HÖHERE MATHEMATIK I+II

Ingo Steinwart

2022-2023

Universität Stuttgart
Fachbereich Mathematik
Institut für Stochastik und Anwendungen

Chapter 1: Grundlagen

Section 1.1
Logik

AUSSAGEN

Die Logik befasst sich mit **Aussagen**, d.h. mit (im weitesten Sinne sprachlichen) Sätzen, die objektiv entweder wahr oder falsch sind. Insbesondere gelten daher folgende Prinzipien:

- **Prinzip vom ausgeschlossenen Dritten ("Tertium non datur"):** Eine Aussage muss wahr oder falsch sein.
- **Prinzip vom ausgeschlossenen Widerspruch:** Eine Aussage darf nicht gleichzeitig wahr und falsch sein.

Im folgenden kürzen wir "wahr" häufig mit "*w*" und "falsch" mit "*f*" ab.

BEISPIELE

Der folgende Satz ist eine Aussage, die wahr ist:

Das Wintersemester 2022/23 an der Universität Stuttgart beginnt im Oktober.

Der folgende Satz ist eine Aussage, die falsch ist:

2 + 2 = 5

Der folgende Satz ist keine Aussage, da sein Wahrheitsgehalt subjektiv ist:

Grün ist schöner als Rot.

Der folgende Satz ist keine Aussage, da kein Wahrheitsgehalt zugewiesen werden kann:

Dieser Satz ist falsch.

Bool'sche Operatoren

Aussagen können miteinander kombiniert werden. Sind beispielsweise p und q zwei Aussagen, so ist die

Negation (Verneinung) von p, schreibe $\neg p$ und sage "nicht p", durch die folgende Wahrheitstabelle gegeben:

p	$\neg p$
w	f
f	w

Konjunktion ("und") von p und q, schreibe $p \wedge q$ und sage "p und q", durch die folgende Wahrheitstabelle gegeben:

p	q	$p \wedge q$
w	w	w
w	f	f
f	w	f
f	f	f

Mit anderen Worten: $p \wedge q$ ist genau dann wahr, wenn sowohl p als auch q wahr sind.

Bool'sche Operatoren

Disjunktion ("oder") von p und q, schreibe $p \vee q$ und sage "p oder q", durch die folgende Wahrheitstabelle gegeben:

p	q	$p \vee q$
w	w	w
w	f	w
f	w	w
f	f	f

Beachte: "oder" ist nicht exklusiv gemeint, d.h. wenn sowohl p als auch q wahr sind, so ist auch "p oder q" wahr.

Bool'sche Operatoren

Implikation von p und q, schreibe $p \Rightarrow q$ und sage "Aus p folgt q.", durch die folgende Wahrheitstabelle gegeben:

p	q	$p \Rightarrow q$
w	w	w
w	f	f
f	w	w
f	f	w

Mit anderen Worten: ist $p \Rightarrow q$ wahr, so können wir:

- aus der Wahrheit von p auf die Wahrheit von q schliessen.
- aus der Wahrheit von q **nicht** auf die Wahrheit von p schliessen.
- aus der Falschheit von q auf die Falschheit von p schliessen.

Wir sagen, dass p **hinreichend für** q **ist** und dass q **notwendig für** p **ist**.

Bool'sche Operatoren

Äquivalenz von p und q, schreibe $p \Leftrightarrow q$ und sage "p ist zu q äquivalent.", durch die folgende Wahrheitstabelle gegeben:

p	q	$p \Leftrightarrow q$
w	w	w
w	f	f
f	w	f
f	f	w

Mit anderen Worten, $p \Leftrightarrow q$ ist genau dann wahr, wenn p und q die gleichen Wahrheitswerte haben.

TAUTOLOGIEN

Eine Verknüpfung von Aussagen, die unabhängig von den Wahrheitswerten der einzelnen Aussagen immer wahr ist, heißt **Tautologie**.

Sind p und q Aussagen, so ist zum Beispiel

$$(p \Rightarrow q) \quad \Leftrightarrow \quad (\neg q \Rightarrow \neg p)$$

eine Tautologie, wie die folgende Wahrheitstabelle zeigt:

p	q	$p \Rightarrow q$	$\neg p$	$\neg q$	$\neg q \Rightarrow \neg p$
w	w	w	f	f	w
w	f	f	f	w	f
f	w	w	w	f	w
f	f	w	w	w	w

Tautologien können also “logischen Rechenregeln” entsprechen. In solchen Fällen schreiben wir daher häufig auch “$\equiv$” statt “$\Leftrightarrow$”. Die obige Tautologie kann deshalb auch als

$$(p \Rightarrow q) \equiv (\neg q \Rightarrow \neg p)$$

geschrieben werden.

EINIGE WICHTIGE TAUTOLOGIEN

Der folgende Satz fasst die wichtigsten Tautologien, die aus einer Aussage gebildet werden können, zusammen.

Theorem 1.1.1

Ist p eine Aussage, so sind die folgenden Aussagen Tautologien:

i). ***Satz vom ausgeschlossenen Dritten****: $p \vee \neg p$.*

ii). ***Satz vom Widerspruch****: $\neg(p \wedge \neg p)$.*

iii). ***Satz von der doppelten Verneinung****: $\neg(\neg p) = p$.*

iv). ***Idempotenz****:*

$$p \wedge p \equiv p\,,$$
$$p \vee p \equiv p\,.$$

EINIGE WICHTIGE TAUTOLOGIEN

Der folgende Satz fasst die wichtigsten Tautologien, die aus mehreren Aussagen gebildet werden, zusammen.

Theorem 1.1.2

Sind p, q und r Aussagen, so sind die folgenden Aussagen Tautologien:

i). ***Rechenregeln von de Morgan****:*

$$\neg(p \wedge q) \equiv \neg p \vee \neg q\,,$$
$$\neg(p \vee q) \equiv \neg p \wedge \neg q\,.$$

ii). ***Kontraposition****:* $(p \Rightarrow q) \equiv (\neg q \Rightarrow \neg p)$.

iii). ***Zerlegung der Äquivalenz****:* $p \Leftrightarrow q \equiv (p \Rightarrow q) \wedge (q \Rightarrow p)$.

iv). ***Transitivität****:* $(p \Rightarrow q) \wedge (q \Rightarrow r) \Rightarrow (p \Rightarrow r)$.

v). ***Distributivgesetze****:*

$$p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r)\,,$$
$$p \vee (q \wedge r) \equiv (p \vee q) \wedge (p \vee r)\,.$$

BEMERKUNGEN

Neben den oben genannten Tautologien gibt es noch weitere Tautologien, wie z.B. die Kummutativitäten $(p \wedge q) \equiv (q \wedge p)$ und $(p \vee q) \equiv (q \vee p)$.

Es gibt genau 16 zweistellige logische Operationen, d.h. Verknüpfungen, die aus 2 Aussagen p und q eine neue Aussage $f(p, q)$ generieren. Jede dieser 16 Operationen und die Verneinung lässt sich durch die Operation $\neg(p \wedge q)$, die wir mit $p \,\mathrm{NAND}\, q$ bezeichnen, darstellen. So gilt zum Beispiel:

$$\begin{aligned} \neg p &\equiv (p \,\mathrm{NAND}\, p)\,, \\ p \wedge q &\equiv (p \,\mathrm{NAND}\, q) \,\mathrm{NAND}\, (q \,\mathrm{NAND}\, p)\,. \end{aligned}$$

Analog lässt sich jede dieser 17 Operationen auch durch die Operation $\neg(p \vee q)$, die wir mit $p \,\mathrm{NOR}\, q$ bezeichnen, darstellen.

Schreiben wir “1” statt “*w*” und “0” statt “*f*”, so entspricht $p \wedge q$ der Multiplikation. Analog entspricht das “exklusive oder”, d.h. $p \,\mathrm{XOR}\, q :\equiv \neg(p \Leftrightarrow q)$, der Addition mit Überlauf.

AUSSAGEN MIT VARIABLEN

Wenn Aussagen Variablen enthalten, sprechen wir von **Aussageformen**.
Einfache Beispiele von Aussageformen sind

$$p_1(x) := \text{“}x \text{ ist 20 Jahre alt.”}\,,$$
$$p_2(x) := \text{“}x < 2\text{”}\,.$$

Durch Einsetzen “erlaubter” Werte für die Variable erhalten wir Aussagen:

$$p_1(\text{“Ingo Steinwart”}) = f\,,$$
$$p_2(1) = w\,,$$
$$p_2(2) = f\,.$$

$p_1(1)$ macht hingegen keinen Sinn, da die Zahl 1 kein Alter hat.

QUANTOREN

Aus einer Aussageform $p(x)$ können mithilfe von **Quantoren** ebenfalls Aussagen gebildet werden. Die folgende Quantoren interessieren uns:

- **All-Quantor:**

  $$\forall x : p(x)$$

  steht für die Aussage, dass für alle erlaubten Werte von x die Aussage $p(x)$ wahr ist.
- **Existenz-Quantor:**

  $$\exists x : p(x)$$

  steht für die Aussage, dass es mindestens einen erlaubten Wert von x gibt, so dass die Aussage $p(x)$ wahr ist.
- **Einzigkeits-Quantor:**

  $$\exists! x : p(x)$$

  steht für die Aussage, dass es genau einen erlaubten Wert von x gibt, so dass die Aussage $p(x)$ wahr ist.

Mit den obigen p_1 und p_2 ist beispielsweise die Aussage $\forall x : p_1(x)$ falsch, während die Aussage $\exists x : p_2(x)$ wahr ist.

VERGLEICH VON AUSSAGEN

Mit Quantoren gebildete Aussagen können wieder verglichen werden. So gelten zum Beispiel die de-Morgan'schen-Regeln

$$\neg(\exists x : p(x)) \equiv \forall x : \neg p(x)\,,$$
$$\neg(\forall x : p(x)) \equiv \exists x : \neg p(x)\,.$$

Aussageformen können von mehreren Variablen abhängen. In diesem Fall kann die Reihenfolge von Quantoren eine wesentliche Rolle spielen. Beispielsweise sind die Aussagen

$$\forall x\, \exists y : p(x, y) \qquad \text{und} \qquad \exists y\, \forall x : p(x, y)$$

im Allgemeinen nicht äquivalent, da im ersten Fall der Wert von y von x abhängen kann, während er im zweiten Fall von x unabhängig sein muss.

BEMERKUNGEN ZUM BEWEIS VON AUSSAGEN

Wollen wir zeigen, dass die Aussage $\forall x : p(x)$ wahr ist, können wir z.B. die Aussage $p(x)$ für jeden einzelnen Wert von x überprüfen. Dies funktioniert aber eigentlich nur, falls es nur endlich viele (oder genauer: wenige) erlaubte Werte von x gibt.

Steht x beispielsweise für alle Mannschaften der 1. Fußball-Bundesliga, so lässt sich durch Überprüfen aller 18 Teams zeigen, dass die Aussage

$$\forall x : \text{“}x \text{ hat mindestens einen Torwart”}$$

wahr ist. Steht x jedoch für alle Zahlen, so ist die Aussage

$$\forall x : x + 1 > x$$

nicht mehr durch Betrachtung jeder einzelnen Zahl als wahr zu identifizieren. Stattdessen ist ein mathematischer Beweis notwendig, der beispielsweise schon bekannte, wahre Aussagen ausnutzt. Wissen wir z.B. dass die beiden Aussagen $1 > 0$ und $\forall x, y, z : x > y \Rightarrow z + x > z + y$ wahr sind, so folgt die obige Aussage durch Betrachtung von $x = 1$ und $y = 0$. In den meisten Fällen ist ein Beweis aber deutlich länger und damit komplizierter.

BEMERKUNGEN ZUM BEWEIS VON AUSSAGEN

Wollen wir zeigen, dass die Aussage $\exists x : p(x)$ wahr ist, müssen wir ein x finden. Wie für den All-Quantor kann dies z.B. dadurch geschehen, dass man hintereinander die Aussage $p(x)$ für jeden einzelnen Wert von x überprüft und stoppt, sowie man einen Wert gefunden hat, für den $p(x)$ wahr ist.

Steht x wieder für alle Mannschaften der 1. Fußball-Bundesliga, so ist schon für $x =$ “Werder Bremen” die Aussage

$$\exists x : \text{“}x \text{ hat mindestens einen Torwart”}$$

wahr und das Überprüfen der restlichen 17 Teams erübrigt sich.

Im Falle von unendlich vielen Werten kann manchmal ebenfalls ein geeigneter Wert von x “erraten” werden, meistens ist aber ebenfalls ein Beweis notwendig.

Schließlich lassen sich mithilfe der de-Morgan’schen-Regeln Aussagen mit All-Quantoren in solche mit Existenz-Quantoren (und umgekehrt) umwandeln. Ein etwas längeres, aber nicht untypisches Beispiel hierzu ist

$$\neg\big(\forall x\, \exists y\, \forall z : p(x, y, z)\big) \quad \equiv \quad \exists x\, \forall y\, \exists z : \neg p(x, y, z)\,.$$

Section 1.2
Mengen

MENGEN

Eine **Menge** ist eine Zusammenfassung von Objekten. Diese werden **Elemente** genannt. Jedes Element kann nur einmal in der Menge vorkommen. Es gibt keine Reihenfolge für die Elemente einer Menge.

Mengen werden in der Regel mit großen Buchstaben bezeichnet, Elemente häufig mit kleinen. Ist M eine Menge und x ein Objekt, so schreiben wir:

$x \in M$	falls x ein Element von M ist.
$x \notin M$	falls x kein Element von M ist.

Ferner bezeichnet $\emptyset$ die **leere Menge**, d.h. die Menge die kein Element enthält.

KONSTRUKTION VON MENGEN

Mengen mit wenigen, explizit bekannten Elementen können durch Auflistung der Elemente beschrieben werden, z.B.

$$M := \{1, 2, 3, 4\}.$$

Ist $p(x)$ eine Aussageform mit Variable x so ist

$$M := \{x : p(x)\} := \{x \mid p(x)\}$$

die Menge aller erlaubten Werte von x, für die Aussage wahr ist[1]. Damit lassen sich auch unendliche Mengen einfach beschreiben, z.B

$$M := \{n : n \text{ ganze Zahl und } n \geq 10\}.$$

[1] Hier muss man allerdings etwas aufpassen, wie in Abschnitt 28 noch gezeigt wird.

INKLUSION UND IDENTITÄT

Seien A und B zwei Mengen, so ist A eine **Teilmenge** von B, genau dann wenn jedes Element von A auch ein Element von B ist, d.h.

$$\forall x : (x \in A \Rightarrow x \in B) .$$

In diesem Fall schreiben wir $A \subset B$.

Ferner sind die **Mengen gleich**, geschrieben $A = B$, genau dann wenn $A \subset B$ und $B \subset A$ gilt, d.h., genau dann wenn

$$\forall x : (x \in A \Leftrightarrow x \in B) .$$

Beispielsweise gilt $\{21, 22, 23\} \subset \{n : n \text{ ganze Zahl und } n \geq 10\}$ und

$$\{1, 2, 3, 4\} = \{4, 3, 2, 1\} = \{1, 2, 3, 4, 4\} .$$

Ferner gilt sowohl $\emptyset \subset A$ als auch $A \subset A$ für jede Menge A. Sind schließlich A, B und C Mengen mit $A \subset B$ und $B \subset C$, so gilt auch $A \subset C$.

OPERATIONEN MIT MENGEN

Sind A und B zwei Mengen, so können wir die folgenden Mengen definieren:

$$\begin{aligned} \textbf{Durchschnitt:} \quad & A \cap B := \{x : x \in A \wedge x \in B\}, \\ \textbf{Vereinigung:} \quad & A \cup B := \{x : x \in A \vee x \in B\}, \\ \textbf{Differenz:} \quad & A \setminus B := \{x : x \in A \wedge x \notin B\}. \end{aligned}$$

Ferner sagen wir, dass A und B **disjunkt** sind, falls $A \cap B = \emptyset$. Gibt es zudem eine feste, aus dem Zusammenhang bekannte Menge M mit $A \subset M$, so schreiben wir auch

$$A^c := M \setminus A.$$

Es gilt beispielsweise

$$\begin{aligned} \{1,2\} \cap \{2,3,4\} &= \{2\}, \\ \{1,2\} \cup \{2,3,4\} &= \{1,2,3,4\}, \\ \{1,2\} \setminus \{2,3,4\} &= \{1\}, \\ \{2,3,4\} \setminus \{1,2\} &= \{3,4\} \end{aligned}$$

und die Mengen $\{1,2\}$ und $\{3,4\}$ sind disjunkt.

Lemma 1.2.1

Sind A und B Mengen, so gelten die folgenden Identitäten:

$$A \cap B = B \cap A\,,$$
$$A \cup B = B \cup A\,,$$
$$A \cap A = A\,,$$
$$A \cup A = A\,,$$
$$A \cap \emptyset = \emptyset\,,$$
$$A \cup \emptyset = A\,.$$

Ist C eine weitere Menge, so gilt zudem

$$A \cap (B \cap C) = (A \cap B) \cap C\,,$$
$$A \cup (B \cup C) = (A \cup B) \cup C\,,$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)\,,$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)\,.$$

Lemma 1.2.2

Sind A, B und C Mengen, so gelten die folgenden Identitäten:

$$
\begin{aligned}
A \setminus A &= \emptyset\,, \\
A \setminus (B \cap C) &= (A \setminus B) \cup (A \setminus C)\,, \\
A \setminus (B \cup C) &= (A \setminus B) \cap (A \setminus C)\,, \\
(A \setminus B) \setminus C &= A \setminus (B \cup C)\,, \\
A \setminus (B \setminus C) &= (A \setminus B) \cup (A \cap C)\,,
\end{aligned}
$$

und damit insbesondere auch $(A \cap B)^c = A^c \cup B^c$ *und* $(A \cup B)^c = A^c \cap B^c$.

KARTESISCHES PRODUKT

Sind x und y zwei Objekte, bezeichnet (x, y) das geordnete Paar von x und y. Insbesondere ist (x, x) ein Paar und es gilt $(x, y) = (y, x)$ genau dann wenn $x = y$.

Sind A und B zwei Mengen, so ist das kartesisches Produkt von A und B

$$A \times B := \{(x, y) : x \in A \wedge y \in B\}.$$

So ist beispielsweise $(1, 2) \neq (2, 1)$ und

$$\{1, 2\} \times \{2, 3\} = \{(1, 2), (1, 3), (2, 2), (2, 3)\}.$$

An dieser Stelle sei bemerkt, dass geordnete Paare auch formal durch

$$(x, y) := \{\{x\}, \{x, y\}\}$$

einführt werden können. Uns wird aber ein informelles Verständnis ausreichen.

OPERATIONEN MIT MEHR ALS ZWEI MENGEN

Ist $I \neq \emptyset$ eine Menge und haben wir für jedes $i \in I$ eine Menge A_i, so definieren wir

$$\bigcap_{i \in I} A_i := \left\{ x \,\middle|\, \forall i \in I : x \in A_i \right\},$$
$$\bigcup_{i \in I} A_i := \left\{ x \,\middle|\, \exists i \in I : x \in A_i \right\}.$$

Sei ferner $I = \{1, 2, \ldots, n\}$ und $x_i \in A_i$ für alle $i \in I$, so bezeichnen wir die geordnete Folge mit Wiederholungen $(x_1, x_2, \ldots, x_n)$ als n-Tupel. Für $n = 3$ sprechen wir auch von Tripeln. Insbesondere ist ein 2-Tupel ein geordnetes Paar und Gleichheit von zwei n-Tupeln ist durch

$$(x_1, x_2, \ldots, x_n) = (y_1, y_2, \ldots, y_n) \qquad :\Leftrightarrow \qquad \forall i \in I : x_i = y_i$$

definiert. Damit definieren wir das kartesisches Produkt von A_1, A_2, ..., A_n durch

$$A_1 \times A_2 \times \cdots \times A_n := \left\{ (x_1, x_2, \ldots, x_n) \,\middle|\, \forall i \in I : x_i \in A_i \right\}.$$

Für $A := A_1 = A_2 = \cdots = A_n$ schreiben wir auch $A^n = A_1 \times A_2 \times \cdots \times A_n$.

POTENZMENGEN

Ist A eine Menge, so ist die Potenzmenge $\mathcal{P}(A)$ von A die Menge aller Teilmengen von A, d.h.

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

Hat A genau n Elemente, so hat $\mathcal{P}(A)$ genau 2^n Elemente.

Für $A := \{1, 2, 3\}$ gilt beispielsweise

$$\mathcal{P}(A) = \big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\big\}.$$

Aussageformen, die nicht zu Mengen führen

Bei der Bildung von Mengen mit Aussageformen muss man (etwas) vorsichtig sein. Für die Aussageform

$$p(x) := (x \notin x)$$

führt z.B. die zugehörige Bildung der Menge

$$M := \{x : p(x)\} = \{x : x \notin x\}$$

bei der Frage nach $M \in M$ zu einem Widerspruch, denn es gilt

$$\begin{aligned} M \in M \quad &\Rightarrow \quad M \notin M\,, \\ M \notin M \quad &\Rightarrow \quad M \in M\,. \end{aligned}$$

Damit kann $M \in M$ weder richtig noch falsch sein, d.h. die Aussage $M \in M$ hat keinen Wahrheitswert! Das Problem liegt in der "Selbstreferenzierung", die vermieden werden kann, wenn Mengen nur innerhalb einer bekannten Obermenge, wie den ganzen Zahlen, den reellen Zahlen, etc. gebildet werden. Dies führt zu einer Beschränkung der erlaubten Werte von Variablen in Aussageformen.

EINIGE WICHTIGE MENGEN

Aus der Schule sollten schon einige wichtige Mengen bekannt sein, wie z.B.

$\mathbb{N} := \{1, 2, \dots\}$	Menge der natürlichen Zahlen
$\mathbb{N}_0 := \{0, 1, 2 \dots\}$	Menge der natürlichen Zahlen inklusive 0
$\mathbb{Z} := \{0, \pm 1, \pm 2, \dots\}$	Menge der ganzen Zahlen
$\mathbb{Q} := \{m/n : m \in \mathbb{Z} \wedge n \in \mathbb{N}\}$	Menge der rationalen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$[a, b] := \{x \in \mathbb{R} : a \leq x \leq b\}$	abgeschlossenes Intervall zwischen a und b
$(a, b) :=]a, b[:= \{x \in \mathbb{R} : a < x < b\}$	offenes Intervall zwischen a und b
$[a, b) := [a, b[:= \{x \in \mathbb{R} : a \leq x < b\}$	halboffenes Intervall zwischen a und b
$\dots$	

Diese werden aber noch ausführlicher betrachtet werden.

Section 1.3
Relationen und Abbildungen

Definition 1.3.1

Seien A und B Mengen. Dann ist eine **Relation** R zwischen A und B eine Teilmenge $R \subset A \times B$.

Ist R eine Relation zwischen A und B so wird $(x, y) \in R$ so interpretiert, dass x und y in Relation zueinander stehen. Dies wird auch häufig durch

$$xRy :\Leftrightarrow (x, y) \in R$$

beschrieben. Für spezielle Relationen wird R dabei durch ein anderes, eingängigeres Symbol ersetzt.

Ist beispielsweise $A = B$, so ist $R := \{(x, y) \in A \times B : x = y\}$ eine Relation und wir schreiben $x = y$ statt xRy. Für $A = B = [0, 1]$ ist auch $R := \{(x, y) \in A^2 : x \leq y\}$ eine Relation und wir schreiben $x \leq y$ statt xRy. Illustrationen dieser beiden Relationen finden sich in Abbildung 1.

Diese Beispiele verdeutlichen schon, dass unsere Definition einer Relation nur beschreibt, was sie "formal" ausmacht, nicht aber was sie "inhaltlich" bedeutet. Im allgemeinen ist nur ein sehr kleiner Anteil aller Relationen zwischen zwei Mengen A und B wirklich "interessant".

EIN WEITERES BEISPIEL

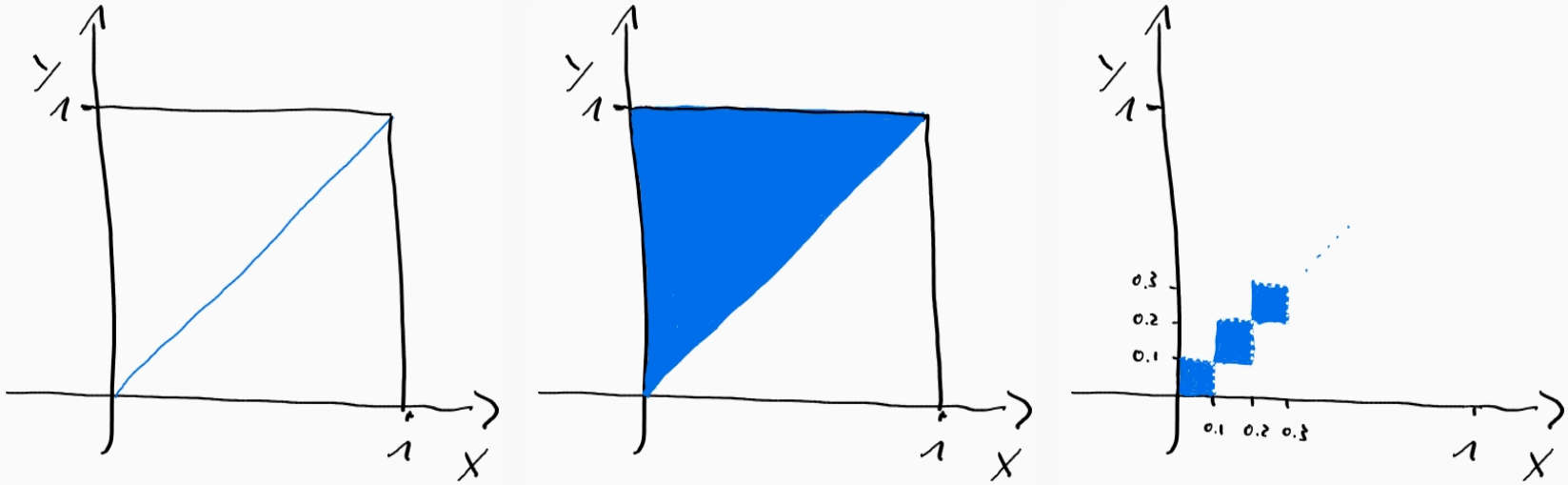


Abbildung: Illustrationen für drei Relationen (jeweils in blau) auf $[0,1]^2$. **Links:** Die Relation "$=$". **Mitte:** Die Relation "$\leq$". **Rechts:** Die Relation R aus (1.3.1).

Für $t \in \mathbb{R}$ sei $\lfloor t \rfloor$ die Abrundung von t auf die nächste kleinere ganze Zahl, d.h.

$$\lfloor t \rfloor = \max\{k \in \mathbb{Z} : k \leq t\}\,.$$

Für $A = B = [0,1]$ ist dann $R := \{(x,y) \in A \times B : \lfloor 10x \rfloor = \lfloor 10y \rfloor\}$ eine Relation und für $x, y \in [0,1]$ gilt

$$xRy \Leftrightarrow \text{1. Nachkommastelle ist gleich} \tag{1.3.1}$$

ÄQUIVALENZRELATIONEN

Das letzte Beispiel gehört zu einer besondere Klasse von Relationen, die wir in der folgenden Definition einführen.

Definition 1.3.2

Sein $A \neq \emptyset$ eine Menge und $R \subset A \times A$ eine Relation. Dann heißt R **Äquivalenzrelation** auf A, genau dann wenn die folgenden drei Eigenschaften erfüllt sind:

i). **Reflexivität:** Für alle $x \in A$ gilt $(x, x) \in R$.

ii). **Symmetrie:** Für alle $x, y \in A$ gilt $(x, y) \in R \Leftrightarrow (y, x) \in R$.

iii). **Transitivität:** Für alle $x, y, z \in A$ gilt

$$\big((x, y) \in R \wedge (y, z) \in R\big) \Rightarrow (x, z) \in R\,.$$

Im folgenden werden Äquivalenzrelationen auch mit $\sim$ oder $\equiv$ bezeichnet. Ist $\sim$ eine Äquivalenzrelation auf A und $x \in A$, dann heißt

$$[x]_\sim := \{y \in A : x \sim y\}$$

die **Äquivalenzklasse** von x und x ein **Repräsentant** dieser Äquivalenzklasse.

Die Relation "$=$" ist neben der in (1.3.1) definierten Relation eine Äquivalenzrelation. Bei der letzteren sieht man auch deutlich, dass z.B. die Elemente in $[0]_\sim = [0, 1/10)$ zueinander äquivalent sind, während keines der Elemente in $[0]_\sim = [0, 1/10)$ äquivalent zu einem Element in $[1/10]_\sim = [1/10, 2/10)$ ist.

Das folgende Resultat zeigt, dass diese Beobachtung nicht zufällig ist.

Lemma 1.3.3

Sei $A \neq \emptyset$ eine Menge, $\sim$ eine Äquivalenzrelation auf A. Dann gilt für alle $x, y \in A$ entweder

$$[x]_\sim = [y]_\sim$$

oder

$$[x]_\sim \cap [y]_\sim = \emptyset\,.$$

Insbesondere ist jedes $x \in A$ in genau einer Äquivalenzklasse, nämlich $[x]_\sim$, enthalten.

BEWEIS

Anhand einer Wahrheitstabelle überprüft man leicht, dass wir für alle $x, y \in A$

$$[x]_\sim = [y]_\sim \qquad \Longleftrightarrow \qquad [x]_\sim \cap [y]_\sim \neq \emptyset$$

zeigen müssen.

Sei dazu zunächst $[x]_\sim = [y]_\sim$. Dann gilt $y \in [y]_\sim = [x]_\sim$ und damit $y \in [x]_\sim \cap [y]_\sim$, d.h. $[x]_\sim \cap [y]_\sim \neq \emptyset$.

Gilt umgekehrt $[x]_\sim \cap [y]_\sim \neq \emptyset$ dann gibt es ein $z \in [x]_\sim \cap [y]_\sim$. Wegen $x \sim z$ und $z \sim y$ folgern wir $x \sim y$.

Sei nun $w \in [x]_\sim$. Dann gilt $w \sim x$ und wegen $x \sim y$ damit auch $w \sim y$, d.h. $w \in [y]_\sim$. Mit anderen Worten haben wir $[x]_\sim \subset [y]_\sim$ gezeigt. Der Beweis von $[y]_\sim \subset [x]_\sim$ ist analog, und damit sehen wir insgesamt $[x]_\sim = [y]_\sim$.

BEMERKUNGEN

Eine Äquivalenzrelation $\sim$ auf A teilt A in die Äquivalenzklassen von $\sim$ auf. Die Elemente einer Äquivalenzklasse $[x]_\sim$ werden dabei als “gleichwertig” im Sinne von $\sim$ aufgefasst und jedes Element $y \in [x]_\sim$ kann als Repräsentant dieser Äquivalenzklasse genommen werden. In diesem Sinne dient eine Äquivalenzklasse zur Fokussierung auf das für einen bestimmten Zweck Wesentliche.

Für die Äquivalenzrelation $\sim := A \times A$ gilt $[x]_\sim = A$ für jedes $x \in A$ und damit ist $[x]_\sim \cap [y]_\sim \neq \emptyset$ nie erfüllt.

Weitere wichtige Relationen werden in der folgenden Definition eingeführt.

Definition 1.3.4

Sein $A \neq \emptyset$ eine Menge und $R \subset A \times A$ eine Relation. Dann heißt R **Halbordnung** auf A, genau dann wenn die folgenden drei Eigenschaften erfüllt sind:

i). **Reflexivität:** Für alle $x \in A$ gilt $(x, x) \in R$.

ii). **Anti-Symmetrie:** Für alle $x, y \in A$ gilt

$$\big((x, y) \in R \wedge (y, x) \in R\big) \Rightarrow x = y\,.$$

iii). **Transitivität:** Für alle $x, y, z \in A$ gilt

$$\big((x, y) \in R \wedge (y, z) \in R\big) \Rightarrow (x, z) \in R\,.$$

Gilt zusätzlich noch die **Vergleichbarkeit**, d.h.

$$\forall x, y \in A : \big((x, y) \in R \vee (y, x) \in R\big)$$

so ist R eine **Totalordnung**.

BEMERKUNGEN

Halbordnungen werden statt mit R in der Regel mit einem zu $\leq$ “ähnlichen” Symbol bezeichnet. Insbesondere ergibt $\leq$ auf jeder Teilmenge von $\mathbb{R}$ eine Totalordnung. Ferner liefert $\subset$ auf $\mathcal{P}(A)$ eine Halbordnung, die im allgemeinen keine Totalordnung ist.

Bis jetzt haben wir nur Relationen auf A^2 kennengelernt. Die folgende Art von Relationen kann auf beliebigen Produkten $A \times B$ betrachtet werden und gehört zu den zentralsten Begriffen der Mathematik.

Definition 1.3.5

Seien A und B nicht leere Mengen und $f \subset A \times B$ eine Relation zwischen A und B. Dann heißt f **Abbildung** oder **Funktion**, genau dann wenn für alle $x \in A$ genau ein $y_x \in B$ gibt, das zu x in Relation steht. In diesem Fall schreiben wir $f : A \to B$ und $f(x) := y_x$, oder auch

$$\begin{aligned} f : A &\to B \\ x &\mapsto f(x)\,. \end{aligned}$$

Die Menge A heißt **Definitionsbereich** der Funktion f und B heißt **Bildbereich** von f.

GLEICHHEIT VON FUNKTIONEN

Zwei Funktionen $f_1 : A_1 \to B_1$ und $f_2 : A_2 \to B_2$ sind **gleich**, geschrieben $f_1 = f_2$, genau dann wenn $A_1 = A_2$, $B_1 = B_2$ und

$$f_1(x) = f_2(x)$$

für alle $x \in A_1$ gilt.

Häufig wird eine Funktion durch eine konkrete Berechnungsvorschrift wie z.B.

$$f(x) := x^2 \tag{1.3.2}$$

angegeben. Ohne Angabe von Definitions- und Wertebereich spezifiziert dies aber noch keine Funktion. Setzen wir jedoch zusätzlich zum Beispiel $A := [0, 1]$ und $B := \mathbb{R}$, so wird aus der obigen Berechnungsvorschrift (1.3.2) eine Funktion. Ändern wir dann B in $B := [0, \infty)$ um, erhalten wir eine andere Funktion, während das Ersetzen von der Berechnungsvorschrift (1.3.2) zu

$$f(x) := (x + 1)^2 - 2x - 1$$

die Funktion auf $A := [0, 1]$ und $B := \mathbb{R}$ nicht ändert.

DIE IDENTITÄT UND WEITERE BEISPIELE

Ist A eine Menge so heißt die Abbildung $\mathrm{id}_A : A \to A, x \mapsto x$ **Identität** auf A. Ist $A \subset B$, so heißt die Abbildung $\mathrm{id}_{A,B} : A \to B, x \mapsto x$ **Inklusionsabbildung**. Obwohl beide Abbildungen die gleiche Berechnungsvorschrift haben sind sie nur im Fall $B = A$ auch tatsächlich gleich.

Ist A eine Menge und $B \subset A$ so heißt die Abbildung $\mathbf{1}_B : A \to \{0, 1\}$

$$\mathbf{1}_B(x) := \begin{cases} 1\,, & \text{falls } x \in B \\ 0\,, & \text{falls } x \notin B \end{cases}$$

Indikatorfunktion von B. Häufig wird auch der Bildbereich $\mathbb{R}$ betrachtet.

Sei A die Menge aller Autos, die am 31.12.2021 in Stuttgart gemeldet waren, und B die Menge aller möglichen Kennzeichenkombinationen. Dann ist $f : A \to B$ mit

$$f(x) := \text{Kennzeichen}(x)\,, \qquad x \in A \tag{1.3.3}$$

eine Abbildung. Umgekehrt können wir aus $g(y) :=$ Auto mit Kennzeichen(y) keine Abbildung $g : B \to A$ erstellen, da nicht jede mögliche Kennzeichenkombination auch vergeben war. An diesem Beispiel sieht man auch, dass die “Berechnungsvorschriften” sehr willkürlich sein kann.

Die folgende Definition führt ein paar weitere wichtige Begriffe für Abbildungen ein.

Definition 1.3.6

Sei $f : A \to B$ eine Abbildung und $A_0 \subset A$, $B_0 \subset B$. Dann heißt

i). $\mathbf{graph}(f) := \{(x, f(x)) : x \in A\}$ der **Graph** von f.

ii). $f(A_0) := \{f(x) : x \in A\}$ das **Bild** von A_0 unter f.

iii). $f^{-1}(B_0) := \{x \in A : f(x) \in B_0\}$ das **Urbild** von B_0 unter f.

Ferner heißt die Funktion $f_{|A_0} : A_0 \to B$, die durch $f_{|A_0}(x) := f(x)$ für alle $x \in A_0$ gegeben ist, die **Einschränkung** von f auf A_0.

GRAFISCHE DARSTELLUNGEN VON FUNKTIONEN

Funktionen können auf verschiedene Arten visualisiert werden, wie in den Abbildungen 43 und 3 illustriert ist.

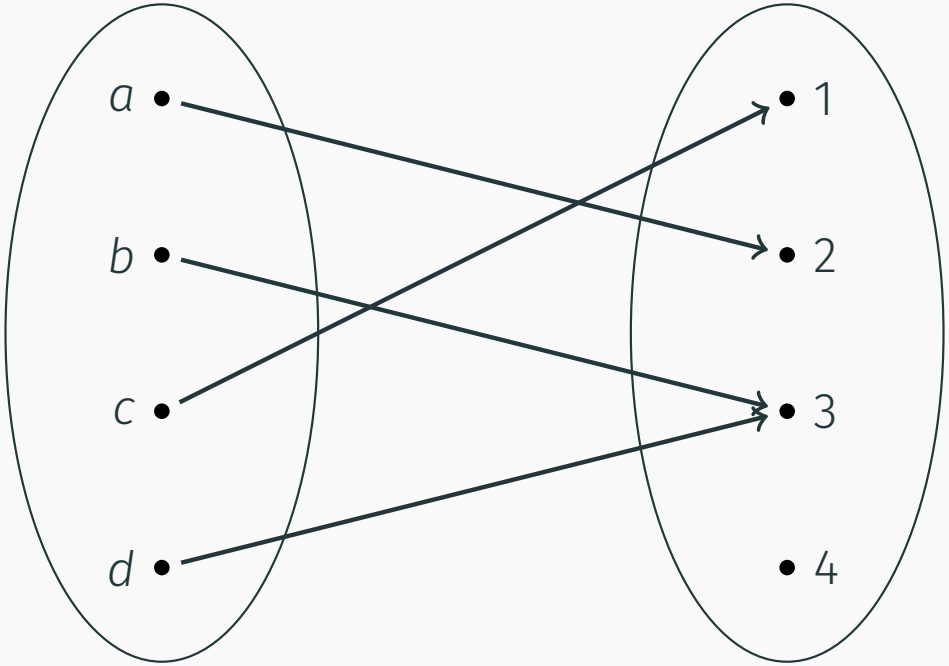


Abbildung: Schematische Darstellung einer Abbildung von einer Menge A (links) in eine Menge B (rechts). Die Pfeile deuten das Verhalten der Abbildung an, wie z.B. $f(b) = 3$.

GRAFISCHE DARSTELLUNGEN VON FUNKTIONEN

Sind $A, B \subset \mathbb{R}$, so lassen sich Funktionen auch einem aus der Schule bekannten Koordinatensystem darstellen. Ein Beispiel einer solchen Darstellung und 2 Beispiele von Relationen, die keine Funktionen sind findet sich in Abbildung 3.

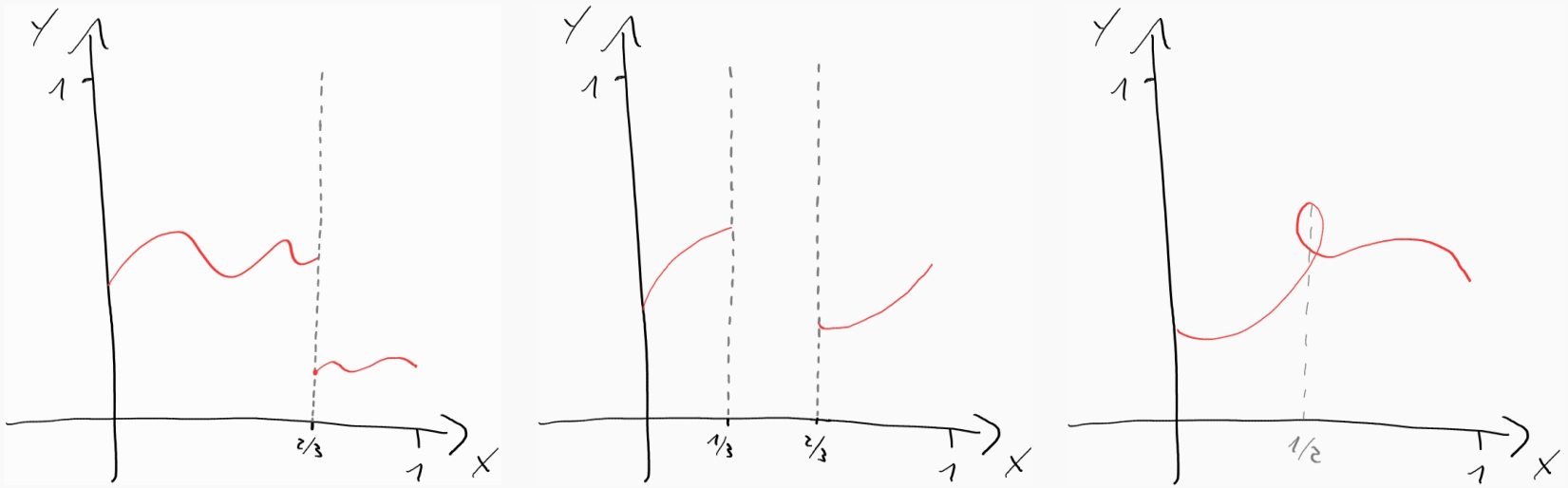


Abbildung: Illustrationen für drei Relationen (jeweils in rot) auf $[0, 1]^2$. **Links:** Eine Relation, die sogar eine Funktion ist. **Mitte:** Eine Relation, die keine Funktion ist, da zwischen 1/3 und 2/3 kein Funktionswert zugewiesen wird. **Rechts:** Eine Relation, die keine Funktion ist, da z.B. für $x = 1/2$ kein eindeutiger Funktionswert existiert.

In vielen Fällen erfüllen Abbildungen zusätzliche Eigenschaften. Die folgende Definition stellt drei grundlegende vor.

Definition 1.3.7

Sei $f : A \to B$ eine Funktion. Dann heißt f:

i). **injektiv**, genau dann wenn gilt:

$$\forall x_1, x_2 \in A : f(x_1) = f(x_2) \qquad \Rightarrow \qquad x_1 = x_2 \,.$$

ii). **surjektiv**, genau dann wenn $f(A) = B$ gilt, d.h.

$$\forall y \in B \exists x \in A : f(x) = y \,.$$

iii). **bijektiv**, genau dann wenn sie injektiv und surjektiv ist, d.h.

$$\forall y \in B \exists! x \in A : f(x) = y \,.$$

BEISPIELE

Für $A = B = \mathbb{R}$ ist die durch (1.3.2) gegebene Funktion weder injektiv noch surjektiv. Betrachten wir jedoch $A = [0, \infty)$ und $B = \mathbb{R}$ so wird sie injektiv, während wir für $A = \mathbb{R}$ und $B = [0, \infty)$ eine surjektive. Funktion bekommen. Für $A = B = [0, \infty)$ erhalten wir schließlich eine bijektive Funktion.

Die Inklusionsabbildung ist immer injektiv und sie ist surjektiv genau dann wenn $B = A$ gilt. Indikatorfunktionen sind “fast nie” injektiv und “meistens” surjektiv für den Bildbereich $\{0, 1\}$.

Die um (1.3.3) beschriebene “Kennzeichenabbildung” ist injektiv aber nicht surjektiv.

Ist $f : A \to B$ bijektiv, so gibt es zu jedem $y \in B$ genau ein $x_y \in A$ mit $f(x_y) = y$. Damit können wir eine neue Abbildung $f^{-1} : B \to A$ durch

$$f^{-1}(y) := x_y\,, \qquad\qquad y \in Y$$

definieren. Die Abbildung f^{-1} heißt **Umkehrabbildung** oder **Umkehrfunktion** von f. Wie man sich leicht überlegt, ist f^{-1} ebenfalls bijektiv und es gilt $(f^{-1})^{-1} = f$.

Die folgende Definition führt eine weitere Operation für Abbildungen ein.

Definition 1.3.8

Seien $f : A \to B$ und $g : B \to C$ Abbildungen, so heißt die durch

$$\begin{aligned} g \circ f : A &\to C \\ x &\mapsto g(f(x)) \end{aligned}$$

die **Komposition** von f und g

RECHENREGELN FÜR KOMPOSITIONEN

Die Komposition von Abbildungen ist assoziativ, d.h. es gilt

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

und aus diesem Grund schreibt man meistens nur $h \circ g \circ f$. Andererseits ist die Komposition im Allgemeinen nicht kommutativ, d.h. es gibt Funktionen f und g mit

$$g \circ f \neq f \circ g\,.$$

Für $A = B = C = \mathbb{R}$ ist dies zum Beispiel für $f(x) := x^2$ und $g(x) := x + 1$ der Fall, denn es gilt:

$$\begin{aligned} g \circ f(x) &= g(x^2) = x^2 + 1\,, \\ f \circ g(x) &= f(x+1) = (x+1)^2 = x^2 + 2x + 1\,, \end{aligned}$$

und damit $g \circ f(1) = 2 \neq 4 = f \circ g(1)$.

Schließlich gilt $f \circ \mathrm{id}_A = f$ und $\mathrm{id}_B \circ f$ für alle $f : A \to B$.

RECHENREGELN FÜR KOMPOSITIONEN

Ist $f : A \to B$ eine bijektive Abbildung so gelten zusätzlich die Formeln

$$f^{-1} \circ f = \mathrm{id}_A \,,$$
$$f \circ f^{-1} = \mathrm{id}_B \,.$$

Ist ferner $g : B \to C$ eine weitere bijektive Abbildung, so ist auch $g \circ f$ bijektiv und es gilt

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

Schließlich gilt immer $\mathrm{id}_A^{-1} = \mathrm{id}_A$.

DEFINITION VON TUPELN UND PRODUKTEN MIT ABBILDUNGEN

Für $n \geq 3$ hatten wir schon in Abschnitt 26 n-Tupel und n-fache Produkte A^n eingeführt. Dies war allerdings nicht wirklich formal sauber. Mit Hilfe von Abbildungen können wir dies jetzt nochmal formal korrekt wiederholen.

Für $n \geq 3$ sei dazu $I := \{1, 2, \ldots, n\}$. Ein n-Tupel $(a_1, \ldots, a_n)$ in A kann dann als Abbildung

$$I \to A$$

$$i \mapsto a_i$$

aufgefasst werden. Umgekehrt kann jede solche Abbildung wegen der Ordnung auf I als n-Tupel aufgefasst werden. Formal werden daher n-Tupel als Abbildungen definiert, wobei im allgemeinen Fall $A_1, \ldots, A_n$ noch etwas mehr Arbeit notwendig ist, um die zunächst betrachteten Abbildungen $I \to A_1 \cup A_2 \cup \ldots A_n$ geeignet einzuschränken.

Diese Einsicht wird uns demnächst nützlich sein, wenn wir unendlich lange Tupel einführen werden.

CHAPTER 2: ZAHLEN

Section 2.1
Natürliche Zahlen

Definition 2.1.1

Eine Menge $\mathbb{N}$ heißt Menge der **natürlichen Zahlen**, falls die folgenden fünf **Peano-Axiome** erfüllt sind:

P1 Es existiert ein besonderes Element $1 \in \mathbb{N}$.

P2 Zu jedem Element $n \in \mathbb{N}$ gibt es genau ein weiteres Element $n' \in \mathbb{N}$, das als der **Nachfolger** von n bezeichnet wird.

P3 Das Element 1 ist selbst kein Nachfolger, d.h.

$$\forall n \in \mathbb{N} : n' \neq 1.$$

P4 Die Nachfolger-Abbildung $n \mapsto n'$ auf $\mathbb{N}$ ist injektiv, d.h. für alle $n, m \in \mathbb{N}$ folgt aus $n' = m'$, dass $n = m$.

P5 **Induktionsaxiom:** Besitzt eine Teilmenge $M \subset \mathbb{N}$ die Eigenschaften

i). $1 \in M$
ii). $\forall n \in M : n' \in M$

so gilt $M = \mathbb{N}$.

Wir nutzen die übliche Notation und bezeichnen $2 := 1'$, $3 = 2'$, etc.

BEMERKUNGEN

Man kann zeigen, dass es tatsächlich eine Menge $\mathbb{N}$ gibt, die alle fünf Peano-Axiome erfüllt.

Ferner bestimmen die Peano-Axiome die Menge $\mathbb{N}$ im wesentlichen eindeutig: Wenn zwei Mengen $\mathbb{N}$ und $\widetilde{\mathbb{N}}$ die Eigenschaften **P1** bis **P5** besitzen, dann kann man zeigen, dass eine eindeutig bestimmte Bijektion $\varphi : \mathbb{N} \to \widetilde{\mathbb{N}}$ zwischen diesen Mengen existiert mit $\varphi(1) = \widetilde{1}$ und $\varphi(n') = \varphi(n)'$. Bis auf eine Umbenennung der Elemente haben wir damit durch die Eigenschaften **P1** bis **P5** die natürlichen Zahlen charakterisiert.

BEMERKUNGEN

Die Axiome charakterisieren die natürlichen Zahlen als eine nicht abbrechende Reihe von Elementen verbunden durch die Nachfolgeroperation,

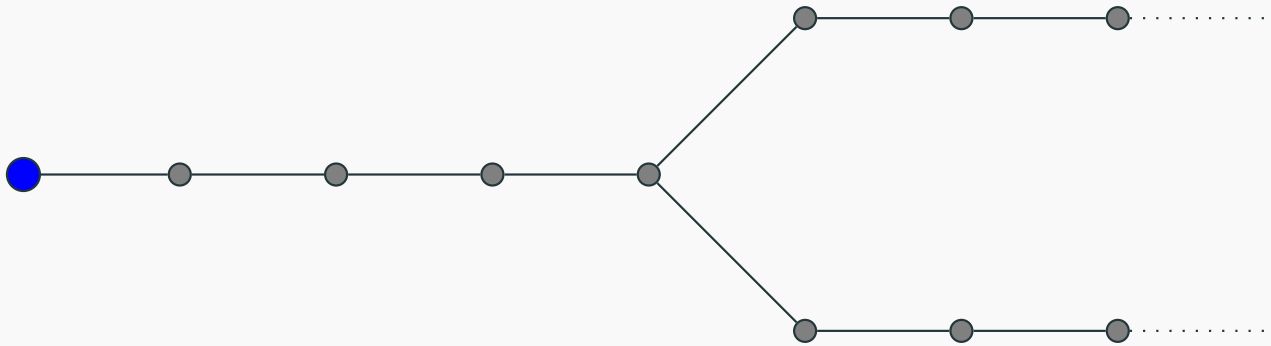
wobei die 1 als blauer Punkt dargestellt ist, und der Nachfolger n' einer Zahl n direkt rechts von ihr steht und mit einer Linie mit ihr verbunden ist.

Wir überlegen uns kurz, dass dieses Bild die natürlichen Zahlen richtig illustriert, indem wir zeigen, was die einzelnen Axiome ausschließen.

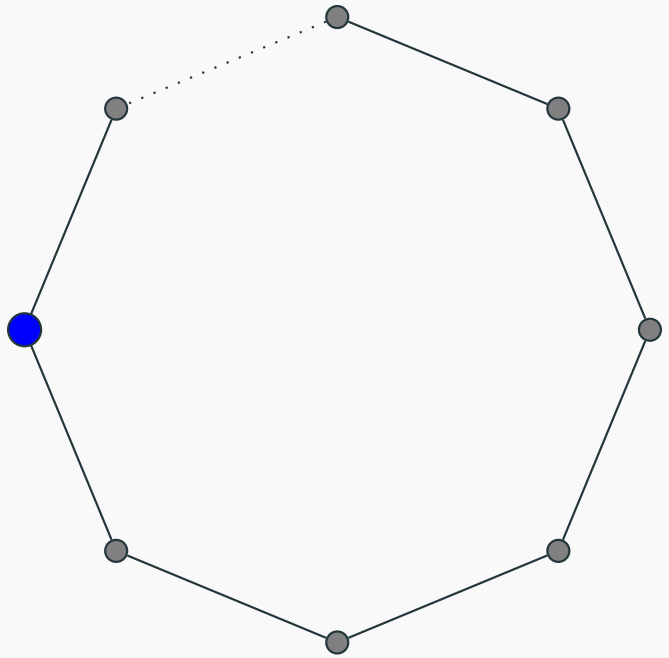
BEMERKUNGEN

Das Axiom **P2** schließt aus, dass die Zahlenreihe aufhört, oder ein “Loch” hat

oder aber sich an einer Stelle verzweigt

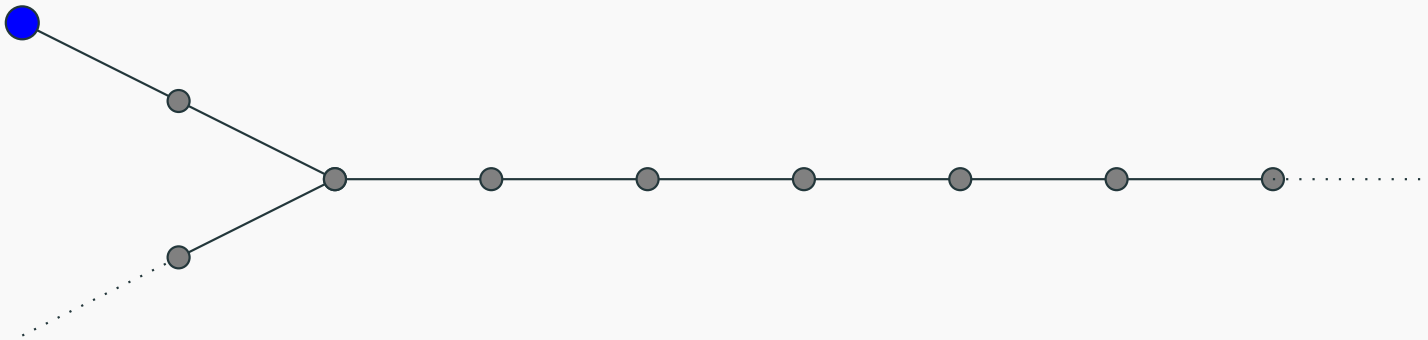


Axiom **P3** verhindert, dass die Zahlen zyklisch verlaufen,



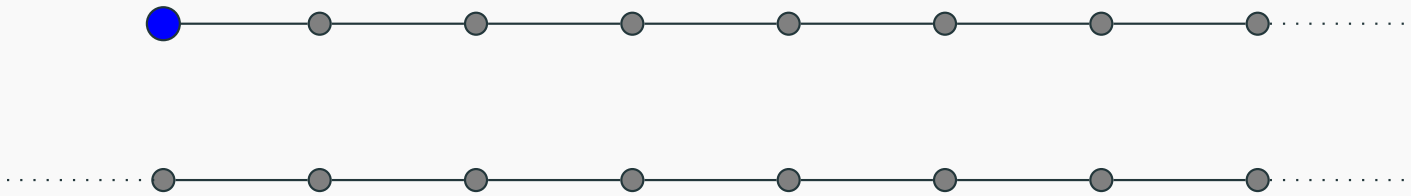
BEMERKUNGEN

Axiom **P4** schließt Zusammenführungen der Form



aus.

