibt sich

$$\frac{x_1^2}{\alpha_1^2} + \frac{x_2^2}{\alpha_2^2} = 0,$$

was genau von den Punkten $(0, 0, x_3) \in \mathbb{R}^3$ mit $x_3 \in \mathbb{R}$ beliebig, erfüllt wird. Geometrisch erhalten wir also eine Gerade (Abb. 25.16), nämlich die x_3–Achse.

Abbildung 25.16. Eine Gerade im $\mathbb{R}^3$ als Quadrik.

6) Selbstverständlich gibt es noch einige weitere Fälle. Beispielsweise ist es uns möglich, auch die leere Menge oder einen Punkt als Quadrik im $\mathbb{R}^3$ erhalten. Diese Fälle wird der Leser ohne Mühe angeben können.

Bemerkung 25.14 (Echtzeit–Visualisierung algebraischer Flächen). Lässt man, statt wie bei Quadriken, auch Polynome höheren Grades in drei Variablen zu, so heißen deren Nullstellenmengen im $\mathbb{R}^3$ **algebraische Flächen**. Deren Visualisierung ist in letzter Zeit auf dem Grenzgebiet zwischen Mathematik und Informatik sehr aktuell, da dies durch Verwendung von schnellen Graphikkarten in Echtzeit möglich ist. Leider ist noch keines der existierenden Programme zufriedenstellend in der Hinsicht, dass es sowohl schnell genug ist als auch korrekte Ergebnisse liefert.

Die Software `surfer` (`http://www.surfer.Imaginary-Exhibition.com`), die eine einfache Benutzerschnittstelle bereit stellt und die vom Mathematischen Forschungsinstitut Oberwolfach gemeinsam mit dem Autor O. Labs für die Wanderausstellung *Imaginary* entwickelt wurde, ermöglicht es, die hier vorgestellten Graphiken ohne großen Aufwand selbst zu erzeugen.

`surfer` (wie auch dessen Vorgänger, das zwar mächtigere, aber nicht so einfach zu installierende, von O. Labs entwickelte Programm `surfex`), benutzt im Hintergrund das Programm `surf`, das wiederum auf der Visualisierungs–Technik des Raytracings basiert. Da hierbei einfach endlich viele Strahlen von einem virtuellen Auge ausgesandt werden, wird es selbstverständlich vorkommen, dass besonders kleine oder dünne Objekte, wie z.B. eine Gerade, meist gar nicht von diesen Strahlen getroffen werden und daher im Bild (unkorrekterweise) weggelassen werden. Für das Bild der Gerade in der obigen Liste (Abb. 25.16) haben wird daher ein wenig geschummelt und statt dessen die Gleichung eines dünnen Kreiszylinders visualisiert. Solche und verwandte Probleme demnächst automatisiert und in Echtzeit lösen zu können, ist noch immer Aufgabe der Forschung auf diesem Gebiet.

25.4 Typen von Quadriken

Mit der Hauptachsentransformation und einigen weiteren Koordinaten–Änderungen kann man, wie wir gesehen haben, die Normalform für jede Quadrik bestimmen.

Manchmal interessiert man sich aber nur für den **Typ einer Quadrik**, d.h. für die Normalform, wenn man auch Streckungen und Stauchungen in den Koordinaten erlaubt, so dass also alle α_i im Klassifikationssatz = 1 gewählt werden können.

Dies kann man oft wesentlich einfacher erreichen, insbesondere ohne das Berechnen der Eigenwerte und –vektoren. Betrachten wir dazu das Beispiel 25.11 von oben noch einmal:

$$q(x, y) = x^2 - xy + y^2 - x - y - 1 = 0.$$

Mit **quadratischer Ergänzung** können wir zunächst den gemischten Term $-xy$ eleminieren:

$$q(x, y) = \Big(x - \frac{1}{2}y\Big)^2 - \frac{1}{4}y^2 + y^2 - x - y - 1.$$

$\frac{1}{4}y^2$ müssen wir wieder abziehen, da wir den Fehler, den wir beim Ersetzen von x^2 durch $(x - \frac{1}{2}y)^2$ gemacht haben, wieder beheben müssen. Führen wir nun neue Koordinaten ein,

$$\tilde{x} = x - \frac{1}{2}y, \quad \tilde{y} = y,$$

so erhalten wir, da dann $x = \tilde{x} + \frac{1}{2}y$ ist, und da wir weiter mit quadratischer Ergänzung die linearen Terme eliminieren können:

$$\begin{aligned} q(x, y) &= \tilde{x}^2 + \frac{3}{4}\tilde{y}^2 - \tilde{x} - \frac{1}{2}\tilde{y} - \tilde{y} - 1 \\ &= \tilde{x}^2 - \tilde{x} + \frac{3}{4}\Big(\tilde{y}^2 - 2\tilde{y}\Big) - 1 \\ &= \Big(\tilde{x} - \frac{1}{2}\Big)^2 - \frac{1}{4} + \frac{3}{4}\Big(\tilde{y} - 1\Big)^2 - \frac{3}{4} - 1 \\ &= x'^2 + \frac{3}{4}y'^2 - 2, \end{aligned}$$

wenn wir $x' = \tilde{x} - \frac{1}{2}$ und $y' = \tilde{y} - 1$ als neue Koordinaten wählen. Die Gleichung $x'^2 + \frac{3}{4}y'^2 - \frac{9}{4} = 0$ beschreibt offenbar eine Ellipse.

Doch warum bedeutet dieses, dass die Ursprungsquadrik in den Koordinaten x, y ebenfalls eine Ellipse ist? Im Gegensatz zu den im Klassifikationssatz erlaubten Koordinatentransformationen, die durch eine orthogonale Matrix gegeben sind, haben wir hier andere lineare Koordinatentransformationen vorgenommen.

Man müsste nun beweisen, dass tatsächlich eine Abbildung $x \mapsto Ax + t$ mit $A \in \mathrm{GL}(2, \mathbb{R})$, $t \in \mathbb{R}^n$, eine Quadrik eines bestimmen Typs wieder auf eine des gleichen Typs abbildet. In einer Geometrie–Vorlesung würde man dies selbstverständlich durchführen, da es nicht sehr schwierig ist, doch hier können wir aus Zeitgründen nicht weiter darauf eingehen.

Wir möchten hier nur noch einmal auf das obige Beispiel zurückkommen. Wir erhalten schließlich eine Ellipse, was beweist, dass die ursprüngliche Quadrik ebenfalls vom Typ *Ellipse* war. Allerdings ist *Kreis* kein Typ einer Quadrik im obigen Sinn, denn eine Veränderung der α_i macht aus einem Kreis ja eine Ellipse. Außerdem lassen allgemeine Koordinatentransformationen (mit $A \in \mathrm{GL}(n, \mathbb{R})$, nicht unbedingt in $\mathrm{SO}(n)$) Abstände nicht unbedingt fest, so dass natürlich dabei problemlos aus einem Kreis eine Ellipse werden kann, die kein Kreis ist, und umgekehrt.

Aufgaben

Aufgabe 25.1 (Kegelschnitte). Stellen Sie fest, zu welchem Typ die folgenden Kegelschnitte (d.h. Quadriken in der Ebene $\mathbb{R}^2$, in zwei Variablen) gehören und zeichnen Sie diese, gemeinsam mit ihren Hauptachsen, jeweils in ein Koordinatensystem ein:

1. $-8x^2 + 12xy - 6x + 8y^2 - 18y + 8 = 0$,
2. $5x^2 - 8xy + 2x + 5y^2 + 2y + 1 = 0$.

Aufgabe 25.2 (Quadriken im $\mathbb{R}^3$). Stellen Sie fest, zu welchem Typ die folgende Quadrik im $\mathbb{R}^3$ gehört und zeichnen Sie sie, gemeinsam mit ihren Hauptachsen, in ein Koordinatensystem ein:

$$2x^2 + 2xy + 2y^2 - 2xz + 2z^2 - 2yz - 1 = 0.$$

Aufgabe 25.3 (Geraden auf Quadriken).

1. Zeigen Sie, dass auf einem einschaligen Hyperboloid sowie auf einem hyperbolischen Paraboloid zwei Scharen von ∞ vielen Geraden liegen und dass diese die folgende Eigenschaft besitzen: Jede Gerade schneidet zwar keine der Geraden der eigenen Schar, schneidet aber mit nur einer Ausnahme jede Gerade der anderen Schar in genau einem Punkt.
2. Zeigen Sie, dass auf einem Ellipsoid und auf einem zweischaligen Hyperboloid keine Geraden liegen.

Aufgabe 25.4 (Drei Windschiefe Geraden definieren eine Quadrik).

1. Es seien 3 paarweise windschiefe Geraden l_1, l_2, l_3 im $\mathbb{R}^3$ gegeben. Zeigen Sie: Es gibt genau eine Quadrik, die diese 3 Geraden enthält; diese ist ein einschaliges Hyperboloid oder ein hyperbolischer Paraboloid. Wie erhält man alle Geraden aus den gegebenen dreien geometrisch?
   Hinweise: Zur Existenz der Quadrik kann man ausnutzen, dass der Raum $\mathbb{R}[x, y, z]_{\leq 2}$ der Quadriken im $\mathbb{R}^3$ 10-dimensional ist. Wie viele lineare Bedingungen an die Koeffizienten einer Quadrik sind es, eine vorgegebene Gerade zu enthalten?

   Um alle Geraden zu finden, betrachten Sie zunächst eine Ebene E, die von l_1 und einem Punkt $p \in l_2$ aufgespannt wird. E schneidet l_3 in einem Punkt q. Nun geht durch p und q eine Gerade. In wievielen Punkten kann eine Gerade eine Quadrik im $\mathbb{R}^3$ maximal schneiden, ohne in ihr zu liegen?
2. Es seien 4 paarweise windschiefe Geraden im $\mathbb{R}^3$ gegeben. Zeigen Sie: Es gibt entweder 0, 1, 2 oder ∞ viele Geraden im $\mathbb{R}^3$, die alle 4 Geraden schneiden.

26

Skalarprodukte

Im Kapitel 25 über die Hauptachsentransformation haben wir einige Resultate unbewiesen benutzt, um zunächst die Geometrie — insbesondere der Quadriken — zu betonen. Hier werden diese schon benutzten Ergebnisse gezeigt und einige verwandte wichtige Begriffe eingeführt: Im ersten Abschnitt über hermitesche Skalarprodukte werden wir (in einer allgemeineren Situation) nachweisen, dass reelle symmetrische Matrizen tatsächlich nur reelle Eigenwerte besitzen. Nach Ausführungen über allgemeinere Skalarprodukte und deren Beziehung zum Vorzeichen von reellen Eigenwerten sowie zu Normen werden wir im Abschnitt über Orthonormalisierung schließlich das Gram–Schmidt–Verfahren erklären, mit Hilfe dessen wir die im Beweis zu Satz 25.4 nicht gelöste Frage klären, wie wir eine Basis aus zueinander orthogonal stehenden Vektoren der Länge 1 explizit ohne großen Aufwand produzieren können.

26.1 Das hermitesche Skalarprodukt

Die komplexen Zahlen sind zwar nicht unser Hauptanwendungsgebiet, doch viele Phänomene lassen sich mit ihrer Hilfe wesentlich besser und konzeptueller verstehen als über den reellen Zahlen. Hierzu gehört die Frage nach der Realität von Eigenwerten symmetrischer Matrizen, die sich im komplexen Fall zu sogenannten hermiteschen Matrizen verallgemeinern, genauso wie die Hauptachsentransformation symmetrischer Matrizen mit Hilfe orthogonaler Matrizen, die sich zu sogenannten unitären Matrizen im Komplexen verallgemeinern.

Wir betrachten also den Körper $\mathbb{C}$ der komplexen Zahlen (siehe auch Abschnitt 7.1), bestehend aus Elementen der Form $z = x + iy$, wobei $x, y \in \mathbb{R}$ und $i \in \mathbb{C} : \; i^2 = -1$. Die komplexe Konjugation eines solchen Elementes $z = x + iy \in \mathbb{C}$ ist definiert als die Abbildung (siehe auch Abb. 26.1):

$$\bar{}: \mathbb{C} \to \mathbb{C},\ z = x + iy \mapsto \overline{z} := x - iy.$$

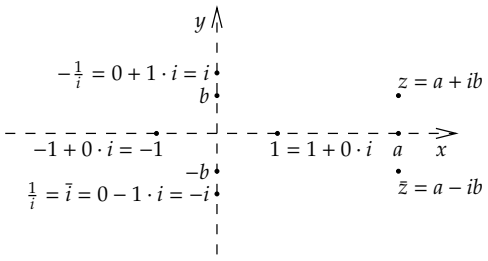


Abbildung 26.1. Die komplexe Konjugation.

Definition 26.1. *Das **hermitesche Skalarprodukt** auf $\mathbb{C}^n$ ist durch*

$$\langle z, w\rangle := \sum_{j=1}^{n} \overline{z}_j w_j, \qquad z, w \in \mathbb{C}^n,$$

definiert, d.h. $\langle z, w\rangle = \overline{z}^{\,\mathrm{t}}\, w$.

Im Gegensatz zum reellen Standard-Skalarprodukt auf $\mathbb{R}^n$ ist das hermitesche Skalarprodukt also nicht symmetrisch. Auch einige andere aus dem reellen Fall bekannten Eigenschaften müssen angepasst werden:

Proposition 26.2 (Eigenschaften des hermiteschen Skalarproduktes). *Es gilt:*

*1) **Additivität**:*

$$\langle z, v + w\rangle = \langle z, v\rangle + \langle z, w\rangle$$
$$\langle z + v, w\rangle = \langle z, w\rangle + \langle v, w\rangle$$

$\forall z, w, v \in \mathbb{C}^n$.

*2) **Sesquilinearität** (d.h. $(1 + \frac{1}{2})$–fache Linearität):*

$$\langle \lambda z, w\rangle = \overline{\lambda}\langle z, w\rangle$$
$$\langle z, \lambda w\rangle = \lambda\langle z, w\rangle$$

$\forall \lambda \in \mathbb{C}, z, w \in \mathbb{C}^n$.

3) ***Hermitesch****:*

$$\langle w, z\rangle = \overline{\langle z, w\rangle} \quad \forall z, w \in \mathbb{C}^n,$$

insbesondere:

$$\langle z, z\rangle \in \mathbb{R} \quad \forall z \in \mathbb{C}^n.$$

4) ***Positiv Definitheit****:*

$$\langle z, z\rangle \geq 0 \quad \forall z \in \mathbb{C}^n$$

und Gleichheit gilt nur für $z = 0$*.*

Beweis. Wir zeigen nur die letzte Eigenschaft, da die anderen nicht allzu schwer nachzuweisen sind. Für $z \in \mathbb{C}^n$ schreiben wir dafür $z_j = x_j + iy_j$ mit $x_j, y_j \in \mathbb{R}$. Damit ist $\overline{z_j} \cdot z_j = (x_j - iy_j)(x + iy_j) = x_j^2 + y_j^2$ und wir erhalten:

$$\langle z, z\rangle = \sum_{j=1}^{n} (x_j^2 + y_j^2) \geq 0$$

und Gleichheit gilt genau dann, wenn: $x_j = y_j = 0 \,\forall\, j \;\Leftrightarrow\; z = 0$. □

Die **zugehörige Norm** (auch: **induzierte Norm**) auf $\mathbb{C}^n$ ist

$$\|z\| = \sqrt{\langle z, z\rangle}.$$

Wie im reellen Fall ist ein **normierter** Vektor $z \in \mathbb{C}^n$ einer mit $\|z\| = 1$. Für $v \in \mathbb{C}^n, v \neq 0$, ist $\frac{v}{\|v\|}$ normiert. Auch hier heißen z und w **senkrecht zueinander**, wenn $\langle z, w\rangle = 0$, in Zeichen $z \perp w$.

Definition 26.3. *Eine Matrix* $A \in \mathbb{C}^{n\times n}$ *heißt* ***hermitesch****, wenn*

$$\langle Az, w\rangle = \langle z, Aw\rangle \quad \forall z, w \in \mathbb{C}^n.$$

Bemerkung 26.4. Da $\langle z, w\rangle = \overline{z}^{\,t} \cdot w$ und da $\overline{Az}^{\,t} w = \overline{z}^{\,t}\,\overline{A}^{\,t} w$, ist

$$\overline{Az}^{\,t} w = \langle Az, w\rangle = \langle z, Aw\rangle = \overline{z}^{\,t} A\, w$$

genau dann für alle z und w erfüllt, wenn $\overline{A}^{\,t} = A$ gilt. Insbesondere sind die Diagonaleinträge a_{kk} einer hermiteschen Matrix $A = (a_{kj})$ reell und $\overline{a_{kj}} = a_{jk}$. Jede reelle symmetrische Matrix ist offenbar auch hermitesch.

Satz 26.5. *Die Eigenwerte einer hermiteschen Matrix sind alle reell.*

Beweis. Sei $A = \overline{A}^{\,t}$ und $v \in \mathbb{C}^n \setminus \{0\}$ ein Eigenvektor von A zum Eigenwert λ, also $A{\cdot}v = \lambda{\cdot}v$. Dann gilt:

$$\lambda\langle v, v\rangle = \langle v, \lambda v\rangle = \langle v, Av\rangle = \langle Av, v\rangle = \langle \lambda v, v\rangle = \overline{\lambda}\langle v, v\rangle.$$

Dies zeigt: $(\lambda - \overline{\lambda})\langle v, v\rangle = 0 \Rightarrow \lambda = \overline{\lambda}$, d.h. $\lambda \in \mathbb{R}$. □

Als direktes Korollar erhalten wir Satz 25.3 über die Realität der Eigenwerte von symmetrischen Matrizen. Nun zur Verallgemeinerung der Hauptachsentransformation symmetrischer Matrizen durch orthogonale auf jene hermitescher durch sogenannte unitäre Matrizen:

Proposition/Definition 26.6. *Eine Matrix $S \in \mathrm{GL}(n,\mathbb{C})$ heißt **unitär**, wenn*

$$\overline{S}^{\,t} \cdot S = E.$$

Mit

$$\mathrm{U}(n) = \left\{ S \in \mathrm{GL}(n,\mathbb{C}) \;\middle|\; \overline{S}^{\,t} \cdot S = E \right\}$$

*bezeichnen wir die **unitäre Gruppe**. Die Gruppe $\mathrm{SU}(n)$ der **speziellen unitären Matrizen** ist:*

$$\mathrm{SU}(n) := \{ S \in \mathrm{U}(n) \mid \det S = 1 \}.$$

Beweis. Wir müssen nachrechnen, dass $\mathrm{U}(n)$ und $\mathrm{SU}(n)$ tatsächlich Gruppen sind. Wir zeigen nur die Abgeschlossenheit der Multiplikation in $\mathrm{U}(n)$: Seien $S, T \in \mathrm{U}(n)$. Nach Definition gilt: $\overline{S}^{\,t} S = E$, $\overline{T}^{\,t} T = E$, also $S^{-1} = \overline{S}^{\,t}$ und $T^{-1} = \overline{T}^{\,t} \Rightarrow (\overline{S\,T})^{t} \cdot (ST) = (\overline{T}^{\,t}\,\overline{S}^{\,t}) \cdot (ST) = \overline{T}^{\,t} \cdot E \cdot T = E$. □

Inzwischen haben wir schon einige Matrixgruppen kennengelernt. $\mathrm{GL}(n,K)$, $\mathrm{SL}(n,K) := \{A \in \mathrm{GL}(n,K) \mid \det A = 1\}$, $\mathrm{SO}(n) = \mathrm{SL}(n,\mathbb{R}) \cap \mathrm{GL}(n,\mathbb{R}) \subseteq \mathrm{O}(n) \subseteq \mathrm{GL}(n,\mathbb{R})$, $\mathrm{SU}(n) \subseteq \mathrm{U}(n) \subseteq \mathrm{GL}(n,\mathbb{C})$.

Beispiel 26.7. Die eindimensionale unitäre Gruppe

$$\mathrm{U}(1) = \left\{ \lambda \in \mathbb{C} \;\middle|\; \overline{\lambda}\lambda = 1 \right\}$$

ist einfach zu verstehen: Da $\overline{\lambda}\lambda = |\lambda|^2$, ist $\mathrm{U}(1) = \left\{ \lambda \in \mathbb{C} \;\middle|\; |\lambda| = 1 \right\}$. Sie besteht also aus allen komplexen Zahlen auf dem Einheitskreis.

Bemerkung 26.8. Eine Matrix $S = (v_1, \ldots, v_n) \in \mathrm{GL}(n,\mathbb{C})$ ist genau dann unitär, wenn die Spaltenvektoren $v_1, \ldots, v_n$ normiert sind und zueinander senkrecht stehen.

Beweis. Es gilt: $\overline{S}^{\,t} \cdot S = (\overline{v_k}^{\,t} v_l)_{k=1,\ldots,n,\ l=1,\ldots,n}$. □

Damit können wir das komplexe Analogon zur Hauptachsentransformation formulieren:

Satz 26.9. *Sei $A \in \mathbb{C}^{n\times n}$ eine hermitesche Matrix. Dann existiert eine unitäre Matrix $S \in \mathrm{U}(n)$, so dass*

$$\overline{S}^{\,t} A\, S = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \qquad \text{mit } \lambda_i \in \mathbb{R}.$$

Beweis. Der Beweis ist analog zum Beweis der Hauptachsentransformation für symmetrische Matrizen $A \in \mathbb{R}^{n\times n}$ (Satz 25.4).

Sei $\lambda \in \mathbb{C}$ ein Eigenwert von A. Nach Satz 26.5 ist $\lambda \in \mathbb{R}$. Sei v ein Eigenvektor zu λ; ohne Einschränkung können wir annehmen, dass $\|v\| = 1$. Nun betrachten wir:

$$v^\perp = W = \{w \in \mathbb{C}^n \mid \langle v, w\rangle = 0\} \cong \mathbb{C}^{n-1},$$

da $\langle v, w\rangle = 0$ eine Ursprungs–Hyperebene im $\mathbb{C}^n$ definiert. Dann gilt für alle $w \in W$:

$$\langle Aw, v\rangle = \langle w, Av\rangle = \langle w, \lambda v\rangle = \lambda\langle w, v\rangle = 0,$$

d.h. wie im Reellen ist die Einschränkung von A auf W, d.h. $A|_W$, tatsächlich eine Abbildung in W, also $A|_W \colon W \to W$. Mit Induktion existiert eine Basis $v_2, \dots, v_n$ von W aus normierten zueinander senkrechten Eigenvektoren von A. Wir setzen $S := (v_1, \dots, v_n)$; damit gilt:

$$\overline{S}^{\,t} A\, S = \overline{S}^{\,t}(Av_1, Av_2, \dots, Av_n) = \overline{S}^{\,t}(\lambda_1 v_1, \dots, \lambda_n v_n) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

da $\overline{v_k}^{\,t} v_l = \delta_{kl}$. □

26.2 Abstrakte Skalarprodukte

Bisher haben wir uns auf das Standard–Skalarprodukt im Reellen (siehe insbesondere Abschnitt 17.2) und das hermitesche Skalarprodukt im Komplexen, das wir im vorigen Abschnitt 26.1 betrachtet haben, beschränkt. Viele der Eigenschaften dieser beiden Skalarprodukte können wir auch in einen allgemeineren Kontext bringen und so auch Begriffe wir Orthogonalität in anderen Räumen definieren. Dies hat sehr interessante Anwendungen in vielen Bereichen der Mathematik. Beispielsweise können wir definieren, wann zwei stetige Funktionen auf einem Intervall (dies findet insbesondere in der Nummerik Anwendung) oder zwei Dichten (dazu kommen wir in der Wahrscheinlichkeitsrechnung im nächsten Semester) senkrecht zueinander stehen.

Im Folgenden bezeichnet $\mathbb{K}$ entweder den Körper $\mathbb{R}$ oder $\mathbb{C}$.

Definition 26.10. *V sei ein $\mathbb{K}$–Vektorraum, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. Ein **Skalarprodukt** auf V ist eine Abbildung*

$$\langle .,. \rangle \colon V \times V \to \mathbb{K}, \quad (v, w) \mapsto \langle v, w\rangle$$

mit folgenden Eigenschaften:

1) ***Additivität***:

$$\langle v_1 + v_2, w\rangle = \langle v_1, w\rangle + \langle v_2, w\rangle$$
$$\langle v, w_1 + w_2\rangle = \langle v, w_1\rangle + \langle v, w_2\rangle$$

$\forall v, w, v_1, v_2, w_1, w_2 \in \mathbb{K}^n$.

2) ***Sesquilinearität*** *(d.h.* $(1 + \frac{1}{2})$*–fache Linearität):*

$$\langle \lambda v, w\rangle = \overline{\lambda}\langle v, w\rangle$$
$$\langle v, \lambda w\rangle = \lambda\langle v, w\rangle$$

$\forall \lambda \in \mathbb{K} \quad \forall v, w \in V$.

3) ***Hermitesch***:

$$\langle v, w\rangle = \overline{\langle w, v\rangle} \quad \forall v, w \in V.$$

4) ***Positiv Definitheit***:

$$\langle v, v\rangle \geq 0 \quad \text{und} \quad \langle v, v\rangle = 0 \Leftrightarrow v = 0$$

$\forall v \in V$.

Für ein Skalarprodukt $\langle ., . \rangle$ *heißt die Abbildung*

$$V \to \mathbb{R}_{\geq 0},\ v \mapsto \|v\| := \sqrt{\langle v, v\rangle}$$

die ***zugehörige Norm***.

Beispiel 26.11. 1. $\mathbb{R}^n$ bzw. $\mathbb{C}^n$, versehen mit dem Standardskalarprodukt bzw. dem hermiteschen Skalarprodukt.

2. Der Raum
$$V = C^\circ[a, b] = \{f : [a, b] \to \mathbb{R} \mid \text{ f ist stetig}\}$$
der stetigen Funktionen (s. Kapitel 8), versehen mit dem Skalarprodukt

$$\langle f, g\rangle = \int_a^b f(t)g(t)\, dt.$$

Additivität, Sesquilinearität und Hermitsch sind offenbar erfüllt wegen der Linearität des Integrals und weil komplexe Konjugation auf reellen Zahlen keine Auswirkungen hat. Um zu erkennen, dass die Formel tatsächlich ein Skalarprodukt definiert, müssen wir also noch die positive Definitheit einsehen. Offenbar gilt:

$$\langle f, f\rangle = \int_a^b f(t)^2\, dt \geq 0.$$

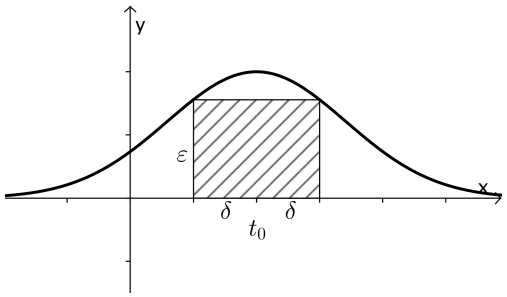


Abbildung 26.2. Zum Skalarprodukt auf dem Raum der stetigen Funktionen.

Ist $f \not\equiv 0$, d.h. f ist nicht die konstante Nullfunktion, dann $\exists\, t_0 \in (a,b)$: $f(t_0) \neq 0$, also wegen der Stetigkeit (s. Abb. 26.2) $\exists\, \varepsilon > 0$, $\delta > 0$ mit $|f(t)| > \varepsilon\ \forall t$ mit $|t - t_0| < \delta$.

Es folgt:

$$\int_a^b |f(t)|^2\, dt \geq \int_{t_0-\delta}^{t_0+\delta} \varepsilon^2\, dt = 2\varepsilon^2\delta > 0.$$

Dieses Skalarprodukt kann man als stetige Variante des Standard - Skalarproduktes auf dem $\mathbb{R}^n$ interpretieren, indem das Summenzeichen durch das Integralzeichen ersetzt wird. Das Integral ist ja nichts anderes als der Grenzwert von Summen von Rechtecksflächen, wobei die Breite der Rechtecke gegen 0 geht (s. Definition 13.1).

3. Die komplexe Variante davon ist:

$$V = \left\{ f\colon [a,b] \to \mathbb{C} \;\middle|\; f \text{ stetig} \right\}, \quad \langle f, g\rangle = \int_a^b \overline{f(t)} g(t)\, dt.$$

4. Sei φ eine **Dichte**, d.h.: $\varphi\colon [a,b] \to \mathbb{R}$ ist strikt positiv und stückweise stetig. Auf dem Vektorraum aller Dichten können wir das folgende Skalarprodukt nutzen:

$$\langle f, g\rangle_\varphi = \int_a^b f(t)g(t)\varphi(t)\, dt.$$

5. Sei $V = \mathbb{K}^n$ endlich–dimensional und

$$\langle .,. \rangle\colon V \times V \to \mathbb{K}$$

ein Skalarprodukt. Die Einheitsvektoren von V bezeichnen wir wie üblich mit: $e_k \in \mathbb{K}^n$. Wir setzen:

$$a_{kj} := \langle e_k, e_j \rangle = \overline{\langle e_j, e_k \rangle} = \overline{a_{jk}}.$$

Sei $A = (a_{kj})$. Dann ist A offenbar hermitesch. Die so definierte Matrix A bestimmt das Skalarprodukt schon eindeutig: Zwei beliebige Vektoren $z, w \in \mathbb{K}^n$ schreiben sich nämlich

$$z = \sum_{k=1}^{n} z_k e_k, \; w = \sum_{j=1}^{n} w_j e_j$$

für gewisse $z_k, w_j \in \mathbb{K}$ und es gilt:

$$\langle z, w \rangle = \sum_{k=1}^{n} \sum_{j=1}^{n} \langle z_k e_k, w_j e_j \rangle = \sum_{k=1}^{n} \overline{z_k} \sum_{j=1}^{n} w_j a_{kj} = \overline{z}^{\,\mathrm{t}} A \, w.$$

Ist umgekehrt $A \in \mathbb{K}^{n \times n}$ eine beliebige hermitesche Matrix, so definiert sie vermöge

$$\langle z, w \rangle_A := \overline{z}^{\,\mathrm{t}} A \, w$$

ein **Skalarprodukt zur hermiteschen Matrix** A auf $\mathbb{K}^n$ genau dann, wenn alle Eigenwerte von A strikt positiv sind: Ist nämlich v ein Eigenvektor von A zum Eigenwert λ, so gilt: $\langle v, v \rangle_A = \overline{v}^{\,\mathrm{t}} A \, v = \overline{v}^{\,\mathrm{t}} \lambda \, v = \lambda \, \overline{v}^{\,\mathrm{t}} \, v$. Aber: $\lambda \, \overline{v}^{\,\mathrm{t}} \, v > 0 \Leftrightarrow \lambda > 0$, da $\overline{v}^{\,\mathrm{t}} \, v > 0$.

Auf einem Vektorraum, der ein solches Skalarprodukt besitzt, können wir explizit Basen konstruieren, so dass deren Vektoren paarweise senkrecht aufeinander stehen und normiert sind:

Beispiel 26.12. Wir betrachten den Untervektorraum

$$U = \left\langle \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix} \right\rangle \subset \mathbb{R}^3$$

und möchten eine Basis von U finden, deren Elemente orthogonal zueinander stehen, d.h. wir suchen z.B. $v = a(1,2,1)^{\mathrm{t}} + b(1,3,0)^{\mathrm{t}}$ mit

$$0 = (1,2,1)^{\mathrm{t}} \cdot v = v_1 + 2v_2 + v_3 = (a+b) + 2 \cdot (2a+3b) + a = 6a + 7b.$$

Offenbar dürfen wir einen der Koeffizienten frei wählen (nur $b = 0$ ist natürlich nicht erlaubt), z.B. $a = t \in \mathbb{R}$, d.h. $b = -\frac{6}{7}t$, um eine Lösung v zu erhalten. Dies sollte uns nicht wundern, da natürlich mit v auch jeder Vektor $s \cdot v$ mit $s \neq 0$ eine Lösung ist.

Das Konstruieren einer Basis eines Untervektorraumes, bei der alle Basiselemente orthogonal zueinander stehen, geht in Dimension zwei auf diese Art noch in vertretbarem Aufwand. Sogar in höherdimensionalen Fällen noch sehr einfach ist es mit folgendem Verfahren möglich:

Satz 26.13 (Gram–Schmidt–Verfahren). *Sei V ein $\mathbb{K}$–Vektorraum mit Skalarprodukt und $w_1, \ldots, w_n$ eine Familie von linear unabhängigen Vektoren (also insbesondere:* $\dim\langle w_1, \ldots, w_k\rangle = k \ \forall k = 1, \ldots, n$*). Dann existieren Vektoren $v_1, \ldots, v_n$ in V mit $\langle v_i, v_j\rangle = \delta_{ij}$ für $i, j \in \{1, 2, \ldots, n\}$, so dass:*

$$\langle v_1, \ldots, v_k\rangle = \langle w_1, \ldots, w_k\rangle \quad \text{für } k = 1, \ldots, n.$$

Beweis. Wir gehen induktiv vor:

1. Zunächst wählen wir $v_1 := \frac{w_1}{\|w_1\|}$. Dann ist v_1 normiert und $\langle v_1\rangle = \langle w_1\rangle$.
2. Sind $v_1, \ldots, v_{k-1}$ schon definiert, dann setzen wir:

   $$u_k := w_k - \sum_{j=1}^{k-1} \langle v_j, w_k\rangle v_j.$$

   Dafür gilt:

   $$\langle v_l, u_k\rangle = \langle v_l, w_k\rangle - \sum_{j=1}^{k-1} \langle v_j, w_k\rangle \langle v_l, v_j\rangle = 0, \quad l = 1, 2, \ldots, k-1,$$

   also: $u_k \perp \langle v_1, \ldots, v_{k-1}\rangle$ und $u_k \neq 0$, da $w_k \notin \langle v_1, \ldots, v_{k-1}\rangle = \langle w_1, \ldots, w_{k-1}\rangle$.
3. Dann setzen wir

   $$v_k := \frac{u_k}{\|u_k\|},$$

   so dass $v_1, \ldots, v_k$ ein sogenanntes **Orthonormalsystem** (siehe auch Def. 26.24) ist. Außerdem gilt: $\langle w_1, \ldots, w_k\rangle = \langle v_1, \ldots, v_k\rangle$.

□

Dieser Satz und sein Beweis liefern also die Bestätigung, dass wir die im Beweis zur Hauptachsentransformation (Satz 25.4) nötige Basis aus orthogonal zueinander stehenden normierten Vektoren tatsächlich explizit und ohne großen Aufwand konstruieren können. Ein Beispiel in einer höheren Dimension, so dass man nicht einfach durch kurzes Überlegen direkt eine Basis hinschreiben kann:

Beispiel 26.14. Wir betrachten $V = \mathbb{R}^4$ mit dem Standard–Skalarprodukt und der Basis:

$$w_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, w_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, w_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, w_4 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Um eine sogenannte **Orthonormalbasis** (siehe auch Def. 26.24) daraus zu machen, müssen wir nach dem Gram–Schmidt–Verfahren setzen:

$$v_1 = \frac{1}{2} w_1 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \text{ da } 2 = \sqrt{4} = \|w_1\|.$$

Damit können wir nun v_2 berechnen:

$$u_2 = w_2 - \langle v_1, w_2 \rangle \cdot v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} - 1 \cdot \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \text{ also } v_2 = \frac{u_2}{\|u_2\|} = u_2.$$

Der dritte Vektor v_3 ergibt sich daraus folgendermaßen:

$$u_3 = w_3 - \langle v_1, w_3 \rangle v_1 - \langle v_2, w_3 \rangle v_2, \text{ also } v_3 = \frac{u_3}{\|u_3\|} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}.$$

Schließlich finden wir für u_4

$$\begin{aligned} u_4 &= w_4 - \langle v_1, w_4 \rangle v_1 - \langle v_2, w_4 \rangle v_2 - \langle v_3, w_4 \rangle v_3 \\ &= \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} - \frac{1}{2} \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} - \frac{1}{2} \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{4} \\ -\frac{1}{4} \\ -\frac{1}{4} \\ \frac{1}{4} \end{pmatrix}, \end{aligned}$$

also:

$$v_4 = \frac{u_4}{\|u_4\|} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Die Matrix der Spaltenvektoren ist demnach:

$$(v_1\ v_2\ v_3\ v_4) = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \in \mathrm{SO}(4) \subseteq \mathrm{GL}(4, \mathbb{R}).$$

Leider muss man feststellen, dass — im Gegensatz zum obigen Beispiel — bei diesem Verfahren wegen der Normierung meist Quadratwurzeln auftreten, die Berechnungen mit Papier und Bleistift etwas anstrengend machen. Außer einigen kleinen Rechnungen zum vertiefenden Verständnis wird man daher zur praktischen Durchführung meist einen Computer verwenden.

26.3 Das Hurwitz–Kriterium

In Beispiel 26.11.5 haben wir gesehen, dass genau solche hermiteschen Matrizen ein Skalarprodukt definieren, die nur strikt positive Eigenwerte besitzen. Wir untersuchen nun, welche hermiteschen Matrizen diese besondere Eigenschaft erfüllen.

Es gibt außerdem noch weitere gute Motivationen, dieses Problem zu studieren. Solche Fragen sind nämlich zentral bei der Untersuchung von Funktionen im Mehrdimensionalen, die wir im nächsten Semester betrachten werden: In Analogie zur Kurvendiskussion in einer reellen Variablen (siehe Kapitel 10), in der ein Punkt mit verschwindender erster und positiver zweiter Ableitung ein Minimum darstellt, ist im Mehrdimensionalen Fall ein Punkt ein Minimum, wenn dort die erste Ableitung verschwindet und die Determinante der sogenannten Hesse–Matrix nur strikt positive Eigenwerte besitzt.

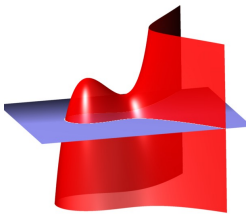


Abbildung 26.3. Eine reelle Funktion im Mehrdimensionalen mit einem Maximum und einem Sattelpunkt. Eine Unterscheidung zwischen Minimum, Maximum, etc. liefert hier die positive bzw. negative Definitheit einer geeigneten Matrix.

Proposition/Definition 26.15. *Eine hermitesche (oder symmetrische) Matrix $A \in \mathbb{K}^{n\times n}$ heißt **positiv definit** (in Zeichen: $A > 0$), wenn folgende äquivalente Bedingungen erfüllt sind:*

1) Alle Eigenwerte von A sind strikt positiv.

2) $\overline{z}^{\,t} A\, z > 0 \quad \forall z \in \mathbb{K}^n \setminus \{0\}$.

3) Durch

$$\langle z, w\rangle_A = \overline{z}^{\,t} A\, w$$

wird ein Skalarprodukt auf $\mathbb{K}^n$ definiert.

Beweis. Wir müssen die Äquivalenz der drei Aussagen zeigen:

2) $\Leftrightarrow$ 3) Zunächst ist 3) $\Rightarrow$ 2) klar, weil 2) eine Teilaussage von 3) ist. Für die andere Richtung ist nur die hermitesche Eigenschaft nicht selbstverständlich. Dies folgt aber aus $\overline{A}^{\,t} = A \iff A^t = \overline{A}$:

$$\langle w, z\rangle_A = \overline{w}^{\,t}(Az) = (z^t A^t)\,\overline{w} \overset{A^t=\overline{A}}{=} z^t\,\overline{A}\,\overline{w} = \overline{(\overline{z}^{\,t} A\, w)} = \overline{\langle z, w\rangle_A}.$$

3) $\Rightarrow$ 1) Dies haben wir schon in Beispiel 26.11.5 gesehen: Sei λ ein Eigenwert, $v \in \mathbb{K}^n$ ein zugehöriger Eigenvektor. Dann gilt:

$$0 < \langle v, v\rangle_A = \overline{v}^{\,t}(Av) = \overline{v}^{\,t}(\lambda v) = \lambda\,\overline{v}^{\,t} v,$$

also: $\lambda > 0$, da $\overline{v}^{\,t} v > 0$.

1) $\Rightarrow$ 2) Nach Satz 26.9 und Satz 24.5 existiert eine Basis von $\mathbb{K}^n$ aus normierten Eigenvektoren von A, etwa $v_1, \ldots, v_n$, die wir nach dem Gram–Schmidt–Verfahren 26.13 auch orthogonal zueinander wählen können. In dieser Basis können wir $z \in \mathbb{K}^n$ schreiben als

$$z = \sum_{j=1}^{n} c_j v_j \neq 0,$$

für gewisse $c_j \in \mathbb{K}$. Damit gilt:

$$\begin{aligned}
\overline{z}^{\,t} A\, z &= \Big(\sum_{j=1}^{n} \overline{c_j v_j}\Big)^t A \Big(\sum_{k=1}^{n} c_k v_k\Big) \\
&= \sum_{j,k=1}^{n} \overline{c}_j c_k\, \overline{v_j}^{\,t}\, \lambda_k v_k \\
&= \sum_{j,k=1}^{n} \overline{c}_j c_k \lambda_k \delta_{jk}, \qquad \text{da } \overline{v_j}^{\,t} \cdot v_k = \delta_{jk} \\
&= \sum_{k=1}^{n} |c_k|^2 \lambda_k > 0
\end{aligned}$$

da alle $\lambda_k > 0$ und wenigstens ein $c_k \neq 0$.

□

Satz 26.16 (Hurwitz–Kriterium). *Eine hemitesche Matrix $A \in (a_{kj}) \in \mathbb{C}^{n\times n}$ ist positiv definit genau dann, wenn sämtliche oberen linken Minoren strikt positiv sind, d.h. wenn:*

$$\det \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix} > 0 \qquad \forall\, k = 1, 2, \dots, n.$$

Beispiel 26.17. Es gilt

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} > 0$$

nach dem Kriterium, da $2 > 0$, $4 - 1 = 3 > 0$, $8 - 2 - 2 = 4 > 0$.

Beweis (des Hurwitz–Kriteriums (Satz 26.16)). Ist zunächst $A > 0$, dann sind nach Definition auch alle Untermatritzen

$$A_k := \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix} > 0.$$

Bedingung 2) in 26.15 ist nämlich auch für $z = (z_1, \dots, z_k, 0, \dots, 0)^t$ erfüllt und

$$(\overline{z_1}, \dots, \overline{z_k}) A_k (z_1, \dots, z_k)^{\mathrm{t}} > 0 \Leftrightarrow (\overline{z_1}, \dots, \overline{z_k}, 0, \dots, 0) A (z_1, \dots, z_k, 0, \dots, 0)^{\mathrm{t}} > 0.$$

Auch diese Untermatrizen haben also nur positive Eigenwerte. Aber die Determinante $\det A_k$ ist das Produkt aller Eigenwerte der Matrix und somit auch strikt positiv. Die Bedingung ist also notwendig.

Wir müssen noch zeigen, dass sie auch hinreichend ist. Dazu verwenden wir Induktion nach n. Der Fall $n = 1$ ist klar. Es bleibt also noch der Induktionsschritt $n - 1 \to n$:

Nach Induktionsvoraussetzung ist:

$$B = \begin{pmatrix} a_{1,1} & \dots & a_{1,n-1} \\ \vdots & \ddots & \vdots \\ a_{n-1,1} & \dots & a_{n-1,n-1} \end{pmatrix} > 0.$$

Da B ebenfalls hermitesch ist, existiert $S \in \mathrm{U}(n-1)$, so dass:

$$\overline{S}^{\mathrm{t}} B S = \begin{pmatrix} \lambda'_1 & & 0 \\ & \ddots & \\ 0 & & \lambda'_{n-1} \end{pmatrix},$$

wobei $0 < \lambda'_1, \ldots, \lambda'_{n-1} \in \mathbb{R}$. Dann gilt:

$$\begin{pmatrix} \overline{S}^{\,t} & 0 \\ 0 & 1 \end{pmatrix} \cdot A \cdot \underbrace{\begin{pmatrix} S & 0 \\ 0 & 1 \end{pmatrix}}_{=:S'} = \left(\begin{array}{ccc|c} \lambda'_1 & & 0 & b_1 \\ & \ddots & & \vdots \\ 0 & & \lambda'_{n-1} & b_{n-1} \\ \hline \overline{b_1} & \ldots & \overline{b_{n-1}} & c \end{array}\right) =: A'$$

für gewisse $b_i \in \mathbb{C}, i = 1, 2, \ldots, n-1, c \in \mathbb{C}$. Wir betrachten nun:

$$T = \begin{pmatrix} 1 & & 0 & \frac{-b_1}{\lambda'_1} \\ & \ddots & & \vdots \\ 0 & & & 1 \end{pmatrix} \in \mathrm{SL}(n, \mathbb{C}).$$

Es gilt, da $\lambda'_i \in \mathbb{R}$:

$$\overline{T}^{\,t} A' T = \begin{pmatrix} \lambda'_1 & & & 0 \\ & \ddots & & \\ & & \lambda'_{n-1} & \\ 0 & & & c' \end{pmatrix} =: D.$$

Da $0 < \det A = \det A' = \det D$, $\det T = \det \overline{T}^{\,t} = 1$ und $\lambda'_1, \ldots, \lambda'_{n-1} > 0$, folgt $c' > 0$. Für $w = (w_1, \ldots, w_n)^t \neq 0$ gilt daher: $\overline{w}^{\,t} D w = \sum_{i=1}^{n-1} w_i^2 \cdot \lambda'_i + w_n^2 c' > 0$. Insgesamt ergibt sich demnach:

$$0 < \overline{w}^{\,t} D w = \overline{w}^{\,t}\, \overline{T}^{\,t}\, \overline{S'}^{\,t} A S' T w.$$

Nun ist aber $S'T \in \mathrm{GL}(n, \mathbb{C})$, d.h. $\forall z \neq 0\ \exists\, w \neq 0$ mit $z = STw$, also:

$$\overline{z}^{\,t} \cdot A \cdot z > 0 \quad \forall z \neq 0 \in \mathbb{K}^n,$$

d.h. A ist positiv definit. □

26.4 Normen

Wir haben bereits in einigen Spezialfällen Normen und damit Abstände definiert (s. beispielsweise Definition 26.10) und einige ihrer Eigenschaften nachgewiesen; nun folgt die allgemeine Definition:

Definition 26.18. *Sei V ein $\mathbb{K}$–Vektorraum. Eine **Norm** auf V ist eine Abbildung*

$$\|.\| \colon V \to \mathbb{R}$$

mit folgenden Eigenschaften:

1) $\|v\| \geq 0$ *und* $\|v\| = 0 \Leftrightarrow v = 0$,

2) $\|\lambda v\| = |\lambda| \cdot \|v\| \quad \forall \lambda \in \mathbb{K},\ \forall v \in V$,

3) ($\triangle$*–Ungleichung)* $\|v + w\| \leq \|v\| + \|w\| \quad \forall v, w \in V$.

Beispiel 26.19. 1. Die **euklidische Norm** eines Skalarproduktes ist:

$$\|v\| := \sqrt{\langle v, v \rangle}.$$

Um zu sehen, dass dies auch wirklich eine Norm in obigem Sinn ist, müssen wir noch die $\triangle$–Ungleichung zeigen, da die anderen Bedingungen offensichtlich erfüllt sind. Dies werden wir erst auf Seite 380 nach einigen Vorbereitungen erledigen.

2. Wir betrachten:

$$V = \mathbb{R}^n \ni x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Dann ist die p**–Norm**, $1 \leq p \leq \infty$, definiert durch:

$$\|x\|_p := \Big(\sum_{k=1}^{n} |x_k|^p\Big)^{\frac{1}{p}}.$$

Die 2–Norm ist die euklidische Norm für das Standard–Skalarprodukt auf $\mathbb{R}^n$.

3. Die **Maximum–Norm** oder ∞**–Norm** auf $\mathbb{R}^n$ ist definiert durch:

$$\|x\|_\infty := \max\{\, |x_k| \mid k = 1, \ldots, n\}.$$

4. Analog zu den p–/∞–Normen auf $\mathbb{R}^n$ definieren wir für den Funktionenraum $V = C^\circ[a, b]$:

$$\|f\|_\infty := \sup_{x \in [a,b]} |f(x)|, \quad \|f\|_p := \left(\int_a^b |f(t)|^p \, dt\right)^{\frac{1}{p}}.$$

Wir haben zwar in diesem Semester gar nicht erklärt, was Konvergenz und Cauchy–Folgen sind, möchten aber trotzdem kurz hierauf eingehen, da doch viele der Hörer im letzten Semester anwesend waren und da es hier sehr gut passt:

Definition 26.20. *Ein* ***normierter Vektorraum*** *ist ein* $\mathbb{K}$*–Vektorraum* V *zusammen mit einer Norm. Die Norm gibt uns den Begriff einer Cauchy–Folge (siehe dazu Definition 5.17 bzw. den ganzen Abschnitt 5.5). Ein normierter Raum, in dem jede Cauchy–Folge konvergiert (siehe Definition 7.8), heißt* ***Banachraum****.*

*Ein **euklidischer Vektorraum** (bzw. **unitärer Vektorraum**) ist ein endlich–dimensionaler $\mathbb{K}$–Vektorraum V für $\mathbb{K} = \mathbb{R}$ (bzw. $\mathbb{K} = \mathbb{C}$) zusammen mit einem (hermiteschen) Skalarprodukt $\langle .,. \rangle \colon V \times V \to \mathbb{K}$.*

*Im Falle eines, möglicherweise unendlich–dimensionalen, Vektorraumes mit Skalarprodukt spricht man von einem **Prä–Hilbertraum**. Ein solcher heißt **Hilbertraum**, falls jede Cauchy–Folge konvergiert.*

Beispiel 26.21. Jeder endlich–dimensionale normierte $\mathbb{K}$–Vektorraum, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$, ist nach dem Vollständigkeitsaxiom (Satz 5.19) ein Banachraum.

Wie viele Aussagen über das Standard–Skalarprodukt, gilt auch die Cauchy–Schwartz–Ungleichung (Satz 17.4) allgemeiner für beliebige Skalarprodukte. Auch für unendliche–dimensionale Prä–Hilberträume ist sie eines der zentralen Hilfsmittel, um Konvergenz nachzuweisen (s. Beispiel 26.23):

Satz 26.22 (Cauchy–Schwarzsche Ungleichung). *Sei $\langle .,. \rangle$ ein Skalarprodukt auf dem $\mathbb{K}$–Vektorraum V. Dann gilt für $x, y \in V$:*

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|$$

und Gleichheit gilt für $x \neq 0$ genau dann, wenn ein $\lambda \in \mathbb{K}$ mit $y = \lambda x$ existiert.

Beweis. Der Beweis ist analog zu jenem für das Standard–Skalarprodukt aus Satz 17.4: Für $x = 0 \in V$ ist nichts zu zeigen. Sei also $x \neq 0$. Wir setzen:

$$\mathbb{K} \ni \mu := \langle x, x\rangle = \|x\|^2 > 0, \varphi = -\langle x, y\rangle \in \mathbb{K}.$$

Wir betrachten nun:

$$\begin{aligned} 0 \leq \langle \varphi x + \mu y, \varphi x + \mu y\rangle &= |\varphi|^2\langle x, x\rangle + \overline{\mu}\varphi\langle y, x\rangle + \overline{\varphi}\mu\langle x, y\rangle + |\mu|^2\langle y, y\rangle \\ &= \underbrace{\mu}_{>0} \cdot \Big(\underbrace{|\langle x, y\rangle|^2 - 2|\langle x, y\rangle|^2 + |\langle x, x\rangle| \cdot |\langle y, y\rangle|}_{\geq 0} \Big). \end{aligned}$$

Wir können also durch μ dividieren und erhalten, nach Ausrechnen des Ausdrucks in Klammern:

$$\|x\|^2 \cdot \|y\|^2 \geq |\langle x, y\rangle|^2.$$

Die Monotonie der Wurzel (Bemerkung 5.30) liefert nun die Ungleichung.

Gilt Gleichheit, dann gilt Gleichheit von Anfang an, d.h. $\varphi x + \mu y = 0 \Rightarrow y = -\frac{\varphi}{\mu}x = -\lambda x$. □

Beweis (der $\triangle$–Ungleichung aus Beispiel 26.19). Auch dieser Beweis ist wörtlich quasi identisch zu jenem für das Standard–Skalarprodukt (Proposition 17.5): Die Cauchy–Schwarzsche Ungleichung liefert:

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle \\ &= \|x\|^2 + \langle x,y\rangle + \langle x,y\rangle + \|y\|^2 \\ &= \|x\|^2 + 2|\langle x,y\rangle| + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\|\cdot\|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2.\end{aligned}$$

Nun folgt, mit der Monotonie der Wurzel (Bemerkung 5.30):

$$\|x+y\| \leq \|x\| + \|y\|.$$

□

Zwar sind die Beweise für Cauchy–Schwartz und die Dreiecks–Ungleichung so gut wie identisch zu jenen im $\mathbb{R}^n$ bzgl. des Standard–Skalarprodukts. Mit Hilfe der abstrakteren Versionen können wir aber auch interessantere Beispiele studieren. Das Folgende ist eines der einfachsten Beispiele eines Hilbertraumes:

Beispiel 26.23. Der Raum $l^2(\mathbb{R})$ ist die Menge

$$l^2(\mathbb{R}) = \Big\{(x_n)_{n\in\mathbb{N}} \text{ relle Folge} \;\Big|\; \sum_{n=0}^{\infty} |x_n|^2 < \infty\Big\}.$$

Wir definieren darauf das Skalarprodukt:

$$\langle (x_n), (y_n)\rangle = \sum_{n=0}^{\infty} x_n y_n.$$

Zu zeigen ist, dass die Reihe konvergiert (absolut!). Die Cauchy–Schwarzsche Ungleichung liefert:

$$\Big|\sum_{n=0}^{N} x_n y_n\Big| \leq \sum_{n=0}^{N} |x_n|\,|y_n| \leq \Big(\sum_{n=0}^{N} |x_n|^2\Big)^{\frac{1}{2}}\Big(\sum_{n=0}^{N} |y_n|^2\Big)^{\frac{1}{2}} \leq \|x\|\cdot\|y\| < \infty.$$

Das Skalarprodukt $\langle .,. \rangle \colon l^2(\mathbb{R}) \times l^2(\mathbb{R}) \to \mathbb{R}$ ist also tatsächlich wohldefiniert.

Wie schon angedeutet, ist $l^2(\mathbb{R})$ sogar ein Hilbertraum. Um dies zu zeigen, betrachten wir eine Cauchy–Folge $(v_k)_{k\in\mathbb{N}}$, $v_k = (x_n^{(k)})_{n\in\mathbb{N}} \in l^2(\mathbb{R})$, von Vektoren in $l^2(\mathbb{R})$, d.h. $\forall \varepsilon > 0\ \exists\, N$:

$$\|v_k - v_l\| < \varepsilon \quad \forall\, k,l \geq N.$$

Nun folgt: $(x_n^{(k)})_{k\in\mathbb{N}}$ bildet für jedes feste n eine Cauchyfolge, denn

$$|x_n^{(k)} - x_n^{(l)}|^2 \leq \|v_k - v_l\|^2 < \varepsilon^2 \quad \forall\, k,l \geq N$$

Da in den reellen Zahlen jede Cauchy–Folge konvergiert (das ist gerade die Aussage des Vollständigkeitsaxioms, Satz 5.19), hat jede dieser Folgen einen Grenzwert in den reellen Zahlen:

$$\exists\, \tilde{y}_n \in \mathbb{R} : \lim_{k\to\infty} x_n^{(k)} = \tilde{y}_n.$$

Weiter existiert daher $\tilde{y} = (\tilde{y}_n)_{n\in\mathbb{N}}$ mit $\lim_{k\to\infty} v_k = \tilde{y}$, so dass wirklich jede Cauchy–Folge in $l^2(\mathbb{R})$ konvergiert.

26.5 Orthogonale Projektion

Wir haben bereits das Gram–Schmidt–Verfahren zur expliziten Konstruktion von Basen, deren Elemente normiert und paarweise orthogonal zueinander sind, kennen gelernt. Dies werden wir nun verwenden, um die geometrische Abbildung der orthogonalen Projektion nun auch für allgemeine Vektorräume mit Skalarprodukt durchführen zu können. Im Abschnitt 17.3.3 hatten wir ja schon den Fall des $\mathbb{R}^n$ gemeinsam mit dem Standard–Skalarprodukt untersucht (siehe auch Abb. 26.4).

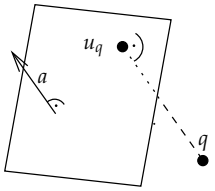


Abbildung 26.4. Die Orthogonale Projektion des Punktes q auf die Hyperebene H.

In diesem Abschnitt bezeichnet $\mathbb{K}$ entweder $\mathbb{R}$ oder $\mathbb{C}$. V ist immer ein $\mathbb{K}$–Vektorraum und

$$\langle .,. \rangle \colon V \times V \to \mathbb{K}$$

ein (euklidisches oder hermitisches) Skalarprodukt.

Definition 26.24. *Sei V ein Vektorraum mit Skalarprodukt $\langle .,. \rangle$ und $\{v_j\}_{j\in J}$ eine Familie von Vektoren.*

1. *$\{v_j\}_{j\in J}$ bildet ein **Orthogonalsystem**, wenn $\langle v_j, v_k \rangle = 0 \;\forall\, j \neq k \in J$ und wenn $v_j \neq 0 \;\forall\, j \in J$.*
2. *$\{v_j\}_{j\in J}$ ist ein **Orthonormalsystem** wenn außerdem: $\langle v_j, v_j \rangle = 1 \forall\, j$ ($\Leftrightarrow \|v_j\| = 1$), wenn also gilt: $\langle v_j, v_k \rangle = \delta_{jk} \;\forall\, j,k \in J$.*

3. *Ein Orthogonal– bzw. –normalsystem heißt* ***Orthogonalbasis*** *bzw.* ***Orthonormalbasis****, wenn die Vektoren des Systems eine Basis bilden.*

Beispiel 26.25. 1. Die Einheitsvektoren $e_j \in \mathbb{K}^n$ bilden eine Orthonormalbasis des $\mathbb{K}^n$ bezüglich des Standard–Skalarprodukts.

2. Die Spalten v_i einer orthogonalen beziehungsweise unitären $n \times n$-Matrix A bilden ein Orthogonalsystem, denn:

$$(\langle v_j, v_k \rangle)_{j=1,\dots,n,\ k=1,\dots,n} = \overline{A}^{\,\mathrm{t}} A = E.$$

3. Wir betrachten den Vektorraum:

$$V = \{f \colon \mathbb{R} \to \mathbb{C} \mid f \text{ ist } 2\pi \text{ –periodisch und stetig } \}.$$

Eine Funktion f heißt 2π**–periodisch**, falls $f(x + 2\pi) = f(x) \quad \forall x \in \mathbb{R}$. Der Grund, hier komplexwertige Funktionen zu betrachten ist, dass sich mit $e^{int} = \cos(nt) + i \sin(nt)$, $n \in \mathbb{N}_0$, leichter rechnen lässt als mit $\sin(nt)$ und $\cos(nt)$, $n \in \mathbb{N}_0$. Beispielsweise gilt $e^{x+iy} = e^x \cdot (\cos y + i \sin y)$, $e^{iy} = \cos y + i \sin y$, $e^{i\pi} = -1$, $e^{2\pi i} = 1$, $e^{z+w} = e^z \cdot e^w$ für $z, w \in \mathbb{C}$ (siehe Abschnitt 7.4 für weitere Informationen zur komplexen Exponentialfunktion).

Wir betrachten also die Funktionen $\{e_n\}_{n\in\mathbb{Z}} \subseteq V$ mit $e_n(t) := e^{int}$. Die $e_n(t)$ sind 2π–periodisch, da $e^{2\pi i} = 1$ und $n \in \mathbb{Z}$, und sie bilden ein Orthonormalsystem bezüglich des Skalarprodukts

$$\langle f, g \rangle = \frac{1}{2\pi} \int_0^{2\pi} \overline{f(t)} g(t)\, dt$$

auf V. Es gilt nämlich:

$$\langle e_m, e_n \rangle = \frac{1}{2\pi} \int_0^{2\pi} \overline{e^{imt}} \cdot e^{int}\, dt = \frac{1}{2\pi} \int_0^{2\pi} e^{i(n-m)t}\, dt$$

$$= \frac{1}{2\pi} \begin{cases} \int_0^{2\pi} 1\, dt, & n = m \\ \left[\frac{e^{i(n-m)t}}{i(n-m)}\right]_0^{2\pi}, & n \neq m \end{cases} = \begin{cases} 1, & n = m \\ 0, & n \neq m \end{cases} = \delta_{mn}.$$

Mit Hilfe des Gram–Schmidt–Verfahrens (Satz 26.13) können wir Orthonormalsysteme bzw. –Basen explizit berechnen. Eine interessante Eigenschaft solcher Basen ist folgende:

Satz 26.26 (Darstellung bzgl. einer Orthonormalbasis). *Sei $\{v_j\}_{j=1,\dots,n}$ eine Orthonormalbasis des Vektorraums V und $w \in V$ ein weiterer Vektor. Dann gilt:*

$$w = \langle v_1, w \rangle v_1 + \cdots + \langle v_n, w \rangle v_n.$$

Beweis. Wir bilden die Differenz u zwischen beiden Seiten,

$$u = w - \sum_{j=1}^{n} \langle v_j, w \rangle v_j,$$

und wenden auf diese Summe das Skalarprodukt $\langle v_k, - \rangle$ an:

$$\begin{aligned}
\langle v_k, u \rangle &= \langle v_k, w \rangle - \sum_{j=1}^{n} \langle v_k, \langle v_j, w \rangle v_j \rangle \\
&= \langle v_k, w \rangle - \sum_{j=1}^{n} \langle v_j, w \rangle \langle v_k, v_j \rangle \\
&= \langle v_k, w \rangle - \langle v_k, w \rangle \\
&= 0.
\end{aligned}$$

Es folgt: u steht senkrecht auf $v_1, \ldots, v_n$ und somit auch senkrecht auf jeder Linearkombination von $v_1, \ldots, v_n$. Da $v_1, \ldots, v_n$ den Vektorraum V erzeugen, gilt insbesondere:

$$\langle u, u \rangle = 0 \Rightarrow \|u\| = 0 \Rightarrow u = 0.$$

□

Wir können, wenn wir eine Orthonormalbasis eines Vektorraumes zur Verfügung haben, für jeden beliebigen Vektor die Koeffizienten einer Darstellung als Linearkombination in der Basis direkt hinschreiben. Außerdem werden wir jetzt gleich sehen, dass wir mit Hilfe dieser Formel auch orthogonale Projektionen auf Untervektorräume explizit angeben können.

Definition 26.27. *Seien V ein Vektorraum mit Skalarprodukt $\langle ., . \rangle$ und $U \subseteq V$ ein Untervektorraum. Dann heißt*

$$U^\perp := \{v \in V \mid \langle v, u \rangle = 0 \; \forall \; u \in U\}$$

der zu U ***orthogonale Untervektorraum*** *oder* ***orthogonales Komplement*** *von U.*

Bemerkung 26.28. Es gilt:

$$U^\perp \cap U = 0,$$

da nur der Nullvektor zu sich selbst senkrecht ist.

Wenn $\dim V < \infty$, gilt außerdem:

$$U^\perp \oplus U = V,$$

da $\dim U^\perp = \dim V - \dim U$. Ferner ist dann

$$(U^\perp)^\perp = U.$$

$U \subseteq (U^\perp)^\perp$ ist auch ohne die Voraussetzung $\dim V < \infty$ klar.

Definition 26.29. *Sei $U \subseteq V$ ein Untervektorraum. Eine Abbildung $\varphi\colon V \to U$ heißt **Projektion von V auf U**, falls für jedes $u \in U$ gilt: $\varphi(u) = u$. Eine Projektion heißt **orthogonale Projektion auf den Untervektorraum U**, falls für jeden Vektor $v \in V$ gilt:*

$$(\varphi(v) - v) \perp U.$$

Dieser Begriff verallgemeinert den bereits für Hyperebenen in Proposition/Definition 17.14 eingeführten.

Satz 26.30. *Sei V ein $\mathbb{K}$–Vektorraum mit Skalarprodukt und $U \subseteq V$ ein endlich–dimensionaler Teilraum. Es sei $\{u_1, \dots, u_k\}$ eine Orthonormalbasis. Dann sind die Abbildungen*

$$\mathrm{pr}_U\colon V \to U,\ v \mapsto \sum_{j=1}^{k} \langle u_j, v\rangle u_j$$

und

$$\mathrm{pr}_{U^\perp}\colon V \to U^\perp,\ v \mapsto v - \sum_{j=1}^{k} \langle u_j, v\rangle u_j$$

orthogonale Projektionen auf die Teilräume.

Beweis. $\mathrm{pr}_U(v) \in U$ ist klar, $v = \mathrm{pr}_U(v) + \mathrm{pr}_{U^\perp}(v)$ ist auch klar. $\mathrm{pr}_U(u) = u\ \forall\ u \in U$ und $\mathrm{pr}_{U^\perp}(w) = w\ \forall\ w \in U^\perp$ ebenfalls, z.B. mit dem Satz über die Darstellung bzgl. einer Orthonormalbasis.

Es bleibt zu zeigen: $w := \mathrm{pr}_{U^\perp}(v) \in U^\perp$. Es gilt aber:

$$\langle u_l, w\rangle = \langle u_l, v\rangle - \sum_{j=1}^{k} \langle u_j, v\rangle \langle u_l, u_j\rangle = 0, \quad l = 1, \dots, k.$$

Also: $w = \mathrm{pr}_{U^\perp}(v) \in U^\perp = \langle u_1, \dots, u_k\rangle^\perp$. Mit $u := \mathrm{pr}_U(v)$ ist $v = u + w \in U \oplus U^\perp$, und $u \perp w$. pr_U und $\mathrm{pr}_{U^\perp}$ sind also orthogonale Projektionen auf U bzw. $U^\perp$. Auch in diesem Fall gilt: $U \oplus U^\perp = V$. □

Beispiel 26.31. 1. Sei $V = \mathbb{R}^n$ und $U = \langle e_1, \dots, e_k\rangle$. Dann ist $U^\perp = \langle e_{k+1}, \dots, e_n\rangle$ und offenbar gilt mit $x = (x_1, \dots, x_n)^t$:

$$\mathrm{pr}_U(x) = (x_1, \dots, x_k, 0, \dots, 0)^t = \sum_{j=1}^{k} \langle e_j, x\rangle e_j,$$

$$\mathrm{pr}_{U^\perp}(x) = (0, \dots, 0, x_{k+1}, \dots, x_n)^t = x - \mathrm{pr}_U(x).$$

2. Sei $L \subset \mathbb{R}^n$ eine Gerade durch den Ursprung mit Richtungsvektor v. Ohne Einschränkung können wir $\|v\| = 1$ wählen. Dann sind die orthogonalen Projektionen auf L bzw. $L^\perp$:

$$\begin{aligned} \mathrm{pr}_L &: \mathbb{R}^n \to L, \;\; x \mapsto \langle v, x\rangle v, \\ \mathrm{pr}_{L^\perp} &: \mathbb{R}^n \to L^\perp, \;\; x \mapsto x - \langle v, x\rangle v. \end{aligned}$$

Betrachten wir ein konkretes Beispiel:

$$L = \langle (1, \ldots, 1)^{\mathrm{t}} \rangle \subset \mathbb{R}^n, \text{ d.h. } v = \frac{1}{\sqrt{n}}(1, \ldots, 1)^{\mathrm{t}}.$$

Es ergeben sich: $L^\perp = \{x \mid \sum x_i = 0\}$ und $\langle v, x\rangle = \frac{1}{\sqrt{n}} \sum x_i = \sqrt{n} \cdot \overline{x}$, wobei $\overline{x} = \frac{1}{n} \sum x_i$ das **arithmetische Mittel** der Komponenten von x ist. Die beiden Projektionen sind also:

$$\begin{aligned} \mathrm{pr}_L &: \mathbb{R}^n \to L, \;\; x \mapsto (\overline{x}, \ldots, \overline{x}), \\ \mathrm{pr}_{L^\perp} &: \mathbb{R}^n \to L^\perp, \;\; x \mapsto x - (\overline{x}, \ldots, \overline{x}). \end{aligned}$$

Tatsächlich ist für $y = \mathrm{pr}_{L^\perp}(x)$:

$$\sum y_i = \sum x_i - n \cdot \overline{x} = \sum x_i - \sum x_i = 0,$$

d.h. $y \in L^\perp$.

Wie bei der Projektion auf Hyperebenen in Abschnitt 17.3.3, ist die Projektion eines Vektors v auf einen Untervektorraum derjenige Punkt darauf, der von v den kleinsten Abstand hat, der also die beste Annäherung von v durch einen Vektor des Unterraumes darstellt:

Korollar 26.32 (Approximationssatz). *V sei ein $\mathbb{R}$–Vektorraum, versehen mit einem Skalarprodukt und der zugehörigen Norm $\|.\|$. Sei ferner U ein Untervektorraum. Zu jedem $v \in V$ ist $\mathrm{pr}_U(v)$ die beste Approximation von v in U, d.h.:*

$$\|v - \mathrm{pr}_U(v)\| < \|v - u\| \;\forall u \in U \text{ mit } u \neq \mathrm{pr}_U(v).$$

Beweis. Da für $x, y \in V$ der verallgemeinerte Satz des Pythagoras (siehe Proposition 17.3) gilt, d.h. $\|x + y\|^2 = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2$, ergibt sich:

$$\begin{aligned} \|v - u\|^2 &= \| \underbrace{v - \mathrm{pr}_U(v)}_{\in U^\perp} + \underbrace{\mathrm{pr}_U(v) - u}_{\in U} \|^2 \\ &= \|v - \mathrm{pr}_U(v)\|^2 + \|\mathrm{pr}_U(v) - u\|^2 \\ &\geq \|v - \mathrm{pr}_U(v)\|^2. \end{aligned}$$

Gleichheit gilt offenbar genau dann, wenn $u = \mathrm{pr}_U(v)$. □

Beispiel 26.33. Wir versehen $V = C^0[0, \frac{\pi}{2}]$, den Raum der stetigen Funktionen auf $[0, \frac{\pi}{2}]$, mit dem Skalarprodukt

$$\langle f, g\rangle = \int_0^{\frac{\pi}{2}} f(t) \cdot g(t)\, dt.$$

Wir möchten die Gerade bestimmen, die $f(t) = \sin t$ auf dem Intervall $[0, \frac{\pi}{2}]$ bzgl. der zugehörigen Norm am Besten approximiert.

Wir betrachten also $U := \langle 1, t\rangle$ der Untervektorraum aller Geraden. Wir suchen: $\mathrm{pr}_U(f(t)) = \lambda_1 \cdot 1 + \lambda_2 \cdot t$ mit

$$\langle f(t) - \lambda_1 \cdot 1 - \lambda_2 \cdot t, 1\rangle = 0, \quad \langle f(t) - \lambda_1 \cdot 1 - \lambda_2 \cdot t, t\rangle = 0.$$

Für λ_1 und λ_2 ergibt sich das Gleichungssystem

$$\langle 1, 1\rangle\lambda_1 + \langle t, 1\rangle\lambda_2 = \langle \sin t, 1\rangle, \quad \langle 1, t\rangle\lambda_1 + \langle t, t\rangle\lambda_2 = \langle \sin t, t\rangle.$$

Die Skalarprodukte sind mit Schulmitteln oder mit der Integrationstheorie aus dem ersten Semester einfach zu berechnen:

$$\langle 1, 1\rangle = \int_0^{\frac{\pi}{2}} 1\, dt = \frac{\pi}{2}, \qquad \langle t, 1\rangle = \langle 1, t\rangle = \int_0^{\frac{\pi}{2}} t\, dt = \frac{\pi^2}{8},$$
$$\langle t, t\rangle = \int_0^{\frac{\pi}{2}} t^2\, dt = \frac{\pi^3}{24}, \qquad \langle \sin t, 1\rangle = \int_0^{\frac{\pi}{2}} \sin t\, dt = 1,$$
$$\langle \sin t, t\rangle = \int_0^{\frac{\pi}{2}} t \cdot \sin t\, dt = \cdots = 1.$$

Das System ist also

$$\begin{pmatrix} \frac{\pi}{2} & \frac{\pi^2}{8} \\ \frac{\pi^2}{8} & \frac{\pi^3}{24} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Es hat die Lösung (siehe auch Abb. 26.5):

$$\lambda_1 = 8 \cdot \frac{\pi - 3}{\pi^2} \approx 0.11, \quad \lambda_2 = 24 \cdot \frac{4 - \pi}{\pi^3} \approx 0.66.$$

Eine ähnliche Vorgehensweise wird bei Wavelets verwendet, auf denen das Bild–Komprimierungsverfahren Jpeg2000 basiert: statt n Koordinaten speichert man nur die $m < n$ Koordinaten der besten Approximation auf einen geeigneten Untervektorraum; je kleiner m gewählt wird, desto stärker ist die Komprimierung, aber desto größer auch der mögliche Verlust an Informationen. Ein erster Schritt zu deren Verständnis liefern Fourierreihen, die wir im nächsten Abschnitt kurz betrachten.

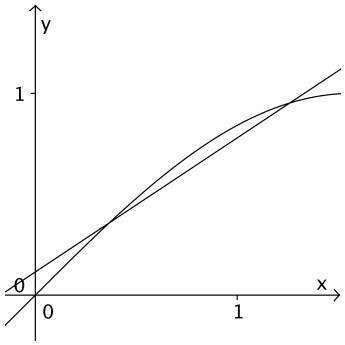


Abbildung 26.5. Approximation von $\sin(t)$ zwischen 0 und $\frac{\pi}{2}$.

Aufgaben

Aufgabe 26.1 (Unitäre Matrizen). Sei $A \in U(n) \subset \mathbb{C}^{n\times n}$ eine unitäre Matrix. Zeigen Sie: $\exists\, S \in U(n)$ mit:

$$\bar{S}^{\mathrm{t}}\, A\, S = D =: \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

wobei $\lambda_i \in \mathbb{C}$ die Eigenwerte von A sind. Zeigen Sie ferner, dass gilt: $|\lambda_i| = 1$.

Hinweis: Zeigen Sie, dass das orthogonale Komplement W eines Eigenvektors v von A von der Matrix A in sich abgebildet wird, d.h. $AW \subset W$.

Aufgabe 26.2 (Orthogonales Komplement). Sei V ein endlich-dimensionaler $\mathbb{R}$-Vektorraum und sei $U \subset V$ ein Untervektorraum von V. Zeigen Sie:

$$(U^{\perp})^{\perp} = U.$$

Aufgabe 26.3 (Orthonormalisierungsverfahren).

1. Berechnen Sie mit dem Gram-Schmidt-Verfahren aus $1, x, x^2, x^3$ eine Orthonormalbasis des Vektorraumes $U = \mathbb{R}[x]_{\leq 3}$ bezüglich der Skalarprodukte

$$\langle p, q\rangle = \int_{-1}^{1} p(x)q(x)x^2 dx,$$
$$\langle p, q\rangle = \int_{-1}^{1} p(x)q(x)(1-x^2) dx.$$

2. Bestimmen Sie bezüglich beider Skalarprodukte aus (a) die orthogonale Projektion $\pi(f)$ von $f = x^2(x^2 - 1) \in \mathbb{R}[x]_{\leq 4}$ auf U und fertigen Sie Zeichnungen (auch mit Maple okay) von $f - \pi(f)$ an.

Aufgabe 26.4 (Pseudoinverse in einem Punkt). Wir betrachten die Abbildung

$$f\colon \mathbb{R}^2 \to \mathbb{R}^3, \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}.$$

1. Berechnen Sie Ker f und Bild f.
2. Sei $P = (1, 1, 0)^t \in \mathbb{R}^3$. Berechnen Sie $Q := \text{Bild}(P)$, das Bild von P unter der orthogonalen Projektion des $\mathbb{R}^3$ auf Bild f.
3. Berechnen Sie das Urbild $f^{-1}(Q) = \{v \in \mathbb{R}^2 \mid f(v) = Q\}$ von Q unter f.

27

Fourierreihen

Fourierreihen und deren Verallgemeinerungen gehen zentral bei der Komprimierung von Bildern im JPEG2000–Format und anderen Bereichen der Bildverarbeitung ein. Bei deren Vorstellung werden wir, wie schon in einigen Beispielen zuvor, intensiv mit Skalarprodukten auf Funktionenräumen arbeiten. Dies macht deutlich, wie wichtig auch solche, auf den ersten Blick vielleicht sinnlos allgemeinen, Konstruktionen sein können.

Außerdem werden wir sehen, dass Fourierreihen einen Bezug zur sogenannten Riemannschen Zeta–Funktion haben; wir werden nämlich mit unseren Mitteln einige besonders interessante Werte dieser Funktion berechnen können. Fourierreihen haben also nicht nur wichtige Anwendungen in der realen Welt, sondern auch in der rein innermathematischen.

Leider werden in diesem Kapitel auch einige Resultate aus der Analysis eingehen, doch wir werden versuchen, die Darstellung immer so zu halten, dass auch die Hörer, die den ersten Teil der Vorlesung nicht gehört haben, die wesentlichen Ideen nachvollziehen können.

27.1 Zur Definition

Wir betrachten im gesamten Kapitel den Vektorraum (für die Definition von *integrierbar* siehe Abschnitt 13.1; man kann auch *stetig* statt dessen denken)

$$V = \left\{ f \colon \mathbb{R} \to \mathbb{C} \;\middle|\; f \text{ ist über } [0, 2\pi] \text{ integrierbar und } 2\pi\text{–periodisch} \right\}$$

mit dem Skalarprodukt

$$\langle f, g \rangle = \frac{1}{2\pi} \int_0^{2\pi} \overline{f(t)} g(t)\, dt.$$

Wie wir in Beispiel 26.25.3 schon gesehen haben, bilden die Funktionen $\{e_n\}_{n\in\mathbb{Z}} \subset V$ mit $e_n(t) := e^{int}$ ein Orthonormalsystem. Die reellen Funktionen

$$\left\{\frac{1}{\pi}\cos(nt)\right\}_{n\geq 0} \cup \left\{\frac{1}{\pi}\sin(nt)\right\}_{n\geq 1}$$

bilden ebenfalls ein Orthonormalsystem, da nämlich

$$\begin{aligned}
&\int_0^{2\pi} \sin(kt)\cos(lt)\,dt = 0 \;\forall k,l,\\
&\int_0^{2\pi} \sin(kt)\sin(lt)\,dt = 0 = \int_0^{2\pi} \cos(kt)\cos(lt)\;\forall k \neq l,\\
&\int_0^{2\pi} \cos^2(kt)\,dt = \pi = \int_0^{2\pi} \sin^2(kt)\,dt\;\forall k \geq 1,
\end{aligned}$$

wie man leicht mit Schulmitteln oder mit Methoden des ersten Semesters berechnen kann. Da außerdem die Beziehung

$$e^{int} = \cos(nt) + i\sin(nt)$$

besteht (siehe Abschnitt 7.4), können wir auch leicht zwischen den beiden Orthonormalsystemen umrechnen, wenn wir zusätzlich die bekannten Eigenschaften $\cos(x) = \cos(-x)$ und $\sin(x) = -\sin(-x)$ verwenden (diese wiederum folgen direkt aus den Formeln in Beispiel 7.3):

$$\begin{aligned}
\cos(nt) &= \frac{\cos(nt)+\cos(-nt)}{2} = \frac{e^{int}+e^{-int}}{2} = \frac{1}{2}(e_n(t)+e_{-n}(t)),\\
\sin(nt) &= \frac{\sin(nt)-\sin(-nt)}{2} = \frac{e^{int}-e^{-int}}{2i} = \frac{1}{2i}(e_n(t)-e_{-n}(t)).
\end{aligned}$$

Definition 27.1. *Sei $f \in V$ eine reellwertige Funktion. Dann heißen*

$$\begin{aligned}
a_k &:= \frac{1}{\pi}\int_0^{2\pi} f(t)\cos(kt)\,dt,\; k = 0,1,\ldots,\\
b_k &:= \frac{1}{\pi}\int_0^{2\pi} f(t)\sin(kt)\,dt,\; k = 1,2,\ldots,
\end{aligned}$$

*die **Fourierkoeffizienten** von f und*

$$\frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k\cos(kt) + b_k\sin(kt))$$

*die **Fourierreihe** von f.*

Ähnlich wie ein Taylorpolynom eine Funktion approximiert (siehe Beispiel 15.5), können wir versuchen, mit einem sogenannten trigonometrischen Polynom eine 2π–periodische Funktion anzunähern. Genauso wie die Taylorreihe zu einem Polynom gerade wieder das ursprüngliche Polynom ist, ist auch die Fourierreihe zu einem trigonometrischen Polynom wieder das ursprüngliche, wie das folgende Beispiel zeigt:

Beispiel/Definition 27.2. Ist

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx)), \quad a_k \in \mathbb{R}, b_k \in \mathbb{R},$$

ein **trigonometrisches Polynom**, d.h. ein Polynom in $(\sin(x), \cos(x))$ vom Grad $\leq n$, so sind

$$a_0, a_1, \ldots, a_n, 0, \ldots \text{ und } b_1, \ldots, b_n, 0 \ldots$$

die Fourierkoeffizienten von f und die Fourierreihe gibt in diesem Fall gerade die Funktion zurück. Setzen wir nämlich $f(x)$ in die Formeln aus der Definition 27.1 ein, so erhalten wir, weil die $\sin(kx)$ und $\cos(kx)$ eine Orthonormalbasis bilden, tatsächlich die angegebenen Werte.

Im Allgemeinen können wir die Partialsumme

$$\frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos(kt) + b_k \sin(kt))$$

der Fourierreihe von f als Bild von f unter der orthogonalen Projektion auf den von den trigonometrischen Polynomen vom Grad $\leq n$ aufgespannten Untervektorraum $U \subset V$ und damit als beste Approximation (bzgl. der zugehörigen Norm) der Reihe durch solch ein trigonometrisches Polynom auffassen.

Häufig schreibt man wegen $e^{int} = \cos(nt) + i \sin(nt)$ die Fourierreihen auch in der Form

$$\sum_{-\infty}^{\infty} c_k \cdot e^{ikx} \text{ mit } c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) \cdot e^{-ikx} \, dx.$$

Eine Fourierreihe in dieser Form konvergiert, wenn die Funktionenfolge

$$s_n(x) = \sum_{-n}^{n} c_k \cdot e^{ikx}$$

konvergiert.

27.2 Fourierreihen und Konvergenz

Im ersten Semester haben wir verschiedene Arten von Konvergenz kennen gelernt. Wir werden sehen, dass einige spezielle Fourierreihen besonders gute Konvergenzeigenschaften besitzen, was uns erlauben wird, einige Grenzwerte konkret zu berechnen. Insbesondere werden wir den Wert einer Reihe ausrechnen können, für dessen Bestimmung Euler unter anderem bekannt ist.

Satz 27.3. *Die Fourierreihe*

$$\sum_{k=1}^{\infty} \frac{\sin(kx)}{k}$$

konvergiert punktweise (d.h. für jedes feste x, siehe Definition 16.1 für Details) gegen die ***Zackenfunktion*** *(siehe Abb. 27.1)*

$$f\colon \mathbb{R} \to \mathbb{R},\ f(x) = \begin{cases} \frac{\pi - x}{2}, & x \in]0, 2\pi[, \\ 0, & x = 0 \end{cases}$$

und $f(x + 2\pi k) = f(x)$ für $k \in \mathbb{Z}$. Für jedes δ mit $0 < \delta < \frac{\pi}{2}$ ist die Konvergenz auf $]\delta, 2\pi - \delta[$ gleichmäßig.

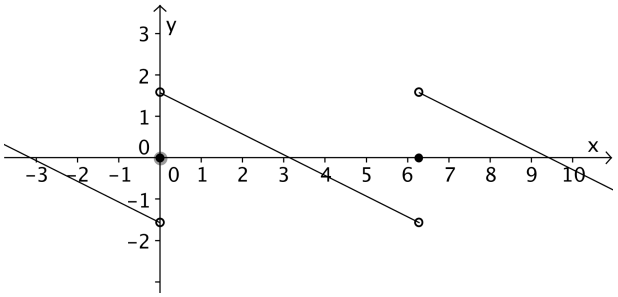


Abbildung 27.1. Eine Zackenfunktion als Grenzfunktion der Fourierreihe.

Bemerkung 27.4. Nach Abschnitt 16.1 heißt eine Folge (f_n) von reellwertigen Funktionen auf einem Intervall I gleichmäßig konvergent gegen eine Grenzfunktion $f\colon I \to \mathbb{R}$, wenn: $\forall \varepsilon > 0\ \exists\ n_0 :\ |f_n(x) - f(x)| < \varepsilon\ \forall n \geq n_0\ \forall\ x \in I$. Die Konvergenz im Satz kann wegen des Satzes 16.4 über die Stetigkeit eines gleichmäßigen Limes stetiger Funktionen nicht überall gleichmäßig sein, da die Grenzfunktion nicht stetig ist.

Wir verwenden für den Beweis des Satzes folgenden Hilfssatz:

Lemma 27.5. *Für $t \in \mathbb{R}$, das kein ganzzahliges Vielfaches von 2π ist, gilt:*

$$\frac{1}{2} + \sum_{k=1}^{n} \cos(kt) = \frac{\sin((n+\frac{1}{2})t)}{2\sin(\frac{t}{2})}.$$

Beweis. Es gilt wegen $\cos(kx) = \frac{1}{2}(e^{ikx} + e^{-ikx})$ und $\sum_{k=0}^{2n} x^k = \frac{1-x^{2n+1}}{1-x}$ (weil $(1 + x + x^2 + \cdots x^k) \cdot (1-x) = 1 + x + x^2 + \cdots + x^k - x - x^2 - \cdots - x^k - x^{k+1}$ ist):

$$\begin{aligned}
\frac{1}{2} + \sum_{k=1}^{n} \cos(kt) &= \frac{1}{2} \cdot \sum_{k=-n}^{n} e^{ikt} \\
&= \frac{1}{2} \cdot e^{-int} \cdot \sum_{k=0}^{2n} e^{ikt} \\
&= \frac{1}{2} \cdot e^{-int} \cdot \frac{1 - e^{(2n+1)it}}{1 - e^{it}} \\
&= \frac{1}{2} \cdot \frac{e^{i(n+\frac{1}{2})t} - e^{-i(n+\frac{1}{2})t}}{e^{i\frac{t}{2}} - e^{-i\frac{t}{2}}} \\
&= \frac{\sin((n+\frac{1}{2})t)}{2\sin(\frac{t}{2})},
\end{aligned}$$

wie behauptet war. □

Beweis (von Satz 27.3). Zunächst zur punktweisen Konvergenz: Für $x = 0, \pi, 2\pi$ ist die Aussage klar. Sei nun $x \in]\pi, 2\pi[$, dann liefert das Lemma, da $\int_\pi^x \cos(kt)\,dt = \left[\frac{1}{k}\sin(kt)\right]_\pi^x = \frac{1}{k}\sin(kx)$:

$$\begin{aligned}
\sum_{k=1}^{n} \frac{\sin(kx)}{k} &= \sum_{k=1}^{n} \int_\pi^x \cos(kt)\,dt \\
&= \int_\pi^x \Big(\frac{\sin((n+\frac{1}{2})t)}{2\sin(\frac{t}{2})} - \frac{1}{2} \Big)\,dt \\
&\xrightarrow[n\to\infty]{} \int_\pi^x -\frac{1}{2}\,dt = \frac{\pi - x}{2},
\end{aligned}$$

weil wir folgenden Hilfssatz für $f(x) = \frac{1}{\sin(\frac{t}{2})}$ anwenden können:

Lemma 27.6. *Sei $f\colon [a,b] \to \mathbb{R}$ eine stetig differenzierbare Funktion. Es gilt*

$$\int_a^b f(x)\sin(kx)\,dx \xrightarrow[k\to\infty]{} 0.$$

Siehe dazu auch Abb. 27.2. Analog erhält man:

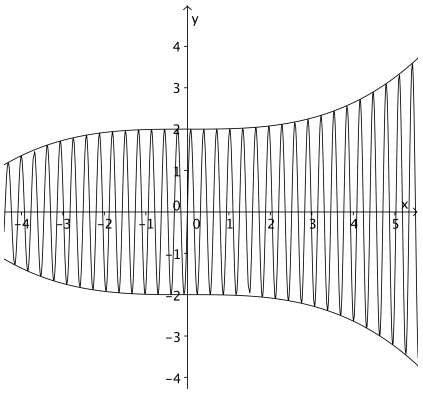


Abbildung 27.2. Der Grenzwert des Integrals $\int_a^b f(x)\sin(kx)\,dx$ für $k \to \infty$ ist die Nullfunktion. Anschaulich erkennt man, dass sich die Flächeninhalte über und unter der x-Achse immer mehr angleichen.

$$\int_a^b f(x)\cos(kx)\,dx \underset{k\to\infty}{\longrightarrow} 0.$$

Beweis. Für $k \neq 0$ ist eine Stammfunktion $F(x)$ des ersten Integranden (mit partieller Integration, Satz 13.22):

$$F(x) = \Big[-f(x)\frac{\cos(kx)}{k}\Big]_a^b + \frac{1}{k}\int_a^b f'(x)\cos(kx)\,dx.$$

Da f und f' stetig sind, existiert wegen des Satzes 8.10 zur Existenz von Maximum und Minimum stetiger Funktionen ein $M \in \mathbb{R}$, so dass

$$|f(x)| \leq M,\ |f'(x)| \leq M\ \forall [a,b].$$

Es folgt: $|F(k)| \leq \frac{2M}{k} + \frac{M(b-a)}{k} \underset{k\to\infty}{\longrightarrow} 0.$ □

Für $x \in]0,\pi[$ zeigt man die Aussage analog.

Nun zur gleichmäßigen Konvergenz auf $]\delta, 2\pi - \delta[$. Es gilt:

$$s_n(x) := \sum_{k=1}^{n} \sin(kx) = \Im\mathfrak{m}\Big(\sum_{k=1}^{n} e^{ikx}\Big).$$

Da $\delta \leq x \leq 2\pi - \delta$, ist:

$$|s_n(x)| \le \Big|\sum_{k1=1}^{n} e^{ikx}\Big| = \Big|\frac{e^{i(n+1)x}-1}{e^{ix}-1}\Big|$$
$$\le \frac{2}{e^{ix/2}-e^{-ix/2}} = \frac{1}{\sin x/2} \le \frac{1}{\sin \delta/2}.$$

Für $m > n > 0$ folgt:

$$\Big|\sum_{k=n}^{m} \frac{\sin kx}{k}\Big| = \Big|\sum_{k=n}^{m} \frac{s_k(x)-s_{k-1}(x)}{k}\Big|$$
$$= \Big|\sum_{k=n}^{m} s_k(x)\Big(\frac{1}{k}-\frac{1}{k+1}\Big) + \frac{s_m(x)}{m+1} - \frac{s_{n-1}(x)}{n}\Big|$$
$$\le \frac{1}{\sin \delta/2} \cdot \Big(\frac{1}{n} - \frac{1}{m+1} + \frac{1}{m+1} + \frac{1}{n}\Big)$$
$$\le \frac{2}{n \sin \delta/2}$$

Es folgt, dass $\sum_{k=1}^{n} \frac{\sin(kx)}{k}$ auf $]\delta, 2\pi - \delta[$ eine gleichmäßige Cauchy–Folge ist; wir erhalten also die gleichmäßige Konvergenz auf $]\delta, 2\pi - \delta[$. □

Dies erlaubt es uns, einige Grenzwerte konkret zu bestimmen:

Korollar 27.7. *Die Fourierreihe*

$$\sum_{k=1}^{\infty} \frac{\cos(kx)}{k^2}$$

konvergiert auf $[0, 2\pi]$ *gleichmäßig gegen die Funktion*

$$F(x) = \Big(\frac{x-\pi}{2}\Big)^2 - \frac{\pi^2}{12}.$$

Beweis. Die Konvergenz ist gleichmäßig, da $\sum_{k=1}^{\infty} \frac{1}{k^2}$ eine konvergente Majorante ist. Die Folge der Ableitungen, $-\sum_{k=1}^{\infty} \frac{\sin(kx)}{k}$, konvergiert auf $]\delta, 2\pi - \delta[$ gleichmäßig gegen $\frac{\pi - x}{2}$. Es folgt mit Korollar 16.8, dass F auf $[0, 2\pi]$ diffbar ist mit $F'(x) = \frac{x-\pi}{2}$. Integration liefert:

$$F(x) = \Big(\frac{x-\pi}{2}\Big)^2 + c$$

für eine Konstante $c \in \mathbb{R}$. Um c zu bestimmen, betrachten wir

$$\int_0^{2\pi} F(x)\,dx = \int_0^{2\pi} \Big(\frac{x-\pi}{2}\Big)^2 dx + \int_0^{2\pi} c\,dx = \frac{\pi^3}{6} + 2\pi c.$$

Andererseits gilt (wegen der gleichmäßigen Konvergenz dürfen wir nach Satz 16.6 Grenzwert und Integral vertauschen):

$$\int_0^{2\pi} \Big(\sum_{k=1}^{\infty} \frac{\cos(kx)}{k^2}\Big)\, dx = \sum_{k=1}^{\infty} \frac{1}{k^2} \int_0^{2\pi} \cos(kx)\, dx = \sum_{k=1}^{\infty} \frac{1}{k^2} \cdot 0 = 0.$$

Es folgt: $c = -\frac{\pi^2}{12}$. □

Insbesondere erhalten wir an der Stelle 0 für $F(x)$ die Folgerung:

Korollar 27.8.

$$\zeta(2) := \sum_{k=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Beweis. Mit der Notation aus dem vorigen Korollar erhalten wir unmittelbar: $\sum_{n=1}^{\infty} \frac{1}{n^2} = F(0) = (\frac{\pi}{2})^2 - \frac{\pi^2}{12} = \frac{\pi^2}{6}$. □

Das Problem, diesen Grenzwert $\sum_{k=1}^{\infty} \frac{1}{n^2}$ zu berechnen, formulierte als erster wohl Pietro Mengoli im Jahr 1644. Erst 1735 fand Euler den Wert heraus. Seitdem wurden sehr viele verschiedene Wege gefunden, dieses Resultat zu erhalten. Besonderes Interesse hat die Summe auch, weil sie der spezielle Wert $\zeta(2)$ der **Riemannschen Zeta–Funktion** für $n \in \mathbb{N}$ ist:

$$\zeta(n) = \sum_{k=1}^{\infty} \frac{1}{k^n}.$$

Die Webseite http://mathworld.wolfram.com/RiemannZetaFunction.html gibt dazu recht viele Hintergrundinformationen.

27.3 Besselsche Ungleichung und Vollständigkeitsrelation

Eine der zentralen Ungleichungen im Zusammenhang mit Fourierreihen und der Funktionalanalysis im Allgemeinen ist die Besselsche Ungleichung. Auch sie und verwandte Sätze werden es uns erlauben, konkret einige Fourierreihen zu berechnen und als Spezialfälle einige berühmte Reihengrenzwerte zu bestimmen. Ein erstes Resultat in diesem Zusammenhang ist folgendes:

Satz 27.9. *Sei $f \in V$ und seien*

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-ikt}\, dt$$

die Fourierkoeffizienten. Dann gilt:

$$\left\|f - \sum_{k=-n}^{n} c_k e_k\right\|^2 = \|f\|^2 - \sum_{k=-n}^{n} |c_k|^2.$$

Beweis. Es bezeichne $s := \sum_{k=-n}^{n} c_k e_k$ die orthogonale Projektion von f auf den Untervektorraum $\langle e_{-n}, \ldots, e_n \rangle$ mit (siehe Satz 26.30; die $e_k := e^{ikt}$ bilden ja nach Beispiel 26.25.3 eine Orthonormalbasis). Dann ist $g := f - s$ die orthogonale Projektion von f auf $\langle e_{-n}, \ldots, e_n \rangle^\perp$ (ebenfalls Satz 26.30). Insbesondere ist $s \perp g$. Mit dem Satz des Pythagoras folgt:

$$\|f\|^2 = \|g\|^2 + \|s\|^2 = \left\|f - \sum_{k=-n}^{n} c_k e_k\right\|^2 + \sum_{k=-n}^{n} |c_k|^2.$$

□

Wir sehen also, dass hier die Theorie der Orthonormalbasen und der orthogonalen Projektion auf Funktionenräumen relevant eingeht. Eine wichtige Folgerung ist:

Korollar 27.10 (Besselsche Ungleichung). *Sei $f : \mathbb{R} \to \mathbb{C}$ eine über $[0, 2\pi]$ integrierbare 2π–periodische Funktion und seien $(c_k)_{k\in\mathbb{Z}}$ die Fourierkoeffizienten. Dann gilt:*

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

Beweis. Nach Proposition 27.9 gilt insbesondere:

$$\|f\|^2 - \sum_{k=-n}^{n} |c_k|^2 \geq 0.$$

Mit $n \to \infty$ folgt die Behauptung. □

Im allgemeineren Kontext eines Hilbertraumes sagt die Besselsche Ungleichung aus, dass ein Vektor mindestens so lang ist wie eine beliebige Projektion auf einen Unterraum, d.h. $\|f\|^2 \geq \sum_{k=1}^{n} |\langle f_n, f \rangle|^2$, wobei f_n ein Orthonormalbasis des Unterraumes ist. Gilt in der Besselschen Ungleichung Gleichheit, so heißt sie **Parsevalsche Gleichung** und ist eine allgemeine Form des Satzes von Pythagoras.

Wir fragen nun, ob die Fourierreihe von f gegen f konvergiert. Es stellt sich heraus, dass im Allgemeinen weder gleichmäßige noch punktweise Konvergenz vorliegt. Den besten Konvergenzbegriff für Fourierreihen gibt Konvergenz im quadratischen Mittel:

Definition 27.11. *Seien $f\colon \mathbb{R} \to \mathbb{C}$ und $f_n \in \mathbb{R} \to \mathbb{C}$ Funktionen aus V. Die Folge $(f_n)_{n\in\mathbb{N}}$* ***konvergiert im quadratischen Mittel*** *gegen f, wenn*

$$\lim_{n\to\infty} \|f - f_n\|_2 = 0.$$

Satz 27.12 (Vollständigkeitsrelation). *Für jede 2π–periodische und über $[0, 2\pi]$ integrierbare Funktion $f\colon \mathbb{R} \to \mathbb{C}$ gilt: Die Fourierreihe von f,*

$$\sum_{k=-\infty}^{\infty} c_k \cdot e^{ikx},$$

konvergiert im quadratischen Mittel gegen f und es gilt:

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = (\|f\|_2)^2.$$

Bemerkung 27.13. Die Vollständigkeitsrelation besagt, dass der $\|.\|_2$–Abschluss (dieser Begriff geht leider über den Inhalt dieser Vorlesung hinaus) des von den e_k erzeugten Unterraum von V, ganz V ist.

Wir werden die Relation gleich erst beweisen. Zunächst aber einige Anwendungen; mit ihrer Hilfe können wir beispielsweise $\zeta(2)$ auch bestimmen:

Korollar 27.14.

$$\zeta(2) = \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

Beweis. Wir betrachten $f(x) := \sum_{k=1}^{\infty} \frac{\sin(kx)}{k} = \frac{\pi - x}{2}$ auf $]0, 2\pi[$ und erhalten:

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) \cdot e^{-ikx}\, dx = \begin{cases} \cdots = -\frac{i}{2k}, & k \neq 0, \\ \frac{\pi}{2} - \frac{1}{4\pi} \cdot [\frac{x^2}{2}]_0^{2\pi} = 0, & k = 0. \end{cases}$$

Mit der Vollständigkeitsrelation folgt:

$$\begin{aligned} \frac{1}{2} \sum_{k=1}^{\infty} \frac{1}{k^2} &= \sum_{k=-\infty, k\neq 0}^{\infty} \left|-\frac{i}{2k}\right|^2 \\ &= \frac{1}{2\pi} \int_0^{2\pi} \Big(\frac{\pi - x}{2}\Big)^2\, dx \\ &= \frac{1}{8\pi} \Big[\Big((\pi - x)^3 \cdot \Big(-\frac{1}{3}\Big)\Big]_0^{2\pi} \\ &= -\frac{1}{24\pi} \cdot (-\pi^3 - \pi^3) = \frac{\pi^2}{12}, \end{aligned}$$

was die Behauptung liefert. □

Korollar 27.15.

$$\sum_{k=1}^{\infty} \frac{1}{k^4} = \frac{\pi^4}{90}.$$

Beweis. Wir wissen:

$$\begin{aligned}\frac{(\pi - x)^2}{4} &= \frac{\pi^2}{12} + \sum_{k=1}^{\infty} \frac{\cos(kx)}{k^2} \\ &= \frac{\pi^2}{12} + \sum_{k=1}^{\infty} \left(\frac{1}{2k^2}(e^{ikx} + e^{-ikx})\right).\end{aligned}$$

Es folgt:

$$\frac{\pi^4}{144} + 2\sum_{k=1}^{\infty} \frac{1}{4k^2} = \frac{1}{2\pi} \int_0^{2\pi} \frac{(\pi - x)^4}{16}\, dx = \frac{\pi^4}{90}.$$

Also:

$$\sum_{k=1}^{\infty} \frac{1}{k^4} = \frac{\pi^2}{90},$$

wie behauptet. □

Nun zum Beweis der Vollständigkeitsrelation (Satz 27.12). Wir beweisen sie zunächst für einen Spezialfall:

Lemma 27.16. *Satz 27.12 gilt für 2π–periodische Treppenfunktionen.*

Beweis. Es genügt, den Spezialfall

$$f(x) = \begin{cases} 1, & 0 \le x < a, \\ 0, & a \le x < 2\pi \end{cases}$$

zu betrachten. Andere Treppenfunktionen entstehen aus solchen nämlich durch Linearkombination und da $\|.\|_2$ der Dreiecksungleichung genügt, reicht dies.

Die Fourierkoeffizienten von f sind offenbar:

$$\begin{aligned} c_0 &= \frac{a}{2\pi}, \\ c_k &= \frac{1}{2\pi} \int_0^a e^{-ikx}\, dx = \frac{i}{2\pi k}(e^{-ika} - 1) \quad \text{für } k \neq 0. \end{aligned}$$

Für $k \neq 0$ gilt:

$$|c_k|^2 = \frac{1}{4\pi^2 k^2}(1 - e^{ika})(1 - e^{-ika}) = \frac{1 - \cos(ka)}{2\pi^2 k^2}.$$

Es folgt:

$$\begin{aligned}\sum_{k=-\infty}^{\infty} |c_k|^2 &= \frac{a^2}{4\pi^2} + \sum_{k=1}^{\infty} \frac{1 - \cos ka}{\pi^2 k^2} \\ &= \frac{a}{4\pi^2} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2} - \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{\cos ka}{k^2} \\ &= \frac{a}{4\pi^2} + \frac{1}{6} - \frac{1}{\pi^2}\left(\frac{(\pi - a)^2}{4} - \frac{\pi^2}{12}\right) \\ &= \frac{a}{2\pi}.\end{aligned}$$

Andererseits ist

$$\|f\|_2 = \frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\, dx = \frac{a}{2\pi}.$$

Die Besselsche Ungleichung ist also eine Gleichheit und es folgt:

$$\|f - \sum_{k=-n}^{n} c_k \cdot e^{ikx}\|_2 \underset{n\to\infty}{\longrightarrow} 0,$$

was zu zeigen war. □

Beweis (allgemeiner Fall der Vollständigkeitsrelation Satz 27.12). Sei $f\colon \mathbb{R} \to \mathbb{C}$ 2π–periodisch und über $[0, 2\pi]$ integrierbar. Dann ist f insbesondere beschränkt. Ohne Einschränkung können wir also annehmen, dass f reellwertig ist und $|f(x)| \le 1$ für jedes x gilt.

Sei nun $\varepsilon > 0$ vorgegeben. Dann existieren Treppenfunktionen φ, ψ auf $[0, 2\pi]$ mit

- $-1 \le \varphi \le f \le \psi \le 1$,
- $\int_0^{2\pi} (\psi(x) - \varphi(x))\, dx \le \frac{\pi}{4}\varepsilon^2$.

Dann gilt für $g := f - \varphi$, dass $|g|^2 \le |\psi - \varphi|^2 \le 2(\psi - \varphi)$, also:

$$\frac{1}{2\pi}\int_0^{2\pi} |g(x)|^2\, dx \le \frac{1}{\pi}\int_0^{2\pi} 2(\psi(x) - \varphi(x))\, dx \le \frac{1}{4}\varepsilon^2.$$

Es seien $S_{f,n}, S_{\varphi,n}, S_{g,n}$ die n-ten Partialsummen der Fourierriehe von f, φ bzw. g, d.h.:

$$S_{f,n} = S_{\varphi,n} + S_{g,n}.$$

Nach Satz 27.9 existiert ein N, so dass

$$\|\varphi - S_{\varphi,n}\|_2 \leq \frac{\varepsilon}{2} \quad \forall n \geq N.$$

Wieder nach dem gleichen Satz 27.9

$$\|g - S_{g,n}\|_2^2 \leq \|g\|_2^2 \leq \frac{1}{4}\varepsilon^2,$$

also $\|g - S_{g,n}\|_2 \leq \frac{1}{2}\varepsilon$. Es folgt:

$$\begin{aligned}\|f - S_{f,n}\|_2 &\leq \|\varphi - S_{\varphi,n}\|_2 g - S_{g,n}\|_2 \\ &\leq \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon \quad \forall n \in N.\end{aligned}$$

□

Die verschiedenen Konvergenzbegriffe hängen folgendermaßen zusammen:

$$\text{gleichmäßige Konvergenz} \quad \begin{array}{ll} \Rightarrow & \text{Konvergenz im quadratischen Mittel} \\ \Rightarrow & \text{punktweise Konvergenz.} \end{array}$$

Weitere Implikationen gelten nicht. Unter gewissen Bedingungen gilt aber sogar gleichmäßige Konvergenz:

Satz 27.17. *Sei $f\colon \mathbb{R} \to \mathbb{C}$ eine stetige, stückweise stetig diffbare Funktion, d.h. es existiert eine Unterteilung*

$$0 = t_0 < t_1 < \cdots < t_r = 2\pi$$

von $[0, 2\pi]$, so dass $f|_{[t_{j-1},t_j]}$ stetig diffbar ist. Dann konvergiert die Fourierreihe gleichmäßig gegen f.

Beweis. Es sei $\varphi_j\colon [t_{j-1}, t_j] \to \mathbb{R}$ die stetige Ableitung von $f|_{[t_{j-1},t_j]}$ und φ die periodische Funktion mit $\varphi|_{[t_{j-1},t_j]} = \varphi_j$.

Für die Fourier-Koeffizienten γ_k von φ gilt nach der Besselschen Ungleichung:

$$\sum_{k=-\infty}^{\infty} |\gamma_k|^2 \leq \|\varphi\|_2^2 < \infty.$$

Für $k \neq 0$ lassen sich die Fourier-Koeffizienten c_k von f aus denen von φ durch partielle Integration gewinnen:

$$\begin{aligned}\int_{t_{j-1}}^{t_j} f(x)e^{ikx}\,dx &= \int_{t_{j-1}}^{t_j} f(x)cos(kx)\,dx - i\int_{t_{j-1}}^{t_j} f(x)sin(kx)\,dx\\ &= \Big[\frac{1}{k}f(x)\sin(kx) + \frac{i}{k}f(x)\cos(kx)\Big]_{t_{j-1}}^{t_j}\\ &\qquad -\frac{1}{k}\int_{t_{j-1}}^{t_j}\varphi_j(x)\sin(kx)\,dx - \frac{i}{k}\int_{t_{j-1}}^{t_j}\varphi_j(x)\cos(kx)\,dx\\ &= \frac{i}{k}\Big[f(x)e^{ikx}\Big]_{t_{j-1}}^{t_j} - \int_{t_{j-1}}^{t_j}\varphi(x)e^{-ikx}\,dx.\end{aligned}$$

Damit folgt:

$$\begin{aligned}c_k &= \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx\\ &= \frac{1}{2\pi}\sum_{j=1}^{r}\int_{t_{j-1}}^{t_j} f(x)e^{-ikx}\,dx\\ &= \frac{-i}{2\pi k}\int_0^{2\pi}\varphi(x)e^{-ikx} = \frac{-i\gamma_k}{k}.\end{aligned}$$

Da $|\alpha\beta| \le 2(|\alpha|^2 + |\beta|^2)$ für $\alpha, \beta \in \mathbb{C}$, folgt:

$$|c_k| \le 2\Big(\frac{1}{k^2} + |\gamma_k|^2\Big).$$

Die Reihen $\sum \frac{1}{k^2} < \infty$ und $\sum |\gamma_k|^2 < \infty$ liefern also eine konvergente Majorante.

Die Reihe $c_k e^{ikx}$ konvergiert deshalb gleichmäßig gegen eine stetige Funktion g. Im Mittel konvergiert die Fourierreihe also gegen g und gegen f, d.h.

$$\|f - g\|_2 = 0.$$

Da aber f und g stetig sind, folgt: $f = g$. □

Wie schon angedeutet, haben Fourierreihen viele Anwendungen:

Signalverarbeitung: Die Fourierkoeffizienten eines Signals geben die Anteile der einzelnen Frequenzen an. a_k und b_k mit kleinem k entsprechen niedrigen Frequenzen, solche mit großem k hohen Frequenzen. Dadurch kann man beispielsweise Filter produzieren, die gewisse Frequenzbereiche dämpfen.

Bildverarbeitung: Ähnlich zur Signalverarbeitung. Hier entsprechen niedrige Frequenzen großräumigen Bildstrukturen und höhere Frequenzen Details.

In beiden genannten Anwendungsbereichen liegen die Daten meist diskret vor (beispielsweise als einzelne Pixel oder abgetastete Signale). Dann verwendet man die sogenannte **diskrete Fouriertransformation**, in der Integrale durch Summen ersetzt werden. Hierfür existieren sehr schnelle Algorithmen (**Fast Fourier Transform** (**FFT**)), die beispielsweise ein Signal mit n Werten mit einer Laufzeit von $O(n \log n)$ in seine Frequenzanteile zerlegen.

Wavelets sind eine Weiterentwicklung der Fourierreihen, die sowohl Frequenz als auch Ort berücksichtigen. Diese liefern das sehr effiziente Verfahren zur Signal– und Bildkompression namens JPEG2000, da viele der Koeffizienten sehr klein sind und ohne großen Nachteil weggelassen werden können. Auch hierfür existieren effiziente Algorithmen mit einer Laufzeit von $O(n)$.

Aufgaben

Aufgabe 27.1 (. . .). . . .

28

Singulärwertzerlegung

Für quadratische Matrizen hatten wir Eigenvektoren und Eigenwerte definiert und festgestellt, dass sie essentielle Informationen zu den durch die Matrix definierten Abbildungen liefern, etwa durch ihren Einfluss beim Diagonalisieren oder der Klassifikation der Quadriken mit Hilfe der Hauptachsentransformation.

In diesem Abschnitt entwickeln wir nun ein ähnliches Konzept für nichtquadratische Matrizen; die auftretenden Zahlen heißen nun Singulärwerte und sind im Gegensatz zu den Eigenwerten immer reell. Sie sind also keine direkte Verallgemeinerung der Eigenwerte.

Wir werden sehen, dass wir mit Hilfe der hier vorgestellten Methoden nummerisch interessante Berechnungen durchf"hren können. Beispielsweise werden wir sehen, dass wir den Rang einer Matrix auf nummerisch stabile Weise ermitteln können.

28.1 Die Singulärwertzerlegung

Satz/Definition 28.1 (Singulärwertzerlegung). *Sei $A \in \mathbb{R}^{m\times n}$. Dann existieren $\sigma_1, \dots, \sigma_p \in \mathbb{R}$ mit $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ sowie $U \in O(m)$ und $V \in O(n)$, so dass*

$$U^t A V = \Sigma := \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \\ \vdots & & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix},$$

*wobei $p = \min(m, n)$. Die σ_i heißen **Singulärwerte** von A. Eine Darstellung der Form $A = U\Sigma V^t$ heißt **Singulärwertzerlegung** (englisch **singular value decomposition**, kurz **SVD**).*

Bevor wir diesen Satz beweisen, zunächst einige Beispiele:

Beispiel 28.2. 1. Als erstes Beispiel nehmen wir eine quadratische Matrix:

$$A_1 = \begin{pmatrix} 4 & 12 \\ 12 & 11 \end{pmatrix} = \begin{pmatrix} 3/5 & 4/5 \\ 4/5 & -3/5 \end{pmatrix} \cdot \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix} \cdot \begin{pmatrix} 3/5 & 4/5 \\ -4/5 & 3/5 \end{pmatrix}$$

2. Die Singulärwertzerlegung von orthogonalen Matrizen ist besonders einfach:
$$A_2 = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

3. Wegen der in der Einleitung erwähnten numerischen Stabilität werden häufig Singulärwertzerlegungen für (auch nicht–quadratische) Matrizen mit Dezimalzahlen als Einträgen berechnet:

$$A_3 = \begin{pmatrix} 0.36 & 1.60 & 0.48 \\ 0.48 & -1.20 & 0.64 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.6 \\ -0.6 & 0.8 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 \\ 0.6 & 0 & 0.8 \\ -0.8 & 0 & 0.6 \end{pmatrix}$$

4. Offenbar gilt
$$\operatorname{rang}\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0, \ \operatorname{rang}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 1.$$
Die beiden Eigenwerte sind in beiden Fällen jeweils $0,0$. Die Singulärwerte sind dagegen, wie man berechnen kann, $0,0$ bzw. $0,1$. In diesem Fall sagen die Eigenwerte also nichts über den Rang der Matrix aus, die Anzahl der Singulärwerte, die nicht 0 sind, ist aber gerade der Rang. Wir werden im Kapitel zur Numerik noch sehen, dass dies kein Zufall und eine sehr wichtige Eigenschaft der Singulärwerte ist.

Einer der Gründe für deren Wichtigkeit ist, dass diese Rang-Betrachtung auch noch funktioniert, wenn die Einträge der Matrix leicht fehlerbehaftet sind, also wenn z.B. einer der Einträge nicht 0, sondern $\varepsilon > 0$ ist:
$$A = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}.$$
Da $\chi_A(t) = t^2 - \varepsilon$ ist, sind die Eigenwerte $\pm\sqrt{\varepsilon}$. Die Singulärwerte sind $\sigma_1 = 1, \sigma_2 = \varepsilon$. Und wenn ε gegen 0 geht, strebt der Rang der Matrix tatsächlich gegen 1, genauso wie die Anzahl der von 0 verschiedenen Singulärwerte.

Beweis (von Satz/Definition 28.1 über die Singulärwertzerlegung). Wir konstruieren eine Singulärwertzerlegung von A:

Zunächst setzen wir $B := A^t A$. Dies ist eine reelle symmetrische $n \times n$–Matrix und hat daher nur reelle Eigenwerte λ_i, die wir so nummerieren, dass $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Mit $\{v_1, \ldots, v_n\}$ bezeichnen wir eine Basis des $\mathbb{R}^n$ aus orthonormalen Eigenvektoren von B zu den entsprechenden Eigenwerten. Tatsächlich sind alle λ_i nicht negativ, denn es gilt einerseits, da die v_i orthonormal sind,

$$v_i^t \cdot B \cdot v_i = \lambda_i \cdot v_i^t \cdot v_i = \lambda_i$$

und andererseits, nach Definition von B und da das Skalarprodukt positiv definit ist:

$$v_i^t \cdot B \cdot v_i = v_i^t \cdot A^t \cdot A \cdot v_i = \langle Av_i, Av_i \rangle \geq 0.$$

Da $r := \operatorname{rang} A = \operatorname{rang} B$, sind genau die ersten r Eigenwerte $\lambda_1, \ldots, \lambda_r$ strikt positiv.

Wir setzen nun für $i = 1, \ldots, r$

$$u_i := \frac{1}{\sqrt{\lambda_i}} A v_i$$

und ergänzen diese durch $m - r$ orthonormale Vektoren $u_{r+1}, \ldots, u_m$, die außerdem zu $u_1, \ldots, u_r$ orthonormal sind, zu einer Basis des $\mathbb{R}^m$.

Wir bilden nun die beiden gesuchten Matrizen U und V aus den Spaltenvektoren u_i bzw. v_i:

$$U = (u_1 \ldots u_m), \quad V = (v_1 \ldots v_n).$$

Die Singulärwerte von A sind $\sigma_i := \sqrt{\lambda_i}$, für $i = 1, 2, \ldots, r$ und $\sigma_i = 0$ für $i = r + 1, \ldots, p$.

Es ist noch zu zeigen, dass $A = U\Sigma V^t$ dann tatsächlich eine Singulärwertzerlegung von A ist. Zunächst ist V orthogonal, da v_i nach Konstruktion eine Orthonormalbasis ist. Die u_i bilden ebenfalls eine Orthonormalbasis, denn für $i, j = 1, \ldots, r$ gilt

$$u_i^t u_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} v_i^t A^t A v_j = \frac{\lambda_j}{\sqrt{\lambda_i \lambda_j}} v_i^t v_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases}$$

und diese Orthonormalität setzt sich nach Konstruktion auf $u_{r+1}, \ldots, u_m$ fort.

Es bleibt also nur noch zu zeigen, dass wirklich $A = U\Sigma V^t$ gilt:

$$U\Sigma V^t = \sum_{i=1}^{r} \sqrt{\lambda_i} u_i v_i^t = \sum_{i=1}^{r} A v_i v_i^t = \sum_{i=1}^{n} A v_i v_i^t = A \cdot \sum_{i=1}^{n} v_i v_i^t = A \cdot I = A.$$

Wir haben also tatsächlich eine Singulärwertzerlegung von A konstruiert. □

In der Praxis führt die im Beweis angegebene Konstruktion der Singulärwertzerlegung auf das Problem, dass wir die Eigenwerte λ_i berechnen müssen,

was sich z.B. für Polynome hohen Grades ≥ 5 recht schwierig gestalten kann. Auch aus anderen Gründen gibt es weitere Methoden zur Berechnung einer Singulärwertzerlegung. G. Golub war einer der ersten, die solche geeigneten Methoden gefunden und damit die Anwendung dieser Zerlegung erst möglich gemacht haben. Leider können wir im Rahmen dieser Vorlesung nicht auf Details eingehen.

Korollar 28.3. *Sei $A = U\Sigma V$ die Singulärwertzerlegung von $A \in \mathbb{R}^{m\times n}$ mit Singulärwerten $\sigma_1 \geq \cdots \geq \sigma_p$ für $p = \min(m, n)$. $u_1, \ldots, u_m$ und $v_1, \ldots, v_n$ bezeichnen die Spalten von U bzw. V. Dann gilt:*

1. *$Av_i = \sigma_i u_i$ und $A^t u_i = \sigma_i v_i$ für $i = 1, 2, \ldots, p$.*
2. *Ist $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, so ist $\operatorname{rang} A = r$. Außerdem ist*

$$\operatorname{Ker} A = \langle v_{r+1}, \ldots, v_n \rangle \text{ und } \operatorname{Bild} A = \langle u_1, \ldots, u_r \rangle.$$

3. *Die Quadrate $\sigma_1^2, \ldots, \sigma_p^2$ der Singulärwerte sind die Eigenwerte von $A^t A$ und von AA^t zu den Eigenvektoren $v_1, \ldots, v_p$ bzw. $u_1, \ldots, u_p$.*

Beweis. Mit dem Satz ist dies recht einfach nachzurechnen und wird hier nicht vorgeführt. □

Bemerkung 28.4. Für symmetrischce Matrizen A sind die Singulärwerte gerade die Beträge der Eigenwerte von A. Sind alle Eigenwerte nicht-negativ, so ist die Hauptachsentransformation $A = S^t DS$ auch die Singulärwertzerlegung.

28.2 Die Pseudoinverse

Wir haben in vielen Situationen gesehen, wie hilfreich die Kenntnis der Inversen Matrix ist. Da diese aber leider nur für invertierbare Matrizen existiert, verallgemeinern wir den Begriff der Inversen nun auf quadratische Matrizen mit Determinante 0 sowie auf nicht-quadratische Matrizen:

Definition 28.5. *Sei $A \in \mathbb{R}^{m\times n}$. Eine Matrix $A^+ \in \mathbb{R}^{n\times m}$ heißt **Pseudoinverse** von A, wenn $\forall b \in \mathbb{R}^m$ der Vektor $x = A^+ b$ die bzgl. der euklidischen Norm kleinste Lösung der Minimierungsaufgabe*

$$\textit{Finde } x, \textit{ so dass } \|b - Ax\| \textit{ minimal ist,}$$

d.h. $A^+ b \in (\operatorname{Ker} A)^\perp$ und $\|b - AA^+ b\| = \min_{x\in\mathbb{R}^n} \|b - Ax\|$.

Es ist klar, dass für eine quadratische invertierbare Matrix A die Pseudoinverse gerade die Inverse ist: $A^+ = A^{-1}$. In diesem Sinne verallgemeinert die Pseudoinverse also den Begriff der Inversen.

Direkt aus der Definition der Pseudoinversen folgt eine ihrer wichtigen Anwendungen:

Anwendung 28.6. Ist ein Gleichungssystem $Ax = b$ nicht lösbar, so können wir mit Hilfe der Pseudoinversen immerhin die beste Näherung $\tilde{x}$ an eine Lösung finden, die nämlich den quadratischen Fehler $\|Ax - b\|$ minimiert: $\tilde{x} = A^+ b$. Sie liefert also eine Lösung des Problems der kleinsten Quadrate.

...

Tatsächlich ist A^+ eine lineare Abbildung und es gilt:

$$AA^+ \colon \mathbb{R}^m \to \operatorname{Bild} A$$

ist die orthogonale Projektion auf das Bild und

$$A^+A \colon \mathbb{R}^n \to (\operatorname{Ker} A)^\perp$$

ist die orthogonale Projektion auf das orthogonale Komplement von A.

Satz 28.7. *Sei $A \in \mathbb{R}^{m\times n}$ und sei $A = U\Sigma V^t$ ihre Singulärwertzerlegung mit Singulärwerten $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$. Dann ist mit der Notation*

$$D^+ = \begin{pmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ & & \frac{1}{\sigma_r} \\ & & \\ 0 & & 0 \end{pmatrix}$$

die Matrix $A^+ = VD^+U^t \in \mathbb{R}^{n\times m}$ die Pseudoinverse von A.

Beweis. ... □

Kennen wir also eine Singulärwertzerlegung, so auch automatisch die Pseudoinverse und können daher, wie oben erwähnt, beispielsweise für nicht lösbare Gleichungssysteme eine Näherungslösung mit kleinstem quadratischem Fehler bestimmen.

Aufgaben

Aufgabe 28.1 (). ...

Teil IV

Mehrdimensionale Analysis

Die mehrdimensionale Analysis ist ein essentielles Hilfsmittel für viele Bereiche der angewandten Mathematik. Reale Objekte sind nämlich meistens Kurve, Oberflächen oder Volumina und wenn wir diese untersuchen möchten, ist die mehrdimensionale Analysis für einige Aspekte ein geeignetes Mittel. Im Bereich der Computergrafik beschäftigt man sich beispielsweise mit solchen geometrischen Objekten.

Wir werden die mehrdimensionale Analysis außerdem in der Wahrscheinlichkeitstheorie und Statistik (Teil V) essentiell benötigen. Beispielsweise lässt sich auch die Fläche unter der bekannten Gaußschen Glockenkurve mit Hilfe mehrdimensionaler Analysis recht einfach berechnen.

Die meisten Grafiken in diesem Abschnitt stammen zwar vom ersten Autor, doch einige wurden dankenswerter Weise von einem Hörer, Felix Freiberger, bereit gestellt.

29

Kurven im $\mathbb{R}^n$

In diesem einleitenden Kapitel zur mehrdimensionalen Analysis betrachten wir einige Objekte, für deren Bearbeitung wir im Wesentlichen nur die Differential– und Integralrechnung in einer Veränderlichen benötigen, nämlich Kurven. Diese beschränken sich jetzt allerdings nicht mehr auf Funktionsgraphen von Funktionen einer Veränderlichen, wie wir sehen werden.

Dieses Kapitel ist auch eine gute Möglichkeit, die Begriffe und Konzepte aus dem Analysis Teil aus dem ersten Semester zu wiederholen und zu vertiefen, bevor wir auf kompliziertere Aspekte der mehrdimensionalen Analysis eingehen. Anwendungen der Kurven im $\mathbb{R}^n$ in der Informatik liegen beispielsweise im Bereich der Computergraphik, der geometrischen Modellierung und auch der Bilderkennung.

29.1 Elementare Definitionen und Beispiele

Definition 29.1. *Seien I ein Intervall und $f_1, \ldots, f_n \colon I \to \mathbb{R}$ stetige Funktionen. Dann heißt*

$$f \colon I \to \mathbb{R}^n,\ t \mapsto f(t) = (f_1(t), \ldots, f_n(t))$$

*eine **Kurve** im $\mathbb{R}^n$. Die Kurve heißt **differenzierbar** (kurz **diffbar**), wenn alle Komponenten f_k differenzierbar sind.*

Wie man hier schon sieht, werden wir Vektoren, wenn keine Verwechselungen möglich sind, häufig aus Platzgründen als Zeilenvektoren schreiben. Häufig werden wir den Begriff des Intervalls in der vorigen Definition ein wenig weiter fassen als in Definition 5.7 und als Intervallgrenzen auch ∞ und $-\infty$ zulassen, so dass die Kurvendefinition insbesondere auch Kurven einbezieht, die auf ganz $\mathbb{R}$ definiert sind.

Für eine Kurve $f\colon I \to \mathbb{R}^n$ hat die Menge $f(I)$ von Punkten im $\mathbb{R}^n$ oft einen Namen, wie z.B. Kreis, Gerade. Wir werden diesen Namen auch für die Abbildung f verwenden, wenn dies nicht zu Verwirrungen führt:

Beispiel 29.2. 1. Sei $r > 0$. Der Kreis mit Radius r (Abb. **??**) ist die Kurve

$$f\colon [0, 2\pi[\to \mathbb{R}^2, t \mapsto (r\cos t, r\sin t).$$

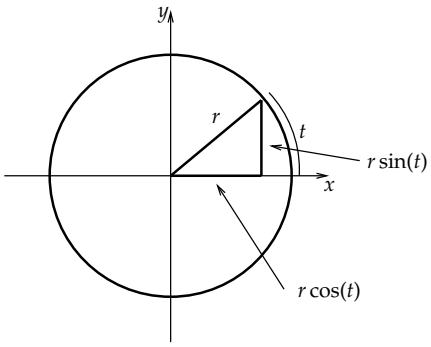


Abbildung 29.1. Die Parametrisierung eines Kreises mit Sinus und Cosinus.

2. Seien $a \in \mathbb{R}^n$ ein Vektor, $v \in \mathbb{R}^n \backslash \{0\}$ ein Richtungsvektor und

$$f\colon \mathbb{R} \to \mathbb{R}^n,\ f(t) = a + v \cdot t.$$

   Das Bild von f ist offenbar eine Gerade im $\mathbb{R}^n$.
3. Eine **Schraubenlinie** (Abb. 29.2): Seien $r > 0, c \neq 0 \in \mathbb{R}$ und

$$f(t) = (r\cos t, r\sin t, ct) \in \mathbb{R}^3.$$

4. Sei $\varphi\colon I \to \mathbb{R}$ eine stetige Funktion. Dann ist der Graph von φ eine Kurve im $\mathbb{R}^2$:

$$f(t) = (t, \varphi(t)) \in \mathbb{R}^2.$$

Wir interpretieren den Parameter t oft als Zeit und die Kurve als Bewegung eines Partikels im Raum. Daran angelehnt ist die folgende Definition.

Definition 29.3. *Seien I ein Intervall, $f\colon I \to \mathbb{R}^n$ eine diffbare Kurve. Dann heißt der Vektor*

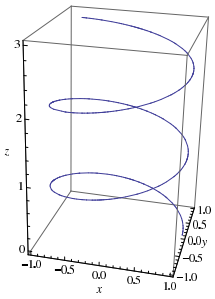


Abbildung 29.2. Eine Schraubenlinie.

$$f'(t) = (f_1'(t), \ldots, f_n'(t)),$$

dessen Komponenten die Ableitungen der Komponenten von f sind, der ***Geschwindigkeitsvektor zum Zeitpunkt*** *t. Seine Länge $\|f'(t)\|$ heißt* ***Geschwindigkeit*** *zum Zeitpunkt t. Ist $f'(t_0) \neq 0$ für ein t_0, so heißt $t\colon \mathbb{R} \to \mathbb{R}^n,\ f(t_0) + t \cdot f'(t_0)$ die* ***Tangente*** *an f in $f(t_0)$.*

Der Geschwindigkeitsvektor ist also:

$$f'(t) = \lim_{h \to 0} \frac{f(t+h) - f(t)}{h}.$$

Eine Kurve $f\colon I \to \mathbb{R}^n$ braucht nicht injektiv zu sein, wie die folgenden Beispiele zeigen:

Beispiel 29.4. 1. Der **Newtonsche Knoten**: $f(t) = (t^2 - 1, t^3 - t)$. Das Bild $f(\mathbb{R}) \subset \mathbb{R}^2$ (Abb. 29.3) kann auch durch eine Gleichung beschrieben werden: $f(\mathbb{R}) = \{(x, y) \mid y^2 = x^2 + x^3\}$. Wir geben hier keinen Beweis; allerdings ist die Inklusion $\subseteq$ mittels Nachrechnen leicht einzusehen: tatsächlich gilt nämlich $(t^3 - t)^2 = (t^2 - 1)^2 + (t^2 - 1)^3$.

2. Die **Neilsche Parabel**: $f\colon \mathbb{R} \to \mathbb{R}^2,\ f(t) = (t^2, t^3)$. Es gilt (auch dies ohne Beweis): $f(\mathbb{R}) = \{(x, y) \mid y^2 = x^3\}$ (Abb. 29.4).

Schneiden sich zwei Kurven in einem Punkt, so können wir mit Hilfe der Geschwindigkeitsvektoren wie in der linearen Algebra (Definition 17.8) einen Winkelbegriff einführen:

Definition 29.5. *Es seien $f\colon I \to \mathbb{R}^n$, $g\colon J \to \mathbb{R}^n$ zwei Kurven mit $f(t_1) = g(t_2)$. Sind die Geschwindigkeitsvektoren $f'(t_1), g'(t_2) \neq 0$, dann ist der* ***Winkel*** *θ* ***zwischen den Kurven*** *im Punkt $f(t_1) = g(t_2)$ definiert durch:*

$$\cos\theta = \frac{\langle f'(t_1), g'(t_2)\rangle}{\|f'(t_1)\| \cdot \|g'(t_2)\|}.$$

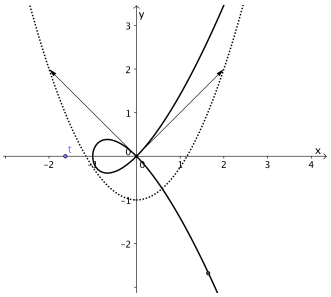


Abbildung 29.3. Der Newtonsche Knoten. Das Bild zeigt auch $f'(t)$ (gepunktet) sowie die beiden Vektoren $f'(1)$ und $f'(-1)$ im Doppelpunkt $(0,0) = f(1) = f(-1)$.

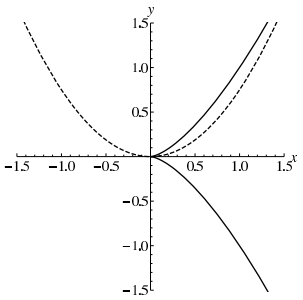


Abbildung 29.4. Die Neilsche Parabel. Das Bild zeigt auch $f'(t)$ (gepunktet).

Bemerkung 29.6. Diese Definition passt mit der Definition der Tangenten zusammen, denn der Winkel zwischen zwei Kurven ist gerade der Winkel zwischen den Richtungsvektoren der Tangenten der beiden Kurven im Schnittpunkt.

Beispiel 29.7. 1. Hat eine Kurve f **Selbstüberschneidungen**, d.h. $f(t_1) = f(t_2)$ für $t_1 \neq t_2$, so kann man die obige Definition des Winkels auch auf eine einzige Kurve (mit $g = f$) anwenden: Beim Newtonschen Knoten aus Beispiel 29.4 können wir für $t_1 = -1$ und $t_2 = 1$ den Winkel $\theta \in [0, \pi[$ zwischen den beiden Geschwindigkeitsvektoren im Ursprung bestimmen:

$$\cos\theta = \frac{\langle(-2,2),(2,2)\rangle}{\|(-2,2)\| \cdot \|(2,2)\|} = \frac{-4+4}{8} = \frac{0}{8} = 0,$$

d.h. $\theta = \frac{\pi}{2}$. Die beiden sogenannten **Kurvenzweige** stehen im Ursprung also senkrecht aufeinander.

2. Die **logarithmische Spirale** (Abb. 29.5

$$l\colon \mathbb{R} \to \mathbb{R}^2,\ l(\varphi) = (e^{\varphi} \cdot \cos\varphi, e^{\varphi} \cdot \sin\varphi)$$

bildet mit jeder Geraden durch den Ursprung an jedem ihrer Schnittpunkte den gleichen Winkel, nämlich 45°.

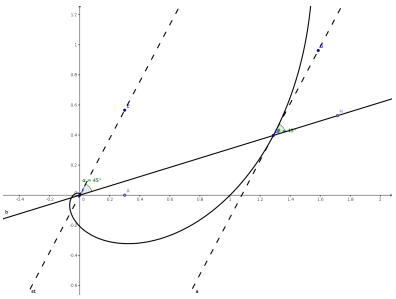


Abbildung 29.5. Eine logarithmische Spirale und deren Schnittwinkel mit einer Geraden durch den Ursprung.

Beweis. Übungsaufgabe. □

3. Alle Ellipsen zu zwei festen Brennpunkten (siehe dazu Bemerkung 25.13) bilden in all ihren Schnittpunkten mit Hyperbeln zu den selben Schnittpunkten jeweils rechte Winkel (siehe Abb. 29.6). Mit der obigen Definition ist dies etwas anstrengend nachzuweisen; wir verweisen auf [HCV32, S. 5] für einen wesentlich einfacheren Zugang zu diesem Spezialfall.

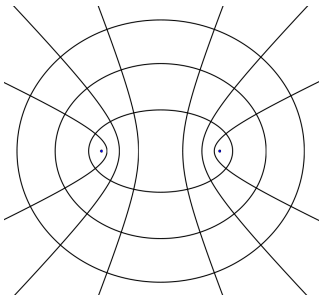


Abbildung 29.6. Ellipsen und Hyperbeln mit gemeinsamen Brennpunkten stehen in allen Schnittpunkten senkrecht aufeinander.

29.2 Rektifizierbarkeit und Bogenlänge

Wir kennen aus der Schule die Umrechnung eines Winkels in sein Bogenmaß. Dieses ist am Einheitskreis einfach die Länge des Kreisbogens, der zu dem gegebenen Winkel gehört.

Wir möchten hier nun einer Kurve bzw. einem Kurvenabschnitt (Bogen) sinnvoll eine Länge zuweisen. Insbesondere soll diese Länge im Fall von Strecken und dem eben angesprochenen Kreisbogen mit dem uns Bekannten zusammenpassen.

Hierzu betrachten wir, wie schon Archimedes vor mehr als 2000 Jahren, zunächst eine Approximation der Kurve durch einen stückweise linearen Linienzug:

Bemerkung 29.8 (Polygonapproximation). Seien $[a,b] \subset \mathbb{R}$ ein abgeschlossenes Intervall, $f\colon [a,b] \to \mathbb{R}^n$ eine Kurve und $a = t_0 < t_1 < \cdots < t_r = b$ eine Unterteilung. Dann können wir den Polygonzug durch die Punkte $f(t_0), f(t_1), \ldots, f(t_r)$ als Approximation der Kurve ansehen (Abb. 29.7). Die Länge des Polygonzuges ist:

$$P_f(t_0,\ldots,t_r) = \sum_{k=1}^{r} \|f(t_k) - f(t_{k-1})\|.$$

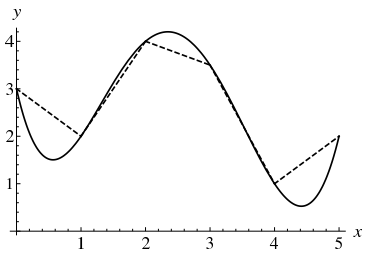


Abbildung 29.7. Polygonapproximation einer Kurve.

Damit können wir nun sinnvoll die Länge einer Kurve definieren:

Definition 29.9. *Eine Kurve $f\colon [a,b] \to \mathbb{R}^n$ heißt **rektifizierbar** mit **Bogenlänge** $L \in \mathbb{R}$, wenn $\forall \varepsilon > 0$ ein $\delta > 0$ existiert, so dass für jede Unterteilung $a = t_0 < t_1 < \cdots < t_r = b$ mit **Feinheit** $\leq \delta$ (d.h. $|t_i - t_{i+1}| \leq \delta \; \forall \; 0 \leq i < r$) gilt:*

$$|P_f(t_0,\ldots,t_r) - L| < \varepsilon.$$

Der Begriff Bogenlänge wird statt Kurvenlänge (wie man hätte vermuten können) verwendet, da sich die Länge eines Bogens auch definieren lässt, wenn die Kurve nicht nur auf einem abgeschlossenen Intervall $[a, b]$, sondern auf ganz $\mathbb{R}$ definiert ist. Den meisten Kurven, die auf einem solchen abgeschlossenen Intervall definiert sind, können wir eine Länge zuweisen, wie der folgende Satz zeigt:

Satz 29.10. *Jede stetig diffbare Kurve* $f : [a, b] \to \mathbb{R}^n$ *ist rektifizierbar mit Bogenlänge*

$$L = \int_a^b \|f'(t)\| \, dt.$$

Bevor wir dies nachweisen, zunächst ein Beispiel:

Beispiel 29.11. Wir betrachten $f : [a, b] \to \mathbb{R}^2$, $f(t) = (\cos t, \sin t)$. Das Bild von f ist bekanntlich ein Kreisbogen (Abb. 29.8).

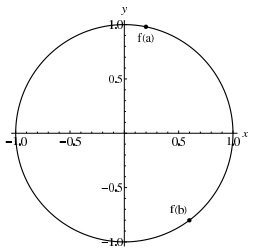


Abbildung 29.8. Zur Berechnung der Bogenlänge eines Kreises.

Die Bogenlänge von f ist:

$$\begin{aligned} L &= \int_a^b \|f'(t)\| \, dt = \int_a^b \|(-\sin t, \cos t)\| \, dt \\ &= \int_a^b \sqrt{\sin^2 t + \cos^2 t} \, dt = \int_a^b 1 \, dt = b - a. \end{aligned}$$

Dies passt mit der aus dem Schule bekannten Bogenmaß eines Winkels zusammen, denn für $a = 0$ und $b = 2\pi$ erhalten wir tatsächlich 2π und entsprechendes für andere Werte von b.

Analog zu obiger Rechnung ergibt sich, dass der Kreis mit Radius r,

$$g : [0, 2\pi] \to \mathbb{R}^2, \; g(t) = (r \cos t, r \sin t),$$

die Bogenlänge $\int_0^{2\pi} r \, dt = 2\pi r$ hat.

Hilfssatz 29.12. *Sei $f\colon [a,b] \to \mathbb{R}^n$ stetig diffbar. Dann gilt: $\forall\ \varepsilon > 0\ \exists\ \delta > 0$, so dass:*

$$\left\| \frac{f(t) - f(\tau)}{t - \tau} - f'(t) \right\| \leq \varepsilon$$

$\forall t, \tau \in [a,b]$ mit $0 < |t - \tau| < \delta$.

Beweis. Die Koordinatenfunktionen von f sind nach Voraussetzung stetig diffbar. Daher sind die $f_i'\colon [a,b] \to \mathbb{R}$ gleichmäßig stetig (siehe Definition 13.8 und Satz 13.10), also: $\forall\ \varepsilon > 0\ \exists\ \delta > 0$, so dass:

$$|f_i'(s) - f_i'(t)| < \varepsilon\ \forall t, s \text{ mit } |t - s| < \delta.$$

Der Mittelwertsatz 10.5 liefert:

$$\frac{f_i(t) - f_i(\tau)}{t - \tau} = f_i'(s)$$

für ein gewisses $s \in [\tau, t]$. Also:

$$\left| \frac{f_i(t) - f_i(\tau)}{t - \tau} - f_i'(t) \right| = |f_i'(s) - f_i'(t)| < \varepsilon.$$

Summation ergibt:

$$\left\| \frac{f(t) - f(\tau)}{t - \tau} - f'(t) \right\| \leq \sqrt{n} \cdot \max_i \left| \frac{f_i(t) - f_i(\tau)}{t - \tau} - f_i'(t) \right| < \sqrt{n} \cdot \varepsilon.$$

Wir hätten oben statt ε auch $\tilde{\varepsilon} = \frac{\varepsilon}{\sqrt{n}}$ wählen können. Damit folgt die Behauptung. □

Damit können wir nun den Satz über die Rektifizierbarkeit stetig diffbarer Kurven angehen:

Beweis (des Satzes 29.10). Sei $\varepsilon > 0$ vorgegeben. Aus der Approximation des Integrals durch Riemannsche Summen (siehe Definition 13.1 und Satz 13.7) wissen wir: $\exists\ \delta_1 > 0$, so dass:

$$\left| \int_a^b \|f'(t)\|\, dt - \sum_{k=1}^{r} \|f'(t_k)\| \cdot (t_k - t_{k-1}) \right| < \frac{\varepsilon}{2}$$

für alle Unterteilungen $a = t_0 < t_1 < \cdots < t_r = b$ mit Feinheit $\leq \delta_1$. Nach dem Hilfssatz 29.12 existiert ein δ mit $0 < \delta \leq \delta_1$ mit:

$$\left\| \frac{f(t_k) - f(t_{k-1})}{t_k - t_{k-1}} - f'(t_k) \right\| \leq \frac{\varepsilon}{2(b-a)}.$$

Dies ergibt:

$$\Big|\, \|f(t_k) - f(t_{k-1})\| - \|f'(t_k)(t_k - t_{k-1})\| \,\Big| \le \frac{\varepsilon}{2(b-a)}(t_k - t_{k-1}).$$

Summation über alle Teilintervalle liefert:

$$\Big|\sum_{k=1}^{r} \|f(t_k) - f(t_{k-1})\| - \sum_{k=1}^{r} \|f'(t_k)(t_k - t_{k-1})\|\Big| \le \frac{\varepsilon}{2(b-a)}(b-a) = \frac{\varepsilon}{2}.$$

Daraus folgt insgesamt mit der Dreiecksungleichung in $\mathbb{R}$:

$$\Big|P_f(t_0, \ldots, t_k) - \int_a^b \|f'(t)\|\, dt\Big| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

für alle Unterteilungen mit Feinheit $< \delta$. □

Korollar 29.13. *Jede stückweise stetig diffbare Kurve ist rektifizierbar.*

Beispiel 29.14. Die Zykloide ist die Kurve:

$$f\colon \mathbb{R} \to \mathbb{R}^2,\ f(t) = (t - \sin t, 1 - \cos t).$$

Sie beschreibt die Bewegung eines festen Punktes auf einem rollenden Rad mit Radius 1 (Abb. 29.9).

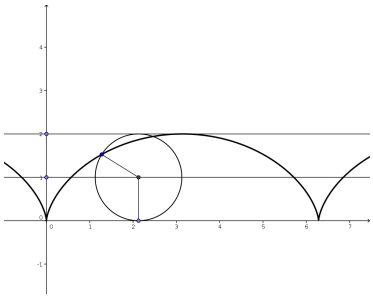


Abbildung 29.9. Die Zykloide.

Wir berechnen die Länge des Bogens der Kurve, die entsteht, wenn sich das Rad genau einmal dreht: $f'(t) = (1 - \cos t, \sin t)$, also

$$\|f'(t)\|^2 = (1 - \cos t)^2 + \sin^2 t = 2 - 2\cos t = 4\sin^2 \frac{t}{2}$$

mit Hilfe einer trigonometrischen Formel. Für die Bogenlänge L ergibt sich damit, da $\sin \frac{t}{2} \ge 0$ für $t \in [0, 2\pi]$:

$$L = \int_0^{2\pi} 2 \sin \frac{t}{2} \, dt = 4 \int_0^{\pi} \sin u \, du = 8$$

mit $u = \frac{t}{2}$, also $du = \frac{dt}{2}$. Bewegt sich ein Auto also um $2\pi \approx 6.28$ Meter, so bewegt sich ein Punkt auf dem Rand eines seiner Räder um 8 Meter.

Es stellt sich die Frage: Ist jede Kurve rektifizierbar? Die Antwort ist nein, wie das folgende Beispiel zeigt:

Beispiel 29.15. Die Kurve

$$\gamma \colon [0,1] \to \mathbb{R}^2, t \mapsto \begin{cases} (0,0), & t = 0, \\ (t, t \cdot \cos(\frac{1}{t}), & t \neq 0, \end{cases}$$

ist stetig, aber nicht rektifizierbar. Formal möchten wir das hier nicht beweisen, doch was sollte die Länge L sein? Jedenfalls gilt (siehe Abb. 29.10):

$$L \geq 2 \cdot \frac{1}{2\pi} + 2 \cdot \frac{1}{3\pi} + 2 \cdot \frac{1}{4\pi} + 2 \cdot \frac{1}{5\pi} + \cdots .$$

Aber der Grenzwert $\sum_{i=1}^{\infty} \frac{1}{i}$ existiert nicht.

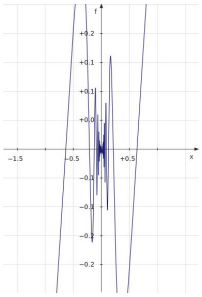


Abbildung 29.10. Eine nicht rektifizierbare Kurve.

Es gibt sogar Kurven, wie die im folgenden Beispiel, die keine Kurven im anschaulichen Sinn sind, die nämlich als Bild im $\mathbb{R}^2$ ein Flächenstück haben.

Beispiel 29.16 (Peano-Kurve). Wir konstruieren eine surjektive stetige Abbildung vom Intervall $[-1,1] \subset \mathbb{R}$ auf das Dreieck im $\mathbb{R}^2$ mit Ecken $(-1,0)$, $(1,0)$, $(0,1)$ vermöge Intervallschachtelung (Abb. 29.11).

Für jede reelle Zahl $r \in [-1,1]$ haben wir dann eine bzw. möglicherweise zwei Intervallschachtelungen und zu diesen eine bzw. zwei Schachtelungen von

Abbildung 29.11. Zur Definition der Peano-Kurve.

Dreiecken. Der Punkt im Durchschnitt der Schachtelungen der Dreiecke sei der Bildpunkt $\varphi(r)$. Eine bemerkenswerte Eigenschaft der Abbildung φ ist: φ ist stetig und surjektiv. Dies zeigen wir hier nicht; in einer Übungsaufgabe werden wir aber nachweisen, dass solche Kurven (genannt **Peanokurven**) nicht rektifizierbar sind.

Definition 29.17. *Sei $f\colon [a,b] \to \mathbb{R}^n$ eine (stetig diffbare) Kurve und $\varphi\colon [\alpha,\beta] \to [a,b]$ eine monoton steigende bijektive (stetig diffbare) Abbildung. Wir sagen, die Kurve $g = f \circ \varphi\colon [\alpha,\beta] \to \mathbb{R}^n$ geht **aus** f **durch Parameterwechsel hervor**.*

Wir interessieren uns für Eigenschaften von Kurven, die nicht von der Parametrisierung abhängen. Beispielsweise gilt (sonst wäre die Definition der Bogenlänge auch nicht wirklich sinnvoll):

Satz 29.18. *Die Bogenlänge einer stetig diffbaren Kurve hängt nicht von der Parametrisierung ab.*

Beweis. Die Bogenlänge ist: $L = \int_a^b \|f'(t)\|\,dt$. Sei $t = \varphi(u)$ ein stetig diffbarer Parameterwechsel. Es gilt:

$$\begin{aligned}
\int_\alpha^\beta \|(f\circ\varphi)'(u)\|\,du &\overset{\text{Kettenregel}}{=} \int_\alpha^\beta \|f'(\varphi(u))\cdot\varphi'(u)\|\,du \\
&\overset{\varphi'(u)\geq 0}{=} \int_\alpha^\beta \|f'(\varphi(u))\|\cdot\varphi'(u)\,du \\
&\overset{\text{Substitutionsregel}}{=} \int_{\varphi(\alpha)}^{\varphi(\beta)} \|f'(t)\|\,dt \\
&= \int_a^b \|f'(t)\|\,dt.
\end{aligned}$$

□

Definition 29.19 (Parametrisierung nach Bogenlänge). *Sei $f\colon [a,b] \to \mathbb{R}^n$ stetig diffbar mit $f'(t) \neq 0\ \forall\, t \in [a,b]$. Dann ist die Funktion*

$$\varphi(t_1) = \int_a^{t_1} \|f'(t)\|\,dt$$

streng monoton und diffbar mit $\varphi'(t) \neq 0\ \forall t \in]a,b[$ und definiert eine Bijektion $\varphi\colon [a,b] \to [0,L]$, wobei L die Bogenlänge von f beschreibt. Die Umkehrabbildung

*$\psi := \varphi^{-1}\colon [0,L] \to [a,b]$ liefert die Parametrisierung $g = f \circ \psi\colon [0,L] \to \mathbb{R}^n$. Diese heißt **Parametrisierung nach Bogenlänge**. Es ist üblich, den Parameter in dieser Parametrisierung mit s zu bezeichnen: $s \mapsto g(s)$.*

Bemerkung 29.20. Ist $g\colon [0,L] \to \mathbb{R}^n$ eine Parametrisierung einer Kurve $f\colon [a,b] \to \mathbb{R}^n$ nach Bogenlänge, so hat für $[s_1,s_2] \subseteq [0,L]$ die Teilkurve $g\big|_{[s_1,s_2]}\colon [s_1,s_2] \to \mathbb{R}^n$ die Bogenlänge $s_2 - s_1$.

Beweis. Seien $\varphi\colon [a,b] \to [0,L]$ und $\psi = \varphi^{-1}$ wie in der Definition. Ferner seien $t_1 = \psi(s_1)$ und $t_2 = \psi(s_2)$. Dann ist auch

$$\tilde{\varphi}\colon [t_1,t_2] \to [0,s_2-s_1],\ t \mapsto \varphi(t) - s_1$$

eine Bijektion mit Umkehrabbildung $\tilde{\psi}$ und $g\big|_{[0,,s_2-s_1]} = f \circ \tilde{\psi}\colon [0,s_2-s_1] \to \mathbb{R}^n$ die Parametrisierung von $f\big|_{[t_1,t_2]}$ nach Bogenlänge. Diese ist daher $s_2 - s_1$. □

Beispiel 29.21. Sei $f\colon [0,2\pi] \to \mathbb{R}^2, t \mapsto f(t) = (r\cos t, r\sin t)$ ein parametrisierter Kreis mit Radius r. Dann ist $\|f'(t)\| = r \cdot \sqrt{\sin^2 t + \cos^2 t} = r$ und:

$$\int_0^{2\pi} \|f'(t)\|\,dt = 2\pi r.$$

Die Bijektion $\varphi\colon [0,2\pi] \to [0,2\pi r], t \mapsto t\cdot r$ hat die Umkehrfunktion $\psi(s) = \frac{1}{r}\cdot s$. Die Parametrisierung nach Bogenlänge ist also:

$$g\colon [0,2\pi r] \to \mathbb{R}^2,\ s \mapsto (f\circ\psi)(s) = f\Big(\frac{s}{r}\Big) = \Big(r\cos\frac{s}{r}, r\sin\frac{s}{r}\Big).$$

Satz 29.22. *Sei $g\colon [0,L] \to \mathbb{R}^n$ eine Parametrisierung nach Bogenlänge (d.h. insbesondere $g'(s) \neq 0\ \forall\, s$). Dann hat der Geschwindigkeitsvektor $T(s) = g'(s)$ die Länge $\|T(s)\| = 1$.*

Beweis. Wegen der Bemerkung 29.20 ist:

$$\int_{s_1}^{s_2} \|g'(s)\|\,ds = s_2 - s_1\ \forall s_1,s_2 \in [0,L].$$

Nach dem Mittelwertsatz der Integralrechnung 13.12 folgt: Es existiert ein $\xi \in [s_1,s_2]$, so dass

$$\int_{s_1}^{s_2} \|g'(s)\|\,ds = \|g'(\xi)\| \cdot (s_2 - s_1),$$

also $\|g'(\xi)\| = \frac{s_2-s_1}{s_2-s_1} = 1$, weil die linke Seite ja, wie gerade gesagt, $s_2 - s_1$ ist. Da dies aber für alle $s_1,s_2 \in [0,L]$ gilt, folgt: $\|g'(\xi)\| = 1\ \forall\ \xi$. □

29.3 Krümmung

Definition 29.23. *Sei $g\colon [0,L] \to \mathbb{R}^n$ eine zweimal stetig diffbare Kurve, parametrisiert nach Bogenlänge. $T(s) = g'(s)$ heißt **Tangentialvektor** der Kurve. $\kappa = \kappa(s) = \|T'(s)\|$ (kappa) heißt **Krümmung** der Kurve im Punkt $g(s)$. $N(s) = \frac{T'(s)}{\kappa(s)}$ heißt **Normalenvektor** (definiert, wenn $\kappa(s) \neq 0$). Also: $T'(s) = \kappa(s) \cdot N(s)$.*

Bemerkung 29.24. Tatsächlich steht der Normalenvektor N, wenn er definiert ist, senkrecht (auch normal genannt) auf dem Tangentialvektor T. Nach Definition der Parametrisierung nach Bogenlänge gilt nämlich $1 = \langle T, T\rangle$, d.h. diese Funktion ist konstant, so dass ihre Ableitung verschwindet. Nach der Produktregel ergibt sich daher: $0 = (\langle T, T\rangle)' = \langle T', T\rangle + \langle T, T'\rangle = 2\kappa\langle T, N\rangle$, d.h. $N \perp T$.

Beispiel 29.25. 1. Ein Kreis mit Radius r: $s \mapsto (r\cos\frac{s}{r}, r\sin\frac{s}{r}) = g(s)$. Dann ist: $T(s) = g'(s) = (-\sin\frac{s}{r}, \cos\frac{s}{r})$ und somit $T'(s) = \frac{1}{r}(-\cos\frac{s}{r}, -\sin\frac{s}{r})$, also $\kappa = \frac{1}{r}$ und $N = (-\cos\frac{s}{r}, -\sin\frac{s}{r})$ (Abb. 29.12).

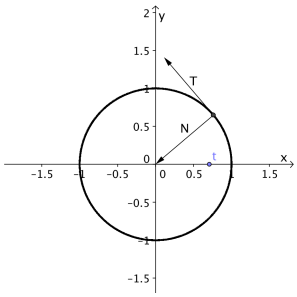


Abbildung 29.12. Normalen- und Tangentialvektor am Kreis.