P5 schließt Bereiche aus, welche nicht nach endlich vielen Schritten von 1 aus gesehen erreichbar sind



REKURSION AM BEISPIEL VON RECHENOPERATIONEN

Mit Hilfe der Axiome von Peano können wir auf $\mathbb{N}$ die Addition definieren. Dazu fixieren wir ein $m \in \mathbb{N}$ und definieren $n \in \mathbb{N}$:

$$m + 1 := m' ,$$
$$m + n' := (m + n)' ,$$

wobei n' der Nachfolger von n ist und $(m + n)'$ der Nachfolger von $m + n$. Aus dem Induktionsaxiom schließen wir nun, dass wir damit $m + n$ für alle $n \in \mathbb{N}$ erklärt haben, und da $m \in \mathbb{N}$ beliebig war, auch für alle $m \in \mathbb{N}$.

Eine solche Definition heißt **rekursiv** und sie ist für $\mathbb{N}$ typisch. Um eine Größe A_n für jede natürliche Zahl n zu definieren, genügen nach dem Induktionsaxiom die folgenden beiden Schritte:

Rekursionsanfang (RA): Festlegung, was A_1 ist.

Rekursionsschritt (RS): Für jedes $n \in \mathbb{N}$ die Größe A_{n+1} mit den Größen $A_1, \ldots, A_n$ auszudrücken.

Rekursive Ansätze spielen aber auch in der Informatik eine wichtige Rolle, z.B. bei der Behandlung von sogenannten Bäumen.

REKURSION AM BEISPIEL VON RECHENOPERATIONEN

Neben der Addition kann auch die Multiplikation rekursiv definiert werden. Dazu fixieren wir wieder ein $m \in \mathbb{N}$ und definieren für $n \in \mathbb{N}$:

$$m \cdot 1 := m\,,$$
$$m \cdot n' := m \cdot n + m\,,$$

wobei wir annehmen, dass wir die Addition schon wie oben definiert haben. Schließlich definieren wir die Relation $<$ auf $\mathbb{N}$ durch

$$m < n \qquad :\Longleftrightarrow \qquad \exists r \in \mathbb{N} : n = m + r\,.$$

Damit können wir auch die Relation $\leq$ auf $\mathbb{N}$ durch

$$m < n \qquad :\Longleftrightarrow \qquad (m < n \vee m = n)\,.$$

Im folgenden setzen wir diese Operationen und Relationen sowie ihre Gesetzmäßigkeiten als bekannt voraus.

VOLLSTÄNDIGE INDUKTION

Ähnlich wie bei der Rekursion kann man die Peano-Axiome auch zum Beweis von Aussagen über naürliche Zahlen nutzen. Sei dazu $q(n)$ eine Aussageform über naturliche Zahlen n. Wenn wir zeigen wollen, dass $q(n)$ für jede natürliche Zahl n wahr ist, d.h., dass

$$\forall n \in \mathbb{N} : q(n) \tag{2.1.1}$$

wahr ist, gehen wir dazu wie folgt vor:

Induktionsanfang (IA): Wir zeigen, dass $q(1)$ wahr ist.

Induktionsschritt (IS): Wir zeigen, dass

$$\forall n \in \mathbb{N} : \big(q(n) \;\Rightarrow\; q(n+1)\big)$$

wahr ist. Dabei wird $q(n)$ Induktionsvoraussetzung (IV) genannt und $q(n+1)$ Induktionsbehauptung.

Mit anderen Worten zeigen wir für die Menge

$$M := \{n \in \mathbb{N} : q(n)\},$$

dass $1 \in M$ und für alle $n \in M$ auch $n' \in M$ gilt. Nach dem Induktionsaxioms wissen wir dann, dass $M = \mathbb{N}$ gilt und damit ist (2.1.1) tatsächlich bewiesen.

VOLLSTÄNDIGE INDUKTION: VARIATIONEN

Alternativ kann man im Induktionsschritt nicht nur $q(n)$ benutzen, sondern sogar $q(1), \ldots, q(n)$, d.h. es reicht zu zeigen, dass

$$\forall n \in \mathbb{N} : \Big(\big(q(1) \wedge \cdots \wedge q(n) \big) \;\Rightarrow\; q(n+1) \Big)$$

wahr ist.

Manchmal sind Aussagen erst ab einer gewissen Zahl n_0 wahr. In diesem Fall ist im Induktionsanfang $q(n_0)$ zu beweisen und im Induktionsschritt zeigt man dann nur $q(n_0) \;\Rightarrow\; q(n+1)$, bzw. $\big(q(n_0) \wedge \cdots \wedge q(n) \big) \;\Rightarrow\; q(n+1)$ für alle $n \geq n_0$. Dies gilt insbesondere, wenn man Aussagen für alle $n \in \mathbb{N}_0$ beweisen möchte.

Die aus der Schule bekannten Gesetzmäßigkeiten für die Addition und Multiplikation in $\mathbb{N}$ oder $\mathbb{N}_0$ können über vollständige Induktion bewiesen werden. Dies wollen wir hier aber nicht vertiefen.

Beispiel für vollständige Induktion

Um die vollständige Induktion an einem Beispiel zu betrachten, wollen wir das folgende Lemma zeigen, zu dem es auch eine Anekdote von Gauss gibt.

Lemma 2.1.2

Für alle $n \in \mathbb{N}$ gilt $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.

Beweis.

Induktionsanfang: Hier müssen die Aussage für $n = 1$ zeigen, d.h. $1 = \frac{1(1+1)}{2}$. Das ist aber offensichtlich wahr.
Induktionsschritt: Hier fixieren wir ein $n \in \mathbb{N}$ und nehmen als Induktionsvoraussetzung an, dass die Aussage für dieses n wahr ist. Dann gilt

$$\begin{aligned} 1 + 2 + \cdots + n + (n+1) = \frac{n(n+1)}{2} + (n+1) &= (n+1) \cdot \left(\frac{n}{2} + 1\right) \\ &= \frac{(n+1)(n+2)}{2} \end{aligned}$$

und damit ist die Aussage auch auch für $n + 1$, d.h. die Induktionsbehauptung gezeigt. □

WEITERE REKURSIVE DEFINITIONEN

Im folgenden sei n immer eine natürliche Zahl. Dann wird die Fakultät von n durch

$$1! := 1\,, \qquad (n+1)! := n! \cdot (n+1)$$

definiert. Haben wir ferner Zahlen $a_k \in \mathbb{N}$ für alle $k \in \mathbb{N}$ so definieren wir für $m \in \mathbb{N}$ und $n \geq m$:

$$\sum_{k=m}^{m} a_k := a_m\,, \qquad \sum_{k=m}^{n+1} a_k = a_{n+1} + \sum_{k=m}^{n} a_k$$

und

$$\prod_{k=m}^{m} a_k := a_m\,, \qquad \prod_{k=m}^{n+1} a_k = a_{n+1} \cdot \prod_{k=m}^{n} a_k\,.$$

Diese Schreibweisen sind im Zweifelsfrei präziser als z.B. $a_1 + \cdots + a_n$, während die letztere suggestiver ist.

Section 2.2
Kombinatorik

MÄCHTIGKEIT VON MENGEN

In diesem Abschnitt wollen wir uns mit Zählen von Objekten in einer Menge beschäftigen.

Eine Menge A heißt **n-elementig**, falls es eine Bijektion $A \to \{1, 2, \ldots, n\}$ gibt. Die Zahl n heißt dann die **Anzahl** der Elemente von A oder auch die **Kardinalität** oder **Mächtigkeit** von A und wir schreiben dafür

$$|A| := \#A := n\,.$$

Die Anzahl ist eindeutig bestimmt, da es für verschiedene natürliche Zahlen $n \neq m$ keine Bijektion $\{1, \ldots, n\} \to \{1, \ldots, m\}$ geben kann.

Eine Menge A, für die eine injektive Abbildung $\mathbb{N} \to A$ existiert, heißt **unendlich**. Existiert keine solche injektive Abbildung, so gibt es ein $n \in \mathbb{N}$ mit $|A| = n$ und wir sagen die Menge A ist **endlich**. Wir vereinbaren weiter, dass $|\emptyset| := 0$ gilt.

Sind A und B zwei n-elementige Mengen und $f : A \to \{1, \ldots, n\}$ und $g : B \to \{1, \ldots, n\}$ zwei Bijektionen, so ist $h := g^{-1} \circ f : A \to B$ eine Bijektion. Gibt es umgekehrt eine Bijektion $h : A \to B$ zwischen zwei beliebigen endlichen Mengen A und B, so gilt $|A| = |B|$.

Theorem 2.2.1

Seien A und B endliche Mengen, Dann gilt:

i). $|A \cup B| = |A| + |B| - |A \cap B|$.

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

i). Sind A und B disjunkt, so gilt offensichtlich $|A \cup B| = |A| + |B|$. Im Allgemeinen Fall betrachten wir zunächst $A \cup B = A \cup (B \setminus A)$. Da A und $(B \setminus A)$ disjunkt sind folgt

$$|A \cup B| = |A \cup (B \setminus A)| = |A| + |B \setminus A| \,. \tag{2.2.1}$$

Schließlich gilt $B \setminus A = B \setminus (A \cap B)$, und da $B \setminus (A \cap B)$ und $A \cap B$ disjunkt sind, folgt wegen $B = (B \setminus (A \cap B)) \cup (A \cap B)$:

$$|B| = |B \setminus (A \cap B)| + |A \cap B| = |B \setminus A| + |A \cap B| \,.$$

Damit haben wir $|B \setminus A| = |B| - |A \cap B|$ und Einsetzen in (2.2.1) ergibt die Behauptung.

ii). Seien $m := |A|$ und $n := |B|$, sowie $f : A \to \{1, \ldots, m\}$ und $g : B \to \{1, \ldots, n\}$ zwei Bijektionen. Dann ist

$$\begin{aligned} A \times B &\to \{1, \ldots, m\} \times \{1, \ldots, n\} \\ (a, b) &\mapsto \big(f(a), g(b)\big) \end{aligned}$$

eine Bijektion. Die Menge $\{1, \ldots, m\} \times \{1, \ldots, n\}$ kann man sich rechteckig angeordnet denken und zeilenweise gezählt ergeben sich mn Elemente.

ANZAHL VON ABBILDUNGEN

Seien A und B Mengen, so schreiben wir $B^A := \{f \mid f : A \to B\}$ für die Menge der Abbildungen von A nach B. Der folgende Satz bestimmt die Mächtigkeit von B^A.

Theorem 2.2.2

Seien A und B endliche Mengen. Dann gilt $|B^A| = |B|^{|A|}$.

BEWEIS

Wir schreiben $m := |A|$ und $n := |B|$ und führen den Beweis über Induktion über m.

Induktionsanfang: Für $m = 1$, haben wir $A = \{a\}$ und jede Abbildung $f : A \to B$ ist durch Angabe von $f(a) \in B$ eindeutig bestimmt. Wegen $|B| = n$ haben wir hierfür n verschieden Werte und damit folgt $|B^A| = |B| = n^1$.

Induktionsschritt: Sei $|A| = m + 1$ und $a \in A$. Wir setzen $A' := A \setminus \{a\}$. Dann gilt $|A'| = m$ nach Satz 2.2.1 und daher wissen wir nach Induktionsvoraussetzung, dass $|B^{A'}| = |B|^{|A'|}$ gilt.

Sei nun $f : A \to B$. Dann ist $f_{|A'} \in B^{A'}$ und $f(a) \in B$ und ferner wird f eindeutig durch Angabe von $f_{|A'}$ und $f(a)$ bestimmt. Umgekehrt kann jede Funktion $g : A' \to B$ eindeutig zu einer Funktion auf A fortgesetzt werden, indem man den Funktionswert $g(a)$ festlegt. Formal erhalten wir also eine Bijektion

$$B^A \to B^{A'} \times B$$
$$f \mapsto (f_{|A'}, f(a))$$

Damit folgt $|B^A| = |B^{A'} \times B| = |B^{A'}| \cdot |B| = n^m \cdot n = n^{m+1} = |B|^{|A|}$ mit Satz 2.2.1.

Mächtigkeit von Potenzmengen

Mit Hilfe des Satzes 2.2.2 können wir jetzt die Mächtigkeit von Potenzmengen bestimmen.

Korollar 2.2.3

Sei A eine endliche Menge. Dann gilt $|\mathcal{P}(A)| = 2^{|A|}$.

Wir setzen $B := \{0, 1\}$ und bemerken, dass $|B| = 2$ gilt. Dann lässt sich jede Teilmenge C von A durch ihre Indikatorfunktion $\mathbf{1}_C$ eindeutig beschreiben. Wir haben daher eine Bijektion

$$\begin{aligned} \mathcal{P}(A) &\to B^A \\ C &\mapsto \mathbf{1}_C \,. \end{aligned}$$

Die Behauptung folgt nun aus Satz 2.2.2.

ANZAHL VON UMORDNUNGEN

Sei A eine endliche Menge, so wissen wir bereits, dass es $|A|^{|A|}$ Abbildungen $A \to A$ gibt? Wir wollen nun untersuchen, wie viele **Permutationen**, d.h. bijektive Abbildungen $A \to A$ es gibt. Im folgenden schreiben wir dazu

$$\mathrm{Perm}(A) := \{f : A \to A \mid f \text{ bijektiv}\}.$$

Gilt $|A| = n$ und sind $a_1, \ldots, a_n$ die Elemente von A, dann ergibt jede Permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ eine andere Anordnung $(a_{\pi(i)}, \ldots, a_{\pi(n)})$ dieser Elemente. Die Anzahl von Umordnungen oder Anordnungen dieser Elemente ist daher gleich der Anzahl von Permutationen $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Theorem 2.2.4

Sei A eine endliche Menge. Dann gilt $|\mathrm{Perm}(A)| = |A|!$.

Informell kann man sich dies am Mischen von Karten klar machen: Für $n = 1$ gilt es nur eine Anordnung. Haben wir schon n Karten gemischt (mit $n!$ Anordnungen nach Induktionsvoraussetzung), dann haben wir genau $n + 1$ Möglichkeiten, eine $(n + 1)$-te Karte in diese n Karten einzusortieren. Zusammen ergibt dies $n! \cdot (n + 1) = (n + 1)!$ Anordnungen.

BEWEIS

Es reicht, die Menge $A_n := \{1, \ldots, n\}$ zu betrachten. Wir führen einen Induktionsbeweis über n.

Induktionsanfang: Für $n = 1$ gibt es genau eine bijektive Abbildung $\{1\} \to \{1\}$.

Induktionsschritt: Für jedes $b \in A_{n+1}$ fixieren wir zunächst eine Bijektion $h_b : A_{n+1} \setminus \{b\} \to A_n$.

Wie im Beweis von Satz 2.2.2 können wir jede Abbildung $f : A_{n+1} \to A_{n+1}$ in $f_{|A_n}$ und $b := f(n+1)$ eindeutig zerlegen. Ist f bijektiv, so gilt $f(n+1) \notin f(A_n) = f_{|A_n}(A_n)$. Ferner ist die Abbildung $f_{|A_n} : A_n \to A_{n+1} \setminus \{b\}$ bijektiv und damit ist $h_b \circ f_{|A_n} \in \mathrm{Perm}(A_n)$. Insgesamt erhalten wir eine Bijektion

$$\begin{aligned} \mathrm{Perm}(A_{n+1}) &\to \mathrm{Perm}(A_n) \times A_{n+1} \\ f &\mapsto \left(h_{f(n+1)} \circ f_{|A_n}, f(n+1)\right). \end{aligned}$$

Nach Induktionsvoraussetzung wissen wir $|\mathrm{Perm}(A_n)| = n!$ und mit $|A_n| = n + 1$ und Satz 2.2.1 folgt dann die Behauptung.

Für $k, n \in\in \mathbb{N}_0$ mit $k \leq n$ definieren wir den **Binomialkoeffizienten** durch

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}, \tag{2.2.2}$$

wobei $0! := 1$ gesetzt wird.

Theorem 2.2.5

Sei A eine n-elementige Menge und $k \in \mathbb{N}_0$ mit $k \leq n$. Da gibt es genau $\binom{n}{k}$ verschiedene k-elementige Teilmengen von A, d.h.

$$\left|\{B \subset A : |B| = k\}\right| = \binom{n}{k}.$$

Beim Lotto “6 aus 49” gibt es beispielsweise $\binom{49}{6} = 13.983.816$ Möglichkeiten, 6 Kugeln aus den 49 Kugeln zu ziehen.

Die Gesamtzahl der möglichen Anordnungen von k verschiedenen Elementen aus A_n ist

$$n(n-1)\dots(n-k+1)\,,$$

wie man sich leicht durch sukzessives Ziehen und Anordnen von Elementen aus A_n ohne Zurücklegen klar machen kann. Ignoriert man Anordnung, so fallen alle Kombinationen zusammen, die durch Umordnung entstehen, d.h., jeweils $k!$ Kombinationen fallen zusammen. Wegen

$$\frac{n(n-1)\dots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

folgt die Behauptung.

Die Berechnung der Binomialkoeffizienten mit Hilfe von (2.2.2) ist im Allgemeinen zu aufwendig. Der folgende Satz bietet eine alternative, rekursive Berechnung, die nur auf Addition beruht.

Theorem 2.2.6

Für $k, n \in \mathbb{N}$ mit $0 < k < n$ gilt $\binom{n}{0} = \binom{n}{n} = 1$ und

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

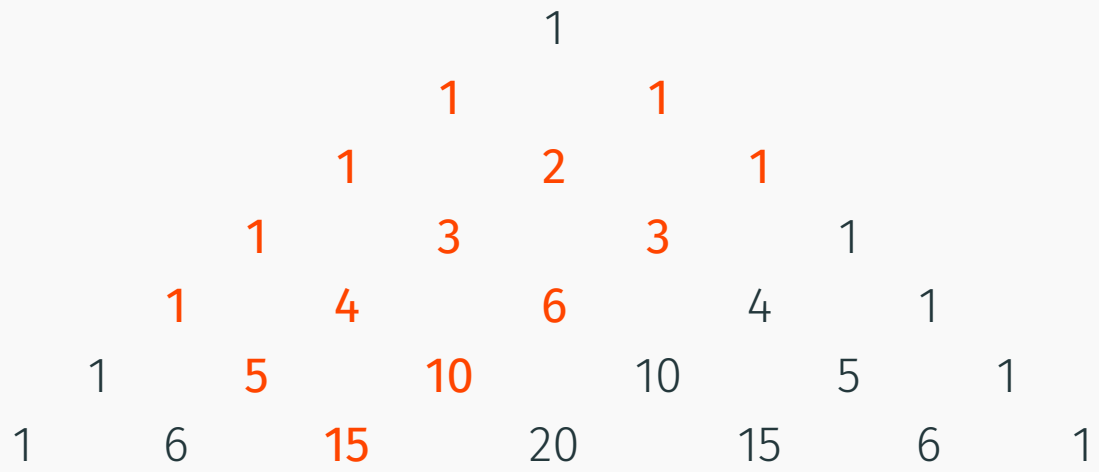
Mit Bruchrechnung erhalten wir

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

Ferner ist $\binom{n}{0}=\binom{n}{n}=1$ offensichtlich.

Pascal'sches Dreieck

Die Rekursionsformel kann man sich am Pascal'schem Dreieck illustrieren, wobei die Zeilen von oben nach unten mit $n = 0, 1, 2, \ldots$ nummeriert sind und die "Spalten" von links nach rechts mit $k = 0, \ldots, n$.



Die hervorgehobenen Zahlen zeigen $\binom{6}{2}$ und die zur rekursiven Berechnung notwendigen Binomialkoeffizienten mit $n < 6$. Die Spitze, die für $\binom{0}{0}$ steht, ist streng genommen zur Berechnung nicht notwendig, da immer schon bei $\binom{1}{0}$ oder $\binom{1}{1}$ abgebrochen werden kann.

Für große n und k kann man Binomialkoeffizienten, oder allgemeiner Fakultäten, auch durch die Stirling'sche Formel approximativ berechnen:

$$\mathrm{e}^{\frac{1}{1+12n}} \leq \frac{n!\mathrm{e}^n}{\sqrt{2\pi n}\, n^n} \leq \mathrm{e}^{\frac{1}{12n}}, \qquad n \geq 1.$$

BINOMISCHER LEHRSATZ

Binomialkoeffizienten sind für eine Vielzahl von Berechnungen nützlich. An dieser Stelle wollen nur den sogenannten binomischen Lehrsatz erwähnen.

Theorem 2.2.7

Für alle $a, b \in \mathbb{R}$ und $n \in \mathbb{N}_0$ gilt

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

BEWEIS

Der Beweis wird durch Induktion über $n \in \mathbb{N}_0$ geführt.

Induktionsanfang: Für $n = 0$ gilt

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

Induktionsschritt: Wir nehmen an, dass die Formel für n wahr ist. Um zu zeigen, dass sie auch für $n+1$ wahr ist, betrachten wir die Rechnung

$$
\begin{aligned}
(a+b)^{n+1} &= (a+b)(a+b)^n \\
&= (a+b)\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k} \\
&= \sum_{k=0}^{n}\binom{n}{k}a^{k+1}b^{n-k} + \sum_{k=0}^{n}\binom{n}{k}a^k b^{n+1-k} \\
&= \sum_{k=1}^{n+1}\binom{n}{k-1}a^k b^{n+1-k} + \sum_{k=0}^{n}\binom{n}{k}a^k b^{n+1-k} \\
&= \sum_{k=1}^{n}\binom{n}{k-1}a^k b^{n+1-k} + \binom{n}{n}a^{n+1}b^0 + \binom{n}{0}a^0 b^{n+1} + \sum_{k=1}^{n}\binom{n}{k}a^k b^{n+1-k} \\
&= a^{n+1} + \sum_{k=1}^{n}\left(\binom{n}{k-1} + \binom{n}{k}\right)a^k b^{n+1-k} + b^{n+1} \\
&= \sum_{k=0}^{n+1}\binom{n+1}{k}a^k b^{n+1-k},
\end{aligned}
$$

wobei wir im letzten Schritt Satz 2.2.6 verwendet haben.

Section 2.3
Ganze und Rationale Zahlen

Ausgehend von den natürlichen Zahlen $\mathbb{N}$ kann man die ganzen Zahlen gewinnen:

$$\mathbb{Z} := \{\dots, -2, -1, 0, 1, 2, \dots\} = \{m - n : m \in \mathbb{N}_0 \wedge n \in \mathbb{N}_0\}.$$

Möchte man dies formal sauber konstruieren, kann man z.B. auf $\mathbb{N}_0 \times \mathbb{N}_0$ die Äquivalenzrelation

$$(m_1, n_1) \sim (m_2, n_2) \qquad :\Longleftrightarrow \qquad m_1 + n_2 = m_2 + n_1 \tag{2.3.1}$$

definieren, die Zahlenpaare zusammenfasst, die die gleiche Differenz haben. Jede Äquivalenzklasse $[(m, n)]_\sim$ entspricht dann einer ganzen Zahl $m - n$.

Die Addition auf diesen "ganzen Zahlen" wird durch

$$[(m_1, n_1)]_\sim + [(m_2, n_2)]_\sim := [(m_1 + m_2, n_1 + n_2)]_\sim$$

definiert, wobei man aufpassen muss, dass dies von der Wahl der Repräsentanten unabhängig ist. Diese Konstruktion wollen wir nicht vertiefen, sondern nur in Abbildung 4 illustrieren.

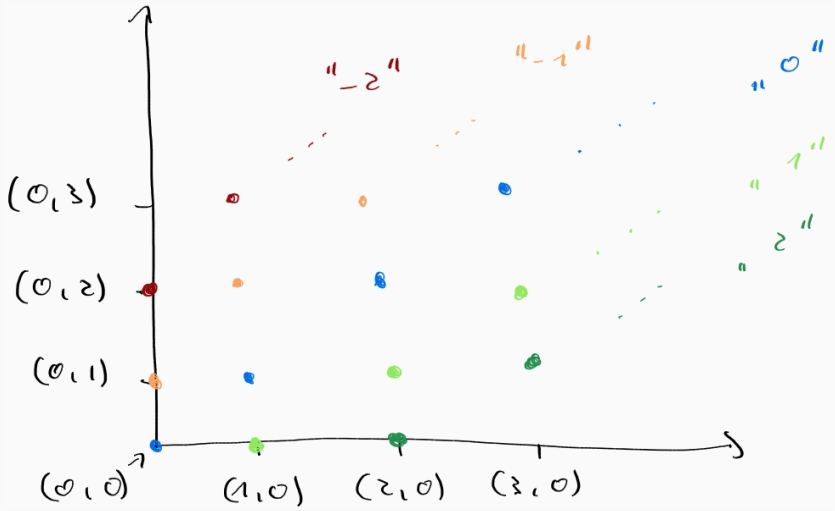


Abbildung: Konstruktion der ganzen Zahlen mit Hilfe der Äquivalenzrelation (2.3.1). Jede Diagonale entspricht einer Äquivalenzklasse und somit einer ganzen Zahl. Naheliegende Repräsentanten liegen auf den Achsen, wobei $(m, 0)$ die natürlichen Zahlen $\mathbb{N}_0$ repräsentieren und $(0, n)$ die "negativen Zahlen" $-\mathbb{N}_0$ repräsentieren.

Es kann gezeigt werden, dass die ganzen Zahlen zusammen mit der Addition eine kommutative Gruppe im Sinne der folgenden Definition bilden.

Definition 2.3.1

Sei G eine Menge und $\cdot : G \times G \to G$ eine Abbildung. Dann heißt $(G, \cdot)$ **Gruppe**, falls die folgenden drei Bedingungen erfüllt sind:

i). **Assoziativität:** Für alle $x, y, z \in G$ gilt $x \cdot (y \cdot z) = (x \cdot y) \cdot z$.

ii). **Neutrales Element:** Es existiert ein Element $1 \in G$, so dass für alle $x \in G$ gilt: $1 \cdot x = x \cdot 1 = x$

iii). **Inverses Element:** Für alle $x \in G$ existiert ein $x^{-1} \in G$ mit $x \cdot x^{-1} = x^{-1} \cdot x = 1$.

Ist ferner $\cdot$ **kommutativ**, d.h. es gilt $x \cdot y = y \cdot x$ für alle $x, y \in G$, dann heißt $(G, \cdot)$ **kommutative Gruppe** oder **Abel'sche Gruppe**.

BEMERKUNGEN UND BEISPIELE

In einer Gruppe $(G, \cdot)$ kann es nur ein neutrales Element geben, denn wenn 1 und $1'$ neutrale Elemente sind, so folgt $1 = 1 \cdot 1' = 1'$. Ferner ist jedes inverse Element eindeutig. Denn wenn es zu $x \in G$ zwei inverse Elemente x^{-1} und $\tilde{x}$ gibt, so folgt

$$\tilde{x} = \tilde{x} \cdot 1 = \tilde{x} \cdot (x \cdot x^{-1}) = (\tilde{x} \cdot x) \cdot x^{-1} = 1 \cdot x^{-1} = x^{-1}\,.$$

Damit hat die Gleichung $a \cdot x = b$ für feste $a, b \in G$ genau eine Lösung x, nämlich $x = a^{-1} \cdot b$. Entsprechend hat die Gleichung $x \cdot a = b$ die eindeutige Lösung $x = b \cdot a^{-1}$. Ähnlich leicht lassen sich die Formeln

$$1^{-1} = 1\,, \qquad (x^{-1})^{-1} = x \qquad \text{und} \qquad (xy)^{-1} = y^{-1}x^{-1}$$

nachweisen.

Ist A eine Menge und $\mathrm{Perm}(A)$ die Menge alle Bijektionen $A \to A$, dann ist $(\mathrm{Perm}(A), \circ)$ eine Gruppe, wobei id_A das neutrale Element ist und die Umkehrabbildung f^{-1} zu einem $f \in \mathrm{Perm}(A)$ ist das inverse Element. Im allgemeinen ist $(\mathrm{Perm}(A), \circ)$ aber nicht kommutativ, vgl. auch Seite 48.

Für $(\mathbb{Z}, +)$ sind die Assoziativität und Kommutativität aus der Schule bekannt. Das neutrale Element ist die 0 und für $k \in \mathbb{Z}$ ist $-k$ das inverse Element. In unserer Konstruktion (2.3.1) gilt $0 = [(0,0)]_\sim$ und $-k = [(n,m)]_\sim$ für $k = [(m,n)]_\sim$. Damit ist $(\mathbb{Z}, +)$ tatsächlich eine kommutative Gruppe.

Für $k, m \in \mathbb{Z}$ definieren wir die Subtraktion als

$$k - m := k + (-m)\,.$$

Schließlich können wir auf $\mathbb{Z}$ auch eine Totalordnung durch

$$[(m_1, n_1)]_\sim \geq [(m_2, n_2)]_\sim \qquad :\Longleftrightarrow \qquad m_1 + n_2 \geq m_2 + n_1$$

definieren, wobei wir die Konstruktion von $\mathbb{Z}$ aus (2.3.1) zugrunde gelegt haben. Die natürlichen Zahlen $\mathbb{N}_0$ entsprechen dann den ganzen Zahlen $[(m,n)]_\sim \geq [(0,0)]_\sim$, was gleichbedeutend mit $m \geq n$ ist.

Während die Multiplikation auf $\mathbb{Z}$ ohne Probleme definiert werden kann, ist das Dividieren von Zahlen in $\mathbb{Z}$ nur sehr eingeschränkt möglich. Um dieses Problem zu beheben, betrachten wir die rationalen Zahlen

$$\mathbb{Q} := \left\{ \frac{m}{n} : m \in \mathbb{Z} \wedge n \in \mathbb{N} \right\}. \tag{2.3.2}$$

Auch hier kann man diese recht informelle Definition formalisieren, indem man z.B. auf $\mathbb{Z} \times \mathbb{N}$ die Äquivalenzrelation

$$(m_1, n_1) \sim (m_2, n_2) \qquad :\Longleftrightarrow \qquad m_1 \cdot n_2 = m_2 \cdot n_1$$

betrachtet, die Paare mit gleichem "Quotienten" zusammenfasst. Der Ausdruck $\frac{m}{n}$ wird entsprechend durch $[(m, n)]_\sim$ formalisiert.

Addition und Multiplikation können durch

$$\frac{m_1}{n_1} + \frac{m_2}{n_2} := \frac{m_1 n_2 + m_2 n_1}{n_1 n_2}$$
$$\frac{m_1}{n_1} \cdot \frac{m_2}{n_2} := \frac{m_1 m_2}{n_1 n_2}$$

definiert werden, wobei man überprüfen muss, dass “äquivalente” Brüche das Ergebnis nicht verändern. Die entsprechenden Rechnungen sind nicht schwierig aber zeitaufwändiger.

In der Darstellung (2.3.2) entsprechen die ganzen Zahlen $\mathbb{Z}$ den Brüchen $\frac{m}{n}$ mit $n = 1$. Es ist offensichtlich, dass auf diese Weise die eben definierte Addition und Multiplikation auf $\mathbb{Q}$ der auf $\mathbb{Z}$ entspricht.

Weitere elementare Rechnungen zeigen, dass $(\mathbb{Q}, +, \cdot)$ ein Körper im Sinne der folgenden Definition bilden.

Definition 2.3.2

Sei K eine Menge und $+ : K \times K \to K$ und $\cdot : K \times K \to K$ zwei Abbildungen. Dann heißt $(K, +, \cdot)$ **Körper**, falls die folgenden Bedingungen erfüllt sind:

1. $(K, +)$ ist eine kommutative Gruppe mit neutralem Element 0.
2. $(K \setminus \{0\}, \cdot)$ ist eine kommutative Gruppe mit neutralem Element 1.
3. **Distributivgesetz:** Für alle $x, y, z \in K$ gilt $x \cdot (y + z) = x \cdot y + x \cdot z$.

In der Darstellung (2.3.2) von $\mathbb{Q}$ ist $\frac{0}{1}$ das neutrale Element der Addition und $\frac{1}{1}$ das neutrale Element der Multiplikation. Ferner ist $\frac{-m}{n}$ das additive Inverse zu $\frac{m}{n}$. Schließlich ist das multiplikative Inverse von $\frac{m}{n}$ mit $m \neq 0$ entweder $\frac{n}{m}$ oder $\frac{-n}{-m}$, je nachdem, ob $m > 0$ oder $m < 0$ ist.

BEMERKUNGEN

Ist $(K, +, \cdot)$ ein Körper, so gilt $1 \neq 0$, denn das neutrale Element 1 der Gruppe $(K \setminus \{0\}, \cdot)$ muss $1 \in K \setminus \{0\}$ erfüllen.

Für $x \in K$ gilt immer $0 \cdot x = x \cdot 0 = 0$, denn das Distributivgesetz zusammen mit der neutralen Eigenschaft der 0 ergibt

$$0 \cdot x + 0 \cdot x = (0 + 0) \cdot x = 0 \cdot x\,.$$

Durch Addition von $-(0 \cdot x)$ auf beiden Seiten bekommen wir dann $0 \cdot x = 0$.

Die Eigenschaft $0 \cdot x = 0$ für alle $x \in K$ zeigt, dass es kein multiplikatives Inverses $0^{-1} \in K$ von 0 geben kann, denn sonst hätten wir ja $0 \cdot 0^{-1} = 1$.

BEMERKUNGEN

Haben wir $x, y \in K$ mit $x \cdot y = 0$, so gilt $x = 0$ oder $y = 0$. Ist beispielsweise $x \neq 0$, so folgt $y = y \cdot (x \cdot x^{-1}) = (y \cdot x) \cdot x^{-1} = 0 \cdot x^{-1} = 0$.

Für $x \in K$ gilt $(-1)x = -x$, wie die folgende Rechnung zusammen mit der Eindeutigkeit von $-x$ zeigt:

$$x + (-1) \cdot x = 1 \cdot x + (-1) \cdot x = (1 + (-1)) \cdot x = 0 \cdot x = 0\,.$$

Hieraus können wir dann auch leicht $(-x) \cdot y = -(x \cdot y) = x \cdot (-y)$ und $(-1) \cdot (-1) = -(-1) = 1$ folgern.

In Körpern wird häufig der Multiplikationspunkt " $\cdot$ " weggelassen, d.h. wir schreiben $xy := x \cdot y$. Außerdem definieren wir die die Subtraktion und die Division durch

$$x - y := x + (-y) \qquad \text{und} \qquad \frac{x}{y} := xy^{-1}\,,$$

wobei für die Division $y = 0$ natürlich ausgeschlossen ist. Die letzte Schreibweise steht dabei nicht im Widerspruch zu unserer Darstellung (2.3.2) von $\mathbb{Q}$.

TOTALORDNUNG AUF DEN RATIONALE ZAHLEN

Neben der Addition und Multiplikation können wir auf $\mathbb{Q}$ auch eine Totalordnung definieren. Benutzen wir die Darstellung (2.3.2), so können wir diese durch

$$\frac{m_1}{n_1} \leq \frac{m_2}{n_2} \qquad :\Longleftrightarrow \qquad m_1 n_2 \leq m_2 n_1$$

definieren. Den leichten Nachweis, dass dies wohldefiniert ist und tatsächlich eine Totalordnung ergibt, überspringen wir wieder aus Zeitgründen.

In der Darstellung (2.3.2) entsprechen, wie schon erwähnt, die ganzen Zahlen $\mathbb{Z}$ den Brüchen $\frac{m}{n}$ mit $n = 1$. Es ist offensichtlich, dass auf diese Weise die oben definierte Totalordnung der auf $\mathbb{Z}$ entspricht.

Die eben definierte Totalordnung verträgt sich mit der Addition und Multiplikation im Sinne der folgenden Definition.

Definition 2.3.3

Sei $(K, +, \cdot)$ ein Körper und $\leq$ eine Totalordnung auf K. Dann heißt $(K, +, \cdot, \leq)$ (an)geordneter Körper, falls für alle $x, y \in K$ mit $x \leq y$ und alle $s \in K$ und $t \geq 0$ gilt:

$$x + s \leq y + s\,, \tag{2.3.3}$$

$$t \cdot x \leq t \cdot y\,. \tag{2.3.4}$$

Im folgenden benutzen wir $\leq$, um den Relationen $<$, $\geq$ und $>$ die üblichen Bedeutungen zu geben. Ist insbesondere $x < y$, so können wir $\leq$ in (2.3.3) durch $<$ ersetzen und im Falle $t > 0$ gilt dies auch für (2.3.4). Außerdem sichert die Anti-Symmetrie und die Vergleichbarkeit, dass für $x, y \in K$ immer genau eine der folgenden drei Aussagen wahr ist: $x < y$, $x = y$ oder $x > y$.

In jedem geordneten Körper gelten die üblichen “Vergleichs-Regeln”. Zum Beispiel gilt

$$x \leq y \qquad \Longrightarrow \qquad -y \leq -x \tag{2.3.5}$$

wie aus (2.3.3) mit $s := -y - x$ leicht folgt. Ferner gilt für alle $y \in K$, dass

$$y \cdot y \geq 0\,, \tag{2.3.6}$$

denn im Fall $y \geq 0$, folgt $0 = y \cdot 0 \leq y \cdot y$ aus (2.3.4) mit $x := 0$ und $t := y$ und im Fall $y \leq 0$ wissen wir schon $-y \geq 0$ und damit folgt ebenfalls $y \cdot y = (-y) \cdot (-y) \geq 0$. Für $y := 1$ erhalten wir insbesondere

$$1 > 0\,, \tag{2.3.7}$$

da $0 \neq 1$ in jedem Körper gilt. Daraus können wir wiederum

$$y > 0 \qquad \Longrightarrow \qquad y^{-1} > 0 \tag{2.3.8}$$

schließen, denn wäre $y^{-1} \leq 0$, so würde aus (2.3.4) mit $x := y^{-1}$ und $t := y$

$$1 = y \cdot y^{-1} \leq y \cdot 0 = 0$$

folgen. Ein doppeltes Anwenden von (2.3.8) und (2.3.4) liefert dann die zu (2.3.5) analoge Implikation $0 < x \leq y \Rightarrow 0 < y^{-1} \leq x^{-1}$.

Wir betrachten ein Beispiel und formen die beiden Ungleichungen

$$-3a - 2 \leq 5 \leq -3a + 4$$

äquivalent um. Mit den oben gezeigten Regeln folgt

$$-3a - 2 \leq 5 \quad \Leftrightarrow \quad -3a \leq 7 \quad \Leftrightarrow \quad -\frac{7}{3} \leq a$$

und

$$5 \leq -3a + 4 \quad \Leftrightarrow \quad 1 \leq -3a \quad \Leftrightarrow \quad a \leq -\frac{1}{3}$$

und die beiden Ungleichungen sind damit zusammen äquivalent zu $-7/3 \leq a \leq -1/3$.

RATIONALE ZAHLEN ERFÜLLEN ARCHIMEDISCHES AXIOM

Neben den Axiomen eines geordneten Körpers erfüllt $\mathbb{Q}$ auch das sogenannte Archimedische Axiom wie der folgende Satz zeigt.

Theorem 2.3.4

Für alle $x, y \in \mathbb{Q}$ mit $x > 0$ und $y > 0$ existiert ein $k \in \mathbb{N}$ mit $k \cdot x > y$.

Beweis.

Wir betrachten die Darstellung (2.3.2), d.h. $x = \frac{m_1}{n_1}$ und $y = \frac{m_2}{n_2}$ mit $m_1, m_2, n_1, n_2 \in \mathbb{N}$. Wir setzen $k := 2m_2n_1$. Aus $2m_1n_2 > 1$ folgt dann

$$y = \frac{m_2}{n_2} < 2m_1n_2 \cdot \frac{m_2}{n_2} = 2m_1m_2 = 2m_2n_1\frac{m_1}{n_1} = 2m_2n_1x$$

und mit der Definition von k erhalten wir $y < kx$. □

Betrachten wir $\varepsilon := x$ und $y := 1$ so impliziert Satz 2.3.4, dass es zu jedem $\varepsilon \in \mathbb{Q}$ mit $\varepsilon > 0$ ein $n \in \mathbb{N}$ gibt mit $1/n < \varepsilon$.

Section 2.4
Reelle Zahlen

RATIONALE ZAHLEN KÖNNEN NICHT ALLES AUSDRÜCKEN

Sind $x, y \in \mathbb{Q}$ mit $x < y$ und setzen wir $z := (x + y)/2$, so gilt $x < z < y$. Mit anderen Worten finden wir zwei beliebig nahe beieinander liegenden, unterschiedlichen rationalen Zahlen immer noch eine rationale Zahl, die zwischen diesen beiden liegt. Damit können wir uns $\mathbb{Q}$ als Zahlengerade vorstellen. Leider hat diese Gerade trotz der obigen Beobachtung Lücken. So ist die Länge der Diagonalen in einem Quadrat der Seitenlänge 1 nicht in $\mathbb{Q}$, wie das folgende Lemma zeigt.

Lemma 2.4.1

Es gibt keine rationale Zahl $r \in \mathbb{Q}$ mit $r^2 = 2$.

BEWEIS

Angenommen, es gäbe eine solche rationale Zahl $r = \frac{m}{n}$, wobei wir wie üblich $m \in \mathbb{Z}$ und $n \in \mathbb{N}$ annehmen. Offensichtlich ist $m = 0$ unmöglich und im Fall $m < 0$ ist $r = \frac{-m}{n}$ eine weitere Zahl, die $r^2 = 2$ erfüllt. Wir können daher zusätzlich $m \in \mathbb{N}$ annehmen. Schließlich können wir ohne Einschränkung annehmen, dass der Bruch $\frac{m}{n}$ vollständig gekürzt ist.

Aus $r^2 = 2$ folgt nun $m^2 = 2n^2$ und damit ist m^2 gerade. Wäre m ungerade, dann gäbe es ein $k \in \mathbb{N}$ mit $m = 2k - 1$. Wegen

$$m^2 = (2k-1)^2 = 4k^2 - 4k + 1 = 2(2k^2 - 2k) + 1$$

wäre dann aber auch m^2 ungerade. Da letzteres falsch ist, muss m gerade sein. Damit gibt es ein $l \in \mathbb{N}$ mit $m = 2l$. Dies ergibt $4l^2 = m^2 = 2n^2$ und damit $2l^2 = n^2$. Mit anderen Worten ist n^2 gerade und ein Wiederholen des obigen Arguments zeigt uns dann, dass auch n gerade ist.

Damit sind sowohl m als auch n gerade, was im Widerspruch zu unserer Annahme des vollständig gekürzten Bruchs steht.

AXIOMATISCHE BESCHREIBUNG DER REELLEN ZAHLEN

Um die Lücken zu schließen, suchen wir eine Menge $\mathbb{R}$ mit

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R},$$

so dass $\mathbb{R}$ ein geordneter Körper ist, der keine “Lücken mehr enthält”. Die Eigenschaften eines geordneten Körpers garantieren uns dabei, dass wir wie gewohnt rechnen und vergleichen zu können.

Dazu könnte man versuchen, Lösungen von “algebraischen” Gleichungen wie die obige zu den rationalen Zahlen “hinzuzufügen”. Da dies aber immer noch nicht Zahlen wie π oder e erzeugen würde, werden wir einen anderen Ansatz verfolgen, wobei wir aber die eigentliche Konstruktion aus Zeitgründen nicht ausführen, sondern nur die gewünschten Eigenschaften von $\mathbb{R}$ aufzählen.

AXIOMATISCHE BESCHREIBUNG DER REELLEN ZAHLEN

Die folgenden beiden Axiome fassen unsere erste Forderung an $\mathbb{R}$ zusammen.

Körperaxiome. Es gibt eine Addition $+$ und eine Multiplikation $\cdot$ auf $\mathbb{R}$, so dass $(\mathbb{R}, +, \cdot)$ ein Körper ist.

Ordnungsaxiome. Es gibt eine Totalordnung $\leq$ auf $\mathbb{R}$, so dass $(\mathbb{R}, +, \cdot, \leq)$ ein geordneter Körper ist.

Man kann zeigen, dass jeder geordnete Körper K die rationalen Zahlen $\mathbb{Q}$ "enthält". Dies ist nicht ganz überraschend, denn $\mathbb{Q}$ ist ja durch, in gewissen Sinne minimale, Ergänzungen aus $\mathbb{N}$ entstanden und da $1 < 1+1 < \dots$ in jedem geordneten Körper K gilt, können wir $\mathbb{N}$ in K "wiederfinden".

Da $\mathbb{Q}$ auch ein geordneter Körper ist, schließen diese Axiome alleine aber noch keine Lücke. Dies soll nun unser nächstes Ziel sein.

Definition 2.4.2

Sei K ein geordneter Körper, $M \subset K$ und $b \in K$. Dann heißt:

i). M **nach oben beschränkt**, falls es ein $b' \in K$ gibt mit $x \leq b'$ für alle $x \in M$. In diesem Fall heißt b' eine **obere Schranke** von M.

ii). b **Supremum** von M, geschrieben $\sup M := b$, falls b eine obere Schranke von M ist und für jede obere Schranke b' von M gilt $b \leq b'$.

Mit anderen Worten kann M höchsten ein Supremum besitzen und falls dieses existiert, ist es gleich der **kleinsten oberen Schranke** von M.

Die Existenz von Suprema in $\mathbb{Q}$ ist nicht automatisch erfüllt, wie das folgende Lemma zeigt.

Lemma 2.4.3

Die Menge $M := \{x \in \mathbb{Q} : x^2 \leq 2\}$ ist nichtleer und nach oben beschränkt, hat aber kein Supremum in $\mathbb{Q}$.

BEWEIS

Offensichtlich gilt $1 \in M$, d.h. M ist nichtleer. Für jedes $x \in M$ gilt ferner $x \leq 2$, denn wäre $x > 2$ wahr, so würde $x^2 = x \cdot x > 2 \cdot 2 > 2$ gelten. Damit ist M nach oben beschränkt.

Als nächstes wollen wir zeigen, dass jede obere Schranke b von M die Ungleichung $b^2 \geq 2$ erfüllt. Dazu nehmen wir $b^2 < 2$ an. Für $\varepsilon := 2 - b^2 > 0$ gibt es dann nach Satz 2.3.4 ein $n \in \mathbb{N}$ mit $3b/n < \varepsilon$. Wir definieren $x := b + 1/n$. Dann gilt

$$x^2 = b^2 + \frac{2b}{n} + \frac{1}{n^2} \leq b^2 + \frac{3b}{n} < b^2 + \varepsilon = 2\,,$$

d.h. $x \in M$. Wegen $x = b + 1/n > b$ ist dann aber b keine obere Schranke von M. Die Annahme $b^2 < 2$ ist daher falsch.

BEWEIS

Schließlich zeigen wir, dass M kein Supremum in $\mathbb{Q}$ hat. Wir nehmen dazu an, dass das Supremum $b := \sup M$ in $\mathbb{Q}$ existiert, d.h. $b \in \mathbb{Q}$. Da dann b eine obere Schranke ist, wissen wir $b^2 \geq 2$ aus unserer Vorüberlegung. Ferner ist $b^2 = 2$ nach Lemma 2.4.1 unmöglich ist, und daher gilt sogar $b^2 > 2$. Wir setzen nun

$$b' := b - \frac{b^2 - 2}{2b}.$$

Wegen $1 \in M$ und $0 < 1 \leq b$ gilt dann $b' < b$, und aus $b^2 > 0 > -2$ können wir wegen

$$b^2 > -2 \quad \Leftrightarrow \quad 2b^2 > b^2 - 2 \quad \Leftrightarrow \quad b > \frac{b^2 - 2}{2b} \quad \Leftrightarrow \quad b - \frac{b^2 - 2}{2b} > 0$$

auf $b' > 0$ schließen. Ferner haben wir

$$(b')^2 = \left(b - \frac{b^2 - 2}{2b}\right)^2 = b^2 - 2b \cdot \frac{b^2 - 2}{2b} + \left(\frac{b^2 - 2}{2b}\right)^2 = 2 + \left(\frac{b^2 - 2}{2b}\right)^2 > 2$$

und damit folgt $x^2 < (b')^2$ für alle $x \in M$. Dies impliziert $x < b'$, denn $x \geq b'$ würde wegen $b' > 0$ die Ungleichung $x^2 \geq (b')^2$ erzwingen. Mit anderen ist b' eine obere Schranke von M mit $b' < b$, und damit kann b kein Supremum von M sein.

Aufgrund der obigen Beobachtung liegt es nun nahe, die Lücken durch die Existenz von Suprema zu erzwingen. Dies ist die Motivation für das folgende, letzte Axiom für $\mathbb{R}$.

Supremumsaxiom. Jede nichtleere, nach oben beschränkte Teilmenge der reellen Zahlen $\mathbb{R}$ besitzt ein Supremum.

Mit dem Supremumsaxiom finden wir in $\mathbb{R}$ nun tatsächlich ein $b \in \mathbb{R}$ mit $b^2 = 2$. Sei dazu $b := \sup M$, wobei M die Menge aus Lemma 2.4.3 ist. In dem Beweis von Lemma 2.4.3 haben wir dann zunächst $b^2 \geq 2$ gesehen. Ferner führte $b^2 > 2$ durch die Konstruktion eines b' zu einem Widerspruch und damit gilt in der Tat $b^2 = 2$.

Die axiomatische Einführung von $\mathbb{R}$ erfordert natürlich einen Existenzbeweis, den wir aber hier nicht durchführen wollen. Das Gleiche gilt für die "Eindeutigkeit", vgl. dazu auch Abschnitt 53. Schließlich kann man auch zeigen, dass $\mathbb{Q} \subset \mathbb{R}$ gilt und dass $\mathbb{R}$ keine Lücken mehr besitzt, so dass unser anfängliches Problem tatsächlich gelöst ist.

Der folgende Satz zeigt, dass das Archimedische Axiom in $\mathbb{R}$ erfüllt ist.

Theorem 2.4.4

Für alle $x, y \in \mathbb{R}$ mit $x > 0$ und $y > 0$ existiert ein $n \in \mathbb{N}$ mit $nx > y$.

Beweis.

Ist $y \leq x$, so können wir $k = 2$ wählen. Es reicht daher den Fall $x < y$ zu betrachten.
Wir nehmen nun an, dass es $0 < x < y$ gibt, so dass $nx \leq y$ für alle $n \in \mathbb{N}$ gilt. Die Menge

$$M := \{nx : n \in \mathbb{N}\}$$

ist dann beschränkt und daher existiert $b := \sup M$. Da b eine obere Schranke von M ist, folgt $nx \leq b$ für alle $n \in \mathbb{N}$ und damit auch $(n+1)x \leq b$ für alle $n \in \mathbb{N}$. Dies impliziert $nx \leq b - x$ für alle $n \in \mathbb{N}$, d.h. $b' := b - x$ ist eine obere Schranke von M. Wegen $x > 0$ gilt aber $b' = b - x < b$, was $b = \sup M$ widerspricht. □

Im folgenden Schreiben wir $\sup \emptyset := -\infty$. Ist ferner $M \subset \mathbb{R}$ nicht beschränkt, so definieren wir $\sup M := \infty$. Damit hat jede Teilmenge M von $\mathbb{R}$ ein Supremum.

Für $M \subset \mathbb{R}$ ist ferner das Infimum von M durch

$$\inf M := -\sup(-M) \tag{2.4.1}$$

definiert. Hierbei ist $-M := \{-x : x \in M\}$ und $-(-\infty) := \infty$. Analog sagen wir, dass M nach unten beschränkt ist, falls $-M$ nach oben beschränkt ist und in diesem Fall ist $-b$ eine untere Schranke von M, falls b eine obere Schranke von $-M$ ist. Insbesondere ist dann $\inf M$ die größte untere Schranke von M.

Gilt $\sup M \in M$, so schreiben wir $\max M := \sup M$ und im Fall $\inf M \in M$ schreiben wir analog $\min M := \inf M$.

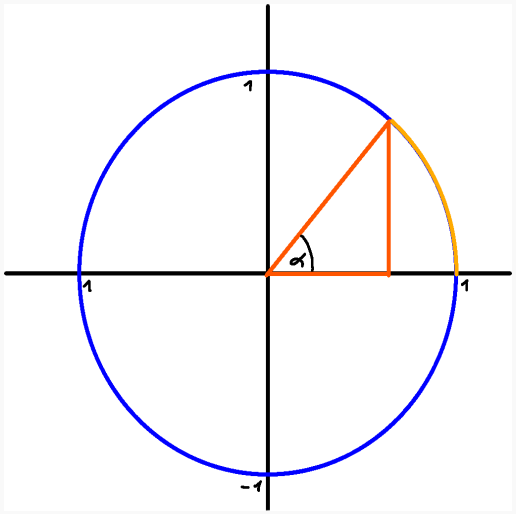
Für die Menge $M := \{1 + 1/n : n \in \mathbb{N}\}$ gilt beispielsweise $\inf M = 1$ und $\inf M \notin M$.

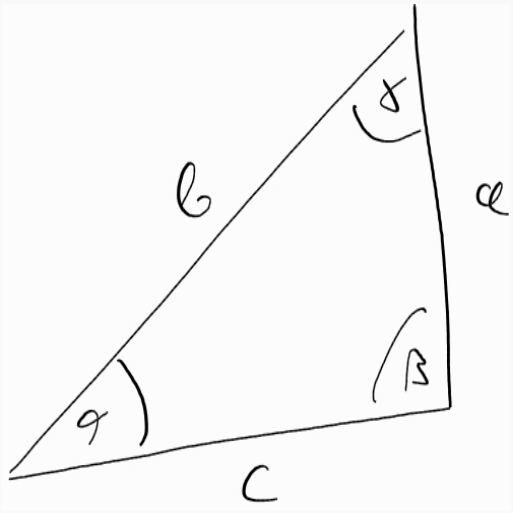
In der Tat ist 1 eine untere Schranke von M und gäbe es eine untere Schranke $b > 1$ von M, so würden wir mit Satz 2.4.4 ein $n \in \mathbb{N}$ mit $\frac{1}{n} < b - 1$ finden. Dies ergäbe $1 + 1/n < b$, d.h. b kann keine untere Schranke sein. Insgesamt haben wir also $\inf M = 1$ gesehen und $1 \notin M$ folgt aus den Körperaxiomen.