2. Sei f eine Kurve mit der Eigenschaft $\kappa \equiv 0$ und $T' \equiv 0$. Dann gilt: Die Kurve ist eine Gerade.

Bemerkung/Definition 29.26. Im Fall von ebenen Kurven kann man κ mit einem Vorzeichen versehen: Wir wählen $N(s)$, so dass $(T(s), N(s)) \in \mathrm{SO}(2)$. Um jetzt die Gleichung $T'(s) = \kappa(s) \cdot N(s)$ immer noch zu erfüllen, muss $\kappa(s)$ jetzt ggf. ein Vorzeichen bekommen. Ist dieses positiv, so heißt die Kurve in diesem Punkt **positiv gekrümmt**; ist es negativ, so heißt sie in diesem Punkt **negativ gekrümmt**. Für ebene Kurven ist der Kreis mit Mittelpunkt $g(s) + \frac{1}{\kappa}N(s)$ und Radius $r = \frac{1}{\kappa}$ der Kreis, der die Kurve in $g(s)$ am Besten approximiert. Er heißt **Krümmungskreis**; siehe Aufgabe 29.5 für ein Beispiel.

Mit Hilfe der Krümmung kann man folgende interessante Kurve definieren:

Definition 29.27. *Ist f eine Kurve mit Krümmung $\kappa(s)$, so heißt*

$$s \mapsto \big(\kappa(s), \kappa'(s)\big)$$

*die **charakteristische Kurve** von f.*

Satz 29.28 (von Cartan, hier ohne Beweis). *Zwei Kurven g und $\tilde{g}$ in $\mathbb{R}^2$ gehen genau dann durch eine euklidische Bewegung auseinander hervor, wenn ihre charakteristischen Kurven übereinstimmen.*

Ähnliche Sätze charakterisieren, ob Kurven durch sogenannte affine oder projektive Transformationen auseinander hervorgehen. Charakteristische Kurven finden in der Bilderkennung Anwendung (siehe beispielsweise [COS+98]).

29.4 Kurven im $\mathbb{R}^3$

Definition 29.29. *Sei $g\colon [0, L] \to \mathbb{R}^3$ eine Kurve, die nach Bogenlänge parametrisiert ist. Dann ist $N \perp T$. Wir wählen nun $B \in \mathbb{R}^3$, so dass (T, N, B) eine orientierte Orthonormalmatrix ($\in \mathrm{SO}(3)$) bilden (Abb. 29.13). B heißt **Binormalenvektor** und das Tripel (T, N, B) **Fresnelsches Dreibein**.*

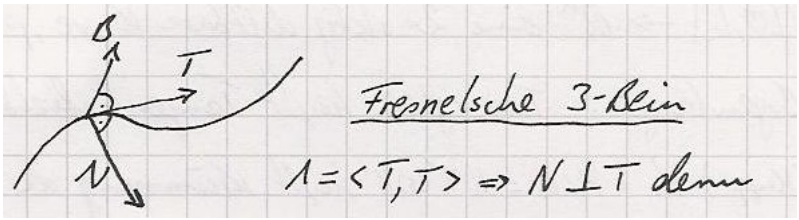


Abbildung 29.13. Das Fresnelsche Dreibein einer Kurve im $\mathbb{R}^3$.

Um einen Binormalenvektor zu berechnen, ist häufig das Folgende hilfreich:

Definition 29.30 (Kreuzprodukt im $\mathbb{R}^3$). *Seien $a = (a_1, a_2, a_3)^{\mathrm{t}}$ und $b = (b_1, b_2, b_3)^{\mathrm{t}} \in \mathbb{R}^3$ zwei Vektoren. Dann ist das **Kreuzprodukt** der Vektor*

$$a \times b = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix}.$$

Die folgenden Eigenschaften sind nicht schwer nachzuprüfen, so dass wir uns auf den Nachweis einer einzigen beschränken:

Elementare Eigenschaften 29.31. *1. $a \times b$ steht senkrecht auf a und b.*

2. *$\|a \times b\|$ ist die Größe (d.h. der Flächeninhalt) des Parallelogrammes, das von a und b aufgespannt wird.*
3. *Es gilt:* $\det(a, b, a \times b) = \|a \times b\|^2$.
4. *Es gelten die folgenden Rechenregeln:*
   - $a \times (b + c) = a \times b + a \times c$,
   - $(a + b) \times c = a \times c + b \times c$,
   - $(\lambda a) \times b = \lambda(a \times b) = a \times (\lambda b)$,
   - $b \times a = -a \times b$.

Beweis. 1. Wir betrachten die Determinante:

$$\begin{aligned} 0 &= \det\begin{pmatrix} a_1 & a_1 & b_1 \\ a_2 & a_2 & b_2 \\ a_3 & a_3 & b_3 \end{pmatrix} \\ &\overset{\text{Entw. nach 1. Spalte}}{=} a_1(a_2b_3 - a_3b_2) - a_2(a_1b_3 - b_1a_3) + a_3(a_1b_2 - a_2b_1) \\ &= \langle a, a \times b\rangle. \end{aligned}$$

Also: $a \perp (a \times b)$ (und b genauso).

□

Die in der folgenden Beziehung zwischen Binormalenvektor und Normalenvektor auftretende Torsion misst, wie weit die Kurve von einer ebenen Kurve entfernt ist:

Proposition/Definition 29.32. *$B'(s) = \tau(s)N(s)$ für eine gewisse Funktion $\tau(s)$. $\tau(s)$ heißt **Torsion** der Kurve im Punkt $g(s)$.*

Beweis. Es gilt: $1 = \langle B, B\rangle$. Dies liefert: $0 = (\langle B, B\rangle)' = \langle B', B\rangle + \langle B, B'\rangle = 2\langle B', B\rangle \Rightarrow B' \perp B$. Da T, N, B eine Orthonormalbasis bilden, folgt: $B' = \alpha(s)T + \tau(s)N$ für gewisse α, τ. Wir müssen noch zeigen, dass gilt: $\alpha \equiv 0$. Dafür bemerken wir zunächst: $0 = \langle T, B\rangle = \langle T, T \times N\rangle$. Dies liefert:

$$\begin{aligned} 0 = (\langle T, B\rangle)' &= \langle T', B\rangle + \langle T, B'\rangle \\ &= \underbrace{\langle \kappa N, B\rangle}_{=0} + \langle T, \alpha T + \tau N\rangle = \alpha\langle T, T\rangle = \alpha, \end{aligned}$$

da auch $\langle T, N\rangle = 0$. □

Aufgaben

Aufgabe 29.1 (Der Newtonsche Knoten). Wir betrachten die Menge $M := \{(x, y) \in \mathbb{R}^2 \mid y^2 = x^2 + x^3\} \subset \mathbb{R}^2$. Zeigen Sie, dass die Kurve $f\colon \mathbb{R} \to \mathbb{R}^2,\ t \mapsto (t^2 - 1, t^3 - t)$ surjektiv auf M abbildet.
Hinweis: Betrachten Sie Geraden durch den sog. Doppelpunkt $(0, 0)$.

Aufgabe 29.2 (Winkel zwischen Kurven). Wir betrachten eine sogenannte logarithmische Spirale

$$l\colon \mathbb{R} \to \mathbb{R}^2,\ l(\varphi) = (e^{\varphi} \cdot \cos\varphi, e^{\varphi} \cdot \sin\varphi)$$

und für jeden festen Winkel φ die Geraden

$$g_{\varphi}\colon \mathbb{R} \to \mathbb{R}^2,\ g(t) = \begin{cases} (0, t), & \text{falls } \varphi = \frac{\pi}{2} + n\pi \text{ für ein } n \in \mathbb{Z}, \\ (t, \tan\varphi \cdot t), & \text{sonst.} \end{cases}$$

Zeigen Sie:

1. Für jedes $\varphi \in \mathbb{R}$ liegt der Punkt $l(\varphi) \in \mathbb{R}^2$ auf der Geraden g_{φ}.
2. Der Winkel $\alpha_{\varphi} \in [0, \pi[$ zwischen den Kurven l und g_{φ} im Punkt $l(\varphi)$ ist unabhängig von φ, nämlich $\alpha_{\varphi} = \frac{\pi}{4}$ für alle $\varphi \in \mathbb{R}$.

Aufgabe 29.3 (Bogenlänge). Seien $r, c \in \mathbb{R}, r > 0$ und $R \in \mathbb{R}, R > 0$. Seien ferner

$$f\colon [0, 2\pi] \to \mathbb{R}^3,\ t \mapsto (r\cos t, r\sin t, ct)$$

und

$$g\colon [0, 2\pi] \to \mathbb{R}^2,\ t \mapsto R \cdot (t - \sin t, 1 - \cos t)$$

gegeben.

1. Berechnen Sie die Bogenlängen von f und g.
2. Parametrisieren Sie f und g nach Bogenlänge.

Aufgabe 29.4 (Krümmung eines Graphen einer Funktion). Wir betrachten den Spezialfall einer ebenen Kurve, die in der Form

$$f : [a, b] \to \mathbb{R}^2,\ x \mapsto (x, y(x))$$

geschrieben werden kann (also einen Graphen einer Funktion).

Zeigen Sie, dass für die Krümmung dann gilt:

$$\kappa(x) = \frac{y''(x)}{(1 + (y'(x))^2)^{3/2}}.$$

Aufgabe 29.5 (Krümmungskreis). Berechnen Sie mit Hilfe des Computeralgebrasystems MAPLE (oder eines anderen) für die ebene Kurve $(x, x^3) \subset \mathbb{R}^2$ die Krümmungen in $x = 0, -1, 1, -\frac{1}{2}, \frac{1}{2}, \frac{1}{24}, -\frac{1}{24}$ und plotten Sie die Kurve und die Krümmungskreise.

Aufgabe 29.6 (Krümmung einer ebenen Kurve). Wir betrachten den Spezialfall einer ebenen Kurve, die parametrisiert gegeben ist:

$$f : [a, b] \to \mathbb{R}^2,\ t \mapsto (x(t), y(t)),$$

wobei x und y jeweils zweimal stetig differenzierbar sind.

Zeigen Sie, dass für die Krümmung dann gilt:

$$\kappa(x) = \frac{x'(t)y''(t) - x''(t)y'(t)}{(x'(t)^2 + y'(t)^2)^{\frac{3}{2}}}.$$

Aufgabe 29.7 (Die charakteristische Kurve einer ebenen Kurve). Wir definieren die *charakteristische Kurve* einer nach Bogenlänge parametrisierten ebenen Kurve $[a, b] \to \mathbb{R}^2,\ s \mapsto (x(s), y(s))$ als die ebene Kurve

$$[a, b] \to \mathbb{R}^2,\ s \mapsto \left(\kappa(s), \tfrac{d\kappa}{ds}(s)\right).$$

1. Sei $r \in \mathbb{R}$. Bestimmen Sie die charakteristische Kurve der nach Bogenlänge parametrisierten Kurve

$$g\colon [0, 2\pi r] \to \mathbb{R}^2,\ t \mapsto (r\cos(\tfrac{t}{r}), r\sin(\tfrac{t}{r})).$$

2. Bestimmen Sie die charakteristische Kurve von

$$h : [0, 2\pi] \to \mathbb{R}^2,\ s \mapsto (\tfrac{1}{2}\sin t, \cos t).$$

30

Funktionen auf $\mathbb{R}^n$

Sei $D \subset \mathbb{R}^n$, $f\colon D \to \mathbb{R}$ eine Funktion. Wie können wir uns f veranschaulichen? Wir stellen hier zwei Möglichkeiten vor: den Graph von f und Niveaumengen.

30.1 Erste Definitionen und Beispiele

Definition 30.1. *Wir nennen*

$$\Gamma_f := \{(x, y) \in D \times \mathbb{R} \mid f(x) = y\}$$

den ***Graph*** *von f.*

Beispiel 30.2. 1. $f(x_1, x_2) = x_2^3 - x_1 x_2$ (s. Abb. 30.1, links).

2. $f(x_1, x_2) = x_1^2 + x_2^2$ (s. Abb. 30.1, rechts).

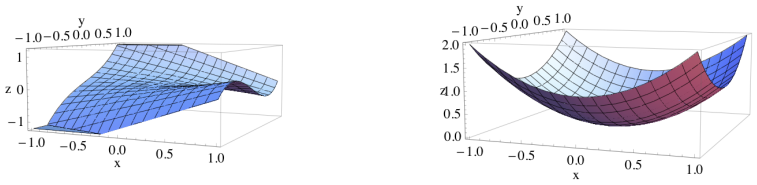


Abbildung 30.1. Zwei Graphen von Funktionen.

Definition 30.3. *Sei $f\colon D \to \mathbb{R}$, $c \in \mathbb{R}$. Wir nennen*

$$N_c(f) := \{x \in D \mid f(x) = c\}$$

*die **Niveaumenge** von f zum **Niveau** c. Falls $n = 2$ (also $D \subset \mathbb{R}^2$), so heißt diese Niveaumenge auch **Niveaulinie**, falls $n = 3$ **Niveaufläche**.*

Beispiel 30.4. 1. $f(x_1, x_2) = x_1^2 - x_2^2 - x_1^4$ (s. Abb. 30.2).

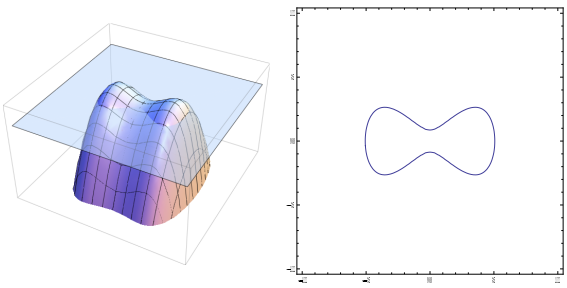


Abbildung 30.2. Niveaulinien einer Funktion.

2. $f(x_1, x_2) = x_1^2 + x_2^2$ (s. Abb. 30.3).

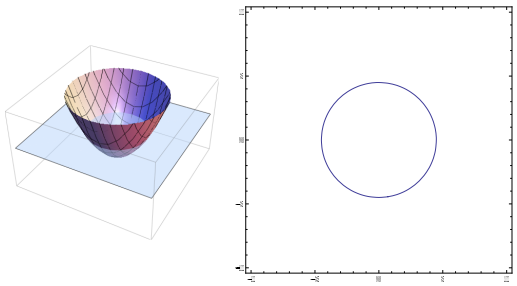


Abbildung 30.3. Niveaulinien einer Funktion.

3. Anschaulich kann man Niveaulinien recht gut verstehen, wenn man das Beispiel $f(x, y) = x^2 - y^2 + x^3$ betrachtet und sich vorstellt, dass mit ihrer Hilfe die Höhe eines Berges über dem Punkt (x, y) der Ebene näherungsweise beschrieben wird (Abb. 30.4). Die Niveaulinien $f(x, y) = c$ sind dann die Linien, die wie die Höhenlinien in Landkarten die Wege angeben, auf denen man weder bergauf noch bergab gehen muss. Oft ist das Zeichnen der Niveauflächen oder -linien ohne Computer nicht einfach.

Bemerkung 30.5. Der Definitionsbereich D von f soll möglichst einfach, z.B. „echt n-dimensional" sein.

Abbildung 30.4. Niveaulinien an einem abstrahierten Berg

30.2 Offene und abgeschlossene Mengen

Definition 30.6. *Zu $a \in \mathbb{R}^n$ und $r > 0$ heißt*

$$B_r(a) := \{x \in \mathbb{R}^n \mid \|x - a\| < r\}$$

*der **offene Ball** mit Radius r.*

$$\overline{B_r(a)} := \{x \in \mathbb{R}^n \mid \|x - a\| \leq r\}$$

*heißt **abgeschlossener Ball** um a mit Radius r.*

Definition 30.7. *Sei $U \subset \mathbb{R}^n$ eine Teilmenge. U heißt **offen**, wenn $\forall\, a \in U \quad \exists\, \varepsilon > 0$, so dass $B_\varepsilon(a) \subset U$. Sei $A \subset \mathbb{R}^n$ eine weitere Teilmenge. A heißt **abgeschlossen**, wenn $\mathbb{R}^n \backslash A$ offen ist.*

Beispiel 30.8. $B_r(a)$ ist offen, $\overline{B_r(a)}$ ist abgeschlossen.

Definition 30.9. *$D \subset \mathbb{R}^n$ heißt **beschränkt**, wenn es ein $r > 0$ gibt, so dass gilt: $D \subset B_r(0)$. Eine abgeschlossene und beschränkte Teilmenge $K \subset \mathbb{R}^n$ heißt **kompakt**.*

Definition 30.10. *Zu einer beliebigen Teilmenge $D \subset \mathbb{R}^n$ bezeichnet*

$$\overset{\circ}{D} := \{x \in \mathbb{R}^n \mid \exists\, \varepsilon > 0\ B_\varepsilon(x) \subset D\}$$

*die Menge der **inneren Punkte** (oder kurz: das **Innere**) von D.*

$$\overline{D} := \{x \in \mathbb{R}^n \mid B_\varepsilon(x) \cap D \neq \emptyset\ \forall\, \varepsilon > 0\}$$

*heißt **Abschluss** von D und*

$$\partial D := \overline{D} \backslash \overset{\circ}{D}$$

***Rand** von D.*

Beispiel 30.11. $\overline{B_r(a)}$ ist der Abschluss von $B_r(a)$ und $\partial B_r(a) = \{x \in \mathbb{R}^n \mid \|x-a\| = r\}$ ist die Kugeloberfläche.

In der Regel werden wir offene Mengen als Definitionsbereich nehmen, eventuell auch Mengen, die sowohl offen als auch abgeschlossen sind. Analog zum univariaten Fall (Def. 8.3) können wir den Begriff einer stetigen Funktion einführen:

Definition 30.12. *Sei $D \subset \mathbb{R}^n$, $f\colon D \to \mathbb{R}$ eine Funktion. f heißt **stetig** in $a \in D \subset \mathbb{R}^n$, wenn*

$$\forall\, \varepsilon > 0 \quad \exists\, \delta > 0\colon \quad |f(x) - f(a)| < \varepsilon \;\forall\, x \in D \text{ mit } \|x - a\| < \delta.$$

Satz 30.13. *Summen, Produkte und Quotienten (wo sie definiert sind) stetiger Funktionen sind stetig.*

Wir geben keinen Beweis, weil dieser analog zum Fall einer Veränderlichen ist. Wenigstens ein kleines Beispiel dazu:

Beispiel 30.14. $f(x_1, x_2) = \frac{x_1^2 + x_2^4}{x_1^2 + x_2^2 + 1}$ ist stetig.

30.3 Differentiation

Sei $f\colon U \to \mathbb{R}$ eine Funktion, wobei U offen ist. Wir stellen nun zwei Konzepte vor, die die Differentiation in einer Variablen auf höhere Dimensionen verallgemeinern: partielle Differentiation und totale Differentiation.

30.3.1 Partielle Differentiation

Definition 30.15. *Seien $f\colon U \to \mathbb{R}$, $a = (a_1, \ldots, a_n) \in U$. Dann heißt f in a **partiell nach x_i differenzierbar** (kurz **partiell nach x_i diffbar**), wenn die Funktion in einer Variablen*

$$x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n)$$

nach x_i differenzierbar ist. Dann bezeichnet

$$\frac{\partial f}{\partial x_i}(a) := \lim_{h \to 0} \frac{f(a_1, \ldots, a_i + h, \ldots, a_n) - f(a_1, \ldots, a_i, \ldots, a_n)}{h}$$

*die **partielle Ableitung** von f nach x_i.*

Beispiel 30.16. $f(x_1, x_2) = x_1^3 - x_1 x_2$. Dann gilt: $\frac{\partial f}{\partial x_1} = 3x_1^2 - x_2$, $\frac{\partial f}{\partial x_2} = -x_1$.

Definition 30.17. *Ist $f\colon U \to \mathbb{R}$, U offen, in jedem Punkt nach allen Variablen partiell diffbar, dann heißt f **partiell differenzierbar auf** U (kurz: **partiell diffbar auf** U).*

Der Vektor

$$\nabla f(a) := (\operatorname{grad} f)(a) := \left(\frac{\partial f}{\partial x_1}(a), \dots, \frac{\partial f}{\partial x_n}(a)\right)$$

*heißt **Gradient** von f im Punkt $a \in U$.*

Definition 30.18 (höhere partielle Ableitungen). *Sei $f\colon U \to \mathbb{R}$ in U nach x_i partiell diffbar und $\frac{\partial f}{\partial x_i}\colon U \to \mathbb{R}$ partiell nach x_j diffbar, dann bezeichnet*

$$\frac{\partial^2 f}{\partial x_j\, \partial x_i}\colon U \to \mathbb{R}$$

*die **j-te partielle Ableitung** von $\frac{\partial f}{\partial x_i}$.*

Das folgende Beispiel zeigt, dass wir im Allgemeinen die Reihenfolge der Ableitungen hierbei nicht vertauschen dürfen:

Beispiel 30.19. Sei

$$f(x_1, x_2) = \begin{cases} x_1 x_2 \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}, & \text{falls } (x_1, x_2) \neq (0,0), \\ 0, & \text{falls } (x_1, x_2) = 0. \end{cases}$$

Im Nullpunkt gilt dann: $\frac{\partial f}{\partial x_1}(0,0) = 0$, $\frac{\partial f}{\partial x_2}(0,0) = 0$, da $f(x_1, 0) \equiv 0$, $f(0, x_2) \equiv 0$. Ferner ist:

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \frac{(x_1^2 + x_2^2)\big(x_2(x_1^2 - x_2^2) + x_1 x_2 2x_1\big) - 2x_1(x_1x_2)(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2} \\ &= \frac{x_2(x_1^4 - x_2^4) + 4x_1^2 x_2^3}{(x_1^2 + x_2^2)^2}. \end{aligned}$$

$$\Rightarrow \frac{\partial^2 f}{\partial x_2\, \partial x_1}(0,0) = \frac{\partial}{\partial x_2}\Big(\frac{\partial f}{\partial x_1}(0, x_2)\Big) = \frac{\partial}{\partial x_2}\Big(\frac{-x_2^5}{x_2^4}\Big) = \frac{\partial}{\partial x_2}(-x_2) = -1.$$

Andererseits: $\frac{\partial^2 f}{\partial x_1\, \partial x_2}(0,0) = 1$ wegen der Symmetrie von f: $f(x_1, x_2) = -f(x_1, x_2)$. Im Allgemeinen gilt also:

$$\frac{\partial^2 f}{\partial x_i\, \partial x_j} \neq \frac{\partial^2 f}{\partial x_j\, \partial x_i}.$$

Glücklicherweise gibt es aber viele Situationen, in denen das Vertauschen der Reihenfolge doch richtig ist:

Satz 30.20. *Sei $f\colon U \to \mathbb{R}$ zweimal stetig partiell diffbar, dann gilt:*

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Beweis. Siehe z.B. Forsters Analysis 2 Buch [For08b]. □

Korollar/Definition 30.21. *Unter der Voraussetzung dieses Satzes ist also die **Hesse-Matrix***

$$\operatorname{Hess}(f) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{ij} := \begin{pmatrix} \frac{\partial^2 f}{(\partial x_1)^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 f}{(\partial x_n)^2} \end{pmatrix}$$

symmetrisch.

Beispiel 30.22. Sei $f(x_1, x_2) = x_1 \sin(x_2 + x_1)$. Dann ist der Gradient

$$\operatorname{grad} f = \Big(\sin(x_1 + x_2) + x_1 \cos(x_1 + x_2), \quad x_1 \cos(x_1 + x_2)\Big)$$

und die Hesse-Matrix:

$$\begin{pmatrix} 2\cos(x_1 + x_2) - x_1 \sin(x_1 + x_2) & \cos(x_1 + x_2) - x_1 \sin(x_1 + x_2) \\ \cos(x_1 + x_2) - x_1 \sin(x_1 - x_2) & -x_1 \sin(x_1 + x_2) \end{pmatrix}.$$

30.3.2 Totale Differentiation

Das zweite Konzept, das die Ableitung in einer Variablen verallgemeinert, beruht auf der Grundidee, die Ableitung als beste lineare Approximation aufzufassen (s. Abb. 30.5). Wir verallgemeinern dieses Konzept gleich auf Abbildungen

$$f\colon U \to \mathbb{R}^n,\ U \subseteq \mathbb{R}^m \text{ offen.}$$

Definition 30.23. *$f\colon U \to \mathbb{R}^n$, $f = (f_1, \ldots, f_n)$ heißt in $a \in U \subseteq \mathbb{R}^m$ **total differenzierbar** (kurz: **totel diffbar**), wenn es eine lineare Abbildung $x \mapsto A{\cdot}x$, $A \in \mathbb{R}^{n\times m}$, gibt, so dass für den Fehler φ, definiert durch*

$$f(a + x) = f(a) + Ax + \varphi(x)$$

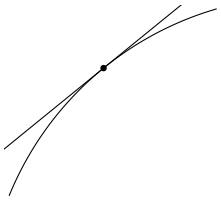


Abbildung 30.5. Ableitung als beste lineare Approximation.

die Bedingung

$$\lim_{x\to 0} \frac{\varphi(x)}{\|x\|} = 0$$

erfüllt ist. $A =: Df(a) =: J_f(a)$ *heißt dann die* ***Jacobimatrix*** *(oder das* ***Differential****) von* f *in* a*.*

Dass die Jacobimatrix tatsächlich eindeutig ist, ist nicht schwierig nachzuweisen. Im Eindimensionalen ist A einfach eine Zahl, und zwar die Ableitung an der Stelle a, und es gilt: $f(a+x) = f(a) + f'(a)\cdot x + \varphi(x)$.

Beispiel 30.24. Sei $q\colon \mathbb{R}^n \to \mathbb{R}$, $q(x) = x^t \cdot C \cdot x$, $C \in \mathbb{R}^{n\times n}$ mit $C = C^t$ symmetrisch. Sei ferner $a \in \mathbb{R}^n$. Dann gilt:

$$q(x+a) = (x+a)^t C(x+a) = \underbrace{a^t Ca}_{q(a)} + 2a^t Cx + \underbrace{x^t Cx}_{\varphi(x)}.$$

Außerdem ist:

$$\|x^t Cx\| = \|\langle x, Cx\rangle\| \le \|x\| \cdot \|Cx\| \le \|C\| \cdot \|x\|^2. \tag{30.1}$$

Hierbei ist $\|C\|$ die sogenannte **Matrixnorm** von C, definiert duch

$$\|C\| := \max_{x\in\mathbb{R}^n,\ \|x\|=1} \|Cx\|.$$

Damit gilt $\left\|C\cdot x\right\| = \left\|C\cdot \frac{x}{\|x\|}\cdot\|x\|\right\| \le \|C\|\cdot\|x\|$, so dass (30.1) tatsächlich erfüllt ist. Insgesamt folgt also:

$$\frac{|\varphi(x)|}{\|x\|} \le |c|\cdot\|x\| \underset{x\to 0}{\to} 0$$

und $2a^t \in \mathbb{R}^{1\times n}$ ist die Jacobimatrix.

Satz 30.25 (Kettenregel). *Seien* $U \subset \mathbb{R}^n$ *offen,* $f\colon U \to \mathbb{R}^m$, $f(U) \subset V \subset \mathbb{R}^m$, V *offen,* $g\colon V \to \mathbb{R}^k$*. Ist* f *in a total diffbar und* g *im Punkt* $b = f(a)$ *total diffbar, dann ist die Komposition* $h = g\circ f\colon U \to \mathbb{R}^k$ *im Punkt a total diffbar und es gilt für das Differential:*

$$Dh(a) = Dg(b)\cdot Df(a).$$

Siehe auch Abbildung 30.6.

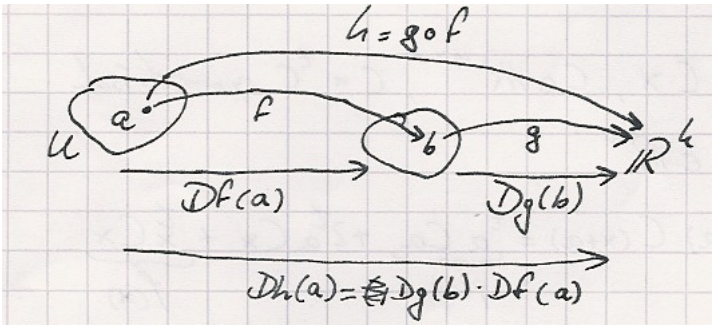


Abbildung 30.6. Die Kettenregel.

Beweis. Sei φ der Fehler für f und ψ der Fehler für g. Dann: $f(x+a) = f(a) + Ax + \varphi(x)$, $g(y+b) = g(b) + By + \psi(y)$ mit $A = Dg(b)$ und $B = Df(a)$. Daraus folgt:

$$\begin{aligned}(g \circ f)(x+a) &= g(b + \underbrace{Ax + \varphi(x)}_{y}) \\ &= g(b) + B(Ax + \varphi(x)) + \psi(Ax + \varphi(x)) \\ &= c + BAx + \underbrace{B\varphi(x) + \psi(Ax + \varphi(x))}_{\eta(x)}.\end{aligned}$$

Es gilt weiterhin:

$$\frac{\|B\varphi(x)\|}{\|x\|} \leq \|B\| \cdot \frac{\|\varphi(x)\|}{\|x\|} \underset{x\to 0}{\to} 0,$$

da f diffbar ist. Außerdem gilt:

$$\frac{\|\psi(Ax+\varphi(x))\|}{\|x\|} = \varepsilon(x) \cdot \frac{\|Ax+\varphi(x)\|}{\|x\|} \leq \varepsilon(x) \cdot \left(\|A\| + \frac{\varphi(x)}{\|x\|}\right) \underset{x\to 0}{\to} 0,$$

wobei $\varepsilon(x) = \frac{\|\psi(Ax+\varphi(x))\|}{\|Ax+\varphi(x)\|} \to 0$, da g diffbar.

Insgesamt folgt also, dass $\eta(x) \to 0$ für $x \to 0$. BA ist die Jacobimatrix von $g \circ f$, weil das Produkt von Matrizen der Hintereinanderausf"hrung der zugehörigen linearen Abbildungen Dg und Df entspricht. □

Korollar 30.26. *Die Einträge der Jacobimatrix $A = Df(a) = (a_{ij})$ sind die partiellen Ableitungen der Komponentenfunktionen: $a_{ij} = \frac{\partial f_i}{\partial x_j}(a)$. Also: $J_f(a) = \left(\frac{\partial f_i}{\partial x_j}(a)\right)$.*

Beweis. Für ein festes Paar (i, j) betrachten wir:

$$e\colon \mathbb{R} \to \mathbb{R}^n,\ x_i \mapsto (a_1, \dots, a_{i-1}, x_i, a_{i+1}, \dots, a_n)$$
$$g\colon \mathbb{R}^m \to \mathbb{R},\ (y_1, \dots, y_m) \mapsto y_j.$$

Dies sind lineare Abbildungen. Es gilt daher: $De = (0, \dots, 1, \dots, 0)^t =: {e_i}^t$ (1 an der i-ten Position) und $Dg = (0, \dots, 1, \dots, 0) =: e_j$ (1 an der j-ten Position). Damit folgt nun mit der Kettenregel:

$$\begin{aligned} D(g \circ f \circ e)(a_i) &= Dg(f(a)) \cdot Df(a) \cdot De(a_i) \\ &= e_j \cdot J_f(a) \cdot {e_i}^t = a_{ji}. \end{aligned}$$

Andererseits ist $g \circ f \circ e$ eine Abbildung von $\mathbb{R}$ nach $\mathbb{R}$, so dass die Jacobimatrix nur einen Eintrag hat, nämlich: $(g \circ f \circ e)(x_i) = f_j(a_1, \dots, a_{i-1}, x_i, \dots, a_n)$; nach x_i ableiten liefert:

$$\frac{\partial}{\partial x_i}(g \circ f \circ e)(a_i) = \frac{\partial f_j}{\partial x_i}(a).$$

□

Beispiel 30.27 (Polarkoordinaten). Sei $P\colon (r, \varphi) \mapsto (r\cos\varphi, r\sin\varphi)$, wobei wir P entweder als Abbildung $P\colon \mathbb{R}^2 \to \mathbb{R}^2$ oder $P\colon [0, 2\pi] \times \mathbb{R}_{\geq 0} \to \mathbb{R}^2$ auffassen (s. Abb. 30.7). $P|_{]0,2\pi[\times]0,\infty[}$ ist injektiv. Es gilt:

$$J_P = DP = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix},\ \det J_P = r.$$

Sei nun: $g\colon \mathbb{R}^2 \to \mathbb{R}$, $g(x, y) = x^2 + y^2$, dann ist $Dg = (2x, 2y)$ und $h := g \circ P = r^2$,

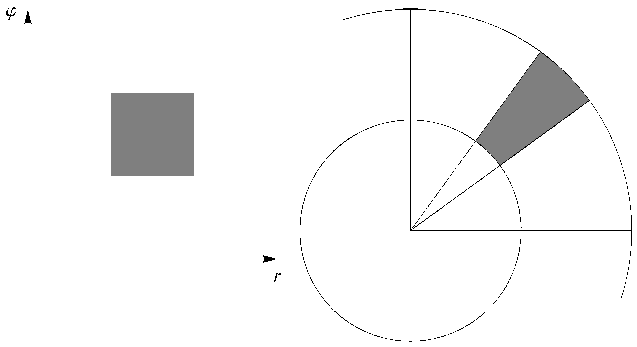


Abbildung 30.7. Polarkoordinaten.

d.h. $\frac{\partial h}{\partial r} = 2r$ und $\frac{\partial h}{\partial \varphi} = 0$. Nun: $Dh \overset{!}{=} Dg(P((r, \varphi))) \cdot DP(r, \varphi) = (2r\cos\varphi, 2r\sin\varphi) \cdot DP = (2r, 0)$, wie erhofft.

Korollar 30.28 (aus Satz 30.25, Korollar 30.26 und Beweis). *Seien $U \subset \mathbb{R}^n$ offen $f\colon U \to \mathbb{R}^m$, $f(U) \subset V \subset \mathbb{R}^m$ und V offen sowie $g\colon V \to \mathbb{R}$, und $h := g \circ f$. Die Koordinaten in U bezeichnen wir mit x_i, jene in V mit y_j. Dann gilt:*

$$\frac{\partial h}{\partial x_i}(x) = \sum_{j=1}^{m} \frac{\partial g}{\partial y_j}(f_1(x), \ldots, f_m(x)) \cdot \frac{\partial f_j}{\partial x_i}(x).$$

Bemerkung 30.29. Der Zusammenhang zwischen den Diffbarkeitsbegriffen ist wie folgt:

$$\text{stetig partiell diffbar} \Rightarrow \text{total diffbar} \begin{array}{l} \Rightarrow \text{partiell diffbar} \\ \Rightarrow \text{stetig.} \end{array}$$

Weitere Implikationen gelten nicht (hier nicht bewiesen).

Wir möchten nun auch Ableitungen in Richtungen definieren, die nicht den Koordinatenrichtungen entsprechen:

Definition 30.30. *Seien $U \subset \mathbb{R}^n$ offen und $f\colon U \to \mathbb{R}$, $a \in U$ und $v \in \mathbb{R}^n$ mit $\|v\| = 1$.*

$$(D_v f)(a) = \lim_{h \to 0} \frac{f(a + h \cdot v) - f(a)}{h}$$

*heißt **Richtungsableitung** von f in Richtung v.*

Satz 30.31. *Sei $f\colon U \to \mathbb{R}$, $U \subseteq \mathbb{R}^n$ offen, total diffbar in a. Dann gilt:*

$$D_v f(a) = \langle (\operatorname{grad} f)(a),\ v \rangle.$$

Insbesondere gilt: Für $v \in S^{n-1} := \{v \in \mathbb{R}^n \mid \|v\| = 1\}$ ist die Richtungsableitung maximal genau dann, wenn der Gradient $\operatorname{grad} f(a)$ in die gleiche Richtung wie v zeigt.

Beweis. Kettenregel und Geometrie der orthogonalen Projektion auf $\mathbb{R}v$. Details, s. [For08b]. □

30.3.3 Taylorformel

In einer Variablen ist die Formel

$$f(x + a) = f(a) + f'(a) \cdot x + \frac{1}{2} f''(a) \cdot x^2 + \varphi(x)$$

nur der erste Fall der Taylorformel (siehe Abb. 30.8).

Wir möchten dies für Funktionen $f\colon U \to \mathbb{R}$, $U \subset \mathbb{R}^n$ offen, verallgemeinern. Wir werden f durch Polynome in $x_1, \ldots, x_n$ approximieren. Wie man vermutlich erwartet, wird sich dabei herausstellen, dass $f'(a)$ durch die Jacobi–Matrix ersetzt wird und die Hesse–Matrix in $\varphi(x)$ auftaucht. Mit *geschickter* Indexnotation ist die Formel am Ende genauso einfach wie in einer Variablen.

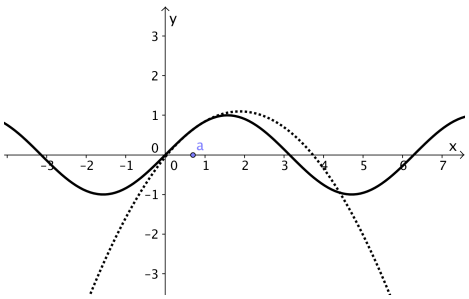


Abbildung 30.8. Approximation des Sinus in einer Variablen x an der Stelle a durch das Taylorpolynom zweiten Grades in a.

Notation 30.32. *$\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ nennen wir einen **Multiindex**. $|\alpha| = \alpha_1 + \cdots + \alpha_n$ heißt **Totalgrad** von α. Wir setzen $x^\alpha := x_1^{\alpha_1} \cdot \cdots \cdot x_n^{\alpha_n}$. Dann bezeichnet*

$$D^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x^\alpha} = \frac{\partial^{|\alpha|} f}{(\partial x_1)^{\alpha_1} \cdot \cdots \cdot (\partial x_n)^{\alpha_n}}$$

die α-te (partielle) Ableitung und $\alpha! := \alpha_1! \cdots \alpha_n!$. Für $|\alpha|$-mal steig partiell diffbare Funktionen kommt es auf die Reihenfolge des partiellen Ableitens nicht an.

Definition 30.33. *Sei $U \subset \mathbb{R}^n$ offen, $a \in U$. Sei ferner $f\colon U \to \mathbb{R}$ k-mal stetig partiell diffbar. Dann heißt das Polynom*

$$\sum_{|\alpha| \le k} \frac{\partial^{|\alpha|} f}{(\partial x)^\alpha}(a) \cdot \frac{(x-a)^\alpha}{\alpha!}$$

*das k-**te Taylorpolynom** von f in a. Es ist das eindeutig bestimmte Polynom, welches die gleichen Werte für die partiellen Ableitungen bis zur Ordnung k an der Stelle a hat wie f.*

Satz 30.34 (Taylorformel). *Sei $f\colon U \to \mathbb{R}$ $(k+1)$-mal stetig partiell diffbar. Dann gibt es für jedes $x \in \mathbb{R}^n$, das so klein ist, dass die Strecke $\{a + tx \mid t \in [-1, 1]\} \subset U$ ist, ein $\vartheta \in [0, 1]$, so dass:*

$$f(a+x) = \sum_{|\alpha| \le k} \frac{D^\alpha f(a)}{\alpha!} x^\alpha + \sum_{|\alpha| = k+1} \frac{D^\alpha f(a)}{\alpha!} (a + \vartheta \cdot x) x^\alpha.$$

Beweis. Wir betrachten die Funktion $g(t) = f(a + t \cdot x)$ in einer Variablen, die Taylorformel dort und die Identität

$$\frac{d^k g}{(dt)^k}(t) = k! \sum_{|\alpha| = k} \frac{D^\alpha f(a + t \cdot x)}{\alpha!} x^\alpha,$$

welche mit Induktion nach k aus der Kettenregel (der Fall $k = 1$) folgt. □

Für eine ausführlichere Darstellung, siehe die Literatur. Hier nur ein Beispiel:

Beispiel 30.35. Wir betrachten einen konkreten und einen etwas allgemeineren Fall:

1. $f(x,y) = e^x \cdot e^y = e^{x+y}$. Das zweite Taylorpolynom in $(0,0)$ berechnet sich folgendermaßen:

   Die Multiindizes α mit $|\alpha| \leq 2$ sind $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$, $(2,0)$, $(0,2)$.

   Wir müssen also die entsprechende partiellen Ableitungen an der Stelle $a = (0,0)$ auswerten. Da $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = e^x e^y = f$ ist, sind alle auch höhreren partiellen Ableitungen identisch:

   $$\begin{aligned} D^{(0,0)}f(a) &= f(a) = e^0 e^0 = 1. \\ D^{(1,0)}f(a) &= \frac{\partial f}{\partial x}(a) = e^0 e^0 = 1. \\ &\cdots\cdots \\ D^{(1,1)}f(a) &= \frac{\partial^2 f}{\partial x \partial y}(a) = e^0 e^0 = 1. \end{aligned}$$

   Die in der Taylorformel auftretenden Fakultäten sind $0! = 1, 1! = 1, 2! = 2$ und die Potenzen von $v := (x,y)$ sind $v^{0,0} = x^0 y^0 = 1$, $v^{1,0} = x^1 y^0 = x, \ldots,$ $v^{0,2} = x^0 y^2 = y^2$, $v^{1,1} = x^1 y^1 = xy$.

   Damit ist das zweite Taylorpolynom in $(0,0)$ (siehe auch Abb. 30.9):

   $$\begin{aligned} \sum_{|\alpha|\leq 2} \frac{D^\alpha f}{\alpha!}(a)v^\alpha &= \frac{D^{(0,0)}f}{0!0!}(a)v^{0,0} + \cdots + \frac{D^{(0,2)}f}{0!2!}(a)v^{0,2} + \frac{D^{(1,1)}f}{1!1!}(a)v^{1,1} \\ &= \frac{1}{1} + \frac{1}{1}\cdot x + \frac{1}{1}\cdot y + \frac{1}{2}\cdot x^2 + \frac{1}{2}\cdot y^2 + 1\cdot xy. \\ &= 1 + x + y + \frac{(x+y)^2}{2}. \end{aligned}$$

2. Sei $f\colon U \to \mathbb{R}$, $U \subset \mathbb{R}^n$ offen, f zweimal stetig partiell diffbar. Wie sieht das zweite Taylorpolynom in $a = 0 \in U$ in diesem allgemeinen Fall aus? Es lässt sich sehr komprimiert schreiben als:

   $$f(0) + \langle \operatorname{grad} f(0), x\rangle + \frac{1}{2} x^{\mathrm{t}} \operatorname{Hess}(f)(0) x.$$

   Für $x = (x_1, \ldots, x_n)$ gilt nämlich: $\frac{\partial f}{\partial x_1}(0) \cdot x^{(1,0,\ldots,0)} + \cdots + \frac{\partial f}{\partial x_n}(0) \cdot x^{(0,\ldots,0,1)} = \langle \operatorname{grad} f(0), x\rangle$. Die Aussage über den Term mit $Hess(f)$ zeigt man analog.

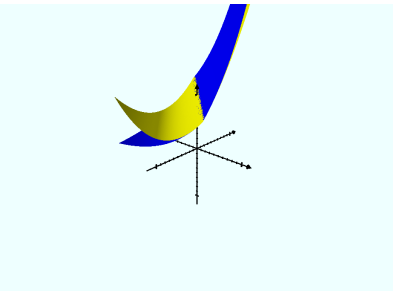


Abbildung 30.9. Eine Approximation von $\exp(x+y)$ durch das Taylorpolynom zweiten Grades im Punkt $(0,0)$.

30.3.4 Extremalstellen

Viele praktische Probleme führen auf Optimierungsaufgaben. Auch in der Situation, dass hierbei mehrere Variablen auftreten, gibt es einen Kalkül, ähnlich der Kurvendiskussion im univariaten Fall, um diese anzugehen.

Definition 30.36. *Sei $f\colon U \to \mathbb{R}$ eine Funktion, $U \subset \mathbb{R}^n$. f hat in a ein **lokales Maximum** (**lokales Minimum**), wenn ein Ball $B_r(a) \subset U$ existiert, so dass $f|_{B_r(a)}$ in a das Maximum (Minimum) hat. f hat in a ein **lokales Extremum**, wenn einer der beiden Fälle eintritt.*

In 26.15 haben wir definiert, wann eine Matrix A positiv definit heißt, geschrieben $A > 0$, nämlich wenn alle Eigenwerte von A strikt positiv sind, d.h. wenn $\overline{z}^{\,t} A z > 0 \ \forall z \in \mathbb{K}^n \setminus \{0\}$. Anschließend haben wir in Satz 26.16 das Hurwitz–Kriterium dafür bewiesen. Um ein hinreichendes Kriterium für Extremstellen geben zu können, benötigen wir nun auch die folgenden verwandten Begriffe:

Definition 30.37. *Eine hermitesche Matrix $A \in \mathbb{C}^{n\times n}$ (symmetrisch über $\mathbb{R}$) heißt **positiv semi–definit**, wenn alle Eigenwerte von A größer oder gleich 0 sind.*

*A heißt **negativ definit** bzw. **negativ semi–definit** (in Zeichen: $A < 0$ bzw. $A \leq 0$), wenn $-A$ positiv definit bzw. positiv semi–definit ist.*

*A heißt **indefinit**, wenn keiner der vorigen Fälle eintritt.*

Es gibt auch für die negative Definitheit ein Kriterium im Stil des Hurwitz–Kriteriums; dieses ist aber ein wenig komplizierter zu formulieren, als man denken könnte. Insbesondere ist die Negativität aller linken oberen Minoren kein Kriterium für die negative Definitheit einer Matrix. Am Einfachsten ist es daher wohl meist, die Eigenwerte zu berechnen.

Satz 30.38. *Sei $U \subset \mathbb{R}^n$ offen, $f\colon U \to \mathbb{R}$ zweimal stetig partiell diffbar.*

1. *Notwendig dafür, dass f in $a \in U$ ein lokales Extremum hat, ist, dass*

$$\frac{\partial f}{\partial x_1}(a) = \cdots = \frac{\partial f}{\partial x_n}(a) = 0.$$

2. *Ist die notwendige Bedingung erfüllt, dann ist hinreichend für ein lokales Minimum, dass die Matrix $A = \operatorname{Hess}(f)(a) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(a)\right)_{ij}$ positiv definit ist. Ist A negativ definit, dann liegt ein lokales Maximum vor. Ist A indefinit, dann ist a kein lokales Extremum.*

Diese Aussage verallgemeinert offenbar direkt die entsprechenden Resultate aus dem univariaten Fall, nämlich Satz 10.2 und Satz 10.10. Bevor wir diesen Satz auf Seite 451 beweisen, zunächst einige Beispiele:

Beispiel 30.39. 1. $f(x,y) = x^2 + y^2$, $\frac{\partial f}{\partial x} = 2x = 0$, $\frac{\partial f}{\partial y} = 2y = 0$. Somit ist $a = (0,0)$ der einzige Kandidat für ein lokales Extremum. Es gilt:

$$\operatorname{Hess}(f)(a) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} > 0.$$

Also hat f in $(0,0)$ ein lokales Minimum (Abb. 30.10).

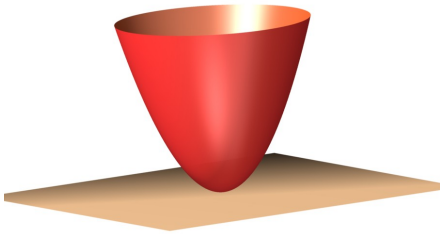


Abbildung 30.10. Ein lokales Minimum.

2. $f(x,y) = x^2 - y^2$. Wieder ist der Nullpunkt der einzige Kandidat für ein lokales Extremum. Es gilt:

$$\operatorname{Hess}(f)(a) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}.$$

Dies ist eine indefinite Matrix. Der Ursprung ist hier also ein sogenannter **Sattelpunkt** (Abb. 30.11).

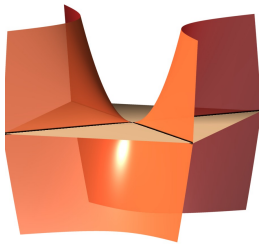


Abbildung 30.11. Ein Sattelpunkt. Es ist auch die Niveaulinie $f(x,y) = x^2 - y^2 = (x-y)\cdot(x+y) = 0$ eingezeichnet.

3. Sei $f(x,y) = x^2 + y^3$. Hier ist ebenfalls $a = (0,0)$ der einzige Kandidat. Dann ist $\frac{\partial f}{\partial x} = 2x$, $\frac{\partial f}{\partial y} = 3y^2$ und

$$\mathrm{Hess}(f)(a) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \text{ (positiv semidefinit).}$$

Der Satz macht für diese Situation keine Aussage. Der einzige Kandidat für ein lokales Extremum, der Ursprung, ist aber keines, da $f|_{\{x=0\}} : \mathbb{R} \to \mathbb{R},\ y \mapsto y^3$ kein Extremum hat (s. auch Abb. 30.12).

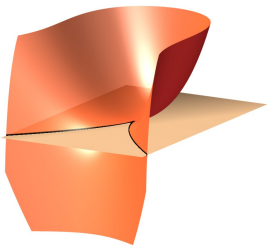


Abbildung 30.12. Die gewöhnliche Spitze $f(x,y) = x^2 + y^3$ als Funktion. Es ist auch die Niveaulinie $f(x,y) = 0$ eingezeichnet.

4. Für $f(x,y) = x^2 + y^4$ erhalten wir ebenfalls $a = (0,0)$ als einzigen Kandidaten und wieder

$$\mathrm{Hess}(f)(a) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \text{ (positiv semidefinit).}$$

Auch hier macht der Satz keine Aussage. In diesem Fall ist a aber offensichtlich ein Minimum (s. auch Abb. 30.13), denn $f(x,y) \geq 0\ \forall (x,y) \in \mathbb{R}^2$ und $f(0,0) = 0$.

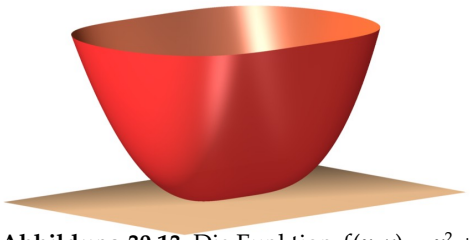


Abbildung 30.13. Die Funktion $f(x,y) = x^2 + y^4$.

Dies ist analog zum Fall einer Variablen, wo die Funktion $f(x) = x^4$ zwar eine verschwindende zweite Ableitung, aber trotzdem ein Minimum in $x = 0$ besitzt. Wir werden hier aber keine analogen Kriterien für höhere Ableitungen aufstellen.

5. Für $f(x,y) = x^2$ erhalten wir ebenfalls $a = (0,y)$ für jedes $y \in \mathbb{R}$ als Kandidaten und wieder ist

$$\operatorname{Hess}(f)(a) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \text{ (positiv semidefinit).}$$

In diesem Fall ist jeder Punkt der Geraden $\{(0,y) \mid y \in \mathbb{R}\}$ ein lokales Minimum. Allerdings sind dies keine **isolierten Minima**, in dem Sinn, dass es keine Umgebung der Punkte gibt, in der die Punkte jeweils das einzige Minimum sind. Analog kann man **isolierte Maxima** definieren.

Zum Satz 8.10 über die Existenz von Maximum und Minimum stetiger Funktionen auf einem abgeschlossenen (d.h. kompakten) Intervall gibt es folgende Verallgemeinerung:

Satz 30.40 (Maximum und Minimum auf einem Kompaktum). *Seien $K \subset \mathbb{R}^n$ eine kompakte Menge und $f\colon K \to \mathbb{R}$ eine stetige Funktion. Dann nimmt f ein Maximum und ein Minimum auf K an (s. Abb. 30.14).*

Beweis. Sei $M := \sup\{f(x) \mid x \in K\} \in]-\infty,\infty]$. Dann existiert eine Folge $(x^\nu)_{\nu\in\mathbb{N}}$, so dass $\lim_{\nu\to\infty} f(x^\nu) = M$. Die Folgen der Komponenten $(x_i^\nu) \subset \mathbb{R}$ sind jeweils beschränkt. Nach Bolzano–Weierstrass 5.33 existiert daher eine Teilfolge $(x^{\nu_k})_{k\in\mathbb{N}}$, die konvergiert. Der Grenzwert $y = \lim_{k\to\infty} x^{\nu_k} \in K$, da K abgeschlossen ist (siehe Aufgabe **??**). Für die Funktionswerte gilt: $f(y) = \lim_{k\to\infty} f(x^{\nu_k}) = M$ wegen des Folgenkriteriums für Stetigkeit (Satz 8.5), da f stetig ist. In $y \in K$ nimmt f daher ein Maximum an. Der Fall des Minimums ist analog. □

Nun zum notwendigen und zum hinreichenden Kriterium für Extremstellen:

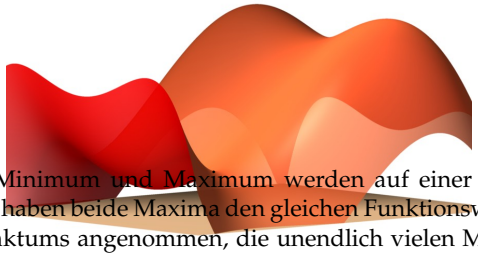


Abbildung 30.14. Minimum und Maximum werden auf einer kompakten Menge angenommen. Hier haben beide Maxima den gleichen Funktionswert und werden im Inneren des Kompaktums angenommen, die unendlich vielen Minima dagegen auf dem Rand.

Beweis (von Satz 30.38).

1. Hat f in a ein lokales Extremum, dann hat die Funktion

$$g_i \colon t \mapsto g_i(t) = f(a + e_i t),$$

wobei e_i der i-te Einheitsvektor ist, eines in $t = 0$. Der entsprechende Satz 10.2 in einer Variablen liefert nun $g_i'(0) = 0$, d.h.:

$$0 = g_i'(0) = \frac{\partial g_i}{\partial t}(0) \overset{\text{Kettenregel}}{=} \frac{\partial f}{\partial x_i}(a).$$

2. Es sei $(\operatorname{grad} f)(a) = 0$ und zunächst einmal $A = \operatorname{Hess}(f)(a) > 0$. Die Taylorformel erster Ordnung für f nahe a ergibt, da $\sum_{|\alpha|\le 1} \frac{D^\alpha f(a)}{\alpha!} x^\alpha = f(a) + \langle \operatorname{grad} f(a), x\rangle$: Es existiert $\vartheta \in [0,1]$, so dass

$$\begin{aligned} f(x+a) &= f(a) + \Big\langle \operatorname{grad} f(a), x \Big\rangle + \frac{1}{2} x^{\mathrm{t}} \Big(\frac{\partial^2 f}{\partial x_i \, \partial x_j}(a + \vartheta \cdot x) \Big) x \\ &= f(a) + \frac{1}{2} x^{\mathrm{t}} A x + \frac{1}{2} x^{\mathrm{t}} B(x) x, \end{aligned}$$

wobei $B(x) \underset{x \to 0}{\to} 0 \in \mathbb{R}^{n \times n}$, da f zweimal stetig diffbar ist. Sei nun

$$\eta := \min_{v \in \mathbb{R}^n \setminus \{0\}} \frac{v^{\mathrm{t}} A \, v}{\|v\|^2} = \min_{v \in \mathbb{R}^n,\, \|v\| = 1} \|Av\|,$$

das wegen des obigen Satzes 30.40 auch existiert, da die 1–Sphäre $\partial B_1(0)$ kompakt ist. Es gilt $\eta > 0$, da A positiv definit ist. Mit $\varepsilon := \frac{\eta}{2}$ folgt:

$$\exists\, \delta > 0, \text{ so dass } \|B(x)\| < \varepsilon \ \forall x \text{ mit } \|x\| < \delta.$$

Wir zeigen, dass a das Minimum von $f|_{B_\delta(a)}$ ist:

$$f(x+a) - f(a) = \frac{1}{2}\left(x^t Ax + x^t B(x)x\right)$$
$$\geq \frac{1}{2}\left(\eta\|x\|^2 - \varepsilon\|x\|^2\right) = \frac{\eta-\varepsilon}{2}\|x\|^2 \geq 0$$

und Gleichheit gilt genau dann, wenn $x = 0$.

Die anderen Fälle gehen analog. Im indefiniten Fall schränkt man f auf $a + \mathrm{Eig}(A, \lambda)$ mit λ positiv bzw. negativ ein.

□

Beispiel 30.41. Wir betrachten $f(x, y) = x^2 - y^2 + x^3$ (Abb. 30.15) und suchen mögliche Extrema.

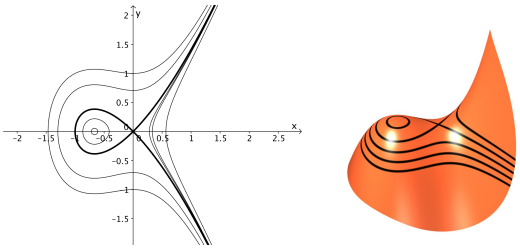


Abbildung 30.15. Der Newtonsche Knoten $\mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto f(x, y) = x^2 - y^2 + x^3$: Die linke Abbildung zeigt einige Niveaulinien, die rechte die Funktion gemeinsam mit einigen eingezeichneten Niveaulinien.

Zunächst berechnen wir dafür die eine partielle Ableitung:

$$\frac{\partial f}{\partial x} = 2x + 3x^2 \stackrel{!}{=} 0, \qquad \text{also: } x = 0 \text{ oder } x = -\frac{2}{3}.$$

Die andere partielle Ableitung liefert:

$$\frac{\partial f}{\partial y} = -2y \stackrel{!}{=} 0 \Rightarrow y = 0.$$

Insgesamt folgt: $a = (0, 0)$ und $a = (-\frac{2}{3}, 0)$ kommen als lokale Extrema in Frage. Um Genaueres herauszufinden, betrachten wir die Hesse-Matrix

$$\mathrm{Hess}(f) = \begin{pmatrix} 6x+2 & 0 \\ 0 & -2 \end{pmatrix}$$

an diesen Stellen:

$$\mathrm{Hess}(f)(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \quad \text{(indefinit)},$$
$$\mathrm{Hess}(f)(-\frac{2}{3},0) = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} \quad \text{(negativ definit)}.$$

Also ist $(-\frac{2}{3},0)$ ein lokales Maximum und $(0,0)$ kein Extremum.

Aufgaben

Aufgabe 30.1 (Offene Mengen). Zeigen Sie:

1. $U_i \subset \mathbb{R}^n$, $i \in I$ eine Familie von offenen Mengen $\Rightarrow \bigcup_{i\in I} U_i$ offen.
2. $U_1, U_2 \subset \mathbb{R}^n$ offen $\Rightarrow U_1 \cap U_2$ offen.
3. $\mathbb{R}^n$ und $\emptyset$ sind offen.

Aufgabe 30.2 (Abgeschlossene Mengen). Im Folgenden bezeichne $\|.\|$ die euklidische Norm auf $\mathbb{R}^n$. Wir sagen, dass eine Folge $(x_\nu)_{\nu\in\mathbb{N}}$ im $\mathbb{R}^n$ zu einem Punkt $p \in \mathbb{R}^n$ *konvergiert* (in Zeichen: $\lim_{\nu\to\infty} x_\nu = p$), falls gilt: Zu jedem $\varepsilon > 0$ gibt es ein $\nu_0 \in \mathbb{N}$, so dass für alle $\nu > \nu_0$ gilt: $\|x_\nu - p\| < \varepsilon$.

1. Zeigen Sie, dass eine Folge $(x_\nu)_{\nu\in\mathbb{N}}$ genau dann gegen p konvergiert, wenn für jedes $j \in \{1,\dots,n\}$ die Folge der j-ten Komponente, $(x_{\nu j})_{\nu\in\mathbb{N}}$, gegen die entsprechende Komponente p_j von p konvergiert, d.h.:
$$\lim_{\nu\to\infty} x_\nu = p \;\Leftrightarrow\; \Big(\lim_{\nu\to\infty} x_{\nu j} = p_j \;\forall j = 1,\dots,n\Big).$$
2. Zeigen Sie: $A \subset \mathbb{R}^n$ ist genau dann abgeschlossen, wenn für jede konvergente Folge $(x_\nu)_{\nu\in\mathbb{N}} \subset A$ gilt: $\lim_{\nu\to\infty} x_\nu \in A$.

Aufgabe 30.3 (Kompakte Mengen). Sei $(x_\nu)_{\nu\in\mathbb{N}}$ eine Folge im $\mathbb{R}^n$. Eine *Teilfolge* ist eine Folge $(a_{\kappa(\nu)})_{\nu\in\mathbb{N}}$, wobei $\kappa : \mathbb{N} \to \mathbb{N}, \nu \mapsto \kappa(\nu)$ eine injektive Abbildung ist.

Zeigen Sie: Eine Menge $K \subset \mathbb{R}^n$ ist genau dann kompakt, wenn jede Folge $(x_n)_{n\in\mathbb{N}}$ mit Werten in K eine in K konvergente Teilfolge besitzt.

Aufgabe 30.4 (Jacobimatrix). In welchen Punkten ist die Jacobi-Matrix der Abbildung

$$f\colon \mathbb{R}^3 \to \mathbb{R}^3,\ (x,y,z) \mapsto (4y,\ 3x^2 - 2\sin(yz),\ 2yz)$$

nicht invertierbar?

Aufgabe 30.5 (Definitheit). Bestimmen Sie, ob die folgenden Matrizen positiv definit, negativ definit oder indefinit sind:

$$A = \begin{pmatrix} 6 & 10 & -1 & 2 \\ 10 & 21 & -7 & -1 \\ -1 & -7 & 15 & -2 \\ 2 & -1 & -2 & 11 \end{pmatrix}, \quad B = \begin{pmatrix} -3 & -4 & -5 & 1 \\ -4 & -12 & -8 & 16 \\ -5 & -8 & -6 & 5 \\ 1 & 16 & 5 & -35 \end{pmatrix} \in \mathbb{R}^{4\times 4}.$$

Aufgabe 30.6 (Taylorpolynom). Bestimmen Sie das Taylorpolynom 3-ter Ordnung im Punkt $(1, 1)$ von:

$$f\colon \mathbb{R}^2_{>0} \to \mathbb{R},\ f(x, y) = x^y.$$

Aufgabe 30.7 (Taylorpolynom). Bestimmen Sie das Taylorpolynom 3-ter Ordnung im Punkt $(1, 1)$ von:

$$f\colon \mathbb{R}_{>0} \times \mathbb{R}_{>0},\ f(x, y) = \frac{x - y}{x + y}.$$

Aufgabe 30.8 (Extremstellen). Berechnen Sie alle Extremstellen der Funktion

$$f\colon \mathbb{R}^2 \to \mathbb{R},\ f(x, y) = \sqrt{x^4 + (y^2 - 1)^2},$$

finden Sie heraus, ob die lokale Minima oder Maxima sind und skizzieren Sie den Graph von f.

Hinweis: Sie dürfen auch ein geeignetes Computer-Algebra-Programm einsetzen.

Aufgabe 30.9 (Ausgleichsgerade).

1. Seien $(x_k, y_k) \in \mathbb{R}^2$, $k = 1, \ldots, n$ Punkte in der Ebene. Bestimmen Sie die Koeffizienten $a, b \in \mathbb{R}$ der Gerade $y = ax + b$, so dass

   $$\sum_{k=1}^{n} (ax_k + b - y_k)^2$$

   minimal wird. Diese erhaltene Gerade heißt **Ausgleichsgerade**.
2. Bestimmen Sie zu folgenden Punkten die Ausgleichgerade und skizzieren Sie die Punkte sowie die Ausgleichsgerade.

x	0	1	2	3	4	5
y	0.9	2.6	5.1	6.2	8.3	8.9

Aufgabe 30.10 (Extremwerte). Bestimmen Sie Lage und Art der lokalen Extrema der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto x^3y^2(1 - x - y).$$

31

Hyperflächen und der Satz über implizite Funktionen

Wir haben schon in der linearen Algebra Beispiele von Flächen kennengelernt, die als Nullstellenmenge von Funktionen in mehreren Variablen definiert sind, nämlich die Quadriken. In der geometrischen Modellierung und Visualisierung werden in letzter Zeit aber auch immer mehr Flächen im $\mathbb{R}^3$ verwendet, die sich zwar immer noch als Nullstellenmenge von Funktionen schreiben lassen, wobei diese Funktionen aber nicht unbedingt nur quadratisch, sondern komplizierter sind. In diesem Abschnitt gehen wir auf Möglichkeiten ein, wie man die Geometrie solcher Flächen, zumindest lokal, untersuchen und oft recht gut beschreiben kann.

31.1 Defintion und Hauptsatz

Definition 31.1. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ eine (diffbare) Funktion. Dann heißt ihre **Nullstellenmenge***

$$N(f) := N_0(f) = \{\, a \in \mathbb{R}^n \mid f(a) = 0 \,\}$$

*die durch f definierte **Hyperfläche**.*

Beispiel 31.2. Neben den bereits genannten Quadriken im $\mathbb{R}^n$ (siehe Abschnitt 25.2) – z.B. $x^2 + y^2 - z^2 = 0$, ein Doppelkegel, $x^2 + y^2 + z^2 - 1 = 0$, eine Kugel – sind Hyperebenen und Niveaumengen Beispiele für Hyperflächen, die wir bereits kennengelernt haben. Außerdem ist jeder Graph $x_{n+1} = f(x_1, \ldots, x_n)$ eine Hyperfläche, da seine Gleichung umgeschrieben werden kann zu $x_{n+1} - f(x_1, \ldots, x_n) = 0$.

Sogar, wenn wir uns auf polynomielle Funktionen beschränken, können Hyperflächen aber eine sehr komplexe Geometrie aufweisen. Etwa der sog. **Whitney Umbrella** (Abb. 31.1), bei dem eine Halbgerade aus der Fläche herausschaut.

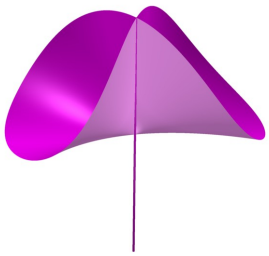


Abbildung 31.1. Der Whitney Umbrella.

Definition 31.3. *Sei $X = N(f)$, f stetig diffbar, eine Hyperfläche und $a \in X$. Ist*

$$f(x) = f(a) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(a)(x_j - a_j) + o(\|x - a\|)$$

die erste Taylorformel (zum Landau–Symbol $o(.)$ siehe Abschnitt 5.3), so heißt

$$T_a X = \left\{ x \in \mathbb{R}^n \;\middle|\; \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(a)(x_j - a_j) = 0 \right\}$$

*der **Tangentialraum** von X im Punkt a (s. Abb. 31.2).*

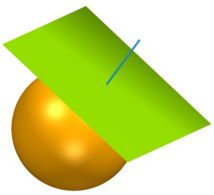


Abbildung 31.2. Tangentialraum und Gradient in einem glatten Punkt einer Fläche.

$T_a X$ ist der um a verschobene Untervektorraum

$$\{ x \in \mathbb{R}^n \mid \langle \operatorname{grad} f(a), x \rangle = 0 \}.$$

*Ist $\operatorname{grad} f(a) \neq 0$, so ist $T_a X$ eine Hyperebene und X heißt **glatt** in a. Andernfalls ist $T_a X$ der ganze Raum und X heißt**singulär** in a.*

Beispiel 31.4. 1. Sei $E = \{(x, y) \in \mathbb{R}^2 \mid x^2 + 2y^2 = 3\} = N(f)$, d.h. $f(x, y) = x^2 + 2y^2 - 3$ (Abb. 31.3). Einsetzen zeigt, dass $(1, 1) \in E$. Der Gradient grad $f = (2x, 4y)$ ist in diesem Punkt grad $f(1, 1) = (2, 4)$. Die Hyperfläche E ist daher glatt im Punkt $(1, 1)$.

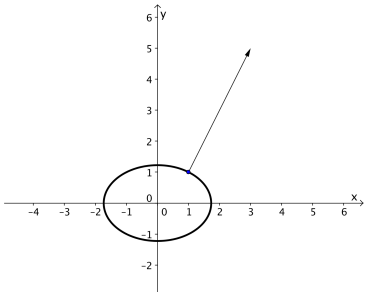


Abbildung 31.3. Tangentialraum und Gradient in einem glatten Punkt einer ebenen Kurve.

2. Der Tangentialraum an einer Hyperfläche im $\mathbb{R}^3$ berührt die Hyperfläche in der Umgebung eines glatten Punktes nicht unbedingt nur in einem einzigen Punkt. Dies liegt daran, wie die Fläche gekrümmt ist. Wir können darauf hier nicht weiter eingehen, sondern zeigen nur ein Beispiel: Der einschalige Hyperboloid $x^2 + y^2 - z^2 = 1$ im Punkt $(1, 0, 0)$ (Abb. 31.4).

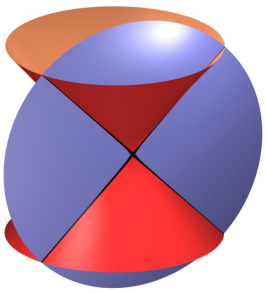


Abbildung 31.4. Eine Tangentialebene an einen einschaligen Hyperboloiden. Auf unserer Webseite gibt es dazu eine Animation: GIF-Format, SWF-Format. Genau wie das Bild wurde die Animation mit unserer Software **surfex** [HLM05] erstellt.

3. Ist $f(x, y) = y^2 - x^3$, so ist grad $f = (-3x^2, 2y)$ im Ursprung $(0, 0)$. Der Tangentialraum ist dort also der ganze Raum $\mathbb{R}^2$, obwohl man anschaulich eine eindeutige Tangente im Ursprung erkennen kann. Diese lässt sich aber nur durch Betrachtung höherer Ableitungen berechnen.

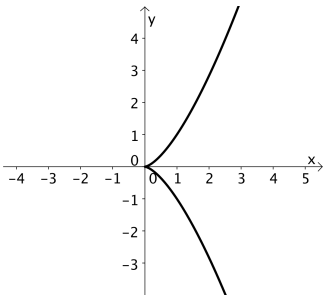


Abbildung 31.5. Eine singuläre Kurve.

Der Tangentialraum ist in glatten Punkten eine oft akzeptable Annäherung an die Hyperfläche.

Frage 31.5. *Gegeben seien $X = N(f)$ und $a \in X$ ein glatter Punkt. Können wir X nahe a besser darstellen?*

Antwort. Ja, wenn wir die Gleichung $f(x_1, \ldots, x_n) = 0$ nach einer Variablen auflösen können. Etwa nach x_n; dann suchen wir eine Funktion $g(x_1, \ldots, x_{n-1})$, so dass

$$f(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1})) = 0.$$

Beispiel 31.6. Wir betrachten nochmals das Beispiel 31.4: $E = \{(x, y) \in \mathbb{R}^2 \mid x^2 + 2y^2 - 3 = 0\}$. Die Gleichung können wir nach y auflösen. Für nicht negative y erhalten wir (s. auch Abb. 31.6):

$$y = \sqrt{\frac{3 - x^2}{2}}.$$

Dies beschreibt die Kurve für $y > 0$ wesentlich genauer als der Tangentialraum in einem der Punkte.

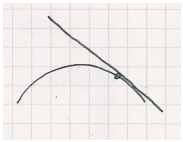


Abbildung 31.6. Eine lokal aufgelöste Kurvengleichung lässt sich sehr leicht visualisieren, wo die Auflösung gilt. Die Tangente liefert dagegen meist nur in einem Punkt eine gute Annäherung.

Der folgende Satz besagt, dass ein solches **Auflösen** zumindest **lokal** in der Umgebung eines Punktes oft möglich ist:

Satz 31.7 (über implizite Funktionen). *Sei $U \subset \mathbb{R}^n$, U offen, $f\colon U \to \mathbb{R}$ k-mal stetig diffbar und $a = (a_1, \dots, a_{n-1}, a_n) \in N(f)$. Gilt $\frac{\partial f}{\partial x_n}(a) \neq 0$, dann existieren offene Umgebungen $V' \subset \mathbb{R}^{n-1}$ von $(a_1, \dots, a_{n-1}) = a'$ und $V'' \subset \mathbb{R}$ von $a_n = a''$ mit $V' \times V'' \subset U$ (s. Abb. 31.7) und es existiert eine Funktion $g\colon V' \to V'' \subset \mathbb{R}$ mit $g(a') = a''$ und*

1. *$f(x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1})) = 0$ $\forall x' = (x_1, \dots, x_{n-1}) \in V'$ und*
2. *$\forall (x', x'') \in (V' \times V'') \cap N(f)$ gilt: $x'' = x_n = g(x')$.*

g ist k-mal stetig diffbar und

$$\frac{\partial g}{\partial x_i}(a') = -\frac{\frac{\partial f}{\partial x_i}(a)}{\frac{\partial f}{\partial x_n}(a)} \text{ für } i = 1, 2, \dots, n-1.$$

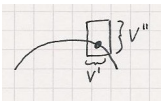


Abbildung 31.7. V' und V'' im Satz über implizite Funktionen.

Mit Hilfe von g wird die implizite Gleichung $f = 0$ also nach $x'' = x_n$ aufgelöst und die partiellen Ableitungen von g in a' können wir sogar explizit ausrechnen. Natürlich können wir nach anderen Variablen auflösen, indem wir die Variablen umnumerieren und dann den Satz anwenden.

Beweis. Der vollständige Beweis folgt später für eine allgemeinere Version in Satz 31.18. Nur die Formel für die partiellen Ableitungen zeigen wir sofort, unter der Annahme, dass der Rest bereits bewiesen ist, also insbesondere $x'' = x_n = g(x')$. Diese folgt aus der Kettenregel

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_i}\Big(f\big(x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1})\big)\Big) \\ &= \frac{\partial f}{\partial x_i}\big(x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1})\big) \\ &\quad + \frac{\partial f}{\partial x_n}\big(x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1})\big) \cdot \frac{\partial g}{\partial x_i}(x_1, \dots, x_{n-1}) \end{aligned}$$

durch Einsetzen von a', da $g(a') = a'' = a_n$. □

Formeln für höhere Ableitungen von g bekommt man, indem man erneut ableitet. In singulären Punkten verschwinden alle partiellen Ableitungen, so dass dort weder der Tangentialraum noch der Satz über implizite Funktionen Auskunft geben. Selbst, wenn man sich auf singuläre Punkte von Hyperflächen beschränkt, die durch Polynome gegeben sind, ist dies ein sehr weites und faszinierendes Gebiet, das zur sogenannten **Singularitätentheorie** gehört, auf die wir hier leider nicht genauer eingehen können.

31.2 Extrema mit Nebenbedingungen

Sei $h\colon U \to \mathbb{R}$, $U \subset \mathbb{R}^n$, eine diffbare Funktion. Wir möchten h unter der Nebenbedingung $f(x) = 0$ maximieren, wobei $f\colon U \to \mathbb{R}$ eine weitere diffbare Funktion ist. In diesem Abschnitt werden die meisten Beispiele für diese Art von Problem aus der Geometrie kommen, wie wir aber später sehen werden (Seite 542), hat diese Methode auch wichtige Anwendungen in der Statistik und vielen anderen Bereichen der Wissenschaft.

Beispiel 31.8. Für alle Punkte auf dem Kreis $f(x, y) = (x-1)^2+y^2-1 = 0$ (Abb. 31.8) möchten wir den Abstand $h(x, y)$ zum Ursprung maximieren. Offenbar ist dies der Punkt des Kreises, der im Bild am weitesten rechts liegt, nämlich $(1, 0)$.

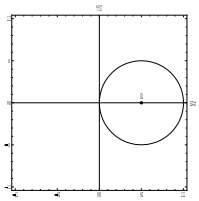


Abbildung 31.8. Eine Extremwertaufgabe mit Nebenbedingungen.

Satz/Definition 31.9. *Sei $f\colon U \to \mathbb{R}$ diffbar, $a \in N(f) = \{x \in U \mid f(x) = 0\}$ mit* grad $f(a) \neq 0$. *Sei $h\colon U \to \mathbb{R}$ eine weitere Funktion. Notwendig dafür, dass $h|_{N(f)}$ im Punkt a ein lokales Extremum hat, ist die Existenz eines $\lambda \in \mathbb{R}$, so dass* grad $h(a) = \lambda$ grad $f(a)$. *Der Faktor λ heißt* ***Lagrangescher Multiplikator****.*

Bevor wir den Satz beweisen, ein Beispiel:

Beispiel 31.10. $f(x, y) = x^2 + \frac{1}{4}y^2 - 1$, $h(x, y) = x + y$ (Abb. 31.9).

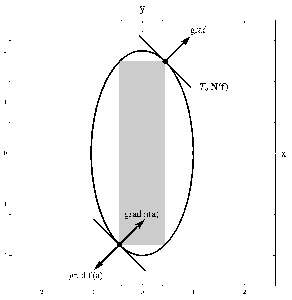


Abbildung 31.9. Eine Extremwertaufgabe mit Nebenbedingungen.

Es gilt: grad $h = (1,1)$, grad $f = (2x, \frac{y}{2})$. Die notwendige Bedingung liefert zwei Gleichungen:

$$1 = \lambda \cdot 2x, \quad 1 = \lambda \cdot \frac{y}{2}.$$

Gemeinsam mit der Gleichung $x^2 + \frac{1}{4}y^2 = 1$ vom Anfang ergibt sich (das sind insgesamt 3 Gleichungen mit 3 Unbekannten): $\lambda = \frac{1}{2x} \Rightarrow 2x = \frac{y}{2} \Rightarrow 0 = x^2 + \frac{(4x)^2}{4} - 1 = 5x^2 - 1$. Das liefert: $x = \pm\frac{1}{5}\sqrt{5}$, $y = \pm\frac{4}{5}\sqrt{5}$.

Da $h|_{N(f)}$ ein Maximum und ein Minimum annimmt, weil hier $N(f)$ kompakt ist und da grad $f(b) \neq 0 \; \forall b \in N(f)$, folgt: In $(\frac{1}{5}\sqrt{5}, \frac{4}{5}\sqrt{5})$ wird $h|_{N(f)}$ maximal und in $(-\frac{1}{5}\sqrt{5}, -\frac{4}{5}\sqrt{5})$ wird $h|_{N(f)}$ minimal.

Beweis (des Satzes 31.9). Da grad $f(a) \neq 0$, können wir wegen des Satzes

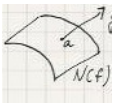


Abbildung 31.10. Zum Beweis des Satzes über Lagrangemultiplikatoren.

über implizite Funktionen die Gleichung nach einer der Variablen auflösen, etwa nach x_n. Der Satz liefert die Existenz einer Funktion $g\colon \mathbb{R}^{n-1} \supset U' \to \mathbb{R}$ mit $g(a') = a_n$ und $0 = f(x_1, \ldots, x_{n-1}, g(x'))$. Somit hat die Funktion $h(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1}))$ in $a' = (a_1, \ldots, a_{n-1})$ ein lokales Extremum, weil nach Voraussetzung h in a eines besitzt. Also:

$$\begin{aligned}
0 &= \Big(\frac{\partial}{\partial x_k}(h(x_1,\dots,x_{n-1},g(x_1,\dots,x_{n-1}))\Big)(a') \\
&= \Big(\frac{\partial h}{\partial x_k}(x_1,\dots,x_{n-1},g(x_1,\dots,x_{n-1})) \\
&\qquad + \frac{\partial h}{\partial x_n}(x_1,\dots,x_{n-1},g(x_1,\dots,x_{n-1}))\cdot\frac{\partial g}{\partial x_k}(x_1,\dots,x_{n-1})\Big)(a') \\
&= \frac{\partial h}{\partial x_k}(a) + \frac{\partial h}{\partial x_n}(a)\cdot\frac{\partial g}{\partial x_k}(a').
\end{aligned}$$

Andererseits liefert der Satz über implizite Funktionen:

$$\frac{\partial g}{\partial x_k}(a') = -\frac{\frac{\partial f}{\partial x_k}(a)}{\frac{\partial f}{\partial x_n}(a)}$$

für jedes k. Insgesamt erhalten wir demnach durch Einsetzen in die vorige Gleichung:

$$\frac{\partial h}{\partial x_k}(a) = \left(\frac{\frac{\partial h}{\partial x_n}(a)}{\frac{\partial f}{\partial x_n}(a)}\right)\cdot\frac{\partial f}{\partial x_k}(a).$$

Der erste Faktor hängt dabei nicht mehr von k ab, so dass wir ihn mit λ bezeichnen können und somit erhalten: $\operatorname{grad} h(a) = \lambda\cdot\operatorname{grad} f(a)$. □

Bemerkung 31.11. Mit Hilfe der Rekursionsformel für höhere partielle Ableitungen von g aus dem Satz über implizite Funktionen lässt sich auch die Hesse–Matrix von $h(x_1,\dots,x_{n-1},g(x_1,\dots,x_{n-1}))$ bestimmen, also ein hinreichendes Kriterium angeben.

31.3 Der Umkehrsatz

Für differenzierbare Funktionen $f\colon\mathbb{R}\to\mathbb{R}$ ist uns bekannt, dass diese in einer Umgebung eines Punkte $a\in\mathbb{R}$ mit $f'(a)\neq 0$ lokal umkehrbar sind, weil sie auf einer geeigneten Umgebung streng monoton und damit bijektiv sind. Diese Aussage verallgemeinern wir nun.

Seien $U\subset\mathbb{R}^n$ offen und $f\colon U\to\mathbb{R}^n$ eine diffbare Abbildung (s. Abb. 31.11); wir sind hier also im Fall, dass sowohl U als auch $f(U)$ Teilmengen des gleichen Raumes $\mathbb{R}^n$ sind. Wir schreiben nun $b=f(a)$ für $a\in U$. Existiert eine offene Umgebung V von a, so dass $f|_V$ bijektiv ist und ist $g=(f|_V)^{-1}$ ebenfalls diffbar, dann ist $(Dg)(b)\circ(Df)(a)=E_n$ die $n\times n$–Einheitsmatrix. Der folgende Satz gibt eine hinreichende Bedingung dafür, dass diese Situation eintritt. Mit Hilfe dieses Satzes werden wir dann eine allgemeinere Variante des Satzes über implizite Funktionen herleiten können.

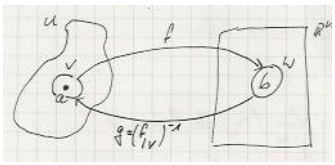


Abbildung 31.11. Zum Umkehrsatz.

Satz 31.12 (Umkehrsatz). *Seien $U \subset \mathbb{R}^n$ offen und $f\colon U \to \mathbb{R}^n$ eine k-mal stetig diffbare Abbildung, $a \in U$ und $b = f(a)$. Ist $\det(Df(a)) \neq 0$, dann existieren Umgebungen V und W mit $a \in V \subset U$ und $b \in W \subset \mathbb{R}^n$, so dass $f|_V\colon V \to W$ bijektiv ist und $g = (f|_V)^{-1}\colon W \to V \subset \mathbb{R}^n$ ebenfalls k-mal stetig diffbar ist mit Jacobimatrix $Dg(b) = (Df(a))^{-1}$.*

Zum Beweis werden wir einen Fixpunktsatz verwenden. Dazu benötigen wir zunächst den Begriff des Fixpunktes:

Definition 31.13. *Sei $\varphi\colon M \to M$ eine Abbildung. $\xi \in M$ ist ein **Fixpunkt** von φ, wenn $\xi = \varphi(\xi)$.*

Über Fixpunkte gibt es viele interessante Sätze, beispielsweise gilt:

Satz 31.14 (von Brouwer). *Sei $\varphi\colon \overline{B_1(0)} \to \overline{B_1(0)}$ eine stetige Abbildung der abgeschlossenen Kugel $\overline{B_1(0)} \subset \mathbb{R}^n$ in sich selbst. Dann hat φ einen Fixpunkt.*

Der Satz hat viele interessante Folgerungen, wie beispielsweise den **Satz vom Igel**, der im Wesentlichen aussagt, dass ein Igel, der sich zu einer Kugel zusammengerollt hat, nicht überall in die gleiche Richtung gekämmt sein kann. Wir können den Satz von Brouwer hier zwar leider nicht beweisen (siehe dazu beispielsweise [Kö02]); für gewisse Arten von Abbildungen ist ein Nachweis aber nicht allzu schwierig.

In Dimension 1 ist der Satz von Brouwer aber sogar anschaulich sehr einsichtig: Für Fixpunkte x von $\varphi\colon [-1,1] \to [-1,1]$ gilt nämlich, dass $(x, \varphi(x))$ auf der ersten Winkelhalbierende liegt. Ist $\varphi(-1) \neq -1$ und $\varphi(1) \neq 1$, dann gibt es wegen des Zwischenwertsatzes eine Stelle x zwischen -1 und 1, für die $(x, \varphi(x))$ auf der Winkelhalbierenden liegt (Abb. 31.12).

Für kontrahierende Abbildung ist der Fixpunkt sogar eindeutig, wie wir gleich sehen werden (Satz 31.16):

Definition 31.15. *Sei $\overline{B_r} \subset \mathbb{R}^n$ eine abgeschlossene Kugel. Eine Abbildung $\varphi\colon \overline{B_r} \to \overline{B_r}$ heißt **kontrahierend**, wenn es ein λ mit $0 \leq \lambda < 1$ gibt, so dass $\|\varphi(x) - \varphi(y)\| \leq \lambda\|x - y\|$ $\forall x, y \in \overline{B_r}$.*

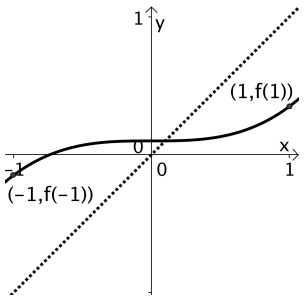


Abbildung 31.12. Satz von Brouwer in Dimension 1.

Man kann leicht nachweisen, dass kontrahierende Abbildungen stetig sind. Wie schon angedeutet, können wir hier für solche Abbildungen einen Fixpunktsatz beweisen; außerdem ist der Fixpunkt in diesem Fall sogar eindeutig und wir können explizit angeben, wie wir diesen Fixpunkt erhalten. Dieser speziellere Fixpunktsatz wird reichen, um den Umkehrsatz zu beweisen.

Satz 31.16 (Banachscher Fixpunktsatz). *Seien $\overline{B_r} \subset \mathbb{R}^n$ eine abgeschlossene Kugel und $\varphi\colon \overline{B_r} \to \overline{B_r}$ eine kontrahierende Abbildung. Dann hat φ genau einen Fixpunkt ξ. Genauer gilt: Für jeden Startpunkt $x_0 \in \overline{B_r}$ konvergiert die Folge $(x_k)_{k\in\mathbb{N}}$ mit $x_{k+1} = \varphi(x_k)$ gegen ξ.*

Beweis. Wegen der Kontraktionseigenschaft gibt es ein $0 \le \lambda < 1$ mit:

$$\|x_m - x_n\| \le \lambda\cdot\|x_{m-1} - x_{n-1}\| \le \lambda^n\cdot\|x_{m-n} - x_0\| \le \lambda^n\cdot 2r \;\forall m \ge n.$$

Die Folge (x_k) ist also eine Cauchy–Folge und konvergiert daher. Für den Grenzwert $\xi := \lim_{k\to\infty} x_k \in \overline{B_r}$ gilt mit dem Folgenkriterium für Stetigkeit (Satz 8.5):

$$\varphi(\xi) = \varphi(\lim_{k\to\infty} x_k) \overset{\varphi \text{ stetig}}{=} \lim_{k\to\infty} \varphi(x_k) = \lim_{k\to\infty} x_{k+1} = \xi.$$

Also ist ξ ein Fixpunkt von φ. Es ist der einzige, da für jeden weiteren Punkt $\eta \in \overline{B_r}$ gilt:

$$\|\varphi(\eta) - \xi\| = \|\varphi(\eta) - \varphi(\xi)\| \le \lambda \cdot \|\eta - \xi\|.$$

Wenn η ebenfalls ein Fixpunkt ist, folgt: $\|\eta - \xi\| \le \lambda \cdot \|\eta - \xi\|$. Dies ist aber nur für $\eta = \xi$ möglich, da $\lambda < 1$. □

Nach diesen Vorbereitungen nun zum Beweis des Umkehrsatzes:

Beweis (des Umkehrsatzes 31.12). Sei $f\colon U \to \mathbb{R}^n$ wie im Satz. Wir dürfen $a = 0$ und $b = f(a) = 0$ annehmen (sonst betrachten wir $\tilde{f}(x) = f(a+x) - f(a)$). Ferner sei $L = (Df(0))^{-1}\colon \mathbb{R}^n \to \mathbb{R}^n$. Dann gilt für $L \circ f\colon D(L \circ f)(0) = E_n$.

Ohne Einschränkung sei also $a = 0 = b$, $Df(0) = E$. Wir wollen die Gleichung $y = f(x)$ für y nahe $b = 0$ mit Hilfe einer Fixpunktabbildung lösen. Dazu betrachten wir $\varphi_y(x) := y + x - f(x)$. Ist ξ ein Fixpunkt von φ_y, dann gilt: $\xi = \varphi_y(\xi) = y + \xi - f(\xi)$, also $y = f(\xi)$. Siehe auch Abb. 31.13.

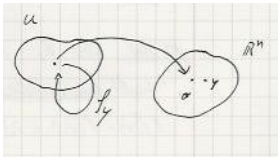


Abbildung 31.13. Zum Beweis des Umkehrsatzes.

Wir müssen zunächst den Definitionsbereich von φ_y festlegen. Dazu wählen wir $r > 0$, so dass

$$\|D\varphi_y(x)\| = \|E_n - Df(x)\| \leq \frac{1}{2} \quad \forall x \in \overline{B_{2r}(0)},$$

wobei $\|A\|$ wieder die Matrixnorm

$$\|A\| = \sup\{\, \|Av\| \mid v \in \mathbb{R}^n \text{ mit } \|v\| = 1 \,\}$$

bezeichnet. So ein r existiert, da f stetig diffbar ist und $Df(0) = E_n$.

Nach der Taylorformel 30.34 existiert daher ein $\vartheta \in [0,1]$, so dass

$$\begin{aligned} \|\varphi_y(x_2) - \varphi_y(x_1)\| &\leq \|(E_n - Df)(x_2 + \vartheta(x_1 - x_2))\| \cdot \|x_1 - x_2\| \\ &\leq \tfrac{1}{2} \cdot \|x_1 - x_2\| \ \forall x_1, x_2 \in \overline{B_{2r}(0)}. \end{aligned} \tag{31.1}$$

Damit gilt für y mit $\|y\| \leq r$, weil $\varphi_y(0) = y$ ist:

$$\begin{aligned} \|\varphi_y(x)\| &= \|\varphi_y(x) - \varphi_y(0) + y\| \\ &\leq \|\varphi_y(x) - \varphi_y(0)\| + \|y\| \leq \tfrac{1}{2}\|x\| + r \leq r + r \\ &= 2r \ \forall x \in \overline{B_{2r}(0)}. \end{aligned} \tag{31.2}$$

Also: $\varphi_y \colon \overline{B_{2r}(0)} \to \overline{B_{2r}(0)}$ nach (31.2). Nach (31.1) ist φ_y außerdem kontrahierend. Nach dem Banachschen Fixpunktsatz 31.16 hat φ_y also für jedes $y \in \overline{B_r(0)}$ genau einen Fixpunkt

$$\xi(y) \in \overline{B_{2r}(0)}$$

und für diesen gilt: $f(\xi(y)) = y$. Dieses $\xi(y)$ ist also der einzige Urbildpunkt von y in $\overline{B_{2r}(0)}$. Wir setzen $W := \overline{B_r(0)}$ und $V := f^{-1}(W) \cap \overline{B_{2r}(0)}$ und definieren $g \colon W \to V$ durch $y \mapsto$ Fixpunkt $\xi(y)$ von φ_y.

Es bleibt zu zeigen, dass g stetig und diffbar ist. Zur Stetigkeit: Seien $y_1, y_2 \in W$ und $x_1 = g(y_1), x_2 = g(y_2)$. Dann ist: $x_2 - x_1 = \varphi_0(x_2) - \varphi_0(x_1) + f(x_2) - f(x_1)$, also, mit der Dreiecksungleichung und (31.1):

$$\begin{aligned}\|x_2 - x_1\| &\leq \|\varphi_0(x_2) - \varphi_0(x_1)\| + \|f(x_2) - f(x_1)\| \\ &\leq \frac{1}{2}\|x_2 - x_1\| + \|f(x_2) - f(x_1)\|.\end{aligned}$$

Es folgt $\|x_2 - x_1\| \leq 2 \cdot \|f(x_2) - f(x_1)\| = 2 \cdot \|y_2 - y_1\|$ und daher $\|g(y_2) - g(y_1)\| = \|x_2 - x_1\| \leq 2 \cdot \|y_2 - y_1\|$. Die Abbildung g ist also stetig.

Nun zur Differenzierbarkeit von g. Zunächst einmal ist $Df(x)$ invertierbar $\forall x \in V$, denn für $v \in \mathbb{R}^n$ gilt:

$$\|(E_n - Df(x)) \cdot v\| \leq \|E_n - Df(x)\| \cdot \|v\| \leq \frac{1}{2} \cdot \|v\|.$$

Andererseits ist $Df(x) \cdot v = 0$; wir haben demnach:

$$\|v\| = \|(E_n - Df(x)) \cdot v\| \leq \frac{1}{2} \cdot \|v\| \Rightarrow v = 0.$$

$Df(x)$ hat also als Kern nur $\{0\}$ und es folgt, dass $Df(x)$ invertierbar ist. Differenzierbarkeit von f in x_0 bedeutet aber:

$$f(x) - f(x_0) = Df(x_0) \cdot (x - x_0) + o(\|x - x_0\|).$$

Dies zeigt:

$$(Df(x_0))^{-1}(f(x) - f(x_0)) = x - x_0 + \underbrace{(Df(x_0))^{-1} o(\|x - x_0\|)}_{o(\|x-x_0\|)=o(\|y-y_0\|)}.$$

Mit $f(x) = y$ und $f(x_0) = y_0$ folgt:

$$g(y) - g(y_0) = (Df(x_0))^{-1}(y - y_0) + \underbrace{o(2\|y - y_0\|)}_{o(\|y-y_0\|}.$$

g ist also differenzierbar und

$$Dg(y_0) = (Df(g(y_0)))^{-1}.$$

Höhere Ablleitungen folgen mit der Kettenregel induktiv. □

Beispiel 31.17. Wir betrachten den Durchschnitt zweier Zylinder: $x^2 + z^2 = 1$, $y^2 + (z-1)^2 = 1$. Können wir diese Kurve, wenigstens nahe dem Punkt $a = (0, 1, 1)$, als eine Funktion von x darstellen? Siehe dazu Abbildung 31.14 und eine Animation auf unserer Webseite: GIF-Format, SWF-Format.

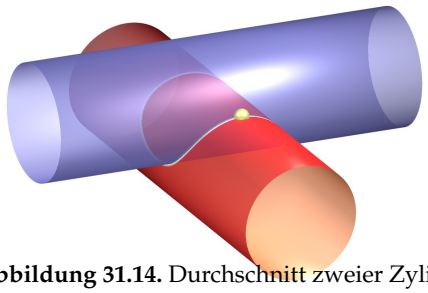


Abbildung 31.14. Durchschnitt zweier Zylinder.

Satz 31.18 (über implizite Funktionen, allgemeiner Fall). *Sei $U \subset \mathbb{R}^n$, $f = (f_1, \ldots, f_m) \colon U \to \mathbb{R}^m, m < n$, eine k-mal stetig diffbare Abbildung und $a \in U$ ein Punkt mit $f(a) = 0$. Angenommen, der letzte Minor der Jacobimatrix erfüllt*

$$\det \begin{pmatrix} \frac{\partial f_1}{\partial x_{n-m+1}} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_{n-m+1}} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} (a) \neq 0.$$

Dann existiert für $a = (a', a'') \in \mathbb{R}^{n-m} \times \mathbb{R}^m$ eine Umgebung $a \in V' \times V'' \subset U \subset \mathbb{R}^{n-m} \times \mathbb{R}^m$ und eine k-mal stetig diffbare Abbildung $g \colon V' \to V''$ mit $g(a') = a''$, so dass

1. $f(x', g(x')) = 0$,
2. $\forall (x', x'') \in V' \times V''$ *mit* $f(x', x'') = 0$ *gilt:* $g(x') = x''$.

Bevor wir dies zeigen, zunächst zurück zum obigen Beispiel:

Beispiel 31.19. Entsprechend der Problemstellung in 31.17 setzen wir

$$f \colon \mathbb{R}^3 \to \mathbb{R}^2, (x, y, z) \mapsto f(x, y, z) = \begin{pmatrix} x^2 + z^2 - 1 \\ y^2 + z^2 - 2z \end{pmatrix}.$$

Für Punkte (x, y, z) auf der Schnittkurve ist dann $f(x, y, z) = (0, 0)^t$. Dann ist

$$Df = \begin{pmatrix} 2x & 0 & 2z \\ 0 & 2y & 2z - 2 \end{pmatrix}, \text{ also } Df(0, 1, 1) = \begin{pmatrix} 0 & \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix} \end{pmatrix}.$$

Für den letzten Minor gilt demnach: $\det(.) \neq 0$. Der Satz über implizite Funktionen liefert nun die Existenz einer Abbildung: $x \mapsto (y(x), z(x)) = g(x)$. Tatsächlich gilt:

$$z(x) = \sqrt{1 - x^2}, \quad y(x) = \sqrt{1 - (\sqrt{1 - x^2} - 1)^2}.$$

Wir sehen am obigen Beispiel, dass der Satz zwar die Existenz sichert, dass die konkrete Berechnung der Abbildung aber schwierig werden kann. Betrachten wir noch ein weiteres Beispiel:

Beispiel 31.20. Wir möchten die Schnittkurve von $x^2 + y^2 + z^2 = 1$ (Kugel) und $(x - \frac{1}{2})^2 + y^2 = (\frac{1}{2})^2$ (Zylinder) im Punkt $a = (1, 0, 0)$ untersuchen, s. Abb. 31.15 und die Animationen auf unserer Webseite: GIF-Format, SWF-Format. Wenn wir f analog zum vorigen Beispiel aufstellen, erhalten wir:

$$Df = \begin{pmatrix} 2x & 2y & 2z \\ 2(x - \frac{1}{2}) & 2y & 0 \end{pmatrix}, \text{ also } Df(1,0,0) = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Alle Minoren sind demnach null und wir können den Satz nicht anwenden.

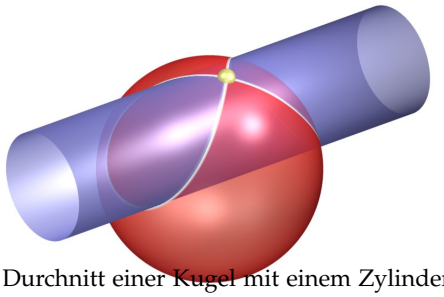


Abbildung 31.15. Durchnitt einer Kugel mit einem Zylinder. Der Punkt $a = (1, 0, 0)$, in dem wir die Schnittkurve (weiß) untersuchen möchten, ist gelb markiert.

Dies liegt an der speziellen Wahl des Punktes (dies ist nämlich ein singulärer Punkt der Kurve, genauer gesagt ist es ein Punkt, in dem sich die Kurve selbst schneidet). Beispielsweise ist aber

$$Df(0,0,1) = \begin{pmatrix} 0 & 0 & 2 \\ -1 & 0 & 0 \end{pmatrix},$$

so dass wir mit dem Satz nach Umnummerieren der Variablen tatsächlich die Existenz von Abbildungen $x(y)$ und $z(y)$ in der Umgebung von $(0, 0, 1)$ garantiert bekommen.

Nun schließlich zum Beweis der allgemeinen Version des Satzes über implizite Funktionen, der dann auch den oben nicht ausgeführten Beweis der anderen Variante (31.7) dieses Satzes liefert. Mit dem Umkehrsatz ist dieser Nachweis nun nicht mehr viel Arbeit:

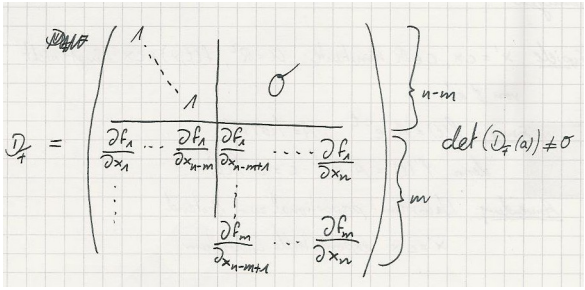


Abbildung 31.16. Eine Anwendung des Umkehrsatzes.

Beweis (des Satzes 31.18 über implizite Funktionen). Sei $f = (f_1, \ldots, f_m)$ wie im Satz. Wir betrachten die Abbildung $F(x) = x_1, \ldots, x_{n-m}, f_1(x), \ldots, f_m(x))$, $F\colon U \to \mathbb{R}^n$. Dann ist DF wie in Abb. 31.16 angegeben.

Wir wenden den Umkehrsatz an (s. Abb. 31.16) und erhalten: $\exists W = V' \times V''$, eine Umgebung von $(a', 0)$, und $G\colon W \to U$ mit $F \circ G = \mathrm{id}_W$. Es folgt:

$$G(x_1, \ldots, x_{n-m}, y_1, \ldots, y_m) = (x_1, \ldots, x_{n-m}, G_{n-m+1}(x', y), \ldots, G_n(x', y)),$$

da $F \circ G = \mathrm{id}_W$. Die Abbildung $g\colon V' \to V''$, definiert durch $g(x') = (G_{n-m+1}(x', 0), \ldots, G_n(x', 0))$, erfüllt dann: $f(x', g(x')) = 0$, da $F \circ G = \mathrm{id}_W$. □

Damit haben wir zwar alle Aussagen dieses Abschnittes bewiesen, doch leider mussten wir auch feststellen, dass die abstrakten Existenzaussagen schon in einfach erscheinenden Beispielen auf schwierige Rechnungen führen, wenn wir uns nicht auf eine reine Existenzaussage beschränken, sondern explizite Ergebnisse erhalten möchten. Daran können wir im Rahmen dieser Veranstaltung nichts ändern. In nicht zu komplizierten Beispielen ist es aber doch noch recht häufig möglich, mit Hilfe von Computeralgebra Software konkrete Formeln oder zumindest gute Annäherungen zu produzieren.

Aufgaben

Aufgabe 31.1 (Lokales Auflösen). Zeigen Sie, dass sich

$$f\colon \mathbb{R}^2 \to \mathbb{R},\ f(x, y) = 1 + x + xy - e^y$$

in einer Umgebung von $(0, 0)$ lokal nach y auflösen lässt, und berechnen Sie die Taylorreihe der Auflösung $y = g(x)$ bis zum Grad 2.

Aufgabe 31.2 (Extremwerte unter Nebenbedingungen). Welcher Punkt der Fläche $z = x^2 + y^2$ liegt dem Punkt $p := \begin{pmatrix} 1 \\ 1 \\ \frac{1}{2} \end{pmatrix}$ am nächsten?

Aufgabe 31.3 (Extremwerte unter Nebenbedingungen). Bestimme die lokalen und globalen Extrema der Funktion $f(x, y) = xy^2$ auf dem Kreis $x^2 + y^2 = 1$ mit Hilfe Lagrangescher Multiplikatoren.

Aufgabe 31.4. Wo ist die Kugelkoordinaten-Abbildung

$$\Psi\colon \mathbb{R}^3 \to \mathbb{R}^3,\ \Psi(r, \varphi, \theta) == (rcos\varphi cos\theta, rsin\varphi cos\theta, rsin\theta)$$

nicht lokal umkehrbar?

Aufgabe 31.5 (Banachscher Fixpunktsatz). Zeigen Sie mit Hilfe des Banachschen Fixpunktsatzes, dass die Abbildung

$$f : [1, 2] \to \mathbb{R}, x \mapsto \frac{1}{3}\sqrt{x} - \frac{1}{24}x^3 - x + 1$$

genau eine Nullstelle besitzt. Bestimmen Sie diese näherungsweise mit Hilfe von Maple.

Aufgabe 31.6 (Brouwerscher Fixpunktsatz). Zeigen Sie, dass der Brouwersche Fixpunktsatz für stetige Abbildungen

$$f : B_1(0) \to B_1(0)$$

mit $B_1(0) =]-1, 1[\subset \mathbb{R}$, die also nur auf dem *Inneren* der Einheitskugel definiert sind, im Allgemeinen falsch ist.

32

Ein Blick auf Differentialgleichungen

Die meisten physikalischen Naturgesetze lassen sich in der Form *Es ist eine bestimmte Differentialgleichung erfüllt* formulieren. In der Informatik sind DGLs nicht so wichtig, außer etwa in der Bildverarbeitung. Daher geben wir hier nur einen sehr knappen Überblick.

32.1 Gewöhnliche Differentialgleichungen erster Ordnung

Definition 32.1. *Eine **DGL** (**Differentialgleichung**, genauer: **gewöhnliche Differentialgleichung**) erster Ordnung ist folgendermaßen gegeben:* $U \subseteq \mathbb{R}^2$ *offen,* $f\colon U \to \mathbb{R}$, $x' = f(t,x)$ *(wobei hier* $x = x(t)$ *und* $x' = x'(t)$ *die Ableitung nach* t *bezeichnet). Eine Lösung von* $x' = f(t,x)$ *ist eine diffbare Funktion* $\varphi\colon I \to \mathbb{R}$ *mit*

1. $(t, \varphi(t)) \in U \; \forall t \in I$,
2. $\varphi'(t) = f(t, \varphi(t)) \; \forall t \in I$.

*Oft werden Lösungen gesucht, die einer **Anfangsbedingung** $\varphi(t_0) = x_0$ genügen.*

Beispiel 32.2. Wir geben zunächst einige Fälle an, in denen wir die Lösung direkt angeben können:

Exponentielles Wachstum: Wir betrachten die Gleichung: $x' = cx$, wobei $c \in \mathbb{R}$ eine Konstante ist, $U \subseteq \mathbb{R}^2$, $f(t,x) = cx$ hängt nicht von t ab.

Lösung: $\varphi(t) = A \cdot e^{ct}$, $A = \varphi(0) \in \mathbb{R}$, denn $\varphi'(t) = A \cdot e^{ct} \cdot c = c \cdot \varphi(t)$.

Anwendung: Wachstum proportional zum Bestand, z.B. für jeweils zwei Kaninchen einer Population kommen vier in der nächsten Generation hinzu.

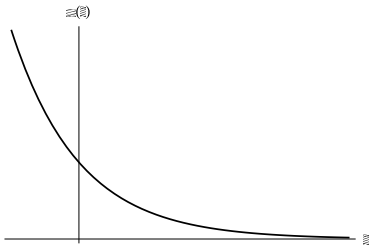


Abbildung 32.1. Radioaktiver Zerfall.

Radioaktiver Zerfall: Strahlung proportional zur radioaktiven Masse (s. Abb. 32.1). $x' = -cx$, $c > 0$. $x(t) = A_0 \cdot e^{-c(t-t_0)}$, $x(t_0) = A_0$ (Masse zum Zeitpunkt t_0). Einfacher: $A_0 \cdot e^{-ct} = x(t)$. Dann heißt t_h = **Halbwertszeit**, definiert durch

$$x(t_h) = \frac{1}{2}x(0) \Leftrightarrow A_0 \cdot e^{-ct_h} = \frac{1}{2}A_0 \Leftrightarrow \frac{1}{2} = e^{-ct_h} \Leftrightarrow ct_h = \ln 2,\ t_h = \frac{\ln 2}{c}.$$

Trennung der Variablen: Wir betrachten die DGL $x' = x \cdot t$.

Lösung (**Trennung der Variablen**): Mit $\frac{dx}{dt} = x' = x \cdot t$ erhalten wir formal $t \cdot dt = \frac{dx}{x}$, d.h. Integrieren liefert:

$$\int t\,dt + c = \int \frac{1}{x}\,dx$$

für eine gewisse Konstante $c \in \mathbb{R}$. Es folgt:

$$\frac{1}{2}t^2 + c = \ln|x|.$$

Dies liefert $|x| = e^{\frac{1}{2}t^2+c}$. Eine Lösung unseres Problems ist also $\varphi(t) = c' \cdot e^{\frac{1}{2}t^2}$ für eine gewisse Konstante $c' > 0$. Tatsächlich erfüllt dies die ursprüngliche Gleichung, denn:

$$\varphi'(t) = c' \cdot e^{\frac{1}{2}t^2} \cdot t = \varphi(t) \cdot t.$$

Das angegebene Verfahren haben wir freilich nicht sauber untermauert. Tatsächlich lässt es sich aber präzisieren, was wir hier aus Zeitgründen aber unterlassen werden.

Im Allgemeinen hat eine Differentialgleichung in getrennten Variablen die Gestalt

$$x' = g(t) \cdot h(x);$$

die rechte Seite lässt sich also in Produktform schreiben, wobei der eine Faktor nur von t und der andere nur von x abhängt. Zur Lösung formt man diese in $\frac{x'}{h(x)} = g(t)$ um und findet die Lösung durch Integration beider Seiten.

Autonome DGLs: Man kann diese Methode auch anwenden, wenn die rechte Seite gar nicht von t abhängt. In diesem Fall heißt die DGL **autonom**.

Ein Beispiel ist die sogenannte **Explosionsgleichung** $x' = x^2$: Es folgt $\frac{dx}{dt} = x^2$, d.h. formal $\frac{dx}{x^2} = dt$. Integrieren ergibt:

$$\int \frac{1}{x^2}\, dx = \int dt \; - c$$

für eine gewisse Konstante $c \in \mathbb{R}$, d.h. $-\frac{1}{x} = t - c$. Wir erhalten $\varphi(t) = \frac{1}{c-t}$. In diesem Fall ist die Lösung nur auf einem endlichen Zeitintervall $[t_0, c[$ gegeben.

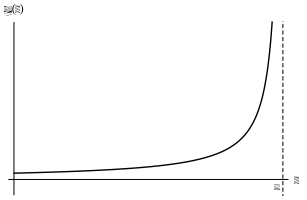


Abbildung 32.2. Die Explosionsgleichung.

Definition 32.3. *Seien $f\colon U \to \mathbb{R}, U \subset \mathbb{R}^2$ und $x' = f(t,x)$ eine DGL. Wir ordnen jedem Punkt (t,x) den Vektor $(1,x') = (1, f(t,x)) \in \mathbb{R}^2$ zu. Dies heißt **Richtungsfeld**. Zur Veranschaulichung zeichnet man die Vektoren, die ja die Steigung einer Lösung $x(t)$ in einem gegebenen Punkt angeben, als Pfeile an die Punkte ein (s. Abb. 32.3). Eine Lösung $x(t)$ für einen gegebenen Anfangswert folgt dann den Pfeilen.*

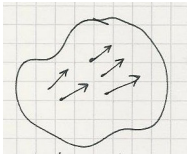


Abbildung 32.3. Ein Richtungsfeld.

Beispiel 32.4. Die logistische Gleichung $x' = x(1-x)$ (Abb. 32.4). Es gilt:

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

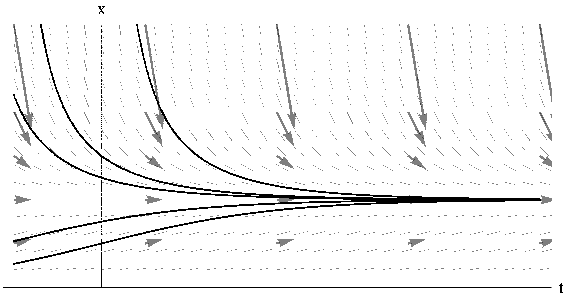


Abbildung 32.4. Das Richtungsfeld der Logistischen Gleichung und einige Lösungen für verschiedene Anfangswerte.

Daher ist, weil $\int dt + c = t + c$: $\frac{x}{x-1} = Ae^t$ mit $A = e^c$. Nach Umformung folgt: $x(t) = \frac{Ae^t}{Ae^t-1} \underset{t\to\infty}{\to} 1$.

Die konstante Funktion $\varphi(t) \equiv 1$ heißt **Gleichgewichtslösung**. Im Allgemeinen heißen so die Nullstellen von $f(x)$, weil an diesen Stellen die Ableitung x' verschwindet.

32.2 Gewöhnliche Differentialgleichungen höherer Ordnung

Definition 32.5 (DGL höherer Ordnung). *Sei $U \subseteq \mathbb{R} \times \mathbb{R}^n$ offen, $f\colon U \to \mathbb{R}$. Mit $x^{(i)}$ bezeichnen wir die i-te Ableitung einer Funktion $\mathbb{R} \to \mathbb{R}$. Dann heißt*

$$x^{(n)} = f(t, x, x', \ldots, x^{(n-1)})$$

*eine **gewöhnliche DGL n-ter Ordnung**. Eine Lösung ist eine n-mal diffbare Funktion $\varphi\colon I \to \mathbb{R}$ mit*

1. $(t, \varphi(t), \varphi'(t), \ldots, \varphi^{(n-1)}(t)) \in U \; \forall t \in I$,
2. $\varphi^{(n)}(t) = f(t, \varphi(t), \ldots, \varphi^{(n-1)}(t))$.

*Ein **Differentialgleichungssystem 1-ter Ordnung** ist gegeben durch $U \subseteq \mathbb{R} \times \mathbb{R}^n$, $f\colon U \to \mathbb{R}$,*

$$\begin{aligned} x_1' &= f_1(t, x) \\ &\vdots \quad \vdots \\ x_n' &= f_n(t, x). \end{aligned}$$

Lösungen sind $\varphi\colon I \to \mathbb{R}^n$, so dass $\varphi_k'(t) = f_k(t, \varphi(t)), k = 1, \ldots, n$.

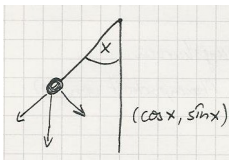


Abbildung 32.5. Das mathematische Pendel.

Beispiel 32.6. Das **Mathematische Pendel**: $x'' = -\sin x$.

Der **Harmonische Oszillator**: $x'' = -x$. Lösung: $x(t) = a\cos t + b\sin t$.

Bemerkung 32.7. Jede DGL n-ter Ordnung ist äquivalent zu einem DGL-System erster Ordnung, nämlich:

$$x^{(n)} = f(t, x, \dots, x^{(n-1)}) \iff x_2 = x_1', \ \dots, \ x_n = x_{n-1}', \ x_n' = f(t, x_1, \dots, x_n).$$

Definition 32.8. *Sei*

$$x' = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = f(x)$$

eine autonome DGL. Dann nennen wir die Menge der Lösungskurven $t \mapsto (x_1(t), \dots, x_n(t))$ *das* ***Phasenportrait*** *der DGL.*

Beispiel 32.9. Räuber-Beute-Modell; s. Lotka-Volterra Modell: Ist x = eine Population von Karpfen, y = Population von Hechten, so beschreiben folgende Gleichungen die Entwicklungen der Populationen näherungsweise:

$$x' = kx - axy, \quad y' = -ly + bxy.$$

Die nichttriviale Gleichgewichtslösung (d.h. $x \not\equiv 0 \not\equiv y$) erfüllt $x' = kx - axy = 0$, $y' = -ly + bxy = 0$, d.h. $y(t) = \frac{k}{a}$ und $x(t) = \frac{l}{b}$. Man kann zeigen, dass die anderen Lösungen konzentrische Kreise um diesen Punkt, der die Gleichgewichtslösung darstellt, beschreiben; s. Abb. 32.6.

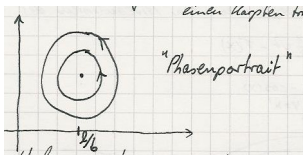


Abbildung 32.6. Das Phasenportrait des Räuber-Beute-Modells.

DGLs vom Typ $x'' = f(x)$: Das Phasenportrait besteht aus Kurven in der (x, x')-Ebene.

Aus der Physik motiviert, setzen wir:

$$E_{kin} := \frac{1}{2}(x')^2, \qquad E_{pot} := -F(x),$$

wobei $F(x)$ eine Stammfunktion von $f(x)$ ist, also $F(x) = \int f(x)\, dx$ und somit $F'(x(t)) = f(x(t)) \cdot x'(t)$.

Mit dieser Notation folgt, dass $E_{tot} := E_{kin} + E_{pot}$ tatsächlich konstant ist, wie sich leicht nachrechnen lässt:

$$E'_{tot} = x'(t) \cdot x''(t) - f(x(t)) \cdot x'(t) = x'(t) \cdot \underbrace{(x''(t) - f(x(t)))}_{=0} = 0.$$

Die Gleichung $E_{tot} = c$ für eine Konstante c zeigt, dass jede Lösungskurve der Differentialgleichung, die wir im $(x, y) := (x, x')$-Koordinatensystem einzeichnen wollen, eine Gleichung

$$\frac{1}{2}y^2 - F(x) = c$$

erfüllen muss, also eine Niveaulinie der Gesamtenergie

$$E_{tot}(x, x') = E_{tot}(x, y) = \frac{1}{2}y^2 - F(x)$$

ist.

Das **Mathematische Pendel** ($x'' = -\sin(x)$): Beispielsweise besteht das Phasenportrait des mathematischen Pendels aus den Niveaukurven von: $\frac{1}{2}y^2 - \cos(x)$ (s. Abb. 32.7).

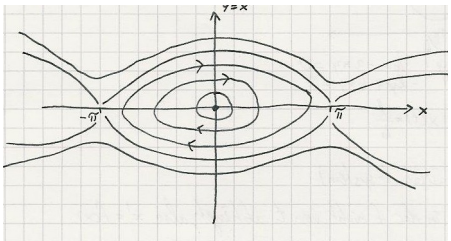


Abbildung 32.7. Das Phasenportrait des mathematischen Pendels.

Wir konnten hier natürlich nur einige wenige Beispiele von DGLs und deren Lösungen vorstellen. Tatsächlich kann man aber in vielen Fällen zeigen, dass Lösungen existieren müssen und sogar eindeutig sind, wenn man den Anfangswert vorgibt:

Satz 32.10 (Existenz und Eindeutigkeit von Lösungen von DGLs). *Sei $U \subseteq \mathbb{R} \times \mathbb{R}^n$ offen, $f\colon U \to \mathbb{R}^n$, $x' = f(t,x)$ ein DGL-System. Ist f stetig partiell diffbar, dann existiert $\forall (t_0, a) \in U$ ein Intervall I mit $t_0 \in I$ und eine Lösung $\varphi\colon I \to \mathbb{R}^n$ mit $\varphi(t_0) = a$ und*

1. *$(t, \varphi(t)) \in U \ \forall t \in I$,*
2. *$\varphi'(t) = f(t, \varphi(t)) \ \forall t \in I$.*

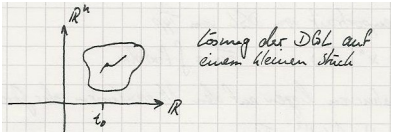


Abbildung 32.8. Skizze einer Lösung einer DGL.

Ferner gilt: Zwei Lösungen $\varphi\colon I \to \mathbb{R}^n$, $\psi\colon J \to \mathbb{R}^n$ mit dem gleichen Anfangswert $\varphi(t_0) = \psi(t_0)$ stimmen auf dem Durchschnitt $I \cap J$ überein. Die Lösung $\varphi = \varphi_{t_0,a}$ hängt stetig von dem Anfangswert t_0, a und den "Koeffizienten" von f ab.

Im Rahmen dieser Veranstaltung können wir leider keinen Beweis hierfür geben. Für die Existenz reicht Stetigkeit (Satz von Peano). Für die Eindeutigkeit nicht, wie das folgende Beispiel zeigt:

Beispiel 32.11. Für die in 0 nicht diffbare Funktion $f(x) = 3x^{2/3}$ hat die DGL $x' = 3x^{2/3}$ als Lösung z.B. $x(t) = t^3$. Aber auch (wie man leicht nachrechnen kann):

$$\varphi(t) = \begin{cases} (t-t_1)^3, & t \le t_1 \\ 0, & t_1 \le t \le t_2 \\ (t-t_2)^3, & t_2 \le t. \end{cases}$$

$f(x) = 3x^{2/3}$ hat partielle Ableitung $\frac{\partial f}{\partial x} = 2x^{-1/3}$; diese hat einen Pol bei $x = 0$.

32.3 Partielle DGL

Partielle DGLs (engl.: **PDE**) beschreiben Funktionen durch Bedingungen an die partiellen Ableitungen. Im Folgenden betrachten wir Abbildungen der Form $u\colon G \to \mathbb{R}$, wobei $G \subset \mathbb{R}^n$ ein **Gebiet** ist, d.h. eine zusammenhängende offene Menge.

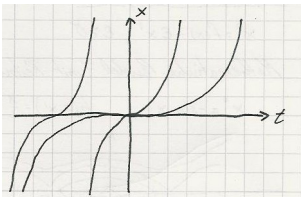


Abbildung 32.9. Skizze von Lösungen einer DGL.

32.3.1 Die Laplacegleichung bzw. die Potentialgleichung

Es seien $G \subset \mathbb{R}^n$ ein Gebiet und $u\colon \bar{G} \to \mathbb{R}$ zweimal stetig differenzierbar. Wir betrachten

$$\frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} = 0$$

auf G. Mit dem **Laplace–Operator**

$$\Delta\colon C^\infty(G,\mathbb{R}) \to C^\infty(G,\mathbb{R}), \quad \Delta = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$$

schreibt sich dies sehr kurz: $\Delta u = 0$. Solche u, die diese **Laplacegleichung** (bzw. **Potentialgleichung**) erfüllen, heißen **harmonisch**.

Die Webseite http://abel.math.upb.de/Beispiele/01/ zeigt ein nettes bebildertes Beispiel hierzu.

32.3.2 Die Wellengleichung

Wir betrachten $\mathbb{R} \times G \subset \mathbb{R} \times \mathbb{R}^n$ mit den Koordinaten $t, x_1, \ldots, x_n$. Mit der Notation $\Delta_{xx} = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$ heißt für $u\colon \mathbb{R} \times G \to \mathbb{R}$ die Gleichung

$$\frac{\partial^2 u}{\partial t^2} = \Delta_{xx} u$$

Wellengleichung. Siehe Abb. 32.10 und die schon eben zitierte Webseite http://abel.math.upb.de/Beispiele/01/. t wird meist als Zeit interpretiert; die Webseite zeigt dementsprechend auch eine Animation.

32.3.3 Wärmeleitungsgleichung bzw. Diffusionsgleichung

Für $G \subset \mathbb{R}^n$ und $\mathbb{R} \times G$ mit Koordinaten t, x bezeichnen wir eine Gleichung der Form

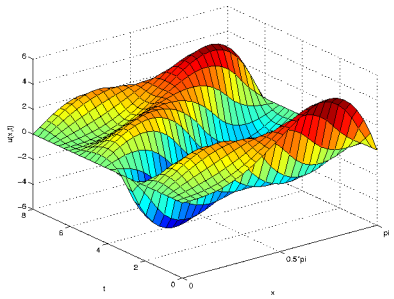


Abbildung 32.10. Skizze zur Wellengleichung.

$$\frac{\partial u}{\partial t} = \Delta_{xx}$$

als **Wärmeleitungsgleichung** oder **Diffusionsgleichung**. Für Illustrationen hierzu siehe wiederum http://abel.math.upb.de/Beispiele/01/. Neben der Modellierung von Wärmeleitung kann man die Gleichung auch verwenden, um andere Ausgleichsprozesse, wie Diffusionsprozesse, zu beschreiben. u ist hierbei die Konzentration bzw. Temperatur und t die Zeit.

In der Bildverarbeitung verwendet man Diffusionsprozesse zum **Entrauschen** (**Diffusionsfilter**). u beschreibt den Grauwert und die Diffusionszeit t ist ein Maß für die Glättung.

Bemerkung 32.12. Alle in diesem Abschnitt vorgestellten partiellen Differentialgleichungen sind durch Linearkombinationen von **Differentialoperatoren** zweiter Ordnung $\partial^2 u \;/\; \partial x_i \partial x_j$ beschreibbar. Analog zur Klassifikation der Quadriken kann man diese dann in entartetete und nichtentartete einteilen und, beispielsweise im Fall von zwei Variablen, die nichtentarteten in **elliptische** (Poissiongleichung bzw. Laplacegleichung), **parabolische** (Wäremeleitungsgleichung bzw. Diffusionsgleichung) und **hyperbolische** (Wellengleichung) Differentialgleichungen einteilen.

Aufgaben

Aufgabe 32.1 (Trennung der Variablen). Seien I ein Intervall, $f_1, f_2 : I \to \mathbb{R}$ stetig und f_2 ohne Nullstellen. Dann heißt

$$y' = \frac{f_1(x)}{f_2(y)}$$

Differentialgleichung mit getrennten Variablen. Schreiben wir formal nun $y' = \frac{dy}{dx}$, so erhalten wir die Gleichung $f_2(y)dy = f_1(x)dx$. Seien F_i Stammfunktionen von $f_i, i = 1, 2$. Dann liefert die Gleichung $F_2(y) = F_1(x) + C$ eine Lösung $y = F_2^{-1}(F_1(x) + C)$ der Differentialgleichung.

1. Zeigen Sie, dass dies tatsächlich eine Lösung der obigen Differentialgleichung ist.
2. Lösen Sie mit dieser Methode die Differentialgleichung $x + yy' = 0$ im Allgemeinen.
3. Wir fordern nun zusätzlich die sogenannte Anfangsbedinung $y(0) = 2$. Wie lautet die Lösung der Aufgabe in diesem speziellen Fall?

Aufgabe 32.2 (Phasenportrait). Skizzieren Sie (ggf. mit Hilfe eines Computeralgebrasystems) die Phasenportraits der Differentialgleichungen $x'' = \frac{x^2}{2}$ und $x'' = \cos(x)$.

Aufgabe 32.3 (Ein Anfangswertproblem). Lösen Sie das Anfangswertproblem $y' = e^y \sin(x)$ für Anfangswerte $y(0) = y_0 < -\log 2$.

33

Integration im $\mathbb{R}^n$

Zum Abschluss der kurzen Einführung in die mehrdimensionale Analysis gehen wir nun noch auf die Integration ein. Sie ist grundlegend für Vieles im Abschnitt über Wahrscheinlichkeitstheorie und Statistik. Aus Zeitgründen müssen wir leider auch hier manche Resultate ohne Beweis akzeptieren, wie beispielsweise den sogenannten Satz von Fubini.

33.1 Integrale über kompakten Mengen

Bemerkung/Definition 33.1. Seien $K \subset \mathbb{R}^n$ eine kompakte Teilmenge und $f\colon K \to \mathbb{R}$ eine stetige Funktion. Dann lässt sich das Integral von f über K, geschrieben $\int_K f\,dx$, wie folgt definieren. Wir setzen f fort:

$$\tilde{f}\colon \mathbb{R}^n \to \mathbb{R},\ \tilde{f}(x) = \begin{cases} f(x), & x \in K, \\ 0, & x \notin K. \end{cases}$$

Sei $\dot{\bigcup}_{i=1}^{N} Q_i \supseteq K$ eine endliche disjunkte Vereinigung von Quadern, die K überdeckt (s. Abb. 33.1). **Endliche Überdeckungen** von kompakten Teilmengen des $\mathbb{R}^n$ existieren immer; wir werden dies hier nicht formal beweisen, sondern verweisen dazu auf die Literatur, wie beispielsweise [For08b, §3] (siehe Satz von Heine–Borel). Allerdings sollte dies anschaulich nicht wirklich erstaunen, wenn man Abb. 33.1 betrachtet.

Wir setzen:

Abbildung 33.1. Endliche Überdeckung eines Kompaktums durch disjunkte Quader.

$$\int_K^* f\,dx = \inf\Big\{ \underbrace{\sum_{i=1}^{N} \max(f|_{\overline{Q_i}}) \cdot \mathrm{Vol}(Q_i)}_{\text{Obersumme}} \;\Big|\; \bigcup Q_i \text{ überdeckt } K \Big\},$$

$$\int_{K_*} f\,dx = \sup\Big\{ \underbrace{\sum_{i=1}^{N} \min(f|_{\overline{Q_i}}) \cdot \mathrm{Vol}(Q_i)}_{\text{Untersumme}} \;\Big|\; \bigcup Q_i \text{ überdeckt } K \Big\}.$$

Es liegt nahe, dass im Fall stetiger Funktionen f auf einem Kompaktum beide Grenzwerte übereinstimmen. Dies gilt tatsächlich und wir können daher definieren:

$$\int_K^* f(x)\,dx = \int_{K_*} f(x)\,dx =: \int_K f(x)\,dx.$$

Bemerkung 33.2. 1. Der Integrationsbereich K wird bei obigem Prozess ebenfalls approximiert (Quader am Rand von K spielen hierbei eine Rolle).

2. Für den Quader $Q = [a_1, b_1] \times \cdots \times [a_n, b_n]$ gilt: $\mathrm{Vol}(Q) = \Pi_{i=1}^n (b_i - a_i)$

3. $\mathrm{Vol}(K) := \int_K 1\,dx$ ist das **Volumen des Kompaktums** K.

Für praktische Zwecke ist das Folgende oft hilfreich (ohne Beweis):

Satz 33.3 (von Fubini). *Sei $f\colon K \to \mathbb{R}$ stetig, $K \subset \mathbb{R}^n$ kompakt, etwa $K \subset [a,b]^n$. Dann gilt (siehe auch Abb. 33.2):*

$$\int_K f(x)\,dx = \int_K f(x)\,dx_1 \cdots dx_n = \int_a^b \Big(\int_{K \cap \mathbb{R}^{n-1} \times \{x_n\}} f(x)\,dx_1 \cdots dx_{n-1} \Big)\,dx_n.$$

Im inneren Integral ist x_n ja eine feste, konstante Zahl, so dass wir damit das Problem der Integralberechnung um eine Variable reduziert haben. Damit kann man prinzipiell Rechnungen schrittweise bis auf Integrale in einer Variablen zurückführen:

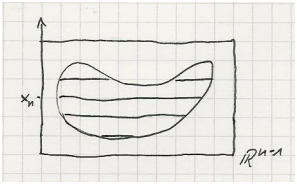


Abbildung 33.2.

Beispiel 33.4. Wir verwenden den Satz, um einige Volumina auszurechnen:

1. Volumen der Kugel

$$K = \overline{B_r(0)} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 \leq r^2\}$$

mit Radius r. Es gilt:

$$\begin{aligned} \mathrm{Vol}(\overline{B_r(0)}) &= \int_K 1 \, dx \, dy \, dz \\ &\overset{\text{Fubini}}{=} \int_{-r}^{r} \Big(\int_{K \cap \mathbb{R}^2 \times \{z\}} dx \, dy \Big) \, dz. \end{aligned}$$

Für jedes feste z gilt aber $x^2 + y^2 = r^2 - z^2$ und dies ist ein Kreis mit Radius $\sqrt{r^2 - z^2}$, dessen Flächeninhalt wir kennen. Also folgt:

$$\begin{aligned} \mathrm{Vol}(\overline{B_r(0)}) &= \int_{-r}^{r} \Big(\pi \cdot (r^2 - z^2) \Big) \, dz \\ &= \pi \cdot \Big(r^2 z - \frac{1}{3} z^3 \Big) \Big|_{-r}^{r} = \pi \cdot \Big(r^3 - \frac{1}{3} r^3 - (-r^3) - \frac{1}{3} r^3 \Big) \\ &= \frac{4}{3} \pi r^3. \end{aligned}$$

2. Wir betrachten nun den halben Paraboloidenstumpf (Abb. 33.3), beschrieben durch $K = \{(x, y, z) \in \mathbb{R}^3 \mid y^2 + z^2 \leq x \leq 1, z \geq 0\}$.

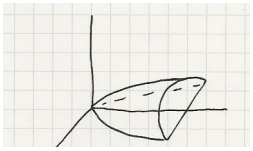


Abbildung 33.3. Skizze zur Volumenberechnung.

Für das Volumen ergibt sich, da offenbar auch $z \leq 1$ ist:

$$\begin{aligned}
\mathrm{Vol}(K) &= \int_K 1 \, dx\, dy\, dz \\
&\overset{\text{Fubini}}{=} \int_0^1 \Big(\int_{-\sqrt{1-z^2}}^{\sqrt{1-z^2}} \Big(\int_{y^2+z^2}^1 1 \, dx \Big) \, dy \Big) \, dz \\
&= \int_0^1 \Big(\int_{-\sqrt{1-z^2}}^{\sqrt{1-z^2}} (1 - y^2 - z^2) \, dy \Big) \, dz \\
&= \int_0^1 \Big(\Big(y - \frac{1}{3} y^3 - y z^2 \Big) \Big|_{-\sqrt{1-z^2}}^{\sqrt{1-z^2}} \Big) \, dz \\
&= 2 \cdot \Big(\int_0^1 \Big((1-z^2)^{3/2} - \frac{1}{3} (1-z^2)^{3/2} \Big) \, dz \Big) \\
&= \dots
\end{aligned}$$

Dies erscheint doch eine etwas komplizierte Rechnung zu werden. Einfacher geht es, wenn wir die Integrationsreihenfolge vertauschen, weil sich für jedes feste x als Schnitt von $\mathbb{R}^2 \times \{x\}$ mit K ein halber Kreis mit Radius $\sqrt{x}$ ergibt:

$$\begin{aligned}
\mathrm{Vol}(K) &= \int_K dy\, dz\, dx = \int_0^1 \Big(\int_{\{(y,z) \in \mathbb{R}^2 \,|\, y^2+z^2 \le x, z \ge 0\}} dy\, dz \Big) \, dx \\
&= \int_0^1 \frac{\pi}{2} (\sqrt{x})^2 \, dx = \frac{\pi}{4} x^2 \Big|_0^1 = \frac{\pi}{4}.
\end{aligned}$$

Analog zur Substitutionsregel in einer Variablen ist die folgende Formel. Die Ableitung wird dabei ersetzt durch die Determinante der Jacobi–Matrix:

Satz 33.5 (Transformationsformel). *Sei $K \subset \mathbb{R}^n$ eine kompakte Teilmenge, $K \subset U$, U offen, $F\colon U \to \mathbb{R}^n$ sei stetig diffbar und auf dem Inneren $\overset{\circ}{K}$ diffbar umkehrbar. Sei ferner $L = F(U)$ und $f\colon L \to \mathbb{R}$ eine stetige Funktion. Dann gilt:*

$$\int_L f(y) \, dy = \int_K \Big(f(F(x)) \cdot \underbrace{|\det DF(x)|}_{\text{Det. d. Jacobi-Matrix}} \Big) \, dx.$$

Beweis (nur Idee). Wir überdecken L mit disjunkten Quadern Q_i: $\dot{\bigcup_{i=1}^r} Q_i \supseteq L$. Dies liefert folgende Approximation:

$$\int_L f(y) \, dy \approx \sum_{i=1}^r \max_{x \in \bar{Q}_i} (f(x)) \cdot \mathrm{Vol}(Q_i).$$

Analog können wir die rechte Seite der Formel approximieren, indem wir K durch disjunkte Quader P_j überdecken:

$$\int_K (f\circ F)(x)\cdot|\det DF(x)|\,dx \approx \sum_{P_j}\max_{x\in\bar{P}_j}\Big((f\circ F(x)\cdot|\det DF(x)|\Big)\cdot\mathrm{Vol}(P_j)$$
$$\approx \sum_{P_j}\max_{x\in\bar{P}_j}(f(x))\cdot|\det DF(p)|\cdot\mathrm{Vol}(P_j),$$

wobei $p\in P_j$ die *linke untere Ecke* ist (f ist nämlich stetig, so dass wir einen beliebigen festen Punkt wählen dürfen, weil die P_j klein sind). Es gilt:

$$|\det DF(p)|\cdot\mathrm{Vol}(P_j) = \mathrm{Vol}(DF(p)\cdot P_j),$$

denn $|\det DF(p)|$ ist das Volumen des Parallelotops, das von den Spaltenvektoren der Matrix $DF(p)$ aufgespannt wird (siehe dazu Abschnitt 22.1). Wenn die Quader P_j gerade die Urbilder der Q_i unter F sind, ist dies aber gerade $\mathrm{Vol}(Q_i)$, so dass sich die Behauptung ergibt. □

Ist F eine diffbare Abbildung einer Teilmenge des $\mathbb{R}^n$ auf eine andere, so ist $|\det DF(x)|$ also der **Volumen–Verzerrungsfaktor** in x.

Beispiel 33.6. Wieder berechnen wir einige Volumina:

1. Sei $E=\{\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}\le 1\}$ ein Ellipsoid (und $a,b,c>0$). E ist das Bild der Einheitskugel unter einer Abbildung F: $E=F(B_1(0))$, wobei

$$F\colon\mathbb{R}^3\to\mathbb{R}^3,\ \begin{pmatrix}x\\y\\z\end{pmatrix}\mapsto\begin{pmatrix}ax\\by\\cz\end{pmatrix}=\begin{pmatrix}a&0&0\\0&b&0\\0&0&c\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}\Rightarrow DF=\begin{pmatrix}a&0&0\\0&b&0\\0&0&c\end{pmatrix}.$$

   Mit der Transformationsformel folgt:

$$\mathrm{Vol}(E)=\int_E 1\,dx\,dy\,dz=\int_{B_1(0)}1\,abc\,dx\,dy\,dz=\mathrm{Vol}\,B_1(0)\,abc=\frac{4}{3}\pi\,abc.$$

2. $\mathrm{Vol}(B_R(0))$: Wir beschreiben die Kugel mit sogenannten **Kugelkoordinaten** (Abb. 33.4), d.h. ein Punkt im Raum wird durch einen Radius und zwei Winkel beschrieben, genauer durch $\Phi\colon[0,R]\times[0,2\pi]\times[-\pi/2,\pi/2]\to\mathbb{R}^3$,

$$\Phi(r,\varphi,\nu)=(r\cos\varphi\cos\nu,\ r\sin\varphi\cos\nu,\ r\sin\nu).$$

   Die Jacobi–Matrix ist:

$$D(\Phi(r,\varphi,\nu))=\begin{pmatrix}\cos\varphi\cos\nu & -r\sin\varphi\cos\nu & -r\cos\varphi\sin\nu\\ \sin\varphi\cos\nu & r\cos\varphi\cos\nu & -r\sin\varphi\sin\nu\\ \sin\nu & 0 & r\cos\nu\end{pmatrix}.$$

   Man kann errechnen, dass: $\det(D(\Phi(r,\varphi,\nu))=r^2\cos\nu\ge 0$. Also:

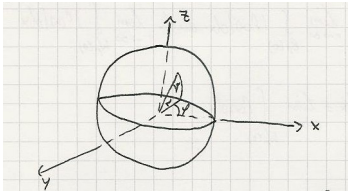


Abbildung 33.4. Kugelkoordinaten.

$$\begin{aligned}\operatorname{Vol}(B_R(0)) &= \int_0^R \Big(\int_0^{2\pi} \Big(\int_{-\pi/2}^{\pi/2} r^2 \cos\nu \, d\nu \Big) \, d\varphi \Big) \, dr \\ &= \int_0^R \Big(\int_0^{2\pi} 2r^2 \, d\varphi \Big) \, dr \\ &= \Big[4\pi \frac{r^3}{3} \Big]_0^R = \frac{4}{3}\pi R^3.\end{aligned}$$

33.2 Uneigentliche Integrale

Bisher haben wir nur über kompakte Mengen integriert. Genauso wie im univariaten Fall des ersten Semesters ist es aber oft nötig, dass wir Bereiche betrachten, die sich ins Unendliche erstrecken. Beispielsweise existiert das Integral $\int_1^\infty \frac{1}{x}\,dx$ aber nicht, obwohl $\int_1^\infty \frac{1}{x^2}\,dx$ bekanntlich existiert (siehe dazu Beispiel 14.2). Nach dem Integralkriterium (Satz 14.3) ist dies äquivalent dazu, dass die entsprechenden Summen $\sum_{i=1}^\infty \frac{1}{i}$ bzw. $\sum_{i=1}^\infty \frac{1}{i^2}$ konvergieren bzw. nicht konvergieren. Eine analoge Problematik existiert im Mehrdimensionalen:

Definition 33.7. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ eine stetige Funktion. Wenn der Grenzwert $\lim_{r\to\infty} \int_{B_r(0)} |f(x)|\,dx$ existiert, dann heißt f **uneigentlich integrierbar** und wir setzen*

$$\int_{\mathbb{R}^n} f(x)\,dx := \lim_{r\to\infty} \int_{B_r(0)} f(x)\,dx = \sum_{r=1}^\infty \int_{\overline{B_r(0)} \setminus \overset{\circ}{B}_{r-1}(0)} f(x)\,dx.$$

Ist $W_r(0)$ der Würfel mit den Ecken $(\pm r, \ldots, \pm r)$, dann existiert mit

$$\lim_{r\to\infty} \int_{B_r(0)} |f(x)|\,dx \quad \text{auch} \quad \lim_{r\to\infty} \int_{W_r(0)} |f(x)|\,dx$$

und es gilt:

$$\int_{\mathbb{R}^n} f(x)\,dx = \lim_{r\to\infty} \int_{W_r(0)} f(x)\,dx.$$

Mit Hilfe von uneigentlichen Integralen in mehreren Variablen können wir beispielsweise die Fläche unter der Funktion e^{-x^2} berechnen, obwohl hier auf den ersten Blick nur eine Variable auftaucht:

Satz 33.8. *Es gilt:* $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$ *(Abb. 33.5).*

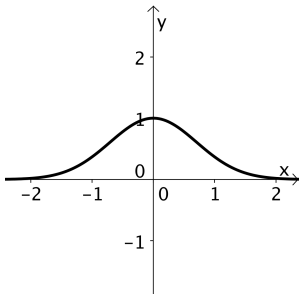


Abbildung 33.5. Graph von e^{-x^2}.

Beweis. Wir berechnen $\int_{\mathbb{R}^2} e^{-(x^2+y^2)}\,dx\,dy$ auf zwei Weisen:

1. Mit dem Satz 33.3 von Fubini:

$$\begin{aligned}
\int_{\mathbb{R}^2} e^{-(x^2+y^2)}\,dx\,dy &= \lim_{r\to\infty}\int_{W_r(0)} e^{-x^2}e^{-y^2}\,dx\,dy \\
&\overset{\text{Fubini}}{=} \lim_{r\to\infty}\int_{-r}^{r}\Big(\int_{-r}^{r} e^{-x^2}e^{-y^2}\,dx\Big)dy \\
&= \lim_{r\to\infty}\int_{-r}^{r}\Big(e^{-y^2}\cdot\Big(\int_{-r}^{r} e^{-x^2}\,dx\Big)\Big)dy \\
&= \lim_{r\to\infty}\Big(\int_{-r}^{r} e^{-x^2}\,dx\cdot\int_{-r}^{r} e^{-y^2}\,dy\Big) \\
&= \Big(\int_{-\infty}^{\infty} e^{-x^2}\,dx\Big)^2.
\end{aligned}$$

2. Mit der Transformationsformel (33.5) und Polarkoordinaten:

$$F(r,\varphi) = (r\cos\varphi, r\sin\varphi),\ DF = \begin{pmatrix}\cos\varphi & -r\sin\varphi\\ \sin\varphi & r\cos\varphi\end{pmatrix} \Rightarrow |\det DF| = r.$$

Also erhalten wir:

$$\begin{aligned}\int_{\mathbb{R}^2} e^{-(x^2+y^2)}\,dx\,dy &= \lim_{R\to\infty}\int_0^R\int_0^{2\pi} e^{-r^2}\cdot r\,d\varphi\,dr\\ &= \lim_{R\to\infty} 2\pi\cdot\left[-\frac{1}{2}e^{-r^2}\right]_0^R = \lim_{r\to\infty}\pi\cdot\left(1-e^{-R^2}\right)\\ &= \pi.\end{aligned}$$

Kombinieren wir die Ergebnisse der beiden Rechnungen, so ergibt sich:

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

□

Dieses Integral wird uns in der Wahrscheinlichkeitstheorie noch häufiger begegnen. Die sogenannte **Dichte** der **Normalverteilung** (siehe Beispiel 34.21) mit **Erwartungswert** $\mu = 0$ und **Standardabweichung** $\sigma = 1$ ist nämlich gerade

$$\varphi_{0;1}(x) = \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}x^2}$$

und man kann unter Ausnutzung der Transformation $x \mapsto \frac{1}{\sqrt{2}}x$ leicht errechnen, dass:

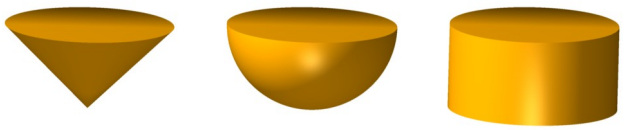
$$\int_{-\infty}^{\infty} \varphi_{0;1}(x)\,dx = 1.$$

Wie wir sehen werden, ist dies die kontinuierliche Variante der Tatsache, dass die Summe über die Wahrscheinlichkeiten aller möglichen Ausgänge eines Experimentes 1 ist.

Aufgaben

Aufgabe 33.1 (Kugeln). Sei $\overline{B_1(0)} \subset \mathbb{R}^n$ eine Einheitskugel. Bestimmen Sie den Radius r, so dass die Kugelschalen $\overline{B_1(0)}\backslash B_r(0)$ und $\overline{B_r(0)}$ gleiches Volumen haben.

Aufgabe 33.2 (Archimedes). Betrachten Sie die Kegel, Halbkugel und Zylinder, deren Grundfläche jeweils ein Kreis mit Radius $r > 0$ und deren Höhe ebenfalls r ist. Zeigen Sie: Die Volumina verhalten sich im Verhältnis



1 : 2 : 3

Aufgabe 33.3 (Rotationskörper). Sei $f\colon [a,b] \to \mathbb{R}$ eine positive Funktion. Leiten Sie die untenstehende Formel für das Volumen V des **Rotationskörpers** her, der durch Rotation des Graphen von f um die x-Achse entsteht:

$$V = \pi \cdot \int_a^b (f(x))^2 \, dx.$$

Aufgabe 33.4 (Torus). Wir berechnen das Volumen eines idealisierten Donuts:

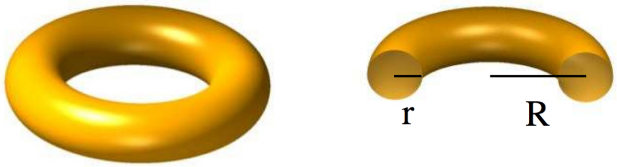
1. Zeigen Sie, dass der **Torus**

$$T_{r,R} = \{(x,y,z) \mid \left(x^2+y^2+z^2+R^2-r^2\right)^2 = 4R^2(x^2+y^2)\}$$

durch

$$\begin{aligned} x &= (R + r\cos v)\cos u, \\ y &= (R + r\cos v)\sin u, \\ z &= r \sin v \end{aligned}$$

$u, v \in [0, 2\pi]$, parametrisiert wird. Anschaulich ist hierbei $r > 0$ der Radius des „Rohres" und $R \geq r$ der Abstand vom Mittelpunkt des „Loches" zu einem Mittelpunkt des „Rohres":



2. Berechnen Sie das Volumen des Torus $T_{r,R}$.

Aufgabe 33.5 (Transformationsformel). Seien $0 < p < q$ und $0 < a < b$. Abschnitte der Parabeln $y^2 = px$, $y^2 = qx$, $x^2 = ay$ und $x^2 = by$ bilden ein krummlinig berandetes Viereck V im $\mathbb{R}^2$.

1. Skizzieren Sie V für die Werte $(p,q,a,b) = (\frac{1}{2}, 1, \frac{1}{2}, 1)$.
2. Berechnen Sie den Flächeninhalt von V.

Teil V

Wahrscheinlichkeitstheorie und Statistik

Einführung

Wahrscheinlichkeitstheorie und Statistik haben sehr viele Anwendungen. Einige davon haben wir bereits in der Einführung zu diesem Semester (Seite **??**) erwähnt. Hier nochmals einige:

Warteschlangen: x Nutzer und Computeserver. $P(x \geq k)$ die Wahrscheinlichkeit, dass mehr als k Nutzer den Service benutzen wollen.

Probabilistische Algorithmen: Wir wollen die erwartete Laufzeit berechnen.

Datenübertragung mit Rauschen: Herausfiltern des Rauschens

Börsenkurse.

Uns ist kein Buch zur Mathematik für Informatiker bekannt, die die in diesem Teil vorgestellten Inhalte vollständig abdeckt. Allerdings gibt es ein gut lesbares Buch, das sich nur diesem Thema widmet: [Kre02]. Dort werden auch einige Resultate bewiesen, die wir im Rahmen dieser Vorlesung nur zitieren können.

34

Grundbegriffe

Zwar kennen viele Hörer die Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik bereits aus der Schule, doch da dies erstens nicht vorausgesetzt werden soll und zweitens sicher einige Notationen und Herangehensweisen in diesem Teil der Vorlesung anders gewählt werden als in der Schule, geben wir zunächst eine ausführliche Einführung in das Thema.

Beispiel 34.1. Wir nehmen einen Würfel und würfeln n-mal. n_i sei die Anzahl, mit der $i \in \{1, 2, \dots, 6\}$ aufgetreten ist.

$$\frac{n_i}{n} \approx \frac{1}{6}$$

ist korrekt für große n, wenn der Würfel fair ist.

Die Wahrscheinlichkeitstheorie gibt einen Rahmen, solche Aussagen zu behandeln und zu erklären.

34.1 Wahrscheinlichkeit von Ereignissen

Definition 34.2. *Ein **Wahrscheinlichkeitsraum** (kurz **W–Raum**) ist ein Tupel $(\Omega, \mathcal{A}, P)$, wobei Ω eine Menge ist, der **Ereignisraum** ist $\mathcal{A} \subset 2^\Omega = \mathcal{P}(\Omega)$, die sogenannte **boolesche Algebra** von beobachtbaren Ereignissen; P ist ein **Wahrscheinlichkeitsmaß** $P \colon \mathcal{A} \to [0, 1]$. Ein solches hat folgende Eigenschaften:*

1. *a) $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cap A_2 \in \mathcal{A}$ und $A_1 \cup A_2 \in \mathcal{A}$,*
   b) $A \in \mathcal{A} \Rightarrow \Omega \backslash A \in \mathcal{A}$
   c) $\emptyset, \Omega \in \mathcal{A}$, und: $A_i \in \mathcal{A}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{A}$.
2. *a) $A, B \in \mathcal{A}$, $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$.*

b) $P(\emptyset) = 0$, $P(\Omega) = 1$, $A \in \mathcal{A} : P(\Omega \backslash A) = 1 - P(A)$,

a′) $A_i \in \mathcal{A}, i \in \mathbb{N}$, A_i *disjunkt, dann:* $P(\dot{\bigcup}_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Beispiel 34.3. Zwei Beispiele, die wir häufiger aufgreifen werden, sind folgende, die wohl jedem geläufig sind:

1. **Würfelmodell**: $\Omega = \{1, \ldots, 6\}$, $\mathcal{A} = 2^{\Omega}$, $P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{6}$.

   Allgemeiner, das **Laplace-Modell**: Ist Ω eine endliche Menge, $\mathcal{A} = 2^{\Omega}$. Aus Symmetrie-Gründen ist klar, dass alle $\omega \in \Omega$ mit gleicher Wahrscheinlichkeit auftreten. Dann gilt: $P(A) = \frac{|A|}{|\Omega|}$ ist ein Wahrscheinlichkeitsmaß, das die Situationen richtig modelliert.
2. **Lotto**: $\Omega = \{\, w \subset \{1, \ldots, 49\} \mid \omega \text{ hat 6 Elemente} \,\}$. Jeder Tipp hat die gleiche Wahrscheinlichkeit. Um dies einzusehen, beschriften wir die Lottokugeln zweifarbig, etwa:

   $$\begin{array}{rl} \text{scharz:} & 1 \;\; 2 \;\; 3 \;\; 4 \;\; 5 \;\; 6 \\ \text{rot:} & 10 \; 11 \; 20 \; 29 \; 30 \; 49. \end{array}$$

   Also: $P(Tipp) = \frac{1}{\binom{49}{6}}$.

Bemerkung 34.4. Die Bedingung 2.*a′*) ist dafür verantwortlich, dass häufig $\mathcal{A} \neq 2^{\Omega}$ gewählt werden muss.

Wir betrachten nun kontinuierliche W-Räume.

Definition 34.5. *Sei* $\Omega = \mathbb{R}$, $[a, b] \in \mathcal{A}$ *ein abgeschlossenes Intervall. Ferner sei* $f \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ *eine stetige Funktion mit* $\int_{-\infty}^{\infty} f(x)\, dx = 1$. *Dann ist mit*

$$P([a, b]) = \int_a^b f(x)\, dx$$

ein Wahrscheinlichsmaß ergeben. Hierbei heißt: $f \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ ***Dichte des Wahrscheinlichkeitsmaßes***.

Beispiel 34.6. Einige häufig verwendete Verteilungen im kontinuierlichen Fall sind die folgenden:

1. **Normalverteilung** (auch **Gaußverteilung**). Die Dichte ist (Abb. 34.1):

   $$f(x) = \varphi_{\mu;\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

   Wir werden sehen, dass $\mu \in \mathbb{R}$ der sogenannte **Erwartungswert** und

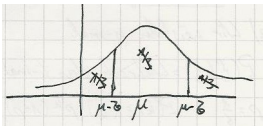


Abbildung 34.1. Die Dichte der Normalverteilung.

$\sigma^2 \in \mathbb{R}_{>0}$ die sogenannte Varianz der Verteilung ist.

Wir zeigen nun, dass tatsächlich $1 = \int_{-\infty}^{\infty} f(x)\,dx$ gilt. Die Substitution $t = \frac{x-\mu}{\sqrt{2\sigma^2}}$, $dt = \frac{dx}{\sqrt{2\sigma^2}}$ liefert:

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt = \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1.$$

Diese Normalverteilung wird folgendermaßen notiert: $\mathcal{N}(\mu, \sigma^2)$. Siehe auch Abb. 34.2.

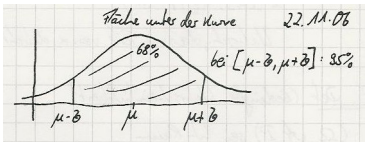


Abbildung 34.2. Die Normalverteilung.

2. **Exponentialverteilung**. Die Dichte ist $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 & x < 0 \end{cases}$ mit $\lambda > 0$ (Abb. 34.3).

Abbildung 34.3. Die Dichte der Exponentialverteilung

Wir rechnen wieder nach, dass dies tatsächlich eine Dichte ist:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_0^{\infty} = 1.$$

Die Exponentialverteilung ist eine typische Lebensdauerverteilung. Beispielsweise sind die **Lebensdauer** von elektronischen Bauelementen und die **Zerfallswahrscheinlichkeit** beim radioaktiven Zerfall annähernd exponentialverteilt.

3. **Gleichverteilung** auf dem Intervall $[a, b]$ (Abb. 34.4). Die Dichte ist hier:

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0 & \text{sonst.} \end{cases}$$

Offenbar ist dies wirklich eine Dichte.

Abbildung 34.4. Die Dichte der Gleichverteilung.