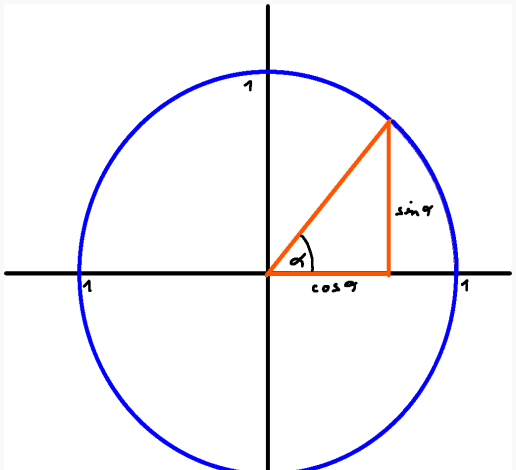
Section 2.5
Winkelfunktionen

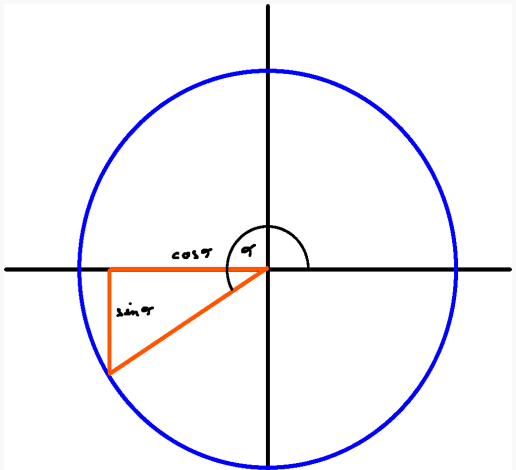
DEFINITION SINUS UND KOSINUS

Winkel messen wir im folgenden im **Bogenmaß**. Die Größe eines Winkels α entspricht damit der Länge des Kreisbogenabschnitts zum Radius 1, der vom Winkel α bestimmt wird. Ein rechter Winkel entspricht damit $\pi/2$, ein Vollkreis 2π, siehe Abbildung 5.









DEFINITION SINUS UND KOSINUS

Betrachten wir ein rechtwinkeliges Dreieck mit den Winkeln, α, $\beta := \pi/2$ und γ, die aufsteigend gegen den Uhrzeigersinn beschriftet sind, und benennen wir die jeweils gegenüberliegenden Seiten mit a, b und c, siehe wieder Abbildung 5, so gilt

$$\cos\alpha := \frac{c}{b} \qquad \text{und} \qquad \sin\alpha := \frac{a}{b}\,.$$

Für $c = 1$ vereinfachen sich die Formeln auf natürliche Weise und eine geometrische Interpretation am Einheitskreis ist möglich, siehe Abbildung 5. Diese erlaubt es auch, $\alpha > \pi/2$ zu betrachten, siehe wieder Abbildung 5.

Einige spezielle Werte der Winkelfunktionen lassen sich direkt aus speziell gewählten Dreiecken ablesen, siehe Abbildung 6.

	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
cos	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
sin	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1

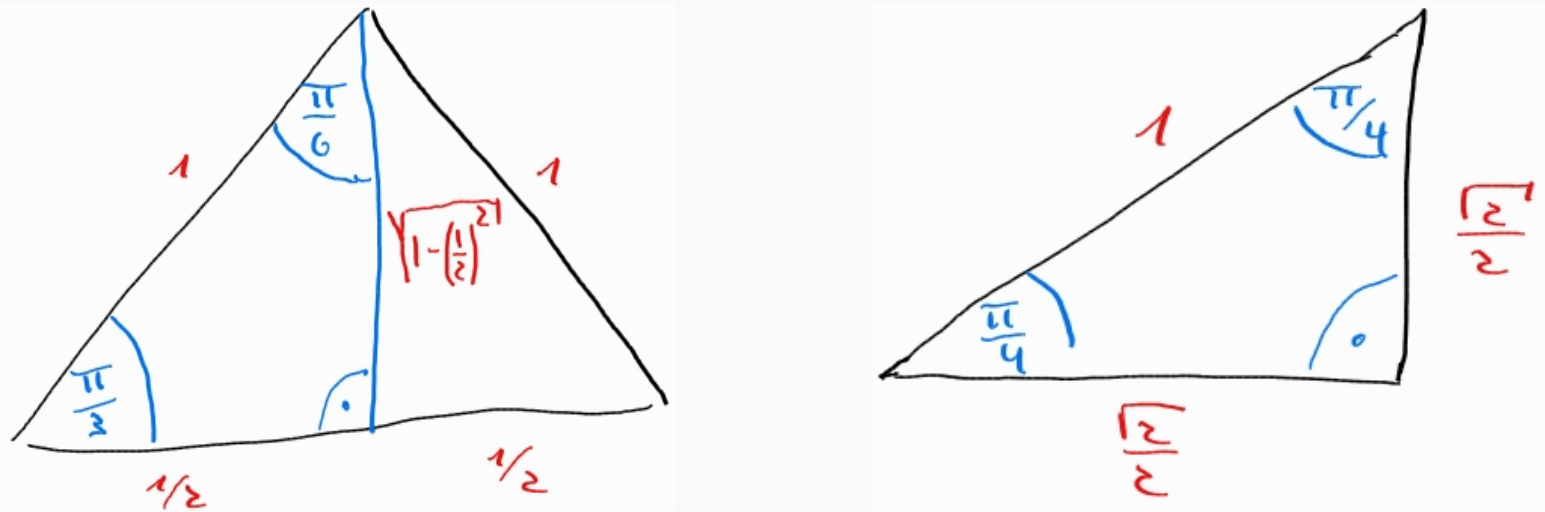


Abbildung: Links: Geometrische Herleitung von **sin** $\frac{\pi}{6}$ an einem gleichseitigen Dreieck der Seitenlänge 1. **Rechts:** Geometrische Herleitung von **sin** $\frac{\pi}{4}$ an einem gleichschenkeligem Dreieck mit rechten Winkel. In beiden Fällen wird der Satz von Pythagoras angewendet.

Mit der geometrischen Interpretation am Einheitskreis und dem Satz des Pythagoras ergibt sich sofort das folgende Lemma.

Lemma 2.5.1

Für alle $\alpha \in \mathbb{R}$ gilt

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

wobei wir die Kurzzschreibweisen $\sin^2 \alpha := (\sin \alpha)^2$ und $\cos^2 \alpha = (\cos \alpha)^2$ verwenden.

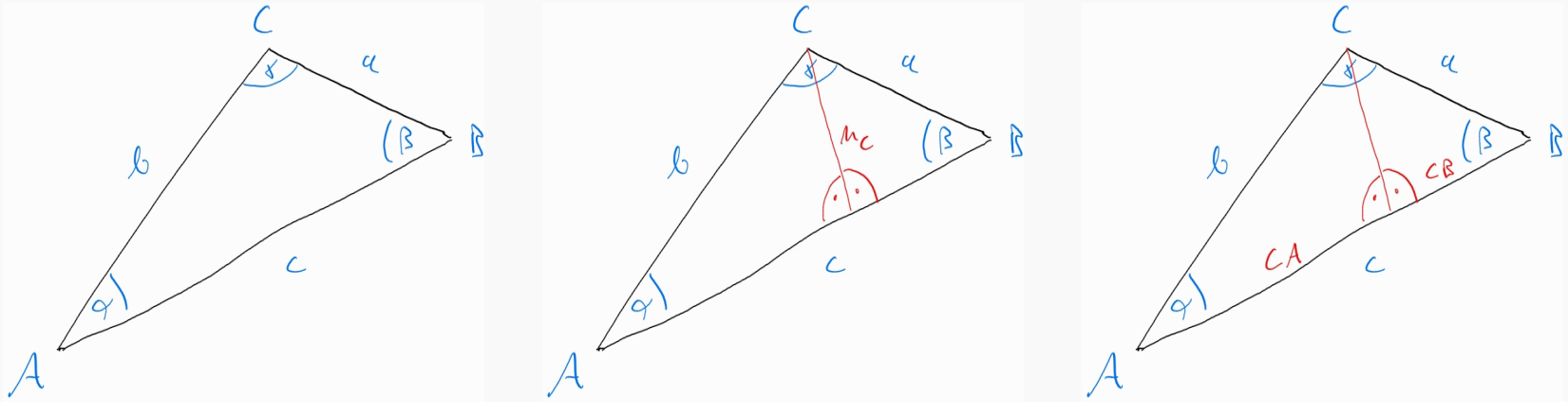


Abbildung: Links: Ein allgemeines Dreieck mit Seiten-, Ecken- und Winkelbezeichnungen **Mitte:** Geometrischer Ansatz für den Sinussatz. **Rechts:** Geometrischer Ansatz für den Kosinussatz.

In den folgenden beiden Sätzen betrachten wir allgemeine Dreiecke, wobei die Seiten-, Ecken- und Winkelbezeichnungen der Abbildung 7 zu entnehmen sind.

Theorem 2.5.2

In jedem Dreieck gilt

$$\frac{\sin \alpha}{a} = \frac{\sin \beta}{b} = \frac{\sin \gamma}{c}\,.$$

Beweis.

Die Höhe h_c zur Grundseite c, siehe Abbildung 7, berechnet sich sowohl durch $h_c = b \sin \alpha$ als auch durch $h_c = a \sin \beta$. Gleichsetzen und Umstellen liefert die erste Gleichheit. Die zweite folgt durch zyklisches Vertauschen der Bezeichnungen. □

Theorem 2.5.3

In jedem Dreieck gilt

$$a^2 = b^2 + c^2 - 2bc\cos\alpha,$$
$$b^2 = c^2 + a^2 - 2ca\cos\beta,$$
$$c^2 = a^2 + b^2 - 2ab\cos\gamma.$$

Beweis.

Seien c_A und c_B für die beiden Abschnitte der Seite c, die durch h_c entstehen, siehe Abbildung 7. Damit haben wir $c_A + c_B = c$. Mit dem Satz des Pythagoras und $c_A = b\cos\alpha$ gilt dann

$$\begin{aligned} a^2 = h_c^2 + c_B^2 &= b^2\sin^2\alpha + (c - b\cos\alpha)^2 \\ &= b^2(\cos^2\alpha + \sin^2\alpha) + c^2 - 2bc\cos\alpha \\ &= b^2 + c^2 - 2bc\cos\alpha. \end{aligned}$$

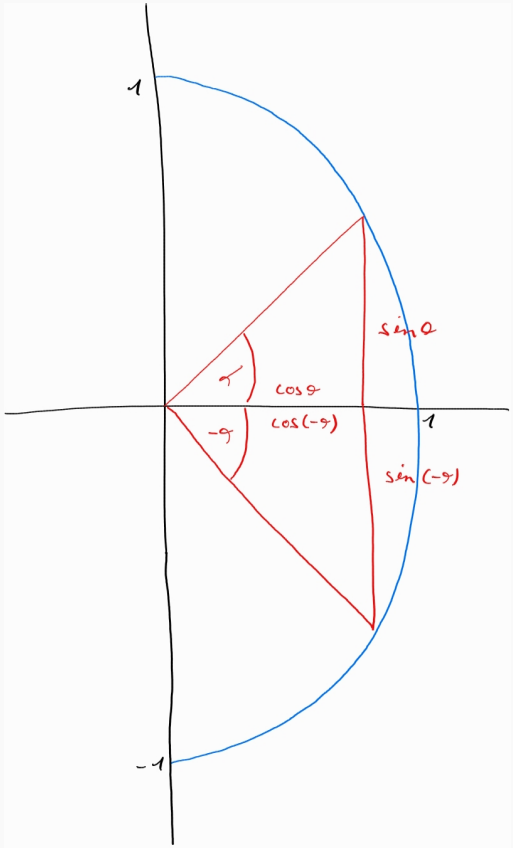
Die weiteren Identitäten folgen durch zyklisches Vertauschen der Bezeichnungen. □

Das folgende Lemma fasst ein paar einfache Formeln für die Sinus- und Kosinus-Funktion zusammen.

Lemma 2.5.4

Sei $\alpha \in \mathbb{R}$. Dann gilt:

$$\sin(-\alpha) = -\sin\alpha \quad \textit{und} \quad \cos(-\alpha) = \cos\alpha\,,$$
$$\sin(\alpha+\pi/2) = \cos\alpha \quad \textit{und} \quad \cos(\alpha+\pi/2) = -\sin\alpha\,,$$
$$\sin(\alpha+\pi) = -\sin\alpha \quad \textit{und} \quad \cos(\alpha+\pi) = -\cos\alpha\,,$$
$$\sin(\alpha+2\pi) = \sin\alpha \quad \textit{und} \quad \cos(\alpha+2\pi) = \cos\alpha\,.$$



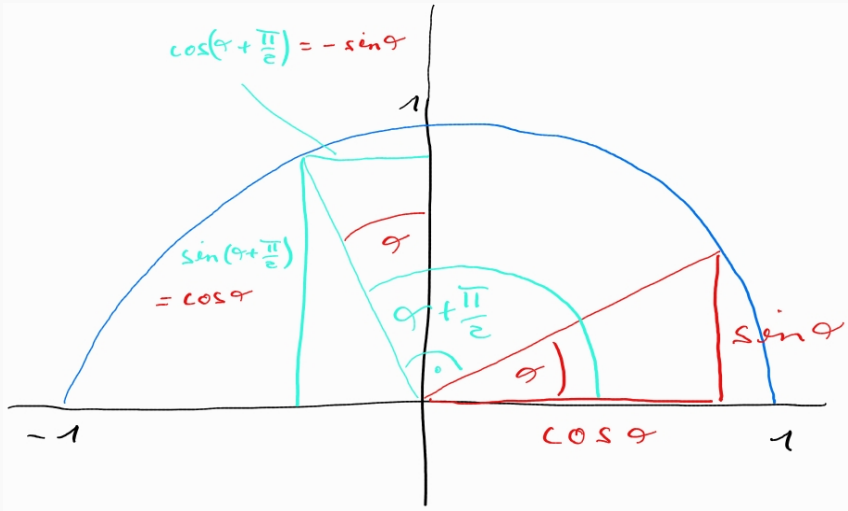


Abbildung: Links: Dreiecke mit Winkel α und α am Einheitskreis. **Rechts:** Drehen des Dreiecks mit Winkel α um $\pi/2$.

Die ersten beiden Zeilen lassen sich leicht am Einheitskreis ablesen, siehe abbildung 8. Die dritte und vierte Zeilen folgen aus der zweiten Zeile durch 2- bzw. viermaliges Anwenden.

Theorem 2.5.5

Für alle $\alpha, \beta \in \mathbb{R}$ gilt:

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta\,,$$
$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta\,.$$

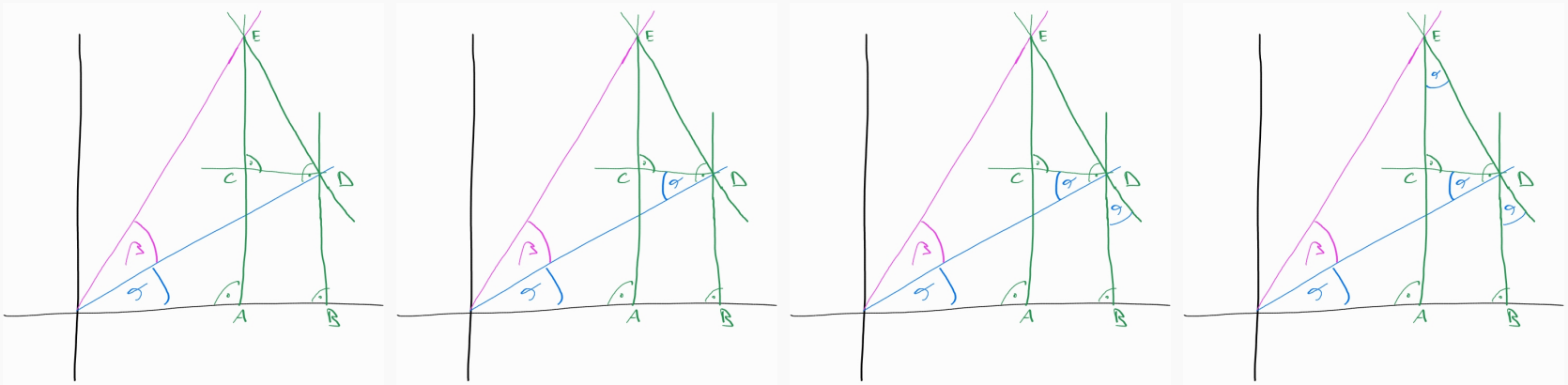


Abbildung: Links: Die Winkel α und β zusammen mit den senkrechten Strecken $\overline{AE}$ und $\overline{BD}$, sowie der waagerechten Strecke $\overline{CD}$. Nachdem die Strahlen in Blau und Magenta gezeichnet sind, wird ein Punkt E fixiert. Dann werden zunächst die Strecken $\overline{AE}$ und $\overline{DE}$ konstruiert. Im Anschluss werden noch die Strecken $\overline{CD}$ und $\overline{BD}$ konstruiert. **Mitte-Links:** Identifikation des Winkels α zwischen den Strecken $\overline{CD}$ und $\overline{0D}$. **Mitte-Rechts:** Identifikation des Winkels α zwischen den Strecken $\overline{BD}$ und der Verlängerung von $\overline{DE}$ nach unten. **Rechts:** Identifikation des Winkels α zwischen den Strecken $\overline{AE}$ und $\overline{DE}$.

BEWEIS

Wir betrachten die Abbildung 9 und bezeichnen die Länge einer Strecke $\overline{XY}$ mit $|\overline{XY}|$. Mit $|\overline{AE}| = |\overline{AC}| + |\overline{CE}|$ und $|\overline{AC}| = |\overline{BD}|$ folgt dann:

$$\begin{aligned}\sin(\alpha+\beta) = \frac{|\overline{AE}|}{|\overline{0E}|} &= \frac{|\overline{BD}| + |\overline{CE}|}{|\overline{0E}|} \\ &= \frac{|\overline{BD}| \cdot |\overline{0D}|}{|\overline{0D}| \cdot |\overline{0E}|} + \frac{|\overline{CE}| \cdot |\overline{ED}|}{|\overline{ED}| \cdot |\overline{0E}|} \\ &= \sin\alpha \cdot \cos\beta + \cos\alpha \cdot \beta\end{aligned}$$

Die restlichen Formeln können analog bewiesen werden.

Section 2.6
Komplexe Zahlen

REELLE ZAHLEN KÖNNEN NICHT ALLES AUSDRÜCKEN

Obwohl wir durch die Hinzunahme von Suprema die rationalen Zahlen $\mathbb{Q}$ erheblich erweitert haben, können gewisse Gleichungen in den resultierenden reellen Zahlen $\mathbb{R}$ nicht gelöst werden. Beispielsweise hat die Gleichung

$$x^2 = -1$$

in $\mathbb{R}$ keine Lösung, da wir einerseits $x^2 \geq 0$ und andererseits $-1 < 0$ haben. Da die beiden letzten Ungleichungen in jedem geordneten Körper gelten, besteht auch keine Hoffnung, einen noch größeren, geordneten Körper $\mathbb{F}$ zu finden, in dem $x^2 = -1$ eine Lösung hat. Wollen wir die Körperaxiome nicht aufgeben, so müssen wir daher zwangsläufig die Ordnungsaxiome aufgeben, um die Gleichung lösen zu können.

REELLE ZAHLEN KÖNNEN NICHT ALLES AUSDRÜCKEN

Nehmen wir nun an, dass wir einen Körper $\mathbb{F}$ haben, in dem es eine solche Lösung $\mathrm{i} \in \mathbb{F}$ gibt, d.h. $\mathrm{i}^2 = -1$. Wegen der Körperaxiome müssen dann für alle $x_1, x_2, y_1, y_2 \in \mathbb{F}$ die beiden Gleichungen

$$(x_1 + \mathrm{i}y_1) + (x_2 + \mathrm{i}y_2) = (x_1 + x_2) + \mathrm{i}(y_1 + y_2)$$

und

$$\begin{aligned}(x_1 + \mathrm{i}y_1)(x_2 + \mathrm{i}y_2) &= x_1x_2 + \mathrm{i}x_1y_2 + \mathrm{i}y_1x_2 - y_1y_2 \\ &= (x_1x_2 - y_1y_2) + \mathrm{i}(x_1y_2 + y_1x_2)\end{aligned}$$

gelten. Zudem wollen wir natürlich $\mathbb{R}$ in $\mathbb{F}$ wiederfinden können, d.h. wir wollen $\mathbb{R} \subset \mathbb{F}$ haben. Die komplexen Zahlen $\mathbb{C}$ werden dieses erfüllen.

Die obigen Gleichungen motivieren dazu, Elemente in $\mathbb{F}$ als Zahlenpaare zu definieren. Die folgende Definition greift diesen Gedanken auf.

Definition 2.6.1

Die Menge $\mathbb{C} := \mathbb{R}^2$ zusammen mit der Addition $+ : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ und Multiplikation $\cdot : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$, die für alle $(x_1, y_1) \in \mathbb{R}^2$ und $(x_2, y_2) \in \mathbb{R}^2$ durch

$$(x_1, y_1) + (x_2, y_2) := (x_1 + x_2, y_1 + y_2)$$
$$(x_1, y_1) \cdot (x_2, y_2) := (x_1 x_2 - y_1 y_2, x_1 y_2 + x_2 y_1)$$

definiert sind, heißt Menge der komplexe Zahlen.

Mit fleißigem Rechnen stellt sich heraus, dass $(\mathbb{C}, +, \cdot)$ ein Körper ist, wobei seine Null durch $(0,0)$ und seine Eins durch $(1,0)$ gegeben sind. Für $(x,y) \in \mathbb{C}$ gilt ferner $-(x,y) = (-x,-y)$ und

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

wobei letzteres natürlich nur im Fall $(x,y) \neq (0,0)$ definiert ist. Ferner gilt $(0,1)\cdot(0,1) = -(1,0)$, d.h. die imaginäre Einheit $\mathrm{i} := (0,1)$ löst tatsächlich die Gleichung $z^2 = -1$. Schließlich liefert die obige Formel $\mathrm{i}^{-1} = -\mathrm{i}$.

Werden die Addition und Multiplikation von $\mathbb{C}$ auf $\mathbf{R} := \{(x,0) : x \in \mathbb{R}\}$ eingeschränkt, so erhalten wir wieder einen Körper, der zudem eine Totalordnung

$$(x_1,0) \leq (x_2,0) \quad :\Longleftrightarrow \quad x_1 \leq x_2$$

hat, mit der er zu einem geordneten Körper wird. Ferner erfüllt diese Totalordnung das Supremumsaxiom, und damit können wir die Menge $\mathbf{R}$ als ein weiteres Modell von $\mathbb{R}$ betrachten. Mit dieser Identifizierung haben wir also $\mathbb{R} \subset \mathbb{C}$ und

$$x + \mathrm{i}y := (x,y), \qquad x,y \in \mathbb{R},$$

ergibt eine verträgliche Notation.

Gauss'sche Zahlenebene

Nach unserer Konstruktion kann $\mathbb{C}$ als **Zahlenebene** gedacht werden, deren x-Achse die reellen Zahlen bildet. Diese Achse heißt aus diesem Grund häufig auch **reelle Achse**. Senkrecht dazu steht die y-Achse $\{(0, y) : y \in \mathbb{R}\}$ die auch als **imaginäre Achse** bezeichnet wird.

Die Addition in $\mathbb{C}$ entspricht offensichtlich der Vektoraddition in $\mathbb{R}^2$ und eine geometrische Interpretation der Multiplikation werden nach den folgenden Definitionen kennenlernen.

Für $z := (x, y) \in \mathbb{C}$ bezeichnen wir die Koordinaten

$$\operatorname{Re} z := x, \qquad \operatorname{Im} z := y$$

als **Real- und Imaginärteil** der Zahl z. Die Zerlegung $z = x + \mathrm{i}y$ wird als **kartesische Darstellung** der Zahl z bezeichnet. Der **Betrag**

$$|z| := \sqrt{x^2 + y^2}$$

entspricht dem Abstand von z zum Ursprung.

MULTIPLIKATION IN POLARKOORDINATEN

Für $z = (x, y) \in \mathbb{C}$ sei nun $r := |z|$ und φ der Winkel zwischen der x-Achse und der Strecke $\overline{0z}$. Dann gilt die Polar-Darstellung von z:

$$z = r(\cos\varphi + \mathrm{i}\sin\varphi)\,.$$

Haben wir nun zwei komplexe Zahlen z_1 und z_2 in Polar-Darstellung (r_1, φ_1) und (r_2, φ_2), so gilt

$$\begin{aligned}
&z_1 \cdot z_2 \\
&= r_1(\cos\varphi_1 + \mathrm{i}\sin\varphi_1) \cdot r_2(\cos\varphi_2 + \mathrm{i}\sin\varphi_2) \\
&= r_1 \cdot r_2\Big((\cos\varphi_1 \cdot \cos\varphi_2 - \sin\varphi_1 \cdot \sin\varphi_2) + \mathrm{i}(\cos\varphi_1 \cdot \sin\varphi_2 + \sin\varphi_1 \cdot \cos\varphi_2)\Big) \\
&= r_1 \cdot r_2\Big(\cos(\varphi_1 + \varphi_2) + \mathrm{i}\sin(\varphi_1 + \varphi_2)\Big)\,, \qquad (2.6.1)
\end{aligned}$$

wobei wir im letzten Schritt die Additionstheoreme aus Satz 2.5.5 benutzt haben. Mit anderen Worten: Die Beträge werden multipliziert und die Winkel werden addiert.

Mit Hilfe der Formel (2.6.1) und Induktion über $n \in \mathbb{N}$ lässt sich nun der folgende Satz, der als **Formel von Moivre** bekannt ist, leicht beweisen.

Theorem 2.6.2

Sei $z = r(\cos\varphi + \mathrm{i}\sin\varphi) \in \mathbb{C}$. Dann gilt für alle $n \in \mathbb{N}$:

$$z^n = r^n \cdot \big(\cos(n\varphi) + \mathrm{i}\sin(n\varphi)\big)\,.$$

Im folgenden definieren wir für $\varphi \in \mathbb{R}$

$$\mathrm{e}^{\mathrm{i}\varphi} := \cos\varphi + \mathrm{i}\sin\varphi\,.$$

Für $\varphi, \varphi_1, \varphi_2 \in \mathbb{R}$ und $n \in \mathbb{N}$ gelten dann die Formeln:

$$\mathrm{e}^{\mathrm{i}\cdot 0} = 1\,, \tag{2.6.2}$$

$$\mathrm{e}^{\mathrm{i}(\varphi_1+\varphi_2)} = \mathrm{e}^{\mathrm{i}\varphi_1}\cdot\mathrm{e}^{\mathrm{i}\varphi_2}\,, \tag{2.6.3}$$

$$\left(\mathrm{e}^{\mathrm{i}\varphi}\right)^n = \mathrm{e}^{\mathrm{i}n\varphi}\,, \tag{2.6.4}$$

$$\mathrm{e}^{-\mathrm{i}\varphi} = \left(\mathrm{e}^{\mathrm{i}\varphi}\right)^{-1}. \tag{2.6.5}$$

Hierbei ist die erste Formel leicht abzulesen und die zweite folgt aus (2.6.1). Die dritte Formel folgt sofort aus der Formel von Moivre und die vierte Formel ist eine einfache Konsequenz aus den ersten beiden Formeln.

Theorem 2.6.3

Es gibt genau n komplexe Zahlen z, die die Gleichung $z^n = 1$ erfüllen. Diese Zahlen sind die sogenannten ***n-ten Einheitswurzeln***

$$z_k := \mathrm{e}^{\mathrm{i}\cdot\frac{2\pi k}{n}}, \qquad k = 0, \ldots, n-1.$$

Bevor wir diesen Satz beweisen, beachten wir, dass die Zahlen

$$0, \frac{2\pi\cdot 1}{n}, \frac{2\pi 2}{n}, \ldots, \frac{2\pi(n-1)}{n}$$

einen ganzen Einheitskreisumfang 2π in n gleich großen Schritte zerlegen. Entsprechend zerteilen die Einheitswurzeln $z_0, \ldots, z_{n-1}$ den Einheitskreis auf der Gauß'schen Zahlenebene und insbesondere gilt $z_i \neq z_j$ für $i \neq j$.

Aus der eben bewiesenen Formel $(\mathrm{e}^{\mathrm{i}\varphi})^n = \mathrm{e}^{\mathrm{i}n\varphi}$ folgt sofort

$$z_k^n = \mathrm{e}^{\mathrm{i}\cdot 2\pi k} = \cos(2\pi k) + \mathrm{i}\sin(2\pi k) = 1\,.$$

Wir müssen daher nur noch zeigen, dass es keine anderen Lösungen der Gleichung gibt. Einen einfachen Beweis werden wir mit Korollar 2.7.6 finden können. Hier werden wir einen etwas geometrischer orientierten Beweis vorstellen.

Sei nun $z \in \mathbb{C}$ mit $z^n = 1$. Wegen der Formel von Moivre wissen wir dann, dass

$$1 = z^n = \cos(n\varphi) + \mathrm{i}\sin(n\varphi)$$

und $|z| = 1$ gilt. Aufgrund der Form der Sinus- oder Kosinus-Funktion ergibt dies $n\varphi = 2\pi m$ für ein $m \in \mathbb{Z}$, d.h. wir haben $z = \mathrm{e}^{\mathrm{i}\cdot\frac{2\pi m}{n}}$. Durch Division mit Rest gibt es dann $l \in \mathbb{Z}$ und $k \in \{0, 1, \ldots, n-1\}$ mit $ln + k = m$. Da dies

$$z = \mathrm{e}^{\mathrm{i}\cdot\frac{2\pi(ln+k)}{n}} = \mathrm{e}^{\mathrm{i}\cdot 2\pi l} \cdot \mathrm{e}^{\mathrm{i}\cdot\frac{2\pi k}{n}} = 1 \cdot z_k$$

ergibt, ist der Beweis dann beendet.

Für eine komplexe Zahl $z = x + \mathrm{i}y$ ist es sinnvoll die dazu **komplex konjugierte Zahl**

$$\bar{z} := x - \mathrm{i}y$$

zu betrachten. Es gilt

$$\overline{z+w} = \bar{z} + \overline{w} \quad \text{und} \quad \overline{zw} = \bar{z}\,\overline{w}.$$

In der Tat ist die Formel für die Addition sofort ersichtlich. Wir rechnen also noch die Produktformel nach. Seien dazu $z = x + \mathrm{i}y$ und $w = u + \mathrm{i}v$. Dann folgt einerseits

$$\overline{zw} = \overline{(x+\mathrm{i}y)(u+\mathrm{i}v)} = \overline{xu - yv + \mathrm{i}(xv+yu)} = xu - yv - \mathrm{i}(xv+uy)$$

und andererseits

$$\bar{z}\,\overline{w} = \overline{x+\mathrm{i}y}\,\overline{u+\mathrm{i}v} = (x-\mathrm{i}y)(u-\mathrm{i}v) = xu - yv - \mathrm{i}(xv+yu).$$

Mit Hilfe der komplex konjugierten Zahl kann der Betrag berechnet werden, denn mit der dritten binomischen Formel und $\mathrm{i}^2 = -1$ gilt

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

Damit können einfache Formeln nachgerechnet werden. So gilt

$$|zw| = |z|\,|w|$$

da

$$|zw|^2 = zw\overline{zw} = z\bar{z}w\overline{w} = |z|^2|w|^2.$$

Beträge sind also multiplikativ und wegen $|1| = 1$ gilt insbesondere $|z^{-1}| = |z|^{-1}$ für alle $z \in \mathbb{C} \setminus \{0\}$.

EXPONENTIALFUNKTION IM KOMPLEXEN

Im folgenden gehen wir davon aus, dass die Exponentialfunktion $x \mapsto \mathrm{e}^x$ für $x \in \mathbb{R}$ bekannt ist. Insbesondere nehmen wir an, dass wir die Gleichung

$$\mathrm{e}^{x_1+x_2} = \mathrm{e}^{x_1} \cdot \mathrm{e}^{x_2}\,, \qquad x_1, x_2 \in \mathbb{R} \tag{2.6.6}$$

kennen. Wir erweitern die Definition nun auf komplexe Zahlen $z := x + \mathrm{i}y$ durch

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

Mit anderen Worten gilt

$$\mathrm{e}^z = \mathrm{e}^{\mathrm{Re}\, z} \big(\cos(\mathrm{Im}\, z) + \mathrm{i} \sin(\mathrm{Im}\, z) \big)\,.$$

Seien nun $z = x + \mathrm{i}y$ und $w = u + \mathrm{i}v$ zwei komplexe Zahlen. Mit der Definition der Exponentialfunktion im Komplexen, der Gleichung (2.6.6) im Reellen und (2.6.3) erhalten wir dann

$$\mathrm{e}^{z+w} = \mathrm{e}^{x+u+\mathrm{i}(y+v)} = \mathrm{e}^{x+u} \cdot \mathrm{e}^{\mathrm{i}(y+v)} = \mathrm{e}^{x}\mathrm{e}^{u}\mathrm{e}^{\mathrm{i}y}\mathrm{e}^{\mathrm{i}v} = \mathrm{e}^{z}\mathrm{e}^{w}. \tag{2.6.7}$$

Mit anderen Worten gilt (2.6.6) nicht nur in $\mathbb{R}$ sondern auch in $\mathbb{C}$. Ebenso gilt nach Definition

$$|\mathrm{e}^{z}| = \mathrm{e}^{\mathrm{Re}\, z}$$

und aus (2.6.7) folgt

$$\mathrm{e}^{0} = 1, \qquad \text{sowie} \qquad \mathrm{e}^{-z} = \frac{1}{\mathrm{e}^{z}}.$$

Wir können die gerade definierte Exponentialfunktion nutzen, um damit die Winkelfunktionen auszudrücken. Wegen

$$\begin{aligned} \mathrm{e}^{\mathrm{i}\varphi} &= \cos\varphi + \mathrm{i}\sin\varphi, \\ \mathrm{e}^{-\mathrm{i}\varphi} &= \cos\varphi - \mathrm{i}\sin\varphi \end{aligned}$$

ergeben sich $\cos\varphi = \frac{1}{2}\left(\mathrm{e}^{\mathrm{i}\varphi} + \mathrm{e}^{-\mathrm{i}\varphi}\right)$ und $\sin\varphi = \frac{1}{2\mathrm{i}}\left(\mathrm{e}^{\mathrm{i}\varphi} - \mathrm{e}^{-\mathrm{i}\varphi}\right)$ für alle $\varphi \in \mathbb{R}$. Dies motiviert die folgenden Definitionen für alle komplexen Zahlen $z \in \mathbb{C}$:

$$\begin{aligned} \cos z &:= \frac{1}{2}\left(\mathrm{e}^{\mathrm{i}z} + \mathrm{e}^{-\mathrm{i}z}\right), \\ \sin z &:= \frac{1}{2\mathrm{i}}\left(\mathrm{e}^{\mathrm{i}z} - \mathrm{e}^{-\mathrm{i}z}\right). \end{aligned}$$

Auch im Komplexen gilt dann

$$\sin z = 0 \quad \Leftrightarrow \quad z = k\pi \text{ für ein } k \in \mathbb{Z}$$

und

$$\cos z = 0 \quad \Leftrightarrow \quad z = \frac{\pi}{2} + k\pi \text{ für ein } k \in \mathbb{Z}.$$

Der Beweis erfolgt durch Nachrechnen, wir zeigen die Aussage für den Sinus. Es gilt

$$\sin z = 0 \quad \Leftrightarrow \quad \mathrm{e}^{\mathrm{i}z} = \mathrm{e}^{-\mathrm{i}z} \quad \Leftrightarrow \quad \left(\mathrm{e}^{\mathrm{i}z}\right)^2 = 1\,.$$

Damit ist $\mathrm{e}^{\mathrm{i}z}$ gleich einer der beiden zweiten Einheitswurzeln, d.h. $\mathrm{e}^{\mathrm{i}z} = \pm 1$. Wegen der Form dieser Einheitswurzeln, siehe Satz 2.6.3, schließen wir auf $z \in \mathbb{R}$ und somit folgt die Äquivalenz aus der entsprechenden Äquivalenz für die reelle Sinus-Funktion.

Wir definieren neben Sinus und Kosinus noch die beiden Funktionen
Tangens

$$\tan z = \frac{\sin z}{\cos z}, \qquad z \neq \frac{\pi}{2} + k\pi,\ k \in \mathbb{Z}$$

und Kotangens

$$\cot z = \frac{\cos z}{\sin z}, \qquad z \neq k\pi,\ k \in \mathbb{Z}.$$

Analog zu den Winkelfunktionen definiert man die **Hyperbelfunktionen**

$$\cosh z := \frac{1}{2}\left(\mathrm{e}^{z} + \mathrm{e}^{-z}\right),$$
$$\sinh z := \frac{1}{2}\left(\mathrm{e}^{z} - \mathrm{e}^{-z}\right),$$
$$\tanh z := \frac{\sinh z}{\cosh z} = \frac{\mathrm{e}^{z} - \mathrm{e}^{-z}}{\mathrm{e}^{z} + \mathrm{e}^{-z}},$$
$$\coth z := \frac{1}{\tanh z} = \frac{\mathrm{e}^{z} + \mathrm{e}^{-z}}{\mathrm{e}^{z} - \mathrm{e}^{-z}}.$$

Für $z \in \mathbb{C}$ gilt somit $\cosh z = \cos(\mathrm{i}z)$ und $\sinh z = -\mathrm{i}\sin(\mathrm{i}z)$. Betrachtet man aber nur die reellen Winkel- und Hyperbelfunktionen, so gibt es keine analogen Gleichungen.

Hyperbelfunktionen können geometrisch interpretiert werden, wenn man statt des Einheitskreises $x^2 + y^2 = 1$ die Einheitshyperbel $x^2 - y^2 = 1$ betrachtet.

Es gilt dann das folgende Lemma, das im Falle der Winkelfunktionen das Lemma 2.5.1 verallgemeinert.

Lemma 2.6.4

Für alle $z \in \mathbb{C}$ gilt:

$$\cos^2 z + \sin^2 z = 1 = \cosh^2 z - \sinh^2 z.$$

Beweis.

Im Falle der Hyperbelfunktionen gilt

$$\begin{aligned}\cosh^2 z - \sinh^2 z &= \left(\frac{\mathrm{e}^z + \mathrm{e}^{-z}}{2}\right)^2 - \left(\frac{\mathrm{e}^z - \mathrm{e}^{-z}}{2}\right)^2 \\ &= \frac{(\mathrm{e}^{2z} + 2\mathrm{e}^z\mathrm{e}^{-z} + \mathrm{e}^{-2z}) - (\mathrm{e}^{2z} - 2\mathrm{e}^z\mathrm{e}^{-z} + \mathrm{e}^{-2z})}{4} \\ &= 1\,.\end{aligned}$$

Der Beweis für die Winkelfunktionen ist analog. □

Section 2.7
Polynome

Eine Funktion $p : \mathbb{C} \to \mathbb{C}$ heißt **Polynom**, falls es ein $n \in \mathbb{N}_0$ und **Koeffizienten** $a_0, \ldots, a_n \in \mathbb{C}$ gibt, so dass

$$p(z) = \sum_{k=0}^{n} a_k z^k \tag{2.7.1}$$

für alle $z \in \mathbb{C}$ gilt. Für Polynome $p \neq 0$ setzen wir

$$\deg p := \min\left\{ n \in \mathbb{N}_0 \;\middle|\; \exists a_0, \ldots, a_n \in \mathbb{C}\, \forall z \in \mathbb{C} : p(z) = \sum_{k=0}^{n} a_k z^k \right\}$$

und für das Polynom $p = 0$ schreiben wir $\deg p := -1$. In beiden Fällen heißt $\deg p$ der **Grad** des Polynoms p.

Ferner sprechen wir von einem **reellen Polynom**, falls wir reelle Koeffizienten finden, d.h. $a_0, \ldots, a_n \in \mathbb{R}$.

Sind p und q zwei Polynome mit $\deg p \geq 0$ und $\deg q \geq 0$, so gilt

$$\deg(p+g) \leq \max\{\deg p, \deg q\}\,, \tag{2.7.2}$$

$$\deg(pq) \leq \deg p + \deg q\,, \tag{2.7.3}$$

wobei beide Formeln durch leichtes Ausrechnen gezeigt werden können.

Später werden wir sehen, dass man mit Hilfe von Ableitungen leicht sehen kann, dass die Koeffizienten eines Polynoms eindeutig sind. Dies wiederum führt dazu, dass die Ungleichung (2.7.3) eine Gleichung ist. Im folgenden werden wir aber die Eindeutigkeit der Koeffizienten nicht benötigen.

Ein $z \in \mathbb{C}$ heißt **Nullstelle** von p, falls $p(z) = 0$ gilt. Für Nullstellen reeller Polynome gilt die folgende Beobachtung:

Lemma 2.7.1

Ist p ein reelles Polynom und $z \in \mathbb{C}$ eine Nullstelle von p, so ist auch $\bar{z}$ eine Nullstelle von p.

Beweis.

Es gilt

$$p(\bar{z}) = \sum_{k=0}^{n} a_k \bar{z}^k = \sum_{k=0}^{n} \overline{a_k} \cdot \bar{z}^k = \overline{\sum_{k=0}^{n} a_k z^k} = \overline{p(z)} = \overline{0} = 0\,,$$

wobei wir im zweiten Schritt $a_0, \ldots, a_n \in \mathbb{R}$ ausgenutzt haben. □

Polynome vom Grad 2 erlauben die Berechnung mit der aus der Schule bekannten “Mitternachts-Formel”, wobei es leicht zu ersehen ist, dass diese auch im Komplexen gilt. Für Polynome vom Grad 3 und 4 gibt es ebenfalls noch Formeln, die aber in der Regel unbrauchbar sind. Ferner kann man zeigen, dass es solche Formeln für allgemeine Polynome vom Grad ≥ 5 nicht geben kann.

Unabhängig davon gilt aber der auf Gauß zurückgehende Fundamentalsatz der Algebra, den wir an dieser Stelle weder beweisen wollen noch können.

Theorem 2.7.2

Jedes Polynom p mit $\deg p \geq 1$ *hat mindestens eine Nullstelle in* $\mathbb{C}$.

Ähnlich wie für ganze Zahlen können wir Polynome mit Rest teilen. Dies ist der Inhalt des folgenden Satzes.

Theorem 2.7.3

Seien p und q Polynome mit $\deg p \geq \deg q \geq 0$. Dann gibt es Polynome r und s mit

$$p(z) = s(z)q(z) + r(z)\,, \qquad z \in \mathbb{C}\,,$$

sowie $\deg s = \deg p - \deg q$ und $\deg r < \deg q$.

Seien $n := \deg p$ und $m := \deg q$. Ferner seien $a_n \neq 0$ und $b_m \neq 0$ entsprechende Koeffizienten von p und q. Dann ist

$$p_1(z) := p(z) - \frac{a_n}{b_m} z^{n-m} q(z)\,, \qquad z \in \mathbb{C}$$

ein Polynom mit $\deg p_1 \leq n - 1$. Wir definieren das Polynom s_1 durch $s_1(z) := \frac{a_n}{b_m} z^{n-m}$ für alle $z \in \mathbb{C}$, so dass dann $p_1 = p - s_1 q$ und $\deg s_1 \leq n - m$ gilt.

Ist $\deg p_1 \geq q$, so wenden wir die obige Konstruktion auf p_1 und q an, so dass wir Polynome p_2 und s_2 mit $p_2 = p_1 - s_2 q$, sowie $\deg s_2 \leq \deg p_1 - m \leq n - m$ und $\deg p_2 \leq \deg(p_1) - 1 \leq n - 2$ gefunden haben.

Dieses Vorgehen wiederholen wir nun solange, bis die gefundenen Polynome p_k und s_k, die $p_k = p_{k-1} - s_k q$ und $\deg s_k \leq n - m$ nach Konstruktion erfüllen, auch $\deg p_k < \deg q$ erfüllen. Da wir $\deg p_i < n - i$ im i-ten Schritt haben, ist dies nach maximal $n - m$ Schritten der Fall. Wir setzen nun $r := p_k$ und $s := s_1 + \cdots + s_k$.

Für $p_0 := p$ liefert unsere Konstruktion dann

$$\begin{aligned} r = p_k &= p_{k-1} - s_k q \\ &= p_{k-2} - s_{k-1} q - s_k q \\ &= \dots \\ &= p_0 - s_1 q - \dots - s_{k-1} q - s_k q \\ &= p - (s_1 + \dots s_k) q \\ &= p - sq \end{aligned}$$

und $\deg r < \deg q$.

Schließlich folgt $\deg s \leq n - m$ aus (2.7.2) in Verbindung mit $\deg s_i \leq n - m$ für alle $1, \ldots, k$. Umgekehrt wissen wir wegen $p = sq + r$ und (2.7.2) auch

$$\begin{aligned} \deg p \leq \max\{\deg(sq), \deg r\} &\leq \max\{\deg(sq), \deg(q) - 1\} \\ &\leq \max\{\deg(sq), \deg(p) - 1\} . \end{aligned}$$

Wäre nun $\deg(p) - 1 \geq \deg(sq)$, so hätten wir $\deg p \leq \deg(p) - 1$, was unmöglich ist. Aus diesem Grund gilt $\deg(p) - 1 < \deg(sq)$. Dies führt zu

$$\deg p \leq \max\{\deg(sq), \deg(p) - 1\} = \deg(sq) \leq \deg s + \deg q ,$$

wobei wir im letzten Schritt (2.7.3) ausgenutzt haben. Dies ergibt $\deg s \geq \deg p - \deg q$.

BEISPIEL

Der Beweis des Satzes 2.7.3 liefert eine explizite Konstruktion für s und r. Wir betrachten dies am Beispiel der Polynome $p(z) := 2z^3 - 3z^2 - 6z + 6$ und $q(z) := z - 2$. Wir gehen dann wie bei der schriftlichen Division vor:

$$
\begin{array}{rl}
\left(\quad 2z^3 - 3z^2 - 6z + 6 \right) \div \left(z - 2 \right) = 2z^2 + z - 4 + \dfrac{-2}{z-2} \\
\underline{-\,2z^3 + 4z^2} \\
\qquad\quad z^2 - 6z \\
\qquad\quad \underline{-\,z^2 + 2z} \\
\qquad\qquad -\,4z + 6 \\
\qquad\qquad \underline{\quad 4z - 8} \\
\qquad\qquad\quad -\,2
\end{array}
$$

Hierbei sind

$$
\begin{aligned}
s_1 &= 2z^2 & p_1 &= z^2 - 6z + 6 \\
s_2 &= z & p_2 &= -4z + 6 \\
s_3 &= 4 & p_3 &= -2\,,
\end{aligned}
$$

und damit erhalten wir $s = 2z^2 + z + 4$ und $r = -2$.

Mit Hilfe der Polynomdivision und dem Fundamentalsatz der Algebra können wir Polynome durch ihre Nullstellen darstellen. Um dies zu zeigen, benötigen wir zunächst das folgende Korollar.

Korollar 2.7.4

Sei p ein Polynom mit $\deg p \geq 1$ und λ eine Nullstelle von p. Dann gibt es ein Polynom s mit $\deg s = \deg(p) - 1$ und

$$p(z) = (z - \lambda) \cdot s(z)\,, \qquad z \in \mathbb{C}.$$

*In diesem Fall heißt das Polynom $(z - \lambda)$ vom Grad 1 **Linearfaktor** von p.*

Wir betrachten das Polynom $q(z) := z - \lambda$. Dies ist eine nicht-konstante, lineare Funktion und deswegen haben wir $\deg q = 1$. Nach Satz 2.7.3 gibt es nun Polynome s und r mit $p = s \cdot q + r$ für alle $z \in \mathbb{C}$, sowie $\deg s = \deg p - \deg q = \deg(p) - 1$ und $\deg r < \deg q = 1$. Damit gibt es ein $a_0 \in \mathbb{C}$ mit $r(z) = a_0$ für alle $z \in \mathbb{C}$. Es folgt

$$0 = p(\lambda) = s(\lambda) \cdot (\lambda - \lambda) + r(\lambda) = r(\lambda) = a_0$$

und damit $r = 0$.

Wir können nun den sogenannten **Hauptsatz der Algebra**, der ein Polynom anhand seiner Nullstellen beschreibt, beweisen.

Theorem 2.7.5

Sei p ein Polynom mit $\deg p = n$ für ein $n \geq 1$. Dann gibt es ein $a \in \mathbb{C}$ mit $a \neq 0$ und $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ mit

$$p(z) = a \cdot (z - \lambda_1) \cdot \cdots \cdot (z - \lambda_n), \qquad z \in \mathbb{C}. \tag{2.7.4}$$

Ferner sind $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ die einzigen Nullstellen von p.

Kommt der Faktor $(z - \lambda_i)$ genau k-mal in der Zerlegung (2.7.4) vor, so sprechen wir von einer **k-fachen Nullstelle** von p. Beispielsweise ist $\lambda = 1$ eine zweifache Nullstelle des reellen Polynoms $p(z) = z^2 - 2z + 1 = (z - 1)^2$.

BEWEIS

Zum Beweis von (2.7.4) benutzen wir Induktion über n.

Für $n = 1$ haben wir dabei $p(z) = a_1 z + a_0$ mit $a_1 \neq 0$, so dass $\lambda_1 := -a_0/a_1$ und $a := a_1$ die gewünschte Darstellung liefert.

Für den Induktionsschritt nehmen wir an, dass die Aussage für alle Polynome q mit $\deg q = n$ wahr ist. Sein nun p ein Polynom mit $\deg p = n + 1$. Nach dem Fundamentalsatz der Algebra, siehe Satz 2.7.2, gibt es dann eine Nullstelle $\lambda_{n+1} \in \mathbb{C}$ von p. Korollar 2.7.4 gibt uns dann ein Polynom q mit $\deg q = n$ und $p(z) = (z - \lambda_{n+1})q(z)$ für alle $z \in \mathbb{C}$. Anwenden der Induktionsvoraussetzung ergibt dann die Behauptung für p.

Aus (2.7.4) lässt sich sofort ablesen, dass $\lambda_1, \ldots, \lambda_n$ Nullstellen von p sind. Sei nun $\lambda \in \mathbb{C}$ eine beliebige Nullstelle. Dann folgt

$$0 = p(\lambda) = a \cdot (\lambda - \lambda_1) \cdot \cdots \cdot (\lambda - \lambda_n)\,.$$

Da $\mathbb{C}$ ein Körper ist, muss einer der Faktoren gleich 0 sein, d.h. es gibt ein i mit $\lambda - \lambda_i = 0$. Dies zeig $\lambda = \lambda_i$, d.h. es gibt keine weitere Nullstelle als die Zahlen $\lambda_1, \ldots, \lambda_n$.

Mit Hilfe des Hauptsatzes der Algebra können wir nun einen Test auf Gleichheit für Polynome formulieren.

Korollar 2.7.6

Seien p und q Polynom mit $\deg p \leq n$ *und* $\deg q \leq n$ *für ein* $n \in \mathbb{N}_0$. *Falls es* $n+1$ *verschiedene Punkte* $z_1, \ldots, z_{n+1} \in \mathbb{C}$ *gibt mit* $p(z_i) = q(z_i)$ *für alle* $i = 1, \ldots, n+1$, *so gilt* $p = q$, *d.h.*

$$p(z) = q(z)$$

für alle $z \in \mathbb{C}$.

Betrachten wir insbesondere $q = 0$ in dem obigen Korollar, so sehen wir, dass die Existenz von $n+1$ verschiedenen Nullstellen von p schon $p = 0$ impliziert. Tatsächlich betrachtet der folgende Beweis des Korollars genau diesen Fall.

BEWEIS

Wir setzen $r := p - q$. Nach (2.7.2) ist dann r ein Polynom mit $k := \mathbf{deg}\, r \leq n$. Ferner sind nach Konstruktion die Zahlen $z_1, \ldots, z_{n+1}$ Nullstellen von r.

Wäre nun $r \neq 0$, so hätte im Fall $k \geq 1$ das Polynom r nach dem Hauptsatz der Algebra genau k Nullstellen. Wegen $k \leq n < n + 1$ führt dies zum Widerspruch.

Im Fall $k = 0$ ist r von der Form $r = a_0$ mit $a_0 \neq 0$ und damit hätte es gar keine Nullstelle, was ebenfalls zum Widerspruch führt.

ZERLEGUNG REELLER POLYNOME

Für reelle Polynome p hatten wir in Lemma 2.7.1 gesehen, dass mit λ auch $\overline{\lambda}$ eine Nullstelle ist. Da $\lambda = \overline{\lambda}$ äquivalent zu $\lambda \in \mathbb{R}$ ist, treten genau zwei Fälle für Nullstellen reeller Polynome auf:

i). Die Nullstelle λ ist reell mit zugehörigem Linearfaktor $z - \lambda$.

ii). Es ist $\lambda = \alpha + \mathrm{i}\beta$ mit $\beta \neq 0$ eine Nullstelle. Dann ist aber ebenso $\overline{\lambda} = \alpha - \mathrm{i}\beta$ Nullstelle und die beiden Linearfaktoren ergeben zusammen den quadratischen Faktor

$$\begin{aligned}(z - \alpha - \mathrm{i}\beta)(z - \alpha + \mathrm{i}\beta) &= (z - \alpha)^2 + \beta^2 \\ &= z^2 - 2\alpha z + \alpha^2 + \beta^2 \,.\end{aligned}$$

Damit zerfällt jedes reelle Polynom in ein Produkt aus Linearfaktoren und quadratischen Faktoren.

BEISPIEL

Das Polynom $p(z) := z^3 + 2z - 3$ besitzt beispielsweise die Nullstelle $\lambda_1 := 1$, wie direkt durch Einsetzen ersichtlich ist. Spaltet man diesen Linearfaktor mittels Polynomdivision ab, so gilt

$$z^3 + 2z - 3 = (z^2 + z + 3)(z - 1)$$

und der entstehende quadratische Faktor $(z^2 + z + 3)$ besitzt keine weiteren reellen Nullstellen. Allerdings besitzt er zwei zu einander komplex konjugierte Nullstellen

$$\lambda_{2,3} := -\frac{1}{2} \pm \sqrt{\frac{1}{4} - 3} = -\frac{1}{2} \pm \mathrm{i}\frac{\sqrt{11}}{2}.$$

ANWENDUNG: FUNKTIONSINTERPOLATION

Im folgenden betrachten wir ein sogenanntes **Interpolationsproblem**, bei dem zu gegebenen **Eingabepunkten** $x_1, \ldots, x_n \in \mathbb{R}$ und **Ausgabepunkten** $y_1, \ldots, y_n \in \mathbb{R}$ eine Funktion $f : \mathbb{R} \to \mathbb{R}$ gesucht ist, die

$$f(x_i) = y_i \,, \qquad\qquad i = 1, \ldots, n$$

erfüllt. Da es naturgemäß sehr viele solche Funktionen gibt, werden typischerweise zusätzliche Forderungen an die gesuchte Funktion f gestellt.