Beispiel 34.7. Wir betrachten **diskrete** W-Räume, d.h. Ω ist endlich oder abzählbar. Für $\omega \in \Omega$ sei $P(\{\omega\}) \in [0, 1]$ vorgegeben, so dass $\sum_{\omega\in\Omega} P(\{\omega\}) = 1$. Dann ist: $P(A) = \sum_{\omega\in A} P(\{\omega\})$.

Anwendung 34.8. 1. Wir würfeln n Mal. $P(\{k\})$ = Wahrscheinlichkeit, dass genau k Mal die 6 auftritt. $p = \frac{1}{6}$.

$$P(\{k\}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

heißt $B_{n,p}$–Verteilung (Binomial–Verteilung , im Fall $n = 1$ auch Bernoulli–Verteilung genannt). Aus der binomischen Formel (Satz 2.7) folgt, dass tatsächlich gilt: $\sum_{k=0}^{n} P(\{k\}) = (p + (1-p))^n = 1^n = 1$.

2. Eine Maschine produziert n Teile. Die Wahrscheinlichkeit, dass ein Teil defekt produziert wird, sei p.

$$P(\{k\}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

ist die Wahrscheinlichkeit, dass genau k Teile defekt sind.

34.2 Bedingte Wahrscheinlichkeit

Definition 34.9. *$(\Omega, \mathcal{A}, P)$ sei ein W-Raum. $\mathcal{A} \ni A_1 \dot\cup \ldots \dot\cup A_n = \Omega$, $B \in \mathcal{A}$, $B = B \cap A_1 \dot\cup \ldots \dot\cup B \cap A_n$. Wir definieren die **bedingte Wahrscheinlichkeit** von $A \in \mathcal{A}$ unter der Annahme B, falls $P(B) > 0$:*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Beispiel 34.10. Wir würfeln 2 Mal. A = wenigstens eine 6. B = Augensumme ≥ 7. Möglichkeiten für B:

erste	zweite
1	6
2	5,6
3	4,5,6
4	3,4,5,6
5	2,3,4,5,6
6	1,2,3,4,5,6

$P(A) = \frac{11}{36} = 1 - \frac{5 \cdot 5}{6 \cdot 6} = 1 - P(\Omega \backslash A)$ und $P(B) = \frac{21}{36}$, $P(A \cap B) = \frac{11}{36}$.

$P(A \mid B) = \frac{11/36}{21/36} = \frac{11}{21} > P(A)$ in diesem Fall.

Beispiel 34.11. Ein Krebstest ist mit 96% Sicherheit positiv, falls der Patient Krebs hat, mit 94% Sicherheit negativ, falls der Patient kein Krebs hat. Bei Patienten in der vorgegebenen Altersgruppe haben 0,5% der Personen Krebs. Wie groß ist die Wahrscheinlichkeit, dass der Patient tatsächlich Krebs hat, bei positivem Testergebnis? T: Test positiv, K: Krebs.

$$P(K \mid T) = \frac{K \cap T}{T} = \frac{0,005 \cdot 0,96}{0,005 \cdot 0,96 + 0,995 \cdot 0,06} \approx 0,074.$$

Dieser Wert erscheint vielen Lesern sicher erstaunlich niedrig.

Satz 34.12 (von der totalen Wahrscheinlichkeit). *Sei $\Omega = \dot{\bigcup}_{i=1}^{n} A_i$ eine Partition und $B \subset \Omega$. Dann:*

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i).$$

Beweis. Es ist: $B = B \cap A_1 \,\dot\cup\, \ldots \,\dot\cup\, B \cap A_n$ eine disjunkte Vereinigung. Für solche addieren sich die einzelnen Wahrscheinlichkeiten, also ergibt sich mit der Definition der bedingten Wahrscheinlichkeit: $P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)$. □

Korollar 34.13 (Formel von Bayes). *Sei $P(B) > 0$. $A_1 \,\dot\cup\, \ldots \,\dot\cup\, A_n = \Omega$. Dann gilt:*

$$P(A_k \mid B) = \frac{P(A_k) \cdot P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i)}.$$

Beweis. Der Satz liefert:

$$P(A_k \mid B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k) \cdot \frac{P(B \cap A_k)}{P(A_k)}}{P(B)} = \frac{P(A_k) \cdot P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i)}.$$

□

Definition 34.14. *Sei $(\Omega, \mathcal{A}, P)$ ein W-Raum, $A, B \in \mathcal{A}$. A und B heißen **unabhängig**, wenn*

$$P(A \cap B) = P(A) \cdot P(B).$$

Falls $P(B) > 0$, so ist dies äquivalent zu:

$$P(A \mid B) = P(A).$$

Beispiel 34.15. Wir betrachten $\Omega = \{1, 2, \ldots, 6\}^2$ und würfeln 2 Mal.

1. $A = \{$ eine 3 im ersten Wurf $\}$, $B = \{$ eine 5 im zweiten Wurf $\}$ sollten hoffentlich nach unserer Definition unabhängige Ereignisse sein. Tatsächlich ergibt sich: $P(A) = \frac{1}{6}$, $P(B) = \frac{1}{6}$ und $P(A \cap B) = \frac{1}{36}$ und daher $P(A \cap B) = P(A) \cdot P(B)$.
2. Weniger offensichtlich ist es, ob die folgenden Ereignisse unabhängig voneinander sind:

   $$A = \{ \text{mindestens eine 6} \}, B = \{ \text{Augensumme ist gerade} \}.$$

   Mögliche Ausgänge für B sind die folgenden:

1. Wurf	2. Wurf
1	1 3 5
2	2 4 6
3	1 3 5
4	2 4 6
5	1 3 5
6	2 4 6

   Also: $P(A \mid B) = \frac{5}{18}$ und $P(A) = \frac{5 \cdot 1 + 1 \cdot 6}{36} = \frac{11}{36}$, $P(B) = \frac{1}{2}$. A und B sind also nicht unabhängig.

34.3 Zufallsvariablen und deren Erwartungswert und Varianz

Definition 34.16. *$(\Omega, \mathcal{A}, P)$ sei ein W-Raum. Eine Abbildung $X \colon \Omega \to \mathbb{R}$ heißt **Zufallsvariable**, wenn $X^{-1}(]-\infty, a]) \in \mathcal{A}$ für jedes $a \in \mathbb{R}$. Es ist: $P(X^{-1}(]-\infty, a])) = P(X \leq a)$.*

*Die **Verteilungsfunktion** von X ist $F_X\colon \mathbb{R} \to [0,1]$, $F_X(a) = P(X \leq a)$. F_X ist monoton steigend; ist F_X stetig diffbar, dann können wir*

$$P(a \leq X \leq b) = \int_a^b F_X'(t)\,dt$$

*als Integral berechnen. Wir schreiben: $f_X := F_X' = \frac{dF}{dX}$ heißt **Wahrscheinlichkeitsdichte** (kurz **W-Dichte** von) X.*

Es gibt viele Zufalls-Variablen, die nur **diskrete Werte** annehmen, d.h. $P(X = x) > 0$ für höchstens abzählbar viele $x \in \mathbb{R}$.

Beispiel 34.17 (faire Münze). Spieler A gewinnt bei Kopf 1 € und verliert bei Zahl 1 €. $S_n \in \mathbb{Z}$ ist der Gewinn nach n Spielen.

$$X = \min_{n \in \mathbb{N}_{\geq 1}} \{n \mid S_n \geq 0\} \in \mathbb{N}_{\geq 1}.$$

In diesem Fall ist X diskret verteilt und F_X ist Treppenfunktion (Abb. 34.5), da $P(X \leq a) = P(X \leq \lfloor a \rfloor)$ für jedes a. Einige Werte: $P(X = 1) = \frac{1}{2}$, $P(X = 2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, usw.

Abbildung 34.5. SKIZZE FEHLT!

Definition 34.18. *1. X sei eine diskrete Zufallsvariable mit endlich vielen Werten $x_i \in \mathbb{R}$, für die $P(X = x_i) > 0$, $i = 1, 2, \ldots, n$. In diesem Fall heißt*

$$E(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$$

*der **Erwartungswert** von X.*

2. *Ist X eine diskrete (nicht notwendig endliche) Zufallsvariable mit Werten x_i, $i \in \mathbb{N}$, dann sei*

$$E(x) = \sum_{i=1}^{\infty} x_i \cdot P(X = x_i).$$

*$E(X)$ heißt **Erwartungswert**, falls $\sum_{i=1}^{\infty} |x_i| \cdot P(X = x_i) < \infty$, d.h. falls die Reihe absolut konvergiert.*

3. *Sei X kontinuierlich verteilte Zufallsvariable mit Dichte f_X.*

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

*heißt **Erwartungswert**, falls $\int_{-\infty}^{\infty} |x| \cdot f_X(x)\, dx < \infty$.*

Beispiel 34.19. 1. Zumindest für endliche Zufallsvariablen entspricht der Erwartungswert unserer Intuition. Würfeln wir z.B. einmal mit einem Würfel, so ergibt sich $E(X) = \sum_{i=1}^{6} i \cdot \frac{1}{6} = \frac{1}{6} \cdot \frac{6 \cdot 7}{2} = \frac{7}{2}$.

2. X sei binomialverteilt ($B_{n,p}$), d.h. $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$. Es gilt:

$$E(X) = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = np.$$

Die letzte Gleichheit werden wir erst später zeigen (Beispiel 35.11).

Bemerkung 34.20 (Linearität des Erwartungswerts). Für zwei Zufallsvariablen X und Y und $\alpha, \beta \in \mathbb{R}$ gilt: $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

Beweis (Idee). Ausnutzen der Linearität von $\sum$, $\int$. □

Beispiel 34.21. X sei eine normal–verteilte Zufalls-Variable, $\mathcal{N}(\mu, \sigma^2)$, d.h.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

Da $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$, folgt:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx \qquad (\text{Subst: } t = x - \mu \Rightarrow dt = dx) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (t+\mu) \cdot e^{\frac{-t^2}{2\sigma^2}}\, dt \\ &= \frac{1}{\sigma\sqrt{2\pi}} \cdot \left(\int_{-\infty}^{\infty} t \cdot e^{\frac{-t^2}{2\sigma^2}}\, dt + \int_{-\infty}^{\infty} \mu \cdot e^{\frac{-t^2}{2\sigma^2}}\, dt \right) \\ &= 0 + \mu \cdot 1 = \mu. \end{aligned}$$

Die 0 kann man hierbei leicht nachrechnen oder einfach mit Punktsymmetrie argumentieren.

Bemerkung 34.22. Sei X eine Zufallsvariable, $\varphi\colon \mathbb{R} \to \mathbb{R}$ eine stetige Funktion. $Y = \varphi \circ X$ ist ebenfalls eine Zufallsvariable mit $E(Y) = \sum_{x\in\mathbb{R}} \varphi(x) \cdot P(X = x)$ im diskreten Fall bzw. $E(Y) = \int_{-\infty}^{\infty} \varphi(x) \cdot f_X(x)\, dx$ im kontinuierlichen Fall. Es ist hierbei nicht klar, dass diese $E(Y)$ endlich sind, d.h. ob dies Erwartungswerte sind.

Definition 34.23. *$\varphi(t) = t^k$, $X^k = \varphi(x)$. Hat X^k einen Erwartungswert, dann heißt $E(X^k)$ k–tes **Moment** von X. Speziell ergibt sich mit der Notation $\mu = E(X)$ wegen der Linearität des Erwartungswertes:*

$$V(X) := E((X - \mu)^2) = E(X^2) - 2 \cdot E(X) \cdot \mu + \mu^2 = E(X^2) - \mu^2.$$

*$V(X)$ heißt **Varianz** von X (falls die ersten beiden Momente existieren). $\sigma := \sqrt{V(X)}$ heißt **Standardabweichung** oder **Streuung** von X.*

Beispiel 34.24. Sei X der Gewinn auf dem Glücksrad

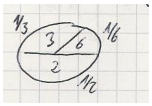


Abbildung 34.6. Ein Glücksrad.

Dann gilt: $\mu = \frac{1}{2} \cdot 2 + \frac{1}{3} \cdot 3 + \frac{1}{6} \cdot 6 = 3$. $E(X^2) = \frac{1}{2} \cdot 4 + \frac{1}{3} \cdot 9 + \frac{1}{6} \cdot 36 = 11$. $\Rightarrow V(X) = \sigma^2 = E(X^2) - \mu^2 = 2 \Rightarrow \sigma = \sqrt{2}$.

Beispiel 34.25. X sei $\mathcal{N}(\mu, \sigma^2)$–verteilt. Dann gilt: $V(X) = \sigma^2$.

Beweis. Übung. □

Bemerkung 34.26 (Eigenschaften der Varianz). Seien X eine reelle Zufallsvariable und $\alpha, \beta \in \mathbb{R}$. Dann gilt:

1. $V(\alpha X) = \alpha^2 \cdot V(X)$,
2. $V(X + \beta) = V(X)$.

Die Varianz ist also nicht linear.

Beweis. Übung. □

Aufgaben

Aufgabe 34.1 (Bedingte Wahrscheinlichkeiten). Zwei Werke sind zu 60% bzw. 40% an der Gesamtproduktion von Transistoren beteiligt. Die Wahrscheinlichkeit, dass ein Transistor mindestens 2000 Stunden betriebsfähig bleibt, ist für das erste Werk 0.8 und für das zweite 0.7.

1. Mit welcher Wahrscheinlichkeit bleibt ein der Gesamtproduktion entnommener Transistor mindestens 2000 Stunden betriebsfähig?
2. Ein beliebig ausgewählter Transistor fiel nach 1200 Stunden aus. Wie groß ist die Wahrscheinlichkeit dafür, dass dieser Transistor aus dem zweiten Werk stammt?

Aufgabe 34.2 (Random Walk). Wir möchten einen eindimensionalen zufälligen Gang simulieren, bei dem wir von 0 ausgehend in jedem Schritt zufällig entweder $1/2$ nach oben oder unten gehen. Nach n Schritten sind wir dann bei einem gewissen Wert $w(n)$ angekommen. Offenbar ist für gerades n der Wert $w(n) \in \mathbb{N}$ und genauer $w(n) \in I := \{-n/2, -n/2+1, \ldots, n/2\} \subset \mathbb{N}$.

1. Benutzen Sie ein Computeralgebra-Programm, wie beispielsweise Maple, um 15 solche Random Walks für $n = 100$ in einem gemeinsamen Koordinatensystem zu visualisieren.
2. Schreiben Sie eine Prozedur, die für N Random Walks, die jeweils n Schritte haben, zählt, wie oft jeder mögliche Ausgang $i \in I$ aufgetreten ist und die diese Anzahlen a_i als Liste oder Array zurückgibt.
3. Visualisieren Sie ein Ergebnis dieser Prozedur für $n = 100$ und $N = 1000$, indem Sie die Werte i gegen a_i in einem Koordinatensystem auftragen.
4. Reskalieren Sie diese Visualisierung, indem Sie nun die Punkte

   $$(i/(2\sqrt{n}),\ 2a_i\sqrt{n}/N)$$

   in ein Koordinatensystem einzeichnen und zeichnen Sie in das selbe Koordinatensystem die Dichte der Gaußschen Normalverteilung $\mathcal{N}(0, \frac{1}{4})$ mit ein. Was fällt auf und wie lässt es sich erklären?

Aufgabe 34.3 (Beim Arzt). Nehmen wir an, dass 1% der Bevölkerung Krankheit X haben. Weiter nehmen wir an, dass es einen Test auf Krankheit X gibt, der in 5% der Fälle als Ergebnis *positiv* liefert, obwohl der Patient die Krankheit nicht hat und in 2% der Fälle als Ergebnis *negativ* liefert, obwohl der Patient die Krankheit hat.

Nehmen wir nun an, dass ein zufälliger Bürger, der getestet wird, sagen wir Herr B, als Resultat *positiv* bekommt. Eine Konsequenz scheint zu sein, dass der Bürger mit einer Wahrscheinlichkeit von 95% die Krankheit hat.

Wie hoch ist diese Wahrscheinlichkeit wirklich? Wie hoch ist umgekehrt die Wahrscheinlichkeit, dass ein *negativ* getesteter Bürger die Krankheit tatsächlich nicht hat?

Bemerkung: Wesentlicher Bestandteil der Berechnungen, die Sie anstellen, ist die zufällige Auswahl des Bürgers. Für einen Patienten, der schon in diversen anderen Tests ein positives Ergebnis bekommen hatte, ist die Sachlage natürlich eine ganz andere.

Aufgabe 34.4 (Unabhängigkeit). Seien X und Y unabhängige, identisch verteilte, kontinuierliche Zufallsvariablen. Wie groß ist $P(X > Y)$?

Aufgabe 34.5 (Varianz bei Gleichverteilung). Sei X in $[a, b]$ gleichverteilt, $a, b \in \mathbb{R}$, $b > a$. Bestimmen Sie die Varianz $V(X)$.

Aufgabe 34.6 (Elementare Wahrscheinlichkeitsrechnung). Bei 4000 Ziehungen im Zahlenlotto 6 aus 49 wurde die Zahlenreihe $15, 25, 27, 30, 42, 48$ zweimal gezogen: am 20.12.1986 und am 21.06.1995. Dies erregte unter Lottospielern ziemliches Aufsehen. Rechnen Sie nach, wie (un)wahrscheinlich das Ereignis, dass mindestens zwei Mal die gleiche Zahlenreihe gezogen wird, wirklich war.

Aufgabe 34.7 (Ausfall von Bauteilen). Ein hochwertig gefertigtes Bauteil hat eine konstante Ausfallrate. D.h. die Wahrscheinlichkeit, dass es innerhalb eines Jahres kaputt geht, bleibt, unabhängig vom Alter des Bauteils, gleich. Aus langjährigen Versuchen ist bekannt, dass am Ende des ersten Jahres 10% der Geräte ausgefallen sind. Wann ist die Hälfte der Bauteile kaputt? *Hinweis:* Treffen Sie eine sinnvolle Verteilungsannahme.

Aufgabe 34.8 (Falsche Übertragung von Nachrichten). In einem Nachrichtenkanal wird ein Zeichen mit der Wahrscheinlichkeit p richtig übertragen. Eine Nachricht bestehe aus acht Zeichen. Mit welcher Wahrscheinlichkeit werden höchstens zwei Zeichen falsch übertragen? Rechnen Sie zuerst allgemein und dann für $p = 0.9$.

35

Kombinatorik und Erzeugende Funktion

In Anwendungen muss man häufig Anzahlen von Möglichkeiten des Auftretens gewisser Ereignisse abzählen. Die Kombinatorik liefert hierzu Methoden. Eine davon ist die sogenannte erzeugende Funktion. Wir werden sehen, dass diese aber noch weitere Anwendungen im Bereich der Wahrscheinlichkeitstheorie besitzt; beispielsweise werden wir mit ihr einige Erwartungswerte ausrechnen können.

Wir beginnen aber mit der Vorstellung zweier Modelle, anhand derer sich viele kombinatorische Probleme erläutern und verstehen lassen, dem Urnen– und dem Schubladenmodell.

35.1 Urnen- und Schubladenmodell

Beispiel 35.1. 1. Das **Urnenmodell**.
Aus einer Urne mit n unterscheidbaren Kugeln (Abb. 35.1) werden k Kugeln gezogen. Dabei kann das Ziehen mit oder ohne Zurücklegen erfolgen und die Reihenfolge eine oder keine Rolle spielen.

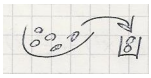


Abbildung 35.1. Das Urnenmodell.

2. Das **Schubladenmodell** (siehe Abb. 35.2). Dieses Modell ist äquivalent zum Urnenmodell (Übungsaufgabe!); der Zusammenhang ist dabei:

Urnenmodell	Schubladenmodell
mit/ohne Zurücklegen	mit/ohne Mehrfachbesetzung
mit/ohne Reihenfolge	unterscheidbare/ununterscheidbare Objekte

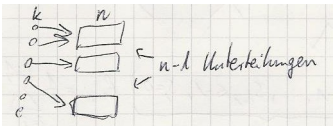


Abbildung 35.2. Das Schubladenmodell.

3. Gegeben n Objekte, von denen wir k auswählen. Wieviele Möglichkeiten gibt es? (Für das Schubladenmodell gelten selbstverständlich die gleichen Zahlen entsprechend.) Für jede der Kombinationen geben wir jeweils eine Kurzschreibweise an, wobei in der Praxis oft nur der Binomialkoeffizient wirklich verwendet wird.

	ohne Zurücklegen	mit Zurücklegen
geordnet	$n(n-1)\cdots(n-k+1) =: (n)_k$ $= \frac{n!}{(n-k)!} = k!\binom{n}{k}$	$n^k = \langle n \rangle_k$
ungeordnet	$\frac{n!}{(n-k)!} \cdot \frac{1}{k!} = \binom{n}{k}$	$\left\langle \binom{n}{k} \right\rangle = \binom{n-1+k}{n-1} = \binom{n-1+k}{k}$

   Außer der Anzahl unten rechts sind alle Anzahlen leicht einzusehen, wenn man sich an den Binomialkoeffizienten aus Abschnitt 2.6 erinnert. Die letzte Anzahl ergibt sich nun folgendermaßen: Sie ist gleich der Anzahl von k–Tupeln $(a_1, \ldots, a_k)$ ganzer Zahlen $1 \leq a_1 \leq a_2 \leq \cdots \leq a_k \leq n \in \mathbb{Z}$. Das ist richtig, weil sie, ohne Beachtung ihrer Reihenfolge, so sortiert sind, dass sie aufsteigend sind. Dies sind aber genauso viele wie die n–Tupel

$$(b_1, \ldots, b_k) := (a_1, a_2 + 1, a_3 + 2, \ldots, a_k + (k-1)),$$

   für die $1 \leq b_1 < b_2 < \cdots < b_k \leq n - 1 + k$ gilt. Deren Anzahl ist aber gerade die Anzahl der Möglichkeiten, k Elemente aus einer $(n-1+k)$–elementigen Menge ohne zurücklegen auszuwählen.

Für Grenzwertbetrachtungen ist häufig der folgende Satz hilfreich, den wir hier leider nicht beweisen können (s. dazu [For08a]):

Satz 35.2 (Stirlingsche Formel, ohne Beweis).

$$n! \approx \sqrt{2n\pi} \cdot \left(\frac{n}{e}\right)^n,$$

wobei $e = \exp(1)$ *die Eulersche Zahl ist.*

Beispiel 35.3 (Fairer Münzwurf). Kopf: +1 €, Zahl: −1 €. S_n: Gewinn nach n Spielen.

$$P(S_{2n}=0)=\frac{\binom{2n}{n}}{2^{2n}}=\frac{(2n)!}{(n!)^2}\cdot 2^{-2n}\approx\frac{\sqrt{4n\pi}\cdot(\frac{2n}{e})^{2n}}{(\sqrt{2n\pi}(\frac{n}{e})^n)^2}\cdot 2^{-2n}=\frac{1}{\sqrt{n\pi}}\underset{n\to\infty}{\to}0.$$

35.2 Abzählen mit erzeugenden Funktionen

Wir betrachten im Folgenden ein längeres Beispiel, anhand dessen die Einführung des Begriffes der erzeugenden Funktion veranschaulicht und dessen Nützlichkeit demonstriert wird:

Beispiel 35.4 (Erste Wechselzeit). Gleiches Spiel, d.h. Kopf: +1 €, Zahl: −1 €. Strategie: Spieler stoppt das Spiel, wenn zum ersten Mal S_n = (Gewinn nach n Würfen) positiv ist (s. Abb. 35.3). $f_n = P(S_1 \le 0, \dots, S_{n-1} \le 0, S_n = 1) = ?$ Einige Werte sind klar: $f_0 = 0$, $f_1 = \frac{1}{2}$, $f_2 = 0 = f_{2n}$ für $n \in \mathbb{N}$.

Abbildung 35.3. Skizze zum Spiel der ersten Wechselzeit.

Wir drücken nun f_n durch f_i aus, für die $i < n$ gilt (s. Abb. 35.3); offenbar muss wenn $S_i = 0$ ist, vorher $S_{i-1} = -1$ gewesen sein, da der Spieler sonst das Spiel schon abgebrochen hätte. Damit ist einsichtig:

$$\begin{aligned} f_n &= P(\text{ an 2-ter Stelle zum ersten Mal wieder 0 })\cdot f_{n-2} \\ &\quad + \cdots + P(\text{ an } (n-1)\text{-ter Stelle zum ersten Mal wieder 0 })\cdot f_1 \\ &= \sum_{i=2}^{n-1} \underbrace{P(S_2 \le -1, \dots, S_{i-1} \le -1, S_i = 0)}_{=\frac{1}{2}\cdot f_{i-1}} \cdot f_{n-i} \\ &= \frac{1}{2}\cdot\sum_{i=2}^{n-1}\cdot f_{i-1}f_{n-i}. \end{aligned}$$

Um nun weiterrechnen zu können, führen wir zunächst einen hilfreichen Begriff ein:

Definition 35.5. *Sei $(f_n)_{n\geq 0}$ eine Folge. Dann heißt*

$$F(x) = \sum_{n=0}^{\infty} f_n x^n$$

***erzeugende Funktion** oder **erzeugende Potenzreihe** von (f_n); diese ist nicht notwendig konvergent. Die Folge $(f_n)\cdot x$ (oder manchmal $(f_n)\cdot z$) heißt **erzeugende Variable** oder **Zählvariable**.*

Beispiel 35.6. Zurück zu obigem Beispiel. Die erzeugende Funktion ist:

$$\begin{aligned} F(x) &= \frac{x}{2} + \sum_{n=2}^{\infty} f_n x^n = \frac{x}{2} + \frac{1}{2}\cdot \sum_{n=2}^{\infty}\Big(x^n\Big(\sum_{i=1}^{n-2} f_i f_{n-(i+1)}\Big)\Big) \\ &\stackrel{!}{=} \frac{x}{2} + \frac{1}{2}\cdot x\cdot F(x)^2. \end{aligned}$$

Um dies einzusehen, berechnen wir zunächst:

$$F(x)^2 = \Big(\sum_{i=0}^{\infty} f_i x^i\Big)\cdot\Big(\sum_{j=0}^{\infty} f_j x^j\Big) = \sum_{n=2}^{\infty}\sum_{i+j=n-1} f_i f_j x^{n-1}.$$

Den Laufindex der inneren Summe können wir zu $j = n-(i+1)$ umschreiben, so dass diese sich als $\sum_{i=1}^{n-2} f_i f_{n-(i+1)} x^{n-1}$ schreiben lässt. Schließlich folgt also:

$$F(x) = \frac{x}{2}(1 + F(x)^2) \;\Rightarrow\; F(x) = \frac{1 \pm \sqrt{1-x^2}}{x}.$$

Da $F(0) = 0$ ist, folgt: $F(x) = \frac{1-\sqrt{1-x^2}}{x}$ (für + ergibt sich $\frac{2}{0} = \infty$ und für − erhält man $\frac{0}{0}$, muss also eine Grenzwertbetrachtung vornehmen). Mit dem binomischen Satz (d.h. die Verallgemeinerung der binomischen Formel 2.7 auf reelle Exponenten; nach Isaac Newton ist nämlich $\sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$ die Taylorreihe von $(1+x)^\alpha$) gilt aber:

$$\sqrt{1-x^2} = (1-x^2)^{\frac{1}{2}} = \sum_{k=0}^{\infty}\binom{\frac{1}{2}}{k}\cdot(-1)^k\cdot x^{2k}.$$

Hierbei ist der Binomialkoeffizient $\binom{r}{k}$ für reelle Zahlen $r \in \mathbb{R}$ in Analogie zum Binomialkoeffizienten für natürliche Zahlen definiert:

$$\binom{r}{k} := \frac{r\cdot(r-1)\cdots(r-(k-1))}{k!}.$$

Damit können wir $F(x)$ nun hinschreiben:

$$F(x) = \binom{\frac{1}{2}}{1} \cdot x - \binom{\frac{1}{2}}{2} \cdot x^3 + \binom{\frac{1}{2}}{3} \cdot x^5 - \dots.$$

Schließlich folgt:

$$f_n = \begin{cases} \binom{\frac{1}{2}}{k} \cdot (-1)^{k-1} & \text{für } n = 2k-1, \\ 0 & \text{für } n = 2k. \end{cases}$$

Kann der Spieler erwarten, etwas zu gewinnen? Wie lange erwartet Spieler A zu spielen, bis er Gewinn macht?

Wir führen dazu eine neue diskrete Zufallsvariable ein: $X = \min\{n \mid S_n \geq 1\}$. Damit gilt: $P(X = n) = f_n$. Wir erhalten

$$E(X) = \sum_{n=1}^{\infty} n \cdot P(X = n) = \sum_{n=1}^{\infty} n \cdot f_n.$$

Es gilt: $F(z) = \sum_{n=0}^{\infty} f_n z^n$, $zF'(z) = \sum_{n=1}^{\infty} n f_n z^n$. Durch Einsetzen von 1 ergibt sich damit::

$$E(X) = (zF')\Big|_{z=1}.$$

Dafür benötigen wir also die Ableitung von F:

$$\begin{aligned} F'(z) &= \frac{1 - \sqrt{1-z^2}}{z} = \frac{z \cdot \frac{1}{2} \cdot \frac{2z}{\sqrt{1-z^2}} - (1 - \sqrt{1-z^2})}{z^2} \\ &= \frac{z^2 - (\sqrt{1-z^2} - (1-z^2))}{z^2\sqrt{1-z^2}} = \frac{1 - \sqrt{1-z^2}}{z^2\sqrt{1-z^2}}. \end{aligned}$$

$$\begin{aligned} \Rightarrow zF'(z) &= \frac{1 - \sqrt{1-z^2}}{z\sqrt{1-z^2}} = \frac{1-(1-z^2)}{z\sqrt{1-z^2}\,(1 + \sqrt{1-z^2})} \\ &= \frac{z}{z\sqrt{1-z^2}\,(1 + \sqrt{1-z^2})}. \end{aligned}$$

Letztendlich ergibt sich also: $E(X) = zF'(z)\Big|_{z=1} = \infty$. Der Erwartungswert existiert also nicht und der Spieler muss erwarten, in endlicher Zeit das Spiel nicht zu beenden.

Wir haben im obigen Beispiel gesehen, dass erzeugende Funktionen hilfreich sein können. Auch bei der Berechnung der Anzahl der Möglichkeiten, Kugeln unter gewissen Nebenbedingungen auf Schubfächer aufzuteilen, sind erzeugende Funktionen hilfreich:

Beispiel 35.7 (Anzahl der Möglichkeiten der Verteilung von Kugeln auf Schubfächer).

1. Wir betrachten die konstante Folge: $f_n = 1 \; \forall n$:

$$(1 + z + z^2 + \cdots + z^k + \cdots) = \frac{1}{1-z}. \tag{35.1}$$

Dies ist korrekt, weil $(1 + z + z^2 + \cdots)(1 - z) = 1 + z - z + z^2 - z^2 \cdots$ eine Teleskopreihe ist.

2. Potenzieren beider Seiten der Gleichung (35.1) liefert:

$$(1 + z + z^2 + \cdots)^n = (1 - z)^{-n}.$$

Was sind die Koeffizienten? Wir betrachten dazu das Schubladenmodell: Wir wollen k identische Kugeln (die k Faktoren von z^k) auf n Schubfächer (die n Faktoren von $(1 + z + z^2 + \cdots)^n$) verteilen. Um die Koeffizienten zu verstehen, stellen wir uns nun vor, dass wir das Produkt $(1 + z + z^2 + \cdots)^n$ nach und nach ausmultiplizieren. Dazu müssen wir aus dem ersten Faktor $(1 + z + z^2 + \cdots)$ eine gewisse Potenz z^{k_1} von z auswählen, dann aus dem zweiten Faktor usw. Im Schubladenmodell heißt dies: Wir legen k_1 Kugeln ins erste Fach, $\ldots$, k_n Kugeln ins n-te Fach mit $k_1 + k_2 + \cdots + k_n = k$: $z^{k_1} \cdot z^{k_2} \cdots z^{k_n} = z^k$. Zählen wir alle diese Möglichkeiten zusammen, so ergibt sich der Koeffizient vor z^k im Produkt. Dieser Koeffizient ist dabei die Anzahl der Möglichkeiten, k identische Kugeln auf n Schubladen zu verteilen; nach der Tabelle in Beispiel 35.1 ist dies gerade $\binom{n-1+k}{n-1}$. Also folgt:

$$(1 + z + z^2 + \cdots)^n = (1 - z)^{-n} = \sum_{k=0}^{\infty} \binom{n-1+k}{n-1} \cdot z^k. \tag{35.2}$$

3. Die Anzahl der Möglichkeiten, k identische Kugeln in n Schubladen zu verteilen, so dass jedes nicht leere Schubfach wenigstens 2 Kugeln enthält, ist der k-te Koeffizient der Potenzreihe

$$\begin{aligned}(1 + z^2 + z^3 + \cdots)^n &= \left(\frac{1}{1-z} - z\right)^n \\ &= \sum_{i=0}^{n} \binom{n}{i} \cdot (-z)^i \cdot (1 - z)^{-(n-i)},\end{aligned}$$

weil wir in der Erläuterung des vorigen Beispiels 35.7.2 immer $k_i \neq 1$ fordern. Wegen der Formel (35.2) folgt:

$$(1+z^2+z^3+\cdots)^n = \sum_{i=0}^{n}\binom{n}{i}(-1)^i z^i \cdot \sum_{j=0}^{\infty}\underbrace{\binom{n-i+j-1}{n-i-1}}_{=\binom{n-i+j-1}{j}} z^j$$
$$= \sum_{k=0}^{\infty}\left(\sum_{i+j=k}^{n}(-1)^i\binom{n}{i}\binom{n-i+j-1}{j}\right) z^k$$
$$= \sum_{k=0}^{\infty}\left(\sum_{i=0}^{\min(n,k)}(-1)^i\binom{n}{i}\binom{n+k-2i-1}{k-i}\right) z^k.$$

Beispielsweise ergibt sich für $n=3, k=4$:

$$\begin{aligned}\# &= \sum_{i=0}^{3}(-1)^i\binom{3}{i}\binom{6-2i}{4-i}\\ &= \binom{6}{4}-\binom{3}{1}\binom{4}{3}+\binom{3}{2}\binom{2}{2}-\binom{3}{3}\binom{0}{1}\\ &= 15-3\cdot 4+3\cdot 1-0=6.\end{aligned}$$

4. k identische Kugeln auf n Schubfächer, in jedem Fall aber höchstens d. In der Notation des obigen Beispiels (2.) entspricht dies der Forderung: $k_i = 0$ für $i > d$. Die Potenzreihe, deren Koeffizienten das zählen, ist also $(1+z+z^2+\cdots+z^d)^n$. Da sich wegen (35.1) durch Durchmultiplizieren mit z^{d+1}
$$(z^{d+1}+z^{d+2}+z^{d+3}+\cdots) = \frac{z^{d+1}}{1-z},$$
ergibt, folgt:
$$(1+z+z^2+\cdots+z^d)^n = \Big(\frac{1}{1-z}-\frac{z^{d+1}}{1-z}\Big)^n = \Big(\frac{1-z^{d+1}}{1-z}\Big)^n.$$

5. Ein Fach mit höchstens 2, eines mit höchstens 3 und ein Fach mit einer geraden Anzahl:
$$(1+z+z^2)(1+z+z^2+z^3)(1+z^2+z^4+\cdots) = \frac{(1-z^3)(1-z^4)}{(1-z)^2(1-z^2)} = \frac{(1-z^3)(1+z^2)}{(1-z)^2}.$$

Einige weitere Beispiele für das Abzählen mit erzeugenden Funktionen:

Beispiel 35.8. Mit $d_k =$ bezeichnen wir die Anzahl der Möglichkeiten, k als Summe von strikt positiven paarweise verschiedenen ganzen Zahlen darzustellen. Ein Beispiel: $k=5 \Rightarrow 5=4+1=3+2$, $d_5=3$. Folgende erzeugende Funktion zählt dies:

$$D(z) = \sum_{k=0}^{\infty} d_k z^k = (1+z)(1+z^2)\cdots = \prod_{n=1}^{\infty}(1+z^n).$$

Beispiel 35.9. P_k = Anzahl der Partitionen von k in einer Summe positiver ganzer Zahlen: 5 = 4+1 = 3+2 = 3+1+1 = 2+2+1 = 2+1+1+1 = 1+1+1+1+1.

$$P_k = \Big|\Big\{\sum_{i=1}^{r} k_i,\ k_1 \geq k_2 \geq \cdots \geq k_r > 0\Big\}\Big|.$$

$$P(z) = \sum_{k=1}^{\infty} P_k z^k = (1+z+z^2+z^3+\cdots)\cdot(1+z^2+z^4+\cdots)\cdot\cdots = \prod_{n=1}^{\infty}\frac{1}{1-z^n}.$$

Erklärungsversuch: Der Summand aus dem ersten Faktor zählt, wie oft wir 1 in Partition nehmen (z^i), der Summand aus dem zweiten Faktor $(z^2)^i$ sagt, dass wir 2 i-mal nehmen usw. $z^k = z^{l_1} z^{2l_2} \cdots z^{sl_s}$, wobei $l_j = |\{i \mid k_i = j\}|$.

35.3 Manipulation erzeugender Funktionen

Wir geben eine knappe Übersicht über mögliche Operationen auf Folgen und den entsprechenden Operationen auf den erzeugenden Funktionen:

Eigenschaft	Folge	Erzeugende Funktion
Definition	(f_i)	$F(z) = \sum_{i=0}^{\infty} f_i z^i$
Summe	$(af_n + bg_n)$	$aF(z) + bG(z)$
Faltung	$h_n = \sum_{k=0}^{n} f_k g_{n-k}$	$H(z) = F(z)\cdot G(z)$

Definition 35.10. *Seien $(f_n), (g_n)$ Folgen, dann heißt $(f_n) * (g_n) = (h_n)$ die **Faltung** der beiden Folgen.*

Weitere Manipulationen:

Eigenschaft	Folge	Erzeugende Funktion
Skalierung: geometrisch	$f_i a^i$	$F(az)$
linear	if_i	$zF'(z)$
faktoriell	$\frac{i!}{(i-k)!} f_i$	$z^k F^{(k)}(z)$
harmonisch	$\frac{f_{i-1}}{i}, i \geq 1$	$\int_0^z F(t)dt$
Summation: kumulativ	$\sum_{j=0}^{i} f_j$	$\frac{F(z)}{1-z}$
vollständig	$\sum_{i=0}^{\infty} f_i$	$F(1)$
alternierend	$\sum_{i=0}^{\infty}(-1)^i f_i$	$F(-1)$
Sequenzwerte: Anfang	f_0	$F(0)$
k-ter Term	f_k	$\frac{F^{(k)}}{k!}(0)$
Grenzwert	$\lim_{k\to\infty} f_k$	$\lim_{z\to 1}(1-z)F(z)$.

Die meisten Aussagen sind klar durch gliedweises Differenzieren bzw. Integrieren in der Reihe. Nur die Ausssagen über die Summationen sind nicht unmittelbar einsichtig und die Grenzwertaussage. Wir zeigen hier nur diese letzte: Es gilt: $(1-z)F(z) = f_0 + (f_1 - f_0)z + (f_2 - f_1)z^2 + \cdots$. Für $z \to 1$ ergibt dies eine Teleskopreihe.

35.4 Anwendung auf eine Erwartungswertberechnung

Mit erzeugenden Funktionen können wir nicht nur abzählen, sondern auch in einigen Fällen Erwartungswerte berechnen:

Beispiel 35.11. Wir berechnen den Erwartungswert einer $B_{n,p}$–verteilten Zufallsvariablen X. Es gilt: $P(X=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$ und daher:

$$E(X) = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Sei $G(z) := G_X(z)$ die **erzeugende Funktion** der Folge $(P(X=k))_{k\in\mathbb{N}}$, d.h.:

$$G(z) = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \cdot z^k = (1-p+pz)^n.$$

Damit ergibt sich für den Erwartungswert von X:

$$E(X) = (zG'(z))\Big|_{z=1} = \left(z \cdot n \cdot (1-p+pz)^{n-1} \cdot p\right)\Big|_{z=1} = n \cdot p,$$

wie wir bereits in Bsp. 34.19 erwähnt haben.

Damit ist es nicht allzu schwierig, die Varianz zu ermitteln, die ja auch ein Erwartungswert ist:

Beispiel 35.12. X sei $B_{n,p}$ verteilt. Dann gilt: $V(X) = np(1-p)$.

Beweis. Übung. □

35.5 Lineare Rekursion

Sei (f_n) eine rekursiv definierte Folge der Form

$$f_{n+1} = af_n + bf_{n-1},$$

wobei f_0 und f_1 vorgegeben sind. Bereits in der linearen Algebra (Abschnitt 24.4.2) haben wir Methoden kennengelernt, um für solche lineare Rekursionen explizite Formeln für die f_n zu berechnen. Erzeugende Funktionen sind hierzu auch ein probates Mittel, wie wir im Folgenden sehen werden:

Beispiel 35.13. Wir betrachten den Fall $a = 2$, $b = 1$ und $f_0 = 0$, $f_1 = 1$, also: $f_{n+1} = 2f_n + f_{n-1}$, d.h. $(f_n) = (0, 1, 2, 5, 12, 29, \dots)$.

Die erzeugende Funktion ist: $F(z) = \sum_{i=0}^{\infty} f_i z^i$. Wegen der Rekursionsgleichung ergibt sich:

$$\begin{aligned} F(z) \cdot (1 - az - bz^2) &= \sum_{n=0}^{\infty} f_n z^n (1 - az - bz^2) \\ &= \sum_{n=0}^{\infty} f_n z^n - \sum_{n=0}^{\infty} a f_n z^{n+1} - \sum_{n=0}^{\infty} b f_n z^{n+2} \\ &= \sum_{n=0}^{\infty} f_n z^n - \sum_{n=1}^{\infty} a f_{n-1} z^n - \sum_{n=2}^{\infty} b f_{n-2} z^n \\ &= \underbrace{f_0 + f_1 z - a f_0 z}_{= f_0 + (f_1 - a f_0) z} + \sum_{n=2}^{\infty} \underbrace{(f_n - a f_{n-1} - b f_{n-2})}_{=0} z^n. \end{aligned}$$

Also folgt: $F(z)(1 - az - bz^2) = c + dz$ mit $c = f_0, d = f_1 - af_0$, also:

$$F(z) = \frac{c + dz}{1 - az - bz^2}.$$

Partialbruchzerlegung (siehe dazu Beispiel 13.27) liefert nun α, β, A, B, so dass

$$\frac{c + dz}{1 - az - bz^2} = \frac{A}{1 - \alpha z} + \frac{B}{1 - \beta z}.$$

Koeffizientenvergleich im Nenner liefert die Bedingung $\beta^2 - a\beta - b = 0$ und aus Symmetriegründen auch $\alpha^2 - a\alpha - b = 0$. Es sind also α und β gerade die Nullstellen von $t^2 - at - b$. Damit können wir nun im Zähler durch Koeffizientenvergleich auch A und B berechnen, was wir hier aber nicht allgemein, sondern nur unten an einem Beispiel vorführen.

Damit folgt nun, da $\frac{1}{1-\alpha z} = \sum_{n=0}^{\infty} \alpha^n z^n$ (geometrische Reihe oder formal nachrechnen):

$$F(z) = A \cdot \sum_{n=0}^{\infty} \alpha^n z^n + B \cdot \sum_{n=0}^{\infty} \beta^n z^n.$$

Somit erhalten wir:

$$f_n = A\alpha^n + B\beta^n.$$

Wir haben also eine Möglichkeit gefunden, explizite Formeln für Werte von rekursiv definierten Folgen zu berechnen.

Beispiel 35.14. In unserem Beispiel von eben: $a = 2, b = 1$: $\alpha, \beta = 1 \pm \sqrt{1 + 1} = 1 \pm \sqrt{2}$, also: $c = 0, d = 1$. Somit folgt: $\frac{z}{1-2z-z^2} = \frac{A}{1-\alpha z} + \frac{B}{1-\beta z}$. Da die Nenner auf

beiden Seiten gleich sind, folgt: $z = A(1-\beta z) + B(1-\alpha z) = (A+B) - (\beta A + \alpha B)z$. Koeffizientenvergleich liefert nun: $A+B = 0 \Rightarrow B = -A$ und $1 = -(\beta A + \alpha B) \Rightarrow A = \frac{1}{4}\sqrt{2}$, also:

$$f_n = \frac{1}{4}\sqrt{2}\left(1+\sqrt{2}\right)^n - \frac{1}{4}\sqrt{2}\left(1-\sqrt{2}\right)^n.$$

35.6 Exkurs: Formale Potenzreihen

Da wir in diesem Kapitel recht intensiv mit erzeugenden Funktionen gearbeitet haben, möchten wir es abschließen mit einigen weiteren Hintergrundinformationen dazu. Eine erzeugende Funktion ist in folgendem Sinn eine formale Potenzreihe:

Definition 35.15. *Sei K ein beliebiger Körper und sei $(f_n)_{n\in\mathbb{N}}$ eine Folge mit Elementen aus K. Dann heißt*

$$F(z) = \sum_{n=0}^{\infty} f_n z^n$$

*eine **formale Potenzreihe**.*

Bemerkung/Definition 35.16. Die **Menge der formalen Potenzreihen** notiert man

$$K[[z]] = \left\{ \sum_{n=0}^{\infty} f_n z^n \mid f_n \in K \right\}.$$

Diese trägt die Struktur eines Ringes (siehe Abschnitt 4.2). Die Addition ist folgendermaßen definiert:

$$\left(\sum_{n=0}^{\infty} f_n z^n\right) + \left(\sum_{n=0}^{\infty} g_n z^n\right) = \sum_{n=0}^{\infty} (f_n + g_n) z^n.$$

Offenbar ist das Negative einer formalen Potenzreihe gerade gegeben durch die Potenzreihe mit den negativen Koeffizienten. Die Multiplikation ist dann naheliegend:

$$\left(\sum_{n=0}^{\infty} f_n z^n\right) \cdot \left(\sum_{n=0}^{\infty} g_n z^n\right) = \sum_{n=0}^{\infty} h_n z^n,$$

wobei $h_n = \sum_{k=0}^{n} f_{n-k} g_k$. Das neutrale Element bzgl. der Mutliplikation ist also $1 = \sum_{n=0}^{\infty} e_n z^n$ mit $e_0 = 1$, $e_i = 0 \; \forall \, i > 0$.

Proposition 35.17. *Ein Element $f(z) = \sum_{n=0}^{\infty} f_n z^n$ hat ein Inverses genau dann, wenn $f_0 \neq 0$.*

Beweis. Die Notwendigkeit dafür ist klar, denn es muss $h_0 = \sum_{k=0}^{0} f_{n-k} g_k = f_0 g_0 = 1$ gelten, was für $f_0 = 0$ nicht möglich ist.

Betrachten wir für die andere Implikation ein Produkt, für das gilt:

$$\Big(\sum_{n=0}^{\infty} f_n z^n\Big) \cdot \Big(\sum_{n=0}^{\infty} g_n z^n\Big) = 1.$$

Dann muss ebenfalls gelten: $f_0 g_0 = 1$. Nach Voraussetzung gilt aber $f_0 \in K \backslash \{0\}$; es gibt also ein Inverses g_0 von f_0, da K ein Körper ist.

Wir müssen nun noch beweisen, dass die Terme höheren Grades in z durch geeignete Wahl von g_n verschwinden. Diese g_n finden wir, indem wir schrittweise das folgende unendliche Gleichunssystem lösen:

$$\begin{aligned} f_0 g_0 = 1 \quad &\Rightarrow \quad g_0 = \tfrac{1}{f_0} \in K \\ f_0 g_1 + f_1 g_0 = 0 \quad &\Rightarrow \quad g_1 = \tfrac{-f_1 g_0}{f_0}, \dots \\ f_0 g_n + \cdots + f_n g_0 = 0 \quad &\Rightarrow \quad \dots \end{aligned}$$

Dazu ist nur notwendig, dass $f_0 \neq 0$, was ja vorausgesetzt war. □

Aufgaben

Aufgabe 35.1 (Erzeugende Funktion). Sei f_i die Anzahl der Möglichkeiten, i als Summe von verschiedenen posititven ganzen Zahlen darzustellen. Sei g_i die Anzahl der Möglichkeiten, i als Summe ungerader positiver ganzer Zahlen darzustellen. Bsp: $f_5 = 3$, nämlich 5, 4 + 1, 3 + 2; $g_5 = 3$, nämlich 5, 3 + 1 + 1, 1 + 1 + 1 + 1 + 1. Euler entdeckte, dass $f_i = g_i \; \forall i$. Zeigen Sie:

1. Die erzeugende Funktion von $(f_i)_{i \in \mathbb{N}}$ ist $F(x) = (1 + x)(1 + x^2)(1 + x^3) \cdots$.
2. Die erzeugende Funktion von $(g_i)_{i \in \mathbb{N}}$ ist $G(x) = \frac{1}{(1-x)(1-x^3)(1-x^5)\cdots}$.
3. Verwenden Sie die Identität $(1 + x^i)(1 - x^i) = (1 - x^{2i})$, um $F = G$ zu zeigen.

Aufgabe 35.2 (Kombinatorik). In einer Urne befinden sich 20 Kugeln, 9 rote, 3 gelbe und 8 blaue. Wir wählen drei davon zufällig aus. Bestimmen Sie die Wahrscheinlichkeit, dass:

1. alle drei gelb sind,
2. mindestens eine gelb ist,
3. eine von jeder Farbe dabei ist.

Aufgabe 35.3 (Lineare Rekursion). Sei $a_0 := 0$, $a_1 := 1$, $a_{n+1} := \frac{a_n + a_{n-1}}{2}$. Berechnen Sie eine nicht-rekursive Formel für a_n mit Hilfe von erzeugenden Funktionen und Partialbruchzerlegung und ermitteln Sie den Grenzwert $\lim_{n \to \infty} a_n$.

36

Summen von Zufallsvariablen

Wir hatten bereits erwähnt, dass aus der Linearität des Summenzeichnens bzw. des Integralzeichens die Linearität des Erwartungswertes, also insbesondere $E(X + Y) = E(X) + E(Y)$ folgt (Bemerkung 34.20). Es wird sich herausstellen, dass eine analoge Aussage für die Varianz nur für unabhängige Zufallsvariablen gilt. Dies werden wir nutzen, um als ein Maß für die Unabhängigkeit von Zufallsvariablen die sogenannte Kovarianz und damit den Begriff der Korrelation zu motivieren.

Um Summen von Zufallsvariablen zu untersuchen, werden wir Verteilungsfunktionen verwenden (Definition 34.16): $F_X \colon \mathbb{R} \to [0,1]$, $F_X(a) = P(X \leq a)$. Dies sind, wie wir oben gesehen haben, monoton wachsende Funktionen, im diskreten Fall monoton wachsende Treppenfunktionen. Für zwei Zufallsvariablen ergibt sich folgender Begriff:

36.1 Gemeinsame Verteilung und Dichte von Summen

Definition 36.1. *Seien X, Y zwei Zufallsvariablen auf Ω. Die **gemeinsame Verteilung** ist*

$$F_{X,Y} \colon \mathbb{R}^2 \to [0,1],\ F_{X,Y}(a,b) = P(X \leq a, Y \leq b).$$

*Im kontinuierlichen Fall sagen wir, dass sie eine **gemeinsame Dichte** $f_{X,Y}$ hat, wenn $f_{X,Y} \colon \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ existiert, so dass*

$$P(X \leq a, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(s,t)\, dt\, ds.$$

Bemerkung 36.2. Ist $f_{X,Y}$ die Dichte des Paares X, Y, dann gilt

$$f_X(s) = \int_{-\infty}^{\infty} f_{X,Y}(s,t)\, dt.$$

Beweis. Es gilt:

$$\int_{-\infty}^{a} f_X(s)\,ds = P(X \le a) = P(X \le a, y \in \mathbb{R}) = \int_{-\infty}^{a} \Big(\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dt \Big)\,ds \quad \forall a.$$

Die Behauptung folgt, da daher

$$\int_{a}^{b} f_X(s)\,ds = \int_{a}^{b} \Big(\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dt \Big)\,ds$$

gilt, durch Grenzwertbildung $b \to a$, da $\int_a^b f_X(s)\,ds \approx (b-a) f_X(s)$ usw. (siehe auch Abb. 36.1). □

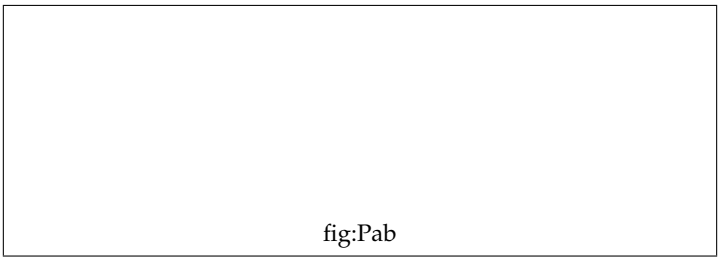


Abbildung 36.1. SKIZZE FEHLT!

Den Begriff der Unabhängigkeit von Ereignissen können wir in naheliegender Weise auf Zufallsvariablen übertragen:

Definition 36.3. *X und Y seien zwei Zufallsvariablen.*

1. *X und Y heißen **unabhängig**, wenn*

$$P(X \le a, Y \le b) = P(X \le a) \cdot P(Y \le b) \quad \forall\, a, b \in \mathbb{R}.$$

2. *Die **bedingte Wahrscheinlichkeit** von $X \le a$ unter der Annahme $Y \le b$, ist*

$$P(X \le a \mid Y \le b) := \frac{P(X \le a, Y \le b)}{P(Y \le b)},$$

   falls der Nenner > 0 ist. Offenbar folgt aus der Unabhängigkeit von X und Y sofort $P(X \le a \mid Y \le b) = P(X \le a)$.

3. *Sind X und Y kontinuierlich verteilt, so lässt sich die **bedingte Wahrscheinlichkeit** $P(X \le a \mid Y = b)$ durch die analoge Formel nicht definieren, da $P(Y = b) = 0$. Die richtige Definition ist folgende:*

$$P(X \le a \mid Y = b) := \frac{\int_{-\infty}^{a} f_{X,Y}(s,b)\,ds}{\int_{-\infty}^{\infty} f_{X,Y}(s,b)\,ds}.$$

Lemma 36.4. *X, Y seien kontinuierlich verteilte Zufallsvariablen. Dann sind X und Y unabhängig genau dann, wenn*

$$f_{X,Y}(s,t) = f_X(s) \cdot f_Y(t)$$

für die Dichten gilt.

Beweis. Unabhängigkeit ist, wie sich recht leicht nachrechnen lässt, äquivalent zu der Bedingung

$$P(a_1 \le X \le a_2, b_1 \le Y \le b_2) = P(a_1 \le X \le a_2) \cdot P(b_1 \le Y \le b_2).$$

Dies ist aber äquivalent zu:

$$\int_{b_1}^{b_2} \int_{a_1}^{a_2} f_{X,Y}(s,t)\, ds\, dt = \int_{a_1}^{a_2} f_X(s)\, ds \cdot \int_{b_1}^{b_2} f_Y(t)\, dt.$$

Multiplizieren mit $\frac{1}{a_2-a_1} \cdot \frac{1}{b_1-b_2}$ und Limes–Bildung $\lim_{a_1 \to a_2, b_1 \to b_2}$ ergibt:

$$f_{X,Y}(a_2,b_2) = f_X(a_2) \cdot f_Y(b_2).$$

Umgekehrt: $f_{X,Y}(s,t) = f_X(s) \cdot f_Y(t) \Rightarrow$ Unabhängigkeit ist klar. □

Satz 36.5. *Seien X, Y kontinuierlich verteilte unabhängige Zufallsvariablen mit Dichten $f = f_X$ und $g = f_Y$. Dann hat die Zufallsvariable $Z = X + Y$ die Dichte*

$$h\colon \mathbb{R} \to \mathbb{R}_{\ge 0},\ h(x) = \int_{-\infty}^{\infty} f(x-y) \cdot g(y)\, dy.$$

Beweis. Wir berechnen (siehe auch Abb. 36.2):

fig:StreifenInt

Abbildung 36.2. SKIZZE FEHLT!

$$\begin{aligned}
P(a_1 \le Z \le a_2) &= P(a_1 \le X+Y \le a_2) \\
&= \int\limits_{\text{Streifen}} \int f_{X,Y}(s,t)\, ds\, dt \\
&= \int_{-\infty}^{\infty} \int_{a_1-t}^{a_2-t} f_{X,Y}(s,t)\, ds\, dt \\
&\overset{\text{Unabhängigkeit}}{=} \int_{-\infty}^{\infty} \int_{a_1-t}^{a_2-t} f(s)g(t)\, ds\, dt \\
&\overset{\text{Subst: } u=s+t}{=} \int_{-\infty}^{\infty} \int_{a_1}^{a_2} f(u-t)g(t)\, du\, dt \\
&= \int_{a_1}^{a_2} \underbrace{\left(\int_{-\infty}^{\infty} f(u-t)g(t)\, dt \right)}_{=h(u)} du.
\end{aligned}$$

Es folgt also: $f_Z(u) = h(u)$. □

Definition 36.6. *Seien f und g Funktionen $\mathbb{R} \to \mathbb{R}$. Dann heißt*

$$f * g \colon \mathbb{R} \to \mathbb{R},\ (f*g)(x) = \int_{-\infty}^{\infty} f(x-y) \cdot g(y)\, dy$$

*die **Faltung der Funktionen** f und g (sofern alle Integrale existieren).*

fig:FaltungAlsFaltung

Abbildung 36.3. SKIZZE FEHLT!

Bemerkung 36.7. 1. Die Faltung von Funktionen ist das kontinuierliche Analogon zur Faltung von Folgen (siehe Definition 35.10).

2. Seien X, Y unabhängige Zufallsvariablen, beide diskret oder beide kontinuierlich mit (diskreten) Dichten $f_i = P(x = i)$, $g_j = P(Y = j)$ bzw. kontinuierlichen Dichten $f(s) = f_X(s)$, $g(t) = f_Y(t)$. Dann wird die Zufallsvariable $Z = X + Y$ durch die Dichte $f * g$ beschrieben.

Bemerkung 36.8. X, Y seien diskret mit Werten $\mathbb{Z}_{\geq 0}$ und unabhängig; die erzeugenden Funktionen der Zufallsvariablen X bzw. Y seien

$$G_X(z) = \sum_{i=0}^{\infty} f_i z^i, \quad G_Y(z) = \sum_{j=0}^{\infty} g_j z^j.$$

Dann gilt:

$$G_X(z) \cdot G_Y(z) = \sum_{k=0}^{\infty} h_k z^k,$$

wobei $h_k = \sum_{i=0}^{k} f_{k-i} g_i$.

Also: Faltung von diskreten Dichten entspricht der Multiplikation der Erzeugenden Funktionen. Im unabhängigen Fall entspricht dies der erzeugenden Funktion der Summe $X + Y$.

Beispiel 36.9. X sei eine $B_{n_1,p}$- und Y eine $B_{n_2,p}$–verteilte Zufallsvariable. X und Y seien unabhängig. $X + Y$ ist dann $B_{n_1+n_2,p}$–verteilt.

Beweis. $G_X(z) = \sum_{i=0}^{n_1} \binom{n_1}{i} p^i (1-p)^{n_1-i} z^i = (1 - p + pz)^{n_1}$, $G_Y(z) = (1 - p + pz)^{n_1}$. Also ist:

$$G_{X+Y}(z) = (1 - p + pz)^{n_1} \cdot (1 - p + pz)^{n_2} = (1 - p + pz)^{n_1+n_2}.$$

□

36.2 Kovarianz und Korrelation

Die Untersuchung der Varianz einer Summe von Zufallsvariablen wird uns auf die Begriffe der Kovarianz und der Korrelation führen, die in vielen Bereichen der Anwendung von Wahrscheinlichkeitstheorie eine wichtige Rolle spielen.

Bemerkung 36.10. 1. X, Y seien unabhängige Zufallsvariablen. Dann gilt:

$$E(X \cdot Y) = E(X) \cdot E(Y),$$

denn (wir zeigen hier nur den kontinuierlichen Fall):

$$E(X \cdot Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s \cdot t \cdot f_X(s) \cdot f_Y(t)\, ds\, dt = \int_{-\infty}^{\infty} s \cdot f_X(s)\, ds \cdot \int_{-\infty}^{\infty} t \cdot f_Y(t)\, dt.$$

2. X, Y seien Zufallsvariablen. Dann gilt, da $V(X) = E(X^2) - E(X)^2$ und da $E(E(X)) = E(X)$ nach Definition der Varianz:

$$\begin{aligned} V(X+Y) &= E\Big(\big[X+Y-E(X+Y)\big]^2\Big) = \\ &= E\Big(X^2 + E(X)^2 + Y^2 + E(Y)^2 + 2XY + 2E(X)E(Y) \\ &\qquad -2XE(X) - 2YE(Y) - 2XE(Y) - 2YE(X)\Big) \\ &= E(X^2) - (E(X))^2 + E(Y^2) - (E(Y))^2 + 2E(XY) - 2E(X)E(Y) \\ &= V(X) + V(Y) + 2\Big(E(X\cdot Y) - E(X)\cdot E(Y)\Big). \end{aligned}$$

Aus den beiden obigen Bemerkungen folgt unmittelbar:

Korollar 36.11. *Sind X und Y unabhängige Zufallsvariablen, dann gilt:*

$$V(X+Y) = V(X) + V(Y).$$

Dies motiviert die Einführung des folgenden Begriffes:

Definition 36.12. *Seien X, Y Zufallsvariablen. Dann heißt*

$$\mathrm{Cov}(X,Y) := E(X\cdot Y) - E(X)\cdot E(Y)$$

*die **Kovarianz** von X und Y.*

Bemerkung 36.13. 1. Es gilt: $\mathrm{Cov}(X,X) = E(X^2) - E(X)^2 = V(X)$.

2. Nach dem obigen Beispiel gilt: X, Y unabhängig $\Rightarrow \mathrm{Cov}(X,Y) = 0$. Deshalb betrachten wir $\mathrm{Cov}(X,Y)$ als ein Maß für die Unabhängigkeit von X und Y. Warnung! Aus $\mathrm{Cov}(X,Y) = 0$ folgt im allgemeinen nicht, dass X, Y unabhängig sind! Dies zeigt das folgende Beispiel.

Beispiel 36.14. Sei X eine Laplace–verteilte Zufallsvariable mit Werten $-1, 0, 1$, also je mit Wahrscheinlichkeit $\frac{1}{3}$. Es gilt: $E(X) = -1\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3} = 0$. Die Zufallsvariable $Y = |X|$ ist determiniert durch X. Mit dieser Definition ergibt sich: $E(X\cdot Y) = -1\cdot 1\cdot\frac{1}{3} + 0\cdot 0\cdot\frac{1}{3} + 1\cdot 1\cdot\frac{1}{3} = 0$ und $E(Y) = 1\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3} = \frac{2}{3}$.
Also: $E(X\cdot Y) = 0 = E(X)\cdot E(Y)$, aber die beiden Zufallsvariablen X, Y sind (offensichtlich) nicht unabhängig.

Definition 36.15. *Die Matrix*

$$C = \begin{pmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} V(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & V(Y) \end{pmatrix}$$

*heißt **Kovarianzmatrix**.*

Bemerkung 36.16. C ist positiv semi–definit, denn

$$(\alpha,\beta)\cdot C\cdot\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \alpha^2 V(X) + 2\alpha\beta\,\mathrm{Cov}(X,Y) + \beta^2 V(Y) = V(\alpha X+\beta Y) \geq 0\ \forall \alpha,\beta\in\mathbb{R}.$$

Definition 36.17. *Der **Korrelationskoeffizient** von X, Y ist*

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)}\,\sqrt{V(Y)}}.$$

In gewissem Sinne misst dieser Wert also den Grad des Zusammenhangs zwischen zwei Zufallsvariablen. Die folgenden Definitheits–Aussagen zeigen, dass alle auftretenden Eigenwerte nicht nur reell (wegen der offensichtlichen Symmetrie der Matrizen), sondern außerdem nicht negativ sind:

Satz/Definition 36.18. *Die **Korrelationsmatrix***

$$\rho := \begin{pmatrix}1 & \rho(X,Y)\\ \rho(X,Y) & 1\end{pmatrix}$$

ist positiv semi-definit und es gilt: $\rho(X,Y)\in[-1,1]$.

Beweis. Man kann sofort nach rechnen, dass sich die Kovarianzmatrix C schreiben lässt als folgendes Produkt:

$$C = \begin{pmatrix}\sqrt{V(X)} & 0\\ 0 & \sqrt{V(Y)}\end{pmatrix}\cdot\begin{pmatrix}1 & \rho(X,Y)\\ \rho(X,Y) & 1\end{pmatrix}\cdot\begin{pmatrix}\sqrt{V(X)} & 0\\ 0 & \sqrt{V(Y)}\end{pmatrix}.$$

Da wir oben (36.16) gesehen haben, dass C positiv semi–definit ist, und da das Konjugieren einer Matrix ihre Eigenwerte nicht ändert, ist ρ ebenfalls positiv semi–definit (weil die linke Matrix bis auf den Faktor $\sqrt{V(X)V(Y)}$ die Inverse der rechten ist).

Es folgt, da die Determinante das Produkt der Eigenwerte ist:

$$\det(\rho) = 1 - (\rho(X,Y))^2 \geq 0$$

also: $\rho(X,Y)\in[-1,1]$. □

Dies gilt auch allgemeiner:

Satz 36.19. *Seien $X_1,\ldots,X_n$ Zufallsvariablen. Dann sind die Kovarianzmatrix $C = (\mathrm{Cov}(X_i,Y_j))\in\mathbb{R}^{n\times n}$ und die Korrelationsmatrix $\rho = (\rho(X_i,X_j))\in[-1,1]^{n\times n}$ positiv semi–definit.*

Beweis. Die positive Semi–Definitheit der Kovarianzmatrix C ergibt sich genauso wie vorher:

$$(a_1, \ldots, a_n) \cdot C \cdot \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = V(a_1 X_1 + \cdots + a_n X_n) \geq 0.$$

Ebenfalls wie vorher folgt die Aussage über die Korrelationsmatrix. □

Eigenwerte der Korrelationsmatrix dicht bei 0 legen nahe, dass einige der Zufallsvariablen sehr stark korrelieren, also wieviele fast kollineare Beziehungen es zwischen den Zufallsvariablen gibt. In der Praxis kann ein Eigenwert nahe 0 also bedeuten, dass man eine Gesetzmäßigkeit entdeckt hat, die einen Zusammenhang zwischen Zufallsvariablen beschreibt. Man kann einen solchen Eigenwert nahe bei 0 auch oft interpretieren als Redundanz in den Daten und daraus folgern, dass man die Anzahl der untersuchten Variablen reduzieren kann, ohne große Informationsverluste befürchten zu müssen. Andererseits legt die Existenz eines dominanten Eigenwertes nahe, dass eine der Untersuchungsrichtungen – die sogenannte **Hauptkomponente** in Richtung des zugehörigen Eigenvektors – vorherrschend ist.

Beispiel 36.20. 1. Zur Reduktion des Rauschens in Bildern kann man beispielsweise untersuchen, ob es eine oder wenige Hauptrichtungen des Rauschens gibt und dann orthogonal dazu projizieren.

2. Zur Untersuchung eines Zusammenhangs verschiedener Kompotenzen im Rahmen der PISA–Studie kann man feststellen, dass zwei Eigenwerte klar größer sind als die anderen, nämlich jene, die zur Mathematik und zum Lesen zugeordnet werden können. Im Gegensatz dazu erscheinen die Naturwissenschaften kein klarer weiterer eigener Einflussfaktor zu sein, sondern eher eine Mischung aus den anderen. Eine solche Untersuchung von Daten bezeichnet man als **Faktorenanalyse**.

Aufgaben

Aufgabe 36.1 (Differenz). Seien X und Y zwei zufällig gewählte Punkte im Intervall $[0, 1]$. Bestimmen Sie die Verteilung der Differenz.

Aufgabe 36.2 (Korrelation). Die Zufallsgrößen X_i, $i = 1, 2, 3$ seien unabhängig und identisch verteilt. Bekannt sind: $E(X_i) = 4$, $V(X_i) = 1$. Es seien $Y_2 = \frac{1}{2}(X_1 + X_2)$ und $Y_3 = \frac{1}{3}(X_1 + X_2 + X_3)$.

Berechnen Sie die Erwartungswerte und die Varianzen der Zufallsgrößen Y_2 und Y_3 sowie die Kovarianz und den Korrelationskoeffizienten zwischen Y_2 und Y_3.

37

Fundamentale Ungleichungen, Gesetz der großen Zahl

Eine der zentralen Aussagen der Wahrscheinlichkeitstheorie ist jene, dass sich bei der Wiederholung eines Experiments nach vielen Versuchen im Wesentlichen das arithmetische Mittel der auftretenden Werte dem Erwartungswert annähert – das sogenannte Gesetz der großen Zahl. Wenn wir häufig genug würfeln, wird sich also tatsächlich im Mittel $\approx 3,5$ ergeben.

Wir werden in diesem Kapitel die Aussage des Gesetzes präzisieren und (die sogenannte schwache Variante davon) beweisen. Dafür benötigen wir zunächst einige Ungleichungen, die allerdings auch unabhängig von dieser Anwendung auf das Gesetz häufig interessant sind.

37.1 Einige Ungleichungen

Die erste Ungleichung, die wir vorstellen, ist die Basis für die weiteren: deren Beweise werden jeweils direkte Anwendungen dieser ersten Ungleichung sein.

Satz 37.1 (Markov-Ungleichung). *X sei eine Zufallsvariable, $h\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ sei eine monoton wachsende Funktion. Dann gilt:*

$$P(X \geq t) \leq \frac{E(h(X))}{h(t)}$$

Beweis. $h(X)$ ist eine neue Zufallsvariable. Wir zeigen die Ungleichung nur im kontinuierlichen Fall:

$$\begin{aligned} E(h(X)) &= \int_{-\infty}^{\infty} h(s)\cdot f_X(s)\,ds \\ &\overset{h\geq 0}{\geq} \int_{t}^{\infty} h(s)\cdot f_X(s)\,ds, \\ &\overset{h \text{ monoton}}{\geq} h(t)\ \cdot \int_{t}^{\infty} f_X(s)\,ds \\ &= h(t)\cdot P(X\geq t). \end{aligned}$$

□

Korollar 37.2 (Chebychev-Ungleichung). *Sei X eine Zufallsvariable, deren Erwartungswert $E(X)$ und Varianz $V(X)$ existieren. Dann gilt:*

$$P(|X - E(X)| \geq t) \leq \frac{V(X)}{t^2}.$$

Beweis. Wir betrachten die Zufallsvariable $Y = (X - E(X))^2$. Dann ist $Y \geq 0$ sicher und der Erwartungswert $E(Y) = V(X)$. Mit $h(t) = t^2$ (diese ist monoton und ≥ 0, so dass wir die Markov Ungleichung anwenden dürfen) erhalten wir:

$$P(|X - E(X)| \geq t) \leq \frac{E(h(|X - E(X)|)}{h(t)} = \frac{E(Y)}{t^2} = \frac{E(Y)}{t^2} = \frac{V(X)}{t^2}.$$

□

Wesentlich bessere Abschätzungen bekommt man, wenn alle Momente $E(X^k)$ von X existieren.

Definition 37.3. *Sei X kontinuierlich verteilt, so dass alle $E(X^k)$ existieren. Dann heißt*

$$\begin{aligned} M_X(\theta) := E(e^{\theta X}) &= \int_{-\infty}^{\infty} e^{\theta s} f_X(s)\,ds \\ &= 1 + E(X)\cdot\theta + E(X^2)\cdot\frac{\theta^2}{2!} + E(X^3)\cdot\frac{\theta^3}{3!} + \cdots \end{aligned}$$

***Momenterzeugende Funktion** von X.*

Bemerkung 37.4. 1. Ist X diskret mit $P(X \in \mathbb{Z}_{\geq 0}) = 1$, dann ist

$$E(e^{\theta X}) = \sum_{k=0}^{\infty} P(X = k)\cdot e^{\theta k} = G_X(e^{\theta}),$$

denn $G_X(z) = \sum_{k=0}^{\infty} P(X = k)\cdot z^k$.

Also: Ist X diskret, dann ist $G_X(z) = M_X(\ln z)$. Die Funktionen M_X und G_X kodieren somit die gleiche Information. M_X verallgemeiert G_X auf den kontinuierlichen Fall.

2. $f\colon \mathbb{R} \to \mathbb{C}$ sei eine komplexwertige, absolut integrierbare Funktion, d.h. $\int_{\mathbb{R}} |f|\, dx < \infty$. Dann ist ihre **Fouriertransformierte**:

$$\hat{f}\colon \mathbb{R} \to \mathbb{C},\ \hat{f}(\theta) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-i\theta s} f(s)\, ds.$$

$\hat{f}$ nennt man auch das **kontinuierliche Spektrum** von f. Man kann zeigen, dass:

$$\hat{\hat{f}}(t) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-it\theta} \hat{f}(\theta)\, d\theta = f(t).$$

Fouriertransformationen verwendet man beispielsweise bei der Bildverarbeitung und beim Lösen von partiellen Differentialgleichungen.

Satz 37.5 (Chernov-Schranke). *X sei eine Zufallsvariable, für die alle Momente existieren, und sei die Momenterzeugende Funktion $M_X(\theta)$ eine konvergente Potenzreihe. Dann gilt:*

$$P(X \geq t) \leq \inf_{\theta \geq 0} (e^{-\theta t} \cdot M_X(\theta)).$$

Beweis. Wir setzen $h(t) = e^{\theta t}$. Dies ist eine monoton wachsende Funktion mit $h(t) \geq 0$. Die Markov-Ungleichung liefert daher:

$$P(X \geq t) \leq \frac{E(e^{\theta X})}{e^{\theta t}} = e^{-\theta t} \cdot M_X(\theta)\ \forall \theta.$$

□

Bei der Chernov–Schranke werden sehr starke Voraussetzungen gestellt. Um zu sehen, was dies bringen kann, vergleichen wir die Güte der verschiedenen Abschätzungen an einem Beispiel:

Beispiel 37.6. Münzwurf: $Y_1, \ldots, Y_n$ seien unabhängige $B_{1,\frac{1}{2}}$-verteilte Zufallsvariablen. Es gilt:

$$E(Y_i) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2},\ V(Y_i) = E\Big(\Big(Y_i - \frac{1}{2}\Big)^2\Big) = \Big(-\frac{1}{2}\Big)^2 \cdot \frac{1}{2} + \Big(\frac{1}{2}\Big)^2 \cdot \frac{1}{2} = \frac{1}{4}.$$

Mit $X_n = \sum_{i=1}^n Y_i$ gilt: $E(X_n) = \frac{n}{2}$ und $V(X_n) = nV(Y_1) = \frac{n}{4}$. Ferner:

$$G_{Y_i}(z) = \frac{1}{2} + \frac{1}{2}z,\ G_{X_n}(z) = \Big(\frac{1+z}{2}\Big)^n, M_{X_n}(\theta) = \Big(\frac{1+e^{\theta}}{2}\Big)^n.$$

Wir schätzen $P(X_n \geq \alpha n)$ mit den verschiedenen Ungleichungen ab:

Markov: $h(t) = \max(t, 0)$. Es gilt:

$$P(X_n \geq \alpha n) \leq \frac{E(X_n)}{\alpha n} = \frac{1}{2\alpha}.$$

Dies liefert nur dann eine nicht-triviale Aussage, wenn $\alpha > \frac{1}{2}$. Wir betrachten daher im Folgenden nur noch $\alpha \in]\frac{1}{2}, 1[$.

Chebychev: Auf $P(X_n \geq \alpha n)$ können wir diese Ungleichung nicht sofort anwenden. Mit $\alpha n = \frac{n}{2} + (\alpha - \frac{1}{2}) \cdot n$ folgt aber, da $E(X_n) = \frac{n}{2}$:

$$P(X_n \geq \alpha n) = P\Big(X_n - \frac{n}{2} \geq \Big(\alpha - \frac{1}{2}\Big) \cdot n\Big)$$

und hierauf können wir die Chebychev–Ungleichung anwenden:

$$\begin{aligned} P\Big(X_n - \frac{n}{2} \geq \Big(\alpha - \frac{1}{2}\Big)n\Big) &\leq P\Big(\Big|X_n - \frac{n}{2}\Big| \geq \Big(\alpha - \frac{1}{2}\Big) \cdot n\Big) \\ &\leq \frac{V(X_n)}{((\alpha - \frac{1}{2}) \cdot n)^2} = \frac{n}{4} \cdot \frac{1}{(\alpha - \frac{1}{2})^2 \cdot n^2} = \frac{1}{4n(\alpha - \frac{1}{2})^2}. \end{aligned}$$

Für große n ist dies deutlich besser als die zuvor gefundene Abschätzung.

Chernov: Es gilt:

$$P(X_n \geq \alpha n) \leq \inf_{\theta} \underbrace{\Big(e^{-\theta \alpha n} \cdot \Big(\frac{1 + e^{\theta}}{2}\Big)^n\Big)}_{= (\frac{1}{2} \cdot (e^{-\theta\alpha} + e^{\theta(1-\alpha)}))^n}.$$

Wir suchen also das Minimum von $g(\theta) = e^{-\theta\alpha} + e^{\theta(1-\alpha)}$. Eine kurze Rechnung ergibt: $\theta_1 = \ln(\frac{\alpha}{1-\alpha})$ ist optimal. Es gilt:

$$g(\theta_1) = \cdots = \frac{\frac{1}{1-\alpha}}{(\frac{\alpha}{1-\alpha})^{\alpha}}.$$

Damit folgt:

$$P(X_n \geq \alpha n) \leq \left(\frac{\frac{1}{2(1-\alpha)}}{\Big(\frac{\alpha}{1-\alpha}\Big)^{\alpha}}\right)^n = \Big(2(1-\alpha) \cdot \Big(\frac{\alpha}{1-\alpha}\Big)^{\alpha}\Big)^{-n} = e^{-n\beta},$$

wobei

$$\beta = \underbrace{\ln(2\underbrace{(1-\alpha)}_{>\frac{1}{2}})}_{>0} + \alpha \underbrace{\ln\Big(\frac{\alpha}{1-\alpha}\Big)}_{>\frac{1}{2}\cdot 2} > 0.$$

Für $n = 100$ ergibt sich die folgende Tabelle, die offenbart, wie stark sich die gefundenen Schranken voneinander unterscheiden:

α	0.55	0.6	0.8
Markov	0.9	0.83	0.62
Chebychev	0.1	0.025	0.002
Chernov	0.006	$1.8 \cdot 10^{-9}$	$1.9 \cdot 10^{-84}$

37.2 Das Gesetz der großen Zahl

Satz 37.7 (Schwaches Gesetz der großen Zahl). *Seien $X_i, 1 \leq i \leq n$, unabhängige identisch verteilte Zufallsvariablen mit $E(X_i) = E(X) < \infty$, $V(X) < \infty$. Dann gilt für $S_n = \sum_{i=1}^n X_i = X_1 + \cdots + X_n$ und für $\varepsilon > 0$: $E(\frac{1}{n}S_n) = E(X)$ und*

$$\lim_{n\to\infty} P\Big(\Big|\frac{1}{n}S_n - E(X)\Big| \geq \varepsilon\Big) = 0.$$

Beweis. Da tatsächlich $E(\frac{1}{n}S_n) = \frac{1}{n}\cdot n \cdot E(X) = E(X)$ und $V(\frac{1}{n}S_n) = \frac{1}{n^2}\cdot(n\cdot V(X)) = \frac{V(X)}{n}$, liefert die Chebychev–Ungleichung:

$$P\Big(\Big|\frac{1}{n}S_n - E(X)\Big| \geq \varepsilon\Big) \leq \frac{V(\frac{S_n}{n})}{\varepsilon^2} = \frac{V(X)}{n\varepsilon^2} \overset{n\to\infty}{\longrightarrow} 0.$$

□

Dies präzisiert nun endlich die Anschauung, dass das arithmetischen Mittel mehrerer Würfe eines Würfels irgendwann im Wesentlichen $\approx 3,5$ ergeben.

Das schwache Gesetz der großen Zahl gilt übrigens auch ohne die Annahme über die Varianz. Dies können wir mit unseren Mitteln hier aber nicht beweisen.

Da wir vorher gesehen haben, dass man wesentlich bessere Abschätzungen als jene mit der Chebychev–Ungleichung erhalten kann, sollte es nicht erstaunen, dass man auch das obigen schwache Gesetz der großen Zahl verschärfen kann. Dies ist tatsächlich der Fall, auch wenn wir dies hier nicht beweisen können:

Satz 37.8 (Starkes Gesetz der großen Zahl). *Sei X_i eine Folge von unabhängigen identisch verteilten Zufallsvariablen mit Erwartungswert. Dann gilt für die Folge $S_n = \sum_{i=1}^n X_i = X_1 + \cdots + X_n$:*

$$P\Big(\limsup_n \Big|\frac{1}{n}S_n - E(X)\Big| \geq \varepsilon\Big) = 0.$$

Mit anderen Worten: Die Folge $(\frac{1}{n}S_n)$ konvergiert **fast sicher** gegen $E(X)$. In [Kre02, §12] ist auch ausgeführt, warum dies eine stärkere Aussage als das von uns bewiesene schwache Gesetz der großen Zahl ist.

37.3 Die Momenterzeugende Funktion

Nachdem wir eben mit dem Gesetz der großen Zahl sicherlich eine der wichtigeren Anwendungen der Chebychev–Ungleichung gegeben haben, möchten wir nun noch einmal etwas detaillierter auf die Momenterzeugende Funktion eingehen, die bei der Chernov–Ungleichung zentral eingeht. Wir haben daran bereits gesehen, wie nützlich die Kenntnis aller Momente sein kann. Und tatsächlich werden wir sehen, dass dies sogar die Verteilung der Zufallsvariablen bereits bestimmt:

Satz 37.9 (Eigenschaften der Momenterzeugenden Funktion).

1. *X sei eine Zufallsvariable mit allen Momenten. Dann gilt:*
$$M_{aX+b}(\theta) = e^{b\theta} \cdot M_X(a \cdot \theta).$$
2. *X, Y seien unabhängige Zufallsvariablen mit allen Momenten. Dann gilt:*
$$M_{X+Y}(\theta) = M_X(\theta) \cdot M_Y(\theta).$$
3. *X, Y seien Zufallsvariablen, deren Momente alle existieren und für die gilt: $M_X(\theta)$, $M_Y(\theta)$ haben Konvergenzradien > 0 und $M_X(\theta)$ und $M_Y(\theta)$ stimmen auf dem gemeinsamen Definitionsbereich überein. Dann folgt: X und Y haben die gleiche Verteilung.*

Beweis. 1. $M_{aX+b}(\theta) = E(e^{(aX+b)\cdot\theta}) = e^{b\theta} \cdot E(e^{aX\theta}) = e^{b\theta} \cdot M_X(a \cdot \theta)$.

2. Mit X und Y sind auch $e^{\theta X}$ und $e^{\theta Y}$ unabhängig für alle θ. Es folgt mit Bemerkung 36.10:
$$M_{X+Y}(\theta) = E(e^{\theta(X+Y)}) = E(e^{\theta X} \cdot e^{\theta Y}) = E(e^{\theta X}) \cdot E(e^{\theta Y}) = M_X(\theta) \cdot M_Y(\theta).$$
3. Für den letzten Teil des Satzes können wir hier keinen Beweis geben. Ohne die Voraussetzung der Konvergenz, etwa nur $E(X^n) = E(Y^n)\ \forall n$ folgt noch nicht, dass X und Y die gleiche Verteilung haben.

□

Definition 37.10. *X, Y seien Zufallsvariablen. Sie heißen **stochastisch gleich** (in Zeichen: $X \underset{st}{=} Y$), wenn*
$$F_X(t) = P(X \leq t) = F_Y(t)\ \forall t \in \mathbb{R}.$$

Bemerkung 37.11. Warnung! Aus $X_1 \underset{st}{=} Y_1$ und $X_2 \underset{st}{=} Y_2$ folgt im allgemeinen nicht: $X_1 + X_2 \underset{st}{=} Y_1 + Y_2$. Dies zeigt das folgende Beispiel.

Beispiel 37.12. X, Y seien zwei unabhängige $B_{1,\frac{1}{2}}$-verteilte Zufallsvariablen. Es gilt also

$$P(X = 0) = P(Y = 0) = \frac{1}{2}, \quad P(X = 1) = P(Y = 1) = \frac{1}{2}$$

und $F_X = F_Y$ (Abb. 37.1), d.h. $X \underset{st}{=} Y$.

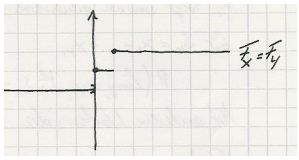


Abbildung 37.1. Summe identisch verteilter Zufallsvariablen (1).

$X + Y$ ist $B_{2,\frac{1}{2}}$-verteilt (Abb. 37.2):

k	0	1	2
$P(X + Y = k)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

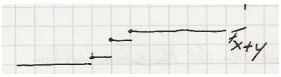


Abbildung 37.2. Summe gleichverteilter Zufallsvariablen (2).

Andererseits sieht dies für $2X$ folgendermaßen aus:

k	0	1	2
$P(2X = k)$	$\frac{1}{2}$	0	$\frac{1}{2}$

Also gilt: $X \underset{st}{=} Y$, aber $X + Y \underset{st}{\neq} 2X$.

Aufgaben

Aufgabe 37.1 (Vergleich der Ungleichungen). Eine unfaire Münze falle mit einer Wahrscheinlichkeit von 1/3 auf Kopf und mit einer Wahrscheinlichkeit von 2/3 auf Zahl. Bestimmen Sie obere Schranken für die Wahrscheinlichkeit, dass die Münze von n Würfen mehr als die Hälfte Mal Kopf zeigt; einmal mit Hilfe der Chebychev-Ungleichung und einmal mit Hilfe der Chernov-Ungleichung. Berechnen Sie konkrete Schranken für $n = 5$ und $n = 20$.

Aufgabe 37.2 (Momente erzeugende Funktion). Eine Zufallsvariable heißt **Poisson–verteilt** zum Parameter λ, wenn

$$P(Y = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

1. Zeigen Sie, dass die Momente erzeugende Funktion der Poisson-Verteilung zum Parameter λ die Funktion $G(\theta) = e^{-\lambda} e^{\lambda e^{\theta}}$ ist.
2. Zeigen Sie damit, dass gilt: $G''(\theta) = \lambda e^{\theta}(G(\theta) + G'(\theta))$.
3. Zeigen Sie, dass Erwartungswert und Varianz den Wert λ haben, dass also gilt: $\sigma^2 = \mu = \lambda$.

38

Der zentrale Grenzwertsatz

Definition 38.1. *Seien (X_n) eine Folge von Zufallsvariablen und X eine weitere Zufallsvariable. X_n **konvergiert in Verteilung** gegen X,*

$$X_n \xrightarrow{\mathcal{D}} X,$$

*(Verteilung heißt auch **Distribution**, daher der Buchstabe $\mathcal{D}$), wenn*

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$$

für alle t, in denen F_X stetig ist.

Beispiel 38.2. 1. X_n sei eine Zufallsvariable, die sicher den Wert $\frac{1}{n}$ annimmt, d.h. $P(X_n = \frac{1}{n}) = 1$ (Abb. 38.1, oben). X sei eine Zufallsvariable mit $P(X = 0) = 1$ (Abb. 38.1, unten). Dann gilt: $X_n \xrightarrow{\mathcal{D}} X$.

Achtung: Da $F_{X_n}(0) = 0\ \forall n$, aber $F_X(0) = 1$, müssen wir diese Sprungstelle herausnehmen.

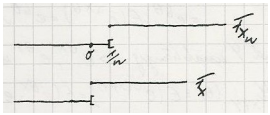


Abbildung 38.1. Ein Beispiel zum zentralen Grenzwertsatz.

2. X_n sei eine Folge unabhängiger gleichverteilter Zufallsvariablen mit $E(X) < \infty$. Dann ist $\frac{1}{n}S_n - E(X)$ eine neue Folge von Zufallsvariablen. Das starke Gesetz der großen Zahl zeigt:

$$\left(\frac{1}{n}S_n - E(X)\right) \xrightarrow{\mathcal{D}} Y,$$

wobei $P(Y = 0) = 1$.

3. X_n sei eine Folge von B_{n,p_n}-verteilten Zufallsvariablen, die alle den gleichen Erwartungswert $np_n = \lambda$ haben. Es gilt also: $p_n = \frac{\lambda}{n}$. Wir berechnen $\lim_{n\to\infty} P(X_n = k)$:

$$\begin{aligned} P(X_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n!}{(n-k)!n^k}\cdot\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^n\cdot\left(1-\frac{\lambda}{n}\right)^{-k} \\ &\xrightarrow{n\to\infty} 1\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda}\cdot 1. \end{aligned}$$

Der Grenzwert des letzten Faktors ist hierbei klar. Die Grenzwerte des ersten und des vorletzten Faktors, 1 und $e^{-\lambda}$, sind Resultate, die man als Übungsaufgabe mit Mitteln aus dem Abschnitt über die Analysis einer Veränderlichen lösen kann. Insgesamt folgt also:

$$X_n \xrightarrow{\mathcal{D}} Y,$$

wobei Y eine Zufallsvariable mit $P(Y = k) = \frac{\lambda^k}{k!}\cdot e^{-\lambda}$ ist.

Definition 38.3. *Eine Zufallsvariable heißt* ***Poisson–verteilt*** *zum Parameter λ, wenn*

$$P(Y = k) = \frac{\lambda^k}{k!}\cdot e^{-\lambda}.$$

Siehe auch Abb. 38.2.

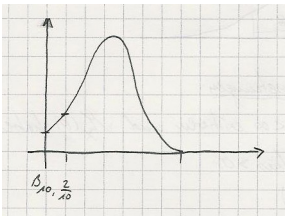


Abbildung 38.2. Die Poisson–Verteilung.

Tatsächlich gilt

$$\sum_{k=0}^{\infty} P(Y = k) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}\cdot e^{\lambda} = 1$$

und daher auch $P(Y \in \mathbb{Z}_{\geq 0}) = 1$.

Nachdem wir nun einige Beispiele der Konvergenz in Verteilung betrachtet haben, kommen wir zu folgendem wichtigen Resultat:

Satz 38.4 (Zentraler Grenzwertsatz). *Sei X_n eine Folge unabhängiger, identisch verteilter Zufallsvariablen mit $E(X), V(X) < \infty$. Wir setzen: $S_n = X_1 + \cdots + X_n$. Dann gilt:*

$$\sqrt{n} \cdot \Big(\frac{1}{n} S_n - E(X)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V(X)).$$

Beweis (Beweisskizze unter zusätzlichen Voraussetzungen). Zusätzlich nehmen wir an, dass alle Momente existieren, d.h. $E(X^n) < \infty$, und dass $M_X(\theta)$ einen Konvergenzradius > 0 hat. Wir setzen: $Z_n := \sqrt{n}(\frac{1}{n} S_n - E(X))$. Dann gilt:

$$M_{Z_n}(\theta) = e^{-\sqrt{n}E(X)\theta} \cdot M_X\Big(\frac{\theta}{\sqrt{n}}\Big)^n = \exp\Big(n \cdot \Big(\ln(M_X(\frac{\theta}{\sqrt{n}})) - E(X) \cdot \frac{\theta}{\sqrt{n}}\Big)\Big).$$

Zur Vereinfachung der Notation setzen wir: $\varepsilon_n = \frac{\theta}{\sqrt{n}}$ und

$$\psi_n(\theta) = n \cdot \Big(\ln(M_X(\varepsilon_n)) - E(X) \cdot \varepsilon_n\Big),$$

also $M_{Z_n}(\theta) = \exp(\psi_n(\theta))$.

Es reicht demnach, $\lim_{n\to\infty} \psi_n(\theta)$ zu bestimmen. Wir formen zunächst um:

$$\psi_n(\theta) = \theta^2 \cdot \frac{\ln(M_X(\varepsilon_n)) - E(X)\varepsilon_n}{(\varepsilon_n)^2}.$$

Mit $n \to \infty$ gilt nach Definition $\varepsilon_n \to 0$ und daher auch $\ln(M_X(\varepsilon_n)) \to \ln(M_X(0)) = \ln(e^0) = 0$. Wir dürfen also den Satz von L'Hospital anwenden:

$$\lim_{n\to\infty} \psi_n(\theta) = \lim_{\varepsilon_n \to \infty} \psi_n(\theta) = \theta^2 \cdot \frac{\frac{M_X'(\varepsilon_n)}{M_X(\varepsilon_n)} - E(X)}{2 \cdot \varepsilon_n}.$$

Da $M_X'(0) = E(X)$ und, wie schon erwähnt, $M_X(0) = 1$, können wir den Satz nochmals anwenden:

$$\begin{aligned} \lim_{n\to\infty} \psi_n(\theta) &= \lim_{\varepsilon_n \to \infty} \psi_n(\theta) = \theta^2 \cdot \frac{M_X(\varepsilon_n) \cdot M_X''(\varepsilon_n) - (M_X'(\varepsilon_n))^2}{2 \cdot M_X(\varepsilon_n)^2} \\ &\xrightarrow{n\to\infty} \theta^2 \cdot \frac{1 \cdot E(X^2) - (E(X))^2}{2} = \frac{\theta^2 \cdot V(X)}{2}. \end{aligned}$$

Also:

$$\lim_{n\to\infty} M_{Z_n}(\theta) = \exp\Big(\frac{\theta^2 \cdot V(X)}{2}\Big).$$

Andererseits ist die Dichte der Normalverteilung mit Erwartungswert 0 bekanntlich

$$f_{\mathcal{N}(0,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}},$$

so dass wir für eine Zufallsvariable Y mit dieser Dichte erhalten:

$$\begin{aligned} M_Y(\theta) = E(e^{\theta Y}) &= \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{\theta t} e^{-\frac{t^2}{2\sigma^2}}\, dt \\ &= \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-\frac{t^2-2\theta\sigma^2 t}{2\sigma^2}}\, dt \\ &= \frac{1}{\sigma\sqrt{2\pi}} \cdot \Big(\int_{-\infty}^{\infty} e^{-\frac{(t-\theta\sigma)^2 t}{2\sigma^2}}\, dt\Big) \cdot e^{\frac{\theta^2\cdot\sigma^2}{2}} \\ &= 1 \cdot e^{\frac{\theta^2\cdot V(X)}{2}}. \end{aligned}$$

Tatsächlich gilt demnach $\lim_{n\to\infty} M_{Z_n}(\theta) = M_Y(\theta)$. Im (von uns nicht bewiesenen) Satz 37.9 hatten wir aber gesehen, dass zwei Zufallsvariablen bereits übereinstimmen, wenn sie die gleiche Momenterzeugende Funktion besitzen und diese einen positiven Konvergenzradius hat. Die letzte Bedingung hatten wir aber zusätzlich oben gefordert, so dass die Behauptung im von uns gewählten Spezialfall bewiesen ist. □

Wir haben für den Beweis den von uns nicht gezeigten Satz 37.9 verwendet. In [Kre02, §12] findet man eine Herleitung des zentralen Grenzwertsatzes, die zwar länger, aber dafür elementarer ist und auf Kersting zurückgeht.

Der zentrale Grenzwertsatz sagt, kurz gesagt, aus, dass das zentrierte arithmetische Mittel identisch verteilter Zufallsvariablen näherungsweise normalverteilt ist, unabhängig von der Ausgangsverteilung der Zufallsvariablen. Dieses Resultat ist von fundamentaler Bedeutung. Bei sehr vielen Problemen lässt sich die Verteilung der interessierenden Zufallsvariablen nicht oder nur mit sehr großem Aufwand bestimmen. Mit Hilfe des Satzes (oder Varianten davon) kann man in solchen Situationen dann oft wenigstens asymptotische Aussagen machen, die für praktische Anwendungen auch häufig ausreichend sind. Beispielsweises kann man recht leicht aus dem Satz folgern:

Satz 38.5 (Moivre/Laplace). *Die Binomialverteilung $B(n,p)$ für $0 < p < 1$ kann näherungsweise durch die Normalverteilung $\mathcal{N}(np, np(1-p))$ beschrieben werden.*

Zum zentralen Grenzwertsatz nun Zahlen aus einem tatsächlich durchgeführten Experiment:

Beispiel 38.6. Länge von Piniennadeln (das Beispiel stammt von der Webseite http://web.neuestatistik.de/demo/Demo_DE/MOD_100238/html/comp_100459.html). Es wurden 3000 Durchschnittswerte der Längen von Piniennadeln ermittelt, wobei jeder Durchschnittswert auf jeweils 250 Messungen beruht (genaue Verteilung siehe Webseite). Dieser Datensatz gibt uns die Möglichkeit, zu überprüfen, ob der Stichprobenumfang schon groß genug ist, um in

diesem Fall die arithmetischen Mittel als normalverteilt ansehen zu können. Wie die Graphik 38.3 zeigt, ist die Näherung schon gar nicht so schlecht.

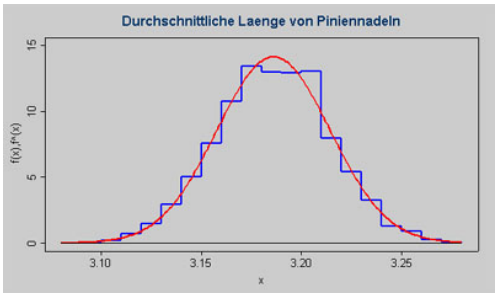


Abbildung 38.3. Der zentrale Grenzwertsatz am Beispiel einer Piniennadelmessung.

Aufgaben

Aufgabe 38.1 (Poissionverteilung). Auf der Erde gibt es pro Jahr im Mittel ein Erdbeben der Stärke 8 oder mehr auf der Richterskala. Wir nehmen an, die Zahl solcher Erdbeben pro Jahr folge der Poisson-Verteilung. Wir gehen davon aus, dass die Anzahlen solcher Erdbeben in den einzelnen Jahren unabhängig voneinander sind.

1. Mit welcher Wahrscheinlichkeit gibt es im nächsten Jahr mehr als ein solches Erdbeben?
2. Wir bezeichen mit Y die Anzahl der Jahre im Zeitraum von 2006 bis 2105 in denen mehr als zwei Erdbeben der Stärke 8 oder mehr auf der Richterskala stattfinden. Welche Verteilung hat Y? Wieviele Jahre mit mehr als zwei Erdbeben solcher Stärke können wir in diesem Zeitraum erwarten?

Aufgabe 38.2 (Flugüberbuchung). Aus jahrelanger Erfahrung weiß ein Flugunternehmen, dass im Mittel 7% der Personen, die ein Flugticket gekauft haben, nicht bzw. zu spät zum Abflug erscheinen. Um die Zahl der somit ungenutzten Plätze nicht zu groß werden zu lassen, werden daher für einen Flug, bei dem 240 Plätze zur Verfügung stehen, mehr als 240 Tickets verkauft.

Wieviele Flugscheine dürfen höchstens verkauft werden, dass mit Wahrscheinlichkeit mindestens 0.99 alle zum Abflug erschienenen Personen, die ein Flugticket haben, auch einen Platz im Flugzeug bekommen?

Zur Modellierung betrachten wir unabhängige $B_{1,p}$-verteilte Zufallsvariablen $X_1, \ldots, X_n$, wobei $X_i = 1$ genau dann gelte, wenn die Person, die das i-te Ticket gekauft hat, tatsächlich mitfliegt. n ist hierbei die Anzahl der verkauften Tickets und $P(X_i = 1) = p = 1 - 0.07$.

Approximieren Sie zur Beantwortung obiger Frage die Verteilung von $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - E(X_1))$ durch die Normalverteilung $\mathcal{N}(0, V(X_1))$.

Aufgabe 38.3 (Salatbar). An der Salatbar einer Mensa kostet der Salat 1 EUR/100 g. Der Salat wird gewogen und der Betrag zwecks leichterer Bezahlbarkeit auf ein Vielfaches von 50 Cent auf- oder abgerundet. Wie hoch ist das Risiko, dass der Student nach 192 maligem Salatessen durch die Rundung einen Nachteil von mehr als 3 EUR hat, wenn er das Salatgewicht nicht vorher abschätzt? (Treffen Sie sinnvolle Annahmen bzgl. der Verteilungen der eingeführten Zufallsgrößen).

Aufgabe 38.4 (Serverperformance). Ein wichtiges Kriterium für die Performance eines Webserves ist die schnelle Bearbeitung von Rechneranfragen. Im folgenden soll die daher die Verteilung von Zeitabständen zwischen Anfragen von Rechnern an den Server untersucht werden. In einem einfachen Modell soll der Server in einem Zeitraum T genau N unabhängige Anfragen erhalten.

- Benutzen Sie ein Computeralgebrasystem (z.B. MAPLE) zur Simulation von $N = 100.000$ Anfragen in $T = 1h = 3.600.000ms$. Plotten Sie die Verteilung der Zeitdifferenzen von zwei zeitlich aufeinanderfolgenden Anfragen und berechnen Sie den Mittelwert dieser Differenzen (in ms).
- Können Sie die Verteilung der Differenzen erraten?
- Wie hoch ist die Wahrscheinlichkeit, dass zwei Anfragen mit weniger als 2 ms Differenzeintreffen? Berechnen Sie den Wert aus der Simulation in (38.4) und mit Hilfe der Verteilung aus (38.4).

39

Statistik

In vielen Anwendungen möchte man Aussagen überprüfen, Parameter von Verteilungen schätzen u.ä. Solche Tätigkeiten gehören in den Bereich der Statistik.

39.1 Testen von Hypothesen

Eine Maschine produziert Teile mit behaupteten Ausschussanteil p. Wir möchten dies überprüfen und nehmen dafür eine Stichprobe von n Teilen. Wir modellieren dies mit Zufallsvariablen X_i, die $B_{1,p}$-verteilt sind, d.h. wenn wir k defekte Teile ziehen, so ist $\frac{k}{n} \approx p$.

Wir können nun verschiedene Hypothesen aufstellen, beispielsweise:

- $H_0 \colon p \leq p_0$, gegen $H_1 \colon p > p_0$ (**einseitiger Test**). Wir werden uns dann für H_0 entscheiden, wenn $\frac{k}{n} - p_0 \leq c$ für ein gewisses c ist und gegen H_0, wenn $\frac{k}{n} - p_0 > c$ für ein gewisses c.
- $H_0 \colon p \geq p_0$, gegen $H_1 \colon p < p_0$ (**einseitiger Test**). Wir werden uns dann für H_0 entscheiden, wenn $\frac{k}{n} - p_0 \geq c$ für ein gewisses c ist und gegen H_0, wenn $\frac{k}{n} - p_0 < c$ für ein gewisses c.
- $H_0 \colon p = p_0$, gegen $H_1 \colon p \neq p_0$ (**zweiseitiger Test**). Wir werden uns dann für H_0 entscheiden, wenn $|\frac{k}{n} - p_0| \leq c$ für ein gewisses c ist und gegen H_0, wenn $|\frac{k}{n} - p_0| > c$ für ein gewisses c.

Da wir nicht sicher sein können, was die Wahrheit ist, versuchen wir, den Fehler, den wir bei einer solchen Entscheidung machen, einzugrenzen. Mögliche Fehlerkategorien werden dabei klassischerweise folgendermaßen eingeteilt:

Entscheidung \ Fakt	H_0	H_1
H_0	okay	Fehler 2. Art
H_1	Fehler 1. Art	okay

Hierbei wird allerdings ein nicht zu unterschätzender Fehler vergessen, der manchmal als Fehler 3. Art bezeichnet wird: das Modell ist komplett falsch. Dieses ist, wie man sieht, ein besonders heimtückischer Fehler.

Das Problem bei unserer Entscheidung ist, dass wir leider nicht den Fehler 1. Art und den Fehler 2. Art gleichzeitig klein halten können. Daher beschränken wir den Fehler 1. Art auf einen relativ kleinen Wert und minimieren unter dieser Nebenbedingung den Fehler 2. Art.

Wenn wir uns also in einem konkreten Beispiel zunächst entscheiden müssen, welche der drei Nullhypothesen H_0 wir verwenden, sollten wir dies so machen, dass der Fehler 1. Art der schlimmere der beiden Fehler ist, da wir die Wahrscheinlichkeit, dass dieser eintritt, ja auf einen beliebig kleinen Wert, etwa 5% oder 1%, beschränken können (im Gegensatz zum Fehler 2. Art). Insbesondere sollten wir die Entscheidung, welche Nullhypothese wir wählen, nicht von den Daten, die ermittelt wurden, abhängig machen.

Beispiel 39.1. Wie können wir als Finanzminister feststellen, ob eine geplante Vereinfachung des Steuerrechts zu Mindereinnahmen des Staates führt oder nicht?

Wir bilden für n Bürger zunächst die Differenzen x_i = Steuer des Bürgers i bei neuen Recht − Steuer des Bürgers i bei altem Recht. Ist hierbei $x_i > 0$, so erhält der Staat bei Bürger i nach neuen Recht mehr Geld als bei altem.

Welche der möglichen Null-Hypothesen über den wahren Erwartungswert μ der Einnahmen sollten wir also verwenden? $H_0 \colon \mu \leq 0$, da nämlich dann der Fehler 1. Art folgender Fall ist: zwar führt die Steuerreform in Wahrheit zu Mindereinnahmen, wir entscheiden uns aber in unserem Test dafür, dass sie zu Mehreinnahmen führt (und führen die Reform also durch). Dieser Fall ist für uns als Finanzminister klarerweise der schlimmere Fall.

Beispiel 39.2. Wir betrachten eine **Stichprobe** $x_1, \ldots, x_n$ von $B_{1,p}$-verteilten Zufallsvariablen. Das **Stichprobenmittel** ist $\bar{x} = \frac{1}{n} \sum x_i$.

Wir testen die Hypothese $H_0 \colon p \leq p_0$ gegen $H_1 \colon p > p_0$. Falls $\overline{X} \leq c$, so nehmen wir H_0 an, sonst lehnen wir H_0 ab. Wie bestimmen wir c?

Die Vorgehensweise ist folgende: Sei $\vartheta = E(X) = E(\overline{X})$ (p ist unbekannt). Wir suchen nun c, so dass die Wahrscheinlichkeit für den Fehler 1. Art durch α beschränkt ist, d.h.

$$P(\overline{X} > c \mid \vartheta \leq p_0) \leq \alpha,$$

etwa $\alpha = 0,05$ oder $0,01$. Unter dieser Nebenbedingung minimieren wir den Fehler 2. Art, d.h.

$$\min_c(P(\overline{X} \leq c \mid \vartheta > p_0)).$$

Wegen der Positivität der Dichten ist die Lösung dieser Minimierungsaufgabe das c mit $P(\overline{X} > c \mid \vartheta = p_0) = \alpha$. Für gewisse Verteilungen können wir für einige α die entsprechenden Werte für c aus Tabellen ablesen oder mit Computeralgebra–Programmen berechnen.

Es ist sogar möglich, Hypothesen "ber die Zufälligkeit einer Stichprobe zu testen:

Beispiel 39.3 (Zufall und Intuition). Die Hörer bekommen die Aufgabe, eine Sequenz $a_1, \ldots, a_{100}$ von 0–en und 1–en auszudenken, die möglichst zufällig sein soll. Außerdem soll eine weitere Sequenz $b_1, \ldots, b_{100}$ durch Wurf einer Münze wirklich zufällig erzeugt werden. Wir zeigen der Illustration halber nur eine Sequenz der Länge 30:

$$0\,0\,0\,1\,0\,1\,1\,1\,0\,0\,1\,1\,0\,1\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,1\,1\,0\,0\,0\,1$$

Üblicherweise kann man recht leicht entscheiden, welche der Folgen wirklich zufällig erzeugt wurde, weil die menschliche Intuition für den Zufall in solchen Fällen meist versagt. Beispielsweise werden von vielen Leuten zu wenige Sequenzen aufeinanderfolgender gleicher Ziffern aufgeschrieben.

Um in der Praxis zu entscheiden, welche der Folgen wirklich zufällig erzeugt wurde, reicht daher folgende Überlegung meist aus: Wir bezeichnen mit einem **Run** der Länge l (kurz l**–Run**) eine maximale Abfolge $a_i, \ldots, a_{i+l}$ von gleichen Ziffern in einer Sequenz, d.h. mit der Eigenschaft $a_{i-1} \neq a_i$ und $a_{i+l+1} \neq a_{i+l}$. Beispielsweise hat obige Folge nur einen 4–Run, aber drei 3–Runs.

Eine zufällige Ziffernreihe mit N Stellen hat dann ungefähr $\frac{N}{2}$ Runs, weil die Wahrscheinlichkeit für einen Wechsel $\frac{1}{2}$ beträgt. Von diesen Runs sind wiederum etwa die Hälfte, also $\frac{1}{2}$ aller Runs, 2–Runs, davon wieder die Hälfte, also $\frac{1}{4}$, sind 3–Runs. Allgemein sollten etwa $\frac{1}{2^{l-1}}$ der Runs die Länge l haben. Diese Information reicht meist schon aus, um zu entscheiden, welche der Folgen wirklich zufällig war, weil wenige Menschen in der Lage sind, dies ähnlich gut zu realisieren wie ein wirklich zufälliger Prozess, wie etwa das Werfen einer Münze.

39.2 Schätzen von Parametern

Definition 39.4. *Sei X eine Zufallsvariable. Wir nehmen eine* ***Stichprobe*** $x_1, \ldots, x_n$. *Das* ***Stichprobenmittel*** *ist*

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Die ***Stichprobenvarianz*** *ist*

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2.$$

Warum dividieren wir bei der Varianz durch $n-1$ und nicht, wie beim Mittel, durch n? Um diese Frage beantworten zu können, führen wir den folgenden Begriff ein:

Definition 39.5. *Sei X eine Zufallsvariable, welche von einem Parameter $\alpha \in \mathbb{R}$ abhängt. Eine Abbildung $h\colon \mathbb{R}^n \to \mathbb{R}$ heißt* ***konsistenter Schätzer*** *für α, wenn für unabhängige Zufallsvariablen X_i mit $X_i \underset{st}{=} X$ der Erwartungswert der Zufallsvariablen $h(X_1, \ldots, X_n)$ gerade α ist:*

$$E(h(X_1, \ldots, X_n)) = \alpha.$$

Beispiel 39.6. 1. $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ ist ein konsistenter Schätzer für $\alpha = E(X)$. Es gilt nämlich:

$$E\Big(\frac{1}{n}\sum_{k=1}^{n} X_k\Big) = \frac{1}{n}\sum_{k=1}^{n} E(X_k) = \frac{n}{n}E(X) = E(X).$$

2. Wir suchen nun einen konsistenten Schätzer für die Varianz. Wegen der Unabhängigkeit der Zufallsvariablen X_i gilt $E(X_iX_k) = E(X_i)\cdot E(X_k)$ für $i \neq k$ und daher:

$$
\begin{aligned}
E\Big(\sum_{k=1}^{n}(X_k-\overline{X})^2\Big) &= E\Big(\sum_{k=1}^{n}X_k^2\Big) - 2E\Big(\sum_{k=1}^{n}X_k\overline{X}\Big) + nE(\overline{X}^2)\\
&= E(n\cdot X^2) - \frac{2}{n}\cdot E\Big(\sum_{k=1}^{n}X_k\sum_{i=1}^{n}X_i\Big) + \frac{n}{n^2}\cdot E\Big(\Big(\sum_{i=1}^{n}X_i\Big)^2\Big)\\
&= n\cdot E(X^2) - \frac{2}{n}\sum_{k=1}^{n}\sum_{i=1,i\neq k}^{n}E(X_kX_i) + \frac{1}{n}\sum_{k=1}^{n}\sum_{k\neq i}^{n}E(X_kX_i)\\
&\qquad -\frac{2}{n}\sum_{k=1}^{n}E(X_k^2) + \frac{1}{n}\sum_{k=1}^{n}E(X_k^2)\\
&= n\cdot E(X^2) - \frac{1}{n}\sum_{k=1}^{n}\sum_{k\neq i}^{n}E(X_k)E(X_i) - 2E(X^2) + E(X^2)\\
&= n\cdot E(X^2) - \frac{1}{n}\cdot(n^2-n)\cdot E(X)^2 - E(X^2)\\
&= (n-1)\cdot(E(X^2)-E(X)^2)\\
&= (n-1)\cdot V(X).
\end{aligned}
$$

Das Ergebnis der beiden Beispiele fassen wir in einem Satz zusammen:

Satz 39.7. *Das Stichprobenmittel und die Stichprobenvarianz sind konsistente Schätzer für den Erwartungswert bzw. die Varianz von X.*

39.3 Parametrisierte Statistik, Konfidenzintervalle

Sei X eine $\mathcal{N}(\mu,\sigma^2)$-verteilte Zufallsvariable. Wir nehmen eine unabhängige Stichprobe $x_1,\ldots,x_n$ für X. Dann schätzt

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n}x_i$$

den Mittelwert μ. Wir suchen eine Zahl $a>0$, so dass μ mit einer Wahrscheinlichkeit von $\gamma = 95\%$ in dem Intervall $[a_1,a_2] = [\overline{x}-a,\overline{x}+a]$ liegt.

Genauer: Liegt $\mu\in[a_1,a_2]$, dann ist die Wahrscheinlichkeit

$$P(\overline{X}\notin[a_1,a_2]) < \alpha = 1-\gamma$$

für die Wahl von Stichproben $x_1,\ldots,x_n$. Das so bestimmte Intervall heißt **Konfidenzintervall** für μ.

Bei der Bestimmung eines Konfidenzintervalles gibt es zwei Fälle: Entweder ist die Standardabweichung σ bekannt oder nicht.

39.3.1 σ bekannt

Wir betrachten zunächst den einfachen Fall, dass σ bekannt ist. Dann ist $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ ebenfalls eine normalverteilte Zufallsvariable, und zwar mit Erwartungswert $E(\overline{X}) = E(X) = \mu$ und der Varianz $V(\overline{X}) = \frac{1}{n}V(X) = \frac{\sigma^2}{n}$, wie man leicht nachrechnen kann.

Wir setzen $U = \overline{x} - \frac{c\cdot\sigma}{\sqrt{n}}$, $W = \overline{x} + \frac{c\cdot\sigma}{\sqrt{n}}$, wobei $c \in \mathbb{R}$ so bestimmt ist, dass

$$\frac{1}{\sqrt{2\pi}} \cdot \int_{-c}^{c} e^{-\frac{x^2}{2}}\, dx = \gamma = 1 - \alpha.$$

Es gibt, beispielsweise im Internet, viele Tabellen, aus denen man für bestimmte α die entsprechenden c ablesen kann. Einige Werte sind:

γ	0.90	0.95	0.975	0.99	0.995
c	1.282	1.645	1.960	2.326	2.576

Dann ist, da X $\mathcal{N}(\mu, \sigma^2)$–verteilt ist, die Zufallsvariable $Y = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu)$ $\mathcal{N}(0,1)$-verteilt und es gilt: $-c \le Y \le c \Leftrightarrow U = \overline{X} - \frac{c\sigma}{\sqrt{n}} \le \mu \le \overline{X} + \frac{c\sigma}{\sqrt{n}} = W$.

Also:

$$\gamma = P(-c \le Y \le c) = P(U \le \mu \le W).$$

Also setzen wir $\overline{X}$ in die Formel für U und W ein.

Beispiel 39.8. $\gamma = 0.95$, $\sigma = 0.03$ sei bekannt. Wir nehmen an, dass folgende Stichprobe $x_1, \ldots, x_8$ vorliegt ($n = 8$):

$$1.21 \;\; 1.15 \;\; 1.18 \;\; 1.23 \;\; 1.24 \;\; 1.19 \;\; 1.18 \;\; 1.20.$$

Es ist: $\overline{x} = 1.1975$, $c = 1.960$ und $U = 1.176$, $W = 1.218$. Mit 95% Wahrscheinlichkeit liegt μ also im Intervall $[1.176; 1.218]$.

Bemerkung 39.9. Möchte man μ genauer wissen, so muss man n vergrößern, denn der Durchmesser des Intervalles ist $2\frac{c\cdot\sigma}{\sqrt{n}}$.

39.3.2 σ unbekannt

Wir sind jetzt in der Situation, dass X $\mathcal{N}(\mu, \sigma^2)$-verteilt ist, dass wir σ^2 aber nicht kennen.

Wieder haben wir eine unabhängige Stichprobe $x_1, \ldots, x_n$. Wir ersetzen das bekannte σ in der Formel des vorigen Abschnitts durch den Schätzer

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Also: $U = \overline{x} - \frac{c \cdot s}{\sqrt{n}}, W = \overline{x} + \frac{c \cdot s}{\sqrt{n}}$.

Wieder müssen wir c geeignet bestimmen, aber wie? Wir dürfen jetzt nicht mehr die Normalverteilungstabelle verwenden, da wir die Varianz ja nicht kennen. Statt dessen müssen wir die sogenannte t_{n-1}-Verteilungstabelle (siehe beispielsweise http://de.wikipedia.org/wiki/Studentsche_t-Verteilung) benutzen. Es gilt nämlich der folgende Satz (Beweisidee folgt gleich):

Satz/Definition 39.10. *Die Zufallsvariable* $Z = \frac{\sqrt{n}}{s}(\overline{X} - \mu)$ *ist eine* t_{n-1}***-verteilte Zufallsvariable*** *(auch* t***-Verteilung mit*** $n-1$ ***Freiheitsgraden*** *genannt), d.h.* Z *hat die Dichte:*

$$\frac{1}{\sqrt{(n-1)\pi}} \cdot \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \cdot \left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}.$$

Bemerkung/Definition 39.11. 1. Im obigen Satz bezeichnet Γ die **Gamma–Funktion**:

$$\Gamma\colon \mathbb{R}_{>0} \to \mathbb{R},\ \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt.$$

Mit partieller Integration kann man einsehen, dass gilt: $\Gamma(x+1) = x \cdot \Gamma(x)$ für $x > 0$. Insbesondere ist, da offenbar $\Gamma(1) = \int_0^\infty e^{-t}\,dt = 1$:

$$\Gamma(n+1) = n! \quad \text{für } n \in \mathbb{N},$$

d.h. die Gamma-Funktion interpoliert die Funktion $n \mapsto (n-1)!$ (siehe Abb. 39.1).

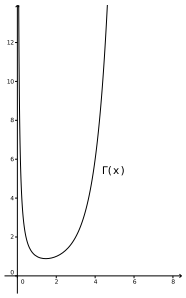


Abbildung 39.1. Die Γ–Funktion im Bereich $[0, 8]$. Da $\Gamma(n) = (n-1)!$ für $n \in \mathbb{N}$, wundert es nicht, dass sie ab $x = 2$ sehr steil ansteigt.

2. Die t_r–Verteilung ist, wie die Normalverteilung, symmetrisch zur y–Achse. Es gilt: $t_\infty = \mathcal{N}(0,1)$. Außerdem ist die t–Verteilung, genauso wie die Normalverteilung, tabelliert; in der Praxis verwendet man häufig schon ab etwa $r > 25$ die Normalverteilung. Abb. 39.2 zeigt die Dichte t_3–Verteilung gemeinsam mit der Dichte der Standard–Normalverteilung in einem Graphen.

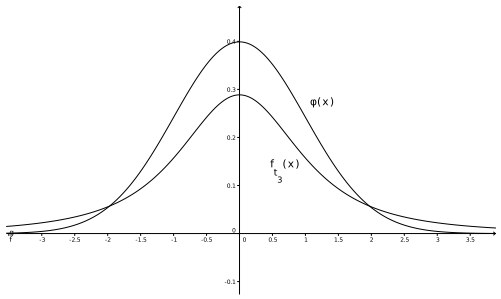


Abbildung 39.2. Die Dichte der t_3–Verteilung gemeinsam mit der Dichte $\varphi(x)$ der Standard–Normalverteilung.

Beispiel 39.12. Wir greifen das Beispiel 39.8 von oben wieder auf. Nun sei σ aber nicht bekannt. Wir müssen daher die t_{n-1}–Verteilung für $n = 8$ verwenden,

$$\begin{array}{r|c|c|c|c|c} \gamma = 1-\alpha & 0,995 & 0,990 & 0,975 & 0,95 & 0,900 \\ \hline c\colon F_{t_7}(c) = \gamma & 3,499 & 2,998 & 2,365 & 1,895 & 1,415 \end{array},$$

und σ schätzen: Da weiterhin $\overline{x} = 1.1975$, ergibt sich

$$s^2 = \frac{1}{7} \cdot \left((1.21 - 1.1975)^2 + \cdots + (1.20 - 1.1975)^2 \right) = \frac{1}{7} \cdot 0.00595.$$

Damit ist $s \approx 0,2915$ und $c = 1,895$ ($< 1,960$), so dass das Konfidenzintervall etwas größer wird, was nicht erstaunt, da wir die Varianz ja nur geschätzt haben und daher die Unsicherheit über unsere Aussage größer wird: $U \approx 1,00199$, $W \approx 1.39301$.

Beweis (des Satzes 39.10, nur die Idee!). Wir setzen $Y = \frac{1}{\sigma}(X - \mu)$. Diese neue Zufallsvariable ist $\mathcal{N}(0,1)$–verteilt. Damit gilt:

$$Z = \frac{\sqrt{n}}{s}(\overline{X} - \mu) = \frac{\sqrt{n} \cdot (\bar{X} - \mu)}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2}} = \frac{\sqrt{n} \cdot \overline{Y}}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}.$$

Die $Y_1, \ldots, Y_n$ sind hierbei $\mathcal{N}(0,1)$–verteilt mit (wie man zeigen kann) gemeinsamer Dichte

$$\frac{1}{(\sqrt{2\pi})^n} \cdot \int \cdots \int e^{-\sum \frac{-y_i^2}{2}} dy_1 \ldots dy_n.$$

Wir wählen eine Orthonormalbasis des $\mathbb{R}^n$ mit $a_0 = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$ und $a_1, \ldots, a_{n-1} \in \mathbb{R}^n$. Damit setzen wir:

$$T_0 = a_0 \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \sqrt{n} \cdot \overline{Y}, \ldots, T_i = a_i \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ldots.$$

Dann sind auch $T_0, \ldots T_{n-1}$ $\mathcal{N}(0,1)$–verteilt, da die gemeinsame Dichte rotationsinvariant ist. Mit diesen neuen Zufallsvariablen schreibt sich Z, wie man leicht nachrechnen kann, als

$$Z = \frac{T_0}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n-1} T_i^2}}.$$

Wir setzen nun:

$$F(z, t_1, \ldots, t_{n-1}) := \left(z \cdot \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n-1} t_i^2}, t_1, \ldots, t_{n-1}\right) = (t_0, t_1, \ldots, t_{n-1}).$$

Die Jacobi–Matrix dieser Abbildung ist

$$DF = \left(\begin{array}{c|cc} \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n-1} t_i^2} & * & \\ \hline 0 & 1 & 0 \\ \vdots & & \ddots \\ 0 & 0 & 1 \end{array}\right).$$

Die Transformationsformel liefert jetzt:

$$\begin{aligned} \text{Dichte} &= \text{konst.} \cdot \int \cdots \int e^{-\frac{1}{2}\sum T_i^2} \, dt_0 \, \cdots \, dt_{n-1} \\ &= \text{konst.} \cdot \left(e^{-\frac{z^2}{n-1}\sum_{i=1}^{n-1} t_i^2 + \sum_{i=1}^{n-1} t_i^2}\right) \cdot \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n-1} t_i^2} \, dz \, dt_1 \, \ldots \, dt_{n-1}. \end{aligned}$$

Wählen wir nun Kugelkoordinaten für $t_1, \ldots, t_{n-1}$, so ergibt sich mit etwas Rechnung:

$$\cdots = \text{konst.} \cdot \int_0^\infty \left(e^{-(\frac{z^2}{n-1}+1)r^2}\right) \cdot r^{n-1}\, dr.$$

Partielle Integration liefert, da $r^{n-1} = r^{n-2} \cdot r$:

$$\begin{aligned}\cdots &= \text{konst} \cdot \int_0^\infty r^{n-3} \cdot \left(1 + \frac{z^2}{n-1}\right)^{-1} \cdot e^{(1+\frac{z^2}{n-1})r^2}\, dr \\ &= \cdots = \text{konst} \cdot \left(1 + \frac{z^2}{n-1}\right)^{-\frac{n}{2}}.\end{aligned}$$

Man kann berechnen, dass dies in der Tat die behauptete Konstante ist. □

Weiß man, dass $c_1 \cdot f(x)$ und $c_2 \cdot f(x)$ beides Dichten sind, so folgt, da das Integral über beide genau 1 ergeben muss, dass die konstanten Vorfaktoren übereinstimmen: $c_1 = c_2$. So muss man beispielsweise im vorigen Beweis die Konstante nicht genau ausrechnen, wenn man nur nachprüft, dass die behauptete Formel eine Dichte liefert.

39.4 Tests auf den Erwartungswert

Leider können wir nur vernünftige Tests durchführen, wenn wir gewisse Annahmen über die Verteilung der gegebenen Daten machen. Eine relativ naheliegende Annahme ist oft, dass die Daten unabhängig identisch normal–verteilt sind (in der englischen Literatur wird **unabhängig identisch verteilt** mit **independent and identically distributed** übersetzt und häufig mit **i.i.d.** abgekürzt), entweder mit einer bekannten Varianz oder (meist realistischer) mit einer unbekannten Varianz.

39.4.1 Zweiseitiger Test

Sei also X eine $\mathcal{N}(\mu, \sigma^2)$–verteilte Zufallsvariable. Wir nehmen unabhängige Stichproben $x_1, \ldots, x_n$ zum Testen der Hypothese $H_0\colon \mu = \mu_0$, $H_1\colon \mu \neq \mu_0$ (**zweiseitiger Test**) mit einer Irrtumswahrscheinlichkeit von α für den Fehler 1. Art und betrachten

$$U = \mu_0 - \frac{c \cdot \sigma_1}{\sqrt{n}},\ W = \mu_0 + \frac{c \cdot \sigma_1}{\sqrt{n}},$$

wobei entweder $\sigma_1 = \sigma$ eine bekannte Streuung oder $\sigma_1 = s$ die **Stichprobenstreuung** (also ein Schätzer für die Streuung) ist. Wir nehmen H_0 an, wenn $U \leq \bar{x} \leq W$ gilt. Dabei bestimmen wir c als sogenanntes $\frac{\alpha}{2}$–**Quantil**, entweder von der $\mathcal{N}(0,1)$– oder von der t_{n-1}–Verteilung, je nachdem ob σ bekannt ist oder nicht.

Dies bedeutet Folgendes: Wir nehmen H_0 an, wenn

$$\left| \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{\sigma_1} \right| \leq |c|,$$

sonst lehnen wir H_0 ab. Dass dies die richtige Wahl von c ist wird klar, wenn wir uns die (symmetrische!) Dichte der Normalverteilung ansehen (Abb. 34.1): Die Wahrscheinlichkeit für den Fehler 1. Art ist nach Definition die Fläche unter den Kurvenbereichen, die einen Abstand von mehr als c von μ haben. Notieren wir die Fläche von $-\infty$ bis c mit $\Phi(c)$, so ist dies wegen der

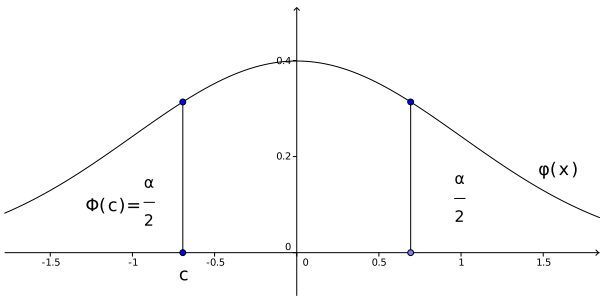


Abbildung 39.3. Der Fehler 1. Art beim zweiseitigen Test.

Symmetrie der Dichte der Normalverteilung gerade $2\Phi(c)$. Wir wollen diese Fläche nun auf α beschränken; wir bestimmen c also so, dass $2\Phi(c) = \alpha$ bzw. $\Phi(c) = \frac{\alpha}{2}$ gilt.

39.4.2 Einseitiger Test

Auf ähnliche Weise können wir auch die einseitigen Tests behandeln. Betrachten wir zunächst den ersten Fall: $H_0 \colon \mu \leq \mu_0$ gegen $H_1 \colon \mu > \mu_0$. Wir bilden dazu die Testgröße

$$t = \sqrt{n} \cdot \frac{\overline{x} - \mu_0}{\sigma_1},$$

wobei σ_1 wie weiter oben ist (d.h. entweder σ oder s).

Wir nehmen H_0 an, wenn $t \leq c$, wobei c das α–Quantil (siehe Abb. 39.4) der entsprechenden Verteilung ($\mathcal{N}(0,1)$ oder t_{n-1}) ist. Andernfalls lehnen wir H_0 ab.

Im umgekehrten Fall, $H_0 \colon \mu \geq \mu_0$ gegen $H_1 \colon \mu < \mu_0$, ergibt sich Folgendes: Wir lehnen die Nullhypothese ab, wenn $t \geq d$ für ein gewisses d. Im

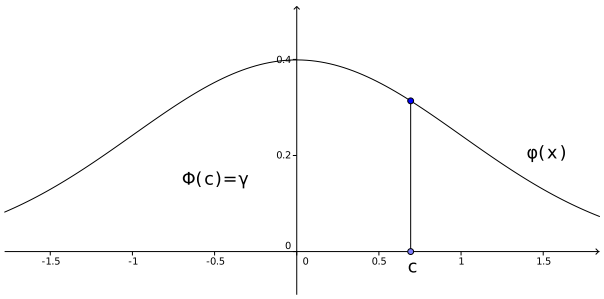


Abbildung 39.4. Der Fehler 1. Art beim einseitigen Test.

Normalverteilungs–Fall ist es gerade das d, für das gilt: $\Phi(d) = \alpha$. Wegen der Symmetrie der Dichte ist dieses aber gerade $d = -c$, da

$$\Phi(d) = \alpha \iff 1 - \Phi(d) = 1 - \alpha = \Phi(c).$$

Das $(1-\alpha)$–Quantil der Standard-Normalverteilung ist also gerade das Negative des α–Quantils der Standard-Normalverteilung.

Da die t_{n-1}–Verteilung die gleiche Symmetrieeigenschaft wie die Standard-Normalverteilung hat, lässt sich die obige Argumentation genauso anwenden, wenn die Varianz nicht als bekannt angenommen wird.

Wollen wir eine Hypothese der Form $H_0 \colon \mu = \mu_0$ gegen $H_1 \colon \mu > \mu_0$ testen (wir interessieren uns also gar nicht für die Möglichkeit $\mu < \mu_0$), so können wir identisch zu den obigen einseitigen Tests vorgehen. Ein Test wird nämlich festgelegt durch seinen Ablehnungsbereich; und dieser ist, wie man sieht, in diesem neuen Test identisch mit dem oben besprochenen.

Beispiel 39.13. ...

39.5 χ^2–Test auf die Varianz

Wieder nehmen wir an, X sei eine $\mathcal{N}(\mu, \sigma^2)$–verteilte Zufallsvariable. Dann sei $x_1, \ldots, x_n$ eine unabhängige Stichprobe. Wir betrachten die Hypothesen $H_0 \colon \sigma = \sigma_0$ gegen $H_1 \colon \sigma > \sigma_0$.

Wir nehmen H_0 zum Niveau $\gamma = 1 - \alpha$ an, wenn die Testgröße

$$Z = \frac{s^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} \cdot \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

die Ungleichung $Z \leq c$ erfüllt, wobei c das α–**Fraktil** der sogenannten χ^2_{n-1} **Verteilung** (auch: χ^2**–Verteilung mit** $n-1$ **Freiheitsgraden** genannt) ist; ein α–Fraktil ist das $(1-\alpha)$–Quantil.

Die Dichte von χ^2_r auf $\mathbb{R}_{\geq 0}$ ist (siehe auch Abb. 39.5):

$$f_{\chi^2_r}(x) = \frac{1}{2^r \Gamma(\frac{r}{2})} \cdot x^{\frac{r}{2}-1} \cdot e^{-\frac{x}{2}}.$$

Auch diese Verteilung ist für verschiedene Werte von r und α tabelliert; siehe z.B. http://de.wikibooks.org/wiki/Mathematik:_Statistik:_Tabelle_der_Chi-Quadrat-Verteilung.

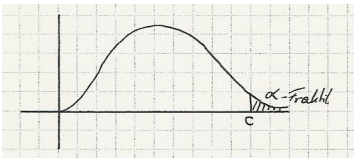


Abbildung 39.5. Das α–Fraktil der χ^2–Verteilung

Bemerkung 39.14. Unter der Annahme H_0 ist $\frac{x_i - \overline{x}}{\sigma_0}$ eine $\mathcal{N}(0,1)$–verteilte Zufallsvariable. Wegen der Abhängigkeit $\sum_{i=1}^n (x_i - \overline{x}) = 0$ ist die Summe

$$\sum_{i=1}^n \frac{(x_i - \overline{x})^2}{\sigma_0^2}$$

also eine Summe von $n-1$ Quadraten von unabhängigen $\mathcal{N}(0,1)$–verteilten Zufallsvariablen.

Lemma 39.15. *Seien $U_1, \ldots, U_n$ unabhängige $\mathcal{N}(0,1)$–verteilte Zufallsvariablen. Dann ist*

$$Y = \sum_{i=1}^n U_i^2$$

eine χ^2_{n-1}–verteilte Zufallsvariable.

Beweis (nur die Idee!). Der Beweis verläuft ähnlich wie bei der t–Verteilung. Wir beginnen mit der gemeinsamen Dichte von $(U_1, \ldots, U_n)$:

$$\Big(\frac{1}{\sqrt{2\pi}}\Big)^n \cdot e^{-\frac{t_1^2}{2}} \cdot e^{-\frac{t_2^2}{2}} \cdots e^{-\frac{t_n^2}{2}}.$$

Für $M \subset \mathbb{R}^n$ ist die Wahrscheinlichkeit

$$P\Big((U_1, \ldots, U_n) \in M\Big) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int \int_M \cdots \int e^{-(t_1^2 + \cdots + t_n^2)} \, dt_1 \, \cdots \, dt_n$$

Der weitere Beweis verläuft ähnlich wie bei der t–Verteilung; insbesondere gehen hier wiederum Kugelkoordinaten und damit die Transformationsformel essentiell ein. □

Beispiel 39.16. ...

39.6 χ^2–Verteilungstest

Jetzt möchten wir überprüfen, ob eine Zufallsvariable tatsächlich eine vermutete Verteilung hat. Wir beschränken uns also hier nicht mehr nur auf die Normalverteilung.

Sei dazu X zunächst eine Zufallsvariable mit nur endlich vielen Werten $\in \{1, \dots, k\}$.

$$H_0 \colon P(X = i) = p_i.$$

Wir wollen H_0 testen mit einem Fehler 1. Art $\leq \alpha$. Wir nehmen dafür eine Stichprobe $X_1, \dots, X_n$ (unabhängige Zufallsvariablen) für X und setzen

$$Z_i = \left|\left\{j \in \{1, \dots, n\} \mid X_j = i\right\}\right|.$$

Dann gilt: $E(Z_i) = n \cdot p_i$, $V(Z_i) = n \cdot p_i(1 - p_i)$, da Z_i eine B_{n,p_i}–verteilte Zufallsvariable ist. Wir bilden

$$Y = \sum_{i=1}^{k} \frac{(Z_i - np_i)^2}{np_i}.$$

Für große n ist unter der Hypothese H_0 die Zufallsvariable $\sqrt{\frac{(Z_i - np_i)^2}{np_i}}$ annähernd $\mathcal{N}(0,1)$–verteilt, nach dem zentralen Grenzwertsatz. Genauer: Die Approximation ist recht gut, falls $np_i \geq 5 \; \forall i$.

Lemma 39.17. *Y ist annähernd χ^2_{k-1}–verteilt.*

Beweis. Obiges Lemma 39.15 zusammen mit Bemerkung 39.14. □

Wir lehnen H_0 also ab, wenn für die Testgröße Y gilt: $Y > c$, wobei c das α–Fraktil der χ^2_{k-1}–Verteilung ist.

Beispiel 39.18. Test auf einen fairen Würfel. X sei eine Zufallsvariable mit Werten in $\{1, 2, \dots, 6\}$. $H_0 \colon P(X = i) = \frac{1}{6}$, $i = 1, \dots, 6$.

Wir machen einen Versuch mit 1020 Würfen:

Augenzahl	1	2	3	4	5	6
Anzahl	195	151	148	189	189	154

Unter H_0 ist $E(Z_i) = np_i = \frac{1020}{6} = 170$. Damit ergibt sich:

$$Y = \frac{(195 - 170)^2 + \dots + (154 - 170)^2}{1020 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 13.2.$$

Nach dem obigen Lemma müssen wir dies mit dem α–Fraktil von χ^2_5 vergleichen:

γ	0,995	0,990	0,975	0,95	0,900
$q\colon F_{\chi^2_5}(q) = \gamma$	16.75	15.09	12.83	11.07	9.24

Für $\alpha = 5\%$ ist $c \approx 11.07 < 13.2$ (wir lehnen H_0 also ab); für $\alpha = 1\%$ ist $c \approx 15.09 > 13.2$ (wir nehmen H_0 also an).

39.7 χ^2–Test auf Unabhängigkeit

X, Y seinen Zufallsvariablen mit Werten in $\{0, 1\}$ (d.h. man teilt die Werte der Zufallsvariablen in zwei **Kategorien** ein). Wir testen diese auf Unabhängigkeit. Gegeben sei dazu eine unabhängige Stichprobe (x_k, y_k), $k = 1, \ldots, n$. Wir setzen $Z_{ij} :=\mid \{k \mid x_k = i, y_k = j\} \mid$, $i = 0, 1$, und die Testgröße

$$W := \frac{n \cdot (Z_{00}Z_{11} - Z_{01}Z_{10})^2}{(Z_{00} + Z_{01})(Z_{00} + Z_{10})(Z_{11} + Z_{01})(Z_{11} + Z_{10})}.$$

Man kann zeigen, dass dann W für große n etwa χ^2_1–verteilt ist; einige Werte dazu:

γ	0,995	0,990	0,975	0,95	0,900
$q\colon F_{\chi^2_1}(q) = \gamma$	7,879	6,635	5,024	3,841	2,706

Dies können wir benutzen, um zu testen.

Beispiel 39.19. Wir teilen die Menschheit unter zwei unterschiedlichen Aspekten jeweils in zwei Kategorien ein: Geschlecht (männlich/weiblich), Rauchgewohnheit (Raucher/Nichtraucher).

Gegeben seien die folgenden Daten:

	Raucher	Nichtraucher	$\sum$
männlich	113	137	250
weiblich	77	73	150
$\sum$	190	210	400

Es ergibt sich:

$$W = \frac{400 \cdot (113 \cdot 73 - 77 \cdot 137)^2}{190 \cdot 210 \cdot 250 \cdot 150} \approx 1.4142.$$

Ist $W \leq c$, so können wir die Hypothese H_0 der Unabhängigkeit der Merkmale Rauchverhalten und Geschlecht nicht ablehnen. Für die Signifikanzniveau $\alpha = 5\%$, also $\gamma = 95\%$, ist dies der Fall, weil $1.4142 < 3.841$.

Für eine andere Stichprobe hat sich folgendes ergeben:

	Raucher	Nichtraucher	$\sum$
männlich	98	152	250
weiblich	77	73	150
$\sum$	175	225	400

In diesem Fall ist

$$W = \frac{400 \cdot (98 \cdot 73 - 77 \cdot 152)^2}{175 \cdot 225 \cdot 250 \cdot 150} \approx 5.608,$$

so dass wir die Nullhypothese der Unabhängigkeit ablehnen müssten, da $5.608 > 3.841$.

Man kann den Test auf Unabhängigkeit auch für beliebig viele Kategorien durchführen, etwa, wenn X in n und Y in r Kategorien eingeteilt werden. Dann wird die Formel für W komplizierter und W ist $\chi^2_{(m-1)(r-1)}$–verteilt. Dies führen wir hier im Detail aber nicht vor. Ein ausführlicheres Beispiel ist auf http://de.wikipedia.org/wiki/Chi-Quadrat-Test zu finden.

Aufgaben

Aufgabe 39.1 (Bolzenmaschine testen). Ein Drehautomat fertigt Bolzen. Es ist bekannt, dass der Durchmesser der von dem Automaten gefertigten Bolzen (in mm) normalverteilt ist mit Varianz $\sigma^2 = 0,26$. Eine Stichprobe von 500 Bolzen ergab einen mittleren Durchmesser von $\overline{x} = 54,03$ mm. Testen Sie mit diesen Daten die Nullhypothese $H_0 : \mu = 54$ auf dem Signifikanzniveau $\alpha = 1\%$.

Aufgabe 39.2 (Tablettengewicht testen). Wir wiegen 8 Tabletten und erhalten die folgenden Massen in Gramm:

$$1.19, 1.23, 1.18, 1.21, 1.27, 1.17, 1.15, 1.14.$$

1. Testen Sie die Hypothese, dass das Durchschnittsgewicht der Tabletten 1.2 g beträgt, zur Irrtums- wahrscheinlichkeit 5%.
2. Es wird vermutet, dass die Tabletten im Mittel weniger als 1.2 g wiegen. Testen Sie auch diese Hypothese zur Irrtumswahrscheinlichkeit 5%.

Aufgabe 39.3 (Test auf Verteilung bei Kreuzungen). Wir kreuzen weiß- und rot-blühende Erbsen, so dass sich rosa-blühende Pflanzen ergeben. Kreuzen wir weiter nun rosa-blühende miteinander, so sollten sich nach den Mendelschen Regeln der Genetik rot-, rosa- und weißblühende Erbsen im Verhältnis $1 : 2 : 1$ ergeben. Unsere 200 Tests ergaben die Häufigkeiten: 52, 107, 41. Liegen die Abweichungen bei einer Irrtumswahrscheinlichkeit von 5% im Zufallsbereich?

40

Robuste Statistik

Wir betrachten eine Zufallsvariable X_ε der Form $X_\varepsilon = X + \varepsilon$, wobei z.B. X $\mathcal{N}(\mu, \sigma^2)$–verteilt ist und ε irgendwie verteilt ist. Dieses ε ist typischerweise ein seltener großer Fehler, der bei der Datenerhebung entstehen kann. Wie schätzt man μ in diesem Fall?

Wir nehmen eine Stichprobe $x_1, \dots, x_n$. Der **Mittelwert** (**arithmetisches Mittel**) $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, der empfindlich auf einzelne große Fehler reagiert, ist nicht robust.

Beispiel 40.1. 29 Schüler möchten die Temperatur der Saar an einer Stelle möglichst genau ermitteln. Alle messen. Bei der Auswertung im Klassenraum stellt sich heraus, dass 26 der Schüler eine Temperatur in der Gegend von 7° ermittelt haben, aber drei Schüler Temperaturen um 17° gemessen haben.

Das arithmetische Mittel aus allen 29 Werten würde durch die drei wohl falsch gemessenen Werte recht weit vom realen Wert entfernt liegen.

Gibt es einen Mittelwert, der weniger stark oder gar nicht auf solche Ausreißer reagiert?

Die robuste Version des Mittelwertes ist der Median:

Definition 40.2. *Sind $x_1, \dots, x_n \in \mathbb{R}$, so ist der **Median** folgendermaßen definiert:*

$$\text{Median}(x_1, \dots, x_n) = \begin{cases} a_{\frac{n+1}{2}}, & n \text{ ungerade,} \\ \frac{a_{\frac{n}{2}} + a_{\frac{n}{2}+1}}{2}, & n \text{ gerade,} \end{cases}$$

wobei $a_1 \leq \dots \leq a_n$ die der Größe nach sortierten Werte $x_1, \dots, x_n$ sind.

Wir können hier nicht beweisen, dass dies manchmal ein sinnvollerer Mittelwert ist als das altbekannte arithmetische Mittel; es hängt selbstverständlich auch von der gegebenen Situation ab. Daher verdeutlichen wir dies im Folgenden nur an einigen Beispielen.

Beispiel 40.3. Wir geben jeweils ein Beispiel mit gerade und mit ungerade vielen Werten:

- Ein Beispiel mit einem Ausreißer bei fünf Werten:

$$\begin{aligned}\text{Median}(9,14,10,12,20) &= \text{Median}(9,10,12,14,20)\\ &= 12\\ &= \text{Median}(9,14,10,12,40).\end{aligned}$$

  Der Median ist also unabhängig vom höchsten Wert der Stichprobe (20 bzw. 40). Ganz im Gegensatz zum arithmetischen Mittel:

$$\frac{9+14+10+12+20}{5} \neq \frac{9+14+10+12+40}{5}.$$

- Ein Beispiel mit einem Ausreißer bei sechs Werten:

$$\begin{aligned}\text{Median}(9,9,14,10,12,20) &= \text{Median}(9,9,10,12,14,20) = \frac{10+12}{2}\\ &= 11\\ &= \frac{10+12}{2} = \text{Median}(9,9,14,10,12,40),\end{aligned}$$

  aber

$$\frac{9+9+14+10+12+20}{6} \neq \frac{9+9+14+10+12+40}{6}.$$

Beispiel 40.4. Bei der Studiendauer berücksichtigt der Mittelwert Langzeitstudenten. Der Median ignoriert diese Ausreißer, was die Studiensituation besser beschreibt (falls es wirklich nur einige wenige Ausreißer sind ;-)). Genauer gesagt: es ist beim Median egal, wie lang die Studiendauer der Langzeitstudenten ist, d.h. ob sie beispielsweise 15, 20 oder 50 Semester beträgt.

Um etwas mehr Informationen auf einen Blick zu geben als nur einen Mittelwert und um auch über die Streuung der Werte grob Auskunft zu geben, wurde der sogenannte Boxplot entwickelt:

Beispiel/Definition 40.5. Abb. 40.1 zeigt eine übliche Visualisierungsmöglichkeit, den sogenannten **Boxplot**, zu den Zahlen aus Beispiel 40.3. Dieser zeigt eine Strecke vom **Minimum der aufgetretenen Werte** bis zum **Maximum der aufgetretenen Werte** sowie eine Box zwischen **unteren Quartil** und **oberen Quartil**, d.h. der Stellen, für die jeweils 25% der Werte $\leq$ bzw. $\geq$ diesen Werten sind. Außerdem wird der Median markiert.

Ein Boxplot hat mehrere Vorteile: Er informiert einerseits über einen robusten Mittelwert, andererseits aber auch über die extremsten Ausreißer und auch darüber, in welchem Bereich die Hälfte der Messwerte lag. Zwar sind Median und Boxplot inzwischen Bestandteil des Schulstoffes, in der öffentlichen Diskussion werden sie allerdings erstaunlich selten verwendet.

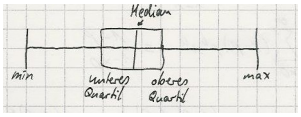


Abbildung 40.1. Der Median ignoriert Ausreißer.

Wir sehen also: Je nach gegebener Situation muss entschieden werden, ob das arithmetische Mittel oder der Median ein sinnvollerer Mittelwert ist.

Aufgaben

Aufgabe 40.1 (Robuste Statistik).

1. Gegeben sei folgende Situation: Eine Gruppe von 6 Studierenden möchte herausfinden, zu welcher Uhrzeit der typische Studierende den Nebenjob beginnt. Es ergeben sich folgende Anfangszeiten:

   9.00 10.00 8.00 9.00 10.00 22.00

   Ist eher das arithmetische Mittel oder der Median geeignet, die typische Anfangszeit zu ermitteln? Wieso?
2. Zeichnen Sie einen Boxplot für die vorige Situation und beschreiben Sie die wesentlichen Eigenschaften des Datensatzes schriftlich.

41

Stochastische Prozesse

Stochastische Prozesse sind grob gesagt Prozesse, die sich zufällig entwickeln. Dabei hängen die Zustände eines Prozesses oft von einem oder mehreren Vorzuständen ab. Wir werden uns fast nur mit Prozessen beschäftigen, bei denen ein Zustand nur vom vorigen Zustand abhängt, sogenannten Markovketten.

Definition 41.1. *Ein **stochastischer Prozess** (X_t) ist eine Familie von reellen Zufallsvariablen, die von einem Parameter $t \in \mathbb{R}$ oder $t \in \mathbb{N}$ abhängt. Wir denken dabei bei t an die Zeit, die kontinuierliche oder diskrete Zeittakte hat. Die Menge der Werte, die die X_t annehmen, heißt die **Menge der Zustände** des Prozesses.*

Beispiel 41.2. Y_k sei B_{n,p_k}–verteilt, z.B.: $Y_k \underset{\text{st}}{=} Y$, wobei Y $B_{1,\frac{1}{2}}$–verteilt ist. Wir setzen $X_n = \sum_{k=1}^{n} Y_k$. Dann ist $(X_n)_{n\in\mathbb{N}}$ ein stochastischer Prozess.

41.1 Markovketten und Stochastische Matrizen

Definition 41.3. *Eine **Markovkette** (oder **Markovscher Prozess**) ist ein diskreter stochastischer Prozess (X_n), der für alle $b_1, \ldots, b_{n-1}$ und c erfüllt:*

$$P(X_n = c \mid X_{n-1} = b_{n-1}, \ldots, X_1 = b_1) = P(X_n = c \mid X_{n-1} = b_{n-1}).$$

Bei einer Markovkette hängt die Wahrscheinlichkeit also immer höchstens vom vorigen Schritt ab. Die wichtigste Klasse sind solche Markovketten (X_n), bei denen alle X_n nur endlich viele Werte $\{1, \ldots, k\}$ (genannt Zustände) annehmen, so dass also $P(X_n \in \{1, \ldots, k\}) = 1$ $\forall n$ gilt.

Beispiel/Definition 41.4. Ein Prozessor bearbeitet pro Zeittakt eine Aufgabe. In jedem Zeitschritt kommen Aufgaben hinzu. In den Cache passen ≤ 2 Aufgaben, bei 2 vorliegenden Aufträgen wird also ein weiterer ignoriert.

A_n sei die Anzahl der zusätzlichen Anfragen im Zeitschritt n. Wir nehmen an, dass $A_n \underset{\text{st}}{=} A$, wobei A eine **geometrisch verteilte Zufallsvariable** ist, d.h.

$$P(A = 0) = 1 - p,\ P(A = 1) = p \cdot (1 - p)$$

und allgemein:

$$P(A = k) = p^k \cdot (1 - p).$$

Es gilt dann

$$P(A \geq 1) = (1 - p) \cdot \sum_{i=1}^{\infty} p^i = 1 - (1 - p) = p,$$

so dass tatsächlich $P(A \in \mathbb{Z}_{\geq 0}) = 1 - p + p = 1$ erfüllt ist. Es ist leicht, andere Werte auszurechnen, z.B.: $P(A \geq 2) = p^2$.

Wir möchten den Erwartungswert von A bestimmen. Wir haben bereits gesehen, dass gilt: $E(A) = (xF'_A(x))_{x=1}$. Um dies zu berechnen benötigen wir:

$$F_A(x) = \sum_{k=0}^{\infty} (1 - p) \cdot p^k \cdot x^k = \frac{1 - p}{1 - xp},$$

wegen der geometrischen Reihe. Damit folgt:

$$E(A) = \left(x \cdot \frac{-(1 - p) \cdot (-p)}{(1 - xp)^2}\right)\Big|_{x=1} = \frac{(1 - p) \cdot p}{(1 - p)^2} = \frac{p}{1 - p}.$$

Im Mittel fallen also in einem Zeitschritt zusätzlich $\approx \frac{p}{1-p}$ Anfragen an. Nur für $p < \frac{1}{2}$ ist die erwartete Anzahl neuer Anfragen pro Zeittakt also kleiner als 1, so dass der Prozessor Chancen hat, die Anfragen zu bearbeiten.

Beispiel 41.5. Manchmal ist ein graphisches Modell hilfreich. Beispielsweise eines wie in Abb. 41.1.

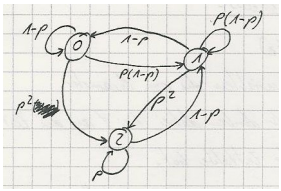


Abbildung 41.1. Graphisches Modell einer Markovkette.

Definition 41.6. *Seien $X_n \in \{1, 2, \dots, k\}$ Zustände einer Markovkette. Wir schreiben für die Übergangswahrscheinlichkeiten $p_{ji} = P(X_n = j \mid X_{n-1} = i)$. Die Matrix*

$$M_n = (p_{ij}^n) \in \mathbb{R}^{k \times k}$$

heißt dann die ***Matrix der Übergangswahrscheinlichkeiten*** *im n–ten Schritt. Hängt M_n nicht von n ab, so heißt (X_n)* ***zeitschrittunabhängige Markovkette****.*

Beispiel 41.7. Zum Beispiel 41.5 von eben ist die Übergangsmatrix

$$M = \begin{pmatrix} 1-p & 1-p & 0 \\ p(1-p) & p(1-p) & 1-p \\ p^2 & p^2 & p \end{pmatrix}.$$

Ist nun $\pi^0 = (\pi_1^0, \pi_2^0, \pi_3^0)^t$ eine Verteillung für X_0. Dann hat X_1 die Verteilung $M\pi^0$, X_2 die Verteilung $M^2\pi^0$ usw.

Ist nun allgemeiner (X_n) ein zeitschrittunabhängiger Markovscher Prozess mit k Zuständen, so notieren wir mit M die Matrix der Übergangswahrscheinlichkeiten und mit $\pi^0 = (\pi_1^0, \dots, \pi_k^0)^t$ die Anfangsverteilung auf den k Zuständen.

Frage 41.8. *1. Konvergiert die Folge von Vektoren $M^n\pi^0$ gegen eine Grenzverteilung*

$$\lim_{n\to\infty} M^n\pi^0 = \pi^\infty \in \mathbb{R}^k \ ?$$

2. Gilt

$$M^n = (\pi^\infty, \dots, \pi^\infty) \in \mathbb{R}^{k\times k}$$

für einen gewissen Vektor $\pi^\infty \in \mathbb{R}^k$?

Klar ist, dass aus einem Bejahen der zweiten Frage auch ein Bejahen der ersten Frage folgt und dass dann außerdem der Grenzwert $\pi^\infty = \lim M^n\pi^0$ nicht von der Ausgangsverteilung π^0 abhängt.

Beispiel 41.9. Abb. 41.2 zeigt einen Graphen, der einen endlichen Markovschen Prozess beschreibt, mit Zuständen, die unterschiedliche Eigenschaften besitzen und die wir nachfolgend definieren werden.

Definition 41.10. *Sei (X_n) eine Markovkette.*

1. Ein Zustand i heißt ***rekurrent****, wenn*

$$P(X_n = i \text{ für } \infty \text{ viele } n) = 1,$$

andernfalls ***transient****. Eine markovsche Kette heißt* ***rekurrent*** *bzw.* ***transient****, wenn jeder Zustand diese Eigenschaft hat.*

Abbildung 41.2. SKIZZE FEHLT!