Eine mögliche Anwendung für dieses Problem besteht darin, dass wir eine komplizierte oder aufwendig zu berechnende Funktion $g : \mathbb{R} \to \mathbb{R}$ haben und diese durch eine "billige" Alternative f ersetzen wollen. Die Hoffnung besteht dann darin, dass aus

$$f(x_i) = y_i := g(x_i) \,, \qquad\qquad i = 1, \ldots, n$$

schon $f(x) \approx g(x)$ für alle $x \in \mathbb{R}$, die uns interessieren, folgt.

Wir wollen im folgenden ein interpolierendes Polynom p mit $\deg p \leq n-1$ konstruieren. Der folgende Satz zeigt, dass dies auf genau eine Art möglich ist.

Theorem 2.7.7

Seien $x_1, \ldots, x_n \in \mathbb{R}$ paarweise verschieden und $y_1, \ldots, y_n \in \mathbb{R}$. Dann gibt es genau ein Polynom p mit $\deg p \leq n-1$ und

$$p(x_i) = y_i\,, \qquad\qquad i = 1, \ldots, n$$

ILLUSTRATION

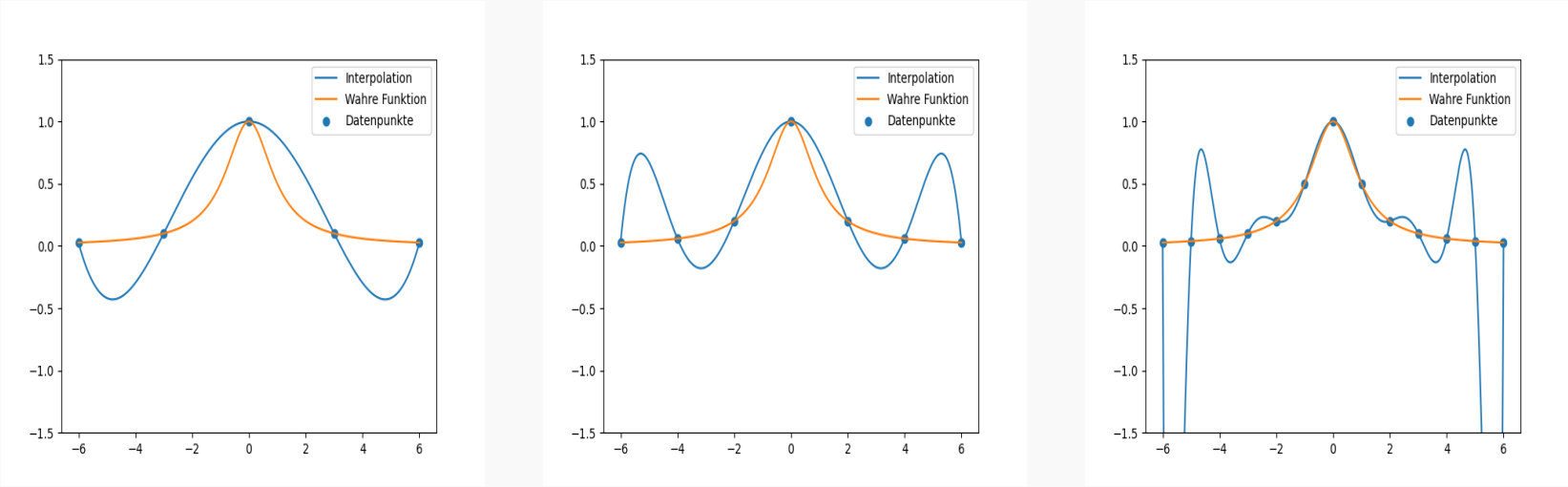


Abbildung: Interpolation des Funktion $g(x) := (1 + x^2)^{-1}$ durch Polynome aus Satz 2.7.7. **Links:** 5 Eingabepunkte mit Abstand 3. **Mitte:** 7 Eingabepunkte mit Abstand 2. **Rechts:** 13 Eingabepunkte mit Abstand 1. Insgesamt nähern sich die interpolierenden Polynome nur "in der Mitte" der Funktion f an.

Zunächst konstruieren wir ein Polynom mit den gesuchten Eigenschaften. Für $k = 1, \ldots, n$ definieren wir dazu die Lagrange-Polynome

$$L_k(x) := \prod_{i=1, i\neq k}^{n} \frac{x - x_i}{x_k - x_i}, \qquad x \in \mathbb{R}.$$

Für $i = 1, \ldots, n$ gilt dann

$$L_k(x_i) = \delta_{i,k} := \begin{cases} 1 & \text{falls } i = k \\ 0 & \text{sonst,} \end{cases}$$

da im ersten Fall alle Brüche des Lagrange-Polynom L_k gleich 1 sind und im zweiten Fall ein Bruch gleich 0 ist. Ferner ist L_k offensichtlich ein Polynom mit $\mathbf{deg}\, L_k \leq n - 1$. Damit ist

$$p := \sum_{k=1}^{n} y_k L_k$$

das gesuchte Polynom.

Die Eindeutigkeit folgt direkt aus Korollar 2.7.6.

BEMERKUNGEN

Der Beweis des Satzes 2.7.7 konstruiert das Polynom explizit. Ferner ist dieses Polynom einfach zu konstruieren und bei Änderung von den Ausgabepunkten y_i lässt es sich leicht anpassen. Ferner ist das Polynom eindeutig bestimmt und sehr “glatt”. Leider werden die Berechnungen teuer, wenn sich die Eingabepunkte x_i ändern sollten. Darüber hinaus können die interpolierenden Polynome “zappeln”, siehe Abbildung 10, was auch als Runges Phänomen bekannt ist.

CHAPTER 3: GEOMETRIE

Section 3.1
Operationen mit Vektoren

Im folgenden wollen wir einige geometrische Aspekte im zwei- und drei-dimensionalen Raum betrachten. Damit diese möglichst gleichzeitig behandelt werden können, betrachten wir zunächst den d-dimensionalen Raum $\mathbb{R}^d$.

Im folgenden nennen wir ein $x \in \mathbb{R}^d$ **Vektor** und schreiben ihn als **Spaltenvektor**

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix},$$

wobei $x_1, \ldots, x_d \in \mathbb{R}$ die **Komponenten** von x sind. In der Literatur findet man auch die Schreibweisen $\vec{x}$ oder $\mathbf{x}$ für Vektoren, diese werden wir aber nicht verwenden.

In den Räumen $\mathbb{R}^2$ und $\mathbb{R}^3$ sehen Vektoren damit so aus

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \qquad \text{und} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \in \mathbb{R}^3 .$$

Vektoren kann man als **Punkte** in dem d-dimensionalen Raum $\mathbb{R}^d$ und als **Richtungspfleile** in dem Raum $\mathbb{R}^d$ interpretieren. Je nachdem in welchem Kontext man sich befindet, ist die eine oder die andere Interpretation hilfreicher.

ADDITION UND SKALARMULTIPLIKATION

Auf $\mathbb{R}^d$ können wir eine **Vektor-Addition** $+ : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ durch

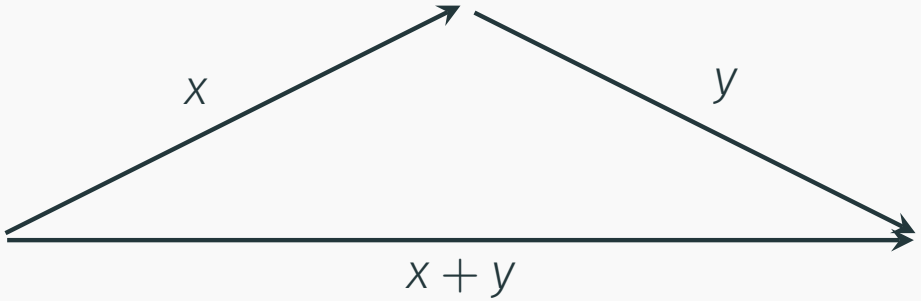
$$\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} := \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_d + y_d \end{pmatrix}$$

definieren. Es ist leicht zu überprüfen, dass $(\mathbb{R}^d, +)$ eine kommutative Gruppe mit neutralem Element

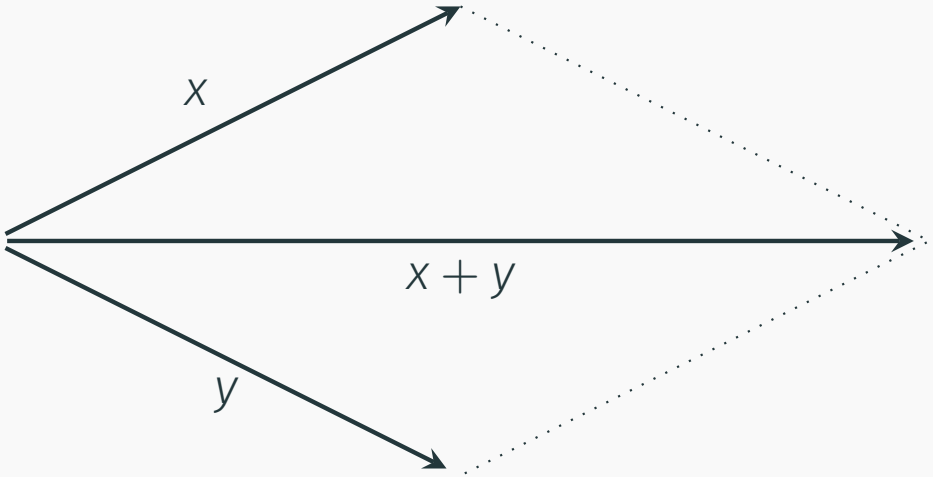
$$0 := \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^d$$

ist. Für $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$ ist zudem $-x := \begin{pmatrix} -x_1 \\ \vdots \\ -x_d \end{pmatrix}$ das inverse Element.

Die Vektoraddition kann man mit Verschiebungen vom Nullpunkt illustrieren:



Die Vektoraddition wird oft auch mit einem Parallelogramm illustriert:



SKALARMULTIPLIKATION

Auf $\mathbb{R}^d$ können wir außerdem eine Skalarmultiplikation $\cdot : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ durch

$$\alpha \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} := \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_d \end{pmatrix}$$

definieren. Häufig wird dann aber der Punkt $\cdot$, wie bei der Multiplikation in $\mathbb{Q}$, $\mathbb{R}$ oder $\mathbb{C}$, weggelassen. Die Skalarmultiplikation entspricht einer Streckung des Vektors x um den Faktor α.

Der Raum $\mathbb{R}^d$ zusammen mit seiner Vektoraddition und Skalarmultiplikation bildet einen Vektorraum im Sinne der folgenden Definition, wie einfache Rechnungen zeigen.

Definition 3.1.1

Sei V eine Menge auf der Operationen $+ : V \times V \to V$ und $\cdot : \mathbb{R} \times V \to V$ definiert sind. Dann heißt $(V, +, \cdot)$ Vektorraum, oder auch $\mathbb{R}$-Vektorraum, falls $(V, +)$ eine kommutative Gruppe mit neutralem Element 0 ist und für alle $x, y \in V$ und $\alpha, \beta \in \mathbb{R}$ gilt:

$$\begin{aligned} \alpha \cdot (\beta \cdot x) &= (\alpha\beta) \cdot x\,, \\ \alpha \cdot (x + y) &= \alpha \cdot x + \alpha \cdot y\,, \\ (\alpha + \beta) \cdot x &= \alpha \cdot x + \beta \cdot x\,, \\ 1 \cdot x &= x\,. \end{aligned}$$

Haben wir stattdessen eine Skalarmultiplikation $\cdot : \mathbb{C} \times V \to V$ die die gleichen 4 Gleichungen für alle $\alpha, \beta \in \mathbb{C}$ erfüllt, so sprechen wir von einem $\mathbb{C}$-Vektorraum.

Aus den 4 Eigenschaften der Skalarmultiplikation kann man z.B. auf $(-1) \cdot x = -x$ und $0 \cdot x = 0$ schließen. Im $\mathbb{R}^d$ kann man dies natürlich auch direkt nachweisen.

Zu einem Vektor $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$ ist die Entfernung dieses Punktes zum Nullpunkt nach Pythagoras durch

$$\|x\|_2 := |x| := \sqrt{x_1^2 + \cdots + x_d^2}$$

gegeben. Im folgenden nennen wir $|x|$ den **Betrag** von x. Der Betrag ist eine Norm im Sinne der folgenden Definition.

Definition 3.1.2

Sei $(V, +, \cdot)$ Vektorraum. Dann heißt eine Abbildung $\|\cdot\| : V \to [0, \infty)$ **Norm** auf V, falls für alle $x, y \in V$ und $\alpha \in \mathbb{R}$ die folgenden drei Eigenschaften erfüllt sind:

i). **Definitheit:** $\|x\| = 0 \Leftrightarrow x = 0$.

ii). **(absolute) Homogenität:** $\|\alpha x\| = |\alpha| \cdot \|x\|$.

iii). **Dreiecksungleichung:** $\|x + y\| \leq \|x\| + \|y\|$.

Vektorräume mit einer Norm heißen **normierte Räume**. Analog können Normen auf $\mathbb{C}$-Vektorräumen definiert werden.

Die ersten beiden Eigenschaften sind dabei für den Betrag leicht nachzurechnen. Für den Betrag entspricht die Dreiecksungleichung der bekannten Tatsache, dass in einem Dreieck jede Seite kürzer als die Summe der beiden anderen Seiten ist. Einen formalen Beweis werden wir später kennenlernen.

Ein Vektor $x \in \mathbb{R}^d$ mit $|x|$ heißt **normiert** oder auch **Einheitsvektor**. Ist $x \neq 0$ ein Vektor im $\mathbb{R}^d$, so ist $|x|^{-1}x$ normiert. Häufig schreiben wir auch $\frac{x}{|x|}$ statt $|x|^{-1}x$. Analoge Definitionen sind in allgemeinen normierten Räumen möglich.

Neben dem Betrag gibt es eine Vielzahl weitere Normen auf dem $\mathbb{R}^d$. Hier wollen wir nur die folgenden zwei Normen erwähnen.

Die **Supremumsnorm**:

$$\|x\|_\infty := \max\{|x_i| : i = 1, \ldots, d\}\,.$$

Die **1-norm**:

$$\|x\|_1 := \sum_{i=1}^{d} |x_i|\,.$$

Ist $\|\cdot\|$ eine Norm auf einem Vektorraum V, so heißt

$$B_V := B_{\|\cdot\|} := \{x \in V : \|x\| \leq 1\}$$

die **Einheitskugel** von $\|\cdot\|$. Im Falle von $V = \mathbb{R}^2$ ist die Einheitskugel bzgl. des Betrages der Kreis mit Mittelpunkt 0 und Radius 1. Für die Supremumsnorm ist die Einheitskugel das Quadrat mit Mittelpunkt 0 und Seitenlänge 2.

Im folgenden heißt der Vektor $e_i \in \mathbb{R}^d$, der an der i-ten Stelle gleich 1 ist und ansonsten nur Nullen enthält, der *i-te Einheitsvektor* im $\mathbb{R}^d$. Im $\mathbb{R}^3$ sind also

$$e_1 := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad e_2 := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad e_3 := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

die drei Einheitsvektoren.

Für $x := \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$ gilt die Darstellung

$$x = \sum_{i=1}^{d} x_i e_i \,, \tag{3.1.1}$$

wie durch einfaches Nachrechnen überprüft werden kann.

Für $x := \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$ und $y := \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} \in \mathbb{R}^d$ ist das Skalarprodukt durch

$$\langle x, y \rangle := \sum_{i=1}^{d} x_i y_i$$

definiert. Damit haben wir eine Abbildung $\langle \cdot\,, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ definiert, die die Gleichungen

$$\begin{aligned} \langle x, x \rangle &= |x|^2 \\ \langle x, y \rangle &= \langle y, x \rangle \\ \langle \alpha x, y \rangle &= \alpha \langle x, y \rangle \\ \langle x + z, y \rangle &= \langle x, y \rangle + \langle z, y \rangle \end{aligned}$$

für alle $x, y, z \in \mathbb{R}^d$ und $\alpha \in \mathbb{R}$ erfüllt. Diese Gleichungen lassen sich jeweils sehr einfach nachrechnen und werden daher übersprungen.

In der Literatur finden sich viele weitere Schreibweisen des Skalarprodukts, wie z.B.

$$x \cdot y := x \circ y := x \bullet y := \langle x, y \rangle .$$

GEOMETRISCHE INTERPRETATION DES SKALARPRODUKTS

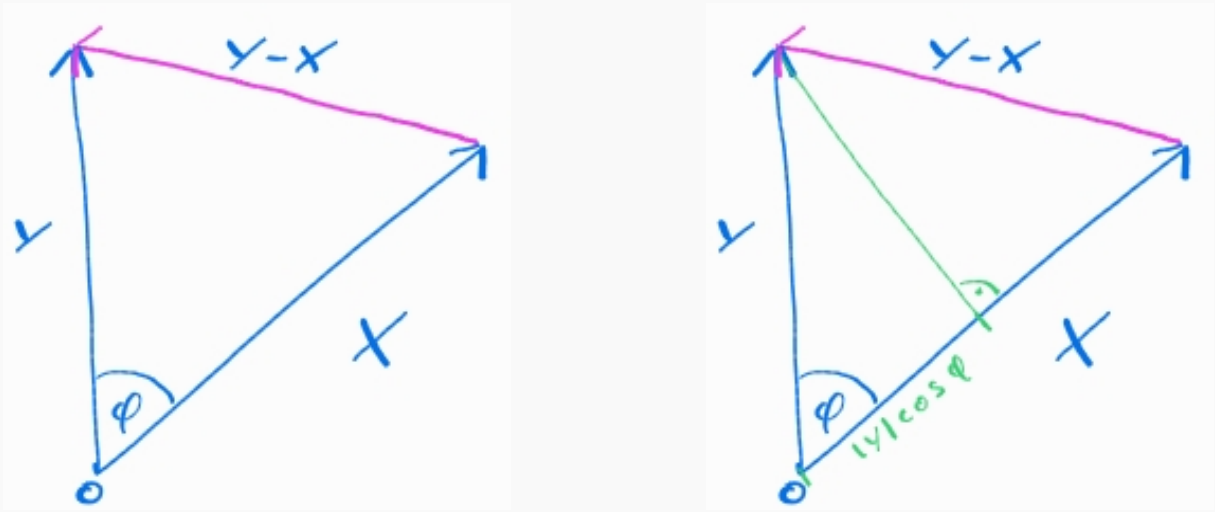


Abbildung: Links: Ansatz für die Anwendung des Kosinussatzes im Beweis von Satz 3.1.3. **Rechts:** Die Länge der Strecke zwischen 0 und der grünen Höhe des Dreiecks ist laut Satz 3.1.3 gleich $|y| \cos \varphi = \langle x, y \rangle \cdot |x|^{-1}$. Der entsprechende Vektor von 0 bis zum Fußpunkt der Höhe ist damit $\langle x, y \rangle \cdot \frac{x}{|x|}$. Dieser Vektor entspricht der Projektion des Vektors y auf die von x aufgespannte Gerade.

Der folgende Satz liefert eine geometrische Interpretation des Skalarprodukts, vgl. auch Abbildung 11.

Theorem 3.1.3

Im Fall $\varphi = 0$ gilt $y = x$ und im Fall $\varphi = \pi$ gilt $y = -x$. In beiden Fällen ist die Behauptung dann offensichtlich.

Für $\varphi \in (0, \pi)$ betrachten wir das Dreieck mit den Ecken 0, x und y. Da die Strecke $\overline{xy}$, die dem Winkel φ gegenüberliegt, durch $y - x$ gegeben ist, haben die Seiten des Dreiecks die Längen $|x|$, $|y|$ und $|y - x|$, siehe auch Abbildung 11. Mit dem Kosinussatz 2.5.3 folgt dann

$$|y - x|^2 = |x|^2 + |y|^2 - 2|x| \cdot |y| \cos \varphi \, .$$

Ferner gilt

$$|y - x|^2 = \sum_{i=1}^{d} (x_i - y_i)^2 = \sum_{i=1}^{d} x_i^2 - 2x_i y_i + y_i^2 = |x|^2 + |y|^2 - 2\langle x, y \rangle \, .$$

Gleichsetzen der beiden Gleichungen ergibt

$$|x|^2 + |y|^2 - 2\langle x, y \rangle = |x|^2 + |y|^2 - 2|x| \cdot |y| \cos \varphi \, ,$$

und einfaches Umformen liefert dann die Behauptung.

Da $|\cos\varphi| \leq 1$ für alle $\varphi \in \mathbb{R}$ gilt, folgt aus Satz 3.1.3 sofort die Cauchy-Schwarz'sche-Ungleichung

$$|\langle x, y\rangle| \leq |x| \cdot |y| \tag{3.1.2}$$

für alle $x, y \in \mathbb{R}^d$. Diese kann auch rein rechnerisch, d.h. ohne den geometrischen Ansatz des Satz 3.1.3, nachgewiesen werden.

Ferner sind zwei Vektoren $x, y \in \mathbb{R}^d$ senkrecht zueinander oder orthogonal, geschrieben $x \perp y$, genau dann wenn $\langle x, y\rangle = 0$ gilt.

Wir sagen ferner, dass zwei Vektoren $x, y \in \mathbb{R}^d$ parallel sind, geschrieben $x||y$, falls es ein $\alpha \in \mathbb{R}$ gibt mit $x = \alpha y$ oder $\alpha x = y$. Die beiden Fälle sind dabei notwendig, um die Fälle $x = 0$ und $y \neq 0$ bzw. $y = 0$ und $x \neq 0$ mit einzuschließen.

Sei $e_i \in \mathbb{R}^d$ der i-te Einheitsvektor. Für $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$ gilt dann

$$\langle x, e_i \rangle = x_i$$

und damit haben wir insgesamt

$$x = \sum_{i=1}^{d} \langle x, e_i \rangle e_i \,. \tag{3.1.3}$$

Ferner sind verschiedene Einheitsvektoren senkrecht zueinander.

KREUZPRODUKT

Für $d = 3$ ist das **Kreuzprodukt** $\times : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ durch

$$x \times y := \begin{pmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{pmatrix}$$

definiert, wobei wir $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ und $y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$ betrachtet haben.

Das Kreuzprodukt erfüllt die Rechenregeln

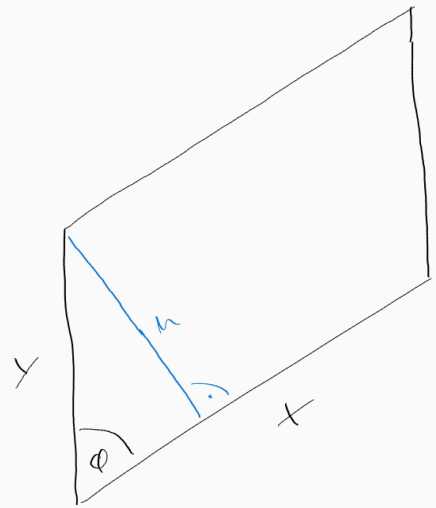
$$\begin{aligned} x \times x &= 0\,, \\ x \times y &= -(y \times x)\,, \\ (\alpha x) \times y &= \alpha \cdot (x \times y)\,, \\ x \times (y+z) &= x \times y + x \times z\,, \\ x \times y = 0 \quad &\Leftrightarrow \quad x = 0 \ \vee \ y = 0 \ \vee \ x \| y, \\ |x \times y|^2 &= |x|^2 \, |y|^2 - \langle x, y \rangle^2\,, \end{aligned}$$

die jeweils elementar nachgerechnet werden können. Das gleiche gilt für die Formeln

$$\begin{aligned} x \times (y \times z) &= \langle x, z \rangle y - \langle x, y \rangle z \\ \langle x \times y, v \times w \rangle &= \langle x, v \rangle \cdot \langle y, w \rangle - \langle y, v \rangle \cdot \langle x, w \rangle\,. \end{aligned}$$

Das Kreuzprodukt hat keine Gruppenstruktur, denn ansonsten gäbe es z.B. ein neutrales Element e. Wegen $e = e \times e = 0$ und $0 \times x = 0$ führt dies für $x \neq 0$ zum Widerspruch. Das Kreuzprodukt ist auch nicht assoziativ!

KREUZPRODUKT: GEOMETRISCHE INTERPRETATION



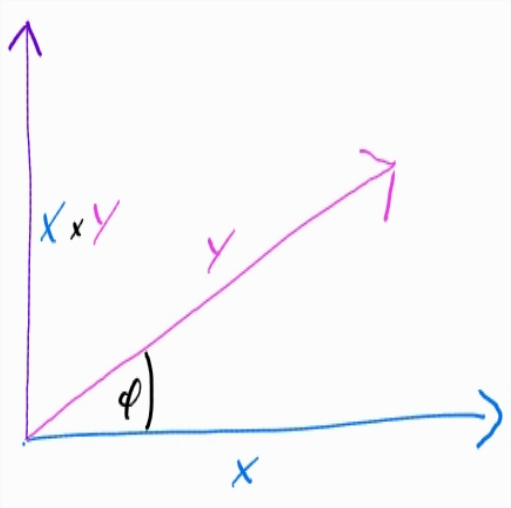


Abbildung: Links: Die Höhe des Parallelogramms ist durch $|y| \sin \varphi$ gegeben. Damit ist seine Fläche gleich $|x|\,|y| \sin \varphi = |x \times y|$. **Rechts:** Die Vektoren x und y liegen in der Zeichnung auf der horizontalen Ebene. Ihr Kreuzprodukt hat die in der Zeichnung angegebene Richtung. Dies entspricht einem Rechts-System.

KREUZPRODUKT: GEOMETRISCHE INTERPRETATION

Für $x, y \in \mathbb{R}^3$ liefert eine weitere einfache Rechnung

$$\langle x \times y, x \rangle = 0 = \langle x \times y, y \rangle .$$

Damit steht $x \times y$ senkrecht zu x und y und damit auch senkrecht zu der von x und y und 0 aufgespannten Ebene $\mathcal{E}$, d.h. zu

$$\mathcal{E} := \left\{ \alpha x + \beta y : \alpha, \beta \in \mathbb{R} \right\} .$$

Umgekehrt ist $\mathcal{E}$ gleich der Menge aller Vektoren, die senkrecht zu $x \times y$ stehen. Neben der geometrischen Anschauung kann man dies z.B. durch Lösen der Gleichung $\langle x \times y, z \rangle = 0$ nach z elementar nachrechnen. Im Sommersemester werden wir sehen, dass dies auch aus einfachen Dimensions-Betrachtungen folgt.

KREUZPRODUKT: GEOMETRISCHE INTERPRETATION

Aus Satz 3.1.3 wissen wir ferner $\langle x, y\rangle^2 = |x|^2 \cdot |y|^2 \cos^2 \varphi$, wobei $\varphi \in [0, \pi]$ der Winkel zwischen den Strecken $\overline{0x}$ und $\overline{0y}$ ist. Aus den Rechenregeln des Kreuzprodukts und dem Satz des Pythagoras folgern wir damit, dass

$$\begin{aligned} |x \times y|^2 &= |x|^2 |y|^2 - \langle x, y\rangle^2 \\ &= |x|^2 |y|^2 (1 - \cos^2 \varphi) \\ &= |x|^2 |y|^2 \sin^2 \varphi \,. \end{aligned}$$

Mit anderen Worten haben wir

$$|x \times y| = |x| \, |y| \sin \varphi \,,$$

wobei der Ausdruck auf der rechten Seite dem Flächeninhalt des von x und y aufgespannten Parallelogramms entspricht, siehe Abbildung 12.

Mit den bisherigen geometrischen Eigenschaften gibt es nur noch zwei geometrische Möglichkeiten für das Kreuzprodukt, die sich auch nur in ihrem Vorzeichen unterscheiden. Da $x, y, x \times y$ immer ein Rechts-System bildet, siehe Abbildung 12, ist das Kreuzprodukt eindeutig geometrisch beschreibbar.

Section 3.2
Geraden und Ebenen

GERADEN IM RAUM

Vektoren im Raum kann man nutzen, um geometrische Konzepte einfach zu beschreiben. Wir sammeln dazu einige Geraden- und Ebenengleichungen und zeigen, wie man mit diesen einfache Aufgabenstellungen der analytischen Geometrie lösen kann.

Eine **Gerade** g im Raum $\mathbb{R}^d$ ist bestimmt durch einen ihrer Punkte $x_0 \in \mathbb{R}^d$ und eine Richtung $v \in \mathbb{R}^d \setminus \{0\}$. Das führt zur Beschreibung

$$g = \{x_0 + tv \mid t \in \mathbb{R}\}.$$

Damit ist jedem Paar (x_0, v) von Vektoren $x_0, v \in \mathbb{R}^d$ eine Gerade zugeordnet. Die Zuordnung ist nicht injektiv, da z.B. (x_0, v) und $(x_0, -v)$ die gleichen Geraden beschreiben. Analog kann man den **Stützpunkt** x_0 durch jeden anderen Punkt x_0' der Geraden ersetzen, ohne dabei die Gerade zu ändern.

Durch den **Richtungsvektor** v haben wir der Gerade eine Richtung gegeben, wir sprechen deswegen auch von einer **gerichteten Geraden**.

Eine **Ebene** $\mathcal{E}$ im Raum $\mathbb{R}^3$ wird analog zu Geraden durch einen ihrer Punkte x_0 und zwei nichtparallele Richtungen $v, w \in \mathbb{R}^3 \setminus \{0\}$ bestimmt, d.h.

$$\mathcal{E} = \{x_0 + sv + tw \mid s, t \in \mathbb{R}\}.$$

Die Forderung, dass v und w nichtparallel sind, kann hierbei durch $v \times w \neq 0$ ausgedrückt werden. Die Darstellung von $\mathcal{E}$ heißt **Punkt-Richtungsform** oder auch **Parameterform**. Wie bei Geraden, kann eine Ebene durch verschiedene **Stützpunkte** x_0 und **Richtungsvektoren** v, w dargestellt werden.

Das folgende Lemma zeigt, dass es einen ausgezeichneten Stützpunkt gibt.

Lemma 3.2.1

Sei $\mathcal{E}$ eine Ebene, die durch den Stützpunkt x_0 und die Richtungsvektoren v, w dargestellt ist. Dann gibt es genau einen Punkt $x_0^* \in \mathcal{E}$ mit $\langle x_0^*, v\rangle = 0$ und $\langle x_0^*, w\rangle = 0$.

Die beiden Gleichungen bedeuten, dass x_0^* senkrecht zu der von v und w aufgespannten Ebene $\mathcal{E}'$ mit Stützpunkt 0 steht. Mit unserer geometrischen Interpretation des Kreuzprodukts gilt daher $x_0^* = \alpha(v \times w)$ für ein $\alpha \in \mathbb{R}$. Damit ist x_0^* auch auf der Geraden g mit Richtungsvektor $v \times w$ und Stützvektor 0. Insgesamt erhalten wir also $x_0^* \in g \cap \mathcal{E}$.

Der folgende Beweis beruht nicht auf diesem geometrischen Argument, sondern auf dem Gauß-Algorithmus zum Lösen eines linearen Gleichungssystems, das in diesem Fall ein einfaches 2×2-System ist, da wir auf diese Weise diesen Algorithmus in Erinnerung rufen können.

BEWEIS

Zunächst stellen wir fest, dass die Ebene $\mathcal{E}$ sich nicht ändert, wenn wir statt v und w die normierten Vektoren $|v|^{-1}v$ und $|w|^{-1}w$ betrachten. Das gleiche gilt für die beiden Orthogonalitätsgleichungen. Im folgenden nehmen wir daher $|v| = |w| = 1$ an. Es gilt dann $0 \neq |v \times w|^2 = |v|^2|w|^2 - \langle v, w\rangle^2 = 1 - \langle v, w\rangle^2$.

Wegen $x_0^* \in \mathcal{E}$ muss es $s, t \in \mathbb{R}$ geben, so dass

$$x_0^* = x_0 + sv + tw$$

gilt. Setzt man dies in die beiden Gleichungen $\langle x_0^*, v\rangle = 0$ und $\langle x_0^*, w\rangle = 0$ ein, ergibt sich

$$\langle x_0, v\rangle + s\langle v, v\rangle + t\langle v, w\rangle = 0$$
$$\langle x_0, w\rangle + s\langle v, w\rangle + t\langle w, w\rangle = 0\,.$$

Wegen $|v| = |w| = 1$ haben wir dann für $a := \langle v, w\rangle$, $c_1 := -\langle x_0, v\rangle$ und $c_2 := -\langle x_0, w\rangle$ die beiden Gleichungen

$$s + at = c_1$$
$$as + t = c_2\,.$$

Multipliziert man die erste Gleichung mit a und zieht man das Ergebnis dann von der zweiten Gleichung ab, so ergibt sich das äquivalente Gleichungssystem

$$\begin{aligned} s + at &= c_1 \\ 0s + t(1 - a^2) &= c_2 - ac_1 \,. \end{aligned}$$

Dazu äquivalent ist das Gleichungssystem

$$\begin{aligned} s + at &= c_1 \\ t &= \frac{c_2 - ac_1}{1 - a^2} \,, \end{aligned}$$

wobei wir $0 \neq 1 - \langle v, w \rangle^2 = 1 - a^2$ benutzt haben. Multipliziert man nun die zweite Gleichung mit a und zieht sie im Anschluss von der ersten Gleichung ab, so ergibt sich

$$\begin{aligned} s &= c_1 - a \cdot \frac{c_2 - ac_1}{1 - a^2} \\ t &= \frac{c_2 - ac_1}{1 - a^2} \,. \end{aligned}$$

Wegen

$$c_1 - a \cdot \frac{c_2 - ac_1}{1-a^2} = \frac{c_1(1-a^2) - ac_2 - a^2c_1}{1-a^2} = \frac{c_1 - ac_2}{1-a^2}$$

haben wir damit insgesamt

$$s = \frac{c_1 - ac_2}{1-a^2} = \frac{-\langle x_0, v\rangle + a\langle x_0, w\rangle}{1-a^2} = \frac{\langle x_0, aw - v\rangle}{1-a^2}$$
$$t = \frac{c_2 - ac_1}{1-a^2} = \frac{-\langle x_0, w\rangle + a\langle x_0, v\rangle}{1-a^2} = \frac{\langle x_0, av - w\rangle}{1-a^2}\,.$$

Damit ist der Punkt x_0^* durch

$$x_0^* = x_0 + sv + tw = x_0 + \frac{\langle x_0, aw - v\rangle}{1-a^2} \cdot v + \frac{\langle x_0, av - w\rangle}{1-a^2} \cdot w\,.$$

Wir beachten dabei, dass wegen $a = \langle v, w\rangle$ und $|v| = |w| = 1$ der Vektor aw der Projektion von v auf w entspricht und der Vektor av der Projektion von w auf v entspricht, vgl. Abbildung 11. Ferner gilt $1 - a^2 = |v \times w|^2$.

Neben der Punkt-Richtungsform gibt es eine weitere Möglichkeit, eine Ebene zu beschreiben. Dies ist der Inhalt des folgenden Satzes.

Theorem 3.2.2

Sei $\mathcal{E}$ eine Ebene, die durch den Stützpunkt x_0 und die Richtungsvektoren v, w dargestellt ist. Wir definieren

$$n := \frac{v \times w}{|v \times w|} \qquad \text{und} \qquad d := \frac{\langle x_0, v \times w \rangle}{|v \times w|} .$$

Dann gilt

$$\mathcal{E} = \{x \in \mathbb{R}^3 : \langle x, n \rangle = d\} .$$

Die in dem obigen Satz gefundene, alternative Beschreibung von $\mathcal{E}$ heißt Hessesche Normalform. Diese kann genutzt werden, um Abstände von Punkten zu Ebenen zu berechnen. Für einen beliebigen Punkt $x \in \mathbb{R}^3$ beschreibt

$$\mathrm{d}(x, \mathcal{E}) := \langle x, n \rangle - d$$

den *orientierten* Abstand von x zu der Ebene $\mathcal{E}$, wobei die Orientierung in Richtung des Normalenvektors gemessen wird. Mit anderen Worten gilt $\mathrm{d}(x, \mathcal{E}) > 0$ genau dann wenn, x auf der Seite der Ebene liegt, in die der Normalenvektor zeigt. Damit ist $|\langle x, n \rangle - d|$ der Abstand zwischen x und $\mathcal{E}$. Für $x := 0$ erhalten wir insbesondere den Wert $-d$, der den orientierten Abstand von $\mathcal{E}$ zum Ursprung wiedergibt und daher ist $|d|$ der Abstand der Ebene zum Ursprung.

BEWEIS

Wir setzen $\mathcal{E}' := \{x \in \mathbb{R}^3 : \langle x, n\rangle = d\}$, so dass wir $\mathcal{E} = \mathcal{E}'$ zeigen müssen.

Sei dazu zunächst $x \in \mathcal{E}$. Dann gibt es $s, t \in \mathbb{R}$ mit $x = x_0 + sv + tw$. Dies ergibt

$$\begin{aligned}\langle x, n\rangle = \frac{\langle x_0 + sv + tw, v \times w\rangle}{|v \times w|} &= \frac{\langle x_0, v \times w\rangle}{|v \times w|} + \frac{s\langle v, v \times w\rangle}{|v \times w|} + \frac{t\langle w, v \times w\rangle}{|v \times w|} \\ &= d\,.\end{aligned}$$

Sei umgekehrt $x \in \mathcal{E}'$. Dann gilt

$$\begin{aligned}\langle x - x_0, v \times w\rangle = \langle x, v \times w\rangle - \langle x_0, v \times w\rangle &= \langle x, v \times w\rangle - d\,|v \times w| \\ &= \big(\langle x, n\rangle - d\big)|v \times w| \\ &= 0\,.\end{aligned}$$

Damit muss nach unserer geometrischen Interpretation des Kreuzprodukts der Punkt $x - x_0$ in der von v und w aufgespannten Ebene sein, d.h. es gibt $s, t \in \mathbb{R}$ mit $x - x_0 = sv + tw$. Dies zeigt $x \in \mathcal{E}$.

ABSTAND WINDSCHIEFER GERADEN

Gegeben seien zwei Geraden

$$g_1 = \{x_1 + tv_1 \mid t \in \mathbb{R}\}$$

und

$$g_2 = \{x_2 + tv_2 \mid t \in \mathbb{R}\}$$

durch Stützpunkte x_i und in Richtungen v_i mit $v_1 \times v_2 \neq 0$.

Wir fragen nach dem Abstand beider Geraden, also der Länge derjenigen Verbindungsstrecke, welche auf beiden Geraden senkrecht steht. Dieser Abstand ist insbesondere gleich dem Abstand der beiden parallelen Ebenen $\mathcal{E}_1$ und $\mathcal{E}_2$ zu den Richtungsvektoren v_1 und v_2, die jeweils x_1 beziehungsweise x_2 enthalten.

Nun ist der orientierte Abstand der Ebene $\mathcal{E}_i$ zum Ursprung durch

$$d_i := \frac{\langle x_i, v_1 \times v_2 \rangle}{|v_1 \times v_2|}$$

gegeben. Der Abstand von $\mathcal{E}_1$ zu $\mathcal{E}_2$ berechnet sich damit durch

$$|d_1 - d_2| = \left| \frac{\langle x_1, v_1 \times v_2 \rangle}{|v_1 \times v_2|} - \frac{\langle x_2, v_1 \times v_2 \rangle}{|v_1 \times v_2|} \right| = \frac{|\langle x_1 - x_2, v_1 \times v_2 \rangle|}{|v_1 \times v_2|}.$$

SCHNITT ZWISCHEN GERADE UND EBENE

Gegeben seien eine Gerade

$$g = \{x_0 + tv \mid t \in \mathbb{R}\}$$

und eine Ebene in Hessescher Normalform

$$\mathcal{E} = \{x \mid \langle x, n\rangle = d\}.$$

Wir fragen nach möglichen Schnittpunkten von der Geraden g und der Ebene $\mathcal{E}$. Angenommen es gibt ein $x \in g \cap \mathcal{E}$. Dann gibt es ein $t \in \mathbb{R}$ mit $x = x_0 + tv$ und es gilt

$$d = \langle x_0 + tv, n\rangle = \langle x_0, n\rangle + t\langle v, n\rangle\,.$$

Im Fall $\langle v, n\rangle \neq 0$ ergibt dies

$$t = \frac{d - \langle x_0, n\rangle}{\langle v, n\rangle}$$

und Einsetzen von t in die Geradengleichung liefert einzigen Schnittpunkt x.

Gilt andererseits $\langle v, n\rangle = 0$, also $v \perp n$, so ist der Richtungsvektor der Geraden parallel zur Ebene und es gilt entweder $g \subset \mathcal{E}$ oder $g \cap \mathcal{E} = \emptyset$.

SCHNITT ZWEIER EBENEN

Wir betrachten nun den Schnitt zweier Ebenen in Hessescher Normalform

$$\mathcal{E}_1 = \{x \in \mathbb{R}^3 \mid \langle x, n_1 \rangle = d_1\},$$
$$\mathcal{E}_2 = \{x \in \mathbb{R}^3 \mid \langle x, n_2 \rangle = d_2\}.$$

Sind die Ebenen parallel, gilt also $n_1 = \pm n_2$, so gilt entweder $\mathcal{E}_1 = \mathcal{E}_2$ oder $\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset$.

Sind die Ebenen nicht parallel, so sind n_1 und n_2 nicht parallel und damit gilt sowohl $v := n_1 \times n_2 \neq 0$ als auch $\langle n_1, n_2 \rangle^2 \neq 1$, wobei letzteres z.B. aus Satz 3.1.3 folgt.

Wir nehmen nun zunächst an, wir haben schon ein $x_0 \in \mathcal{E}_1 \cap \mathcal{E}_2$ gefunden. Wir wollen dann

$$\mathcal{E}_1 \cap \mathcal{E}_2 = g := \{x_0 + tv \mid t \in \mathbb{R}\}$$

zeigen. Dazu sei zunächst $t \in \mathbb{R}$ und $x := x_0 + tv$. Dann gilt

$$\langle x, n_i \rangle = \langle x_0, n_i \rangle + t\langle v, n_i \rangle = \langle x_0, n_i \rangle + t\langle n_1 \times n_2, n_i \rangle = \langle x_0, n_i \rangle = d_i \,.$$

Damit haben wir $\{x_0 + tv \mid t \in \mathbb{R}\} \subset \mathcal{E}_1 \cap \mathcal{E}_2$ gezeigt. Die andere Inklusion kann man jetzt auch formal beweisen, da aber der Schnitt von $\mathcal{E}_1$ und $\mathcal{E}_2$ aus geometrischer Anschauung im Fall $n_1 \times n_2 \neq 0$ eine Gerade ist, kann es keinen weiteren Punkt $x \in (\mathcal{E}_1 \cap \mathcal{E}_2) \setminus g$ geben. Aus diesem Grund überspringen wir den formalen Beweis der anderen Inklusion.

Um g vollständig zu bestimmen, müssen wir also noch ein $x_0 \in \mathcal{E}_1 \cap \mathcal{E}_2$ finden. Dies könnte man z.B. dadurch erreichen, dass man eine Lösung x_0 des linearen Gleichungssystems

$$\begin{aligned} \langle x, n_1 \rangle &= d_1 \,, \\ \langle x, n_2 \rangle &= d_2 \end{aligned}$$

bestimmt. Hier wollen wir stattdessen zeigen, dass

$$x_0 := d_1 n_1 + \frac{d_2 - d_1 \langle n_1, n_2 \rangle}{\langle n_1, n_2 \rangle^2 - 1} \cdot \Big(\langle n_1, n_2 \rangle n_1 - n_2 \Big)$$

eine explizite Lösung ist. Dazu betrachten wir zunächst

$$\begin{aligned} \langle x_0, n_1 \rangle &= \langle d_1 n_1, n_1 \rangle + \frac{d_2 - d_1 \langle n_1, n_2 \rangle}{\langle n_1, n_2 \rangle^2 - 1} \cdot \Big\langle \langle n_1, n_2 \rangle n_1 - n_2, n_1 \Big\rangle \\ &= d_1 + \frac{d_2 - d_1 \langle n_1, n_2 \rangle}{\langle n_1, n_2 \rangle^2 - 1} \cdot \big(\langle n_1, n_2 \rangle \cdot \langle n_1, n_1 \rangle - \langle n_1, n_2 \rangle \big) \\ &= d_1 \,, \end{aligned}$$

wobei wir zweimal $\langle n_1, n_1 \rangle = 1$ ausgenutzt haben.

Analog gilt

$$\begin{aligned}\langle x_0, n_2\rangle &= \langle d_1 n_1, n_2\rangle + \frac{d_2 - d_1\langle n_1, n_2\rangle}{\langle n_1, n_2\rangle^2 - 1} \cdot \big\langle \langle n_1, n_2\rangle n_1 - n_2, n_2 \big\rangle \\ &= d_1\langle n_1, n_2\rangle + \frac{d_2 - d_1\langle n_1, n_2\rangle}{\langle n_1, n_2\rangle^2 - 1} \cdot \big(\langle n_1, n_2\rangle^2 - \langle n_2, n_2\rangle\big) \\ &= d_2\,,\end{aligned}$$

wobei wir im letzten Schritt $\langle n_2, n_2\rangle = 1$ ausgenutzt haben.

Chapter 4: Grenzwerte

Section 4.1
Konvergenz von Folgen

Im $\mathbb{R}^d$ haben wir neben dem Betrag schon weitere Normen kennengelernt. Wir wollen nun diese Betrachtungen noch etwas weiter verallgemeinern, indem wir einen allgemeinen Abstandsbegriff für Paare von Objekten einführen.

Definition 4.1.1

Sei X eine nichtleere Menge. Dann heißt eine Abbildung $\mathsf{d} : X \times X \to [0, \infty)$ Metrik auf X, falls für alle $x, y, z \in X$ die folgenden Eigenschaften erfüllt sind:

i). Definitheit: $\mathsf{d}(x, y) = 0 \Leftrightarrow x = y$.

ii). Symmetrie: $\mathsf{d}(x, y) = \mathsf{d}(y, x)$.

iii). Dreiecksungleichung: $\mathsf{d}(x, y) \leq \mathsf{d}(x, z) + \mathsf{d}(z, y)$.

In diesem Fall heißt (X, d) metrischer Raum.

Ist $\|\cdot\|$ eine Norm auf dem Vektorraum V, so definiert $\mathsf{d}(x, y) := \|x - y\|$ eine Metrik auf V. In diesem Sinne ist jeder normierte Raum auch ein metrischer Raum. Die Umkehrung gilt im Allgemeinen jedoch nicht. Eine einfaches Beispiel hierfür sind echte Teilmengen $X \subset V$, die ebenfalls mit Hilfe der Norm zu einem metrischen Raum werden. Es gibt aber auch Beispiele von Metriken, die keinen Zusammenhang mit Normen besitzen.

WEITERE BEGRIFFE FÜR METRISCHE RÄUME

Ist (X, d) ein metrischer Raum, so gilt auch die **umgekehrte Dreiecksungleichung**

$$\big|\mathrm{d}(x,z) - \mathrm{d}(z,y)\big| \leq \mathrm{d}(x,y)$$

für alle $x, y, z \in X$.

Ist (X, d) ein metrischer Raum, $x \in X$ und $r > 0$, so heißt

$$U(x,r) := \{y \in X : \mathrm{d}(x,y) < r\}$$

die **offene Kugel** um x mit Radius r. Analog bezeichnen wir

$$B(x,r) := \{y \in X : \mathrm{d}(x,y) \leq r\}$$

als die **abgeschlossene Kugel** um x mit Radius r.

In $\mathbb{R}$ gilt $U(x,r) = (x-r, x+r)$ und $B(x,r) = [x-r, x+r]$. Im $\mathbb{R}^2$ und $\mathbb{C}$ sind die offenen, bzw. abgeschlossenen Kugeln bezüglich der Betragsmetrik die Kreise ohne bzw. mit Rand.

Lemma 4.1.2

Sei $(X, \mathbf{d})$ *ein metrischer Raum,* $x, y \in X$*,* $r > 0$ *und* $s \geq r + \mathbf{d}(x, y)$*. Dann gilt*

$$U(y, r) \subset U(x, s) \qquad \textit{und} \qquad B(y, r) \subset B(x, s)\,.$$

Beweis.

Sei $z \in U(y, r)$. Dann gilt $\mathbf{d}(y, z) < r$ und daher folgt

$$\mathbf{d}(x, z) \leq \mathbf{d}(x, y) + \mathbf{d}(y, z) < \mathbf{d}(x, y) + r \leq s\,.$$

Dies zeigt $z \in U(x, s)$. Die zweite Inklusion kann analog gezeigt werden. □

Abstände zwischen Mengen

Ist (X, d) ein metrischer Raum, $x \in X$ und $A \subset X$ nichtleer, so heißt

$$\mathbf{dist}(x, A) := \mathbf{inf}\{\mathrm{d}(x, y) : y \in A\}$$

der Abstand von x zu der Menge A. Ist $B \subset X$ eine weitere, nichtleere Menge, so heißt

$$\mathbf{dist}(A, B) := \mathbf{inf}\{\mathrm{d}(x, y) : x \in A, y \in B\}$$

der Abstand zwischen den Mengen A und B.

Offensichtlich gilt $\mathbf{dist}(x, A) = \mathbf{dist}(\{x\}, A)$ und $\mathbf{dist}(A, B) = \mathbf{dist}(B, A)$. Ist ferner $A \subset B$, so gilt

$$\mathbf{dist}(x, B) \leq \mathbf{dist}(x, A)\,, \qquad x \in X.$$

Ist $x \in A$, so gilt $\mathbf{dist}(x, A) = 0$. Dies ist aber nicht der einige Fall, in dem der Abstand verschwinden kann. So gilt beispielsweise in $\mathbb{R}$ mit der Betragsmetrik für $A = (0, 1)$ und $x = 0$ ebenfalls $\mathbf{dist}(x, A) = 0$.

Analog gilt $\mathbf{dist}(A, B) = 0$ für Mengen A und B mit $A \cap B \neq \emptyset$, aber dies ist wieder nicht der einzige Fall, wie man schon aus der Kombination der vorherigen Bemerkungen schließen kann.

Im folgenden wollen wir die Konvergenz von Folgen untersuchen. Dazu führen wir den Begriff einer Folge zunächst formal ein:

Definition 4.1.3

Sei X eine nichtleere Menge. Dann ist eine **Folge** $(x_n)_{n\in\mathbb{N}}$ in X eine Abbildung $\mathbb{N} \to X$, deren Funktionswert im Punkt $n \in \mathbb{N}$ mit x_n bezeichnet wird.

Manchmal werden wir auch Folgen mit Indexbereich $\mathbb{N}_0$ betrachten. Folgen werden auch mit $(x_n)_{n\geq 1}$ bzw. $(x_n)_{n\geq 0}$ bezeichnet. Wir schreiben auch manchmal $(x_n)_{n\geq 1} \subset X$, um auf kurze Weise zu sagen, dass die Folge $(x_n)_{n\geq 1}$ in X liegt.

Die folgenden Definitionen sind von grundlegender Bedeutung für den Rest des Vorlesungszyklus:

Definition 4.1.4

Sei $(X, \mathbf{d})$ ein metrischer Raum und $(x_n)_{n\in\mathbb{N}}$ eine Folge in X. Dann heißt die Folge:

i). **beschränkt**, falls es eine $B > 0$ und ein $y \in X$ gibt mit $\mathbf{d}(x_n, y) \leq B$ für alle $n \geq 1$.

ii). **Cauchyfolge**, falls gilt:

$$\forall \varepsilon > 0 \exists n_\varepsilon \geq 1 \forall n, m \geq n_\varepsilon : \mathbf{d}(x_n, x_m) < \varepsilon \,.$$

iii). **konvergent**, falls es ein $x \in X$ gibt, so dass gilt:

$$\forall \varepsilon > 0 \exists n_\varepsilon \geq 1 \forall n \geq n_\varepsilon : \mathbf{d}(x_n, x) < \varepsilon \,.$$

In diesem Fall heißt x **Grenzwert** oder **Limes** der Folge und wir sagen, dass die Folge **gegen x konvergiert**.

KONVERGENZ VON FOLGEN: BEISPIEL

In $\mathbb{R}$ ist die Folge, die durch $x_n := 1/n$ definiert ist, wegen

$$|x_n - 0| = |1/n| \leq 1$$

beschränkt, da wir $y := 0$ und $B := 1$ wählen können. Sei nun $\varepsilon > 0$. Nach dem Archimedischen Axiom, siehe Satz 2.4.4, finden wir dann ein $n_\varepsilon \in \mathbb{N}$ mit $n_\varepsilon^{-1} < \varepsilon$. Für $n \geq n_\varepsilon$ folgt dann mit $x := 0$:

$$|x_n - x| = |1/n| \leq |1/n_\varepsilon| < \varepsilon\,.$$

Mit anderen Worten ist die Folge auch konvergent, und nach dem etwas späteren Satz 4.1.6 damit auch eine Cauchyfolge.

Schließlich bemerken wir, dass die Folge $x_n := n^{-1/k}$ für $k \in \mathbb{N}$ ebenfalls gegen 0 konvergiert. Für den Beweis nehmen wir mit dem Archimedischen Axiom ein $n_\varepsilon \in \mathbb{N}$ mit $n_\varepsilon^{-1} < \varepsilon^k$ und wiederholen die obige Abschätzung.

Die Folge, die durch $x_n := n^2$ definiert ist, weder beschränkt noch konvergent, und wie wir in Satz 4.1.6 sehen werden, auch keine Cauchy-Folge.

Reelle Folgen, die gegen 0 konvergieren, werden auch **Nullfolgen** genannt. Analog gilt dies für Folgen in $\mathbb{C}$ oder in normierten Räumen.

Eine Folge $(x_n)_{n\in\mathbb{N}}$ in einem normierten Raum $(V, \|\cdot\|)$ konvergiert gegen ein x, genau dann wenn

$$y_n := \|x_n - x\|$$

eine reelle Nullfolge definiert. Dies folgt sofort aus der Gleichung

$$\|x_n - x\| = \big|\|x_n - x\| - 0\big| .$$

Im Fall $x = 0$ sehen wir damit, dass $(x_n)_{n\in\mathbb{N}} \subset V$ eine Nullfolge ist, genau dann wenn $(\|x_n\|)_{n\geq 1}$ eine reelle Nullfolge ist. Dies gilt insbesondere für die Betragsmetrik auf $\mathbb{R}$.