2. *Eine Teilmenge I der Zustände heißt **absorbierend**, wenn für jedes n gilt:*
$$P(X_{n+1} \in I \mid X_n \in I) = 1.$$
3. *Ein Zustand i heißt **periodisch** mit **Periode** (oder Periode der **Länge**) $l > 0$, wenn für jedes n gilt:*
$$P(X_{n+l} = i \mid X_n = i) = 1.$$

Definition 41.11. *Eine Matrix $M = (p_{ij}) \in \mathbb{R}^{k\times k}$ mit $p_{ij} \in [0,1]$ und $\sum_{i=1}^{k} p_{ij} = 1$ für jedes j heißt **stochastische Matrix**.*

Die Summe der Einträge einer Spalte einer solchen Matrix ist also 1. Sie beschreibt einen endlichen zeitschrittunabhängigen Markovschen Prozess.

Satz 41.12. *Sei $M = (p_{ij})$ eine stochastische Matrix. Dann gilt:*

1. *$\lambda = 1$ ist ein Eigenwert von M.*
2. *$|\lambda| \leq 1$ für alle Eigenwerte von M.*
3. *Ist $d = \min_i(p_{ii}) > 0$, dann sind alle Eigenwerte von M in dem Kreis*
$$\left\{ z \in \mathbb{C} \;\middle|\; |z - d| \leq 1 - d \right\}$$
enthalten (Abb. 41.3). Insbesondere ist in diesem Fall $\lambda = 1$ der einzige Eigenwert λ mit $|\lambda| = 1$.
4. *Ist λ ein Eigenwert mit $|\lambda| = 1$, dann haben alle Jordankästchen von M zu λ die Größe 1 und daher ist $\dim \mathrm{Eig}(M,\lambda) = \mathrm{mult}_\lambda(\det(M - tE_s))$, wobei $E_s \in \mathbb{R}^{s\times s}$ eine Einheitsmatrix und $s = \dim \mathrm{Eig}(M,\lambda) = \mathrm{mult}(\chi_M,\lambda)$ die algebraische Vielfachheit des Eigenwertes λ ist ($\chi_M = \det(M - tE)$ ist das charakteristische Polynom von M).*
5. *Ist $\lambda \in \mathbb{C}$ ein Eigenwert mit $|\lambda| = 1$, so existiert ein m, so dass $\lambda^m = 1$, d.h. λ ist eine sogenannte m–te **Einheitswurzel**.*

Um diesen Satz beweisen zu können, müssen wir zunächst noch einige Eigenwertabschätzungen herleiten.

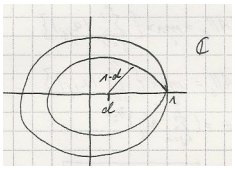


Abbildung 41.3. Die Eigenwerte einer stochastischen Matrix.

41.2 Einschub: Matrixnormen und Eigenwertabschätzungen

Wir möchten Eigenwerte mit möglichst geringem Aufwand abschätzen. Dazu benötigen wir zunächst einiges Wissen über Matrixnormen.

41.2.1 Matrixnormen

Wir haben bereits in der mehrdimensionalen Analysis Matrixnormen verwendet, etwa in Beispiel 30.24. Hier geben wir nun einen detaillierteren Einblick, da wir dies für die folgenden Eigenwertabschätzungen benötigen.

Definition 41.13. *Unter einer **Matrixnorm** verstehen wir eine Abbildung*

$$\|.\|\colon \mathbb{R}^{n\times n} \to \mathbb{R}$$

mit folgenden Eigenschaften:

1. $\|A\| > 0\ \forall A \in \mathbb{R}^{n\times n}$ *und es gilt* $\|A\| = 0 \iff A = 0$,
2. $\|\lambda A\| = |\lambda|\|A\|\ \forall \lambda \in \mathbb{R}, \forall A \in \mathbb{R}^{n\times n}$,
3. $\|A + B\| \leq \|A\| + \|B\|\ \forall A, B \in \mathbb{R}^{n\times n}$,
4. $\|A \cdot B\| \leq \|A\| \cdot \|B\|\ \forall A, B \in \mathbb{R}^{n\times n}\ \forall A, B \in \mathbb{R}^{n\times n}$.

Einige häufig verwendete Matrixnormen sind folgende:

Beispiel/Definition 41.14. Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$.

1. Die **Gesamtnorm**: $\|A\|_G = n \cdot \max_{i,j}\{|a_{ij}|\}$.
2. Die **Zeilensummennorm**: $\|A\|_Z = \max_i\{\sum_{j=1}^n |a_{ij}|\}$.
3. Die **Spaltensummennorm**: $\|A\|_S = \max_j\{\sum_{i=1}^n |a_{ij}|\}$.
4. Die **Frobeniusnorm**: $\|A\|_F = \left(\sum_{i,j=1}^n a_{ij}^2\right)^{\frac{1}{2}}$
5. Die **Spektralnorm**: $\|A\|_2 = \left(\max EW(A^t A)\right)^{\frac{1}{2}}$, wobei $EW(A^t A)$ die Menge der Eigenwerte von $A^t A$ bezeichnet.

Matrixnormen betrachtet man meist im Zusammenhang mit Vektornormen auf $\mathbb{R}^n$. Beide Normen müssen aber zueinander passen:

Definition 41.15. *Sei* $\|.\|_V : \mathbb{R}^n \to \mathbb{R}$ *eine Norm. Die Matrixnorm* $\|.\|_M : \mathbb{R}^{n\times n} \to \mathbb{R}$ *heißt mit der Vektornorm* $\|.\|_V$ ***verträglich****, wenn*

$$\|Ax\|_V \leq \|A\|_M \cdot \|x\|_V \ \forall x \in \mathbb{R}^n, \forall A \in \mathbb{R}^{n\times n}.$$

Beispiel/Definition 41.16. Zu den p-Normen, $p \in [1, \infty]$, auf $\mathbb{R}^n$, d.h.

$$\|x\|_p := \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, & p \in [1, \infty[, \\ \max |x_i|, & p = \infty, \end{cases}$$

sind folgende Matrixnormen verträglich:

1. $\|A\|_G$ und $\|A\|_S$ sind zur **Betragssummennorm** $\|x\|_1$ $(= \sum |x_i|)$ verträglich.
2. $\|A\|_G$, $\|A\|_F$ und $\|A\|_2$ sind zur **euklidischen Norm** $\|x\|_2$ verträglich $(\|x\|_2 = \sqrt{\sum |x_i|^2})$.
3. $\|A\|_G$, $\|A\|_Z$ sind zur **Maximumnorm** $|x|_\infty = \max_i |x_i|$ verträglich.

Beweis. Wir zeigen nur: $\|A\|_G$ und $\|x\|_\infty$ sind verträglich:

$$\begin{aligned} \|Ax\|_\infty &= \max_i \Big|\sum_{j=1}^n a_{ij}x_j\Big|, \\ &\overset{\Delta\text{-Ungl.}}{\leq} \max_i \Big(\sum_{j=1}^n |a_{ij}| \cdot |x_j|\Big) \\ &\leq \sum_{j=1}^n \max_{i,j} |a_{ij}| \cdot \max_i |x_i| \\ &= n \cdot \max_{i,j} |a_{i,j}| \cdot \max_i |x_i| \\ &= \|A\|_G \cdot \|x\|_\infty. \end{aligned}$$

Die anderen Fälle sind ähnlich zu beweisen. □

Satz 41.17. *Sei* $\|.\|_M$ *eine Matrixnorm, die zu einer Vektornorm verträglich ist. Dann gilt für die Eigenwerte* λ *einer Matrix* $A \in \mathbb{R}^{n\times n}$*:*

$$|\lambda| \leq \|A\|_M.$$

Beweis. $\|x\|_V$ sei die Vektornorm, $x \neq 0$ sei ein Eigenvektor zu A mit Eigenwert λ, also: $\lambda x = Ax \Rightarrow \|\lambda x\|_V \leq \|A\|_M \cdot \|x\|_V$. Da $\|\lambda \cdot x\|_V = |\lambda| \cdot \|x\|_V$ ist, folgt: $|\lambda| \leq \|A\|_M$. □

Definition 41.18. *Sei $\|.\|_V$ eine Vektornorm auf $\mathbb{R}^n$. Dann heißt die Matrixnorm*

$$\|A\| = \max_{x\in\mathbb{R}^n, \|x\|_V=1} \|Ax\|_V$$

*die **zu** $\|.\|_V$ **gehörige Matrixnorm**.*

Bemerkung 41.19. Man kann zeigen, dass dies die kleinste Matrixnorm ist, die zu $\|.\|_V$ verträglich ist.

Beispiel 41.20. Wir geben einige Vektornormen und deren zugehörige Matrixnorm an, ohne dies nachzuprüfen:

Vektornorm	zugehörige Matrixnorm
Betragssummennorm $\|x\|_1$	Spaltensummennorm $\|A\|_A$
euklidische Norm $\|x\|_2$	Spektralnorm $\|A\|_2$
Supremumsnorm $\|x\|_\infty$	Zeilensummennorm $\|A\|_Z$

41.2.2 Eigenwertabschätzung

Für jede mit einer Vektornorm verträgliche Matrixnorm $\|.\|\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ und jeden Eigenwert λ von $A \in \mathbb{R}^{n\times n}$ gilt: $|\lambda| \le \|A\|$. Wir werden sehen, dass es noch bessere Abschätzungen für λ gibt.

Beispiel 41.21. Wir betrachten die Matrix

$$A = \begin{pmatrix} 1 & 0.1 & -0.1 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{pmatrix}.$$

Es gilt:

$$\begin{aligned} \|A\|_G &= 3 \max |a_{ij}| = 9, \\ \|A\|_Z &= \max_i \Big\{ \sum_{j=1}^{n} |a_{ij}| \Big\} = \max\{1.2,\ 2.4,\ 3.2\} = 3.2, \\ \|A\|_S &= \max_j \Big\{ \sum_{i=1}^{n} |a_{ij}| \Big\} = \max\{1.2,\ 2.1,\ 3.5\} = 3.5. \end{aligned}$$

Je nach Norm geben sich also sehr unterschiedliche Abschätzungen für λ.

Geometrisch liefert jede Matrixnorm die Information, dass λ in einem Kreis um den Ursprung mit Radius $\|A\|$ liegt. Eine ähnliche, meist bessere, Abschätzung ist folgende:

Satz 41.22 (Gerschgorin). *Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$.*

1. *Die Vereinigung der Kreisscheiben (sogenannte **Gerschgorin–Kreise**)*

$$K_i = \left\{ \mu \in \mathbb{C} \;\middle|\; |\mu - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}$$

   enthält sämtliche Eigenwerte von A.

2. *Jede Zusammenhangskomponente der Vereinigung von genau k dieser Kreise enthält genau k Eigenwerte, gezählt mit Vielfachheit.*

Beweis. Siehe beispielsweise [DS05] oder [SB80, Theorem 6.9.4]. Als Übungsaufgabe zeigen wir eine Variante des Satzes. □

Beispiel 41.23 (zu Bsp. 41.21). Die Gerschgorin–Kreise sind:

$$\begin{aligned} K_1 &= \{\mu \in \mathbb{C} \mid |\mu - 1| \leq 0.2\}, \\ K_2 &= \{\mu \in \mathbb{C} \mid |\mu - 2| \leq 0.4\}, \\ K_3 &= \{\mu \in \mathbb{C} \mid |\mu - 3| \leq 0.2\}. \end{aligned}$$

In jedem der Kreise befindet sich genau ein Eigenwert (s. Abb. 41.4).

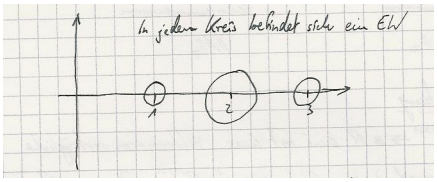


Abbildung 41.4. Drei Gerschgorin-Kreise.

Korollar/Definition 41.24 (Invertierbarkeit von strikt diagonaldominanten Matrizen). *Ist $A \in \mathbb{R}^{n\times n}$ eine Matrix mit der Eigenschaft*

$$|a_{ii}| \;>\; \sum_{k=1, k\neq i}^{n} |a_{ik}| \;\text{ für } i = 1, 2, \ldots, n,$$

*dann ist A invertierbar. Solche Matrizen heißen **strikt diagonaldominant**.*

Beweis. Die 0 ist nach Voraussetzung in keinem der Gerschgorin–Kreise enthalten. □

41.3 Markovketten und Stochastische Matrizen (Teil 2)

Wir sind mit den Vorbereitungen des letzten Abschnittes nun in der Lage, Satz 41.12 über die stochastischen Matrizen zu beweisen.

Beweis (zu Satz 41.12).

1. $\lambda = 1$ ist Eigenwert, da $(1, \ldots, 1)^t \in \mathbb{R}^k$ ein Eigenvektor zum Eigenwert $\lambda = 1$ ist, weil nämlich nach Definition einer stochastischen Matrix die Spaltensummen von M jeweils 1 sind.
2. Es gilt $\|M\|_S = 1$ nach Definition. Daraus folgt die Behauptung.
3. Gerschgorins Satz liefert, da $\sum_{j=1, j\neq i}^{n} p_{ij} = 1 \cdot p_{ii}$, dass die Eigenwerte λ in der Vereinigung der Kreise
$$\left\{ \mu \;\middle|\; |\mu - p_{ii}| \leq 1 - p_{ii} \right\}$$
liegen. Der größte dieser Kreise ist offenbar jener mit $d = \min_i(p_{ii})$ (Abb. 41.5).

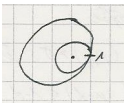


Abbildung 41.5. Der Eigenwert $\lambda = 1$.

Insbesondere ist $\lambda = 1$ der einzige Eigenwert mit $|\lambda| = 1$, da wir oben schon gezeigt haben, dass $|\lambda| \leq 1$ für jeden Eigenwert gilt.
4. Sei $M \in \mathbb{R}^{k\times k}$. Dann existiert nach Satz 24.24 über die Jordansche Normalform eine invertierbare Matrix $T \in \mathrm{GL}(n, \mathbb{C})$, so dass
$$TMT^{-1} = J = \begin{pmatrix} J_{r_1}(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J_{r_s}(\lambda_s) \end{pmatrix}, \text{ wobei } J_r(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \in \mathbb{C}^{r\times r}.$$
Es gilt: $J^N = (T \cdot M \cdot T^{-1})^N = T \cdot M^N \cdot T^{-1}$. Sei nun B ein Jordankästchen von J zum Eigenwert λ. Dann gilt:
$$B = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \Rightarrow B^2 = \begin{pmatrix} \lambda^2 & 2\lambda & 1 & & 0 \\ & \lambda^2 & 2\lambda & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & 2\lambda \\ 0 & & & & \lambda^2 \end{pmatrix}.$$

Allgemeiner ist:

$$B^N = \begin{pmatrix} \lambda^N & N\lambda^{N-1} & & & 0 \\ & \lambda^N & N\lambda^{N-1} & & \\ & & \ddots & \ddots & \\ & & & \ddots & N\lambda^{N-1} \\ 0 & & & & \lambda^N \end{pmatrix}.$$

Falls nun $|\lambda| = 1$, so gilt $|N\lambda^{N-1}| + |\lambda^N| = N + 1$ und daher $N + 1 \leq \|B^N\|_S \leq \|J^N\|_S$. Es folgt:

$$\begin{aligned} N + 1 &\leq \|J^N\|_S \\ &= \|TM^NT^{-1}\|_S = \|T\|_S \cdot \|M\|_S^N \cdot \|T^{-1}\|_S \\ &= \|TM^NT^{-1}\|_S = \|T\|_S \cdot 1^N \cdot \|T^{-1}\|_S = \|T\|_S \cdot \|T^{-1}\|_S \\ &= 1. \end{aligned}$$

Dies ist aber ein Widerspruch, wenn nicht $B = (\lambda)$ eine 1×1–Matrix und damit $B^N = (\lambda^N)$ ist. Alle Jordanblöcke haben also Größe 1 und es gilt

$$\dim \operatorname{Eig}(M, \lambda) = \operatorname{mult}_\lambda(\det(M - tE))$$

für alle λ mit $|\lambda| = 1$.

5. Für diesen Beweis benötigen wir einige Notationen. Es sei $K = \{1, 2, \ldots, 6\}$. Für $j \in K$ sei $M_1(j) = \{i \mid p_{ij} > 0\}$ die Menge der von j in einem Schritt erreichbaren Zustände.

   Für $K_0 \subset K$ sei

   $$M_1(K_0) = \bigcup_{j \in K_0} M_1(j)$$

   die von der Knotenmenge K_0 in einem Schritt erreichbare Knotenmenge. Rekursiv definieren wir nun

   $$M_t(K_0) = M_1(M_{t-1}(K_0))$$

   als die Menge, der in genau t Schritten von K_0 erreichbaren Knoten.

   Sei jetzt λ ein Eigenwert mit $|\lambda| = 1$ und sei $x^t = (x_1, \ldots, x_k)$ ein zugehöriger Eigenvektor von M^t. Seien ferner $y = \max\{|x_j|\}$ und $K_0 = \{j \mid |x_j| = y\} \subset \{1, \ldots, k\}$. Dann gilt für $j \in K_0$:

   $$y = |x_j| = |x_j \cdot \lambda| = \Big|\sum_{i=1}^{k} x_i \cdot p_{ij}\Big| \leq \sum_{i=1}^{k} |x_i| \cdot p_{ij} \leq y \cdot \sum_{i=1}^{k} p_{ij} = y,$$

   so dass überall Gleichheit gelten muss. Insbesondere ist

$$\Big|\sum_{i=1}^{k} x_i \cdot p_{ij}\Big| = \sum_{i=1}^{k} |x_i| \cdot p_{ij}$$

und daher sind für $i, h \in M_1(\{j\})$ (d.h. $p_{ji} \neq 0 \neq p_{jh}$) die Zahlen x_i, x_h komplexe Zahlen, die in die gleiche Richtung zeigen. Also: $x_i = x_h \; \forall\, i, h \in M_1(j)$. Genauer:

$$x_j \cdot \lambda = \sum_{i=1}^{k} x_i \cdot p_{ij} = x_k \cdot \sum_{i=1}^{k} p_{ij} = x_h.$$

Es sei jetzt für $j \in K_0$ fest gewählt. Es gilt: $R = \{k \mid x_k = x_j\} \neq \emptyset$. Dann ist $M_1(R) \neq \emptyset$ und $x_h = \lambda x_j \; \forall h \in M_1(R)$. Für $i \in M_t(R) \neq \emptyset$ gilt $x_i = \lambda x_j{}^{\mathrm{t}}$. Da alle $M_t(R) \neq \emptyset$ und in K enthalten sind, kann die Vereinigung $M_1(R) \cup \cdots \cup M_k(R)$ nicht disjunkt sein. Also existieren Indizes s, t, $s \neq t$, so dass

$$\emptyset \neq M_s(R) \cap M_t(R).$$

Für $i \in M_s(R) \cap M_t(R)$ gilt: $x_i = \lambda^t \cdot x_j = \lambda^s \cdot x_j \;\Rightarrow\; \lambda^{t-s} = 1$.

□

Im Allgemeinen konvergiert die Folge $(M^N)_{N\in\mathbb{N}}$ nicht. In jedem Fall stellt sich heraus, dass $\frac{1}{N}\sum_{k=1}^{N} M^k$ konvergiert. Dies ist Gegenstand des folgenden Satzes.

Satz 41.25 (Ergodensatz). *Sei M eine stochastische Matrix.*

1. *Dann existiert der Grenzwert*

$$Q = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} M^k \;\in \mathbb{R}^{k\times k}$$

und es gilt: $Q^2 = Q = QM = MQ$.

2. *Der Rang $s = \operatorname{rang} Q$ ist die Dimension $\dim \operatorname{Eig}(M, 1)$. Q beschreibt die Projektion auf diesen Eigenraum.*
3. *Ist $\lambda = 1$ der einzige Eigenwert von M mit $|\lambda| = 1$, dann gilt: $Q = \lim_{n\to\infty} M^n$.*
4. *Ist $\lambda = 1$ der einzige Eigenwert von M mit $|\lambda| = 1$ und gilt $\dim \operatorname{Eig}(M, 1) = 1$, so ist*

$$Q = \lim_{n\to\infty} M^n = \begin{pmatrix} z_1 & \cdots & z_1 \\ \vdots & & \vdots \\ z_k & \cdots & z_k \end{pmatrix}$$

mit $z = (z_1, \ldots, z_k)^{\mathrm{t}}$ eine Wahrscheinlichkeitsverteilung mit der Zustandsmenge $\{1, 2, \ldots, k\}$. z ist dabei der eindeutig bestimmte Eigenvektor zum Eigenwert $\lambda = 1$ mit $\sum_{i=1}^{k} z_k = 1$.

Liegt der letzte Fall vor, so nennt man M ***eigenwerteinfach*** *und der Markovsche Prozess konvergiert für jede beliebige Anfangsverteilung $\pi = (\pi_1, \ldots, \pi_k)^t$ gegen die Verteilung z.*

Beweis. 1. Sei $J = T \cdot M \cdot T^{-1}$ die Jordansche Normalform von M. Dann gilt:

$$\frac{1}{n} \cdot \sum_{i=0}^{n-1} J^i = \frac{1}{n} \cdot T \cdot \Big(\sum_{i=0}^{n-1} M^i\Big) \cdot T^{-1}.$$

Wir betrachten daher die Folge $\frac{1}{n} \cdot \sum_{i=0}^{n-1} J^i$. Sei dazu

$$B = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}$$

ein Jordankästchen von J. Wegen der Struktur der Jordanmatrix genügt es nämlich offenbar, die Konvergenz von

$$\lim_{n\to\infty} \frac{1}{n} \cdot \sum_{i=0}^{n-1} B^i$$

zu untersuchen: Wir wissen bereits, dass $|\lambda| \leq 1$, da M stochastisch ist, und betrachten zunächst den Fall $|\lambda| = 1$. Dann ist $B = (\lambda)$ eine 1×1–Matrix nach Teil 4 von Satz 41.12. Im Fall $\lambda = 1$ ergibt sich daher: $\frac{1}{n} \sum_{i=0}^{n-1} \lambda^i = 1$. Ist aber λ eine andere m–te Einheitswurzel, also $\lambda \neq 1$, $\lambda^m = 1$, so ist

$$\sum_{i=0}^{m-1} \lambda^i = \frac{\lambda^m - 1}{\lambda - 1} = 0$$

und daher

$$\Big|\frac{1}{n} \sum_{i=0}^{n-1} \lambda^i\Big| = \frac{1}{n} \cdot \Big|\sum_{i=0}^{m-1} \lambda^i + \sum_{i=m}^{2m-1} \lambda^i + \cdots + \sum_{i=\ldots}^{n-1} \lambda^i\Big| \leq \frac{1}{n} \cdot \Big|0 + \cdots + \sum_{i=n-m}^{n-1} \lambda^i\Big| \leq \frac{m-1}{n}.$$

Die obige Summe konvergiert also gegen 0 für $n \to \infty$.

Wir haben nun noch den letzten Fall $|\lambda| < 1$ zu untersuchen. Es gilt:

$$B = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}, \; B^2 = \begin{pmatrix} \lambda^2 & 2\lambda & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & \ddots & 2\lambda \\ 0 & & & & \lambda^2 \end{pmatrix}$$

und allgemein:

$$B^l = \Big(\lambda E_r + \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}\Big)^l,$$

also:

$$B^l = \lambda^l E_r + l\lambda^{l-1}\begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} + \binom{l}{2}\lambda^{l-2}\begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}^2$$
$$+ \cdots + \binom{l}{r}\lambda^{l-r}\begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}^r.$$

Schließlich ergibt sich:

$$B^l = \begin{pmatrix} \lambda^l & \binom{l}{1}\lambda^{l-1} & \binom{l}{2}\lambda^{l-2} & \cdots & \binom{l}{r-1}\lambda^{l-r+1} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \vdots \\ 0 & & & & \lambda^l \end{pmatrix}.$$

Nun gilt für einen der Einträge, z.B. einen auf der j–ten Nebendiagonalen:

$$\Big|\sum_{l=0}^{n}\binom{l}{j}\lambda^{l-j}\Big| \le \sum_{l=0}^{n}\binom{l}{j}|\lambda|^{l-j}$$
$$\le \sum_{l=0}^{\infty}\binom{l}{j}|\lambda|^{l-j} = \Big[\frac{1}{j!}\frac{d^j}{dx^j}\Big(\frac{1}{1-x}\Big)\Big]_{x=|\lambda|} =: L_j \in \mathbb{C}$$

(da $\frac{1}{1-x} = 1+x+\cdots+x^l+\cdots$ und somit $\frac{d^j}{dx^j}\Big(\frac{1}{1-x}\Big) = \cdots+l(l-1)\cdots(l-j+1)x^{l-j}$).
Diese Summe ist also beschränkt und es folgt:

$$\Big|\frac{1}{n+1}\sum_{l=0}^{n}\binom{l}{j}\lambda^{l-j}\Big| \le \frac{1}{n+1}L_j \underset{n\to\infty}{\longrightarrow} 0.$$

Letztlich ergibt sich somit im Grenzwert also nur für $\lambda = 1$ eine Matrix, die nicht die Nullmatrix ist:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{l=0}^{n-1}J^l=\begin{pmatrix}E_s & 0\\ 0 & 0\end{pmatrix},$$

wobei $s=\dim\mathrm{Eig}(M,1)$ und E_s die $s\times s$-Einheitsmatrix ist. Damit gilt:

$$\begin{aligned}Q&=T^{-1}\Big(\frac{1}{n}\sum_{l=0}^{n-1}J^l\Big)T=T^{-1}\begin{pmatrix}E_s & 0\\ 0 & 0\end{pmatrix}T,\\ Q^2&=T^{-1}\begin{pmatrix}E_s & 0\\ 0 & 0\end{pmatrix}T\,T^{-1}\begin{pmatrix}E_s & 0\\ 0 & 0\end{pmatrix}T\\ &=T^{-1}\begin{pmatrix}E_s & 0\\ 0 & 0\end{pmatrix}^2T=T^{-1}\begin{pmatrix}E_s & 0\\ 0 & 0\end{pmatrix}T=Q.\end{aligned}$$

Schließlich ergibt sich für QM:

$$\begin{aligned}QM&=\lim_{n\to\infty}\Big(\frac{1}{n}\lim_{l=0}^{n-1}M^l\Big)M=\lim_{n\to\infty}\frac{1}{n}\sum_{l=0}^{n-1}M^{l+1}\\ &=\lim_{n\to\infty}\frac{1}{n+1}\Big(\sum_{l=0}^{n}M^l\Big)=Q.\end{aligned}$$

Die vorletzte Gleichheit folgt dabei aus $\lim_{n\to\infty}\frac{n}{n+1}=1$ und $\lim_{n\to\infty}\frac{1}{n}M^0=0$. Die umgekehrte Richtung $MQ=Q$ lässt sich analog zeigen.

2. $\operatorname{rang}Q=\dim\mathrm{Eig}(M,1)=s$ ist klar. $MQ=Q$ besagt, dass die Spalten Eigenvektoren von M zum Eigenwert $\lambda=1$ sind. Insbesondere ist also $\mathrm{Bild}\,Q\subset\mathrm{Eig}(M,1)$ und es gilt Gleichheit, da wir ja wissen, dass $\dim\mathrm{Bild}\,Q=\operatorname{rang}Q=\dim\mathrm{Eig}(M,1)$.
3. Ist $\lambda=1$ der einzige Eigenwert von M mit $|\lambda|=1$, so haben wir $B^n\to 0$ für alle Jordanblöcke zu Eigenwerten λ, $\lambda\neq 1$, wie wir bereits im 1. Teil dieses Beweises gesehen haben. Daher gilt: $M^n\to Q$.
4. Da M eine stochastische Matrix ist, ist auch M^l für jedes l eine stochastische Matrix, also auch der Mittelwert $\frac{1}{n}\sum_{l=0}^{n-1}M^l$ und der Grenzwert Q. Ist nun $\lambda=1$ der einzige Eigenwert mit $|\lambda|=1$ und $\dim\mathrm{Eig}(M,1)=1$, dann sind wegen $QM=Q\iff M^{\mathrm{t}}Q^{\mathrm{t}}=Q^{\mathrm{t}}$ die Zeilen von Q Eigenvektoren von M^{t} zum Eigenwert 1. Dies sind aber Vielfache von $(1,\dots,1)^{\mathrm{t}}$, denn

$$(m_{1j},\dots,m_{kj})\cdot\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}=\sum_{i=1}^{k}m_{ij}=s\cdot 1,$$

weil die Spaltensummen von M bekanntlich 1 ergeben, so dass insgesamt

$$M^t\cdot\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}=\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}.$$

Damit folgt:

$$Q = \begin{pmatrix} z_1 & \cdots & z_1 \\ \vdots & & \vdots \\ z_k & \cdots & z_k \end{pmatrix}.$$

Natürlich gilt außerdem $\sum_{i=1}^{k} z_i = 1$. Also ist $z = (z_1, \ldots, z_k)^t$ ein Eigenvektor von M zum Eigenwert $\lambda = 1$ von M und somit auch die stationäre Limesverteilung auf dem Zustandsraum.

□

Der Ergodensatz liefert gemeinsam mit dem vorigen Satz ein recht gutes Verständnis des Verhaltens von Markovketten. Als Illustration betrachten wir zum Abschluss der Behandlung dieses Themas noch ein Beispiel:

Beispiel 41.26. Wir betrachten das graphische Modell in Abb. 41.6.

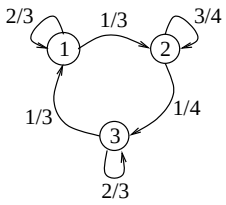


Abbildung 41.6. Eine Markovkette, die durch ein graphisches Modell gegeben ist.

Die Matrix der Übergangswahrscheinlichkeiten ist:

$$M = \begin{pmatrix} \frac{2}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{3}{4} & 0 \\ 0 & \frac{1}{4} & \frac{2}{3} \end{pmatrix}.$$

Es ist leicht nachzurechnen, dass M nur einen Eigenwert λ mit $|\lambda| = 1$ hat, auch wenn Multiplizitäten gezählt werden, und zwar den Eigenwert 1.

Nach dem Ergodensatz hat demnach $\lim_{n\to\infty} M^n$ die Gestalt (z, z, z), wobei $z = (z_1, z_2, z_3)^t$ der eindeutige Eigenvektor zum Eigenwert 1 ist, für den $z_1 + z_2 + z_3 = 1$ gilt. Eine kurze Rechnung liefert für den Eigenraum:

$$\mathrm{Eig}(M, 1) = \left\{ (v_1, v_2, v_3) \;\middle|\; v_1 = v_3,\ v_2 = \frac{4}{3} v_1 \right\}.$$

Die Bedingung $z_1 + z_2 + z_3 = 1$ ergibt damit: $\frac{10}{3} z_1 = 1$, für die Grenzverteilung erhalten wir also schließlich:

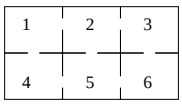
$$(z_1, z_2, z_3)^t = \left(\frac{3}{10}, \frac{4}{10}, \frac{3}{10} \right)^t.$$

Aufgaben

Aufgabe 41.1 (Matrixnormen). Zeigen Sie, dass die zur Supremumsnorm gehörige Matrixnorm die Zeilensummennorm ist, also dass für alle $A \in \mathbb{R}^{n\times n}$ gilt:

$$\max_{x\in\mathbb{R}^n,\|x\|_\infty=1} \|Ax\|_\infty = \max_{i=1,2,\dots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Aufgabe 41.2 (Labyrinth). Betrachten Sie das folgende Labyrinth, in dem sich eine Maus bewegt:



Befindet sich die Maus in Kammer j, so bleibt sie mit Wahrscheinlichkeit $\frac{1}{2}$ dort und wechselt mit Wahrscheinlichkeit $\frac{1}{2\omega_j}$ in die Kammer i, falls von Kammer j genau ω_j Türen abgehen und eine davon in Kammer i führt. Stellen Sie die Übergangsmatrix $A = (a_{ij})_{i,j=1,2,\dots,6}$ auf, zeigen Sie, dass der Grenzwert $\lim_{k\to\infty} A^k$ existiert und bestimmen Sie diesen.

Aufgabe 41.3 (Markovketten). In einer Fabrik arbeiten 5 Maschinen des gleichen Typs. Intakte Maschinen fallen pro Tag mit Wahrscheinlichkeit p aus. Maschinen, die am Anfang eines Tages defekt waren, sind bis zum nächsten Tag wieder repariert. Wir beschreiben das System durch die Anzahl x der zu Beginn eines Tages intakten Maschinen. Stellen Sie die Übergangsmatrix A zwischen den möglichen Zuständen des Systems auf, zeigen Sie die Existenz des Grenzwertes $\lim_{k\to\infty} A^k$ und berechnen Sie diesen.

Aufgabe 41.4 (Grenzverteilung von Markovketten). Wir betrachten eine Markovkette mit der folgenden Matrix der Übergangswahrscheinlichkeiten:

$$M = (p_{ij}) = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ 0 & \frac{1}{3} & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} \end{pmatrix}.$$

Zeigen Sie, dass eine eindeutige Grenzverteilung existiert und bestimmen Sie diese.

Aufgabe 41.5 (Zum Satz von Gerschgorin).

1. Zeigen Sie die folgende Version des Satzes von Gerschgorin: Sei $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ und sei b ein Eigenwert von A mit Eigenvektor ${}^t(v_1, \ldots, v_n)$. Sei i_0 der Index, für den $|v_j|$ maximal wird. Dann gilt: $|b - a_{i_0 i_0}| \leq \sum_{j=1, j\neq i_0}^{n} |a_{i_0 j}|$, d.h. der Eigenwert b liegt in einem sog. Gerschgorin-Kreis um $a_{i_0 i_0}$ mit Radius $\sum_{j=1, j\neq i_0}^{n} |a_{i_0 j}|$.

   Hinweis: Betrachten Sie $|v_{i_0}| \cdot |b - a_{i_0 i_0}|$.

2. Benutzen Sie ein Computer Algebra System (z.B. Maple), um die Eigenwerte und Eigenvektoren der Matrix

$$\begin{pmatrix} -5 & -0.1 & -4 \\ 0 & 2 & 0.1 \\ 0.1 & -0.85 & 3 \end{pmatrix}$$

   zu berechnen. Zeichnen Sie die Gerschgorin-Kreise und die Eigenwerte in ein gemeinsames Koordinatensystem.

42

Hidden Markov Models

Hidden Markov Models (versteckte Markov Modelle) sind beispielsweise wichtig in der Bioinformatik, der Sprach- und Mustererkennung, maschinellem Lernen, Spamfilter, Gestenerkennung sowie Schrifterkennung.

Mit Hilfe der Theorie aus dem Abschnitt über Markovketten werden wir u.a. nach der wahrscheinlichsten Zustandsfolge suchen, die eine gegebene Beobachtung hervorgerufen haben könnte.

42.1 Grundlegende Fragen

Definition 42.1. *Ein **Hidden Markov Model** (kurz **HMM**) ist ein Tupel*

$$\Lambda = (M, B, \pi, V)$$

*aus einer stochastischen $(k \times k)$–Matrix $M = (p_{ij})$ der **Übergangswahrscheinlichkeiten** zwischen den Zuständen $z \in \{1, 2, \ldots, k\}$, einem sogenannten Alphabet V mit v Zeichen, einer $(k \times v)$–Matrix $B = (b_{ix})$ mit $b_{ix} \geq 0$ und $\sum_{x=1}^{v} b_{ix} = 1$ (den sogenannten **Emissionswahrscheinlichkeiten**) und einer **Anfangsverteilung** π, einem $k \times 1$–Vektor.*

Der stochastische Prozess startet, indem wir gemäß π einen Anfangszustand zufällig wählen und weiter zufällig gemäß M eine Folge von Zuständen generieren. Für insgesamt T Schritte ergeben sich somit Zustände $z_1, \ldots, z_T$. In jedem Zustand z_t schreibt das Markovmodell einen Buchstaben $S_t \in V$ und zwar den x–ten Buchstaben mit der Wahrscheinlichkeit b_{ix}, falls $z_t = i$. Für den Beobachter ist nur die Zeichenfolge $S_1, \ldots, S_T$ sichtbar (darauf bezieht sich das Attribut versteckt *im Namen).*

Bzgl. Hidden Markov Models gibt es einige besonders interessante Fragen:

Frage 42.2 (Die grundlegenden Fragen bei Hidden Markov Models).

1. *Das Modell $\Lambda = (M, B, \pi, V)$ sei bekannt. Wie können wir zu einem gegebenen $S = S_1, \ldots, S_T$ die Wahrscheinlichkeit $P(S \mid \Lambda)$ berechnen?*
2. *Gegeben S und Λ. Welches ist die Folge von Zuständen $z_1, \ldots, z_T$, die am wahrscheinlichsten diese Zeichenkette generiert hat?*
3. *Gegeben seien nur S und lediglich einige grundlegende Annahmen über das Modell, etwa die Anzahl der Zustände oder der Graph. Welches ist das Modell (M, B, π, V), das S am wahrscheinlichsten generiert hat?*

In den folgenden Abschnitten geben wir algorithmische Lösungen für all diese Probleme.

Beispiel 42.3. Ein Spieler besitzt eine faire und eine gezinkte Münze, bei der *Zahl* wahrscheinlicher ist. Er setzt sie allerdings nur manchmal ein, damit es nicht zu sehr auffällt.

- Das Alphabeth ist $V = \{0, 1\}$ (0: Kopf, 1: Zahl).
- Es gibt $k = 2$ Zustände (1: faire Münze, 2: gezinkte Münze).
- $M = (p_{ij}) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$, die Übergangsmatrix zwischen den Zuständen.
- $B = (b_{ix}) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$, Matrix der Emissionswahrscheinlichkeiten (erste Zeile für die faire Münze, zweite für die gezinkte).
- $\pi = (\frac{1}{2}, \frac{1}{2})$ (zu Anfang wählt der Spieler die Münzen mit gleicher Wahrscheinlichkeit).

Die vom Spieler tatsächlich verwendete Münzfolge ist $z = (z_1, \ldots, z_T)$; sie ist dem Beobachter nicht bekannt. Er sieht nur die tatsächlichen Ergebnisse S_t der Würfe.

42.2 Die Vorwärtsmethode

Für einen Zeitpunkt $t \in \{1, 2, \ldots, T\}$ und einen Zustand $i \in \{1, 2, \ldots, k\}$ setzen wir:

$$\alpha_t(i) := P(S_1, \ldots, S_t, z_t = i \mid \Lambda).$$

Dann gilt nach Definition zunächst $\alpha_1(i) = P(S_1, z_1 = i \mid \Lambda) = \pi_i \cdot B(i, S_1)$ und weiterhin

$$\alpha_{t+1}(i) = \sum_{j=1}^{k} B(i, S_{t+1}) \cdot p_{ij} \cdot \alpha_t(j)$$

für $t \in \{1, 2, \ldots, T-1\}$, da die Markovschritte und die Zeichenwahl unabhängig voneinander sind. Schließlich ergibt sich:

$$P(S \mid \Lambda) = \sum_{i=1}^{k} \alpha_T(i).$$

Wir verdeutlichen dieses Verfahren an einem Beispiel, ähnlich dem einfach zu durchschauenden Münzwurf–Problem von oben:

Beispiel 42.4. Wir betrachten ein Würfelspiel im Casino mit einem gelegentlich verwendeten unfairen Würfel (Abb. 42.1). Es gibt also genau zwei Zustände: entweder verwenden wir den fairen ($z = 1$) oder den unfairen ($z = 2$) Würfel.

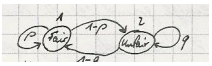


Abbildung 42.1. Würfelspiel mit einem gelegentlich verwendeten unfairen Würfel.

Die Anfangsverteilung sei: $\pi = (1, 0)^t$. Ferner nehmen wir an, dass die Übergangswahrscheinlichkeiten zwischen den Zuständen der folgenden Matrix genügen:

$$M = \begin{pmatrix} p & 1-q \\ 1-p & q \end{pmatrix} = \begin{pmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{pmatrix}.$$

Eine Möglichkeit, einen unfairen Würfel zu basteln, ist es, einfach statt der 6 eine 1 aufzudrucken. Dies führt zu den Wahrscheinlichkeiten:

	1	2	3	4	5	6
fair (1)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
unfair (2)	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0

$$\Rightarrow B = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{2}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \end{pmatrix}.$$

Dies ist allerdings etwas offensichtlich. Ein zwar schwieriger herzustellender, aber nicht so leicht zu enttarnender Würfel ist der folgende:

$$B = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{10} & \frac{1}{10} \end{pmatrix}.$$

Für die nachfolgende Beispielrechnung nehmen wir an, dass der Croupier entweder den fairen Würfel oder aber den zweiten (schwierig zu bauenden) unfairen Würfel verwendet. Er würfelt drei Mal und erhält die Zahlen 1, 2, 6. Die Wahrscheinlichkeit für diesen Ausgang ($S = 126$) ist: $P(126 \mid \Lambda)$. Nach obigem Verfahren müssen wir, um diese zu ermitteln, die $\alpha_t(i)$ berechnen:

$$\alpha_1(1) = 1 \cdot \tfrac{1}{6} = \tfrac{1}{6}, \qquad \alpha_1(2) = 0 \cdot \tfrac{1}{5} = 0,$$
$$\alpha_2(1) = \tfrac{1}{6} \cdot 0.9 \cdot \tfrac{1}{6} + 0 = \tfrac{0.9}{6^2}, \qquad \alpha_2(2) = \tfrac{1}{5} \cdot 0.1 \cdot \tfrac{1}{6} = \tfrac{0.1}{5 \cdot 6}.$$

Daraus erhalten wir

$$\alpha_3(1) = \frac{1}{6} \cdot 0.9 \cdot \frac{0.9}{6^2} + \frac{1}{6} \cdot 0.5 \cdot \frac{0.1}{5 \cdot 6} = \frac{4.35}{5 \cdot 6^3} \approx 0.00403,$$
$$\alpha_3(2) = \frac{1}{10} \cdot 0.1 \cdot \frac{0.9}{6^2} + \frac{1}{10} \cdot 0.5 \cdot \frac{0.1}{5 \cdot 6} = \frac{0.75}{10 \cdot 5 \cdot 6^2} \approx 0.000417$$

und damit letztendlich:

$$P(126 \mid \Lambda) = \alpha_3(1) + \alpha_3(2) = \frac{48}{10 \cdot 5 \cdot 6^3} = \frac{1}{5^2 \cdot 3^2} = \frac{1}{225} \approx 0.0044.$$

Diese sogenannte **Vorwärtsmethode** beantwortet die erste Frage aus 42.2. Wir werden sehen, dass die $\alpha_t(i)$ aber außerdem für weitere Rechnungen nützlich sind.

42.3 Rückwärtsmethode

Eine andere Möglichkeit zur Berechnung der gleichen Wahrscheinlichkeit $P(S \mid \Lambda)$ ist die **Rückwärtsmethode**.

Wir setzen dazu für $i \in \{1, 2, \ldots, k\}$ und $t \in \{1, 2, \ldots, T-1\}$:

$$\beta_t(i) := P(S_{t+1}, \ldots, S_T, z_t = i \mid \Lambda) \text{ und } \beta_T(i) := 1.$$

Diese können wir rekursiv für $t \in \{T-1, \ldots, 2, 1\}$ durch

$$\beta_t(i) = \sum_{j=1}^{k} B(j, S_{t+1}) \cdot p_{ji} \cdot \beta_{t+1}(j)$$

berechnen und erhalten schließlich $P(S \mid \Lambda) = \sum_{j=1}^{k} B(j, S_1) \cdot \pi_j \cdot \beta_1(j)$.

Für das Beispiel von oben ergibt sich natürlich auch mit dieser Methode das gleiche Ergebnis: $P(126 \mid \Lambda) = \frac{1}{225}$.

Die Rechnung verläuft wie folgt:

$$\beta_3(1) = 1, \qquad \beta_3(2) = 1,$$
$$\beta_2(1) = \tfrac{1}{6} \cdot 0.9 \cdot 1 + \tfrac{1}{10} \cdot 0.1 \cdot 1 = \tfrac{9.6}{6 \cdot 10}, \qquad \beta_2(2) = \tfrac{1}{6} \cdot 0.5 \cdot 1 + \tfrac{1}{10} \cdot 0.5 \cdot 1 = \tfrac{8}{6 \cdot 10},$$

Daraus ergibt sich:

$$\beta_1(1) = \frac{1}{6} \cdot 0.9 \cdot \frac{9.6}{60} + \frac{1}{5} \cdot 0.1 \cdot \frac{8}{60} = \frac{43.2 + 4.8}{5 \cdot 6^2 \cdot 10} = \frac{2}{5^2 \cdot 3}$$
$$\beta_1(2) = \frac{1}{6} \cdot 0.5 \cdot \frac{9.6}{60} + \frac{1}{5} \cdot 0.5 \cdot \frac{8}{60} = \frac{8}{5 \cdot 6 \cdot 10}$$

Schließlich erhalten wir:

$$P(126 \mid \Lambda) = \frac{1}{6} \cdot 1 \cdot \beta_1(1) + 0 \cdot \beta_1(2) = \frac{1}{5^2 \cdot 3^2} = \frac{1}{225} \approx 0.0044,$$

was wir mit der Vorwärtsmethode ja bereits berechnet hatten.

42.4 Raten der Zustandsfolge

Gegeben seien $S_1, \ldots, S_T$ und Λ. Wir wollen die Zustandsfolge $z_1, \ldots, z_T$ raten, die mit größter Wahrscheinlichkeit zu $S_1, \ldots, S_T$ geführt hat. In den obigen Beispielen entspricht das der Suche nach den Zeitpunkten, zu denen der Spieler den falschen Würfel bzw. die falsche Münze eingesetzt hat.

Dazu setzen wir

$$\begin{aligned}\gamma_t(i) &:= P(z_t = i \mid S, \Lambda) \\ &= P(z_t = i \mid S_1, \ldots, S_t, \Lambda) \cdot P(z_t = i \mid S_{t+1}, \ldots, S_T, \Lambda).\end{aligned}$$

Dies ist die Wahrscheinlichkeit, dass die Zustandsfolge $z_1, \ldots, z_T$ im Zeitpunkt t im Zustand i war, unter der Annahme des Modells Λ und der Beobachtung S. Da $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, lässt sich dieser Ausdruck wie folgt schreiben:

$$\gamma_t(i) = \frac{P(z_t = i, S_1, \ldots, S_t \mid \Lambda)}{P(S_1, \ldots, S_t \mid \Lambda)} \cdot \frac{P(z_t = i, S_{t+1}, \ldots, S_T \mid \Lambda)}{P(S_{t+1}, \ldots, S_T \mid \Lambda)} = \frac{\alpha_t(i) \cdot \beta_t(i)}{P(S \mid \Lambda)}.$$

Mit Hilfe der Werte $\alpha_t(i)$, $\beta_t(i)$ aus der Vorwärts- und der Rückwärtsmethode können wir also die $\gamma_t(i)$ berechnen. Ein Ansatz, um daraus die wahrscheinlichste Zustandsfolge $z_1, \ldots, z_T$ zu erhalten, ist nun folgender: Wir wählen z_t so, dass $\gamma_t(z_t) = \max_j(\gamma_t(j))$.

Beispiel 42.5. ...

Die beschriebene Methode zum Raten der Zustandsfolge ist nur eine „lokale Optimierung". Sie kann sogar Zustandsfolgen $z_1, \ldots, z_T$ auswählen, die unmöglich sind, d.h. für die $p_{z_{t+1}, z_t} = 0$ für ein t gilt. Es ist also noch keine perfekte Antwort auf die zweite Frage aus 42.2.

Eine Alternative ist der **Viterbi–Algorithmus**, der im Allgemeinen algorithmisch weniger aufwendig ist und der in der Praxis sehr häufig eingesetzt wird. Beispielsweise kann man mit seiner Hilfe nämlich den Optimalempfänger für verzerrte und gestörte Kanäle berechnen.

Der Viterbi–Algorithmus wird daher heutzutage in Handys und Wireless LANs zur **Entzerrung** oder **Fehlerkorrektur** der Funkübertragung verwendet. Siehe auch [PS05, S. 57] für eine originelle Sichtweise auf den Algorithmus. Im Internet gibt es ebenfalls viele Informationen darüber. Auch in der Geometrie kann man den Viterbi–Algorithmus einsetzen, beispielsweise bei der Entstörung von Punkten, die beim Einscannen mit einem $3d$–Scanner anfallen.

42.5 Baum–Welch: Verbessern des Modells

Wir möchten jetzt die dritte Frage aus 42.2 angehen. Gegeben sei dazu ein Modell $\Lambda = (M, B, \pi, V)$ und eine Zeichenkette S. Wir wollen die Parameter des Modells verbessern, so dass S mit größerer Wahrscheinlichkeit ausgegeben wird („aus Beobachtungen lernen"). Der dafür verwendete **Baum–Welch–Algorithmus** ist ein Spezialfall des **EM–Algorithmus** (Expectation–Maximation), siehe [PS05, Theorem 1.15, S. 19].

Expectation Step:

Zunächst bestimmen wir mit der Vorwärts–Rückwärts–Methode die Wahrscheinlichkeiten $\alpha_t(i)$ und $\beta_t(i)$ dafür, dass die verborgene Zustandsfolge $z_1, \ldots, z_T$ im t–ten Schritt im Zustand i war, unter der Annahme, dass das Modell Λ ist. Dann ist die Wahrscheinlichkeit, im t–ten Schritt vom Zustand j in den Zustand i zu wechseln durch

$$\begin{aligned}\xi_t(i,j) &:= P(z_t = j, z_{t+1} = i \mid S, \Lambda) \\ &= \frac{\beta_{t+1}(i) \cdot B(i, S_{t+1}) \cdot p_{ij} \cdot \alpha_t(j)}{P(S \mid \Lambda)}\end{aligned}$$

gegeben. Nach Definition gilt: $\gamma_t(j) = P(z_t = i \mid S, \Lambda) = \sum_{i=1}^{k} \xi_t(i,j)$.

Maximation Step:

Wir verwenden die Wahrscheinlichkeiten aus dem Expectation Step, um aus der Beobachtung S die Parameter des zugrunde liegenden Modells Λ zu schätzen. Das wahrscheinlichste Modell $\overline{\Lambda} = (\overline{M}, \overline{B}, \overline{\pi}, V)$ für die Beobachtung S hat, unter der Annahme, dass die Wahrscheinlichkeiten $\xi_t(i,j)$ und $\gamma_t(j)$ zutreffen, die Parameter:

$$\overline{p}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(j)}, \quad \overline{B}(j,x) = \frac{\sum_{t=1,S_t=x}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)}, \quad \overline{\pi_j} = \gamma_1(j).$$

Für das intuitive Verständnis dieser Formeln mag es helfen, Zähler und Nenner jeweils als Mittelwerte zu sehen, z.B.:

$$\overline{p}_{ij} = \frac{1}{T-1} \cdot \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\frac{1}{T-1} \cdot \sum_{t=1}^{T-1} \gamma_t(j)}.$$

Nach Theorem 1.15 aus [PS05] gilt tatsächlich, dass sich das Modell hierbei höchstens verbessert hat:

$$P(S \mid \overline{\Lambda}) \geq P(S \mid \Lambda).$$

Mit dem verbesserten Modell $\overline{\Lambda}$ können wir zurück in den Expectation Step gehen und das Verfahren so iterieren.

Aufgaben

Aufgabe 42.1 (Wettervorhersage). Wir betrachten eine etwas vereinfachte Klassifikation des Wetters in die drei Zustände S_1 = regnerisch, S_2 = bewölkt, S_3 = sonnig. Die Übergangswahrscheinlichkeiten zwischen den Zuständen seien bekannt und durch die folgende Matrix gegeben:

$$\begin{pmatrix} 0.4 & 0.2 & 0.1 \\ 0.3 & 0.6 & 0.1 \\ 0.3 & 0.2 & 0.8 \end{pmatrix}.$$

Diese Matrix ist so zu verstehen, dass die Wahrscheinlichkeit, dass auf einen regnerischen Tag ein sonniger folgt, 0.3 ist. Tag 0 sei sonnig. Wie groß ist die Wahrscheinlichkeit, dass an den Tagen $1, 2, \ldots, 7$ das Wetter sonnig, sonnig, regnerisch, regnerisch, sonnig, bewölkt, sonnig auftritt?

Aufgabe 42.2 (Wettervorhersage). Wir betrachten das Wetter-Modell aus der vorigen Aufgabe. Was ist, falls Tag 0 bewölkt ist, das wahrscheinlichste Wetter für die Folge der Tage $1, 2, \ldots, 4$ bzw. für die Folge der Tage $1, 2, \ldots, 5$?

Aufgabe 42.3 (Baum–Welch–Algorithmus). Ein Hidden-Markov-Modell mit 3 Zuständen und 3 Buchstaben hat die folgenden Sequenzen der Länge 20 erzeugt:

$$(a,a,a,b,b,b,b,b,b,b,b,c,c,a,b,b,b,b,b,b)$$
$$(b,b,c,c,c,c,a,a,a,b,b,c,c,a,a,a,b,c,c,a)$$
$$(a,b,c,c,a,b,b,b,b,b,b,c,c,c,a,a,b,c,c,c)$$
$$(b,b,b,c,c,a,a,a,a,b,b,b,b,c,c,c,a,a,a,b)$$

Finden Sie mit Hilfe des Baum-Welch-Algorithmus ein Hidden-Markov-Modell, das diese Sequenzen mit möglichst großer Wahrscheinlichkeit reproduziert.

43

Pseudozufallszahlen und Monte–Carlo–Simulation

Es gibt viele Situationen, in denen probabilistische (das heißt von Zufällen abhängige) Algorithmen wesentlich besser geeignet sind, ein Problem zu behandeln, als deterministische.

Hierfür benötigt man allerdings Zufallszahlen; da Computer heutzutage üblicherweise aber deterministisch arbeiten, muss man auf sogenannte Pseudozufallszahlen ausweichen. Tatsächlich gibt es deterministische Algorithmen, die Zahlenfolgen liefern, die recht zufällig aussehen und in vielen Anwendungsbereichen auch gut an Stelle von Zufallszahlen verwendet werden können.

In sicherheitsrelevanten Situationen muss man allerdings oft auf tatsächlich zufällige Zahlen zurückgreifen, wie sie beispielsweise bei Signalrauschen auftreten: z.B. thermisches Rauschen von Widerständen (in Serie auf Intels i810 Chipsatz). Solche Hardware-Lösungen besprechen wir hier allerdings nicht.

43.1 Lineare Kongruenzgeneratoren

Definition 43.1. *Eine Folge von **Pseudozufallszahlen** ist eine Folge von Zahlen, die zwar mit einem deterministischen Algorithmus generiert werden, die aber **zufällig aussehen**. Hierbei ist* zufällig aussehen *nicht exakt definiert, sondern meint nur, dass es möglichst wenig Tests geben soll, die die Determiniertheit der Folge erkennen.*

Ein erstes Kriterium für einen guten Pseudozufallszahlengenerator ist sicherlich, dass die auftretenden Zahlen gleichverteilt sind. Dieses erfüllt schon der einfache **lineare Kongruenzgenerator**: Wir wählen Zahlen $a, c, M \in \mathbb{N}$, sowie eine Anfangszahl (auch **Saat**, engl. **seed**, genannt) $x_0 \in \{0, 1, \ldots, M-1\}$. Die

weiteren Pseudozufallszahlen $x_i \in \{0, 1, \ldots, M\}$ werden nun folgendermaßen berechnet:

$$x_{i+1} = a \cdot x_i + c \mod M,$$

d.h. Rest der Division der Zahl $a \cdot x_i + c$ durch M (siehe dazu Abschnitt 3.2).

Es ist klar, dass sich die Folge spätestens nach M Schritten wiederholt, da es ja höchstens M verschiedene x_i gibt. Die Anzahl der Iterationen, nach denen sich eine Wiederholung einstellt, heißt **Periode des Pseudozufallszahlengenerators** (manchmal auch **Periodenlänge** genannt). Offenbar möchte man diese möglichst groß machen, wenn man gute Pseudozufallszahlen haben möchte. Auf einem 32–Bit–System wird man M daher möglichst nahe an 2^{32} wählen, allerdings unter Beachtung der Tatsache, dass die erzeugte Folge möglichst zufällig erscheint.

Beispiel 43.2. 1969 wurde für das IBM System/360 ein Zufallsgenerator entwickelt (der sogenannte **minimal standard**), für den $M = 2^{31} - 1$, $a = 7^5 = 16807$, $c = 0$ gilt; dies ist eine recht brauchbare Wahl. Im Taschenrechner TI–59 wird $a = 24298$, $c = 99991$, $M = 199017$ verwendet.

Bei der Wahl der Parameter ist Vorsicht geboten; so sollten beispielsweise a und c im Verhältnis zu M nicht zu klein gewählt werden, weil dann immer lange aufsteigende Sequenzen von Zahlen erzeugt werden. Doch es gibt auch weniger offensichtliche Fehlerquellen, die ein unbedachter Einsatz der Algorithmen durch den Menschen hervorrufen kann:

Beispiel 43.3. Zwecks einer schnellen Laufzeit wurde in den 70er Jahren für einen IBM–Rechner der sogenannte **RANDU** Algorithmus verwendet, für den $M = 2^{31}$ und $a = 2^{16} + 3$ gilt. Teilt man die davon erzeugten Zahlen in 3–er–Blöcke ein und fasst die Zahlen als Punkte im Dreidimensionalen auf, so liegen alle auf nur 15 Ebenen, sind also gar nicht zufällig verteilt.

43.2 Der Mersenne–Twister

Der **Mersenne–Twister** ist ein Pseudozufallszahlengenerator, der 1997 von Matsumoto und Nishimura entwickelt wurde. Die verbreitetste Variante ist der MT 19937, der eine Periodenlänge von $2^{19937} - 1$ aufweist (das ist eine sogenannte **Mersenne–Primzahl**, d.h. eine Primzahl der Form $M^p - 1$ mit $M \in \mathbb{N}$ und p prim, daher der Name des Generators). Er ist sehr schnell und liefert sehr gleichverteilte Zahlenfolgen. Außerdem sind alle Bits für sich genommen gleichverteilt. Dieser Pseudozufallszahlengenerator ist also in nahezu jeder Hinsicht besser als die Kongruenzgeneratoren. Recht detaillierte Information dazu finden sich beispielsweise auf der Webseite http://de.wikipedia.org/wiki/Mersenne-Twister.

43.3 Testen von Zufallsfolgen

Im Laufe der Jahre wurden verschiedene Tests entwickelt, um für eine gegebene Folge von Zahlen zu überprüfen, ob sie zufällig aussieht oder nicht.

43.3.1 χ^2–Test

Mit dem χ^2–Test können wir bekanntlich die Gleichverteilung einer Menge von Zahlen x_i überprüfen (siehe Abschnitt 39.6). Wir teilen die Zahlen dazu wieder in k Kategorien ein, und zwar indem wir sagen, dass x_i zur Kategorie s gehört, wenn

$$\frac{s-1}{k} < x_i \leq \frac{s}{k}.$$

Die Zahlen Z_j, $j = 1, 2, \ldots, k$ geben dann an, wieviele Zahlen zu welcher Kategorie gehören. Die Zufallsvariable

$$Y = \sum_{i=1}^{k} \frac{(Z_i - np_i)^2}{np_i}$$

ist dann wieder annähernd χ^2_{k-1}–verteilt, wobei n die Anzahl der Zufallszahlen und $p_j = \frac{1}{k}$ für jedes j ist.

Leider ist der χ^2–Test auf Gleichverteilung kein Test, aus dem man unbedingt auf die gute Qualität des Pseudozufallszahlengenerators schließen könnte, wie das folgende Beispiel zeigt:

Beispiel 43.4. Für $M = 5$, $a = 1$, $c = 1$, also $x_{i+1} = x_i + 1 \mod 5$, sind die erzeugten Zahlen perfekt gleichverteilt, doch sicherlich keine guten Pseudozufallszahlen.

43.3.2 Run–Test

Mit einem Run–Test kann man die erzeugten Zahlen x_i auf Unabhängigkeit prüfen. Ein **Run** ist eine Teilsequenz aufeinanderfolgender Zahlen, die gewisse Eigenschaften erfüllen. Oft verwendet man hierfür den Run–Up–Test und den Run–Down–Test; für den Run–Up–Test verwendet man die Bedingung

$$x_{i+1} \geq x_i$$

und für den Run–Down–Test

$$x_{i+1} < x_i.$$

Beispielsweise beginnt die Folge

$$(3, 3, 6, 7, 8, 3, 2, 1)$$

mit einem Run–Up der Länge 5 und endet mit einem Run–Down der Länge 4.

Für die Auftrittswahrscheinlichkeit von Runs der Länge $l \geq 2$ innerhalb einer Folge von gleichverteilten Zahlen in einem festen Bereich gilt nun

$$P(\text{ Run der Länge } l\,) \approx \frac{l}{(l+1)!},$$

weil, falls bereits $l-1$ Zahlen aus dem Bereich gewählt wurden, nur noch etwa $\frac{1}{l}$ der Zahlen existieren, die größer als die größte bzw. kleiner als die kleinste Zahl sind und schließlich $\frac{l}{l+1}$ der Zahlen den Run beenden können. Beispielsweise beenden bei bereits zwei gewählten Zahlen etwa $\frac{2}{3}$ der Zahlen den Run, so dass $P(2-\text{ Run }) \approx \frac{2}{3}$ gilt.

Wendet man diesen Test auf die triviale Folge aus Beispiel 43.4 ein, so zeigt sich die mangelnde Qualität sofort, da es nur aufsteigende Runs der Länge 5 gibt.

Eine Variante dieses Tests ist der in Beispiel 39.3 verwendete, wo wir in einer 0–1–Folge Runs betrachtet haben, die nur aus gleichen Ziffern bestanden.

43.3.3 Spektraltest

Der **Spektraltest** liefert als Ergebnis, wie viele aufeinanderfolgende Zahlen noch als unabhängig gelten können und wie gut sie unabhängig sind. Eine möglichst große Zahl ist hier also günstig. Das geschieht durch Zahlen

$$\nu_2, \nu_3, \ldots,$$

die für jeweils $2, 3, \ldots$ aufeinander folgende Pseudozufallszahlen deren Qualität angeben.

Beispiel 43.5. Der Spektraltest liefert für den schlechten Generator 43.3 auch schlechte Werte:

$$\nu_2 = 23171, \nu_3 \approx 10, \nu_4 = \cdots = \nu_9 \approx 10.$$

Leider können wir diesen Test, der derzeit wohl einer der besten Tests ist, aus Zeitgründen nicht im Detail beschreiben und verweisen daher auf Literatur, wie [Knu99].

43.4 Fehlerquelle Mensch

Wie so oft in der Informatik ist auch bei der Verwendung von Pseudozufallszahlengeneratoren durch den Menschen Vorsicht geboten.

Beispiel 43.6 (Online-Casino geknackt). Beispielsweise wurde bei einem Online–Casino die Anzahl der Millisekunden seit Mitternacht zum Zeitpunkt des Einloggens des Spielers als Startwert für einen Zufallszahlengenerator verwendet, dessen Algorithmus veröffentlicht wurde. Eine Gruppe von Mathematikern hat es mit dieser Information geschafft, die Kartenfolge bei einem Kartenspiel vorherzusagen, nachdem sie nur die ersten wenigen Karten gesehen hatten, weil sie sich einfach möglichst Punkt Mitternacht einloggten und daher nur wenige Pseudozufallsfolgen möglich waren.

Ein anderer Fall ist ebenfalls sehr amüsant:

Beispiel 43.7 (Systemuhr blieb stehen). Bei einem Glücksspiel, dessen Glückszahlen ebenfalls von einem Pseudozufallszahlengenerator erzeugt wurden, und dessen Startwert die aktuelle Systemzeit des Computers war, blieb eben diese Uhr unbemerkt stehen. Dementsprechend war die Zufallsfolge am nächsten Tag wieder exakt die gleiche wie am Vortag. Dieses (ohne den Systemfehler) sehr unwahrscheinliche Ereignis führte dazu, dass einige Mathematiker errieten, dass obiges Uhr–Problem vorlag, und die daher genau die gleiche Zufallsfolge am folgenden Tag wieder tippten. Tatsächlich hatten die Glücksspiel–Betreiber das Problem nicht realisiert und daher die Systemzeit des Computers nicht verändert, so dass auch an diesem Tag die gleiche Zahlenfolge erschien (jetzt schon bei drei aufeinander folgenden Ziehungen!). Erst dann wurde das Phänomen geklärt.

43.5 Anwendungen

Im Gegensatz zu einem deterministischen Algorithmus verwendet ein **probabilistischer** oder **randomisierter Algorithmus** Zufallszahlen (oder Pseudozufallszahlen), um den Ablauf zu steuern. Solche probabilistischen Algorithmen, die eher den Charakter einer Simulation haben und nicht zwingend zu einem korrekten Ergebnis führen sollen, werden oft auch **Monte–Carlo–Simulationen** genannt.

43.5.1 Quicksort

Damit beim Sortier–Algorithmus **Quicksort** der Worst Case möglichst nicht auftritt, ist es sinnvoll, auch hier die zu sortierende Liste von Zahlen an einer

pseudozufälligen Stelle in zwei Teillisten aufzuteilen und diese dann wiederum rekursiv mit Quicksort zu sortieren. Dies ergibt eine mittlere Laufzeit von $O(n \log n)$.

43.5.2 Buffons Nadelexperiment

Eine der geschichtlich ersten Anwendungen einer Monte–Carlo–Methode ist **Buffons Nadelexperiment** aus dem 18. Jahrhundert: Lässt man eine Nadel der Länge 1 (idealerweise mit Dicke 0 und ohne Kopf) auf ein liniertes Blatt Papier mit Linienabstand 1 fallen, so schneidet die Nadel eine Linie oder auch nicht.

Es gibt dabei zwei Variablen: den Winkel θ, in dem die Nadel fällt, und den Abstand d des Mittelpunktes der Nadel von der nächsten Linie (Abb. 43.1).

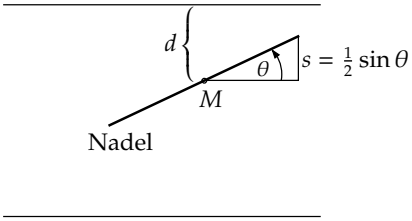


Abbildung 43.1. Bei Buffons Nadelexperiment wird eine Nadel der Länge 1 auf ein liniertes Blatt Papier mit Linienabstand 1 geworfen. Hier sind zwei der Linien gezeigt.

θ kann zwischen 0° und 180° (bzw. 0 und π im Bogenmaß) variieren und d kann nicht mehr als die Hälfte des Linienabstandes betragen. Die Nadel schneidet die Linie, falls

$$d \leq \frac{1}{2} \sin \theta.$$

Wie oft wird dies auftreten? Abbildung 43.2 zeigt den Graph von $\frac{1}{2} \sin \theta$ gemeinsam mit einem umrandenden Rechteck. Punkte auf oder unter der Kurve bedeuten, dass die Nadel die Linie trifft. Die Wahrscheinlichkeit dafür ist das Verhältnis der Fläche unter dem Graphen zu der Fläche des Rechtecks. Die Fläche unter der Kurve ist

$$\int_0^\pi \frac{1}{2} \sin \theta \, d\theta = \frac{1}{2} \cdot \Big[-\cos \theta \Big]_0^\pi = \frac{1}{2} \cdot \Big(-(-1) - (-1) \Big) = 1,$$

während die Fläche des Rechtecks $\frac{1}{2}\pi$ beträgt. Die Wahrscheinlichkeit, dass eine Nadel eine Linie schneidet, ist also

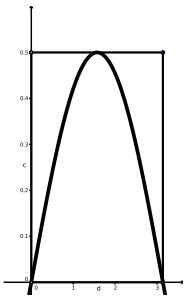


Abbildung 43.2. Zu Buffons Nadelexperiment.

$$P(\text{ Nadel schneidet eine Linie }) = \frac{1}{\frac{1}{2}\pi} = \frac{2}{\pi} \approx 0.63662.$$

Demnach ist

$$\frac{2 \cdot \text{ Anzahl der Nadelwürfe}}{\text{Anzahl der Linienschneidungen}} \approx \pi$$

nach dem Gesetz der großen Zahl. Auf einigen Webseiten kann man dieses Experiment simulieren, z.B. http://mste.illinois.edu/reese/buffon/bufjava.html. Dabei stellt man fest, dass man für eine gute Annäherung der Kreiszahl meist doch mehrere tausend Nadelwürfe benötigt, was Buffon (freilich ohne Computereinsatz) wohl nicht allzu häufig ausprobiert haben dürfte.

43.5.3 Numerische Integration

Die Kreiszahl π kann man mit einem Computer auch folgendermaßen annähern: Wir wählen pseudozufällig Punkte

$$P = (x, y) \in [-1, 1] \times [-1, 1]$$

im Quadrat mit Seitenlänge 2 und dem Ursprung als Mittelpunkt. Dann überprüfen wir jeweils, ob P im Einheitskreis enthalten ist, ob also $x^2 + y^2 \leq 1$ gilt. Da für die Wahrscheinlichkeit

$$P(\text{ Punkt im Kreis }) = \frac{\text{Kreisfläche}}{\text{Quadratfläche}} = \frac{\pi}{4}$$

gilt, ist

$$\frac{\text{Anzahl der Punkte innerhalb des Kreises}}{\text{Anzahl der gewählten Punkte insgesamt}}$$

eine Annäherung für $\frac{\pi}{4}$. Wir können mit genügend Punkten π also näherungsweise berechnen.

Dieses Beispiel zur Bestimmung von π liefert eine Methode zur näherungsweisen Berechnung eines Flächenintegrals. Man kann analog auch Integrale höherdimensionaler Funktionen berechnen. Dies kann, insbesondere in höheren Dimensionen, tatsächlich eine praktikable Methode zur näherungsweisen Berechnung des Integrals sein.

Aufgaben

Aufgabe 43.1 (Näherungsweise Integralberechnung). Benutzen Sie ein geeignetes Computeralgebra–Programm oder eine geeignete Programmiersprache, um mit Hilfe von Monte–Carlo–Simulation das Integral

$$\int_{-3}^{3} \frac{1}{2} \cdot e^{-x^2} \, dx$$

näherungsweise zu berechnen. Den dafür vom benutzten System bereitgestellten Pseudozufallszahlengenerator dürfen Sie hierbei verwenden.

Teil VI

Numerik

Einführung

Numerische Methoden werden es uns schließlich erlauben, einige der in den vorigen Kapiteln vorgestellten Methoden tatsächlich am Rechner umzusetzen — unter Berücksichtigung möglicher Rundungsfehler, die durch die Darstellung von Zahlen im Computer hervorgerufen werden. Beispielsweise sind die Methoden, die im Abschnitt zur linearen Algebra zur Diagonalisierung symmetrischer Matrizen gegeben wurden, nicht wirklich praktikabel und wir geben hier eine gute Alternative.

Wegen der mangelnden Zeit können wir freilich nicht auf alle Aspekte eingehen. Für wesentlich detaillierte Informationen sei daher auf die Literatur verwiesen, wie beispielsweise [FH07], [DS05]. Trotzdem werden wir zumindest einige wesentliche Bereiche beleuchten, insbesondere jene, die mit der Berechnung von Eigenwerten zu tun haben, was eines der grundlegenden Problem für viele Algorithmen ist.

Nicht diskutieren werden wir leider die immer zentraler werdende Tatsache, dass Parallelisierbarkeit von Algorithmen eine immer größere Rolle spielt. Prozessoren mit vier Kernen sind derzeit schon selbstverständlich und Graphikprozessoren mit sogar 256 Kernen wegen des großen Spielemarktes nahezu flächendeckend im Einsatz. Da diese allerdings nur mit einer Rechengenauigkeit von sehr wenigen Ziffern arbeiten können (z.B. 7 oder 12), ist hierfür eine genaue Analyse der auftretenden Rundungsfehler besonders wichtig.

44

Rundungsfehler und grundlegende Algorithmen

Bei einer Verwendung von Fließkommazahlen sind Rundungsfehler unvermeidlich. Wir werden an einigen Beispielen sehen, wo Rundungsfehler auftreten und wie man sie, wenn möglich, vermeiden kann. Außerdem gehen wir schon auf erste wesentliche Algorithmen ein. Für wesentlich mehr Details zum Thema Numerik als wir sie hier liefertn können, siehe [FH07], [DS05].

44.1 Der Gaußalgorithmus mit Spaltenpivotierung

Wir beginnen unsere Untersuchungen zu Rundungsfehlern mit einem Beispiel aus der linearen Algebra. Da sich sehr viele algorithmische Probleme letztendlich auf lineare Gleichungssysteme reduzieren lassen, ist dies ein typisches Problem:

Beispiel 44.1. Wir nehmen an, dass wir nur mit einer **Rechengenauigkeit** von 3 Dezimalstellen arbeiten. Gegeben sei folgendes lineares Gleichungssystem:

$$\begin{pmatrix} 1.00 \cdot 10^{-4} & 1.00 \\ 1.00 & 1.00 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1.00 \\ 2.00 \end{pmatrix}.$$

Die exakte Lösung (auf 5 Stellen genau) ist

$$x_1 = 1.0001, \qquad x_2 = 0.9999.$$

Auf 3 Stellen gerundet ergibt sich:

$$x_1 = 1.00, \qquad x_2 = 1.00.$$

Gewöhnlicher Gaußalgorithmus: Welches Ergebnis liefert der gewöhnliche Gaußalgorithmus? Wir stellen die erweiterte Matrix

$$\left(\begin{array}{cc|c} 1.00 \cdot 10^{-4} & 1.00 & 1.00 \\ 1.00 & 1.00 & 2.00 \end{array}\right)$$

auf und eliminieren die linke untere Position, indem wir das $(1.00 \cdot 10^{-4})^{-1}$–fache, also das 10^4–fache, der oberen Zeile von der unteren abziehen. Exakt würde sich hierbei

$$\left(\begin{array}{cc|c} 1.00 \cdot 10^{-4} & 1.00 & 1.00 \\ 0 & -9999 & -9998 \end{array}\right)$$

ergeben, doch, da wir mit nur 3 Stellen Genauigkeit arbeiten, erhalten wir:

$$\left(\begin{array}{cc|c} 1.00 \cdot 10^{-4} & 1.00 & 1.00 \\ 0 & -1.00 \cdot 10^4 & -1.00 \cdot 10^4 \end{array}\right).$$

Hieraus ergibt sich als Lösung $x_2 = 1.00$ und damit $x_1 \cdot 1.00 \cdot 10^{-4} + 1.00 = 1.00$, d.h. $x_1 \cdot 1.00 \cdot 10^{-4} = 0.00$, also $x_1 = 0.00$, was stark von der oben berechneten tatsächlichen Lösung abweicht.

Gaußalgorithmus mit Zeilenvertauschung: Hätten wir allerdings vorher die beiden Zeilen der erweiterten Matrix vertauscht, so hätten wir

$$\left(\begin{array}{cc|c} 1.00 & 1.00 & 2.00 \\ 1.00 \cdot 10^{-4} & 1.00 & 1.00 \end{array}\right)$$

und daraus

$$\left(\begin{array}{cc|c} 1.00 & 1.00 & 2.00 \\ 0.00 & 1.00 & 1.00 \end{array}\right)$$

erhalten, da $1.00 - 1.00 \cdot 10^{-4} = 0.9999$ auf drei Stellen wieder 1.00 ergibt genauso wie $1.00 - 2.00 \cdot 10^{-4} = 0.9998$. Damit finden wir $x_2 = 1.00$ und $x_1 = 1.00$, was die richtige Lösung ist.

Dieses Beispiel suggeriert, dass ein geschicktes Vertauschen der Zeilen numerisch wesentlich stabiliere Ergebnisse liefern kann. Dies hatten wir im Kapitel zur Linearen Algebra zwar schon kurz erwähnt; nun formalisieren wir dies aber:

Algorithmus 44.2 (Gaußalgorithmus mit Spaltenpivotierung).

Input: Ein Gleichungssystem $Ax = b$, wobei $A \in \mathbb{R}^{n \times n}$ eine quadratische Matrix und $b \in \mathbb{R}^n$ ein Vektor ist.

*Output: Ein äquivalentes Gleichungssystem $Rx = \tilde{b}$, wobei $R \in \mathbb{R}^{n \times n}$ eine **rechte obere Dreiecksmatrix** und $\tilde{b} \in \mathbb{R}^n$ ein Vektor ist.*

Wir setzen $A =: A^{(1)} := (a_{ij}^{(1)})$ und berechnen schrittweise $A^{(k+1)}$ aus $A^{(k)}$ durch Elimination nach geschickter Zeilenvertauschung:

1. *Wähle im k–ten Eliminationsschritt $A^{(k)} \to A^{(k+1)}$ ein $p \in \{k, \dots, n\}$, so dass*

$$\left|a_{pk}^{(k)}\right| \geq \left|a_{jk}^{(k)}\right| \quad \forall\, j = k, \dots, n.$$

2. *Vertausche die p–te und die k–te Zeile. Dann heißt das neue Element in der p–ten Spalte der k–ten Zeile* ***Pivotelement****.*
3. *Eliminiere die Elemente unterhalb des k–ten Eintrags in der k–ten Spalte durch Zeilenoperationen.*

Die Zeilenvertauschung im k–ten Schritt wird hierbei realisiert durch Multiplikation von links mit der Permutationsmatrix

$$P_k = \begin{pmatrix} 1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & \cdots & \cdots & 0 & 1 & \\ & & & \vdots & 1 & & & 0 & \\ & & & \vdots & & \ddots & & \vdots & \\ & & & 0 & & & 1 & \vdots & \\ & & & 1 & 0 & \cdots & \cdots & 0 & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ 0 & & & & & & & & & 1 \end{pmatrix},$$

die von der Einheitsmatrix nur dadurch abweicht, dass in den Spalten k und p die 1 *jeweils in der p–ten bzw. k–ten Zeile steht.*

Die Elimination kann realisiert werden durch die Multiplikation von links mit der ***linken unteren Dreiecksmatrix***

$$L_k = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & l_{k+1,k} & 1 & & & \\ & & \vdots & 0 & \ddots & & \\ & & \vdots & \vdots & \ddots & \ddots & \\ 0 & & l_{n,k} & 0 & \cdots & 0 & 1 \end{pmatrix},$$

wobei $|l_{ij}| \leq 1$ ist, weil im ersten Schritt des Algorithmus ja der größte Wert in der Spalte als Pivotelement ausgewählt wurde.