BESCHRÄNKTE UND KONSTANTE FOLGEN

In einem normierten Raum ist eine Folge $(x_n)_{n\in\mathbb{N}}$ beschränkt genau dann, wenn es ein $\tilde{B} \geq 0$ gibt mit $\|x_n\| \leq \tilde{B}$ für alle $n \geq 1$. Die nicht ganz triviale Richtung folgt hierbei aus $\|x_n\| \leq \|x_n - y\| + \|y\| \leq B + \|y\| =: \tilde{B}$.

Eine konstante Folge $(x_n)_{n\in\mathbb{N}}$ in einem metrischen Raum $(X, \mathbf{d})$, d.h. eine Folge mit $x_n = x^*$ für ein $x^* \in X$ und alle $n \geq 1$, konvergiert immer gegen x^*, da $\mathbf{d}(x_n, x^*) = 0$ für alle $n \geq 1$ gilt.

Das folgende Lemma zeigt, dass eine konvergente Folge genau einen Grenzwert hat.

Lemma 4.1.5

Sei (X, d) *ein metrischer Raum und* $(x_n)_{n\in\mathbb{N}}$ *eine konvergente Folge in* X*. Dann gibt es genau einen Grenzwert der Folge.*

Wegen des Lemmas 4.1.5 bezeichnen wir den eindeutigen Grenzwert einer Folge $(x_n)_{n\in\mathbb{N}}$ mit $\lim_{n\to\infty} x_n$. Ferner schreiben wir auch $x_n \to x$ für $n \to \infty$, falls die Folge $(x_n)_{n\in\mathbb{N}}$ gegen x konvergiert.

BEWEIS

Wir nehmen an, es gibt zwei verschiedene Grenzwerte x und y. Damit gilt $\varepsilon := \mathsf{d}(x, y)/2 > 0$. Da die Folge gegen x konvergiert, gibt es dann ein $n_\varepsilon \geq 1$ mit

$$\mathsf{d}(x_n, x) < \varepsilon$$

für alle $n \geq n_\varepsilon$. Analog gibt es ein $m_\varepsilon \geq 1$ mit

$$\mathsf{d}(x_n, y) < \varepsilon$$

für alle $n \geq m_\varepsilon$. Für $n := n_\varepsilon + m_\varepsilon$ haben wir dann $n \geq \mathsf{max}\{n_\varepsilon, m_\varepsilon\}$ und damit folgt

$$\mathsf{d}(x, y) \leq \mathsf{d}(x, x_n) + \mathsf{d}(x_n, y) < 2\varepsilon = \mathsf{d}(x, y)\,.$$

Damit haben wir einen Widerspruch gefunden.

Der folgende Satz stellt einen Zusammenhang zwischen den 3 Begriffen für Folgen her.

Theorem 4.1.6

Sei (X, d) ein metrischer Raum und $(x_n)_{n\in\mathbb{N}}$ eine Folge in X. Dann gilt:

i). Ist $(x_n)_{n\in\mathbb{N}}$ konvergent, so ist $(x_n)_{n\in\mathbb{N}}$ eine Cauchyfolge.

ii). Ist $(x_n)_{n\in\mathbb{N}}$ eine Cauchyfolge, so ist $(x_n)_{n\in\mathbb{N}}$ beschränkt.

i). Sei x der Grenzwert der Folge und $\varepsilon > 0$. Dann gibt es ein $n_\varepsilon \geq 1$, so dass für alle $n \geq n_\varepsilon$ die Ungleichung $\mathbf{d}(x, x_n) < \varepsilon/2$ gilt. Für $m, n \geq n_\varepsilon$ haben wir damit

$$\mathbf{d}(x_n, x_m) \leq \mathbf{d}(x_n, x) + \mathbf{d}(x, x_m) < \varepsilon/2 + \varepsilon/2 = \varepsilon\,.$$

ii). Für $\varepsilon := 1$ gibt es ein $n_\varepsilon \geq 1$ so dass $\mathbf{d}(x_n, x_m) < 1$ für alle $n, m \geq n_\varepsilon$ gilt. Wir definieren $y := x_{n_\varepsilon}$ und

$$B := 1 + \mathbf{max}\{\mathbf{d}(x_n, y) : 1 \leq n < n_\varepsilon\}\,.$$

Für $n < n_\varepsilon$ haben wir damit $\mathbf{d}(x_n, y) \leq \mathbf{max}\{\mathbf{d}(x_k, y) : 1 \leq k < n_\varepsilon\} < B$ und für $n \geq n_\varepsilon$ gilt $\mathbf{d}(x_n, y) = \mathbf{d}(x_n, x_{n_\varepsilon}) < 1 \leq B$.

Die Umkehrungen des Satzes 4.1.6 sind im Allgemeinen falsch.

So ist beispielsweise die Folge $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$, die durch $x_n := (-1)^n$ für alle $n \geq 1$ definiert ist, beschränkt mit $y := 0$ and $B := 1$, sie ist aber keine Cauchyfolge, da für alle $n \geq 1$

$$|x_n - x_{n+1}| = |(-1)^n| \cdot |1 - (-1)^1| = 2$$

gilt.

Betrachten wir wieder die Folge, die durch $x_n := 1/n$ definiert ist, so haben wir bereits gesehen, dass $\lim x_n = 0$ gilt. Damit ist die Folge nach Satz 4.1.6 eine Cauchyfolge. Betrachten wir diese Folge nun in dem Intervall $X := (0, \infty)$, das wieder mit der Betragsmetrik $\mathbf{d}$ ausgestattet ist, so ist die Folge weiterhin eine Cauchyfolge. Sie ist aber in $(X, \mathbf{d})$ nicht mehr konvergent, da der einzig mögliche Grenzwert $x := 0$ nicht in X liegt!

Wir werden später allerdings sehen, dass, im Unterschied zu $\mathbb{Q}$, in $\mathbb{R}$ jede Cauchyfolge konvergent ist.

Mit Grenzwerten reeller Zahlenfolgen kann man Rechnen, wie der folgende Satz zeigt. Man beachte, dass in diesen Satz jeweils gleichzeitig Konvergenz gezeigt und ein Grenzwert berechnet wird.

Theorem 4.1.7

Seien $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ und $(y_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ zwei konvergente Folgen. Dann gilt:

i). Die Folge $(x_n + y_n)_{n\in\mathbb{N}}$ konvergiert und es gilt

$$\lim_{n\to\infty} (x_n + y_n) = \lim_{n\to\infty} x_n + \lim_{n\to\infty} y_n \,.$$

ii). Die Folge $(x_n \cdot y_n)_{n\in\mathbb{N}}$ konvergiert und es gilt

$$\lim_{n\to\infty} (x_n \cdot y_n) = \lim_{n\to\infty} x_n \cdot \lim_{n\to\infty} y_n \,.$$

iii). Ist $x_n \neq 0$ für alle $n \geq 1$ und $\lim_{n\to\infty} x_n \neq 0$, so konvergiert $(x_n^{-1})_{n\in\mathbb{N}}$ und es gilt

$$\lim_{n\to\infty} \frac{1}{x_n} = \frac{1}{\lim_{n\to\infty} x_n} \,.$$

iv). Ist $x_n \leq y_n$ für alle $n \geq 1$, so gilt $\lim_{n\to\infty} x_n \leq \lim_{n\to\infty} y_n$.

Wir wollen exemplarisch nur die Aussagen *ii)* und *iv)* zeigen. Dazu seien $x := \lim_{n\to\infty} x_n$ und $y := \lim_{n\to\infty} y_n$.

ii). Da $(x_n)_{n\in\mathbb{N}}$ konvergiert, gibt es nach Satz 4.1.6 ein $B > 0$ mit $|x_n| \leq B$ für alle $n \geq 1$. Sei nun $\varepsilon > 0$. Dann gibt es $n_\varepsilon \geq 1$ und $m_\varepsilon \geq 1$ mit $|x_n - x| < \varepsilon/(2|y| + 1)$ für alle $n \geq n_\varepsilon$ bzw. $|y_n - y| < \varepsilon/(2B)$ für alle $n \geq m_\varepsilon$. Für $n \geq \max\{n_\varepsilon, m_\varepsilon\}$ folgt dann

$$\begin{aligned} |x_n y_n - xy| \leq |x_n y_n - x_n y| + |x_n y - xy| &\leq |x_n| \cdot |y_n - y| + |x_n - x| \cdot |y| \\ &\leq B \cdot \frac{\varepsilon}{2B} + \frac{\varepsilon}{2|y| + 1} \cdot |y| \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\,, \end{aligned}$$

wobei wir $|y|/(2|y| + 1) \leq 1/2$ für alle $y \in \mathbb{R}$ ausgenutzt haben.

iv). Wir nehmen an, dass stattdessen $x > y$ gilt. Wir betrachten dann $\varepsilon := (x-y)/2$. Dann gibt es $n_\varepsilon \geq 1$ und $m_\varepsilon \geq 1$ mit $|x_n - x| < \varepsilon$ für alle $n \geq n_\varepsilon$ bzw. $|y_n - y| < \varepsilon$ für alle $n \geq m_\varepsilon$. Für $n \geq \max\{n_\varepsilon, m_\varepsilon\}$ folgt dann einerseits

$$x - x_n \leq |x_n - x| < \varepsilon = \frac{x-y}{2}$$

und damit $x_n > (x+y)/2$, und andererseits

$$y_n - y \leq |y_n - y| < \varepsilon = \frac{x-y}{2}$$

und damit $y_n < (x+y)/2$. Zusammen ergibt dies $y_n < (x+y)/2 < x_n$ und damit haben wir einen Widerspruch gefunden.

Da konstante Folgen konvergent sind, folgt aus *ii)* für eine konvergente Folge $(x_n)_{n\in\mathbb{N}}$ und $\alpha \in \mathbb{R}$ insbesondere, dass $(\alpha x_n)_{n\in\mathbb{N}}$ konvergent ist mit

$$\lim_{n\to\infty} (\alpha \cdot x_n) = \alpha \cdot \lim_{n\to\infty} x_n \,.$$

Diese Aussage und *i)* gelten sogar in allen normierten Räumen, wobei die Beweise analog zu denen in $\mathbb{R}$ sind.

Sind die Annahmen $x_n \neq 0$ in *iii)* und $x_n \leq y_n$ in *iv)* nur für alle $n \geq n_0$ erfüllt, so gelten weiterhin die Aussagen.

Das folgende Korollar ist manchmal ebenfalls nützlich.

Korollar 4.1.8

Sei $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ *und* $k \in \mathbb{N}$. *Dann gilt* $x_n \to 0$ *genau dann, wenn* $x_n^k \to 0$.

Beweis.

Die Konvergenz $x_n^k \to 0$ folgt durch $(k-1)$-maliges Anwenden von Satz 4.1.7.
Sei umgekehrt $\varepsilon > 0$. Dann gibt es ein $n_\varepsilon \geq 1$ mit $|x_n^k| < \varepsilon^k$ für alle $n \geq n_\varepsilon$.
Dies ergibt $|x_n| < \varepsilon$ und damit haben wir $x_n \to 0$ gezeigt. □

Der folgende Satz ist als Vergleichskriterium bekannt. Man beachte dabei, dass die wesentliche Aussage die Konvergenz der Zwischenfolge $(z_n)_{n\in\mathbb{N}}$ ist: In der Tat folgt die Gleichheit der Grenzwerte schon direkt aus zweimaligem Anwenden von Satz 4.1.7 *iv)*, falls die Konvergenz von $(z_n)_{n\in\mathbb{N}}$ zusätzlich bekannt ist.

Theorem 4.1.9

Seien $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ und $(z_n)_{n\in\mathbb{N}}$ Folgen, für die es ein $n_0 \geq 1$ gibt mit

$$x_n \leq z_n \leq y_n$$

für alle $n \geq n_0$. Konvergieren dann die Folgen $(x_n)_{n\in\mathbb{N}}$ und $(y_n)_{n\in\mathbb{N}}$ mit $\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n$, so konvergiert auch die Folge $(z_n)_{n\in\mathbb{N}}$ und es gilt

$$\lim_{n\to\infty} z_n = \lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n \,.$$

Wir setzen $c := \lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n$. Für $\varepsilon > 0$ gibt es dann ein $n_\varepsilon \geq 1$ mit $|x_n - c| < \varepsilon$ und $|y_n - c| < \varepsilon$ für alle $n \geq n_\varepsilon$. Für diese n gilt dann auch

$$z_n - c \leq y_n - c \leq |y_n - c| < \varepsilon$$

und analog $c - z_n \leq c - x_n \leq |x_n - c| < \varepsilon$. Wegen $|a| = \max\{a, -a\}$ folgt dann zusammengenommen

$$|z_n - c| < \varepsilon$$

für alle $n \geq 1$.

Das folgende Lemma führt die Konvergenz im $\mathbb{R}^d$ auf die Konvergenz im Reellen zurück. Analoge Aussagen gelten auch für die Supremumsnorm und die 1-Norm:

Lemma 4.1.10

Sei $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^d$ eine Folge und $x^* \in \mathbb{R}^d$. Dann sind die folgenden Aussagen äquivalent:

1. *$x_n \to x^*$ bezüglich der Betragsmetrik.*
2. *Für alle $i = 1, \ldots, d$ konvergiert die Folge $(x_n^{(i)})_{n\in\mathbb{N}}$ der i-ten Koordinaten der Folge $(x_n)_{n\in\mathbb{N}}$ gegen die i-te Koordinate x_i^* von x^*.*

Bevor wir das Lemma beweisen, bemerken wir noch dass damit in $\mathbb{C}$ eine Folge $(z_n)_{n\in\mathbb{N}}$ konvergiert genau dann, wenn die Folgen der Real- und Imaginärteile konvergieren.

Unsere Notationen liefern

$$x_n = \begin{pmatrix} x_n^{(1)} \\ \vdots \\ x_n^{(d)} \end{pmatrix} \qquad \text{und} \qquad x^* = \begin{pmatrix} x_1^* \\ \vdots \\ x_d^* \end{pmatrix}.$$

Damit gilt

$$|x_n - x^*|^2 = (x_n^{(1)} - x_1^*)^2 + \cdots + (x_n^{(d)} - x_d^*)^2 \tag{4.1.1}$$

i) $\Rightarrow$ *ii)*. Wir betrachten die i-te Koordinate. Aus der angenommenen Konvergenz $x_n \to x^*$ folgt $|x_n - x^*| \to 0$ und mit Satz 4.1.7 dann auch $|x_n - x^*|^2 \to 0$. Unsere anfängliche Betrachtung (4.1.1) liefert ferner

$$0 \leq (x_n^{(i)} - x_i^*)^2 \leq |x_n - x^*|^2 .$$

Das Vergleichskriterium aus Satz 4.1.9 sichert dann $(x_n^{(i)} - x_i^*)^2 \to 0$ und wegen Korollar 4.1.8 erhalten wir $x_n^{(i)} - x_i^* \to 0$, d.h. $x_n^{(i)} \to x_i^*$ nach Satz 4.1.7.

ii) $\Rightarrow$ *i)*. Die Gleichung (4.1.1) zusammen mit Korollar 4.1.8 und Satz 4.1.7 ergibt $|x_n - x^*|^2 \to 0$. Mit Korollar 4.1.8 erhalten wir $|x_n - x^*| \to 0$ und damit $x_n \to x^*$ bezüglich der Betragsmetrik.

BEISPIELE KONVERGENTER FOLGEN IN R

Wir fixieren zunächst ein $k \in \mathbb{Z}$ und definieren die Folge $(x_n)_{n\in\mathbb{N}}$ durch

$$x_n := n^{-k}\,.$$

Für $k = 0$ gilt dann $x_n = 1$ für alle $n \geq 1$ und damit konvergiert diese konstante Folge gegen 1. Ist $k \geq 1$, so gilt wegen $\lim_{n\to\infty} n^{-1} = 0$ und Korollar 4.1.8

$$\lim_{n\to\infty} x_n = 0\,. \tag{4.1.2}$$

Mit dem Archimedischen Axiom kann man schließlich schnell zeigen, dass im Fall $k < 0$ die Folge $(x_n)_{n\in\mathbb{N}}$ unbeschränkt ist.

Betrachten wir nun die Folge $(x_n)_{n\in\mathbb{N}}$, die durch $x_n := n^{-1} \sin n$ definiert ist. Wegen

$$-\frac{1}{n} \leq \frac{\sin n}{n} \leq \frac{1}{n}\,,$$

folgt dann mit dem Vergleichskriterium aus Satz 4.1.9 $\lim_{n\to\infty} \frac{\sin n}{n} = 0$.

Lemma 4.1.11

Es gilt $\lim_{n\to\infty} \sqrt[n]{n} = 1$.

Beweis.

Wir betrachten die Folge $a_n := \sqrt[n]{n} - 1$. Es gilt $a_n \geq 0$ für alle $n \geq 1$ und nach dem binomischen Satz gilt für alle $n \geq 2$

$$n = (a_n + 1)^n = 1 + na_n + \frac{n(n-1)}{2} a_n^2 + \sum_{k=3}^{n} \binom{n}{k} a_n^k$$
$$\geq 1 + \frac{n(n-1)}{2} a_n^2 ,$$

wobei der letzte Schritt wahr ist, da wir nichtnegative Summanden weggelassen haben. Es folgt

$$0 \leq a_n \leq \frac{\sqrt{2}}{\sqrt{n}}$$

für $n \geq 2$ folgt. Damit erhalten wir $a_n \longrightarrow 0$ mit Hilfe von $\lim_{n\to\infty} 1/\sqrt{n} = 0$, von Satz 4.1.7, und dem Vergleichskriterium aus Satz 4.1.9.
Satz 4.1.7 liefert dann die Behauptung. □

Mit Satz 4.1.7 folgt

$$\lim_{n\to\infty} \frac{n+1}{n} = \lim_{n\to\infty} \left(1+\frac{1}{n}\right) = 1 + \lim_{n\to\infty} \frac{1}{n} = 1\,,$$

wobei die Gleichung von rechts nach links gelesen einen Beweis der Existenz des Grenzwertes auf der linken Seite liefert, und das anschließende Lesen der Gleichung von links nach rechts die Berechnung des Grenzwertes ergibt.

Die Kombination von Satz 4.1.7 mit (4.1.2) ergibt

$$\begin{aligned}\lim_{n\to\infty} \left(\frac{6n^4+3n^2+2}{7n^4+12n^3+6}\right)^2 &= \left(\lim_{n\to\infty} \frac{6n^4+3n^2+2}{7n^4+12n^3+6}\right)^2 \\ &= \left(\lim_{n\to\infty} \frac{6+3n^{-2}+2n^{-4}}{7+12n^{-1}+6n^{-4}}\right)^2 \\ &= \left(\frac{6}{7}\right)^2 \\ &= \frac{36}{49}\,,\end{aligned}$$

wobei die Existenz und die Berechnung des Grenzwertes wie im Vorherigen Beispiel zu verstehen ist.

Bevor wir noch die Konvergenz von ein paar weiteren Folgen betrachten können, benötigen wir die Folgende Ungleichung, die Bernoulli-Ungleichung genannt wird.

Lemma 4.1.12

Für alle $x \geq -1$ und alle $n \in \mathbb{N}$ gilt

$$(1+x)^n \geq 1 + nx.$$

BEWEIS

Wir beweisen die Ungleichung per Induktion. Der Induktionsanfang ist durch $(1+x)^1 = (1+x) \geq 1+x$ gegeben. Angenommen, die Ungleichung wurde schon für ein n gezeigt. Dann folgt

$$\begin{aligned}(1+x)^{n+1} &= (1+x)^n(1+x) \\ &\geq (1+nx)(1+x) \\ &= 1+nx+x+nx^2 \\ &\geq 1+(n+1)x\end{aligned}$$

und nach dem Induktionsprinzip gilt dann die Aussage für alle $n \in \mathbb{N}$.

Lemma 4.1.13

Sei $a \in \mathbb{R}$ und $(x_n)_{n\in\mathbb{N}}$ die durch

$$x_n := a^n$$

definierte Folge. Dann gilt:

i). Für $|a| > 1$ ist die Folge unbeschränkt.

ii). Für $|a| < 1$ konvergiert die Folge gegen 0.

iii). Für $a = 1$ konvergiert die Folge gegen 1.

iv). Für $a = -1$ ist die Folge beschränkt aber nicht konvergent.

GEOMETRISCHE FOLGEN: BEWEIS

Die Fälle *iii)* und *iv)* wurden schon behandelt.

i). Wir setzen $y := |a| - 1$. Dann gilt $y > 0$ und mit der Ungleichung von Bernoulli folgt

$$|a^n| = (1+y)^n \geq 1 + ny\,.$$

Geben wir uns nun ein $M > 0$ vor, so gilt

$$1 + ny > M \qquad \Longleftrightarrow \qquad n > \frac{M-1}{y}$$

und damit $|a^n| > M$ für ebensolche n.

ii). Im Fall $a = 0$ ist die Aussage trivial und daher betrachten wir nur noch den Fall $a \neq 0$.

Wir setzen nun $y := \frac{1}{|a|} - 1$. Dann gilt $y > 0$ und somit folgt mit der Bernoulli-Ungleichung

$$0 \leq |a^n| = |a|^n = \frac{1}{(1+y)^n} \leq \frac{1}{1+ny} =: y_n.$$

Wegen $y_n \to 0$ folgt dann mit dem Vergleichskriterium aus Satz 4.1.9, dass $|a^n - 0|$ eine Nullfolge ist. Dies impliziert $a_n \to 0$.

Lemma 4.1.14

Sei $a \geq 1$. Dann gilt

$$\lim_{n \to \infty} \sqrt[n]{a} = 1 .$$

BEWEIS

Wir definieren uns eine Folge $(y_n)_{n\in\mathbb{N}}$ durch

$$y_n := \sqrt[n]{a} - 1\,.$$

Es gilt $y_n \geq 0$ und die Bernoulli-Ungleichung ergibt

$$a = (1 + y_n)^n \geq 1 + ny_n\,.$$

Insgesamt haben wir damit

$$\frac{a-1}{n} \geq y_n \geq 0\,.$$

Wegen $\frac{a-1}{n} \to 0$ folgt dann $y_n \to 0$ mit dem Vergleichskriterium aus Satz 4.1.9. Somit erhalten wir $\sqrt[n]{a} \to 1$.

Wir hatten gesehen, dass die Implikationen des Satzes 4.1.6 im Allgemeinen keine Äquivalenzen sind. Im folgenden wollen wir jedoch eine Klasse von beschränkten Folgen kennenlernen, für die aus der Beschränktheit schon die Konvergenz folgt.

Wir beginnen mit der folgenden Definition.

Definition 4.1.15

Eine reelle Folge $(x_n)_{n\in\mathbb{N}}$ heißt **monoton wachsend**, falls für alle $n \in \mathbb{N}$ die Ungleichung $x_{n+1} \geq x_n$ gilt. Sie heißt **monoton fallend**, falls $x_{n+1} \leq x_n$ für alle $n \in \mathbb{N}$ gilt. Ist eine Folge monoton wachsend oder monoton fallend, so heißt sie **monoton**.

Die Folge, die durch $x_n := n^{-k}$ für $k \in \mathbb{N}$ definiert ist, ist monoton fallend, während die durch $y_n := (-1)^n n^{-1}$ definierte Folge nicht monoton ist.

Haben wir eine beschränkte Folge $(x_n)_{n\in\mathbb{N}}$, dann existieren nach dem Supremums-Axiom sowohl

$$\inf\{x_n : n \geq 1\} \qquad \text{als auch} \qquad \sup\{x_n : n \geq 1\}.$$

Ist die Folge zusätzlich monoton, so zeigt der folgende Satz, dass einer der beiden Werte der Grenzwert von $(x_n)_{n\in\mathbb{N}}$ ist.

Theorem 4.1.16

Sei $(x_n)_{n\in\mathbb{N}}$ eine monoton wachsende und beschränkte Folge in $\mathbb{R}$. Dann konvergiert sie gegen $x := \sup\{x_n : n \geq 1\}$.

Ist die Folge $(y_n)_{n\in\mathbb{N}}$ monoton fallend und beschränkt, so ist die durch $x_n := -y_n$ definierte Folge monoton wachsend und beschränkt. Sie konvergiert nach Satz 4.1.16 gegen $\sup\{x_n : n \geq 1\}$. Damit folgt

$$\begin{aligned}\lim_{n\to\infty} y_n = -\lim_{n\to\infty} x_n = -\sup\{x_n : n \geq 1\} &= \inf\{-x_n : n \geq 1\} \\ &= \inf\{y_n : n \geq 1\}.\end{aligned}$$

Sei $\varepsilon > 0$. Da x die kleinste obere Schranke von $\{x_n : n \geq 1\}$ ist, kann $x - \varepsilon$ keine obere Schranke dieser Menge sein. Damit gibt es ein $n_\varepsilon \in \mathbb{N}$ mit $x_{n_\varepsilon} > x - \varepsilon$. Für $n \geq n_\varepsilon$ folgt dann

$$x - \varepsilon < x_{n_\varepsilon} \leq x_n \leq x < x + \varepsilon$$

und damit haben wir $|x - x_n| < \varepsilon$ gezeigt.

BEISPIEL

Die Folge $(x_n)_{n\in\mathbb{N}}$, die durch

$$x_n := \sum_{k=1}^{n} \frac{1}{k^2}$$

definiert ist, ist monoton wachsend, da alle Summanden nichtnegativ sind. Wegen

$$\frac{1}{k^2} \leq \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}, \qquad k \geq 2,$$

und der daraus folgenden Abschätzung

$$\begin{aligned} x_n = \sum_{k=1}^{n} \frac{1}{k^2} \leq 1 + \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) &\leq 1 + \sum_{k=2}^{n} \frac{1}{k-1} - \sum_{k=2}^{n} \frac{1}{k} \\ &\leq 1 + \sum_{k=1}^{n-1} \frac{1}{k} - \sum_{k=2}^{n} \frac{1}{k} \\ &= 2 - \frac{1}{n} \end{aligned}$$

erhalten wir auch $0 \leq x_n \leq 2$ für alle $n \geq 2$. Damit ist die Folge auch beschränkt und damit nach Satz 4.1.16 konvergent. Allerdings werden wir den Grenzwert dieser Folge erst viel später ausrechnen können.

BEISPIEL: EULERSCHE ZAHL

Die Folge $(x_n)_{n\in\mathbb{N}}$, die durch

$$x_n := \sum_{k=0}^{n} \frac{1}{k!}$$

definiert ist, ist monoton wachsend. Wegen

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

für alle $k \geq 2$ erhalten wir

$$x_n \leq 1+1+\sum_{k=2}^{n} \frac{1}{(k-1)k} \leq 2+\sum_{k=2}^{n} \frac{1}{k^2} \leq 3\,,$$

wobei wir letzten Schritt das vorherige Beispiel benutzt haben. Damit gilt $0 \leq x_n \leq 3$ für alle $n \geq 1$, d.h. die Folge ist beschränkt. Sie ist damit auch konvergent und wir bezeichnen ihren Grenzwert als die Eulersche Zahl e,

$$\mathrm{e} := \sum_{k=0}^{\infty} \frac{1}{k!} := \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!}\,. \tag{4.1.3}$$

Es gilt, wie wir gerade gezeigt haben, $2 \leq \mathrm{e} \leq 3$. Die Eulersche Zahl ist uns schon bei der ersten informellen Einführung der Exponentialfunktion begegnet.

Die Folge $(x_n)_{n\in\mathbb{N}}$, die durch

$$x_n := \left(1+\frac{1}{n}\right)^n \tag{4.1.4}$$

definiert ist, ist ebenfalls monoton wachsend und beschränkt. Die Monotonie folgt dabei aus der Bernoulli-Ungleichung mit $x := -\frac{1}{n^2} \geq -1$ und

$$\begin{aligned}\frac{x_n}{x_{n-1}} = \left(1+\frac{1}{n}\right)^n \cdot \left(1+\frac{1}{n-1}\right)^{-(n-1)} &= \left(1+\frac{1}{n}\right)^n \cdot \frac{n}{n-1} \cdot \left(\frac{n-1}{n}\right)^n \\ &= \frac{n}{n-1} \cdot \left(1-\frac{1}{n^2}\right)^n \\ &\geq \frac{n}{n-1} \cdot \left(1-\frac{1}{n}\right) \\ &= 1\,.\end{aligned}$$

Um zu zeigen, dass die Folge beschränkt ist, betrachten wir zunächst die Abschätzung

$$\begin{aligned}\binom{n}{k}\frac{1}{n^k} &= \frac{n(n-1)\cdots(n-k+1)}{k!n^k}\\ &= \frac{1}{k!}\cdot 1\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) \qquad (4.1.5)\\ &\leq \frac{1}{k!}\,.\end{aligned}$$

Damit folgt mit dem binomischen Lehrsatz 2.2.7 mit $a := \frac{1}{n}$ und $b := 1$

$$x_n = \left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}\frac{1}{n^k} \leq \sum_{k=0}^{n}\frac{1}{k!} < \mathrm{e}\,. \qquad (4.1.6)$$

Das folgende Lemma zeigt nun, dass die Folge sogar gegen e konvergiert.

Lemma 4.1.17

Es gilt

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \mathrm{e}\,.$$

Sei $(x_n)_{n\in\mathbb{N}}$ die durch (4.1.4) definierte Folge. Wir wissen schon $\lim_{n\to\infty} x_n \leq \mathrm{e}$ aus unseren Vorüberlegungen.

Für $1 \leq N \leq n$ gilt wegen (4.1.6) und (4.1.5):

$$\begin{aligned} x_n = \sum_{k=0}^{n} \binom{n}{k} \frac{1}{n^k} &\geq \sum_{k=0}^{N} \binom{n}{k} \frac{1}{n^k} \\ &= \sum_{k=0}^{N} \frac{1}{k!} \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) \\ &\geq \sum_{k=0}^{N} \frac{1}{k!} \left(1-\frac{N}{n}\right)^N . \end{aligned}$$

Damit erhalten wir

$$\lim_{n\to\infty} x_n \geq \lim_{n\to\infty} \sum_{k=0}^{N} \frac{1}{k!} \left(1-\frac{N}{n}\right)^N = \sum_{k=0}^{N} \frac{1}{k!}$$

für alle $N \geq 1$. Für $N \to \infty$ ergibt dies $\lim_{n\to\infty} x_n \geq \mathrm{e}$.

Ist $(x_n)_{n\in\mathbb{N}}$ eine reelle Folge, so ist die Folge $(y_n)_{n\in\mathbb{N}}$, die durch

$$y_n := \sup_{m\geq n} x_m := \sup\{x_m : m \geq n\}\,, \qquad n \geq 1 \tag{4.1.7}$$

definiert ist, monoton fallend. Ist $(x_n)_{n\in\mathbb{N}}$ nun beschränkt, so gibt es ein $B \geq 0$ mit $|x_n| \leq B$ für alle $n \geq 1$. Damit folgt

$$|y_n| \leq \sup_{m\geq n} |x_m| \leq \sup_{m\geq 1} |x_m| \leq B\,,$$

d.h. die Folge $(y_n)_{n\in\mathbb{N}}$ ist ebenfalls beschränkt. Nach Satz 4.1.16 konvergiert dann $(y_n)_{n\in\mathbb{N}}$ und wir definieren den Limes Superior als

$$\limsup_{n\to\infty} x_n := \lim_{n\to\infty} y_n = \lim_{n\to\infty} \sup_{m\geq n} x_m = \inf_{n\geq 1} \sup_{m\geq n} x_m\,.$$

Definieren wir nun die Folge $(z_n)_{n\in\mathbb{N}}$ durch

$$z_n := \inf_{m\geq n} x_m\,, \qquad\qquad n \geq 1\,, \tag{4.1.8}$$

so ist diese monoton wachsend und, falls $(x_n)_{n\in\mathbb{N}}$ beschränkt ist, ist auch $(z_n)_{n\in\mathbb{N}}$ beschränkt. Dies erlaubt dann die Definition des Limes Inferior

$$\liminf_{n\to\infty} x_n := \lim_{n\to\infty} z_n = \lim_{n\to\infty} \inf_{m\geq n} x_m = \sup_{n\geq 1} \inf_{m\geq n} x_m\,.$$

Die Definitionen (4.1.8) und (4.1.7) ergeben sofort

$$z_n \leq x_n \leq y_n\,, \qquad\qquad n \geq 1.$$

Satz 4.1.7 liefert dann $\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n$ und im Falle der Gleichheit ergibt das Vergleichskriterium aus Satz 4.1.9 sofort die eine Implikation des folgenden Lemmas:

Lemma 4.1.18

Sei $(x_n)_{n\in\mathbb{N}}$ eine beschränkte, reelle Folge. Dann konvergiert die Folge $(x_n)_{n\in\mathbb{N}}$ genau dann, wenn $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$. In diesem Fall gilt

$$\lim_{n\to\infty} x_n = \liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n\,.$$

Die Implikation "$\Leftarrow$" und die Gleichung folgt, wie schon erwähnt, aus Satz 4.1.9.

Für den Beweis von "$\Rightarrow$' nehmen wir $\liminf_{n\to\infty} x_n < \limsup_{n\to\infty} x_n$ an. Für

$$\varepsilon := (\limsup_{n\to\infty} x_n - \liminf_{n\to\infty} x_n)/4 > 0$$

gibt es dann ein $n_\varepsilon \geq 1$ mit

$$\inf_{k\geq n} x_n < \liminf_{n\to\infty} x_n + \varepsilon \tag{4.1.9}$$

und

$$\sup_{k\geq n} x_n > \limsup_{n\to\infty} x_n - \varepsilon$$

für alle $n \geq n_0$.

Damit existieren zu jedem $n \geq n_\varepsilon$ zwei Indices $m, k \geq n$ mit

$$x_k < \liminf_{n\to\infty} x_n + \varepsilon$$

und

$$x_m > \limsup_{n\to\infty} x_n - \varepsilon\,,$$

da z.B. $\liminf_{n\to\infty} x_n + \varepsilon$ wegen (4.1.9) keine untere Schranke der Menge $\{x_k : k \geq n\}$ ist. Dies impliziert

$$|x_m - x_k| \geq x_m - x_k > \limsup_{n\to\infty} x_n - \varepsilon - (\liminf_{n\to\infty} x_n + \varepsilon) = 2\varepsilon\,.$$

Somit ist $(x_n)_{n\in\mathbb{N}}$ keine Cauchyfolge und damit auch nicht konvergent.

UNEIGENTLICHE KONVERGENZ

Wir vereinbaren noch einige Sprechweisen zu Zahlenfolgen. Wir sagen, dass eine Folge $(x_n)_{n\in\mathbb{N}}$ **divergent** ist, falls sie nicht konvergent ist. Wir nennen eine reelle Zahlenfolge $(x_n)_{n\in\mathbb{N}}$ **bestimmt divergent gegen ∞**, falls

$$\forall R > 0 \exists n_R \geq 1 \forall n \geq n_R : x_n > R$$

gilt. In diesem Fall schreiben wir $x_n \to \infty$ oder auch $\lim_{n\to\infty} x_n = \infty$. Analog heißt die Folge **bestimmt divergent gegen $-\infty$**, falls

$$\forall R > 0 \exists n_R \geq 1 \forall n \geq n_R : x_n < -R\,.$$

In diesem Fall schreiben wir $x_n \to -\infty$ oder auch $\lim_{n\to\infty} x_n = -\infty$.

Wir hatten in Satz 4.1.6 gesehen, dass jede konvergente Folge eine Cauchyfolge ist und anhand von Beispielen haben wir ferner gesehen, dass die Umkehrung im Allgemeinen falsch ist. Der folgende Satz ist daher bemerkenswert.

Theorem 4.1.19

Jede reelle Cauchyfolge konvergiert in $\mathbb{R}$.

Sei $(x_n)_{n\in\mathbb{N}}$ eine reelle Cauchyfolge. Nach Satz 4.1.6 ist die Folge beschränkt, und nach Lemma 4.1.18 reicht es daher

$$\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$$

zu zeigen. Wie in (4.1.8) und (4.1.7) schreiben wir dazu

$$y_n := \sup_{m\geq n} x_m$$

$$z_n := \inf_{m\geq n} x_m \,.$$

Damit reicht es $\lim_{n\to\infty} z_n = \lim_{n\to\infty} y_n$, d.h.

$$\lim_{n\to\infty} (y_n - z_n) = 0 \tag{4.1.10}$$

zu zeigen.

Sei dazu $\varepsilon > 0$. Dann gibt es ein $n_\varepsilon \geq 1$ mit $|x_n - x_m| < \varepsilon/3$ für alle $n \geq n_\varepsilon$. Sei nun $n \geq n_\varepsilon$ fixiert. Da y_n die kleinste obere Schranke der Menge $\{x_m : m \geq n\}$ ist, gibt es ein $m \geq n$ mit

$$x_m > y_n - \varepsilon/3\,.$$

Da z_n die größte untere Schranke der Menge $\{x_m : m \geq n\}$ ist, gibt es ein $k \geq n$ mit

$$x_k < z_n + \varepsilon/3\,.$$

Damit folgt

$$0 \leq y_n - z_n \leq x_m + \frac{\varepsilon}{3} - x_k + \frac{\varepsilon}{3} \leq |x_m - x_k| + \frac{2\varepsilon}{3} < \varepsilon\,.$$

Dies ergibt $|y_n - z_n| < \varepsilon$ und damit haben wir (4.1.19) gezeigt.

VOLLSTÄNDIGKEIT

Ein metrischer Raum $(X, \mathbf{d})$ heißt vollständig, falls jede Cauchyfolge in X konvergiert.

Der Satz 4.1.19 zeigt, dass $\mathbb{R}$ vollständig ist. Analog zu Lemma 4.1.10 kann man mit Hilfe von Satz 4.1.19 dann auch zeigen, dass $\mathbb{R}^d$ bezüglich der Betragsmetrik vollständig ist, und das gleiche gilt auch für die Supremumsnorm und die 1-Norm. Im Fall $d = 2$ sehen wir zudem, dass $\mathbb{C}$ vollständig ist.

Teilmengen eines vollständigen metrischen Raums sind im Allgemeinen *nicht* vollständig. Wir hatte dies schon für die Teilmenge $(0, \infty) \subset \mathbb{R}$ im Anschluss an Satz 4.1.6 gesehen.

Section 4.2
Konvergenz von Reihen

Sei $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}$ eine Folge. Dann heißt die durch

$$s_n := \sum_{k=1}^{n} x_k\,, \qquad\qquad n \geq 1$$

definierte Folge $(s_n)_{n\in\mathbb{N}}$ **Folge der Partialsummen** oder auch **Reihe**.
Konvergiert die Folge $(s_n)_{n\in\mathbb{N}}$, so schreiben wir

$$\sum_{k=1}^{\infty} x_k := \lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{k=1}^{n} x_k$$

und sprechen vom dem **Wert der Reihe**, oder oder auch nur von der **Reihe**.

In einigen Fällen betrachten wir auch z.B. Folgen $(x_k)_{k\in\mathbb{N}_0}$. Die Definition von

$$\sum_{k=0}^{\infty} x_k$$

ist dann analog zu verstehen. Analoge Definitionen sind in allen normierten Räumen möglich.

Häufig kann der Grenzwert von Reihen leider nicht explizit berechnet werden. Die Konvergenz kann aber meistens nachgewiesen werden. Ein erstes Resultat in diese Richtung liefert der folgende Satz, dessen erster Teil als Cauchy-Kriterium bekannt ist.

Theorem 4.2.1

Seien $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}$ und $N \in \mathbb{N}$. Dann gelten die folgenden Aussagen:

i). Die Reihe $\sum_{k=1}^{\infty} x_k$ konvergiert genau dann, wenn gilt:

$$\forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \forall n, m \geq n_\varepsilon : \left| \sum_{k=n}^{m} x_k \right| < \varepsilon \,. \tag{4.2.1}$$

ii). Konvergiert die Reihe $\sum_{k=1}^{\infty} x_k$, so gilt $x_k \to 0$.

iii). Die Reihe $\sum_{k=1}^{\infty} x_k$ konvergiert genau dann, wenn der Reihenrest $\sum_{k=N}^{\infty} x_k$ konvergiert und in diesem Fall gilt

$$\sum_{k=1}^{\infty} x_k = \sum_{k=1}^{N-1} x_k + \sum_{k=N}^{\infty} x_k \,.$$

i). Für $m, n \in \mathbb{N}$ mit $m \geq n$ gilt:

$$|s_m - s_{n-1}| = \left|\sum_{k=1}^{m} x_k - \sum_{k=1}^{n-1} x_k\right| = \left|\sum_{k=n}^{m} x_k\right|$$

Damit ist die Folge $(s_n)_{n\in\mathbb{N}}$ eine Cauchyfolge genau dann, wenn (4.2.1) gilt. Die Implikation "$\Rightarrow$" folgt dann aus Satz 4.1.6 und die Implikation "$\Leftarrow$" aus Satz 4.1.19.

ii). Folgt aus *i)* durch Betrachtung von $n = m$.

iii). Für $n \geq N$ gilt

$$\sum_{k=1}^{n} x_k = \sum_{k=1}^{N-1} x_k + \sum_{k=N}^{n} x_k \,.$$

Für $n \to \infty$ folgt dann die Behauptung.

Für die Implikation "$\Leftarrow$" in *i)* wurde die Vollständigkeit von $\mathbb{R}$ benutzt. Diese Implikation gilt daher auch/nur in vollständigen normierten Räumen, während die anderen Implikation in allen normierten Räumen gelten. Insbesondere gilt der Satz ohne Einschränkung auch für Reihen in $\mathbb{C}$.

Der Teil *ii)* aus Satz 4.2.1 ist manchmal nützlich, um zu zeigen, dass eine Reihe *nicht* konvergiert.

Außerdem gelten die folgenden Rechenregeln für Reihen, die eine einfache Konsequenz aus Satz 4.1.7 sind.

Lemma 4.2.2

Seien $(x_k)_{k\in\mathbb{N}}$ und $(y_k)_{k\in\mathbb{N}}$ reelle Folgen, deren Reihen konvergieren. Dann gilt:

i). Die Reihe $\sum_{k=1}^{\infty}(x_k + y_k)$ konvergiert und es gilt

$$\sum_{k=1}^{\infty}(x_k + y_k) = \sum_{k=1}^{\infty} x_k + \sum_{k=1}^{\infty} y_k \,.$$

ii). Für alle $\alpha \in \mathbb{R}$ konvergiert die Reihe $\sum_{k=1}^{\infty}(\alpha x_k)$ und es gilt

$$\sum_{k=1}^{\infty}(\alpha x_k) = \alpha \sum_{k=1}^{\infty} x_k \,.$$

Viele der Konvergenzkriterien für Reihen sind eine direkte Konsequenz des Cauchy-Kriteriums für die Konvergenz von reellen (oder komplexen) Zahlenfolgen. Eines der konzeptionell einfachsten dieser Kriterien ist das Leibniz-Kriterium, das wir allerdings, wie unten zu sehen ist, nicht über das Cauchy-Kriterium beweisen.

Theorem 4.2.3

Sei $(x_k)_{k\in\mathbb{N}} \subset [0,\infty)$ *monoton fallend mit* $\lim_{k\to\infty} a_k = 0$. *Dann konvergiert die Reihe*

$$\sum_{k=1}^{\infty}(-1)^k x_k\,.$$

Auf die Monotonie der Folge $(x_k)_{k\in\mathbb{N}}$ kann nicht verzichtet werden. Dies kann man z.B. an der durch $x_{2k} = \frac{1}{k}$ und $x_{2k-1} := 0$ definierten Folge $(x_n)_{n\in\mathbb{N}}$ im Zusammenspiel mit der etwas später gezeigten Divergenz der harmonischen Reihe (4.2.2) leicht sehen.

BEWEIS

Wir betrachten die Folge $(s_n)_{n\in\mathbb{N}}$ der Partialsummen

$$s_n := \sum_{k=1}^{n} (-1)^k x_k \,.$$

Für $k \in \mathbb{N}$ gilt dann $x_{2k+2} \leq x_{2k+1}$ und damit

$$s_{2(k+1)} = s_{2k} - x_{2k+1} + x_{2k+2} \leq s_{2k} \,.$$

Damit ist die Folge $(s_{2k})_{k\in\mathbb{N}}$ monoton fallend. Analog zeigt man mit $s_{2k+1} = s_{2k-1} + x_{2k} - x_{2k+1} \geq s_{2k-1}$, dass die Folge $(s_{2k-1})_{k\in\mathbb{N}}$ monoton wächst.

Mit dieser Monotonie von $(s_{2k-1})_{k\in\mathbb{N}}$ und $x_k \geq 0$ folgt nun

$$s_{2k} = s_{2k-1} + x_k \geq s_1 + x_k \geq s_1 = x_1$$

Damit ist die Folge $(s_{2k})_{k\in\mathbb{N}}$ nach unten beschränkt, und somit konvergiert sie nach Satz 4.1.16 gegen ein $s^+ \in \mathbb{R}$. Analog zeigt man mit der Monotonie von $(s_{2k})_{k\in\mathbb{N}}$ und $x_k \geq 0$:

$$s_{2k-1} = s_{2k} - x_k \leq s_{2k} \leq s_2 = x_1 - x_2 \,.$$

Damit ist die Folge $(s_{2k-1})_{k\in\mathbb{N}}$ nach oben beschränkt und nach Satz 4.1.16 gegen ein $s^- \in \mathbb{R}$.

Ferner gilt mit Satz 4.1.7

$$s^+ - s^- = \lim_{k\to\infty} s_{2k} - \lim_{k\to\infty} s_{2k-1} = \lim_{k\to\infty} (s_{2k} - s_{2k-1}) = \lim_{k\to\infty} x_{2k} = 0\,.$$

Wir schreiben daher $s := s^+ = s^-$ für den gemeinsamen Grenzwert.

Sei nun $\varepsilon > 0$. Dann gibt es ein $k_\varepsilon \geq 1$ mit

$$|s - s_{2k}| < \varepsilon \qquad \text{und} \qquad |s - s_{2k-1}| < \varepsilon$$

für alle $k \geq k_\varepsilon$. Für $n_\varepsilon := 2k_\varepsilon$ und $n \geq n_\varepsilon$ erhalten wir dann $|s - s_n| < \varepsilon$ und damit konvergiert $(s_n)_{n\in\mathbb{N}}$ gegen s.

Es gilt die **geometrische Summenformel**

$$\sum_{k=0}^{n} z^k = \frac{z^{n+1} - 1}{z - 1},$$

wie man leicht durch Ausmultiplizieren

$$(z-1)\sum_{k=0}^{n} z^k = \sum_{k=0}^{n} z^{k+1} - \sum_{k=0}^{n} z^k = \sum_{k=1}^{n+1} z^k - \sum_{k=0}^{n} z^k$$
$$= z^{n+1} - 1$$

zeigt. Wir bilden den Grenzwert für $n \to \infty$ und erhalten:

Lemma 4.2.4

*Die **geometrische Reihe** $\sum_{k=0}^{\infty} z^k$ ist konvergent für alle $|z| < 1$ und erfüllt*

$$\sum_{k=0}^{\infty} z^k = \frac{1}{1 - z}.$$

Ebenso ist die geometrische Reihe für alle $|z| \geq 1$ divergent.

Für die **harmonische Reihe** gilt

$$\sum_{k=1}^{\infty} \frac{1}{k} = \infty\,, \tag{4.2.2}$$

d.h. sie ist bestimmt divergent gegen ∞. Dazu schätzen wir bestimmte Partialsummen nach unten ab:

$$\sum_{k=0}^{2^m} \frac{1}{k} = 1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>\frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>\frac{1}{2}} + \cdots + \underbrace{\frac{1}{2^{m-1}+1} + \cdots + \frac{1}{2^m}}_{>\frac{1}{2}}$$

$$> 1 + \frac{m}{2} \longrightarrow \infty$$

für $m \to \infty$. Da die Folge aller Partialsummen monoton wächst, folgt die gewünschte bestimmte Divergenz. Insbesondere impliziert die Konvergenz $x_k \to 0$ nicht die Konvergenz der zugehörigen Reihe!

Die **alternierende harmonische Reihe**

$$\sum_{k=1}^{\infty} \frac{(-1)^k}{k} \tag{4.2.3}$$

konvergiert nach dem Leibniz-Kriterium 4.2.3.

Die beiden Reihen

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \qquad \text{und} \qquad \sum_{k=0}^{\infty} \frac{1}{k!}$$

konvergieren. Dies hatten wir schon im Abschnitt 242 gesehen.

Die folgende Definition führt eine weitere Konvergenz von Reihen ein, die sich als wichtig herausstellen wird.

Definition 4.2.5

Sei $(x_k)_{k\in\mathbb{N}}$ eine reelle Folge. Dann sagen wir, dass die Reihe $\sum_{k=1}^{\infty} x_k$ **absolut** konvergiert, falls die Reihe

$$\sum_{k=1}^{\infty} |x_k|$$

der Beträge konvergiert.

Konvergiert die Reihe $\sum_{k=1}^{\infty} x_k$ absolut, so gibt es nach dem Cauchy-Kriterium, siehe Satz 4.2.1, zu jedem $\varepsilon > 0$ ein $n_\varepsilon \geq 1$, so dass für $m \geq n \geq n_\varepsilon$ gilt

$$\left|\sum_{k=n}^{m} x_k\right| \leq \sum_{k=n}^{m} |x_k| < \varepsilon\,.$$

Erneutes Anwenden von Satz 4.2.1 zeigt dann, dass die Reihe $\sum_{k=1}^{\infty} x_k$ auch konvergiert.

Mit anderen Worten: **Jede absolut konvergente Reihe ist auch konvergent.** Die Umkehrung gilt nicht, wie man anhand der alternierenden harmonischen Reihe sehen kann.

Der folgende Satz ist als Majoranten-Kriterium bekannt.

Theorem 4.2.6

Seien $(x_k)_{k\in\mathbb{N}}$ und $(a_k)_{k\in\mathbb{N}}$ zwei reelle Folgen mit $|x_k| \leq a_k$ für alle $k \in \mathbb{N}$. Falls die Reihe

$$\sum_{k=1}^{\infty} a_k$$

konvergiert, konvergiert die Reihe

$$\sum_{k=1}^{\infty} x_k$$

absolut und es gilt

$$\left|\sum_{k=1}^{\infty} x_k\right| \leq \sum_{k=1}^{\infty} |x_k| \leq \sum_{k=1}^{\infty} a_k\,.$$

Der Satz gilt analog für Reihen $\sum_{k=1}^{\infty} x_k$ in $\mathbb{C}$. Der Beweis ist analog zu führen.

Wir setzen $a := \sum_{k=1}^{\infty} a_k$. Dann existiert zu jedem $\varepsilon > 0$ ein $n_\varepsilon \in \mathbb{N}$ mit

$$\left| \sum_{k=1}^{n} a_k - a \right| < \varepsilon,$$

für alle $n \geq n_\varepsilon$. Mit Teil *iii)* aus Satz 4.2.1 folgt für solche n:

$$0 \leq \sum_{k=n+1}^{\infty} a_k < \varepsilon.$$

Damit folgt aber für $m, n > n_\varepsilon$ die Abschätzung

$$\sum_{k=n+1}^{m} |x_k| \leq \sum_{k=n+1}^{m} a_k \leq \sum_{k=n+1}^{\infty} a_k < \varepsilon\,,$$

da die Folge der Partialsummen der Reihe $\sum_{k=n+1}^{\infty} a_k$ monoton wachsend ist. Das Cauchy-Kriterium aus Satz 4.2.1 liefert dann die absolute Konvergenz der Reihe $\sum_{k=1}^{\infty} x_k$.

Weiterhin gilt

$$\left| \sum_{k=1}^{n} x_k \right| \leq \sum_{k=1}^{n} |x_k| \leq \sum_{k=1}^{n} a_n \leq \sum_{k=1}^{\infty} a_k$$

und für $n \to \infty$ folgen dann die beiden Ungleichungen

BEISPIEL

Um ein Beispiel zu betrachten, fixieren wir ein $\alpha \geq 2$. Dann gilt $k^{-\alpha} \leq k^{-2}$ für jedes $k \geq 1$ und damit konvergiert die Reihe

$$\sum_{k=1}^{\infty} \frac{1}{k^{\alpha}}$$

absolut nach dem Majoranten-Kriterium und der schon im Abschnitt 242 bewiesenen Konvergenz von $\sum_{k=1}^{\infty} \frac{1}{k^2}$.

Sei $(x_k)_{k\in\mathbb{N}}$ eine reelle Folge. Dann heißt die zugehörige Reihe **unbedingt konvergent**, falls für jede Bijektion $\pi : \mathbb{N} \to \mathbb{N}$ die **umgeordnete Reihe**

$$\sum_{k=1}^{\infty} x_{\pi(k)}$$

konvergiert. Man kann zeigen, dass in $\mathbb{R}$ eine Reihe genau dann absolut konvergent ist, falls sie unbedingt konvergent ist. Der folgende Satz liefert insbesondere die eine Richtung dieser Äquivalenz:

Theorem 4.2.7

Sei $(x_k)_{k\in\mathbb{N}}$ eine reelle Folge, deren Reihe absolut konvergiert. Dann konvergiert die Reihe auch unbedingt und für jede Bijektion $\pi : \mathbb{N} \to \mathbb{N}$ gilt

$$\sum_{k=1}^{\infty} x_k = \sum_{k=1}^{\infty} x_{\pi(k)} \,. \tag{4.2.4}$$

Sei $\pi : \mathbb{N} \to \mathbb{N}$ eine Bijektion. Für $n \in \mathbb{N}$ setzen wir $M := \max\{\pi(1), \ldots, \pi(n)\}$. Dann gilt

$$\sum_{k=1}^{n} |x_{\pi(k)}| \leq \sum_{k=1}^{M} |x_k| \leq \sum_{k=1}^{\infty} |x_k|$$

und damit folgt auch $\sum_{k=1}^{\infty} |x_{\pi(k)}| \leq \sum_{k=1}^{\infty} |x_k| < \infty$. Damit konvergiert die umgeordnete Reihe absolut und ihre Konvergenz folgt.

Anwenden der letzten Ungleichung auf die Inverse $\pi^{-1} : \mathbb{N} \to \mathbb{N}$ liefert

$$\sum_{k=1}^{\infty} |x_k| = \sum_{k=1}^{\infty} |x_{\pi^{-1} \circ \pi(k)}| \leq \sum_{k=1}^{\infty} |x_{\pi(k)}|\,,$$

so dass wir insgesamt

$$\sum_{k=1}^{\infty} |x_k| = \sum_{k=1}^{\infty} |x_{\pi(k)}| \tag{4.2.5}$$

gezeigt haben.

Um (4.2.4) zu zeigen, definieren wir jetzt

$$x_k^+ := \max\{x_k, 0\}\,,$$
$$x_k^- := -\min\{x_k, 0\}\,.$$

Dies ergibt $x_k^\pm \geq 0$, und die Zerlegungen $x_k = x_k^+ - x_k^-$ und $|x_k| = x_k^+ + x_k^-$.

Nach dem Majoranten-Kriterium mit $a_k := |x_k|$ konvergieren dann die Reihen $\sum_{k=1}^\infty x_k^+$ und $\sum_{k=1}^\infty x_k^-$ absolut und wegen $|x_k^\pm| = x_k^\pm$ liefert doppeltes Anwenden unsere Vorüberlegung (4.2.5):

$$\sum_{k=1}^\infty x_{\pi(k)}^+ - \sum_{k=1}^\infty x_{\pi(k)}^- = \sum_{k=1}^\infty x_k^+ - \sum_{k=1}^\infty x_k^-\,.$$

Ferner gilt

$$\sum_{k=1}^\infty x_{\pi(k)} = \sum_{k=1}^\infty (x_{\pi(k)}^+ - x_{\pi(k)}^-) = \sum_{k=1}^\infty x_{\pi(k)}^+ - \sum_{k=1}^\infty x_{\pi(k)}^-$$

und eine analoge Rechnung liefert $\sum_{k=1}^\infty x_k = \sum_{k=1}^\infty x_k^+ - \sum_{k=1}^\infty x_k^-$.

Falls eine Reihe $\sum_{k=1}^{\infty} x_k$ allerdings *nicht* unbedingt konvergiert, zeigt der Riemann'sche Umordnungssatz, dass für jedes $x \in \mathbb{R} \cup \{-\infty, \infty\}$ eine Bijektion $\pi : \mathbb{N} \to \mathbb{N}$ existiert, so dass die umgeordnete Reihe gegen x konvergiert, d.h.

$$\sum_{k=1}^{\infty} x_{\pi(k)} = x\,.$$

Mit anderen Worten: Das Umordnen von Reihen nur für absolut konvergente Reihen ungefährlich.

Absolute Konvergenz kann man auch in normierten Räumen definieren, indem man Beträge durch Normen ersetzt. In vollständigen, normierten Räumen ist dann jede absolut konvergente Reihe auch unbedingt konvergent, die Umkehrung gilt allerdings nur noch im Falle endlicher Dimension des Raums, wie aus dem Satz von Dvoretzky-Rogers folgt.

Typische Vergleichsreihen für das Majoranten-Kriterium sind geometrische Reihen, siehe Lemma 4.2.4. Mit diesen erhalten wir zwei oft nutzbare Konvergenzkriterien. Das erste ist als Quotienten-Kriterium bekannt.

Theorem 4.2.8

Sei $(x_k)_{k\in\mathbb{N}}$ eine reelle Folge mit $x_k \neq 0$ für alle $k \geq 1$. Gilt dann

$$\limsup_{k\to\infty} \frac{|x_{k+1}|}{|x_k|} < 1\,,$$

so konvergiert die Reihe $\sum_{k=1}^{\infty} x_k$ absolut. Gilt andererseits

$$\liminf_{k\to\infty} \frac{|x_{k+1}|}{|x_k|} > 1\,,$$

so divergiert die Reihe $\sum_{k=1}^{\infty} x_k$.

Der Satz gilt analog für Reihen in $\mathbb{C}$. Der Beweis ist ebenfalls analog zu führen.

BEWEIS 1/2

Wir zeigen zunächst die erste Aussage. Sei dazu $\tilde{q} := \limsup_{k\to\infty} \frac{|x_{k+1}|}{|x_k|}$. Wir schreiben $q := \frac{\tilde{q}+1}{2}$ und $\varepsilon := q - \tilde{q}$. Dies ergibt $q \in (\tilde{q}, 1)$ und $\varepsilon > 0$. Damit gibt es ein $n_\varepsilon \geq 1$ mit

$$\sup_{k\geq n} \frac{|x_{k+1}|}{|x_k|} \leq \tilde{q} + \varepsilon = q < 1$$

für alle $n \geq n_\varepsilon$. Mit Induktion ergibt dies

$$|x_{n_\varepsilon+k}| \leq |x_{n_\varepsilon}| \cdot q^k =: a_k$$

für alle $k \geq 1$. Da die Reihe $\sum_{k=1}^{\infty} a_k$ nach Lemma 4.2.4 konvergiert, folgt nach Satz 4.2.6 die absolute Konvergenz der Reihe $\sum_{k=n_\varepsilon}^{\infty} x_k$. Dies impliziert die absolute Konvergenz der Reihe $\sum_{k=1}^{\infty} x_k$, siehe Satz 4.2.1.

BEWEIS 2/2

Für die zweite Aussage nutzen wir entsprechend, dass $q > 1$ und $n_\varepsilon \geq 1$ existieren mit

$$\inf_{k \geq n} \frac{|x_{k+1}|}{|x_k|} \geq q$$

für alle $n \geq n_\varepsilon$. Mit Induktion erhalten wir dann

$$|x_{n_\varepsilon + k}| \geq |x_{n_\varepsilon}| \cdot q^k$$

für alle $k \geq 1$. Wegen $x_{n_\varepsilon} \neq 0$ ist daher $(x_n)_{n \in \mathbb{N}}$ keine Nullfolge und Satz 4.2.1 ergibt die Divergenz.

BEISPIEL: EXPONENTIALREIHE

Die Reihe

$$\sum_{k=0}^{\infty} \frac{z^k}{k!} \tag{4.2.6}$$

konvergiert absolut für alle $z \in \mathbb{C}$. Für $z = 0$ ist dabei nichts zu zeigen, und für alle anderen $z \in \mathbb{C}$ nutzen wir das Quotienten-Kriterium: Es gilt

$$\frac{|z|^{k+1}}{(k+1)!} \cdot \frac{k!}{|z|^k} = \frac{|z|}{k+1} \to 0$$

für $k \to \infty$ und damit gilt dies auch für den Limes Superior, siehe Lemma 4.1.18. Die absolute Konvergenz der Reihe folgt dann aus dem Quotienten-Kriterium.

Das zweite Kriterium, das im folgenden Satz vorgestellt wird, ist als Wurzel-Kriterium bekannt. Auch diese Kriterium ist nur für $\mathbb{R}$ formuliert, es gilt aber ohne Einschränkung auch in $\mathbb{C}$.

Theorem 4.2.9

Sei $(x_k)_{k\in\mathbb{N}}$ eine reelle Folge. Gilt dann

$$\limsup_{k\to\infty} \sqrt[k]{|x_k|} < 1\,,$$

so konvergiert die Reihe $\sum_{k=1}^{\infty} x_k$ absolut. Gilt andererseits

$$\liminf_{k\to\infty} \sqrt[k]{|x_k|} > 1$$

so divergiert die Reihe $\sum_{k=1}^{\infty} x_k$.

Wir zeigen wiederum zuerst die erste Aussage. Nach Voraussetzung existiert ein $n_\varepsilon \geq 1$ und ein $q < 1$ mit

$$\sqrt[k]{|x_k|} \leq q$$

für alle $k \geq n_\varepsilon$. Damit folgt aber direkt $|x_k| \leq q^k$ für alle $k \geq n_\varepsilon$. Aus der Konvergenz der geometrischen Reihe folgt dann die absolute Konvergenz von $\sum_{k=n_\varepsilon}^{\infty} x_k$ mit Hilfe des Majoranten-Kriteriums.

Für die zweite Aussage nutzen wir entsprechend, dass $n_\varepsilon \geq 1$ und $q > 1$ existieren, so dass $\sqrt[k]{|x_k|} \geq q$ für alle $k \geq n_\varepsilon$ gilt. Damit folgt $|x_k| \geq q^k$ für alle $k \geq n_\varepsilon$ und die Reihenglieder bilden keine Nullfolge. Also folgt die Divergenz.

BEISPIEL

Für $z \in \mathbb{C}$ konvergiert die Reihe

$$\sum_{k=1}^{\infty} kz^k$$

genau dann, wenn $|z| < 1$. Dazu bemerken wir zunächst, dass Lemma 4.1.11

$$\sqrt[k]{k|z|^k} = |z|\sqrt[k]{k} \to |z|, \qquad k \to \infty$$

liefert. Das Wurzel-Kriterium ergibt Konvergenz für $|z| < 1$ und Divergenz für $|z| > 1$. Für $|z| = 1$ liefert das Kriterium keine Aussage, jedoch sieht man den Reihengliedern direkt an, dass diese auch hier keine Nullfolge bilden.

WIE GROSS IST R?

Wir hatten in Kapitel 2 die Größe von endlichen Mengen beschrieben. Insbesondere haben wir dort definiert, dass eine Menge A unendlich ist, falls es eine injektive Abbildung $\mathbb{N} \to A$ gibt.

Im folgenden heißt eine unendliche Menge A **abzählbar**, falls es eine Bijektion $\mathbb{N} \to A$ gibt, ansonsten heißt sie **überabzählbar**. Schließlich heißt eine Menge **höchstens abzählbar**, falls A entweder endlich, oder abzählbar ist.

Offensichtlich ist $\mathbb{N}$ abzählbar, und es ist auch $\mathbb{Z}$ abzählbar, denn wir können die z.B. Bijektion

$$n \mapsto \begin{cases} 0 & \text{falls } n = 1 \\ k & \text{falls } n = 2k \text{ für } k \in \mathbb{N} \\ -k & \text{falls } n = 2k+1 \text{ für } k \in \mathbb{N} \end{cases}$$

betrachten.

Man kann zeigen, dass jede Teilmenge einer höchstens abzählbaren Menge wieder höchstens abzählbar ist. Zusammen mit dem folgenden Satz zeigt dies, dass $\mathbb{Q}$ abzählbar ist.

Theorem 4.2.10

Seien A und B abzählbare Mengen. Dann ist $A \times B$ abzählbar.

Beweis.

In Abbildung 13 ist eine Bijektion $\mathbb{N} \to A \times B$ skizziert. Dabei werden die Bijektionen $\mathbb{N} \to A$ mit $i \mapsto a_i$ und $\mathbb{N} \to B$ mit $j \mapsto b_j$ bezeichnet. □

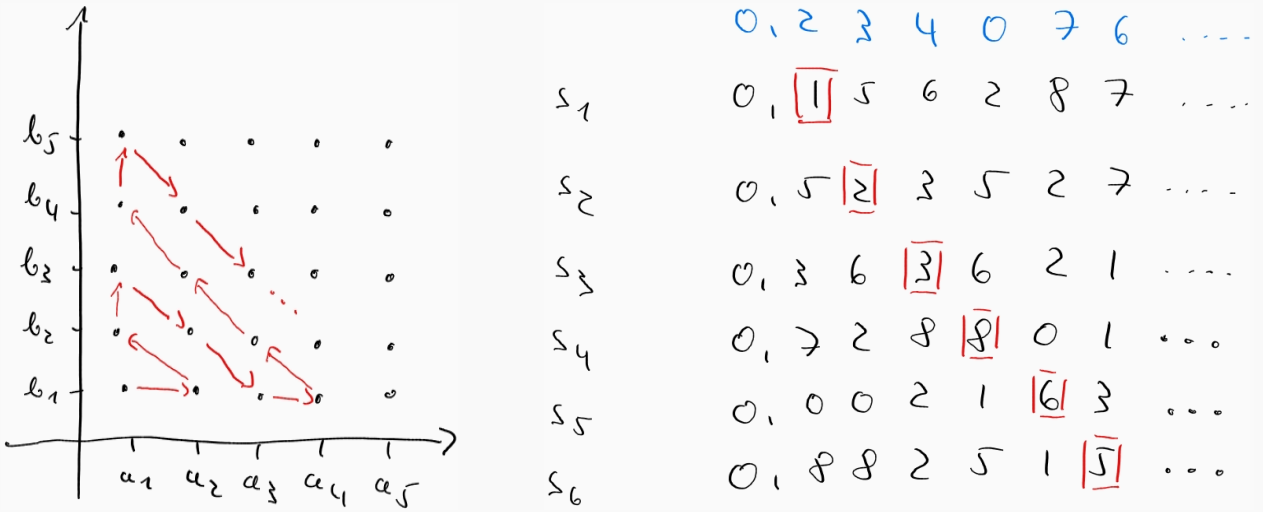


Abbildung: Links: Konstruktion einer Bijektion $\mathbb{N} \to A \times B$. Die 1 wird auf (a_1, b_1) abgebildet, danach werden die Tupel (a_i, b_j) gemäß der roten Pfeilrichtungen abgezählt, d.h. wir betrachten die Abbildung $\mathbb{N} \to A \times B$ mit $2 \mapsto (a_2, b_1)$, $3 \mapsto (a_1, b_2)$, $4 \mapsto (a_1, b_3)$ usw.. **Rechts**: Beweisidee von Satz 4.2.11. Die Zahlen $s_1, s_2, \ldots$ werden in ihrer Dezimaldarstellung untereinander aufgeschrieben. Dann wird die n-Ziffer z_n der n-ten Zahl s_i verändert. Beispielsweise wird die Ziffer um 1 erhöht, falls sie kleiner als 8 ist und ansonsten auf 0 gesetzt. Dies ergibt die blaue Zahl in der obersten Reihe. Diese kann nicht in der Menge $\{s_n : n \geq 1\}$ enthalten sein, weil sie sich nach Konstruktion von jeder dieser Zahlen an mindestens einer Nachkomma-Stelle unterscheidet.

R IST NICHT ABZÄHLBAR

Die rationalen Zahlen hatten wir mit Hilfe des Supremums-Axioms zu den reellen Zahlen aufgefüllt, wobei es unser Ziel war, die "Lücken" zu schließen. Waren es wenige Lücken, d.h. sind die reellen Zahlen auch abzählbar?

Um diese Frage zu beantworten, betrachten wir Reihen der Form

$$\sum_{k=1}^{\infty} z_k \cdot 10^{-k}\,, \tag{4.2.7}$$

wobei $z_k \in \{0, 1, \ldots, 9\}$. Setzen wir $a_k := 9 \cdot 10^{-k}$, so erhalten wir $z_k \cdot 10^{-k} \leq a_k$ für alle $k \geq 1$ und das Majoranten-Kriterium ergibt

$$0 \leq \sum_{k=1}^{\infty} z_k \cdot 10^{-k} \leq 9 \sum_{k=1}^{\infty} \left(\frac{1}{10}\right)^k = 9\left(\frac{1}{1-1/10} - 1\right) = 1\,.$$

Mit anderen Worten hat jede dieser Reihen einen Wert zwischen 0 und 1, und man kann sogar zeigen, dass sich jede reelle Zahl in $[0, 1]$ auf diese Weise darstellen lässt.

Theorem 4.2.11

Das Intervall $[0,1]$ *und* $\mathbb{R}$ *sind nicht abzählbar.*

Beweis.

Wir betrachten die Menge M aller Zahlen der Form (4.2.7), wobei wir nur die "Ziffern" $z_k \in \{0,1,\ldots,8\}$ zulassen. Dies sichert die Eindeutigkeit der Darstellung jeder Zahl in M, den Beweis hiervon lassen wir aber aus. Wenn $\mathbb{R}$ abzählbar wäre, wäre $[0,1]$ abzählbar und damit auch M. Wir nehmen dann eine beliebige Abbildung $\mathbb{N} \to M$ mit $n \mapsto s_n$. In Abbildung 13 ist dann illustriert, warum diese Abbildung nicht surjektiv ist. □

Chapter 5: Stetige Funktionen

Section 5.1
Stetigkeit

Funktionen, bei denen ähnliche Eingabewerte zu ähnlichen Ausgabewerten führen, werden stetig genannt. Die folgende Definition führt diesen intuitiven Ansatz mathematische präzise ein:

Definition 5.1.1

Seien $(X, \mathbf{d}_1)$ und $(Y, \mathbf{d}_2)$ zwei metrische Räume, $f : X \to Y$ eine Abbildung und $x_0 \in X$. Dann heißt f **stetig in** x_0, falls gilt:

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in X : \mathbf{d}_1(x, x_0) < \delta \implies \mathbf{d}_2\big(f(x), f(x_0)\big) < \varepsilon \,.$$

Ist f in allen Punkten x_0 stetig, so heißt f **stetig**.

BEISPIELE

Ist $X \subset Y$ und stimmen die beiden Metriken $\mathbf{d}_1$ und $\mathbf{d}_2$ auf X überein, so ist die Inklusionsabbildung $\mathbf{id}_{X,Y} : X \to Y$ stetig, denn wir können zu jedem $\varepsilon > 0$ einfach $\delta := \varepsilon$ wählen.

Ist $f : X \to Y$ eine konstante Funktion, d.h. es gibt ein $y_0 \in Y$ mit $f(x) = y_0$ für alle $x \in X$, so ist f stetig, denn wir können zu jedem $\varepsilon > 0$ ein beliebiges, von ε unabhängiges $\delta > 0$ wählen.

Für $A \subset \mathbb{R}$ ist $\mathbf{1}_A$ i.A. nicht stetig. Für $A := [0, \infty)$ ist z.B. $x := 0$ eine Unstetigkeitsstelle. Für $A := \mathbb{Q}$ ist die Funktion $\mathbf{1}_A$ in *keinem* Punkt stetig!

Die Stetigkeit kann auch mit Hilfe von Folgen ausgedrückt werden. Dies ist der Inhalt des folgendes Satzes:

Theorem 5.1.2

Seien $(X, \mathbf{d}_1)$ und $(Y, \mathbf{d}_2)$ zwei metrische Räume, $f : X \to Y$ eine Abbildung und $x_0 \in X$. Dann sind äquivalent:

i). f ist stetig in x_0.

*ii). f ist **folgenstetig in** x_0, d.h. für alle Folgen $(x_n)_{n\in\mathbb{N}} \subset X$ mit $x_n \to x_0$ gilt $f(x_n) \to f(x_0)$.*

Ist f stetig, so lässt sich die Aussage des Satzes 5.1.2 informell auf die Formel

$$f(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} f(x_n)$$

für alle konvergenten Folgen $(x_n)_{n\in\mathbb{N}} \subset X$ bringen. Mit anderen Worten: **Die Stetigkeit erlaubt es, Grenzwert-Bildung und Funktions-Anwendung zu vertauschen**.

Entsprechend eingeschränktere Aussagen gelten natürlich auch für die Stetigkeit in einem Punkt x_0.

i) $\Rightarrow$ ii). Sei $\varepsilon > 0$ und $(x_n)_{n\in\mathbb{N}} \subset X$ eine Folge mit $x_n \to x_0$. Da f stetig in x_0 ist, gibt es dann ein $\delta > 0$ mit

$$\mathbf{d}_1(x, x_0) < \delta \implies \mathbf{d}_2\big(f(x), f(x_0)\big) < \varepsilon$$

für alle $x \in X$. Wegen $x_n \to x_0$ gibt es dann ein $n^* \geq 1$ mit $\mathbf{d}_1(x_n, x_0) < \delta$ für alle $n \geq n^*$. Damit haben wir also $\mathbf{d}_2(f(x_n), f(x_0)) < \varepsilon$ für alle $n \geq n^*$, d.h. wir haben $f(x_n) \to f(x_0)$ gezeigt.

ii) $\Rightarrow$ i). Hier zeigen wir die äquivalente Aussage *$\neg$i) $\Rightarrow$ $\neg$ii)*.
Dementsprechend ist f nicht in x_0 stetig und damit ist die folgende Aussage wahr:

$$\exists \varepsilon > 0 \forall \delta > 0 \exists x \in X : \mathbf{d}_1(x, x_0) < \delta \wedge \mathbf{d}_2\big(f(x), f(x_0)\big) \geq \varepsilon\,.$$

Wir fixieren dieses ε. Für $n \in \mathbb{N}$ gibt es dann zu $\delta_n := 1/n$ ein $x_n \in X$ mit $\mathbf{d}_1(x_n, x_0) < \delta_n$ und $\mathbf{d}_2\big(f(x_n), f(x_0)\big) \geq \varepsilon$.

Wegen $\delta_n \to 0$ sichert unsere Konstruktion $\mathbf{d}_1(x_n, x_0) \to 0$ nach dem Vergleichskriterium, siehe Satz 4.1.9, d.h. $x_n \to x$. Andererseits gilt aber auch $\mathbf{d}_2\big(f(x_n), f(x_0)\big) \geq \varepsilon$ für alle $n \geq 1$, und damit kann $(f(x_n))_{n\in\mathbb{N}}$ nicht gegen $f(x_0)$ konvergieren. Damit ist f nicht folgenstetig in x_0.

Die Kombination der Sätze 5.1.2 und 4.1.7 liefert sofort das folgende Korollar:

Korollar 5.1.3

Sei (X, d) *ein metrischer Raum,* $f, g : X \to \mathbb{R}$ *Funktionen, die in* $x_0 \in X$ *stetig sind und* $\alpha, \beta \in \mathbb{R}$*. Dann sind auch die Funktionen* $\alpha f + \beta g$ *und* $f \cdot g$ *in* x_0 *stetig.*

Das obige Korollar zeigt insbesondere, dass die Menge

$$\mathcal{C}(X, \mathsf{d}) := \{f : X \to \mathbb{R} \mid f \text{ stetig}\}$$

ein Vektorraum ist. Da wir ferner schon wissen, dass die Abbildungen $x \mapsto a$ und $x \mapsto x$ als Abbildungen $\mathbb{R} \to \mathbb{R}$ stetig sind, sind alle reellen Polynome ebenfalls stetig. Die gleiche Aussage gilt für komplexe Polynome.

Die Kombination der Sätze 5.1.2 und 4.1.7 liefert zudem das folgende Korollar:

Korollar 5.1.4

Seien $(X_1, \mathbf{d}_1)$, $(X_2, \mathbf{d}_2)$ *und* $(X_3, \mathbf{d}_3)$ *metrische Räume, sowie* $f : X_1 \to X_2$ *und* $g : X_2 \to X_3$. *Ist dann* f *in* x_0 *stetig und* g *in* $y_0 := f(x_0)$ *stetig, so ist* $g \circ f : X_1 \to X_3$ *in* x_0 *stetig.*

Laut Definition ist eine Funktion $X \to Y$ stetig, falls die folgende Aussage wahr ist:

$$\forall x \in X \forall \varepsilon > 0 \exists \delta > 0 \forall x' \in X : \mathrm{d}_1(x, x') < \delta \implies \mathrm{d}_2\big(f(x), f(x')\big) < \varepsilon\,.$$

Damit hängt δ potentiell sowohl von x als auch von ε ab. Die folgenden Definitionen heben die Abhängigkeit von x auf und spezifizieren die Abhängigkeit von ε genauer.

Definition 5.1.5

Seien (X, d_1) und (Y, d_2) zwei metrische Räume und $f : X \to Y$ eine Abbildung und $\alpha \in (0, 1]$. Dann heißt f:

i). gleichmäßig stetig, falls gilt:

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x, x' \in X : \mathrm{d}_1(x, x') < \delta \implies \mathrm{d}_2\big(f(x), f(x')\big) < \varepsilon\,.$$

ii). α-Hölder-stetig, falls es eine Konstante $c \geq 0$ gibt, so dass für alle $x, x' \in X$ gilt:

$$\mathrm{d}_2\big(f(x), f(x')\big) \leq c\, \mathrm{d}_1^{\alpha}(x, x')\,.$$

iii). Lipschitz-stetig, falls f 1-Hölder-stetig ist.

Hölder- und Lipschitz-Stetigkeit

Ist eine Funktion α-Hölder-stetig, so ist sie auch gleichmäßig stetig, denn definieren wir für $\varepsilon > 0$ die Größe $\delta_\varepsilon > 0$ durch $c\delta_\varepsilon^\alpha := \varepsilon$, so gilt für $x, x' \in X$ mit $\mathbf{d}_1(x, x') < \delta_\varepsilon$ die Abschätzung

$$\mathbf{d}_2\big(f(x), f(x')\big) \leq c\, \mathbf{d}_1^\alpha(x, x') < c\delta_\varepsilon^\alpha = \varepsilon$$

Es gibt aber gleichmäßig stetige Funktionen, die nicht Hölder-stetig sind. Als Beispiel kann die Funktion $f : [0, 1/2] \to \mathbb{R}$ mit $f(0) := 0$ und $f(x) := -1/\ln(x)$ für $x \in (0, 1/2]$ dienen. Einen Beweis überspringen wir.

Ist eine Funktion gleichmäßig stetig, so ist sie offensichtlich auch stetig. Die Umkehrung gilt im allgemeinen nicht, wie z.B. $f : \mathbb{R} \to \mathbb{R}$ mit $f(x) := x^2$ zeigt. Wir werden später aber noch sehen, dass für stetige Funktionen, die auf abgeschlossenen Intervallen definiert sind, die Umkehrung gilt.

Ist $(V, \|\cdot\|)$ ein normierter Raum, so ist die Norm $\|\cdot\| : V \to [0, \infty)$ Lipschitz-stetig, denn für $x, x' \in V$ zeigt ja die umgekehrte Dreiecksungleichung:

$$\big|\|x\| - \|x'\|\big| \leq \|x - x'\|\,.$$

Seien $(X, \mathbf{d}_1)$ und $(Y, \mathbf{d}_2)$ zwei metrische Räume, $f : X \to Y$ eine Abbildung und $x_0 \in X$. Konvergiert dann für jede Folge $(x_n)_{n\in\mathbb{N}}$ mit $x_n \to x_0$ und $x_n \neq x_0$ für alle $n \geq 1$ die Folge $(f(x_n))_{n\geq 1}$ gegen ein und dasselbe $y_0 \in Y$, so schreiben wir

$$\lim_{x\to x_0} f(x) = y_0 \,.$$

Äquivalent dazu ist die Aussage

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in U(x_0, \delta) \setminus \{x_0\} : f(x) \in U(f(x_0), \varepsilon) \,.$$

Der Beweis, der analog zu dem von 5.1.2 zu führen ist, wird übersprungen. Man beachte, dass i.A. nicht $y_0 = f(x_0)$ gelten muss. Genauer gesagt ist

$$\lim_{x\to x_0} f(x) = f(x_0)$$

äquivalent zur Stetigkeit von f in x_0.

In $\mathbb{R}$ können wir diese Definitionen noch etwas verfeinern. Sei dazu z.B. $f : [a, b) \to \mathbb{R}$ eine Funktion. Gibt es ein $y \in \mathbb{R}$, so dass für jede Folge $(x_n)_{n\in\mathbb{N}} \subset [a, b)$ mit $x_n \to b$ gilt $f(x_n) \to y$, so ist der **linksseitige Grenzwert**

$$\lim_{x\to b^-} f(x) = y\,.$$

Analog definieren wir den **rechtsseitigen Grenzwert** $\lim_{x\to a^+} f(x) = y$ für Funktionen $f : (a, b] \to \mathbb{R}$. Die gleichen Notationen werden auch für Punkte x_0 im Definitionsbereich von f verwendet, wenn nur eine einseitige Approximation von x_0 erlaubt sein soll. Es gilt

$$\lim_{x\to x_0} f(x) = y$$

genau dann, wenn die links- und rechtsseitige Grenzwerte existieren und $\lim_{x\to x_0^-} f(x) = y = \lim_{x\to x_0^+} f(x)$.

Section 5.2
Hauptsätze über stetige Funktionen

Im folgenden wollen wir den Wertebereich von stetigen Funktionen $f : [a, b] \to \mathbb{R}$ untersuchen. Wir beginnen mit dem sogenannten Nullstellensatz von Bolzano.

Theorem 5.2.1

Sei $f : [a, b] \to \mathbb{R}$ stetig mit

$$f(a) < 0 < f(b)\,.$$

Dann existiert ein $x_0 \in (a, b)$ mit $f(x_0) = 0$.

BEWEIS 1/2

Wir betrachten einen Beweis, der auch als Grundlage für einen Algorithmus zur Nullstellensuche dienen kann. Dazu definieren wir zwei Folgen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ rekursiv wie folgt: Für $n = 1$ setzen wir $a_1 := a$ und $b_1 := b$. Es gilt dann offensichtlich $f(a_1) < 0$ und $f(b_1) > 0$.

Haben wir die beiden Folgen für $n \geq 1$ schon definiert, so betrachten wir zunächst den Mittelpunkt $c_n := (a_n + b_n)/2$ des aktuellen Intervalls $[a_n, b_n]$. Gilt $f(c_n) = 0$, so haben wir eine Nullstelle gefunden, und eine weitere Konstruktion erübrigt sich. Im Fall $f(c_n) < 0$ setzen wir

$$a_{n+1} := c_n \qquad \text{und} \qquad b_{n+1} := b_n$$

und im Fall $f(c) > 0$ setzen wir

$$a_{n+1} := a_n \qquad \text{und} \qquad b_{n+1} := c_n\,.$$

BEWEIS 2/2

Da der Mittelpunkt c_n immer $a_n \leq c_n \leq b_n$ erfüllt, ist die Folge $(a_n)_{n\in\mathbb{N}}$ monoton wachsend und die Folge $(b_n)_{n\in\mathbb{N}}$ ist monoton fallend. Beide Folgen sind außerdem beschränkt, da sie im Intervall $[a, b]$ liegen, siehe auch Abbildung 14. Nach Satz 4.1.16 konvergieren daher beide Folgen. Ferner gilt

$$b_n - a_n = (b - a) \cdot 2^{n-1}, \qquad n \geq 1,$$

wie eine einfache Induktion zeigt. Damit folgt $\lim_{n\to\infty}(b_n - a_n) = 0$ und mit Satz 4.1.7 schließen wir $x_0 := \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$.

Offensichtlich gilt $x_0 \in [a, b]$. Ferner sichert unsere Konstruktion sowohl $f(a_n) < 0$ als auch $f(b_n) > 0$ für alle $n \geq 1$. Mit der Folgen-Stetigkeit und Satz 4.1.7 erhalten wir daher $f(x_0) = \lim_{n\to\infty} f(a_n) \leq 0$ und $f(x_0) = \lim_{n\to\infty} f(b_n) \geq 0$ und damit $f(x_0) = 0$. Dies schließt $x_0 = a$ und $x_0 = b$ aus.

BEWEIS MIT INTERVALL-HALBIERUNG

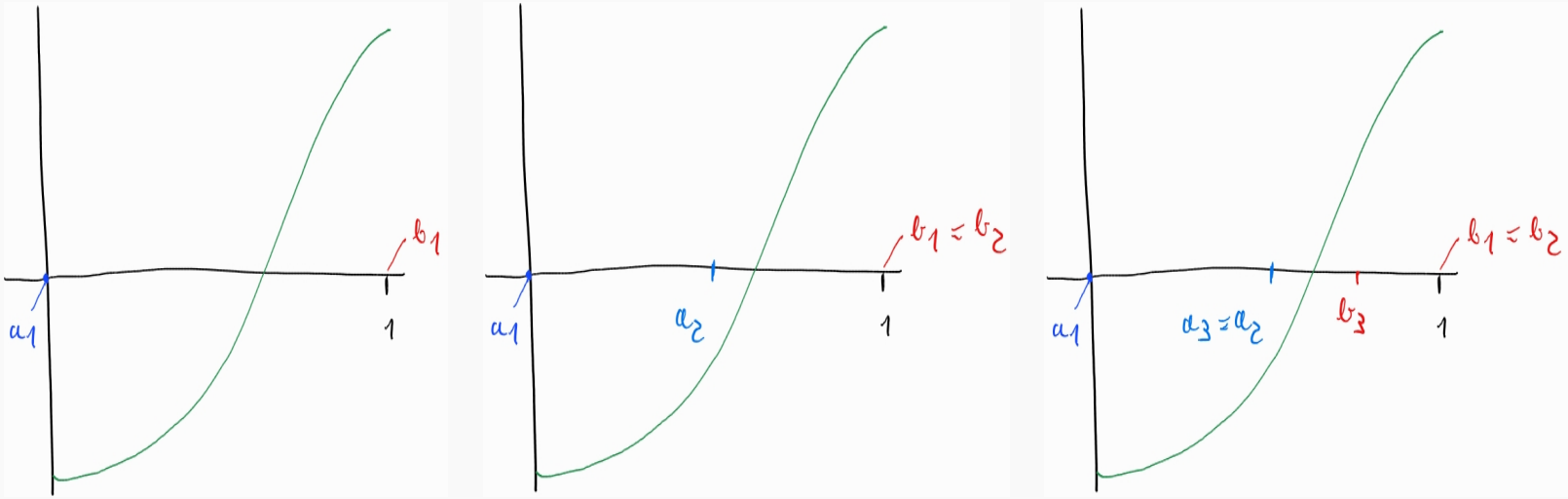


Abbildung: Beweis von Satz 5.2.1 mit Intervall-Halbierung für die grüne Funktion $f : [0, 1] \to \mathbb{R}$. **Links:** Das Anfangsintervall mit $a_1 := a = 0$ und $b_1 := b = 1$. **Mitte:** Die erste Intervall-Halbierung: Es wird $f(c_1)$ für den Mittelpunkt des aktuellen Intervalls $c_1 := (a_1 + b_1)/2$ betrachtet. Da $f(c_1) < 0$ ist, wird $a_2 := c_1$ und $b_2 := b_1$ gesetzt. **Rechts:** Die nächste Intervall-Halbierung: Es wird $f(c_2)$ für den Mittelpunkt des aktuellen Intervalls $c_2 := (a_2 + b_2)/2$ betrachtet. Da $f(c_2) > 0$ ist, wird diesmal $a_3 := a_2$ und $b_3 := c_2$ gesetzt.

Mit Hilfe des Nullstellensatzes können wir nun den folgenden Zwischenwertsatz einfach beweisen.

Theorem 5.2.2

Sei $f : [a, b] \to \mathbb{R}$ *eine stetige Funktion und seien*

$$y_* := \min\{f(a), f(b)\} \qquad \textit{und} \qquad y^* := \max\{f(a), f(b)\}\,.$$

Dann gibt es zu jedem $y \in (y_*, y^*)$ *ein* $x \in (a, b)$ *mit* $f(x) = y$.

Sei $y \in (y_*, y^*)$. Im Fall $f(a) < f(b)$ erfüllt die Funktion $g : [a, b] \to \mathbb{R}$, die durch $g(x) := f(x) - y$ definiert ist, sowohl

$$g(a) = f(a) - y = y_* - y < 0$$

als auch $g(b) = f(b) - y = y^* - y > 0$. Da g auch stetig ist, gibt es nach Satz 5.2.1 ein $x \in (a, b)$ mit $g(x) = 0$. Dies ergibt $f(x) = y$.

Im Fall $f(a) < f(b)$ betrachtet man stattdessen die durch $h(x) := y - f(x)$ definierte Funktion und wiederholt die Argumentation.

Um weitere Eigenschaften von stetigen Funktionen herzuleiten, ist der Begriff von Teilfolgen sehr nützlich. Konzeptionell passt dieser eher ins Kapitel 4.

Definition 5.2.3

Sei $(a_n)_{n\in\mathbb{N}}$ eine Folge und $(n_k)_{k\in\mathbb{N}} \subset \mathbb{N}$ eine streng wachsende Folge. Dann heißt $(a_{n_k})_{k\in\mathbb{N}}$ Teilfolge von $(a_n)_{n\in\mathbb{N}}$.

Jede Folge ist Teilfolge von sich, es gibt aber natürlich auch "echte" Teilfolgen: Für die durch $a_n := (-1)^n$ definierte Folge ist beispielsweise $(a_{2n})_{n\in\mathbb{N}}$ eine echte Teilfolge. Schließlich ist jede Teilfolge einer Teilfolge auch eine Teilfolge der ursprünglichen Folge.

Konvergiert eine Folge $(a_n)_{n\in\mathbb{N}}$ gegen ein a, so konvergiert auch jede Teilfolge von ihr gegen a, wie ein einfaches Anwenden der Definitionen zeigt. Haben wir umgekehrt eine Folge $(a_n)_{n\in\mathbb{N}}$, für die jede Teilfolge konvergiert, so ist die Folge selber schon konvergent, da sie ja eine Teilfolge von sich selbst ist.

Für reelle Folgen können wir Teilfolgen mit zusätzlichen Eigenschaften konstruieren. Dies ist das Ergebnis der folgenden beiden Sätze:

Theorem 5.2.4

Sei $(a_n)_{n\in\mathbb{N}}$ eine reelle Folge. Dann existiert eine monotone Teilfolge von $(a_n)_{n\in\mathbb{N}}$.

BEWEIS

Ein Folgenglied a_m heißt Gipfelstelle, falls $a_m \geq a_n$ für alle $n > m$ gilt.

Falls es unendlich viele verschiedene Gipfelstellen a_{n_k} gibt, können wir durch sukzessives Wählen der kleinsten verbleibenden Indizes n_k die Sortierung $n_1 < n_2 < n_3 < \dots$ herstellen. Die resultierende Folge $(a_{n_k})_{k \in \mathbb{N}}$ ist dann monoton fallend.

Falls es nicht unendlich viele Gipfelstellen gibt, gibt es ein n_1, so dass a_m für alle $m \geq n_1$ keine Spitze ist. Da a_{n_1} keine Gipfelstelle ist, gibt es ein $n_2 > n_1$ mit $a_{n_1} < a_{n_2}$. Da a_{n_2} auch keine Gipfelstelle ist, gibt es ein $n_3 > n_2$ mit $a_{n_2} < a_{n_3}$. Rekursiv erhalten wir somit eine monoton wachsende Teilfolge $(a_{n_k})_{k \in \mathbb{N}}$.

Der folgende Satz ist als Satz von Bolzano-Weierstraß bekannt.

Theorem 5.2.5

Sei $(a_n)_{n\in\mathbb{N}}$ eine reelle und beschränkte Folge. Dann existiert eine konvergente Teilfolge von $(a_n)_{n\in\mathbb{N}}$.

Beweis.

Nach Satz 5.2.4 existiert eine monotone Teilfolge von $(a_n)_{n\in\mathbb{N}}$. Diese ist ebenfalls beschränkt, und damit konvergent nach Satz 4.1.16. □

Ein metrischer Raum $(X, \mathbf{d})$ heißt **folgenkompakt**, falls es zu jeder Folge $(x_n)_{n\in\mathbb{N}} \subset X$ ein $x \in X$ und eine Teilfolge $(x_{n_k})_{k\in\mathbb{N}}$ gibt mit $x_{n_k} \to x$.

Satz 5.2.5 zeigt, dass die Intervalle $[a, b]$ folgenkompakt sind, da jede Folge $(x_n)_{n\in\mathbb{N}}$ in einem solchen Intervall beschränkt ist und Grenzwerte x von konvergenten Teilfolgen $(x_{n_k})_{k\in\mathbb{N}}$ wegen Satz 4.1.7 die Ungleichungen $a \leq x \leq b$ erfüllen müssen.

Durch sukzessives, komponentenweises Anwenden der obigen Argumentation kann man zeigen, dass z.B. auch Megen der Form $[-a, a]^d$ oder $B(0, r)$ folgenkompakt sind. Offene, nichtleere Intervalle sind dagegen *nie* folgenkompakt.

Die folgende Definition beschreibt das Verhalten bestimmter reeller Funktionen.

Definition 5.2.6

Eine Funktion $f : [a, b] \to \mathbb{R}$ heißt:

i). **monoton wachsend**, falls für alle $x_1, x_2 \in [a, b]$ gilt

$$x_1 < x_2 \quad \Rightarrow \quad f(x_1) \leq f(x_2)\,.$$

Entsprechend sagen wir sie sei **monoton fallend**, falls

$$x_1 < x_2 \quad \Rightarrow \quad f(x_1) \geq f(x_2)\,.$$

ii). **streng monoton wachsend**, falls für alle $x_1, x_2 \in [a, b]$ gilt

$$x_1 < x_2 \quad \Rightarrow \quad f(x_1) < f(x_2)\,.$$

Entsprechend sagen wir sie sei **streng monoton fallend**, falls

$$x_1 < x_2 \quad \Rightarrow \quad f(x_1) > f(x_2)\,.$$

Konstante Funktionen sind sowohl monoton wachsend als auch fallend, aber weder streng wachsend noch streng fallend. Die Funktion $x \mapsto x$ ist auf jedem Intervall $[a, b]$ streng wachsend, und die Funktion $x \mapsto x^2$ ist auf $[-1, 1]$ weder wachsend noch fallend.

Ist f (streng) wachsend, so ist $-f$ (streng) fallend, und umgekehrt. Ferner sind streng monotone Funktionen automatisch injektiv.

Der folgende Umkehrsatz zeigt, dass streng monotone, stetige Funktionen eine *stetige* Umkehrfunktion besitzen.

Theorem 5.2.7

Sei $f : [a, b] \to \mathbb{R}$ streng monoton wachsend und stetig. Dann existiert zu jedem $y \in [f(a), f(b)]$ genau ein $x \in [a, b]$ mit $f(x) = y$. Insbesondere existiert damit die Umkehrfunktion

$$f^{-1}(y) := x$$

auf $[f(a), f(b)]$. Diese ist streng monoton wachsend und stetig.

BEWEIS 1/2

Da $f(a) < f(b)$ ist, ist die Existenz von $x \in [a, b]$ mit $f(x) = y$ nach dem Zwischenwertsatz 5.2.2 gesichert. Die Eindeutigkeit von x folgt aus der Injektivität von f. Also folgt die Existenz der Umkehrfunktion

$$f^{-1} : [f(a), f(b)] \to [a, b] .$$

Um die strenge Monotonie von f^{-1} zu überprüfen, wählen wir $y_1 < y_2$. Wäre dann $x_1 := f^{-1}(y_1) \geq f^{-1}(y_2) =: x_2$, so würde $y_1 = f(x_1) \geq f(x_2) = y_2$ folgen. Also ist f^{-1} streng wachsend.

BEWEIS 2/2

Gäbe es nun ein $y \in [f(a), f(b)]$, so dass f^{-1} nicht stetig in y wäre, so finden wir eine Folge $(y_n)_{n\in\mathbb{N}}$ mit $y_n \to y$ und $f^{-1}(y_n) \not\to f^{-1}(y)$. Damit gibt es ein $\varepsilon > 0$ und eine Teilfolge $(y_{n_k})_{k\geq 1}$ mit

$$\left|f^{-1}(y_{n_k}) - f^{-1}(y)\right| > \varepsilon \tag{5.2.1}$$

für alle $k \geq 1$. Mit Satz 5.2.4 können wir zusätzlich annehmen, dass diese Teilfolge monoton ist. Wir setzen $x_k := f^{-1}(y_{n_k})$. Da f^{-1} wachsend ist, ist dann auch die Folge $(x_k)_{k\in\mathbb{N}}$ monoton, und da sie auch beschränkt ist, konvergiert sie gegen ein $x \in [a, b]$. Damit gilt aber mit der Stetigkeit von f

$$y = \lim_{k\to\infty} y_{n_k} = \lim_{k\to\infty} f(x_k) = f(\lim_{k\to\infty} x_k) = f(x)$$

und somit

$$f^{-1}(y) = x = \lim_{k\to\infty} x_k = \lim_{k\to\infty} f^{-1}(y_{n_k})\,.$$

Dies widerspricht (5.2.1).

BEISPIELE: WURZELFUNKTION

Für $x \geq 0$ ist die Funktion $f(x) := x^2$ streng monoton wachsend und stetig. Monotonie folgt dabei direkt aus

$$x_2 > x_1 \quad \Rightarrow \quad x_2^2 - x_1^2 = (x_2 - x_1)(x_2 + x_1) > 0\,.$$

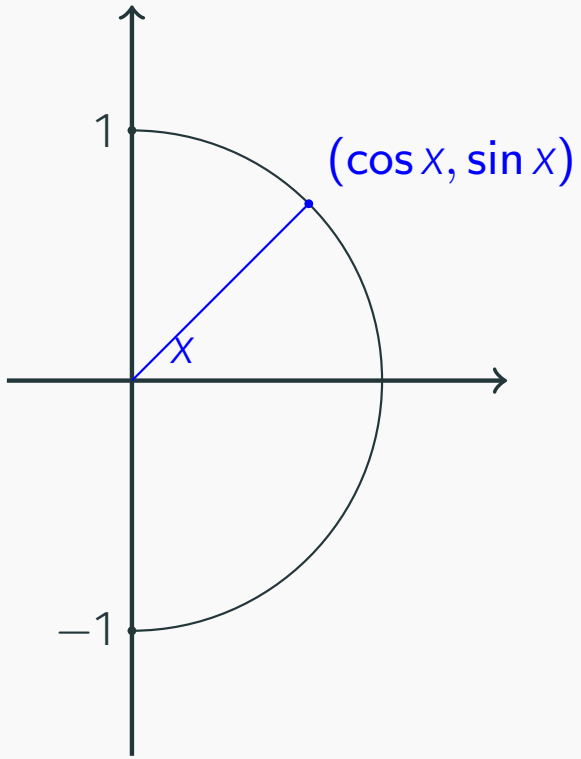
Damit existiert auf jedem Intervall $[0, b^2]$ die durch $f^{-1}(y) = \sqrt{y}$ gegebene Umkehrfunktion

$$f^{-1} : [0, b^2] \to [0, b]$$

und diese ist auch stetig. Da b beliebig war, ist auch die Wurzelfunktion $\sqrt{\cdot} : [0, \infty) \to [0, \infty)$ stetig.

BEISPIELE: ARKUSSINUS

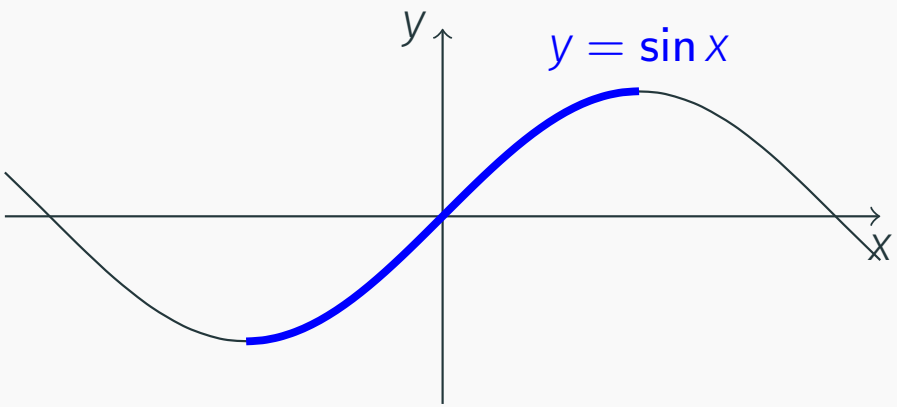
Auf $[-\frac{\pi}{2}, \frac{\pi}{2}]$ ist die Funktion $f(x) = \sin x$ stetig und streng monoton wachsend, wie die folgende Zeichnung illustriert:

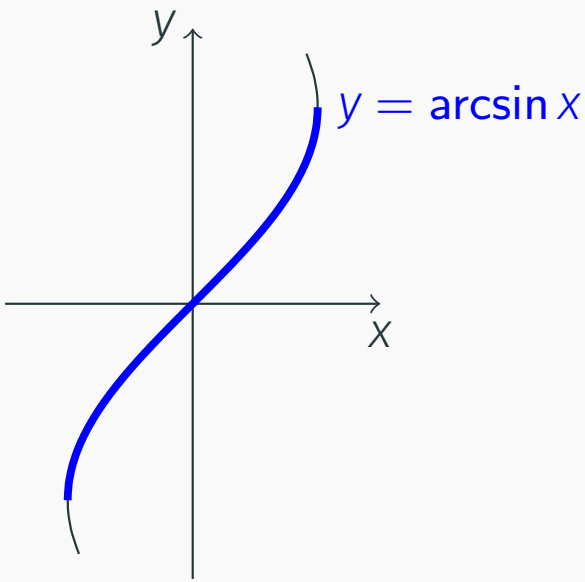


Damit existiert die als Arkussinus bezeichnete, streng monotone und stetige Umkehrfunktion

$$\arcsin : [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}] .$$

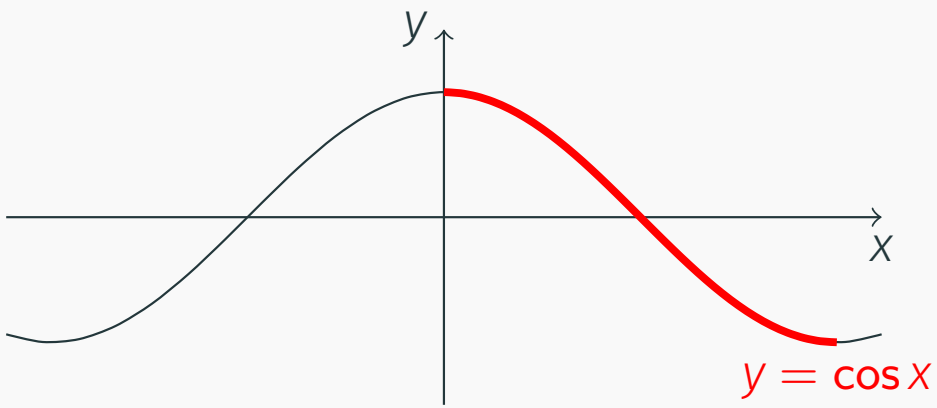
BEISPIELE: ARKUSSINUS

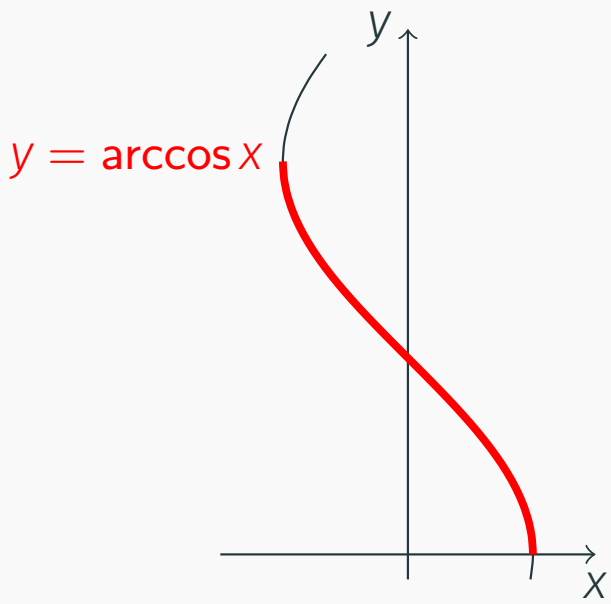




BEISPIELE: ARKUSKOSINUS

Entsprechend ist $\cos : [0, \pi]$ streng monoton fallend und damit existiert die stetig und monoton fallende Umkehrfunktion $\arccos : [-1, 1] \to [0, \pi]$, die als Arkuskosinus bezeichnet wird.





Der folgende Satz zeigt, das stetige Funktionen auf abgeschlossenen, beschränkten Intervallen beschränkt sind und Maximum und Minimum annehmen.

Theorem 5.2.8

Sei $f : [a, b] \to \mathbb{R}$ stetig. Dann existieren Punkte $x_*, x^* \in [a, b]$ mit

$$f(x_*) = \inf\{f(x) \mid x \in [a, b]\}$$

und

$$f(x^*) = \sup\{f(x) \mid x \in [a, b]\}\,.$$

Insbesondere ist f beschränkt und es werden das Minimum und Maximum der Funktionswerte angenommen.

Aus dem Beweis des Satzes 5.2.8 wird schnell ersichtlich, dass er für stetige Funktionen, die auf folgenkompakten Räumen definiert sind, ebenfalls gilt.

Für die Existenz von x^* setzen wir $W := \{f(x) \mid x \in [a,b]\}$ und $M := \sup\{f(x) \mid x \in [a,b]\}$. Da für $n \geq 1$ nach Definition $M - 1/n$ keine obere Schranke von W ist, gibt es ein $x_n \in [a,b]$ mit

$$M - \frac{1}{n} \leq f(x_n) \leq M.$$

Dies ergibt $f(x_n) \to M$. Ferner ist die Folge $(x_n)_{n\in\mathbb{N}}$ beschränkt und damit gibt es nach Satz 5.2.5 eine Teilfolge $(x_{n_k})_{k\in\mathbb{N}}$, die gegen ein $x^* \in \mathbb{R}$ konvergiert. Wegen $(x_{n_k})_{k\in\mathbb{N}} \subset [a,b]$ gilt zudem $x^* \in [a,b]$. Dies ergibt

$$M = \lim_{n\to\infty} f(x_n) = \lim_{k\to\infty} f(x_{n_k}) = f(x^*)\,.$$

Die Existenz von x_* folgt durch die Betrachtung von $-f$.

Im Folgenden schreiben wir

$$\mathcal{C}([a,b]) := \{f : [a,b] \to \mathbb{R} \mid f \text{ stetig}\}.$$

Wir wissen schon, dass dies ein Vektorraum ist. Ferner ist nach Satz 5.2.8

$$\|f\|_\infty := \sup_{x\in[a,b]} |f(x)| < \infty$$

für alle $f \in \mathcal{C}([a,b])$. Eine einfache Rechnung zeigt zudem, dass $\|\cdot\|_\infty$ eine Norm auf $\mathcal{C}([a,b])$ definiert. Konvergiert eine Folge $(f_n)_{n\in\mathbb{N}} \subset \mathcal{C}([a,b])$ bezüglich dieser Norm, so sprechen wir von gleichmäßiger Konvergenz.

Der folgende Satz untersucht die gleichmäßige Konvergenz.

Theorem 5.2.9

Seien $(f_n)_{n\in\mathbb{N}} \subset \mathcal{C}([a,b])$ und $f : [a,b] \to \mathbb{R}$ mit $\|f_n - f\|_\infty \to 0$. Dann gilt $f \in \mathcal{C}([a,b])$.

Konvergiert die Folge $(f_n)_{n\in\mathbb{N}}$ nur **punktweise**, d.h. $f_n(x) \to f(x)$ für alle $x \in [a, b]$, so ist die Grenzwertfunktion im Allgemeinen *nicht* stetig.

Dies kann man an dem Beispiel $f_n(x) := x^n$ für $x \in [0, 1]$ sehen, da in diesem Fall $f = \mathbf{1}_{\{1\}}$ gilt, siehe Lemma 4.1.13.

Ist $(x_k)_{k\in\mathbb{N}} \subset [a, b]$ eine Folge mit $x_k \to x$ und $(f_n)_{n\in\mathbb{N}}$ gemäß Satz 5.2.9, so gilt

$$\lim_{n\to\infty} \lim_{k\to\infty} f_n(x_k) = \lim_{n\to\infty} f_n(x) = f(x) = \lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} \lim_{n\to\infty} f_n(x_k)$$

Man darf also in diesem Fall die Grenzwertbildungen vertauschen. Wenn man $x := 1$ und $x_k := 1 - 1/k$ im obigen Beispiel wählt, so sehen wir, dass diese Vertauschung bei punktweiser Konvergenz im Allgemeinen nicht möglich ist.