Insgesamt erhalten wir also die Matrix im $(k+1)$–ten Schritt wie folgt:

$$A^{(k+1)} = L_k \cdot P_k \cdot A^{(k)}.$$

Wie im Beispiel vorher wird also in k-ten Schritt die aktuelle Zeile mit jener vertauscht, die in der k-ten Spalte unterhalb der k-ten Zeile den betragsmäßig größten Eintrag enthält.

44.2 Matrix–Zerlegungen

Für mehrfach auftretende Rechnungen mit Matrizen ist es oft hilfreich, diese vorher in eine geeignete Form zu bringen. Ein Beispiel hierfür ist die Zerlegung einer Matrix in ein Produkt aus Matrizen mit speziellen Eigenschaften, wie etwa oberen Dreiecks-Matrizen oder Diagonalmatrizen.

Lemma 44.3. *Wir verwenden die Notationen aus dem vorigen Algorithmus 44.2: Für $j < k$ gilt*

$$P_k \cdot L_j \cdot P_k = \tilde{L}_j,$$

wobei sich $\tilde{L}_j$ von L_j nur durch die Anordnung der Elemente in der j–ten Spalte unterscheidet.

Beweis. Einfaches Nachrechnen. □

Als Folgerung aus dem obigen Gaußalgorithmus erhalten wir die Existenz einer Zerlegung einer gegebenen invertierbaren Matrix in eine linke untere und eine rechte obere Dreiecksmatrix, genauer:

Korollar 44.4 (LR–Zerlegung). *Sei $A \in \mathrm{GL}(n, \mathbb{R})$ eine invertierbare Matrix. Dann existiert eine Permutationsmatrix P, eine **unipotente** untere Dreiecksmatrix L (d.h. mit 1–en auf der Diagonalen),*

$$L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{ij} & \ddots & \\ & & 1 \end{pmatrix},$$

mit $|l_{ij}| \leq 1$, sowie eine obere Dreiecksmatrix

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix},$$

so dass

$$L \cdot R = P \cdot A.$$

Beweis. Nach dem Gaußalgorithmus mit Spaltenpivotierung 44.2 ergibt sich im n–ten Eliminationsschritt eine rechte obere Dreiecksmatrix $R = A^{(n)}$ mit

$$A^{(n)} = L_{n-1}P_{n-1}L_{n-2}P_{n-2}\cdots L_1P_1A.$$

Setzen wir

$$\tilde{L}_{n-1} := L_{n-1}, \quad \tilde{L}_k := P_{n-1}\cdots P_{k+1}L_kP_{k+1}\cdots P_{n-1},$$

so sind die $\tilde{L}_k$ nach dem Lemma fast wieder die L_k (es sind nur zwei Werte in der k–ten Spalte vertauscht) und es gilt:

$$\tilde{L}_k \cdot P_{n-1}\cdots P_{k+1} = P_{n-1}\cdots P_{k+1}\cdot L_k,$$

da ja $(P_j)^{-1} = P_j$. Damit können wir $R = A^{(n)}$ schrittweise umformen:

$$\begin{aligned} R = A^{(n)} &= \tilde{L}_{n-1}P_{n-1}L_{n-2}P_{n-2}\cdots L_1P_1A \\ &= \tilde{L}_{n-1}\tilde{L}_{n-2}P_{n-1}P_{n-2}L_{n-3}\cdots L_1P_1A \\ &= \tilde{L}_{n-1}\tilde{L}_{n-2}\tilde{L}_{n-3}P_{n-1}P_{n-2}P_{n-3}L_{n-4}\cdots L_1P_1A \\ &= \cdots \\ &= \tilde{L}_{n-1}\tilde{L}_{n-2}\cdots\tilde{L}_1P_{n-1}\cdots P_2P_1A. \end{aligned}$$

Setzen wir nun $\tilde{L} := \tilde{L}_{n-1}\tilde{L}_{n-2}\cdots\tilde{L}_1$ und $P := P_{n-1}\cdots P_2P_1$, so ergibt sich $R = \tilde{L}\cdot P\cdot A$ und mit $L := (\tilde{L})^{-1}$ schließlich

$$L\cdot R = P\cdot A,$$

wie behauptet war. □

Gibt es eine solche Zerlegung, für die P die Einheitsmatrix ist, so sagt man, A **besitzt eine** LR**–Zerlegung**. Ein Vorteil einer existierenden LR–Zerlegung einer Matrix ist beispielsweise, dass das Invertieren der Dreiecksmatrizen einfacher ist als das Invertieren der ursprünglichen Matrix A.

Wie wir im folgenden Satz sehen werden, besitzen symmetrische positiv definite Matrizen eine LR-Zerlegung. Doch nicht nur das; sie ist sogar symmetrisch:

Satz 44.5 (Cholesky Zerlegung). *Sei $A > 0$, $A \in \mathrm{GL}(n,\mathbb{R})$, eine symmetrische positiv definite Matrix.*

1. *Dann exisitert eine unipotente untere Dreiecksmatrix L und eine Diagonalmatrix D mit positiven Einträgen, so dass*

$$A = L\cdot D\cdot L^{\mathrm{t}}.$$

2. *Setzen wir*

$$D^{\frac{1}{2}} := \sqrt{D} := \begin{pmatrix} \sqrt{d_{11}} & & \\ & \ddots & \\ & & \sqrt{d_{nn}} \end{pmatrix}$$

und $\tilde{L} = L\sqrt{D}$*, so gilt:*

$$A = \tilde{L} \cdot \tilde{L}^{\mathrm{t}}.$$

Beweis. Die zweite Aussage folgt sofort aus der ersten. Wir beweisen also nur diese erste. Man kann zeigen, dass eine positiv definite Matrix $A = (a_{ij})$ positive Diagonaleinträge besitzt, d.h. $a_{ii} > 0$ für alle i. Für solch eine Matrix A gilt nämlich nach Definition $v^{\mathrm{t}} A v > 0$ für alle Vektoren $0 \neq v \in \mathbb{R}^n$, insbesondere also auch für $v = e_i$, den i-ten Einheitsvektor. Für diesen ergibt sich: $0 < {e_i}^{\mathrm{t}} A e_i = {e_i}^{\mathrm{t}}(a_{1i}, \ldots, a_{ni})^{\mathrm{t}} = a_{ii}$.

Elimination der anderen Einträge der ersten Spalte von A wird realisiert von einer Matrix L_1. Mit $z := (a_{21}, \ldots, a_{n1})^{\mathrm{t}}$ schreibt sich

$$A = \begin{pmatrix} a_{11} & z^{\mathrm{t}} \\ z & * \end{pmatrix}$$

und

$$L_1 \cdot A = \begin{pmatrix} a_{11} & z^{\mathrm{t}} \\ 0 & \\ \vdots & * \\ 0 & \end{pmatrix} \text{ mit } L_1 = \begin{pmatrix} 1 & & 0 \\ -\frac{a_{21}}{a_{11}} & \ddots & \\ \vdots & & \ddots \\ -\frac{a_{n1}}{a_{11}} & & 1 \end{pmatrix}.$$

Die oberste Zeile können wir wegen der Symmetrie von A ebenfalls mit L_1 eliminieren:

$$L_1 \cdot A \cdot {L_1}^{\mathrm{t}} = \begin{pmatrix} a_{11} & 0 \\ 0 & \\ \vdots & \tilde{A} \\ 0 & \end{pmatrix},$$

wobei, wie man zeigen kann, $\tilde{A}$ wieder symmetrisch und positiv definit ist. So können wir fortfahren und erhalten damit die Behauptung. □

44.3 Fehleranalyse

Die numerische Behandlung eines Problems besteht in drei Schritten aus:

$$\text{Eingabe} \longrightarrow \text{Algorithmus} \longrightarrow \text{Ausgabe}.$$

Fehler im Resultat können drei Ursachen haben

Fehler in Eingabe $\longrightarrow$ Fehler im Algorithmus $\longrightarrow$ Fehler in Ausgabe.

Wie stark der Fehler bei exaktem Algorithmus von den Eingabefehlern abhängt, misst die **Kondition des Problems**. Fehler, die sich wegen der Realisierungen eines Algorithmus ergeben oder fortsetzen, werden durch die **Stabilität** gemessen.

Für eine Zahl $x \in \mathbb{R}$ hat ihre **Fließkommadarstellung** (auch **Gleitkommadarstellung** genannt) $fl(x)$ einen **relativen Fehler**

$$\frac{|x - fl(x)|}{|x|} \leq \frac{d^{-k+1}}{2} =: \text{eps},$$

wenn wir auf k Ziffern im d–adischen System genau rechnen. Meist ist $d = 2$ oder $d = 10$. Es ist

$$fl(x) = \pm a \cdot d^e \text{ mit } a = \sum_{i=0}^{k} a_i d^{-i},\ 0 \leq a_i < d.$$

a ist eine d**–adische Zahl** mit einer Vorkomma- und k Nachkommaziffern und für den Exponenten e gilt $e \in \{e_{min}, \ldots, e_{max}\} \subset \mathbb{Z}$. Je nach Compiler ist das oben definierte eps $\approx 10^{-7}$ oder auch kleiner. Alternativ könnte man freilich auch **Fixkommazahlen** einsetzen, bei denen also die Anzahl der Nachkommastellen fest gehalten wird und nicht die Genauigkeit. Oft kommen allerdings Fließkommazahlen nach dem IEEE 754 Standard zum Einsatz.

44.3.1 Kondition eines Problems

Wir fassen einen Algorithmus als eine **Realisierung einer Abbildung**

$$f\colon E \to R,$$

von einer Eingabemenge $E \subset \mathbb{R}^N$ in eine Resultatmenge $R \subset \mathbb{R}^M$ auf.

Definition 44.6. *Die **absolute Kondition** eines Problems $f\colon E \to R$ im Eingabepunkt $x \in E \subset \mathbb{R}^N$ ist die kleinste Zahl*

$$\kappa_{abs} \geq 0,$$

so dass für $\tilde{x} \to x$ gilt:

$$\|f(\tilde{x}) - f(x)\| \leq \kappa_{abs} \cdot \|\tilde{x} - x\| + o(\|\tilde{x} - x\|).$$

*Die **relative Kondition** κ_{rel} eines Problems $f\colon E \to R$ ist die kleinste Zahl $\kappa_{rel} \geq 0$, so dass für $\tilde{x} \to x$:*

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq \kappa_{rel} \cdot \left(\frac{\|\tilde{x} - x\|}{\|x\|}\right) + o\left(\frac{\|\tilde{x} - x\|}{\|x\|}\right).$$

Offenbar sollte die relative Kondition möglichst klein sein. Bei Werten, die viel größer als 1 sind, spricht man von einer **schlechten Kondition** und bei Problemen, für die die Kondition ∞ ist, von **schlecht gestellten Problemen**.

Bemerkung 44.7. Ist f total diffbar in x, so gilt nach Definition des Differentials als beste lineare Approximation von f:

$$\kappa_{abs} = \|Df(x)\|,$$

die Matrixnorm des Differentials $Df(x)$, und

$$\kappa_{rel} = \frac{\|x\|}{\|f(x)\|} \cdot \|Df(x)\|.$$

Beispiel 44.8 (Kondition der Addition). Wir betrachten die Addition:

$$f\colon \mathbb{R}^2 \to \mathbb{R},\ f(a,b) = a + b.$$

f ist diffbar mit $(Df)(a,b) = (1,1)$. Verwenden wir im $\mathbb{R}^2$ die Betragssummennorm und für Df die induzierte Matrixnorm (d.h. die Spaltensummennorm), so ergibt sich demnach

$$\kappa_{abs} = \|(1,1)\| = 1 \ \text{ und } \ \kappa_{rel} = \frac{|a| + |b|}{|a+b|} \cdot 1 = \frac{|a| + |b|}{|a+b|}.$$

Falls $a \approx -b$ ist, ist also $\kappa_{rel} \gg 1$; man spricht daher bei der Subtraktion fast gleich großer Zahlen von **Auslöschung**; diese sollte man vermeiden.

Betrachten wir beispielsweise $\pi = 3.14159265358\ldots$ und 3.141 bei einer Fließkommarechnung auf 4 Stellen genau: Die Subtraktion $\pi - 3.141$ liefert nicht $0.0005927 = 5.927 \cdot 10^{-4}$, sondern $3.142 - 3.141 = 1.000 \cdot 10^{-3}$, was fast doppelt so viel ist wie das erhoffte Ergebnis.

Bemerkung 44.9. Wir haben schon im einführenden Beispiel gesehen, dass bei der Addition von Fließkommazahlen weitere Gesetze der üblichen Arithmetik nicht mehr gelten. Beispielsweise ändern die Addition oder Subtraktion einer betragsmäßig viel kleineren Zahl eine gegebene Zahl gar nicht (dieses Phänomen heißt auch **Absorption**):

$$\begin{aligned} 1.000 \cdot 10^2 + 1.000 \cdot 10^{-3} &= 1.000 \cdot 10^2 + 0.000010 \cdot 10^2 \\ &= 1.000 \cdot 10^2 + 0.000 \cdot 10^2 = 1.000 \cdot 10^2. \end{aligned}$$

Ebenso gelten im Allgemeinen weder das Assoziativgesetz noch das Distributivgesetz.

Beispiel 44.10 (Lösung quadratischer Gleichungen). Wir betrachten die quadratische Gleichung in x:

$$x^2 + px + q = 0$$

mit $p, q \in \mathbb{R}$. Bekanntlich hat sie wegen $0 = x^2 + px + q = (x + \frac{p}{2})^2 - (\frac{p}{2})^2 + q \iff x + \frac{p}{2} = \pm\sqrt{(\frac{p}{2})^2 - q}$ die beiden Lösungen

$$x_{1,2} = \frac{-p \pm \sqrt{p^2 - 4q}}{2}.$$

Ist eine der Lösungen nahe bei Null, also $p \approx \pm\sqrt{p^2 - 4q}$, so gibt diese Formel keine guten Ergebnisse.

Eine bessere Formel erhält man folgendermaßen: Zunächst ist

$$x_1 = \frac{-p - \operatorname{sign}(p) \cdot \sqrt{p^2 - 4q}}{2}$$

eine **auslöschungsfreie Formel** für x_1. Wegen

$$(x - x_1)(x - x_2) = x^2 - (x_1 + x_2) \cdot x + x_1 x_2$$

liefert eine Koeffizientenvergleich mit $x^2 + px + q$ die Identitäten: $p = -(x_1 + x_2)$ und $q = x_1 x_2$. Diese Eigenschaft heißt auch **Satz von Vieta**; mit ihr können wir nun x_2 ebenfalls auslöschungsfrei berechnen:

$$x_2 = \frac{q}{x_1}.$$

Beispiel 44.11 (Kondition eines quadratischen linearen Gleichungssystems). Wir betrachten $Ax = b$, wobei A eine invertierbare quadratische Matrix ist. Es gibt zwei Fälle:

- A fest, b variabel: $f\colon \mathbb{R}^n \to \mathbb{R}^n$, $b \mapsto A^{-1}b = x$. Diese Abbildung ist linear und daher diffbar mit $Df = A^{-1}$, also $\kappa_{abs} = \|A^{-1}\|$ und $\kappa_{rel} = \frac{\|b\|}{\|A^{-1}b\|} \cdot \|A^{-1}\| = \frac{\|Ax\|}{\|x\|} \cdot \|A^{-1}\|$.
- b fest, A variabel: Die Abbildung ist nun $\mathbb{R}^{n\times n} \supset \mathrm{GL}(n, \mathbb{R}) \to \mathbb{R}^n$, $A \mapsto A^{-1}b$, zusammensetzbar aus $A \mapsto A^{-1} \mapsto A^{-1}b$. Die Differenzierbarkeit des ersten Teils ist hier aber nicht klar.

Lemma 44.12. *Die Abbildung* $g\colon \mathrm{GL}(n, \mathbb{R}) \to \mathrm{GL}(n, \mathbb{R}) \subset \mathbb{R}^{n\times n} = \mathbb{R}^{n^2}$, $A \mapsto A^{-1}$ *ist diffbar mit Differential*

$$Dg\colon \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}, (Dg)(C) = -A^{-1}CA^{-1}.$$

Beweis. ... □

Direkt ergibt sich daraus:

Proposition 44.13. *Die Abbildung* $f\colon \mathrm{GL}(n,\mathbb{R}) \to \mathbb{R}^n$, $f(A) = A^{-1}b$, *hat das Differential* $Df(C) = -A^{-1}C(A^{-1}b) = -A^{-1}Cx$.

Für $A \in \mathrm{GL}(n,\mathbb{R}) \subset \mathbb{R}^{n^2}$ gilt daher

$$\kappa_{abs} = \sup_{\|C\|=1} \|A^{-1}Cx\| \leq \|A^{-1}\| \cdot \|x\|,$$

also:

$$\kappa_{rel} = \frac{\|A\|}{\|x\|} \cdot \|\kappa_{abs}\| \leq \frac{\|A\|}{\|x\|} \cdot \|A^{-1}\| \cdot \|x\| = \|A^{-1}\| \cdot \|A\|.$$

Deshalb definieren wir:

Definition 44.14. *Die* ***Kondition einer invertierbaren Matrix*** $A \in \mathrm{GL}(n,\mathbb{R})$ *ist*

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|.$$

Bemerkung 44.15. 1. Insbesondere gilt nach dem Vorgehenden für $A \in \mathrm{GL}(n,\mathbb{R})$:

$$\kappa_{rel} \leq \kappa(A) \in [1,\infty[.$$

2. Nach Definition der zugehörigen Matrixnorm ist für $A \in \mathrm{GL}(n,\mathbb{R})$:

$$\|A\| = \max_{\|x\|=1} \|Ax\|.$$

Man kann zeigen (siehe Aufgabe 44.4), dass

$$\|A^{-1}\| = \max_{\|x\|=1} \|A^{-1}x\| = \frac{1}{\min_{\|x\|=1} \|Ax\|}.$$

Bemerkung 44.16. Die zugrundeliegende Vektornorm sei die euklidische Norm $\|.\| = \|.\|_2$. Die zugehörige Matrixnorm ist also die Spektralnorm.

1. Es gilt:

$$\kappa(A) = 1 \iff A = \lambda B$$

für gewisse $\lambda \in \mathbb{R}^*$ und $B \in O(n)$. Mit anderen Worten: Genau die orthogonalen Matrizen (also jene mit $BB^t = E$) sind **optimal konditioniert**, d.h. die Kondition ist die der Einheitsmatrix, also 1.

2. Ist $A \in \mathrm{GL}(n, \mathbb{R})$ symmetrisch, so ist

$$\kappa(A) = \frac{\lambda_{max}}{\lambda_{min}},$$

wobei

$$\lambda_{max} := \|A\| = \max\{|\lambda| \mid \lambda \text{ Eigenwert von } A\},$$
$$\lambda_{min} := \|A^{-1}\| = \min\{|\lambda| \mid \lambda \text{ Eigenwert von } A\}.$$

Beweis. 1. Es gilt wegen Bemerkung 44.15:

$$\begin{aligned}\kappa(A) = 1 &\iff \max_{\|x\|=1} \|Ax\| = \min_{\|x\|=1} \|Ax\| =: \lambda \\ &\iff B := \frac{1}{\lambda} \cdot A \text{ erfüllt } \|Bx\| = 1 \ \forall x \text{ mit } \|x\| = 1 \\ &\iff B \in \mathrm{O}(n).\end{aligned}$$

Die letzte Äquivalenz ergibt sich, weil die vorletzte Bedingung bedeutet, dass die Länge aller Vektoren unter B gleich bleibt (weil $\|B_{\cdot}x\| = |k| \cdot \|Bx\| = \|x\|$ für jedes $k \in \mathbb{R}$ gilt und jeder beliebiger Vektor v ein Vielfaches eines Vektors mit Norm 1 ist), aber solche Matrizen sind gerade die orthogonalen, siehe S. 3.

2. Klar mit Bemerkung 44.15.

□

44.3.2 Stabilität eines Algorithmus

Wir betrachten eine Gleitkommarealisierung

$$\tilde{f} \colon \tilde{E} \to \tilde{R}$$

eines Algorithmus $f \colon E \to R$. Wir messen die Stabilität der Realisierung $\tilde{f}$ im Punkt $\tilde{x} \in \tilde{E}$ dadurch, in wieweit er bei einem unvermeidlichen Eingabefehler eps den unvermeidlichen Ausgabefehler $\kappa_{rel} \cdot$ eps noch verschärft. Im Folgenden schreiben wir der Kürze halber $\kappa := \kappa_{rel}$.

Der Stabilitätsindex

Definition 44.17. *Der **Stabilitätsindex** σ von $\tilde{f}$ in $\tilde{x}$ ist die kleinste Zahl $\sigma \geq 0$, so dass*

$$\frac{\|\tilde{f}(\tilde{x}) - f(\tilde{x})\|}{\|f(\tilde{x})\|} \leq \sigma \cdot \kappa \cdot \text{eps} + o(\text{eps}) \textit{ für } \text{eps} \geq \tilde{x} - x \longrightarrow 0.$$

Lemma/Definition 44.18. *Für eine **Elementaroperation** $o \in \{+, -, \cdot, \div\}$ und ihre Gleitkommarealisierung $\tilde{o} \in \{\tilde{+}, \tilde{-}, \tilde{\cdot}, \tilde{\div}\}$ gilt:*

$$\sigma \cdot \kappa \leq 1.$$

Beweis. $\tilde{o} \in \{\tilde{+}, \tilde{-}, \tilde{\cdot}, \tilde{\div}\}$. Dann gilt wegen der Definition von eps:

$$a \,\tilde{o}\, b = (a \circ b) \cdot (1 + \varepsilon)$$

für ein ε mit $0 \leq |\varepsilon| \leq \text{eps}$. Also:

$$\left| \frac{a \,\tilde{o}\, b - a \circ b}{a \circ b} \right| = \frac{|(1 + \varepsilon) \cdot (a \circ b) - (a \circ b)|}{|a \circ b|} = |\varepsilon| \leq \text{eps}.$$

Damit folgt, weil dies σ ja die kleinste Zahl ist, so dass dies $\leq \sigma\kappa\text{eps}$, ist, dass für den Vorfaktor gilt: $\sigma \cdot \kappa \leq 1$. □

Beispiel 44.19. Für die Subtraktion fast gleich großer Zahlen gilt $\kappa \gg 0$, also, nach dem Lemma, $\sigma \ll 1$.

Zusammengesetzte Algorithmen

Bei der Vorwärtsanalyse eines Algorithmus $f\colon E \to R$ zerlegt man den Algorithmus häufig in Schritte:

$$f = g \circ h\colon \mathbb{R}^n \xrightarrow{h} \mathbb{R}^l \xrightarrow{g} \mathbb{R}^m.$$

Lemma 44.20. *Sei $\tilde{f} = \tilde{h} \circ \tilde{g}$ eine Gleitkommarealisierung des zusammengesetzten Algorithmus $f = h \circ g$. Dann gilt:*

$$\sigma_f \cdot \kappa_f \leq \sigma_h \cdot \kappa_h + \sigma_g \cdot \kappa_g \cdot \kappa_h.$$

Beweis. ... □

Folgerung 44.21. *Unvermeidliche Subtraktionen möglichst an den Anfang eines Algorithmus stellen.*

Beispiel 44.22 (Summation). Wir setzen

$$s_n\colon \mathbb{R}^n \to \mathbb{R},\ (x_1, \ldots, x_n) \mapsto \sum_{i=1}^{n} x_i.$$

Dies berechnen wir rekursiv durch $s_n = s_{n-1} \circ \alpha_n$, wobei

$$\alpha_n\colon \mathbb{R}^n \to \mathbb{R}^{n-1},\ (x_1, \ldots, x_n) \mapsto (x_1 + x_2, x_3, \ldots, x_n).$$

Kondition und Stabilitätsindex von α_n stimmen mit denen der Addition überein: $\kappa_{\alpha_n} = \kappa_+,\ \sigma_{\alpha_n} = \sigma_+$. Wir schreiben $\kappa_j := \kappa_{s_j}, \sigma_j := \sigma_{s_j}$. Damit ist

$$\sigma_n\kappa_n \leq (\sigma_{n-1} + \kappa_+\sigma_+)\kappa_{n-1} \leq (1 + \sigma_{n-1})\kappa_{n-1}$$

wegen Lemma 44.20 und Lemma 44.18. Auf der anderen Seite ist nach der Definition der relativen Kondition und der Dreiecksungleichung:

$$\kappa_n = \frac{\sum_{i=1}^n |x_i|}{|\sum_{i=1}^n x_i|} \cdot 1 \geq 1 \quad \text{und} \quad \kappa_{n-1} = \frac{|x_1 + x_2| + \sum_{i=3}^n |x_i|}{|\sum_{i=1}^n x_i|} \leq \kappa_n.$$

Damit folgt mit der Ungleichung weiter oben:

$$\sigma_n \leq (1 + \sigma_{n-1}).$$

Außerdem ist, wieder wegen 44.18, $\sigma_2 = \sigma_+ \leq \frac{1}{\kappa_+} \leq 1$ (wegen $|\kappa_+| \geq 1$). Induktiv erhalten wir also:

$$\sigma_n \leq n - 1.$$

Definition 44.23. *Eine Gleitkommarealisierung $\tilde{f}$ eines Algorithmus f heißt **numerisch stabil**, wenn $\sigma \leq n$, wobei n die Anzahl der Elementaroperationen im Algorithmus ist.*

Beispiel 44.24. Summation ist numerisch stabil (siehe Beispiel 44.22).

Für Funktionen in einer Variablen lässt sich die Abschätzung für σ_f für zusammengesetzte f, die in Lemma 44.20 gegeben wurde, verbessern:

Bemerkung 44.25. Ist $f\colon \mathbb{R} \xrightarrow{g} \mathbb{R} \xrightarrow{h} \mathbb{R}$ zusammengesetzt und diffbar, so gilt:

$$\sigma_f \leq \frac{\sigma_h}{\kappa_g} + \sigma_g.$$

Beweis. Nach Definition der relativen Kondition ist

$$\begin{aligned}\kappa_f &= \frac{|x|}{f(x)} \cdot |Df(x)| = \frac{|x| \cdot |h'(g(x))| \cdot |g'(x)|}{|h(g(x))|} \\ &= \frac{|g(x)| \cdot |h'(g(x))|}{|h(g(x))|} \cdot \frac{|x| \cdot |g'(x)|}{|g(x)|} \\ &= \kappa_h \cdot \kappa_g.\end{aligned}$$

Einsetzen in Lemma 44.20 liefert die Behauptung. □

Beispiel 44.26 (Auswertung trigonometrischer Polynome). Wir betrachten sogenannte **trigonometrische Polynome**:

$$f(x) = \sum_{k=1}^{n} \Big(a_k \cos(kx) + b_k \sin(kx) \Big).$$

Wir verwenden die Rekursionsformeln:

$$\cos((k+1)x) = 2\cos(x) \cdot \cos(kx) - \cos((k-1)x),$$

die aus den Additionstheoremen 7.25 für den Cosinus folgen, weil danach $\cos((k+1)x) = \cos(kx + x) = \cos(kx)\cos(x) - \sin(kx)\sin(x)$, etc. Um den Stabilitätsindex für die Auswertung von f in x abschätzen zu können, betrachten wir wegen der Bemerkung den Kehrwert der Kondition κ_g für $g(x) = \cos(x)$. Per Definition der relativen Kondition ist dies:

$$\frac{1}{\kappa_g} = \frac{|\cos(x)|}{|x|} \cdot \frac{1}{|\sin(x)|} = \Big| \frac{1}{x \tan x} \Big| \xrightarrow[x \to 0]{} \infty.$$

Im Gegensatz dazu erhalten wir für folgende rekursive Formel eine wesentlich bessere Abschätzung: ...

Für $|x| \ll 1$ ist dies tatsächlich besser: Mit $\bar{g}(x) = \sin^2(\frac{x}{2})$ ist $\bar{g}'(x) = \sin\frac{x}{2} \cdot \cos\frac{x}{2}$ und daher:

$$\frac{1}{\kappa_{\bar{g}}} = \left| \frac{\sin^2 \frac{x}{2}}{x} \right| \cdot \frac{1}{\left| \sin\frac{x}{2} \cdot \cos\frac{x}{2} \right|} = \left| \frac{\tan\frac{x}{2}}{x} \right| \xrightarrow[x \to 0]{} \frac{1}{2}.$$

Aufgabe 44.1 (Gauß–Algorithmus mit Pivotierung). Gegeben sei das Gleichungssystem

$$\begin{pmatrix} 1 & 200 \\ 1 & 1 \end{pmatrix} x = \begin{pmatrix} 100 \\ 1 \end{pmatrix}$$

1. Bestimmen Sie die exakte Lösung des Gleichungssystems.
2. Rechnen Sie nun mit 2 signifikanten Dezimalstellen. Bestimmen Sie die Lösung ohne Pivotsuche und mit vollständiger Pivotsuche.

Aufgabe 44.2 (Cholesky–Zerlegung). Bestimmen Sie die Cholesky-Zerlegung der Matrix

$$A = \begin{pmatrix} 6 & -2 & 2 \\ -2 & 5 & 0 \\ 2 & 0 & 7 \end{pmatrix}$$

Lösen Sie mit Hilfe dieser Zerlegung das lineare Gleichungssystem $Ax = b$ für $b = (3, -4, 13)$.

Aufgabe 44.3 (*LR*–Zerlegung). Berechnen Sie mit vollständiger Pivotsuche die *LR*-Zerlegung der Matrix

$$A = \begin{pmatrix} -2 & 7 & -2 \\ 0 & 2 & -1 \\ -4 & 15 & 0 \end{pmatrix}.$$

Lösen Sie mit Hilfe dieser Zerlegung das Gleichungssystem $Ax = b$ für $b = (1, 2, 1)$.

Aufgabe 44.4 (Konditionszahl). Es sei $A \in \mathbb{R}^{n \times n}$ eine invertierbare Matrix. Zeigen Sie:

$$\kappa(A) = \frac{\max_{x \in \mathbb{R}^n, \|x\|=1} \|Ax\|}{\min_{x \in \mathbb{R}^n, \|x\|=1} \|Ax\|}.$$

45

Iterationsverfahren für Eigenwerte und Rang

. . . schneller, genauer, einfach besser als die eher theoretischen Methoden aus dem Abschnitt über lineare Algebra. . .

45.1 Die QR–Zerlegung

Grundlegend für viele der folgenden Verfahren ist die Zerlegung einer Matrix in eine spezielle orthogonale Matrix Q und eine rechte obere Dreiecksmatrix R. Der QR–Algorithmus ist ein Verfahren, eine solche zu berechnen.

Definition 45.1. *Eine orthogonale Matrix der Gestalt*

$$\begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & c & \cdots & d & \\ & & \vdots & & \vdots & \\ & & -d & \cdots & c & \\ & & & & & \ddots \\ & & & & & & 1 \end{pmatrix} \in \mathrm{SO}(n)$$

*mit $c^2 + d^2 = 1$ heißt **Givensrotation**.*

Satz/Definition 45.2 (QR–Zerlegung). *Sei $A \in \mathbb{R}^{n\times n}$. Dann existiert ein Produkt Q von $\binom{n}{2}$ Givensrotationen, so dass*

$$A = Q \cdot R,$$

*wobei R eine obere Dreiecksmatrix ist. Dieses Produkt heißt auch QR–**Zerlegung**.*

Anwendung 45.3. Möchten wir $Ax = b$ lösen, so betrachten wir

$$Rx = Q^{\mathrm{t}} b$$

(da $Q \in \mathrm{SO}(n)$ ist $Q^{-1} = Q^{\mathrm{t}}$). Für $Q \in \mathrm{O}(n)$ sind optimal konditioniert. Die Gleichung $Rx = Q^{\mathrm{t}} b$ lässt sich dann durch rückwärts einsetzen lösen.

Beweis (von Satz 45.2). Wir betrachten zunächst 2×2–Matrizen: Um $(a, b)^{\mathrm{t}} \in \mathbb{R}^2$ auf ein Vielfaches von $(1, 0)^{\mathrm{t}}$ zu drehen, ... □

Statt Rotationen kann man auch Spiegelungen verwenden. Householder hat dies als erster in der Numerik eingeführt:

Definition 45.4. *Sei $v \neq 0 \in \mathbb{R}^n$ ein Vektor. Die Abbildung*

$$Q_v \colon \mathbb{R}^n \to \mathbb{R}^n,\ y \mapsto y - 2\frac{\langle v, y\rangle}{\langle v, v\rangle} \cdot v$$

heißt ***Householder–Reflexion*** *(an der Hyperebene $H_v = \{y \in \mathbb{R}^n \mid \langle v, y\rangle = 0\}$, siehe dazu auch Abschnitt 17.3).*

Die Matrix der Householder–Reflexion ist durch

$$Q_v = E_n - 2\frac{v \cdot v^{\mathrm{t}}}{v^{\mathrm{t}} \cdot v}$$

gegeben, da $\langle v, y\rangle = v^{\mathrm{t}} \cdot y$ und daher $Q_v(y) = (E_n - 2\frac{v \cdot v^{\mathrm{t}}}{v^{\mathrm{t}} \cdot v}) \cdot y$.

Wir wissen aus der linearen Algebra:

1. Die Matrix Q_v ist symmetrisch.
2. $Q_v^2 = E$.
3. $Q_v^{-1} = Q_v = {Q_v}^{\mathrm{t}}$ ist eine orthogonale Matrix.

Sei nun $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$. Analog zu vorher gehen wir rekursiv vor und bezeichnen die Spalten von A mit $a_i \in \mathbb{R}^m$. Wir suchen $Q_1 := Q_v$, so dass:

$$Q_1 \cdot A = \left(\begin{array}{c|c} \alpha_1 & * \\ \hline 0 & \tilde{\alpha}_2 \cdots \tilde{\alpha}_n \end{array}\right).$$

Da $Q_1 \in \mathrm{O}(n) \setminus \mathrm{SO}(n)$ und orthogonale Abbildungen Längen nicht ändern, ist $\alpha_1 = \pm\|a_1\|_2$.

Ferner soll für die erste Spalte dieses Produktes gelten:

$$\alpha \cdot e_1 = \begin{pmatrix} \alpha_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \stackrel{!}{=} Q_1 \cdot a_1 = \left(E_n - 2\frac{v \cdot v^{\mathrm{t}}}{v^{\mathrm{t}} \cdot v}\right) \cdot a_1 = a_1 - 2\frac{\langle v, a_1 \rangle}{v^{\mathrm{t}} \cdot v} \cdot v.$$

Der gesuchte Vektor v liegt wegen dieser linearen Abhängigkeit also in der von e_1 und a_1 aufgespannten Ebene des $\mathbb{R}^n$.

Man kann leicht nachrechnen, dass tatsächlich $v = a_1 - \alpha_1 e_1$ die gewünschte Eigenschaft hat:

$$\begin{aligned} \langle v, v \rangle &= \cdots \\ &= 2\alpha_1(\alpha_1 - a_{11}) \end{aligned}$$

und

$$\begin{aligned} a_1 \mapsto a_1 - 2\frac{\langle v, a_1 \rangle}{\langle v, v \rangle} \cdot v &= \ldots \\ &= \alpha_1 e_1. \end{aligned}$$

Numerisch ist es hier am Günstigstens, $\alpha_1 = -\operatorname{sign}(a_{11}) \cdot \|a_1\|$ zu wählen, um in $v_1 = a_1 - \alpha_1 e_1$ Auslöschung zu vermeiden.

Nach diesem Schritt haben wir mit $Q_1 \cdot A$ das Problem auf eine Matrix $\tilde{A}_2 \in \mathbb{R}^{(m-1)\times(n-1)}$ reduziert. Rekursiv fortgesetzt liefert dies:

Satz 45.5. *Mit $n-1$ Householder–Reflexionen lässt sich eine Matrix $A \in \mathbb{R}^{m\times n}$ QR–zerlegen: $A = Q \cdot R$.*

Bemerkung 45.6. Um aus A die Matrizen Q und R zu berechnen, benötigt man nur unwesentlich mehr Speicherplatz also für A alleine, denn: ...

Bemerkung 45.7 (Eindeutigkeit der QR–Zerlegung). Sei $A \in \mathrm{GL}(n, \mathbb{R})$ und seien $A = QR = \tilde{Q}\tilde{R}$ zwei QR–Zerlegungen. Dann gibt es $\varepsilon_i \in \{\pm 1\}$, so dass für die Matrix

$$\varepsilon = \begin{pmatrix} \varepsilon_1 & & 0 \\ & \ddots & \\ 0 & & \varepsilon_n \end{pmatrix}$$

gilt:

$$\varepsilon \cdot R = \tilde{R},\ \tilde{Q}^{\mathrm{t}} Q = \varepsilon.$$

Beweis. Wir beginnen mit $\varepsilon := \tilde{R} \cdot R^{-1}$ und bezeichnen die Spalten von ε mit $(q_1, \ldots, q_n)$. Da wir aus der Konstruktion wissen, dass der linke obere Eintrag von R und $\tilde{R}$ jeweils $\pm\|a_1\|_2$ ist und da beides rechte obere Dreiecksmatrizen sind, ist $q_1 = \pm e_1$. Analog hat $q_2 \in q_1^{\perp} = e_1^{\perp}$ die Gestalt $\pm e_2$ usw. Tatsächlich kann man nachprüfen, dass diese auch die Eigenschaften mit den Qs erfüllen. □

45.2 Das *QR*–Verfahren

Oft ist man in Anwendungen an den Eigenwerten von Matrizen interessiert, die spezielle Struktur aufweisen. Beispielsweise kommt es häufig vor, dass solche Matrizen symmetrisch sind. In diesem Fall kann man die Eigenwerte mit dem *QR*–Verfahren recht schnell und numerisch stabil berechnen.

Für eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ existiert nach Satz 25.4 über die Hauptachsentransformation eine orthogonale Matrix $S \in \mathrm{O}(n)$, so dass

$$S^{\mathrm{t}} A\, S = D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

eine Diagonalmatrix ist. Um diese mit den Methoden aus dem Kapitel über lineare Algebra zu bestimmen, berechnet man zunächst

$$\chi_A(t) = \det(tE - A) \in \mathbb{R}[t]$$

und eine Nullstelle $\lambda_1 \in \mathbb{R}$. Dann löst man $Ax = \lambda_1 x$, um einen Eigenvektor v_1 mit $\|v_1\| = 1$ zu erhalten und geht induktiv zu $H_1 = \langle v_1^{\perp} \rangle$, dem zu v_1 orthogonalen Untervektorraum, über.

Hierbei hat man das ernsthafte Problem, dass man eine Nullstelle eines Polynoms berechnen muss, das möglicherweise großen Grad hat. Zwar haben wir in der Analysis das Newtonverfahren zur Berechnung von Nullstellen (Abschnitt 10.3) kennengelernt, doch dies ist leider nur ein lokales Verfahren, das nur unter gewissen Voraussetzungen eine Nullstelle liefert.

Schneller, numerisch stabiler und ohne das Problem der Nullstellenberechnung kommt das folgende Verfahren aus, das wegen seiner Struktur auch als **Iterationsverfahren** bezeichnet wird:

Algorithmus 45.8 (*QR*–Verfahren). *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Wir berechnen induktiv eine Folge (A_k) von $n \times n$–Matrizen durch*

1. $A_1 := A$,
2. $A_k := Q_k R_k$, *wobei dies eine QR–Zerlegung von A_k sei.*
3. $A_{k+1} := R_k Q_K$.

Bemerkung 45.9. Wegen $A_{k+1} = R_k Q_K$ und $R_k = {Q_k}^{\mathrm{t}} A_k$ (weil die Q_k orthogonal sind, also $Q_k^{-1} = {Q_k}^{\mathrm{t}}$ gilt) folgt:

$$A_{k+1} = {Q_k}^{\mathrm{t}} A_k\, Q_k.$$

Alle Matrizen A_k sind also zu A mit orthogonalen Matrizen konjugiert und daher auch symmetrisch.

Satz 45.10. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch mit n vom Betrag her verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_n$, die betraglich der Größe nach sortiert sind, d.h.*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0.$$

Dann konvergiert die Folge (A_k) aus dem QR–Verfahren gegen eine Diagonalmatrix mit Einträgen $\lambda_1, \ldots, \lambda_n$, die in der Regel der Größe nach sortiert sind. Ist Letzteres der Fall, so gilt für die anderen Einträge von $A_k = (a_{ij}^{(k)})$:

$$a_{ij}^{(k)} \longrightarrow 0 \text{ und } a_{ij}^{(k)} \in o\Big(\Big|\frac{\lambda_j}{\lambda_i}\Big|^k\Big) \text{ für } k \to \infty \text{ und } j > i.$$

Bemerkung 45.11. Man kann zeigen, dass bei mehrfachen Eigenwerten, etwa $\lambda_k = \lambda_{k+1}$, auch noch Konvergenz vorliegt. Bei $\lambda_k = -\lambda_{k+1}$ können 2×2–Blöcke stehen bleiben.

Beweis (von Satz 45.10). Wir zeigen zunächst für die Potenzen von A:

$$A^k = \underbrace{Q_1 \cdots Q_k}_{=:P_k} \underbrace{R_k \cdots R_1}_{=:U_k}$$

mit Induktion nach k. Für $k = 1$ ist nichts zu zeigen: $A = Q_1 R_1$.

Für den Induktionsschritt betrachten wir A_{k+1}: Wegen Bemerkung 45.9 ist

$$\begin{aligned} A_{k+1} &= Q_{k+1}R_{k+1} = R_k Q_k \\ &= {Q_k}^{\mathrm{t}} A_k \, Q_k \\ &= {Q_k}^{\mathrm{t}} {Q_{k-1}}^{\mathrm{t}} \cdots {Q_1}^{\mathrm{t}} A \, Q_1 \, \cdots \, Q_k \\ &= {P_k}^{\mathrm{t}} A \, P_k. \end{aligned}$$

Damit folgt mit der Induktionsvoraussetzung:

$$\begin{aligned} A^{k+1} = A \cdot A^k \overset{\text{I.-V.}}{=} & \; A P_k U_k = P_k \, {P_k}^{\mathrm{t}} A \, P_k \, U_k \\ = & \; P_k A_{k+1} U_k = P_k Q_{k+1} R_{k+1} U_k = P_{k+1} U_{k+1}. \end{aligned}$$

Die Aussage über die Potenzen A^k ist damit bewiesen.

Wir möchten nun noch eine weitere QR–Zerlegung von A^k herleiten und betrachten dazu eine Diagonalisierung von A:

$$S \, D \, S^{\mathrm{t}} = A$$

mit $S \in \mathrm{O}(n)$ und D die Diagonalmatrix mit $\lambda_1, \ldots, \lambda_n$ auf der Diagonalen. Da $S^{\mathrm{t}} S = E$, erhalten wir eine weitere Darstellung von A^k:

$$A^k = (S \, D \, S^{\mathrm{t}})^k = S \, D^k \, S^{\mathrm{t}}.$$

Nehmen wir nun an, dass S^t eine LR–Zerlegung hat, dass also insbesondere $S^t = LR$ (andernfalls müssen wir eine Permutation der Zeilen von S vornehmen, was zu einer Permutation der Diagonalelemente $\lambda_1, \ldots, \lambda_n$ führt), so folgt:

$$A^k = S\, D^k\, L\, R = S\, D^k\, L\, D^{-k}\, D^k\, R.$$

Da aber $L = (l_{ij})$ eine linke untere Dreiecksmatrix ist, gilt:

$$(D^k\, L\, D^{-k})_{ij} = l_{ij} \cdot \frac{\lambda_i^k}{\lambda_j^k} = l_{ij} \cdot \Big(\frac{\lambda_i}{\lambda_j}\Big)^k \text{ für } i > j.$$

Somit gilt, da $|\lambda_i| < |\lambda_j|$ für $i > j$ und da die Diagonalelemente offenbar 1 sind (weil L unipotent ist):

$$D^k\, L\, D^{-k} = E + F_k \underset{k\to\infty}{\longrightarrow} E \quad (\text{also } F_k \to 0\,).$$

Wir erhalten also:

$$A^k = S\,(E + F_k)\, D^k\, R.$$

Ist $E + F_k = \widetilde{Q}_k \widetilde{R}_k$ eine QR–Zerlegung von $E + F_k$, wobei $\widetilde{R}_k$ strikt positive Diagonalelemente hat (die Vorzeichen kann man in Q_k unterbringen), so ist

$$A^k = S\,(\widetilde{Q}_k \widetilde{R}_k)\, D^k\, R) = (S\, \widetilde{Q}_k) \cdot (\widetilde{R}_k\, D^k\, R)$$

eine weitere QR–Zerlegung von A^k.

Wir hatten zuvor gezeigt, dass $A^k = P_k U_k$ ebenfalls eine solche ist — bis auf Vorzeichen (die wir leicht ändern könnten) ist eine QR–Zerlegung nach Bemerkung 45.7 aber eindeutig, so dass folgt:

$$S\widetilde{Q}_k = P_k \text{ und } U_k = \widetilde{R}_k D^k R.$$

Wegen $E + F_k \to E$ für $k \to \infty$ gilt aber

$$\widetilde{Q}_k \to E \text{ und } \widetilde{R}_k \to E \text{ für } k \to \infty,$$

so dass folgt:

$$\begin{aligned} Q_k &= {P_{k-1}}^t\, P_k = {\widetilde{Q}_{k-1}}^t\, S^t\, S\, \widetilde{Q}_k = {\widetilde{Q}_{k-1}}^t\, {\widetilde{Q}_k}^t \longrightarrow E, \\ R_k &= U_k\, U_{k-1}^{-1} = \widetilde{R}_k\, D^k\, R\, R^{-1}\, D^{-k+1}\, \widetilde{R}_{k-1}^{-1} = \widetilde{R}_k\, D\, \widetilde{R}_{k-1}^{-1} \underset{k\to\infty}{\longrightarrow} D. \end{aligned}$$

Schließlich folgt:

$$A_k = Q_k R_k \underset{k\to\infty}{\longrightarrow} ED = D.$$

Die genauere Ausssage über das Konvergenzverhalten ergibt sich aus dem von $D^k\, L\, D^{-k} \longrightarrow E$. Dies führen wir hier aber nicht aus. □

45.3 Vektoriteration

Für symmetrische Matrizen gibt es auch ein iteratives Verfahren, bei dem nicht Folgen von Matrizen, sondern Folgen von Vektoren berechnet werden. Leider liefert es nur den größten Eigenwert der Matrix; doch manchmal ist dies genau die benötigte Information.

Satz 45.12 (Vektoriteration). *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix mit Eigenwerten λ_i, für die $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$ gilt. Ist $x_0 \notin \operatorname{Eig}(A,\lambda_1)^\perp$, also nicht senkrecht zum Eigenraum zu λ_1, so konvergiert die Folge von Vektoren*

$$x_{k+1} = \frac{Ax_k}{\|Ax_k\|} \in \mathbb{R}^n,$$

falls $\lambda_1 > 0$ ist, gegen einen normierten Eigenvektor zu λ_1. Ist $\lambda_1 < 0$, so konvergiert die Teilfolge (x_{2k}) gegen einen normierten Eigenvektor zu λ_1.

Beweis. Sei $v_1,\dots,v_n$ eine Orthonormalbasis aus Eigenvektoren zu $\lambda_1,\dots,\lambda_n$ (die nach Satz 25.4 über die Hauptachsentransformation existiert). Dann lässt sich x_0 schreiben als

$$x_0 = \alpha_1 v_1 + \cdots + \alpha_n v_n$$

für gewisse $\alpha_i \in \mathbb{R}$ und es gilt $\alpha_1 \neq 0$ nach Voraussetzung. Dann ist auch $\|A^k x_0\| = \|\sum_{i=1}^n \alpha_i \lambda_i^k v_i\| \neq 0$ (da $v_1 \perp \langle v_2,\dots,v_n\rangle$) und wir können $A^k x_0$ normieren:

$$x_k = \frac{A^k x_0}{\|A^k x_0\|}.$$

Da außerdem nach Voraussetzung $|\lambda_1| > |\lambda_i|$ für $i > 1$ ist, gilt

$$A^k x_0 = \sum_{i=1}^n \alpha_i \cdot \lambda_i^k \cdot v_i = \alpha_1 \cdot \lambda_1^k \cdot \Big(v_1 + \underbrace{\sum_{i=2}^n \Big(\frac{\lambda_i}{\lambda_1}\Big)^k \cdot \frac{\alpha_i}{\alpha_1} \cdot v_i}_{\longrightarrow 0 \text{ für } k\to\infty}\Big),$$

so dass sich für x_k ergibt:

$$\begin{aligned} x_k &= \frac{\alpha_1 \cdot \lambda_1^k \cdot \Big(v_1 + \sum_{i=2}^n \Big(\frac{\lambda_i}{\lambda_1}\Big)^k \cdot \frac{\alpha_i}{\alpha_1} \cdot v_i\Big)}{|\alpha_1| \cdot |\lambda_1^k| \cdot \Big\|v_1 + \sum_{i=2}^n \Big(\frac{\lambda_i}{\lambda_1}\Big)^k \cdot \frac{\alpha_i}{\alpha_1} \cdot v_i\Big\|} \\ &= \operatorname{sign}(\alpha_1) \cdot \operatorname{sign}(\lambda_1) \cdot \frac{v_1 + \sum_{i=2}^n \Big(\frac{\lambda_i}{\lambda_1}\Big)^k \cdot \frac{\alpha_i}{\alpha_1} \cdot v_i}{\Big\|v_1 + \sum_{i=2}^n \Big(\frac{\lambda_i}{\lambda_1}\Big)^k \cdot \frac{\alpha_i}{\alpha_1} \cdot v_i\Big\|} \\ &\xrightarrow[k\to\infty]{} \operatorname{sign}(\alpha_1) \cdot \operatorname{sign}(\lambda_1) \cdot \frac{v_1}{\|v_1\|} = \operatorname{sign}(\alpha_1) \cdot \operatorname{sign}(\lambda_1) \cdot v_1, \end{aligned}$$

da v_1 bereits normiert war. □

Bemerkung 45.13. Der Nachteil der Vektoriteration ist, dass wir nur den größten Eigenwert bestimmen können. Eine Variante liefert auch Eigenwerte in der Mitte: Ist A symmetrisch, λ_i ein einfacher Eigenwert und ist $\tilde{\lambda} \approx \lambda_i$ eine Approximation. Dann ist $(A - \tilde{\lambda}E)$ fast singulär und $(\lambda_i - \tilde{\lambda})^{-1}$ der größte Eigenwert von $(A - \tilde{\lambda}E)^{-1}$. Für einen allgemeinen Vektor x_0 konvergiert daher die durch

$$(A - \tilde{\lambda}E)y_k = x_{k-1},\ x_k = \frac{y_k}{\|y_k\|}$$

iterative definerte Folge x_k bis auf ein Vorzeichen gegen einen Eigenvektor von A zu λ_i. Dieses Verfahren heißt **inverse Vektoriteration**.

Wir bemerkgen dazu noch, dass wir, um $(A-\tilde{\lambda}E)y_k = x_{k-1}$ zu lösen, die Matrix $(A-\tilde{\lambda}E)$ nur einmal LR– oder QR–zerlegen müssen und dass, obwohl $(A-\tilde{\lambda}E)$ fast singulär ist, die inverse Vektoriteration numerisch stabil ist.

45.4 Numerisches Lösen partieller Differentialgleichungen

Wir haben bereits erwähnt, dass viele Probleme auf (partielle) Differentialgleichungen führen. Auch diese möchte man numerisch lösen.

Beispiel 45.14. . . .

noch konkreter:

Beispiel 45.15. Millimeterpapier für $1m^2$, d.h. $10^3 \cdot 10^3 = 10^6$ Stützstellen. . . Laufzeit. . .

dünn besetzt (engl. **sparse**). . . **sparse solver**. . .

45.5 Allgemeine Iterationsverfahren

Bisher haben wir nur Iterationsverfahren für symmetrische Matrizen betrachtet, doch auch im allgemeinen Fall sind solche Verfahren einsetzbar.

. . .

Satz 45.16 (Konvergenzkriterium für Iterationsverfahren). . . .

Beispiel 45.17. 1. $Q = E$: . . .

2. Das **Jacobiverfahren**: $Q = D$, wobei $A = L + D + R$. . .

Satz 45.18 (Konvergenz des Jacobiverfahrens). *Das Jacobiverfahren konvergiert für $A = L + D + R$ für jeden Startwert x_0 gegen die Lösung von $Ax = b$, wenn die Matrix A strikt diagonaldominant ist, d.h.*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \ i = 1, \ldots, n.$$

Beispiel 45.19. . . . von oben. . .

Satz 45.20 (Gauß–Seidel–Verfahren). *Sei $A = L + D + R$ symmetrisch zerlegt wie oben. Dann konvergiert die Folge*

$$x_{k+1} = -(L + D)^{-1} \cdot R \cdot x_k + (D + L)^{-1} \cdot b$$

für jeden Startwert gegen die Lösung $Ax = b$, falls A positiv definit ist.

Beispiel 45.21. . . .

45.6 Numerischer Rang und Singulärwertzerlegung

Bis jetzt haben wir numerische Methoden beschrieben, lineare Gleichungssysteme zu lösen oder Eigenwerte zu berechnen. In vielen Problemstellungen interessiert aber nur der Rang einer Matrix oder eine Approximation eines Problems durch eine Matrix von kleinerem Rang. Solche kann die Singulärwertzerlegung liefern.

45.6.1 Einleitung

In Kapitel 28 über die Singulärwertzerlegung haben wir bereits gesehen, dass es für jede Matrix $A \in \mathbb{R}^{m \times n}$ sogenannte Singulärwerte $\sigma_1, \ldots, \sigma_p \in \mathbb{R}$ mit $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ sowie $U \in O(m)$ und $V \in O(n)$ gibt, so dass

$$U^t A V = \Sigma := \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

Die Quadrate σ_i^2 der Singulärwerte sind die Eigenwerte von $A^t A$. Außerdem ist $\operatorname{rang}(A) = \#\{\text{ Singulärwert von } A \neq 0\}$.

Wir werden sehen, dass diese Aussage über den Rang auch für die Numerik Auswirkungen hat. Den Rang kann man prinzipiell freilich auch an der Jordanschen Normalform ablesen, doch dies ist nicht numerisch stabil und die Eigenwerte alleine reichen (auch, wenn man exakt arbeitet) nicht aus, um den Rang zu berechnen:

Beispiel 45.22. Offenbar gilt

$$\operatorname{rang}\begin{pmatrix}0&0\\0&0\end{pmatrix}=0,\ \operatorname{rang}\begin{pmatrix}0&1\\0&0\end{pmatrix}=1.$$

Die beiden Eigenwerte sind in beiden Fällen jeweils $0,0$. Die Singulärwerte sind dagegen $0,0$ bzw. $0,1$, so dass sich nach der obigen Formel tatsächlich die Ränge ergeben.

Wie im zitierten Satz ist also im Beispiel tatsächlich die Anzahl der von 0 verschiedenen Singulärwerte gerade der Rang. Die Eigenwerte lassen keine solche Aussage zu. Man weiß nur, dass der Rang genau dann voll ist, wenn kein Eigenwert verschwindet.

Auch wenn die Einträge der Matrix fehlerbehaftet sind, bestätigt sich dies:

Beispiel 45.23. Wir betrachten

$$A=\begin{pmatrix}0&1\\\varepsilon&0\end{pmatrix}.$$

Da $\chi_A(t)=t^2-\varepsilon$ ist, sind die Eigenwerte $\pm\sqrt{\varepsilon}$.

Die Singulärwerte sind die Wurzeln der Eigenwerte von

$$B=A^{\mathrm{t}}A=\begin{pmatrix}\varepsilon^2&0\\0&1\end{pmatrix},$$

also $\sigma_1=1,\sigma_2=\varepsilon$.

Eine naheliegende Idee ist es nun, sehr kleine Singulärwerte als 0 anzusehen und damit einen numerisch sinnvollen Rang zu definieren.

Übrigens: Schon an diesem einfachen Beispiel sieht man, dass es beim Rechnen mit einer festen Stellenanzahl passieren kann, dass die Eigenwerte von $A^{\mathrm{t}}A$ zwar numerisch 0 sind, die Singulärwerte es aber nicht sind. Auch aus anderen Gründen ist es meist wesentlich besser, die Singulärwerte auf anderem Weg direkt zu berechnen und nicht über die Eigenwerte von $A^{\mathrm{t}}A$.

45.6.2 Berechnung der Singulärwerte

Golub und Reinsch haben 1971 einen schnellen und stabilen Algorithmus angegeben. Er ist ebenfalls ein iteratives Verfahren und ist eng mit der QR–Methode verwandt. Siehe [SB80, S. 377ff] oder [GL96, S. 452ff] für eine detaillierte Ausführung.

45.6.3 Zum größten Singulärwert

Zur Vorbereitung auf das Hauptresultat dieses Abschnittes über die Approximation von Matrizen durch solche von kleinerem Rang benötigen wir noch ein paar Hilfsmittel.

Wir haben bereits gesehen, dass für symmetrische Matrizen

$$\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \lambda_{max} = \max\{|\lambda| \mid \lambda \text{ Eigenwert von } A\}.$$

Für solche Matrizen sind die Eigenwerte gerade die Singulärwerte: $\lambda_i = \sigma_i$. Ein allgemeineres Resultat ist daher:

Satz 45.24. *Es gilt:*

$$\max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sigma_{max} = \sigma_1,$$
$$\min_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sigma_{min} = \sigma_p.$$

Beweis. Die Matrix $B = A^t A$ ist symmetrisch. Also exisitiert mit der Hauptachsentransformation $U \in \mathrm{O}(n)$, so dass

$$U^t A\, U = D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

mit $\lambda_1 \geq \cdots \geq \lambda_n$. Für $x \in \mathbb{R}^n$ ist daher:

$$\frac{x^t B\, x}{x^t x} = \frac{(x^t U)\,(U^t B\, U)\,(U^t x)}{(x^t U)\,(U^t x)} = \frac{y^t D\, y}{y^t y} = \frac{\sum_i \lambda_i y_i^2}{\sum_i y_i^2} \leq \frac{\sum_i \lambda_1 y_i^2}{\sum_i y_i^2} = \lambda_1.$$

Speziell für einen Eigenvektor x von B zu λ_1 ist $\frac{x^t B\, x}{x^t x} = \lambda_1$, also:

$$\lambda_1 = \max_{0\neq x} \frac{x^t Bx}{x^t x} = \max_{0\neq x} \frac{x^t A^t A\, x}{x^t x},$$

wie behauptet. Das Minimum ergibt sich ähnlich. □

Satz 45.25. *Sei $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, $\sigma_{max} := \sigma_1 = \max\{\sigma \mid \sigma$ Singulärwert von $A\}$. Dann gilt:*

$$|a_{ij}| \leq \sigma_{max}.$$

Beweis. Zunächst zeigt man für $I_{ij} = e_i \cdot e_j^{\mathrm{t}}$, dass $\sigma_{max}(I_{ij}) = 1$; dies führen wir hier nicht aus, es ist nicht schwierig.

Damit gilt $I_{ij}AI_{ij} = a_{ij}I_{ij}$ und

$$||a_{ij}I_{ij}|| = |a_{ij}| \cdot ||I_{ij}|| = |a_{ij}|.$$

Also:

$$|a_{ij}| = ||a_{ij}I_{ij}|| = ||I_{ij}AI_{ij}|| \leq ||A|| \cdot ||I_{ij}||^2 = ||A|| = \sigma_{max}.$$

□

45.6.4 Optimale Rang k Approximation

Wir möchten $A \in \mathbb{R}^{m\times n}$ durch eine Matrix $A_k \in \mathbb{R}^{m\times n}$ vom Rang k approximieren. Dazu betrachten wir die Singulärwertzerlegung $A = U\Sigma V^{\mathrm{t}}$ mit den Singulärwerten $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$. Wir setzen:

$$\Sigma_k = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ & & 0 & & & \end{pmatrix} \in \mathbb{R}^{m\times n}$$

und

$$A_k := U\Sigma_k V^{\mathrm{t}}.$$

Satz 45.26 (Rang k Approximation von Matrizen). *Es gilt:*

$$\min_{B:\ \mathrm{rang}(B)=k} ||A - B|| = ||A - A_k|| = \sigma_{k+1}$$

Beweis. Siehe auch [GL96, S. 73]. Zunächst ist

$$A - A_k = U(\Sigma - \Sigma_k)V^{\mathrm{t}}, \quad \Sigma - \Sigma_k = \begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{k+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_n \\ & & 0 & & & \end{pmatrix}$$

und daher: $\|A - A_k\| = \sigma_{k+1}$ nach Satz 45.24.

Wir müssen also noch sehen, dass alle anderen B mindestens diesen Abstand besitzen. Dazu schreiben wir $U(u_1, \ldots, u_m)$, $V = (v_1, \ldots, v_n)$ und damit:

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^{\mathrm{t}}, \quad A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^{\mathrm{t}}.$$

Ist nun $B \in \mathbb{R}^{m\times n}$ eine beliebige Matrix vom Rang $\operatorname{rang}(B) = k$, so ist $\dim(\ker(B)) = n-k$. Da die Spalten v_i linear unabhängig sind, ist $\dim(\operatorname{Spann}(v_1, \ldots, v_{k+1})) = k+1$, so dass ein

$$z \in \ker(B) \cap \operatorname{Spann}(v_1, \ldots, v_{k+1}) \neq \emptyset$$

existiert. Wir betrachten einen solchen Vektor z mit $\|z\| = 1$. Dieser lässt sich schreiben als

$$z = \sum_{i=1}^{k+1} \lambda_i v_i$$

für gewisse λ_i mit $\sum_{i=1}^{k+1} \lambda_i^2 = 1$.

Nach Definition von z ist $Bz = 0$ und daher:

$$Az = \Big(\sum_{i=1}^{n} \sigma_i u_i v_i^{\mathrm{t}}\Big) \cdot \Big(\sum_{j=1}^{k+1} \lambda_j v_j\Big) = \sum_{i=1}^{k+1} \sigma_i \lambda_i u_i,$$

weil ja $v_i{}^{\mathrm{t}} v_j = 1$, falls $i = j$ und 0 sonst. Es folgt:

$$\begin{aligned}\|A - B\|^2 \geq \|(A-B)z\|^2 = \|Az\|^2 = \Big\|\sum_{i=1}^{k+1} \sigma_i \lambda_i u_i\Big\|^2 &\overset{*}{=} \sum_{i=1}^{k+1} (\sigma_i \lambda_i)^2 \\ \geq \sum_{i=1}^{k+1} (\sigma_{k+1} \lambda_i)^2 = \sigma_{k+1}^2 \cdot \sum_{i=1}^{k+1} \lambda_i^2 &= \sigma_{k+1}^2.\end{aligned}$$

Hierbei gilt $(*)$, weil die u_i orthonormal zueinander stehen. □

Die Matrix A_k besitzt also unter allen Matrizen mit Rang k den kleinsten Abstand von A.

Bemerkung 45.27. Auch bezüglich der Frobeniusnorm

$$\|A\|_F = \sqrt{\sum_{ij} a_{ij}^2} \quad \Big(= \sqrt{\operatorname{tr}(A^{\mathrm{t}} A)}\Big)$$

ist A_k die beste Rang k Approximation:

$$\min_{B:\ \mathrm{rang}(B)=k} \|A-B\|_F = \|A-A_k\|_F = \|\Sigma-\Sigma_k\|_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}.$$

Dies ist auch nicht schwer zu zeigen; es findet sich bereits 1936 bei Eckart und Young sowie 1965 wieder bei Golub und Kahan.

Daher definieren wir:

Definition 45.28. *Für eine Schranke $\varepsilon > 0$ ist der **numerische Rang** $\mathrm{numrang}(A)$ einer Matrix $A \in \mathbb{R}^{m\times n}$ die Zahl*

$$\mathrm{numrang}(A) := \#\{\, i \mid \sigma_i \geq \varepsilon\}.$$

Wie bereits erwähnt, kann man nach Golub und Reinsch den numerischen Rang schnell und stabil berechnen.

Beispiel 45.29. Sei $\varepsilon > 0$ und $A = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}$. Dann gilt:

$$B = A^{\mathrm{t}} A = \begin{pmatrix} \varepsilon^2 & 0 \\ 0 & 1 \end{pmatrix},$$

so dass ε^2 und 1 die Eigenwerte von B und damit ε und 1 die Singulärwerte von A sind. Demnach ist $\mathrm{numrang}(A) = 1$, falls ε klein genug ist.

Ein etwas komplizierteres Beispiel ist folgendes:

Beispiel 45.30. Siehe Abschnitt 9.e) in
http://epub.ub.uni-muenchen.de/4400/5/tr031.pdf.

45.6.5 Anwendungen der optimalen Rang k Approximation

Statistik

Gegeben sei eine Datenmatrix $A \in \mathbb{R}^{m\times n}$, wobei m die Anzahl der Beobachtungen und n die Anzahl der Variablen sei.

Kennt man die Werte a_{ij} nur auf drei Stellen genau, so kann man sich fragen, ob eine Matrix $\tilde{A} \in \mathbb{R}^{m\times n}$ existiert mit $\|A-\tilde{A}\| < 0.001$ und $\mathrm{rang}(\tilde{A}) < \mathrm{rang}(A)$. Ein solches $\tilde{A}$ beschreibt die Situation möglicherweise wesentlich besser.

Computeralgebra und Geometrie

Auch bei algebraischen und geometrischen Berechnungen am Computer hat die Approximation durch eine Matrix von kleinerem Rang viele Anwendungen. Einige Beispiele:

- Nullstellen eines Polynoms berechnen, bei dem die Koeffizienten nur ungefähr bekannt sind,
- Nullstellen von polynomiellen Gleichungssystemen berechnen; insbesondere solche, bei denen die Koeffizienten nur ungefähr bekannt sind,
- **fast singuläre Punkte** geometrischer Objekte bestimmen (siehe dazu Abb. **??**); insbesondere solcher, die durch Gleichungen beschrieben werden, die mit Fehlern behaftet sind. Für eine ebene Kurve $f(x, y) = 0$ sind fast singuläre Punkte beispielsweise Punkte, für die gilt:

$$|f(x, y)| < \varepsilon, \ \left|\frac{\partial f}{\partial x}(x, y)\right| < \varepsilon, \ \left|\frac{\partial f}{\partial x}(x, y)\right| < \varepsilon.$$

Auch die Berechnung des numerischen Ranges hat viele Anwendungen. Beispielsweise in den ebne erwähnten Kontexten:

- Berechnung der Anzahl der Nullstellen eines Polynoms, bei dem die Koeffizienten nur ungefähr bekannt sind,
- Berechnung der Anzahl der fast singulären Punkte bestimmen.

All dies sind Anwendungen der Singulärwertzerleung, die zur aktuellen Forschung gehören. In den meisten Fällen ist es noch nicht klar, welche Herangehensweise an ein Problem sich letztendlich durchsetzen wird. Singulärwerte sind hier nur eine Möglichkeit.

Aufgabe 45.1 (. . .). . . .

Literatur

COS+98. Calabi, E. ; Olver, P.J. ; Shakiban, C. ; Tannenbaum, A. ; Haker, S.: Differential and Numerically Invariant Signature Curves Applied to Object Recognition. In: *Int. J. Comp. Vision* 26 (1998), Nr. 2, S. 107–135

DS05. Durdel, A. ; Stoer, J.: *Stoer/Burlisch: Numerische Mathematik 2*. 5. Aufl. Springer, 2005

FH07. Freund, R. ; Hoppe, R.: *Stoer/Burlisch: Numerische Mathematik 1*. 10. Aufl. Springer, 2007

Fis01. Fischer, G.: *Analytische Geometrie*. 7. Vieweg, 2001

Fis08. Fischer, G.: *Lineare Algebra*. 15. Vieweg, 2008

For08a. Forster, O.: *Analysis 1*. 9. Vieweg, 2008

For08b. Forster, O.: *Analysis 2*. 8. Vieweg, 2008

For08c. Forster, O.: *Analysis 3*. 5. Vieweg, 2008

GL96. Golub ; van Loan: *Matrix Computations*. 3. Johns Hopkins University Press, 1996

HCV32. Hilbert, D. ; Cohn-Vossen, S.: *Anschauliche Geometrie*. Berlin : Verlag von Julius Springer, 1932

HLM05. Holzer, S. ; Labs, O. ; Morris, R. surfex – *Visualization of Real Algebraic Surfaces*. www.surfex.AlgebraicSurface.net. 2005

Knu99. Knuth, D.: *The Art of Computer Programming 2*. 3. Adison Wesley, 1999

Kö02. Königsberger, K.: *Analysis 2*. 4. Springer, 2002

Kre02. Krengel, U.: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. 6. Vieweg, 2002

Pena. Penrose, R.: *A Complete Guide to the Laws of the Universe*. ????, ????

Penb. Penrose, R.: *The Emperor's New Mind. Concerning Computers, Minds, and the Laws of Physics*. ????, ????

PS05. Pachter, L. (Hrsg.) ; Sturmfels, B. (Hrsg.): *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005

SB80. Stoer, J. ; Burlisch: *Introduction to Numerical Analysis*. Springer, 1980

Symbolverzeichnis

2^M	Potenzmenge, 13
$f : M \to N$	Abbildung von M nach N, 29
$x + y$	Addition von Vektoren, 208
$a \sim b$	a ist aquivalent zu b bzgl. einer Aquivalenzrelation, 38
$\iff$	Aquivalenz, 8
a^x	Exponentiation zu einer beliebigen Basis, 156
A_{ij}	Eine gewisse Unterdeterminante von A., 311
A'_{ij}	Eine gewisse Unterdeterminante von A., 312
$A < 0$	Negative Definitheit einer Matrix., 445
$A \leq 0$	Negative Semi–Definitheit einer Matrix., 445
$\|M\|$	Anzahl der Elemente einer Menge M, 13
$A > 0$	Positive Definitheit einer Matrix., 373
arcsin	Arcussinus, Umkehrfunktion von Sinus sin, 161
arctan	Arcustangens, Umkehrfunktion von Tangens tan, 160
Bild A	Bild einer Matrix A., 309
Bild f	Bild eines Homomorphismus., 253
$\binom{r}{k}$	Binomialkoeffizient mit reeelem Eintrag, 508
$\mathbb{C}$	komplexe Zahlen, 110
$\mathbb{C}[z]$	Menge aller Polynome in z mit Koeffizienten in $\mathbb{C}$, 114
χ_A	charakteristische Funktion, 29
$\chi_A(t)$	Charakteristisches Polynom einer Matrix A, 327
$\chi_f(t)$	Charakteristisches Polynom einer linearen Abbildung f, 329
cos	Cosinus, 109
cot	Cotangens cot = cos / sin, 160
$\deg(p)$	Grad des Polynoms p, 114
$\det A$	Determinante einer Matrix A., 298

$d(H,q)$	Abstand eines Punktes von einer Hyperebene., 216
Δf	Differenzfunktion, 15
$\dot\cup$	Disjunkte Vereinigung, 13
$d(L_1,L_2)$	Abstand zweier Geraden., 217
$d(L,q)$	Abstand von Punkt zu Gerade., 215
$\cap$	Durchschnitt zweier Mengen, 22
$\bigcap_{i\in I}$	Durchnitt aller Mengen einer Familie, 34
$d(x,y)$	Abstand zweier Punkte, 209
e	Eulersche Zahl, 119
$\mathrm{Eig}(A,\lambda)$	Eigenraum von A zum Eigenwert λ., 327
$\varphi\big\vert_{\mathbb{Q}}$	Einschrankung von φ auf $\mathbb{Q}$, 82
$a \in M$	a ist ein Element der Menge M, 22
$a \notin M$	a ist kein Element der Menge M, 22
$M \ni a$	M enthalt das Element a, 22
$E(X)$	Erwartungswert, 499, 500
exp	Exponentialfunktion, 109
exp	komplexe Exponentialfunktion, 118
e^z	komplexe Exponentialfunktion, 119
$(A_i)_{i\in I}$	Familie von Mengen, indiziert durch I, 34
$\frac{f}{g}$	rationale Funktion, 128
f_n	n–te Fibonacci–Zahl, 17
$f\colon D \to \mathbb{R}$	reellwertige Funktion, 125
f^{-1}	Umkehrfunktion von f, 131
$(f^{-1})'$	Ableitung der Umkehrfunktion von f, 139
$f^{-1}(B)$	Urbild von B unter f, 30
(a_n)	eine Folge mit Gliedern $a_1, a_2, \ldots$, 71
$(a_n)_{n\in\mathbb{N}}$	eine Folge mit Gliedern $a_1, a_2, \ldots$, 71
$\hat f$	Fouriertransformierte von f, 527
$f'(x)$	Ableitung der Funktion f, 135
F_X	Verteilungsfunktion von X, 498
$\Gamma(x)$	Die Gamma–Funktion., 545
$\lfloor x \rfloor$	entier, ganzzahliger Anteil von x, 29
(a,b)	geordnetes Paar zweier Elemente, 26
G_f	Graph der Funktion f, 125
G_f	Graph der Funktion f, 29
$\langle g \rangle$	Gruppe, die von einem Element g erzeugt wird., 288
$G_X(z)$	Erzeugende Funktion der Folge $(P(X=k))_{k\in\mathbb{N}}$, 513
$\mathrm{Hess}(f)$	Hesse-Matrix der Funktion f., 438

$\neg A$	logische Negation, 8
$A \vee B$	logisches oder, 7
$A \wedge B$	logisches und, 7
$\max_{x\in[a,b]} f(x)$	Maximum einer stetigen Funktion, 130
$\{\dots\}$	Menge, spezifiziert durch Aufzahlen der Elemente, 21
$A \backslash B$	Differenzmenge von A und B, 24
$A - B$	Differenzmenge von A und B, 24
$\{\dots \mid \dots\}$	Die Menge der Elemente $\dots$ mit der Eigenschaft $\dots$, 21
$\bar{A}$	Komplement einer Menge, 22
N^M	Menge aller Abbildungen von der Menge M in die Menge N, 32
$\min_{x\in[a,b]} f(x)$	Minimum einer stetigen Funktion, 130
$a \mod d$	a modulo d, der Rest der Division von a durch d, 38
$m(P, \lambda)$	Vielfachheit von λ als Nullstelle von P, 331
$\lambda \cdot x$	Multiplikation eines Vektors x mit einem Skalar λ, 208
$M_X(\theta)$	Momenterzeugende Funktion von X, 526
$\mathbb{N}$	Menge der naturlichen Zahlen, 11
$N_c(f)$	Niveaumengen der Funktion f zum Niveau c., 433
$N(f)$	Nullstellenmenge (oder Hyperflache) einer Funktion., 453
$n!$	n Fakultat, 14
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Erwartungswert μ und Standardabweichung σ., 495
$n \to n + 1$	Induktionsschritt, 12
$\|\|.\|\|$	Allgemeine Definition einer Norm auf einem VR., 376
$\sqrt[n]{}$	n–te Wurzel, 101
$\binom{n}{k}$	Binomialkoeffizient, n uber k, 26
$O(.)$	O–Notation fur Folgen, 76
$o(.)$	o–Notation fur Folgen, 77
$f \in O(g)$	f liegt in gros O von g, 166
$f \in o(g)$	f liegt in klein o von g, 166
$\mathrm{O}(n)$	Orthogonale Gruppe, 276
$\mathrm{ord}(g)$	Ordnung eines Elementes g, 288
$\mathrm{ord}(G)$	Ordnung der Gruppe G, 288
$U^\perp$	Zu U orthogonaler Untervektorraum., 382
$v^\perp$	Zu $v \in V$ orthogonaler Untervektorraum von V., 343
$P(A \mid B)$	Bedingte Wahrscheinlichkeit, 496
$P_f(t_0, \dots, t_r)$	Lange des Polygonzugs zu einer Unterteilung, 420
π	Kreiszahl π, 159
$\mathcal{P}(M)$	Potenzmenge, 13
$\prod_{k=1}^{n}$	endliches Produkt, 14

$\bigcup_{i \in I}$	Vereinigung aller Mengen einer Familie, 34
$g \circ f$	g verknupft mit f, g nach f, 33
$\mathrm{Vol}(K)$	Volumen eines Kompaktums, 480
$V(X)$	Varianz der Zufallsvariable X, 501
$\sqrt{b}$	Quadratwurzel aus einer positiven reellen Zahl., 83
$\bar{x}$	Stichprobenmittel, 541
$\mathbb{Z}$	Menge der ganzen Zahlen, 15
$\zeta(2)$	Die Riemannsche Zeta–Funktion fur $n = 2$, 396
$\zeta(n)$	Die Riemannsche Zeta–Funktion fur naturliche Zahlen, 396
$\zeta(s)$	Riemannsche Zetafunktion, 188
$\bar{z}$	konjugiert komplexe Zahl zu z, 111
$(i_1\ i_2\ \ldots\ i_k)$	Zykelschreibweise fur eine Permutation, 279

Sachverzeichnis