Aus dem Beweis des Satzes 5.2.9 wird wieder schnell ersichtlich, dass er für stetige Funktionen, die auf folgenkompakten Räumen definiert sind, ebenfalls gilt.

Sei $x_0 \in [a, b]$ und $\varepsilon > 0$. Dann gibt es ein $n \geq 1$ mit

$$\sup_{x \in [a,b]} |f_n(x) - f(x)| = \|f_n - f\|_\infty < \frac{\varepsilon}{3} .$$

Da f_n stetig in x_0 ist, gibt es zudem ein $\delta > 0$, so dass für alle $x \in [a, b]$ mit $|x - x_0| < \delta$ die Abschätzung

$$\left|f_n(x) - f_n(x_0)\right| < \frac{\varepsilon}{3}$$

gilt. Für solche x folgt dann

$$\begin{aligned} \left|f(x) - f(x_0)\right| &\leq \left|f(x) - f_n(x)\right| + \left|f_n(x) - f_n(x_0)\right| + \left|f_n(x_0) - f(x_0)\right| \\ &< \varepsilon , \end{aligned}$$

und damit ist f stetig.

Wir hatten schon gesehen, dass gleichmäßig stetige Funktionen stetig sind, die Umkehrung im Allgemeinen aber falsch ist. Der folgende Satz liefert daher eine bemerkenswerte Aussage.

Theorem 5.2.10

Sei $f : [a, b] \to \mathbb{R}$ stetig. Dann ist f gleichmäßig stetig.

BEWEIS 1/2

Wir nehmen an, dass f nicht gleichmäßig stetig wäre. Nach der Definition von gleichmäßiger Stetigkeit gibt es dann ein $\varepsilon > 0$, so dass es zu jedem $n \geq 1$ Punkte $x_n, z_n \in [a, b]$ gibt mit

$$|x_n - z_n| < \frac{1}{n} \qquad \text{und} \qquad |f(x_n) - f(z_n)| \geq \varepsilon \,.$$

Nach dem Satz von Bolzano-Weierstraß, siehe Satz 5.2.5, gibt es dann ein $x \in [a, b]$ und eine Teilfolge $x_{n_k} \to x$. Wendet man das gleiche Argument auf die Folge $(z_{n_k})_{k \in \mathbb{N}}$ an, so erhalten wir eine Teilfolge $(z_{m_k})_{k \in \mathbb{N}}$ von $(z_{n_k})_{k \in \mathbb{N}}$, die gegen ein $z \in [a, b]$ konvergiert. Für die Teilfolge $(x_{m_k})_{k \in \mathbb{N}}$ von $(x_{n_k})_{k \in \mathbb{N}}$ gilt dann $x_{m_k} \to x$ und wegen

$$|x_{m_k} - z_{m_k}| < \frac{1}{m_k}$$

gilt außerdem $x = z$.

BEWEIS 2/2

Da f stetig ist, folgt

$$\lim_{k\to\infty} f(x_{m_k}) = f(x) = \lim_{k\to\infty} f(z_{m_k})\,.$$

Damit gibt es ein $k_0 \geq 1$, so dass für alle $k \geq k_0$ gilt

$$|f(x_{m_k}) - f(x)| < \frac{\varepsilon}{2} \qquad \text{und} \qquad |f(z_{m_k}) - f(x)| < \frac{\varepsilon}{2}\,.$$

Dies ergibt

$$|f(x_{m_k}) - f(z_{m_k})| \leq |f(x_{m_k}) - f(x)| + |f(x) - f(z_{m_k})| < \varepsilon\,,$$

was im Widerspruch zu dem obigen $|f(x_{m_k}) - f(z_{m_k})| \geq \varepsilon$ für alle $k \geq 1$ steht.

Section 5.3
Exponentialfunktion

Wir hatten in (4.1.3) die Euler'sche Zahl durch die Reihe $\sum_{k=0}^{\infty} \frac{1}{k!}$ definiert, und in (4.2.6) hatten wir gesehen, dass die Reihe $\sum_{k=0}^{\infty} \frac{z^k}{k!}$ für alle $z \in \mathbb{C}$ absolut konvergiert. Diese Einsichten wollen wir jetzt nutzen, um die Exponentialfunktion als Reihe zu definieren.

Definition 5.3.1

Die komplexe Exponentialfunktion $\exp : \mathbb{C} \to \mathbb{C}$ ist durch

$$\exp(z) := \mathrm{e}^z := \sum_{k=0}^{\infty} \frac{z^k}{k!}, \qquad z \in \mathbb{C}$$

definiert.

Betrachtet man die Exponentialfunktion nur für reelle Argumente $x \in \mathbb{R}$, so ergibt die Reihendarstellung sofort $\exp(x) \in \mathbb{R}$. Die resultierende Einschränkung $\exp : \mathbb{R} \to \mathbb{R}$ wird reelle Exponentialfunktion genannt.

Einsetzen von $z = 0$ und $z = 1$ in die Reihendarstellung der Exponentialfunktion ergibt

$$\exp(0) = 1 \qquad \text{und} \qquad \exp(1) = \mathrm{e}\,,$$

wobei für die zweite Formel die Definition (4.1.3) ausgenutzt wurde. Die folgende Proposition präsentiert zwei wichtige Eigenschaften der Exponentialfunktion.

Proposition 5.3.2

Die Exponentialfunktion $\exp : \mathbb{C} \to \mathbb{C}$ *ist stetig mit*

$$\exp(z_1 + z_2) = \exp(z_1) \cdot \exp(z_2)\,, \qquad z_1, z_2 \in \mathbb{C}.$$

BEWEIS 1/2

Der Einfachheit halber betrachten wir hier nur den reellen Fall. Der komplexe Fall ist aber wegen diverser Bemerkungen zu absoluter Konvergenz und Folgenkompaktheit im Skript komplett analog beweisbar.

Um die Stetigkeit in einem $x \in \mathbb{R}$ zu beweisen, setzen wir $a := |x| + 1$. Für $n \geq 1$ definieren wir weiter

$$f_n(y) := \sum_{k=0}^{n} \frac{y^k}{k!}\,, \qquad\qquad y \in [-a, a]\,.$$

Offensichtlich ist jede Funktion $f_n : [-a, a] \to \mathbb{R}$ stetig und ferner gilt

$$\| \exp_{|[-a,a]} - f_n \|_\infty = \sup_{y \in [-a,a]} \left| \exp(y) - f_n(y) \right| = \sup_{y \in [-a,a]} \left| \sum_{k=n+1}^{\infty} \frac{y^k}{k!} \right| \leq \sum_{k=n+1}^{\infty} \frac{a^k}{k!}$$

wobei wir nacheinander die Sätze 4.2.1 und 4.2.6 angewendet haben. Da die Reihe $\sum_{k=1}^{\infty} \frac{a^k}{k!}$ nach (4.2.6) konvergiert, zeigt Satz 4.2.1

$$\lim_{n \to \infty} \sum_{k=n+1}^{\infty} \frac{a^k}{k!} = 0$$

und damit haben wir $\| \exp_{|[-a,a]} - f_n \|_\infty \to 0$ gezeigt. Satz 5.2.9 liefert daher die Stetigkeit von $\exp_{|[-a,a]}$, und dies ergibt die Stetigkeit von $\exp$ in x.

BEWEIS 2/2

Da die Reihendarstellung der Exponentialfunktion in jedem Punkt absolut konvergiert, siehe (4.2.6), dürfen wir nach Satz 4.2.7 die Reihen umordnen. Mit dem Binomischen Lehrsatz 2.2.7 gilt dann

$$\begin{aligned}
\exp(x+y) = \sum_{n=0}^{\infty} \frac{(x+y)^n}{n!} &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k \\
&= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{x^{n-k}}{(n-k)!} \frac{y^k}{k!} \\
&= \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \frac{x^m}{m!} \frac{y^k}{k!} \\
&= \left(\sum_{k=0}^{\infty} \frac{y^k}{k!} \right) \cdot \left(\sum_{m=0}^{\infty} \frac{x^m}{m!} \right) \\
&= \exp(y) \exp(x) \,,
\end{aligned}$$

wobei wir in der vierten Gleichung das Cauchy-Produkt von Reihen ausgenutzt haben, dass in Abbildung 15 illustriert ist.

CAUCHY-PRODUKT

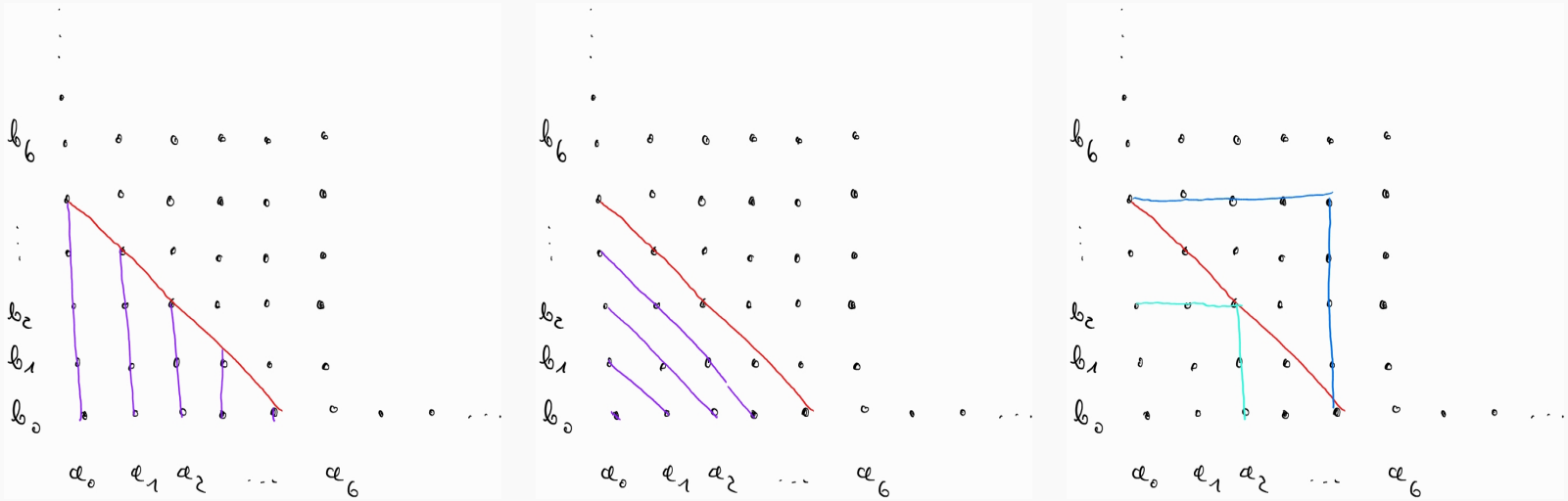


Abbildung: Cauchy-Produkt $(\sum_{i=0}^{\infty} a_i) \cdot (\sum_{j=0}^{\infty} b_j) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k}$, wobei die absolute Konvergenz der beiden linken Reihen gefordert wird und die absolute Konvergenz der rechten Reihe folgt. **Links:** Spaltenweise (lila) Summation $\sum_{i=0}^{n} \sum_{j=0}^{n-i} a_i b_j$ über das rote Dreieck. **Mitte:** Äquivalente, diagonale Summation $\sum_{i=0}^{n} \sum_{j=0}^{i} a_i b_{i-j}$ über das Dreieck. **Rechts:** Einschachtelung der Dreiecks-Summe durch 2 Quadratsummen $(\sum_{i=0}^{m} a_i) \cdot (\sum_{j=0}^{m} b_j) = \sum_{i=0}^{m} \sum_{j=0}^{m} a_i b_j$ für $m = \lfloor n/2 \rfloor$ (türkis) und $m = n$ (blau). Die Differenz zwischen blauen Quadrat und dem Dreieck kann nach oben durch die Differenz zwischen blauen und türkisen Quadrat abgeschätzt werden.

Ist $f : \mathbb{R} \to \mathbb{R}$ eine Funktion mit $f(1) = \mathrm{e}$ und $f(x+y) = f(x)f(y)$ für alle $x, y \in \mathbb{R}$, so gilt zunächst

$$\mathrm{e} = f(1) = f(m/m) = f(1/m)^m\,, \qquad m \in \mathbb{N}$$

und damit $f(1/m) = \mathrm{e}^{1/m}$. Für $q = k/m \in \mathbb{Q}$ mit $k \in \mathbb{Z}$ und $m \in \mathbb{N}$ folgt

$$f(q) = f(k/m) = f(1/m)^k = \mathrm{e}^{k/m} = \mathrm{e}^q\,. \tag{5.3.1}$$

Die obigen Funktionalgleichungen erzwingen also ein eindeutiges Verhalten der Funktion f auf $\mathbb{Q}$. Da die Exponentialfunktion beide Gleichungen erfüllt, muss also $\mathbf{exp}(q) = \mathrm{e}^q$ für alle $q \in \mathbb{Q}$ gelten, was zumindest für rationale Argumente unsere Notation $\mathbf{exp}(x) = \mathrm{e}^x$ rechtfertigt. Ferner ist die Exponentialfunktion stetig und man kann mit (5.3.1) und der Approximation von $\mathbb{R}$ durch $\mathbb{Q}$ zeigen, dass es nur ein stetiges f, das die beiden Gleichungen erfüllt, gibt. Daher ist unsere Definition der Exponentialfunktion identisch zu den in den vorherigen Kapiteln benutzte, heuristische Herangehensweise.

Ferner lässt sich das Lemma 4.1.17 zu der Gleichung

$$\exp(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n, \qquad x \in \mathbb{R}$$

verallgemeinern, die wir hier aber nicht beweisen wollen.

Korollar 5.3.3

Es gilt $\exp(x) > 0$ *für alle* $x \in \mathbb{R}$ *sowie* $\exp(z) \neq 0$ *und* $\exp(\overline{z}) = \overline{\exp(z)}$ *für alle* $z \in \mathbb{C}$.

Da $\exp(0) = 1$ gilt, folgt für jedes $z \in \mathbb{C}$ schon

$$1 = \exp(z)\,\exp(-z)$$

und damit $\exp(z) \neq 0$. Weiter ist für jedes $x \geq 0$ auch jede Partialsumme der Reihe positiv und es folgt $\exp(x) > 0$ für $x \geq 0$. Zusammen mit $\exp(-x) = \frac{1}{\exp(x)}$ folgt damit $\exp(x) > 0$ für alle $x \in \mathbb{R}$.

Ferner gilt

$$\exp(\bar{z}) = \sum_{k=0}^{\infty} \frac{\bar{z}^k}{k!} = \sum_{k=0}^{\infty} \frac{\overline{z^k}}{k!} = \overline{\sum_{k=0}^{\infty} \frac{z^k}{k!}} = \overline{\exp(z)}\,,$$

wobei wir im vorletzten Schritt die Stetigkeit der komplexen Konjugation $z \mapsto \bar{z}$ benutzt haben.

Wir wollen nun die Logarithmusfunktion als Umkehrfunktion der Exponentialfunktion einführen. Dies geschieht in dem folgenden Lemma.

Lemma 5.3.4

Die Funktion $\exp : \mathbb{R} \to \mathbb{R}$ *ist streng monoton wachsend und erfüllt*

$$\lim_{x \to +\infty} \exp(x) = +\infty,$$

$$\lim_{x \to -\infty} \exp(x) = 0.$$

Damit existiert ihre Umkehrfunktion

$$\ln : (0, \infty) \to \mathbb{R}\,,$$

die durch

$$y = \exp(x) \quad \Leftrightarrow \quad x = \ln y$$

charakterisiert ist, und diese ist ebenso streng monoton wachsend und stetig.

Für $x \geq 0$ liefert die Reihendarstellung

$$\exp(x) \geq 1 + x \tag{5.3.2}$$

damit $\lim_{x\to+\infty} \exp(x) = +\infty$. Damit folgt auch

$$\lim_{x\to-\infty} \exp(x) = \lim_{x\to-\infty} \frac{1}{\exp(-x)} = \frac{1}{\lim_{x\to+\infty} \exp(x)} = 0\,.$$

Für $x > 0$ liefert unsere anfängliche Abschätzung zudem $\exp(x) \geq 1 + x > 1$ und für $x_2 > x_1$ ergibt dies

$$\exp(x_2) - \exp(x_1) = \exp(x_1)\big(\exp(x_2 - x_1) - 1\big) > 0.$$

Damit folgt die Behauptung aus dem Umkehrsatz 5.2.7.

Chapter 6: Differentialrechnung

Differentialrechnung dient unter anderem dazu

- Extremwerte zu berechnen (Stetigkeit lieferte uns die Existenz aber keine Methode zur Berechnung der Extremstellen)
- Geschwindigkeiten als Änderungsraten von Größen mathematisch zu beschreiben
- Gleichungen für Änderungsraten als Differentialgleichungen für Funktionen zu verstehen.

Wir werden in diesem Kapitel werden wir daher die Differentialrechnung rigoros aufbauen und beweisen. Viele Aspekte werden dabei aus der Schule in der einen oder anderen Form schon bekannt sein.

Section 6.1
Differenzierbarkeit

Definition 6.1.1

Eine Funktion $f : [a, b] \to \mathbb{R}$ heißt **differenzierbar im Punkt** $x_0 \in (a, b)$, falls der Grenzwert

$$f'(x_0) := \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

existiert. Sie heißt **differenzierbar** auf (a, b), falls sie in jedem $x_0 \in (a, b)$ differenzierbar ist. In diesem Fall bezeichnet man die Funktion $f' : (a, b) \to \mathbb{R}$ als die **Ableitung** von f.

Analog sind alle diese Begriffe für $f : \mathbb{R} \to \mathbb{R}$, $f : (a, b) \to \mathbb{R}$ etc. definiert, wobei x_0 immer im "Inneren" des Definitionsbereichs von f liegen muss.

Jede konstante Funktion $f(x) = c$ auf einem Intervall $[a, b]$ ist differenzierbar mit $f'(x_0) = 0$ für alle $x_0 \in (a, b)$, denn es gilt

$$\lim_{x \to x_0} \frac{c - c}{x - x_0} = 0\,.$$

Die Funktion $f(x) = x$ ist in jedem $x_0 \in \mathbb{R}$ differenzierbar mit $f'(x_0) = 1$, denn offensichtlich gilt

$$\lim_{x \to x_0} \frac{x - x_0}{x - x_0} = 1\,.$$

Beschreibt $t \mapsto s(t)$ den zurückgelegten Weg s in Abhängigkeit von der Zeit t, so entspricht die Ableitung

$$\lim_{\Delta t \to 0} \frac{s(t+\Delta t) - s(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t}$$

der Geschwindigkeit zum Zeitpunkt t. Dabei haben wir Δt für kleine Zeitdifferenzen und Δs für die in diesen Zeiten zurückgelegten Wegdifferenzen geschrieben. Der Grenzwert des Differenzenquotienten wird oft als Differentialquotient bezeichnet und entsprechend suggestiv als

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t}$$

geschrieben, und als $\mathrm{d}s$ nach $\mathrm{d}t$ gelesen. Während es sich bei Δs und Δt jedoch um Zahlen handelt, haben hierbei $\mathrm{d}s$ und $\mathrm{d}t$ keine eigenständige Bedeutung und sie sind daher als rein formale Symbole zu verstehen.

Wir sagen $f : [a, b] \to \mathbb{R}$ ist **linksseitig differenzierbar** in $x_0 \in [a, b]$, falls der einseitige Grenzwert

$$\lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0}$$

existiert. Analog ist f **rechtsseitig differenzierbar** in $x_0 \in [a, b]$, falls der einseitige Grenzwert

$$\lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}$$

existiert. Eine Funktion ist in x_0 differenzierbar, genau dann, wenn sie links- und rechtsseitig differenzierbar in x_0 ist *und* die beiden einseitigen Ableitungen gleich sind.

Analog sind alle diese Begriffe für $f : \mathbb{R} \to \mathbb{R}$ etc. definiert.

Die Abbildung $x \mapsto |x|$ ist in 0 stetig, aber nicht differenzierbar, da die linksseitige Ableitung -1 ist, die rechtsseitige aber 1 ist.

Lemma 6.1.2

Die Exponentialfunktion $\exp : \mathbb{R} \to \mathbb{R}$ *ist differenzierbar und es gilt* $\exp' = \exp$.

BEWEIS

Wir zeigen die Aussage zunächst für $x_0 = 0$. Für $x > 0$ gilt wegen $\exp(x) \geq 1 + x$, siehe auch (5.3.2), die Abschätzung

$$1 \leq \frac{\exp(x) - 1}{x} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{n!} = \sum_{k=0}^{\infty} \frac{x^k}{(k+1)!} \leq \exp(x)\,.$$

Mit der Stetigkeit der Exponentialfunktion, siehe Proposition 5.3.2, und $\exp(0) = 1$ folgt

$$\lim_{x \to 0^+} \frac{\exp(x) - 1}{x - 0} = 1.$$

Damit folgt aber ebenso

$$\lim_{x \to 0^-} \frac{\exp(x) - 1}{x} = \lim_{x \to 0^-} \exp(x) \frac{1 - \exp(-x)}{x} = \lim_{x \to 0^-} \exp(x) \frac{\exp(-x) - 1}{-x} = 1\,.$$

Zusammen ergibt dies die Differenzierbarkeit in $x_0 = 0$ mit $\exp'(x_0) = 1$. Für allgemeine x_0 gilt nun

$$\lim_{x \to x_0} \frac{\exp(x) - \exp(x_0)}{x - x_0} = \exp(x_0) \lim_{x \to x_0} \frac{\exp(x - x_0) - 1}{x - x_0} = \exp(x_0)\,,$$

womit die Behauptung bewiesen ist.

Das folgende Lemma bestimmt die Ableitungen der geometrisch eingeführten Winkelfunktionen.

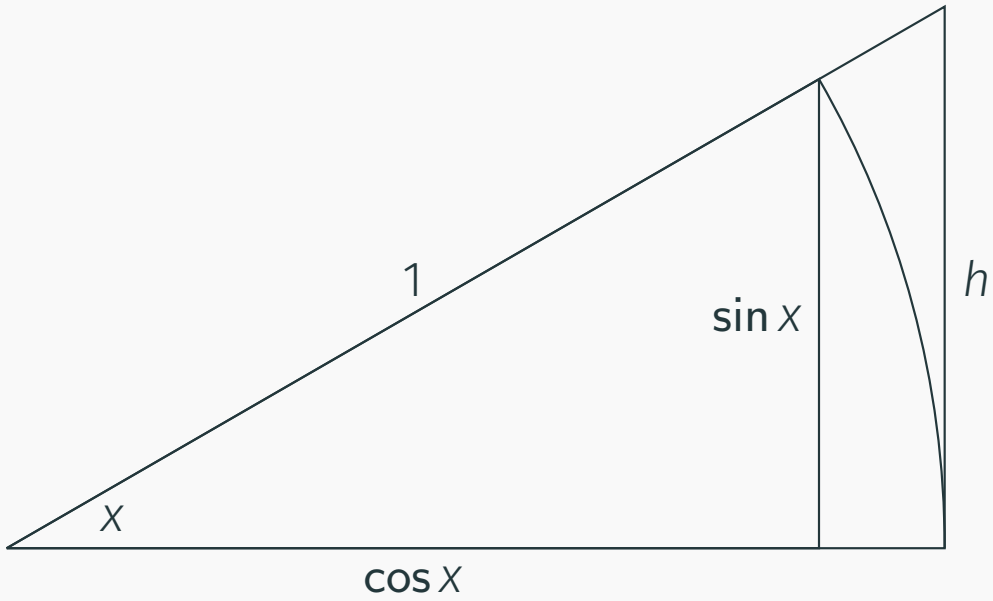
Lemma 6.1.3

Die Winkelfunktionen $\sin$ *und* $\cos$ *sind differenzierbar und für alle* $x \in \mathbb{R}$ *gilt:*

$$\sin' x = \cos x \qquad \textit{und} \qquad \cos' x = -\sin x.$$

BEWEIS 1/3

Wir zeigen zunächst $\sin'(0) = 1$. Dazu betrachten wir zunächst für $x \in (0, \pi/2)$ die folgende Skizze am Einheitskreis



BEWEIS 2/3

Die Fläche A_1 des kleineren Dreiecks lässt sich dann durch

$$A_1 = \frac{1}{2} \cos x \sin x$$

berechnen. Nach dem Strahlensatz gilt ferner $h = \frac{h}{1} = \frac{\sin x}{\cos x}$ und damit ist die Fläche A_3 des größeren Dreiecks

$$A_3 = \frac{1}{2} \cdot \frac{\sin x}{\cos x}.$$

Schließlich ist die Fläche A_2 des Kreissegments proportional zum Winkel, und da der Einheitskreis die Fläche π und den Umfang 2π hat, gilt folglich

$$\frac{A_2}{\pi} = \frac{x}{2\pi}.$$

Aus $A_1 \leq A_2 \leq A_3$ schließen wir nun $\cos x \sin x \leq x \leq \frac{\sin x}{\cos x}$ und damit auch

$$\cos x \leq \frac{\sin x}{x} \leq \frac{1}{\cos x}.$$

Wegen $\lim_{x \to 0^+} \cos x = 1$ folgt

$$\lim_{x \to 0^+} \frac{\sin x - \sin 0}{x - 0} = \lim_{x \to 0^+} \frac{\sin x}{x} = 1$$

und wegen $\frac{\sin(-x) - \sin 0}{-x - 0} = \frac{\sin x}{x}$ gilt das gleiche für die linksseitige Ableitung.

BEWEIS 3/3

Als nächstes zeigen wir $\cos'(0) = 0$. Dazu betrachten wir zunächst

$$\frac{\cos x - 1}{x} = \frac{(\cos x - 1)(\cos x + 1)}{x(\cos x + 1)} = \frac{-\sin^2 x}{x(\cos x + 1)} = \frac{\sin x}{x} \cdot \frac{-\sin x}{\cos x + 1} \to 1 \cdot 0$$

für $x \to 0$.

Mit diesen Vorbereitungen und den Additionstheoremen können wir nun schließen:

$$\begin{aligned}
\lim_{h \to 0} \frac{\sin(x+h) - \sin x}{h} &= \lim_{h \to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h} \\
&= \sin x \lim_{h \to 0} \frac{\cos h - 1}{h} + \cos x \lim_{h \to 0} \frac{\sin h}{h} \\
&= \cos x
\end{aligned}$$

Der Beweis für $\cos' x = -\sin x$ ist analog.

Aus der Schule ist bekannt, dass die Ableitung $f'(x_0)$ als Anstieg der Tangente am Graphen der Funktion im Punkt $(x_0, f(x_0))$ interpretiert werden. Diese Gerade g ist durch die Gleichung

$$g(x) = f'(x_0)(x - x_0) + f(x_0)\,, \qquad x \in \mathbb{R}$$

gegeben. Bildet man die Differenz zwischen beiden Funktionen, so erhält man den Approximationsfehler

$$R_1(x) := g(x) - f(x) = f'(x_0)(x - x_0) + f(x_0) - f(x)$$

und die Definition der Differenzierbarkeit bedeutet gerade, dass

$$\lim_{x \to x_0} \frac{R_1(x)}{x - x_0} = \lim_{x \to x_0} \left(f'(x_0) - \frac{f(x) - f(x_0)}{x - x_0} \right) = 0 \tag{6.1.1}$$

gilt. Man sagt, der Fehler sei **klein-o** von $x - x_0$ und schreibt

$$R_1(x) = o(x - x_0)\,, \qquad x \to x_0,$$

Der Approximationsfehler konvergiert also schneller als linear gegen 0. Ferner zeigt eine analoge Rechnung, dass jede andere lineare Approximation

$$a(x) = m \cdot (x - x_0) + f(x_0)$$

von f im Punkt x_0 *keinen* $o(x - x_0)$ Fehler hat.

Ferner zeigt der folgende Satz, dass die Differenzierbarkeit in x_0 sogar äquivalent zur linearen Approximierbarkeit in x_0 mit $o(x - x_0)$ Fehler ist.

Theorem 6.1.4

Seien $f : [a, b] \to \mathbb{R}$ und $x_0 \in (a, b)$. Dann sind äquivalent:

i). f ist in x_0 differenzierbar

ii). Es existiert eine Konstante $c \in \mathbb{R}$ und eine Funktion $\varphi : [a, b] \to \mathbb{R}$ mit $\varphi(x) = o(x - x_0)$ für $x \to x_0$, so dass

$$f(x) = f(x_0) + c \cdot (x - x_0) + \varphi(x)\,, \qquad x \in [a, b]. \tag{6.1.2}$$

In diesem Fall ist $c = f'(x_0)$.

BEWEIS

i) $\Rightarrow$ *ii)*. Für $c = f'(x_0)$ und $\varphi(x) := R_1(x)$ haben wir dies in (6.1.1) gezeigt.

ii) $\Rightarrow$ *i)*. Einsetzen von (6.1.2) in den Differenzenquotienten ergibt

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{c \cdot (x - x_0) + \varphi(x)}{x - x_0} = c + \frac{\varphi(x)}{x - x_0}$$

Für $x \to x_0$ folgt dann die Behauptung.

Korollar 6.1.5

Ist $f : [a, b] \to \mathbb{R}$ in $x_0 \in (a, b)$ differenzierbar und $\varphi : [a, b] \to \mathbb{R}$ gemäß (6.1.2)*. Dann ist f in x_0 stetig und es gilt*

$$|f(x) - f(x_0)| \leq \left| c + \frac{\varphi(x)}{x - x_0} \right| \cdot |x - x_0|, \qquad x \in [a, b].$$

Beweis.

Mit (6.1.2) gilt

$$|f(x) - f(x_0)| = |c \cdot (x - x_0) + \varphi(x)| = \left| \left(c + \frac{\varphi(x)}{x - x_0} \right) \cdot (x - x_0) \right|.$$

Damit folgt die Ungleichung und die Stetigkeit ist eine direkte Konsequenz der Ungleichung und der Eigenschaften von φ. □

Man kann explizit Funktionen angeben, die stetig aber in keinem Punkt differenzierbar sind. Ein Beispiel hierfür ist die sogenannte Weierstraß-Funktion. Die vom Wiener-Prozess, der auch als Brown'sche Bewegung bekannt ist, erzeugten zufälligen Funktionen sind mit Wahrscheinlichkeit 1 in keinem Punkt differenzierbar.

Wie schon bei Grenzwerten, Reihen und der Stetigkeit kann die Betrachtung von Ableitungen erheblich durch einige wichtige Rechenregeln vereinfacht werden.

Theorem 6.1.6

Seien $f, g : [a, b] \to \mathbb{R}$ in $x_0 \in (a, b)$ differenzierbar und $\alpha, \beta \in \mathbb{R}$. Dann gilt:

i). ***Linearität:*** *Die Linearkombination $\alpha f + \beta g$ ist in x_0 differenzierbar und es gilt*

$$(\alpha f + \beta g)'(x_0) = \alpha f'(x_0) + \beta g'(x_0)\,.$$

ii). ***Produktregel:*** *Das Produkt fg ist in x_0 differenzierbar und es gilt*

$$(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)\,.$$

iii). ***Quotientenregel:*** *Falls $g(x_0) \neq 0$ gilt, so ist der Quotient $\frac{f}{g}$ in x_0 differenzierbar mit*

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2}\,.$$

i). Folgt sofort aus der Linearität des Grenzwertes, siehe Satz 4.1.7.

ii). Addieren von $0 = -f(x_0)g(x) + f(x_0)g(x)$ im Zähler ergibt

$$\begin{aligned}
&\lim_{x\to x_0} \frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0} \\
&= \lim_{x\to x_0} \frac{f(x)g(x) - f(x_0)g(x) + f(x_0)g(x) - f(x_0)g(x_0)}{x - x_0} \\
&= \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} g(x) + f(x_0) \lim_{x\to x_0} \frac{g(x) - g(x_0)}{x - x_0} \\
&= f'(x_0)g(x_0) + f(x_0)g'(x_0)\,,
\end{aligned}$$

da g als in x_0 differenzierbare Funktion dort auch stetig ist, siehe Korollar 6.1.5.

iii). Für den Spezialfall $f = 1$ gilt

$$\lim_{x\to x_0} \frac{\frac{1}{g(x)} - \frac{1}{g(x_0)}}{x - x_0} = \lim_{x\to x_0} \frac{g(x) - g(x_0)}{x - x_0} \frac{-1}{g(x)g(x_0)} = g'(x_0) \frac{-1}{(g(x_0))^2}\,,$$

wobei wegen $g(x_0) \neq 0$ und der Stetigkeit von g in x_0 auch $g(x) \neq 0$ für alle hinreichend kleinen Abstände $|x - x_0|$ gilt. Mit *ii)* folgt der allgemeine Fall.

BEISPIELE: MONOME UND POLYNOME

Für $k \in \mathbb{N}_0$ ist die $g_k(x) := x^k$ differenzierbar und es gilt

$$g_k'(x) = kx^{k-1}\,. \tag{6.1.3}$$

Für $k \in \{0, 1\}$ hatten wir dies schon gesehen, und ist die Aussage für k schon bewiesen, so folgt mit der Produktregel:

$$g_{k+1}'(x) = (g_k \cdot g_1)'(x) = g_k'(x)g_1(x) + g_k(x)g_1'(x) = kx^{k-1}x + x^k = (k+1)x^k\,.$$

Mit der Linearität sind damit auch Polynome $p(x) = \sum_{k=0}^{n} a_k x^k$ differenzierbar mit

$$p'(x) = \sum_{k=1}^{n} k a_k x^{k-1}\,, \qquad x \in \mathbb{R}. \tag{6.1.4}$$

Zudem gilt mit der Quotientenregel, dass für $k \in \mathbb{N}$ die Funktion $f_k(x) = x^{-k} = \frac{1}{g_k(x)}$ in $x \neq 0$ differenzierbar ist mit

$$f_k'(x) = \left(\frac{1}{g_k}\right)'(x) = -\frac{g_k'(x)}{(g_k(x))^2} = -\frac{kx^{k-1}}{x^{2k}} = -kx^{-k-1}\,.$$

Insgesamt gilt also (6.1.3) für alle $k \in \mathbb{Z}$.

BEISPIELE: TANGENS UND KOTANGENS

Die Funktion $\tan x = \frac{\sin x}{\cos x}$ ist nach der Quotientenregel in allen $x \in \mathbb{R}$ mit $\sin x \neq 0$ differenzierbar und mit wegen $\sin^2 x + \cos^2 x = 1$ gilt

$$\tan' x = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = 1 + \tan^2 x\,. \tag{6.1.5}$$

Analog ist der Kotangens $\cot x = \frac{\cos x}{\sin x}$ in allen $x \in \mathbb{R}$ mit $\cos x \neq 0$ differenzierbar und es gilt

$$\cot' x = \frac{-1}{\sin^2 x} = -1 - \cot^2 x\,. \tag{6.1.6}$$

Das folgende Resultat, das als Kettenregel bekannt ist, betrachtet die Differenzierbarkeit von Kompositionen:

Theorem 6.1.7

Seien $f : \mathbb{R} \to \mathbb{R}$ und $g : \mathbb{R} \to \mathbb{R}$ Funktionen, so dass f in x_0 differenzierbar und g in $f(x_0)$ differenzierbar ist. Dann ist $x \mapsto g \circ f(x)$ in x_0 differenzierbar und es gilt

$$(g \circ f)'(x_0) = g'(f(x_0)) \cdot f'(x_0)\,.$$

Intuitiv folgt die Aussage sofort aus

$$\begin{aligned}\lim_{x \to x_0} \frac{g(f(x)) - g(f(x_0))}{x - x_0} &= \lim_{x \to x_0} \frac{g(f(x)) - g(f(x_0))}{f(x) - f(x_0)} \cdot \frac{f(x) - f(x_0)}{x - x_0} \\ &= g'(f(x_0))f'(x_0).\end{aligned}$$

Das Problem an dieser Argumentation ist nur, dass aus $x \neq x_0$ nicht $f(x) \neq f(x_0)$ folgt, und somit der erste Differenzenquotient nicht definiert sein muss.

Wir fixieren eine Folge $(x_n)_{n\in\mathbb{N}}$ mit $x_n \to x_0$ und $x_n \neq x_0$ für alle $n \geq 1$. Ferner schreiben wir $y_n := f(x_n)$ und $y_0 := f(x_0)$. Um mit dem Fall $f(x_n) = f(x_0)$ umgehen zu können, definieren wir $g^* : [a,b] \to \mathbb{R}$ durch

$$g^*(y) := \begin{cases} \frac{g(y)-g(y_0)}{y-y_0} & \text{falls } y \neq y_0\,, \\ g'(y_0) & \text{falls } y = y_0\,. \end{cases}$$

Damit gilt $g^*(y_n) \to g'(y_0) = g'(f(x_0))$ für $n \to \infty$ und

$$g(y) - g(y_0) = g^*(y) \cdot (y - y_0)\,, \qquad y \in \mathbb{R}.$$

Für $y = y_n = f(x_n)$ ergibt dies

$$\begin{aligned} \lim_{n\to\infty} \frac{g(f(x_n)) - g(f(x_0))}{x_n - x_0} &= \lim_{n\to\infty} \frac{g^*(y_n)\cdot(f(x_n) - f(x_0))}{x_n - x_0} \\ &= \lim_{n\to\infty} g^*(y_n) \cdot \lim_{n\to\infty} \frac{f(x_n) - f(x_0)}{x_n - x_0} \\ &= g'(f(x_0)) \cdot f'(x_0)\,. \end{aligned}$$

Da die Wahl der Folge beliebig war, folgt die Behauptung.

Theorem 6.1.8

Ist $f : [a, b] \to \mathbb{R}$ streng monoton und differenzierbar und gilt $f'(x) \neq 0$ auf (a, b), so ist die Umkehrfunktion $g := f^{-1}$ ebenso differenzierbar und es gilt

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

für alle $y \in f\big((a, b)\big)$, wobei $f\big((a, b)\big)$ das Bild von (a, b) unter f ist.

Falls wir schon wissen, dass g differenzierbar ist, folgt die Formel aus $g(f(x)) = x = \mathsf{id}(x)$ und der Kettenregel

$$1 = \mathsf{id}'(x) = (g \circ f)'(x) = g'(f(x)) \cdot f'(x) = g'(y) \cdot f'(g(y)) \,,$$

wobei im letzten Schritt $y := f(x)$, d.h. $g(y) = f^{-1}(f(x)) = x$ gesetzt wurde.

Sei $(y_n)_{n\in\mathbb{N}}$ mit $y_n \neq y$ und $y_n \to y$. Wir schreiben $x_n := g(y_n)$ und $x := g(y)$. Wegen $y_n \neq y$ gilt dann $x_n \neq x$ und da g nach Satz 5.2.7 stetig ist haben wir auch $x_n \to x$. Es folgt dann

$$\lim_{n\to\infty} \frac{g(y_n) - g(y)}{y_n - y} = \lim_{n\to\infty} \frac{x - x_n}{f(x_n) - f(x)} = \frac{1}{f'(x)} = \frac{1}{f'(g(y))}\,.$$

Da die Wahl der Folge beliebig war, folgt die Behauptung.

Da die Logarithmusfunktion $\ln : (0,\infty) \to \mathbb{R}$ die Umkehrfunktion der Exponentialfunktion $\exp : \mathbb{R} \to (0,\infty)$ ist und diese differenzierbar mit $\exp'(x) = \exp(x) \neq 0$ für alle $x \in \mathbb{R}$ ist, siehe Lemma 6.1.2 und Korollar 5.3.3, ist die Logarithmusfunktion differenzierbar und es gilt

$$\ln'(y) = \frac{1}{\exp'(\ln(y))} = \frac{1}{\exp(\ln(y))} = \frac{1}{y}\,. \tag{6.1.7}$$

Die Arkussinusfunktion $\arcsin : [-1,1] \to [-\pi/2, \pi/2]$ ist differenzierbar

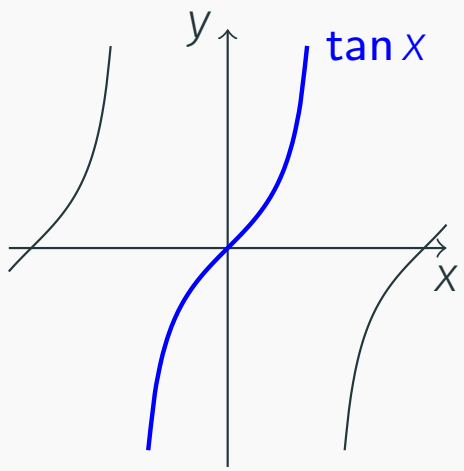
$$\begin{aligned}\arcsin'(y) = \frac{1}{\sin'(\arcsin y)} = \frac{1}{\cos(\arcsin y)} &= \frac{1}{\sqrt{1-\sin^2(\arcsin y)}} \\ &= \frac{1}{\sqrt{1-y^2}}\,,\end{aligned} \tag{6.1.8}$$

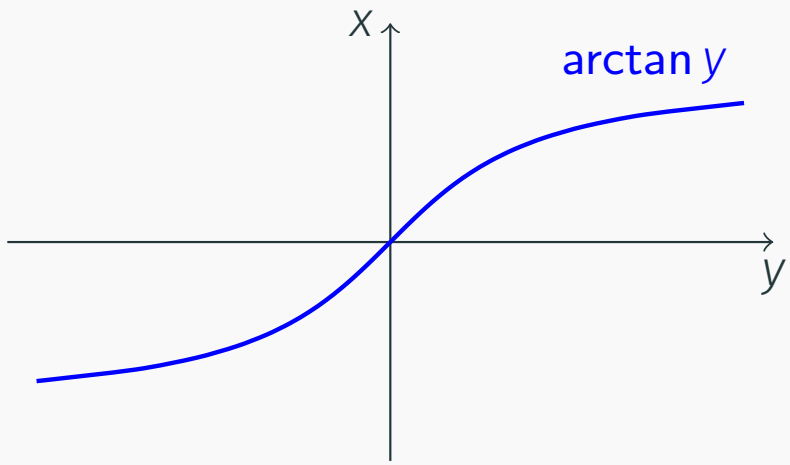
wobei wir $\cos x \geq 0$ für $x \in [-\pi/2, \pi/2]$ ausgenutzt haben. Analog gilt:

$$\arccos' y = -\frac{1}{\sqrt{1-y^2}} \tag{6.1.9}$$

ABLEITUNG DES ARKUSTANGENS

Der **Arkustangens** $\arctan : \mathbb{R} \to \mathbb{R}$ ist die Umkehrfunktion von $\tan$ eingeschränkt auf $[-\frac{\pi}{2}, \frac{\pi}{2}]$.





Wegen der zweiten Identität in (6.1.5) ist die Ableitung durch

$$\arctan'(y) = \frac{1}{\tan'(\arctan y)} = \frac{1}{1+\tan^2(\arctan y)} = \frac{1}{1+y^2}$$

für alle $y \in \mathbb{R}$ gegeben.

Section 6.2
Hauptsätze

Definition 6.2.1

Sei I ein Intervall, $f : I \to \mathbb{R}$ und $x_0 \in I$. Dann hat f in x_0 ein:

i). **lokales Maximum**, falls es ein $\delta > 0$ gibt mit, so dass für alle $x \in I$ mit $|x - x_0| < \delta$ gilt

$$f(x) \leq f(x_0)\,.$$

Gilt diese Ungleichung sogar für alle $x \in I$, so hat f in x_0 ein **globales Maximum**.

ii). **lokales Minimum**, falls $-f$ ein globales Maximum in x_0 hat, d.h. es gibt ein $\delta > 0$, so dass für alle $x \in I$ mit $|x - x_0| < \delta$ gilt

$$f(x) \geq f(x_0)\,.$$

Gilt diese Ungleichung sogar für alle $x \in I$, so hat f in x_0 ein **globales Minimum**

In beiden Fällen sprechen wir von **lokalen** bzw. **globalen Extrema**.

Offensichtlich ist jedes globale Maximum auch ein lokales Maximum, und die Umkehrung ist im Allgemeinen falsch.

Der folgende Satz liefert ein notwendiges Kriterium zum Finden eines lokalen Extremums.

Theorem 6.2.2

Sei $f : (a, b) \to \mathbb{R}$ differenzierbar. Hat f in $x_0 \in (a, b)$ ein lokales Extremum, dann gilt $f'(x_0) = 0$.

Es reicht den Fall eines Maximums in x_0 zu betrachten. Sei nun $\delta > 0$, so dass $(x_0 - \delta, x_0 + \delta) \subset (a, b)$ und

$$f(x) \leq f(x_0)$$

für alle $x \in (x_0 - \delta, x_0 + \delta)$ gilt. Für $x < x_0$ gilt dann $\frac{f(x)-f(x_0)}{x-x_0} \geq 0$, und daher folgt

$$f'(x_0) = \lim_{x \to {x_0}^-} \frac{f(x) - f(x_0)}{x - x_0} \geq 0.$$

Analog finden wir

$$f'(x_0) = \lim_{x \to {x_0}^+} \frac{f(x) - f(x_0)}{x - x_0} \leq 0$$

und zusammen ergibt dies $f'(x_0) = 0$.

BEISPIEL

Dieses notwendige Kriterium für Extrema erlaubt es, Kandidaten für Extrema als Nullstellen der Ableitung zu bestimmen. Wir bezeichnen Nullstellen der Ableitung von f deswegen als kritische Punkte.

Um ein Beispiel zu betrachten, sei $f : \mathbb{R} \to \mathbb{R}$ die durch Wir suchen Extrema der Funktion

$$f(x) = \frac{x}{1+x^2}\,, \qquad\qquad x \in \mathbb{R}$$

definierte Funktion. Etwas Rechnen zeigt, dass die Ableitung dann

$$f'(x) = \frac{1-x^2}{(1+x^2)^2}$$

ist. Potentielle Extrema liegen daher nur in $x = \pm 1$ vor, alle anderen Punkte sind ausgeschlossen. Aus $(1-x)^2 \geq 0$ folgt nun $1 + x^2 \geq 2x$ und damit

$$f(x) = \frac{x}{1+x^2} \leq \frac{1}{2} = f(1)\,,$$

d.h. f hat ein globales Maximum in 1. Analog kann man zeigen, dass f in -1 ein globales Minimum hat.

Der folgende Satz ist als Mittelwertsatz bekannt und spielt eine wichtige Rolle in unseren weiteren Überlegungen.

Theorem 6.2.3

Sei $f : [a, b] \to \mathbb{R}$ stetig und differenzierbar. Dann existiert ein $x_0 \in (a, b)$ mit

$$f'(x_0) = \frac{f(b) - f(a)}{b - a}.$$

Wir betrachten die Funktion $F : [a, b] \to \mathbb{R}$, die durch

$$F(x) = f(x) - \frac{f(b) - f(a)}{b - a} \cdot (x - a), \qquad x \in [a, b]$$

definiert ist. Dann ist F stetig und differenzierbar mit

$$F'(x) = f'(x) - \frac{f(b) - f(a)}{b - a} \tag{6.2.1}$$

und es gilt $F(a) = f(a) = F(b)$.

Falls nun $F(x) = f(a)$ für alle $x \in (a, b)$ gilt, so hat z.B. F ein lokales Maximum in $x_0 := (a + b)/2$ und nach Satz 6.2.2 gilt $F'(x_0) = 0$. Mit (6.2.1) folgt die Behauptung.

Falls es ein $x \in (a, b)$ mit $F(x) \neq f(a)$ gibt, betrachten wir zunächst den Fall $F(x) > f(a)$. Da F nach Satz 5.2.8 sein globales Maximum in einem $x_0 \in [a, b]$ annimmt, gilt dann $F(x_0) \geq F(x) > f(a) = F(a) = F(b)$, und damit $x_0 \in (a, b)$. Mit Satz 6.2.2 folgt $F'(x_0) = 0$ und (6.2.1) liefert die Behauptung.

Der Fall $F(x) < f(a)$ ist analog über globale Minima zu zeigen.

Der folgende Spezialfall des Mittelwertsatzes mit $f(a) = f(b)$ wird als Satz von Rolle bezeichnet.

Korollar 6.2.4

Sei $f : [a, b] \to \mathbb{R}$ stetig und differenzierbar. Gilt nun $f(a) = f(b)$, so existiert ein $x_0 \in (a, b)$ mit $f'(x_0) = 0$.

Eine weitere wichtige Folgerung betrifft das Monotonieverhalten differenzierbarer Funktionen.

Korollar 6.2.5

Sei $f : (a, b) \to \mathbb{R}$ differenzierbar. Dann gelten die folgenden Aussagen:

i). Es gilt $f'(x) \geq 0$ für alle $x \in (a, b)$ genau dann, wenn f monoton wachsend ist.

ii). Falls $f'(x) > 0$ für alle $x \in (a, b)$, dann ist f streng monoton wachsend.

iii). Es gilt $f'(x) = 0$ für alle $x \in (a, b)$ genau dann, wenn f konstant ist.

Analog lassen sich durch Betrachtung von $-f$ auch fallende Funktionen durch negative Ableitungen beschreiben. Die Rückrichtung in *ii)* ist falsch, wie z.B. das Beispiel $x \mapsto x^3$ in $x = 0$ zeigt.

"$\Rightarrow$"). Wir wählen $x_1, x_2 \in (a, b)$ mit $x_1 < x_2$ und betrachten die eingeschränkte Funktion $f_{|[x_1,x_2]}$. Diese ist differenzierbar und ihre Ableitung gleich $f'_{|(x_1,x_2)}$. Insbesondere ist die Einschränkung nach Korollar 6.1.5 auch stetig. Der Mittelwertsatz 6.2.3 liefert daher ein $\xi \in (x_1, x_2)$ mit

$$f(x_2) - f(x_1) = f'(\xi) \cdot (x_2 - x_1)\,.$$

Wegen $x_2 - x_1 > 0$ können dann die einzelnen Aussagen "abgelesen" werden.

"$\Leftarrow$"). Für konstante Funktionen wissen wir bereits, dass ihre Ableitung gleich 0 ist. Ist wiederum f monoton wachsend, so gilt

$$\frac{f(x) - f(x_0)}{x - x_0} \geq 0$$

für alle $x, x_0 \in (a, b)$ mit $x \neq x_0$. Daraus folgt $f'(x_0) \geq 0$ für alle $x_0 \in (a, b)$.

BEISPIEL

Sei $f : \mathbb{R} \to \mathbb{R}$ die durch

$$f(x) = \frac{x}{\sqrt{1+x^2}}, \qquad x \in \mathbb{R}$$

definierte Funktion. Dann gilt

$$f'(x) = \frac{\sqrt{1+x^2} - x\frac{2x}{2\cdot\sqrt{1+x^2}}}{1+x^2} = \frac{1}{(1+x^2)^{3/2}} > 0$$

und damit ist f streng wachsend.

Mit diesen Einsichten können wir nun die folgende, hinreichende Bedingung für ein lokales Maxima beweisen.

Theorem 6.2.6

Sei $f : (a, b) \to \mathbb{R}$ differenzierbar und es gelte $f'(x_0) = 0$ für ein $x_0 \in (a, b)$. Gibt es nun ein $\delta > 0$ mit

$$f'(x) > 0\,, \qquad x \in (x_0 - \delta, x_0)$$

und

$$f'(x) < 0\,, \qquad x \in (x_0, x_0 + \delta),$$

so besitzt f in x_0 ein lokales Maximum.

Beweis.

Nach Korollar 6.2.5 ist f auf $(x_0 - \delta, x_0)$ streng wachsend, d.h. $f(x) < f(x_0)$ für alle $x \in (x_0 - \delta, x_0)$.
Nach Korollar 6.2.5 ist f auf $(x_0, x_0 + \delta)$ streng fallend, d.h. $f(x_0) > f(x)$ für alle $x \in (x_0, x_0 + \delta)$.
Insgesamt hat f daher ein lokales Maximum in x_0. □

BEISPIEL: DIFFERENTIALGLEICHUNG

Wir hatten schon gesehen, dass die Exponentialfunktion die Gleichung $\exp' = \exp$ erfüllt. Der folgende Satz zeigt, dass die Exponentialfunktion die einzige Funktion f mit $f' = f$ und $f(0) = 1$ ist.

Theorem 6.2.7

Sei $f : \mathbb{R} \to \mathbb{R}$ eine differenzierbare Funktion, für die es ein $c \in \mathbb{R}$ gibt mit

$$f'(x) = cf(x)\,, \qquad x \in \mathbb{R}.$$

Dann gilt $f(x) = f(0) \cdot \exp(cx)$ für alle $x \in \mathbb{R}$.

BEWEIS

Wir betrachten die Funktion $g(x) := f(x) \cdot \exp(-cx)$ für $x \in \mathbb{R}$. Es gilt dann mit der Produktregel

$$g'(x) = f'(x) \cdot \exp(-cx) - cf(x) \cdot \exp(-cx) = \big(f'(x) - cf(x)\big) \cdot \exp(-cx) = 0$$

für alle $x \in \mathbb{R}$. Damit ist g nach Korollar 6.2.5 konstant. Wegen $g(0) = f(0)$ folgt $g(x) = f(0)$ für alle $x \in \mathbb{R}$ und ein einfaches Umstellen dieser Identität liefert die Behauptung.

STETIGE DIFFERENZIERBARKEIT

Die folgende hinreichende Bedingung für lokale Extrema ist schon aus der Schule bekannt. Für ihre Formulierung sagen wir, dass eine Funktion $f : (a, b) \to \mathbb{R}$ **stetig differenzierbar** ist, falls f differenzierbar ist und f' stetig ist. Analog ist f **zweimal stetig differenzierbar**, falls f stetig differenzierbar ist und f' auch stetig differenzierbar ist.

Schließlich sagen wir, dass eine Funktion $f : [a, b] \to \mathbb{R}$ stetig differenzierbar ist, wenn f stetig ist, $f_{|(a,b)}$ stetig differenzierbar ist und es eine stetige Funktion $g : [a, b] \to \mathbb{R}$ gibt mit $g(x) = f'(x)$ für alle $x \in (a, b)$. Insbesondere ist also die Ableitung f' beschränkt nach Satz 5.2.8. Analog ist zweimal stetig differenzierbar definiert.

Theorem 6.2.8

Sei $f : (a, b) \to \mathbb{R}$ zweimal stetig differenzierbar. Ferner sei $x_0 \in (a, b)$ mit $f'(x_0) = 0$. Dann gilt:

i). Im Fall $f''(x_0) < 0$ besitzt f in x_0 ein lokales Maximum.

ii). Im Fall $f''(x_0) > 0$ besitzt f in x_0 ein lokales Minimum.

Der Vorteil des Satzes 6.2.8 gegenüber Satz 6.2.6 ist, dass er nur mit dem Vorzeichen der zweiten Ableitung im Punkt x_0 arbeitet. Dieses ist häufig einfach zu bestimmen. Jedoch benötigen wir zweifache stetige Differenzierbarkeit der Funktion f um ihn anzuwenden und er liefert auch keine Aussage, falls $f''(x_0) = 0$ gilt.

Wir zeigen nur *ii)*, da *i)* durch die Betrachtung von $-f$ folgt.

Da f' stetig ist mit $f'(x_0) > 0$, existiert für $\varepsilon := f'(x_0)/2 > 0$ ein $\delta > 0$ mit $|f'(x_0) - f'(x)| < \varepsilon$ für alle $x \in (x_0 - \delta, x_0 + \delta)$. Für solche x folgt

$$f'(x_0) < f'(x) + \varepsilon = f'(x) + f'(x_0)/2$$

und damit $f'(x) > 0$. Damit ist f' nach Korollar 6.2.5 streng wachsend auf dem Intervall $(x_0 - \delta, x_0 + \delta)$.

Da $f'(x_0) = 0$ vorausgesetzt ist, folgt also $f'(x) < 0$ auf $(x_0 - \delta, x_0)$ und $f'(x) > 0$ auf $(x_0, x_0 + \delta)$ und mit Satz 6.2.6 angewendet auf $-f$ folgt die Behauptung.

Zum Schluss geben wir noch eine Verallgemeinerung des Mittelwertsatzes an, die für einige spätere Beweise wichtig sein wird.

Theorem 6.2.9

Seien $f, g : [a, b] \to \mathbb{R}$ stetig und differenzierbar. Sei weiter $g'(x) \neq 0$ auf (a, b). Dann existiert ein $\xi \in (a, b)$ mit

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$

Wäre $g(b) = g(a)$, so gäbe es nach dem Satz von Rolle 6.2.4 ein $g'(x) = 0$. Da wir dies ausschließen, folgt $g(b) \neq g(a)$. Damit ist die Funktion

$$F(x) = f(x) - \frac{f(b) - f(a)}{g(b) - g(a)} g(x)$$

auf ganz $[a, b]$ definiert und auf (a, b) differenzierbar. Wegen

$$\begin{aligned} F(a) - F(b) &= f(a) - \frac{f(b) - f(a)}{g(b) - g(a)} g(a) - f(b) + \frac{f(b) - f(a)}{g(b) - g(a)} g(b) \\ &= f(a) - f(b) - \frac{f(b) - f(a)}{g(b) - g(a)} \cdot (g(a) - g(b)) \\ &= 0\,, \end{aligned}$$

d.h. $F(a) = F(b)$, existiert nach dem Satz von Rolle 6.2.4 ein $\xi \in (a, b)$ mit $F'(\xi) = 0$ und damit folgt

$$f'(\xi) = \frac{f(b) - f(a)}{g(b) - g(a)} g'(\xi).$$

Ein einfaches Umformen liefert dann die Behauptung.

Definition 6.2.10

Sei $I \subset \mathbb{R}$ ein Intervall. Dann heißt $f : I \to \mathbb{R}$ **konvex**, falls für alle $x_0, x_1 \in I$ und alle $t \in [0, 1]$ gilt

$$f\big((1-t)x_0 + tx_1\big) \leq (1-t)f(x_0) + tf(x_1)\,. \tag{6.2.2}$$

Die Funktion heißt **strikt konvex**, falls die obige Ungleichung strikt ist. Schließlich heißt f **(strikt) konkav**, falls $-f$ (strikt) konvex ist.

Die Definition lässt sich von Intervallen auf Vektorräume verallgemeinern. Offensichtlich sind affin linear Funktionen sowohl konvex als auch konkav, aber in beiden Fällen ist dies nicht strikt. Schließlich ist die Ungleichung (6.2.2) sowohl für $x_0 = x_1$ als auch für $t \in \{0, 1\}$ immer erfüllt.

Konvexe Funktionen spielen in vielen Bereichen eine wichtige Rolle, u.a. weil es für solche Funktionen effektive Algorithmen zum Finden von Minima gibt.

Der folgende Satz liefert ein einfaches hinreichendes Kriterium für die Konvexität.

Theorem 6.2.11

Sei $f : (a, b) \to \mathbb{R}$ differenzierbar. Dann ist f genau dann konvex, wenn f' monoton wachsend ist.

Man beachte, dass sich Satz 6.2.11 mit Korollar 6.2.5 angewendet auf f' verbinden lässt. Dies zeigt, dass eine zweimal differenzierbare Funktion $f : (a, b) \to \mathbb{R}$ genau dann konvex ist, falls $f''(x) \geq 0$ für alle $x \in (a, b)$ gilt.

BEWEIS 1/2

Seien f konvex, $x_0, x_1 \in (a, b)$ mit $x_0 < x_1$ und $t \in (0, 1)$. Dann gilt

$$f\big((1-t)x_0 + tx_1\big) \leq (1-t)f(x_0) + tf(x_1)$$

und damit auch

$$\frac{f\big(x_0 + t(x_1 - x_0)\big) - f(x_0)}{t \cdot (x_1 - x_0)} \cdot (x_1 - x_0) \leq f(x_1) - f(x_0)\,.$$

Für $t \to 0$ folgt

$$f'(x_0) \leq \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0) - f(x_1)}{x_0 - x_1}$$

und für $x_0 \to x_1$ damit auch $f'(x_0) \leq f'(x_1)$.

BEWEIS 2/2

Sei nun f' monoton wachsend. Für $x_0, x_1 \in (a, b)$ mit $x_0 < x_1$ und $t \in (0, 1)$ schreiben wir $x_t := (1-t)x_0 + tx_1$. Es gilt dann $x_t \in (x_0, x_1)$ und der Mittelwertsatz 6.2.3 liefert damit $\xi_0 \in (x_0, x_t)$ und $\xi_1 \in (x_t, x_1)$ mit

$$\frac{f(x_t) - f(x_0)}{x_t - x_0} = f'(\xi_0) \leq f'(\xi_1) = \frac{f(x_1) - f(x_t)}{x_1 - x_t}\,.$$

Wegen $x_t - x_0 = t(x_1 - x_0)$ und $x_1 - x_t = (1-t)(x_1 - x_0)$ folgt

$$\frac{f(x_t) - f(x_0)}{t} \leq \frac{f(x_1) - f(x_t)}{1-t}$$

und damit $(1-t)(f(x_t) - f(x_0)) \leq t(f(x_1) - f(x_t))$. Dies wiederum impliziert

$$f(x_t) \leq (1-t)f(x_0) + tf(x_1)\,,$$

was der Konvexitäts-Ungleichung (6.2.2) entspricht.

Die Regel von de L'Hospital

Die Regel von de L'Hospital liefert eine elegante Möglichkeit zur Berechnung von Funktionsgrenzwerten, die in Form unbestimmter Ausdrücke

$$\frac{0}{0}, \qquad \frac{\infty}{\infty}, \qquad 0 \cdot \infty, \qquad \infty - \infty, \qquad \infty^0, \qquad 0^0, \qquad 1^\infty$$

erscheinen. Die Regel von L'Hospital für Quotienten ergibt sich als Anwendung des verallgemeinerten Mittelwertsatzes. Wie bei den elementaren Grenzwertsätzen liefert sie eine Methode zur Berechnung von Grenzwerten, die gleichzeitig deren Existenz mit beweist. Es ist jedoch Vorsicht geboten, die Voraussetzungen der Regel sind in jedem Schritt zu prüfen.

Theorem 6.2.12

Seien $f, g : (a, b) \to \mathbb{R}$ differenzierbar und es gelte

$$\lim_{x \to b^-} f(x) = \lim_{x \to b^-} g(x) = 0 \tag{6.2.3}$$

und $g'(x) \neq 0$ für $x < b$. Existiert dann der Grenzwert $\lim_{x \to b^-} \frac{f'(x)}{g'(x)}$, so gilt

$$\lim_{x \to b^-} \frac{f(x)}{g(x)} = \lim_{x \to b^-} \frac{f'(x)}{g'(x)} .$$

Eine analoge Aussage gilt für rechtsseitige Grenzwerte. Ferner gilt die Aussage auch, falls $\lim_{x \to b^-} \frac{f'(x)}{g'(x)}$ ein uneigentlicher Grenzwert ist.

Wegen (6.2.3) können wir f und g auf $(a, b]$ durch $f(b) := g(b) := 0$ stetig fortsetzen. Im folgenden nehmen wir an, dass wir dies getan haben.

Sei nun $(x_n)_{n\in\mathbb{N}} \subset (a, b)$ mit $x_n \to b$. Wenden wir den verallgemeinerten Mittelwertsatz 6.2.9 auf die Einschränkungen von f und g auf $[x_n, b]$ an, so existiert ein $\xi_n \in (x_n, b)$ mit

$$\frac{f(x)}{g(x)} = \frac{f(x_n) - f(b)}{g(x_n) - g(b)} = \frac{f'(\xi_n)}{g'(\xi_n)}.$$

Ferner impliziert $x_n \to b$ die Konvergenz $\xi_n \to b$ und damit folgt

$$\lim_{x\to b-} \frac{f(x)}{g(x)} = \lim_{\xi\to b-} \frac{f'(\xi)}{g'(\xi)}.$$

Da die Folge $(x_n)_{n\in\mathbb{N}}$ beliebig war, folgt die Behauptung.

Theorem 6.2.13

Seien $f, g : (a, b) \to \mathbb{R}$ differenzierbar und es gelte

$$\lim_{x \to b^-} f(x) = \lim_{x \to b^-} g(x) = \infty$$

und $g'(x) \neq 0$ für $x < b$. Existiert dann der Grenzwert $\lim_{x \to b^-} \frac{f'(x)}{g'(x)}$, so gilt

$$\lim_{x \to b^-} \frac{f(x)}{g(x)} = \lim_{x \to b^-} \frac{f'(x)}{g'(x)} .$$

Eine analoge Aussage gilt für rechtsseitige Grenzwerte. Ferner gilt die Aussage auch, falls $\lim_{x \to b^-} \frac{f'(x)}{g'(x)}$ ein uneigentlicher Grenzwert ist oder $b = \infty$ ist.

Der Beweis ist etwas aufwendiger als der von Satz 6.2.12 und wird aus Zeitgründen daher übersprungen.

Betrachten wir schließlich ein paar Beispiele. Mit Satz 6.2.13 finden wir

$$\lim_{x\to 0^+} x \ln x = \lim_{x\to 0^+} \frac{\ln x}{\frac{1}{x}} = \lim_{x\to 0^+} \frac{\frac{1}{x}}{\frac{-1}{x^2}} = \lim_{x\to 0^+} (-x) = 0. \tag{6.2.4}$$

Dabei wurde der gegebene unbestimmte Ausdruck der Form $0 \cdot \infty$ in einen Quotienten der Form $\frac{\infty}{\infty}$ umgeschrieben.

Mit Satz 6.2.12 finden wir z.B.:

$$\begin{aligned}\lim_{x\to\infty}\left(x\ln\frac{x-1}{x+1}\right) = \lim_{x\to\infty}\frac{\ln(x+1)-\ln(x-1)}{\frac{1}{x}} &= \lim_{x\to\infty}\frac{\frac{1}{x+1}-\frac{1}{x-1}}{-\frac{1}{x^2}}\\ &= \lim_{x\to\infty}\left(\frac{x^2}{x-1}-\frac{x^2}{x+1}\right)\\ &= \lim_{x\to\infty}\frac{2x^2}{x^2-1}\\ &= 2.\end{aligned}$$

Hier wurde der gegebene unbestimmte Ausdruck der Form $0\cdot\infty$ in einen Quotienten der Form $\frac{0}{0}$ umgeschrieben.

Manchmal führt die Anwendung einer der Regeln leider nicht zum Erfolg. Der Versuch Satz 6.2.12 wiederholt zu benutzen führt in diesem Beispiel zu

$$\begin{aligned}
&\lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}}{x} \\
&= \lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}\frac{1}{x^2}}{1} = \lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}}{x^2} \\
&= \lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}\frac{1}{x^2}}{2x} = \lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}}{2x^3} \\
&= \lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}\frac{1}{x^2}}{6x^2} = \lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}}{6x^4} \\
&= \cdots .
\end{aligned}$$

Der Ausdruck wird also mit jeder Anwendung von Satz 6.2.12 komplizierter. Wendet man stattdessen Satz 6.2.13, so ergibt sich

$$\lim_{x\to 0^+} \frac{\mathrm{e}^{-1/x}}{x} = \lim_{x\to 0^+} \frac{\frac{1}{x}}{\mathrm{e}^{1/x}} = \lim_{x\to 0^+} \frac{-\frac{1}{x^2}}{-\mathrm{e}^{1/x}\frac{1}{x^2}} = \lim_{x\to 0^+} \frac{1}{\mathrm{e}^{1/x}} = 0.$$

ANWENDUNG: KURVENDISKUSSION

1. Definitions- und Wertebereich
2. Symmetrie-Eigenschaften (gerade oder ungerade Funktionen, Spiegelsymmetrien zu anderen Punkte oder senkrechten Linien, Periodizität)
3. Stetigkeit und Stetigkeitsintervalle, eventuelle Definitionslücken, Polstellen
4. Nullstellen, Schnittpunkt mit der y-Achse
5. Extremalstellen und Monotonieintervalle
6. Konvexitätsintervalle. Ist die Funktion zweifach differenzierbar, so sind die Konvexitätsintervalle gerade die Monotonieintervalle der Ableitung.
7. Wendepunkte (und Wendetangenten), dabei sind Wendepunkte gerade die Punkte, in welchen sich das Konvexitätsverhalten ändert, die Funktion also von konvex zu konkav wechselt
8. asymptotisches Verhalten an den Rändern des Definitionsbereiches, damit meint man Grenzwerte der Funktion oder einfachere Funktionen, welche den Verlauf der gegebenen Funktion asymptotisch beschreiben
9. eine Skizze der Funktion und ggf. ihrer Ableitungen.

BEISPIEL

Wir geben ein Beispiel und untersuchen die Funktion

$$f(x) = \frac{x^3 - 2x^2 - x + 2}{x + 2}$$

definiert auf der Menge $D_f = \mathbb{R} \setminus \{-2\}$.

Sie ist auf $\mathbb{R} \setminus \{-2\}$ stetig und besitzt in $x = -2$ eine Polstelle erster Ordnung, da

$$\lim_{x \to -2} (x + 2) f(x) = \lim_{x \to -2} (x^3 - 2x^2 - x + 2) = -12 \neq 0$$

gilt.

Wegen $f(0) = 1$ schneidet der Graph der Funktion die y-Achse im Punkt $(0, 1)$ und die Funktion f erfüllt

$$\begin{aligned} f(x) = 0 \quad \Leftrightarrow \quad x^3 - 2x^2 - x + 2 &= (x - 1)(x + 1)(x - 2) \\ &= 0 \end{aligned}$$

und besitzt damit die Nullstellen $x = \pm 1$ und $x = 2$.

BEISPIEL

Das asymptotische Verhalten der Funktion $f(x)$ für große Werte von x ergibt sich durch Polynomdivision

$$f(x) = \frac{x^3 - 2x^2 - x + 2}{x + 2} = x^2 - 4x + 7 - \frac{12}{x + 2},$$

für $x \to \pm\infty$ nähert sich der Graph der Funktion der Parabel $g(x) = x^2 - 4x + 7$. Für $x \to -2$ verhält sich die Funktion wie $h(x) = 19 - \frac{12}{x+2}$.

Zum Bestimmen kritischer Punkte, sowie der Monotonie- und Konvexitätsintervalle, leiten wir die Funktion f ab. Das ergibt

$$f'(x) = \frac{(3x^2 - 4x - 1)(x + 2) - (x^3 - 2x^2 - x + 2)}{(x + 2)^2} = \frac{2x^3 + 4x^2 - 8x - 4}{(x + 2)^2}$$

und

$$\begin{aligned} f''(x) &= \frac{(6x^2 + 8x - 8)(x + 2)^2 - 2(2x^3 + 4x^2 - 8x - 4)(x + 2)}{(x + 2)^4} \\ &= \frac{2x^3 + 12x^2 + 24x - 8}{(x + 2)^3}. \end{aligned}$$

BEISPIEL

Potentielle Extremalstellen ergeben sich aus den Nullstellen der ersten Ableitung. Es gilt

$$f'(x) = 0 \quad \Leftrightarrow \quad 2x^3 + 4x^2 - 8x - 4 = 0$$

Dies liefert drei Nullstellen, da $x^3 + 2x^2 - 4x + 2$ für $x \to \pm\infty$ gegen $\pm\infty$ strebt und an der Stelle $x = -1$ positiv und in $x = 1$ negativ ist. Bestimmt man diese näherungsweise, so ergibt sich

$$x_1 \approx -3.1\,,$$

$$x_2 \approx -0.43\,,$$

$$x_3 \approx 1.51\,.$$

Aufgrund der Vorzeichen ist die Funktion auf $(-\infty, x_1)$ streng monoton fallend, auf $(x_1, -2)$ streng monoton steigend, auf $(-2, x_2)$ streng monoton steigend, auf (x_2, x_3) streng monoton fallend und auf (x_3, ∞) wiederum streng monoton steigend. Bei x_1 und x_3 liegen also lokale Minima vor, bei x_2 ein lokales Maximum.

Alternativ kann man natürlich auch einfach $f''(x_i)$ betrachten, um diese Extrema zu bestimmen.

Die zweite Ableitung besitzt nur eine reelle Nullstelle

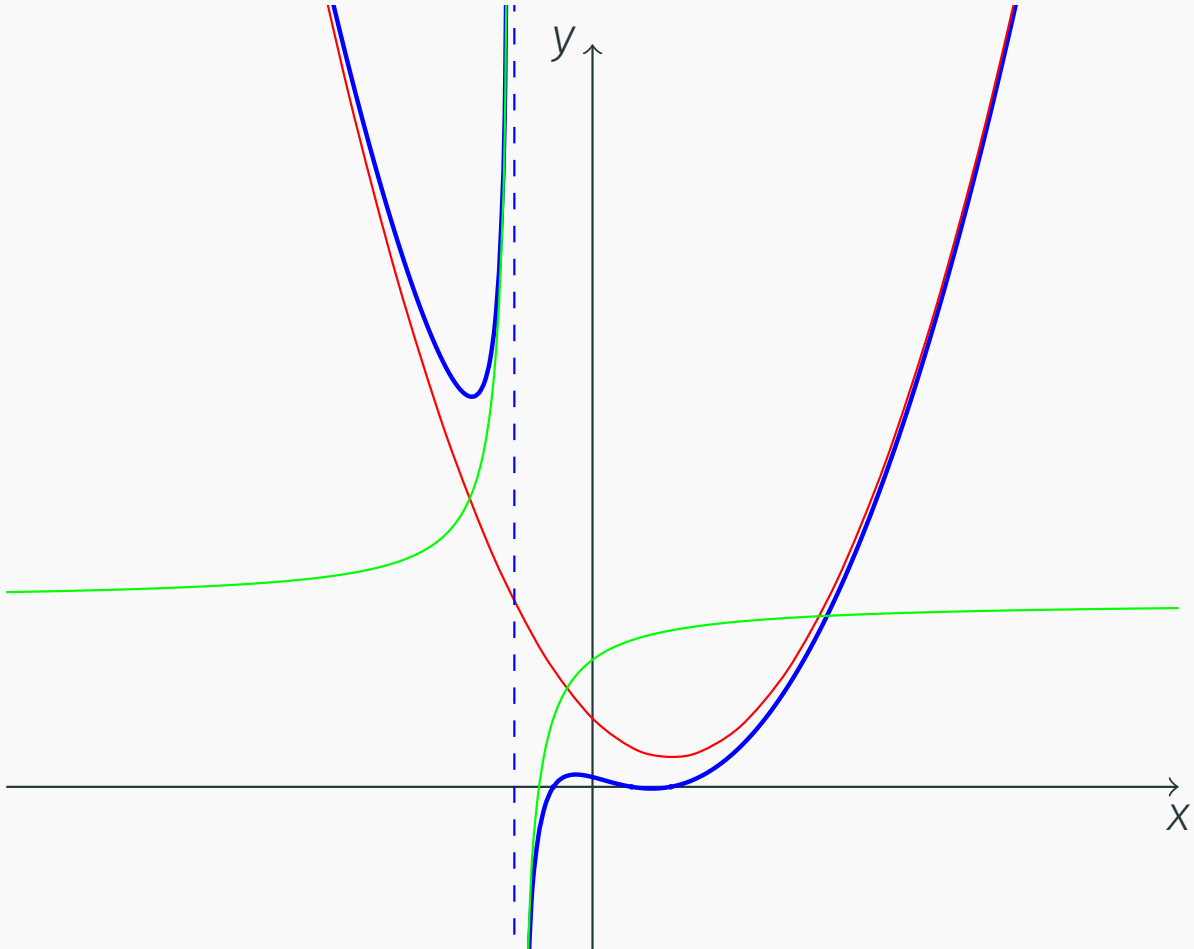
$$f''(x) = 0 \quad \Leftrightarrow \quad x^3 + 6x^2 + 12x - 8 = 0,$$

da ihr Zähler wegen

$$\left(x^3 + 6x^2 + 12x - 8\right)' = 3x^2 + 12x + 12 = 3(x+2)^2 > 0$$

für alle $x \in \mathbb{R} \setminus \{2\}$ streng monoton wachsend ist. Diese eine Nullstelle befindet näherungsweise bei $x_* \approx 0,3$ und es gilt für den Zähler von f'', dass $x^3 + 6x^2 + 12x - 8 < 0$ genau dann, wenn $x < x_*$. Der Nenner $(x+2)^3$ von f'' ändert wiederum sein Vorzeichen von negativ auf positiv bei $x = -2$. Damit ist $f'' \geq 0$ auf $(-\infty, -2) \cup (x_*, \infty)$ und $f'' \leq 0$ auf $(-2, x_*)$. Insgesamt ist daher die Funktion f auf $(-\infty, -2)$ und auf (x_*, ∞) konvex und auf $(-2, x_*)$ konkav.

Skizze: Dargestellt sind in blau die Funktion $f(x)$, in rot die für große $|x|$ asymptotisch äquivalente Funktion $g(x)$ und in grün die den Pol beschreibende Funktion $h(x)$. Die beiden Achsen haben verschiedene Maßstäbe, um das Verhalten der Funktion über einem größerem Bereich darstellen zu können.



Chapter 7: Integration

Section 7.1
Das Riemann-Integral

UNTER- UND OBER-SUMMEN

Wir wollen für eine stetige Funktion $f : [a, b] \to \mathbb{R}$ den Flächeninhalt zwischen dem Graphen und der x-Achse bestimmen. Dabei sollen Flächen über der Achse positiv und Flächen unter der Achse als negativ gezählt werden.

Die Idee ist dabei die folgende: Wir zerlegen das Intervall $[a, b]$ mit Hilfe von Zwischenstellen

$$\mathcal{Z} : \qquad a = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = b \tag{7.1.1}$$

und approximieren den Flächeninhalt von unten und von oben durch entsprechende Summen von Rechtecksflächen.

Im folgenden bezeichnen wir eine solche Zerlegung von $[a, b]$ mit dem Buchstaben $\mathcal{Z}$ und nennen

$$\delta(\mathcal{Z}) = \max_{k=1,\ldots,N} |x_k - x_{k-1}|$$

die Feinheit der Zerlegung $\mathcal{Z}$.

Betrachten wir beispielsweise die durch $x_i := a + i \cdot \frac{b-a}{N}$ für $i = 0, \ldots, N$ gegebene äquidistante Zerlegung $\mathcal{Z}$, so gilt $\delta(\mathcal{Z}) = (b - a)/N$.

Haben wir zwei Zerlegungen $\mathcal{Z}_1$ und $\mathcal{Z}_2$, die durch $x_0, \ldots, x_N$ und $y_0, \ldots, y_M$ beschrieben sind, so schreiben wir

$$\mathcal{Z}_1 \subset \mathcal{Z}_2 \qquad :\Longleftrightarrow \qquad \{x_0, \ldots, x_N\} \subset \{y_0, \ldots, y_M\}\,.$$

In diesem Fall heißt $\mathcal{Z}_2$ **Verfeinerung** von $\mathcal{Z}_1$ und es gilt $\delta(\mathcal{Z}_1) \geq \delta(\mathcal{Z}_2)$.

Eine Möglichkeit, eine Verfeinerung zu konstruieren ist die **Vereinigung $\mathcal{Z}_1 \cup \mathcal{Z}_2$ der Zerlegungen** $\mathcal{Z}_1$ und $\mathcal{Z}_2$, die durch die Vereinigung

$$\{z_0, \ldots, z_K\} := \{x_0, \ldots, x_N\} \cup \{y_0, \ldots, y_M\}$$

der Zwischenstellen entsteht. Hierbei erfüllen die Zwischenstellen der Vereinigung wieder $a = z_0 < z_1 < \cdots < z_{K-1} < z_K = b$, was durch Sortieren und eliminieren doppelter Elemente immer erreichbar ist, vgl. Abbildung 16.

Offensichtlich gilt $\mathcal{Z}_1 \subset \mathcal{Z}_1 \cup \mathcal{Z}_2$ und $\mathcal{Z}_2 \subset \mathcal{Z}_1 \cup \mathcal{Z}_2$. Die Vereinigung zweier äquidistanter Zerlegungen ist im Allgemeinen keine äquidistante Zerlegung mehr, siehe wieder Abbildung 16.

Darboux'sche Unter- und Ober-Summen

Sei nun $f : [a, b] \to \mathbb{R}$ eine Funktion und $\mathcal{Z}$ eine Zerlegung mit den Bezeichnungen aus (7.1.1). Dann betrachten wir **Darboux'sche Obersumme**

$$\mathrm{I}^{\mathcal{Z}}(f) = \sum_{k=1}^{N} (x_k - x_{k-1}) \sup_{\xi \in [x_{k-1}, x_k]} f(\xi)$$

und die **Darboux'sche Untersumme**

$$\mathrm{I}_{\mathcal{Z}}(f) = \sum_{k=1}^{N} (x_k - x_{k-1}) \inf_{\xi \in [x_{k-1}, x_k]} f(\xi)\,.$$

Offensichtlich gilt, siehe auch Abbildung 17

$$\mathrm{I}_{\mathcal{Z}}(f) \leq \mathrm{I}^{\mathcal{Z}}(f)\,. \tag{7.1.2}$$

Wegen $\inf(-A) = -\sup A$, siehe (2.4.1), gilt ferner

$$-\mathrm{I}_{\mathcal{Z}}(f) = \mathrm{I}^{\mathcal{Z}}(-f)\,. \tag{7.1.3}$$

Haben wir zwei Zerlegungen $\mathcal{Z}_1 \subset \mathcal{Z}_2$ so gilt ferner

$$\mathrm{I}_{\mathcal{Z}_1}(f) \leq \mathrm{I}_{\mathcal{Z}_2}(f)\,, \tag{7.1.4}$$

$$\mathrm{I}^{\mathcal{Z}_1}(f) \geq \mathrm{I}^{\mathcal{Z}_2}(f)\,, \tag{7.1.5}$$

siehe wieder Abbildung 17

Darboux'sche Unter- und Ober-Summen

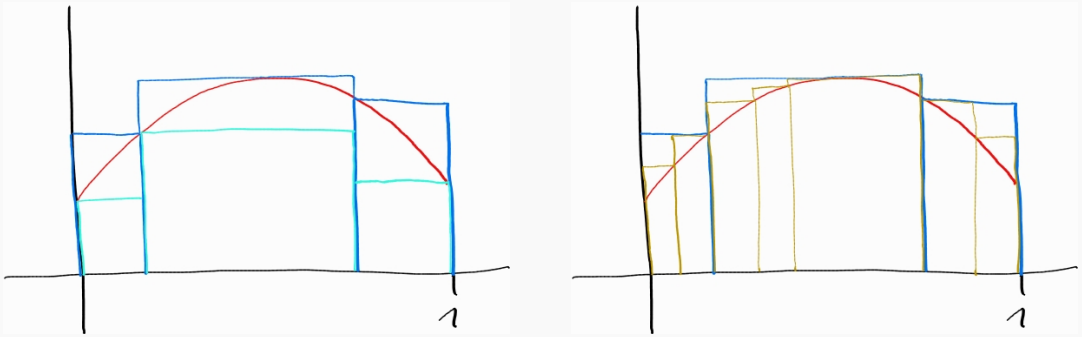


Abbildung: Links: Darboux'sche Unter- und Ober-Summen für eine feste Zerlegung. Die Untersumme entspricht der Fläche der drei türkisfarbenen Rechtecke während die Obersumme der Fläche der drei blauen Rechtecke entspricht. **Rechts:** Die gleiche Obersumme und die Obersumme einer verfeinerten Zerlegung. Die resultierende Fläche der 7 braunen Rechtecke ist kleiner als die der drei blauen Rechtecke.

VERGLEICH VON UNTER- UND OBER-INTEGRAL

Das folgende Lemma nutzt die letzten beiden Beobachtungen aus, um (7.1.2) zu verschärfen.

Lemma 7.1.1

Sei $f : [a, b] \to \mathbb{R}$. Dann gilt

$$\sup\{\mathrm{I}_{\mathcal{Z}'}(f) \mid \mathcal{Z}' \text{ Zerlegung von } [a, b]\} \leq \inf\{\mathrm{I}^{\mathcal{Z}''}(f) \mid \mathcal{Z}'' \text{ Zerlegung von } [a, b]\}\,.$$

BEWEIS

Seien $\mathcal{Z}'$ und $\mathcal{Z}''$ zwei Zerlegungen von $[a, b]$. Für $\mathcal{Z} := \mathcal{Z}' \cup \mathcal{Z}''$ folgt aus (7.1.4), (7.1.2) und (7.1.5) dann

$$\mathrm{I}_{\mathcal{Z}'}(f) \leq \mathrm{I}_{\mathcal{Z}}(f) \leq \mathrm{I}^{\mathcal{Z}}(f) \leq \mathrm{I}^{\mathcal{Z}''}(f)\,.$$

Nimmt man dann zunächst das Supremum über alle $\mathcal{Z}'$ und dann das Infimum über alle $\mathcal{Z}''$, so ergibt sich die Behauptung.

Mit diesen Betrachtungen können wir nun die Integrierbarkeit von Funktionen definieren.

Definition 7.1.2

Eine beschränkte Funktion $f : [a, b] \to \mathbb{R}$ heißt **Riemann-integrierbar**, falls für jede Folge $(\mathcal{Z}_n)_{n\in\mathbb{N}}$ von Zerlegungen $\mathcal{Z}_n$ mit $\delta(\mathcal{Z}_n) \to 0$ die Grenzwerte der zugeordneten Unter- und Obersummen existieren mit

$$\lim_{n\to\infty} \mathrm{I}_{\mathcal{Z}_n}(f) = \lim_{n\to\infty} \mathrm{I}^{\mathcal{Z}_n}(f)\,.$$

Im folgenden sprechen wir häufig auch nur kurz von **R-integrierbar**.

Ist $f : [a, b] \to \mathbb{R}$ eine konstante Funktion mit $f(x) = c$ für alle $x \in [a, b]$, so gilt für jede Zerlegung $\mathrm{I}_{\mathcal{Z}}(f) = \mathrm{I}^{\mathcal{Z}}(f) = c \cdot (b - a)$. Damit ist f Riemann-integrierbar.

Das folgende Lemma stellt die Riemann-Integrierbarkeit in Bezug zu unseren anfänglichen Beobachtungen.

Lemma 7.1.3

Sei $f : [a, b] \to \mathbb{R}$ Riemann-integrierbar und $(\mathcal{Z}_n)_{n\geq 1}$ eine Folge von Zerlegungen von $[a, b]$ mit $\delta(\mathcal{Z}_n) \to 0$. Dann gilt

$$\lim_{n\to\infty} \mathrm{I}_{\mathcal{Z}_n}(f) = \sup\{\mathrm{I}_{\mathcal{Z}}(f) \mid \mathcal{Z} \text{ Zerlegung von } [a, b]\},$$
$$\lim_{n\to\infty} \mathrm{I}^{\mathcal{Z}_n}(f) = \inf\{\mathrm{I}^{\mathcal{Z}}(f) \mid \mathcal{Z} \text{ Zerlegung von } [a, b]\}$$

und damit sind auch das Infimum und das Supremum gleich.

Wir schreiben

$$S := \sup\{\mathrm{I}_{\mathcal{Z}}(f) \mid \mathcal{Z} \text{ Zerlegung von } [a,b]\}\,,$$
$$I := \inf\{\mathrm{I}^{\mathcal{Z}}(f) \mid \mathcal{Z} \text{ Zerlegung von } [a,b]\}\,.$$

Für $n \geq 1$ gelten dann sofort die Ungleichungen

$$\mathrm{I}_{\mathcal{Z}_n}(f) \leq S\,,$$
$$I \leq \mathrm{I}^{\mathcal{Z}_n}(f)\,,$$

und für $n \to \infty$ erhalten wir damit

$$I \leq \lim_{n\to\infty} \mathrm{I}^{\mathcal{Z}_n}(f) = \lim_{n\to\infty} \mathrm{I}_{\mathcal{Z}_n}(f) \leq S\,.$$

Nach Lemma 7.1.1 gilt zudem $S \leq I$, was die Behauptung zeigt.

In der Definition von Riemann-Integrierbarkeit muss die Existenz zweier Grenzwerte gesichert werden. Dieses ist aber nicht notwendig, wenn wir für eine Funktion $f : [a, b] \to \mathbb{R}$ stattdessen

$$\lim_{n\to\infty} \left(\mathrm{I}^{\mathcal{Z}_n}(f) - \mathrm{I}_{\mathcal{Z}_n}(f) \right) = 0 \tag{7.1.6}$$

für alle Folgen $(\mathcal{Z}_n)_{n\geq 1}$ von Zerlegungen $\mathcal{Z}_n$ von $[a, b]$ mit $\delta(\mathcal{Z}_n) \to 0$ wissen. Um dies zu sehen, fixieren wir eine solche Folge $(\mathcal{Z}_n)_{n\in\mathbb{N}}$. Würde dann die Folge der Obersummen nicht konvergieren, gäbe es ein $\varepsilon > 0$ und eine Teilfolge $(\mathcal{Z}_{n_k})_{k\in\mathbb{N}}$ mit

$$\mathrm{I}^{\mathcal{Z}_{n_k}}(f) > I + \varepsilon\,, \qquad\qquad k \geq 1.$$

Dies impliziert mit (7.1.2) und Lemma 7.1.1

$$\begin{aligned} \left| \mathrm{I}^{\mathcal{Z}_{n_k}}(f) - \mathrm{I}_{\mathcal{Z}_{n_k}}(f) \right| = \mathrm{I}^{\mathcal{Z}_{n_k}}(f) - \mathrm{I}_{\mathcal{Z}_{n_k}}(f) &> I + \varepsilon - \mathrm{I}_{\mathcal{Z}_{n_k}}(f) \\ &\geq S + \varepsilon - \mathrm{I}_{\mathcal{Z}_{n_k}}(f) \\ &\geq \varepsilon\,, \end{aligned}$$

wobei wir im letzten Schritt die triviale Ungleichung $\mathrm{I}_{\mathcal{Z}_{n_k}}(f) \leq S$ benutzt haben. Dies widerspricht (7.1.6) und damit muss die Folge der Obersummen konvergieren. Dies wiederum ergibt wegen der Linearität des Limes

STETIGE FUNKTIONEN SIND R-INTEGRIERBAR

Der folgende Satz zeigt uns, dass die meisten Funktionen, für die wir uns interessieren, Riemann-integrierbar sind.

Theorem 7.1.4

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

BEWEIS 1/2

Nach Satz 5.2.8 ist f beschränkt und nach Satz 5.2.10 ist f gleichmäßig stetig. Für $\varepsilon > 0$ existiert daher ein $\delta > 0$, so dass für alle $x, x' \in [a, b]$ mit $|x - x'| \leq \delta$ auch $|f(x) - f(x')| \leq \varepsilon$ gilt.

Wir wählen uns jetzt eine Zerlegung $\mathcal{Z} : x_0, \ldots, x_N$ mit Feinheit $\delta(\mathcal{Z}) \leq \delta$. Da f auf $[x_{k-1}, x_k]$ stetig ist, gibt es nach Satz 5.2.8 dann $x_{k,*}, x_k^* \in [x_{k-1}, x_k]$ mit

$$f(x_{k,*}) = \inf_{x \in [x_{k-1}, x_k]} f(x)\,,$$

$$f(x_k^*) = \sup_{x \in [x_{k-1}, x_k]} f(x)\,.$$

Wegen $|x_{k,*} - x_k^*| \leq |x_k - x_{k-1}| \leq \delta$ folgt dann $|f(x_{k,*}) - f(x_k^*)| \leq \varepsilon$ und damit auch

$$\begin{aligned} \left|\mathrm{I}^{\mathcal{Z}}(f) - \mathrm{I}_{\mathcal{Z}}(f)\right| &= \sum_{k=1}^{N} |x_k - x_{k-1}| \cdot \left|f(x_k^*) - f(x_{k,*})\right| \\ &\leq \varepsilon \sum_{k=1}^{N} (x_k - x_{k-1}) \qquad (7.1.7) \\ &= \varepsilon(b - a)\,. \end{aligned}$$

BEWEIS 2/2

Haben wir nun eine Folge $(\mathcal{Z}_n)_{n\in\mathbb{N}}$ von Zerlegungen mit $\delta(\mathcal{Z}_n) \to 0$, so gibt es ein $n_0 \geq 1$ mit $\delta(\mathcal{Z}_n) \leq \delta$ für alle $n \geq n_0$. Wendet man dann die obige Argumentation auf solche $\mathcal{Z}_n$ an, so ergibt sich

$$\left|\mathrm{I}^{\mathcal{Z}_n}(f) - \mathrm{I}_{\mathcal{Z}_n}(f)\right| \leq \varepsilon(b-a)$$

und dies zeigt $\lim_{n\to\infty}\left(\mathrm{I}^{\mathcal{Z}_n}(f) - \mathrm{I}_{\mathcal{Z}_n}(f)\right) = 0$, d.h. (7.1.6).

Nicht jede beschränkte Funktion $f : [a, b] \to \mathbb{R}$ ist Riemann-integrierbar. Betrachten wir beispielsweise die Indikatorfunktion $f := \mathbf{1}_{[a,b]\setminus\mathbb{Q}} : [a, b] \to \mathbb{R}$, so gibt es für jedes Teil-Intervall $[x_{k-1}, x_k]$ eine rationale Zahl $q \in \mathbb{Q}$ und eine irrationale Zahl $r \in \mathbb{R} \setminus \mathbb{Q}$ mit $q, r \in [x_{k-1}, x_k]$. Es folgt

$$\inf_{\xi\in[x_{k-1},x_k]} f(\xi) = 0\,,$$

$$\sup_{\xi\in[x_{k-1},x_k]} f(\xi) = 1$$

und damit $\mathrm{I}_{\mathcal{Z}}(f) = 0$ und $\mathrm{I}^{\mathcal{Z}}(f) = b - a$ für *jede* Zerlegung $\mathcal{Z}$ von $[a, b]$. Damit kann f nicht Riemann-integrierbar sein.

Es gibt auch unstetige Funktionen, die Riemann-integrierbar sind. Betrachten wir dazu für $c \in [a, b]$ die Funktion $\mathbf{1}_{[c,b]} : [a, b] \to \mathbb{R}$. Es ist dann nicht schwierig zu zeigen, dass diese Funktion Riemann-integrierbar ist. Aus Zeitgründen überspringen wir aber den Beweis.

Mit Hilfe des Lemmas 7.1.3 können wir nun das Riemann-Integral definieren:

Definition 7.1.5

Sei $f : [a, b] \to \mathbb{R}$ Riemann-integrierbar und $(\mathcal{Z}_n)_{n\geq 1}$ eine Folge von Zerlegungen von $[a, b]$ mit $\delta(\mathcal{Z}_n) \to 0$. Dann ist das Riemann-Integral von f durch

$$\int_a^b f(x)\,\mathrm{d}x := \lim_{n\to\infty} \mathrm{I}_{\mathcal{Z}_n}(f) = \lim_{n\to\infty} \mathrm{I}^{\mathcal{Z}_n}(f)\,.$$

definiert.

Ist $f : [a, b] \to \mathbb{R}$ eine konstante Funktion mit $f(x) = c$ für alle $x \in [a, b]$, so hatten wir schon $\mathrm{I}_{\mathcal{Z}}(f) = \mathrm{I}^{\mathcal{Z}}(f) = c \cdot (b - a)$ für jede Zerlegung $\mathcal{Z}$ gesehen. Dies ergibt

$$\int_a^b f(x)\,\mathrm{d}x = c \cdot (b - a)\,. \tag{7.1.8}$$

Man beachte, dass das Riemann-Integral nicht nur im Fall von (7.1.8) sondern *immer* unabhängig von der Wahl der Zerlegungen $(\mathcal{Z}_n)_{n\geq 1}$ ist, da die Grenzwerte nach Lemma 7.1.3 gleich den dort erwähnten Supremum, bzw. Infimum sind. Diese Beobachtung zeigt auch

$$\mathrm{I}_{\mathcal{Z}}(f) \leq \int_a^b f(x)\,\mathrm{d}x \leq \mathrm{I}^{\mathcal{Z}}(f) \tag{7.1.9}$$

für jede Zerlegung $\mathcal{Z}$.

Tatsächlich ist es auch überflüssig, die Suprema und Infima in jedem Teil-Intervall $[x_{k-1}, x_k]$ zu bestimmen. Wählt man nämlich zu einer Zerlegung $\mathcal{Z} : x_0, \ldots, x_N$ von $[a, b]$ beliebige Stützstellen $\xi_k \in [x_{k-1}, x_k]$, so gelten für jedes beschränkte $f : [a, b] \to \mathbb{R}$ die Ungleichungen

$$\mathrm{I}_{\mathcal{Z}}(f) \leq \sum_{k=1}^{N} f(\xi_k)(x_k - x_{k-1}) \leq \mathrm{I}^{\mathcal{Z}}(f)\,. \tag{7.1.10}$$

Wie wollen diese Summen als Riemann-Summen bezeichnen. Ist f Riemann-integrierbar, so nähern sich die Unter- und Obersummen für zunehmende Feinheit der Zerlegungen immer weiter an, und diese Einsicht ergibt den folgenden Satz.

Theorem 7.1.6

Sei $f : [a, b] \to \mathbb{R}$ Riemann-integrierbar. Dann gilt für jede Folge $(\mathcal{Z}_n)_{n\in\mathbb{N}}$ von Zerlegungen $\mathcal{Z}_n : x_0^{(n)}, \dots, x_{N^{(n)}}^{(n)}$ mit $\delta(\mathcal{Z}_n) \to 0$ und jede zugehörige Wahl von Stützstellen $\xi_k^{(n)} \in [x_{k-1}^{(n)}, x_k^{(n)}]$

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{n\to\infty} \sum_{k=1}^{N^{(n)}} f(\xi_k^{(n)})(x_k^{(n)} - x_{k-1}^{(n)})\,.$$

Insbesondere gilt also bei äquidistanter Stützstellenwahl

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{n\to\infty} \frac{b-a}{n} \sum_{k=1}^{n} f\left(a + \frac{k(b-a)}{n}\right). \tag{7.1.11}$$

BEISPIEL

Riemann-Summen können helfen, Integrale explizit zu bestimmen. Um ein Beispiel zu geben, betrachten wir die Funktion $f(x) = x$. Diese ist auf jedem Intervall $[0, b]$ stetig und damit Riemann-integrierbar. Durch Betrachten des Grenzwert einer Riemann-Summe mit äquidistanten Stützstellen, siehe (7.1.11), erhalten wir

$$\int_0^b x \, \mathrm{d}x = \lim_{n\to\infty} \frac{b}{n} \sum_{k=1}^{n} f\left(\frac{kb}{n}\right) = \lim_{n\to\infty} \frac{b^2}{n^2} \sum_{k=1}^{n} k = \lim_{n\to\infty} \frac{b^2}{n^2} \frac{n(n+1)}{2} = \frac{1}{2} b^2.$$

BEMERKUNGEN

Die Güte der Approximation eines Integrals, durch Unter-, Ober-, oder Riemann-Summen kann für bestimmte Funktionen einfach quantifiziert werden. Ist $f : [a, b] \to \mathbb{R}$ beispielsweise α-Hölder-stetig mit Konstante c, so zeigt eine zu (7.1.7) analoge Abschätzung

$$\left|\mathrm{I}^{\mathcal{Z}}(f) - \mathrm{I}_{\mathcal{Z}}(f)\right| \leq c \cdot (b-a) \cdot \delta^{\alpha}(\mathcal{Z})$$

für jede Zerlegung $\mathcal{Z}$ von $[a, b]$. Wegen (7.1.9) hat die Feinheit damit einen *kontrollierbaren* Einfluss auf die Güte der Approximationen

$$\mathrm{I}_{\mathcal{Z}}(f) \approx \int_a^b f(x)\,\mathrm{d}x \approx \mathrm{I}^{\mathcal{Z}}(f)\,.$$

Das gleiche gilt für die Approximation des Integrals durch eine Riemann-Summe $R_{\mathcal{Z}}$, denn wegen (7.1.9) und (7.1.10) haben wir

$$\left|\int_a^b f(x)\,\mathrm{d}x - R_{\mathcal{Z}}\right| \leq \left|\mathrm{I}^{\mathcal{Z}}(f) - \mathrm{I}_{\mathcal{Z}}(f)\right| \leq c \cdot (b-a) \cdot \delta^{\alpha}(\mathcal{Z})\,.$$

Für äquidistanten Stützstellen, siehe (7.1.11), ist die Feinheit der zugehörigen Zerlegung $\mathcal{Z}$ durch $\delta(\mathcal{Z}) = n^{-1}$ gegeben. Dies ergibt die Fehlerabschätzung $c \cdot (b-a)n^{-\alpha}$. Für glattere Funktionen gibt es aber deutlich bessere Approximationsverfahren.

Auch für das Riemann-Integral gibt es Rechenregeln. Der folgende Satz fasst die einfachsten zusammen.

Theorem 7.1.7

Seien $f, g : [a, b] \to \mathbb{R}$ Riemann-integrierbar und $\alpha, \beta \in \mathbb{R}$. Dann gelten die folgenden Aussagen:

i). ***Linearität:*** *$\alpha f + \beta g$ ist Riemann-integrierbar und es gilt*

$$\int_a^b \alpha f(x) + \beta g(x)\,\mathrm{d}x = \alpha \int_a^b f(x)\,\mathrm{d}x + \beta \int_a^b g(x)\,\mathrm{d}x\,.$$

ii). ***Zerlegung des Definitionsbereichs:*** *Für jedes $c \in (a, b)$ ist f auch auf den Teil-Intervallen $[a, c]$ und $[c, b]$ integrierbar und es gilt*

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x\,.$$

BEWEIS 1/2

i). Seien $\alpha, \beta \geq 0$ und $\mathcal{Z} : x_0, \ldots, x_N$ eine Zerlegung von $[a, b]$. Dann gilt

$$\sup_{\xi \in [x_{k-1}, x_k]} \big(\alpha f(\xi) + \beta g(\xi)\big) \leq \alpha \sup_{\xi \in [x_{k-1}, x_k]} f(\xi) + \beta \sup_{\xi \in [x_{k-1}, x_k]} g(\xi) \tag{7.1.12}$$

und damit folgt $\mathrm{I}^{\mathcal{Z}}(\alpha f + \beta g) \leq \alpha \mathrm{I}^{\mathcal{Z}}(f) + \beta \mathrm{I}^{\mathcal{Z}}(g)$. Da für Infima die Ungleichung (7.1.12) umgekehrt gilt, kann man analog auf $\mathrm{I}_{\mathcal{Z}}(\alpha f + \beta g) \geq \alpha \mathrm{I}_{\mathcal{Z}}(f) + \beta \mathrm{I}_{\mathcal{Z}}(g)$ schließen.

Haben wir nun eine Folge $\mathcal{Z}_n$ von Zerlegungen mit $\delta(\mathcal{Z}_n) \to 0$ so ergeben beide Abschätzungen zusammen

$$\alpha \mathrm{I}_{\mathcal{Z}_n}(f) + \beta \mathrm{I}_{\mathcal{Z}_n}(g) \leq \mathrm{I}_{\mathcal{Z}_n}(\alpha f + \beta g) \leq \mathrm{I}^{\mathcal{Z}_n}(\alpha f + \beta g) \leq \alpha \mathrm{I}^{\mathcal{Z}_n}(f) + \beta \mathrm{I}^{\mathcal{Z}_n}(g)\,.$$

Aus der Riemann-Integrierbarkeit von f und g erhalten wir dann die Behauptung.

BEWEIS 2/2

Um die Aussage auch für negative Koeffizienten zu zeigen, reicht es den Fall $\alpha = -1$ und $\beta = 0$ zu betrachten. In diesem Fall folgt die Riemann-Integrierbarkeit aber leicht aus der Identität

$$\mathrm{I}^{\mathcal{Z}}(-f) - \mathrm{I}_{\mathcal{Z}}(-f) = -\mathrm{I}_{\mathcal{Z}}(f) + \mathrm{I}^{\mathcal{Z}}(f)\,,$$

wobei wir zweimal die Gleichung (7.1.3) angewendet haben. Das Integral von $-f$ kann ebenfalls mit (7.1.3) bestimmt werden.

ii). Seien $\mathcal{Z}' : x_0, \ldots, x_N$ eine Zerlegung von $[a, c]$ und $\mathcal{Z}'' : y_0, \ldots, y_M$ eine Zerlegung von $[c, b]$. Dann ist $\mathcal{Z} : x_0, \ldots, x_N, y_1, \ldots, y_M$ eine Zerlegung von $[a, b]$ mit $\delta(\mathcal{Z}) \leq \delta(\mathcal{Z}') + \delta(\mathcal{Z}'')$. Dies ergibt

$$\mathrm{I}^{\mathcal{Z}'}(f_{|[a,c]}) - \mathrm{I}_{\mathcal{Z}'}(f_{|[a,c]}) \leq \mathrm{I}^{\mathcal{Z}}(f) - \mathrm{I}_{\mathcal{Z}}(f)\,,$$
$$\mathrm{I}^{\mathcal{Z}''}(f_{|[c,b]}) - \mathrm{I}_{\mathcal{Z}''}(f_{|[c,b]}) \leq \mathrm{I}^{\mathcal{Z}}(f) - \mathrm{I}_{\mathcal{Z}}(f)\,.$$

Durch Betrachten von Folgen $(\mathcal{Z}'_n)_{n\in\mathbb{N}}$ und $(\mathcal{Z}''_n)_{n\in\mathbb{N}}$ mit $\delta(\mathcal{Z}'_n) \to 0$ und $\delta(\mathcal{Z}''_n) \to 0$ ergibt sich dann die Riemann-Integrierbarkeit von $f_{|[a,c]}$ und $f_{|[c,b]}$.

Ist $f : [a, b] \to \mathbb{R}$ Riemann-integrierbar, so setzen wir im folgenden

$$\int_a^a f(x)\,\mathrm{d}x := 0$$

und

$$\int_b^a f(x)\,\mathrm{d}x := -\int_a^b f(x)\,\mathrm{d}x\,.$$

Das Riemann-Integral erfüllt außerdem einige Ungleichungen, die im folgenden Satz dargestellt werden.

Theorem 7.1.8

Seien $f, g : [a, b] \to \mathbb{R}$ Riemann-integrierbar. Dann gilt:

i). ***Beschränktheit:***

$$\inf_{x\in[a,b]} f(x) \leq \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x \leq \sup_{x\in[a,b]} f(x)\,.$$

ii). ***Monotonie:*** *Gilt $f(x) \leq g(x)$ für alle $x \in [a, b]$ so folgt*

$$\int_a^b f(x)\,\mathrm{d}x \leq \int_a^b g(x)\,\mathrm{d}x\,.$$

iii). ***Vergleich mit Supremums-Norm:*** *Die Funktion $|f|$ ist Riemann-integrierbar und es gilt:*

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leq \int_a^b |f(x)|\,\mathrm{d}x \leq (b-a)\cdot\|f\|_\infty\,. \tag{7.1.13}$$

BEMERKUNG

Haben wir eine Funktionenfolge $(f_n)_{n\in\mathbb{N}}$ von Riemann-integrierbaren Funktionen $f_n : [a,b] \to \mathbb{R}$, die *gleichmäßig* gegen eine Riemann-integrierbare Funktion $f : [a,b] \to \mathbb{R}$ konvergiert, so gilt mit *iii)* aus Satz 7.1.8

$$\left| \int_a^b f_n(x)\,\mathrm{d}x - \int_a^b f(x)\,\mathrm{d}x \right| \leq (b-a) \cdot \|f_n - f\|_\infty \to 0\,.$$

BEWEIS 1/3

i). Durch Betrachten der trivialen Zerlegung $\mathcal{Z} : a, b$ folgt dies sofort aus (7.1.9).

ii). Wir definieren $h(x) := g(x) - f(x)$ für alle $x \in [a, b]$. Dann gilt $\inf_{x \in [a,b]} h(x) \geq 0$, und mit *i)* erhalten wir

$$0 \leq \frac{1}{b-a} \int_a^b g(x) - f(x) \, \mathrm{d}x \, .$$

Die Linearität des Riemann-Integrals ergibt dann die Behauptung.

iii). Wir setzen $f^+ := \max\{0, f\}$ und $f^- := -\min\{0, f\} = \max\{0, -f\}$. Es gilt dann $|f| = f^+ + f^-$ und damit muss für die Riemann-Integrierbarkeit von $|f|$ nur die von f^+ überprüft werden. Sei dazu $\mathcal{Z} : x_0, \ldots, x_N$ eine Zerlegung von $[a, b]$. Wir betrachten dann ein Teil-Intervall $[x_{k-1}, x_k]$.

$$\sup_{\xi \in [x_{k-1}, x_k]} f^+(\xi) - \inf_{\xi \in [x_{k-1}, x_k]} f^+(\xi) = 0 \leq \sup_{\xi \in [x_{k-1}, x_k]} f(\xi) - \inf_{\xi \in [x_{k-1}, x_k]} f(\xi)$$

Gibt es umgekehrt ein $\xi^* \in [x_{k-1}, x_k]$ mit $f(\xi^*) > 0$, so folgt

$$\sup_{\xi \in [x_{k-1}, x_k]} f^+(\xi) = \sup_{\xi \in [x_{k-1}, x_k]} f(\xi)$$

und wegen $f \leq f^+$ damit auch wieder

$$\sup_{\xi \in [x_{k-1}, x_k]} f^+(\xi) - \inf_{\xi \in [x_{k-1}, x_k]} f^+(\xi) \leq \sup_{\xi \in [x_{k-1}, x_k]} f(\xi) - \inf_{\xi \in [x_{k-1}, x_k]} f(\xi)\,.$$

BEWEIS 3/3

Die Definitionen der Unter- und Obersummen impliziert dann

$$\left|\mathrm{I}^{\mathcal{Z}}(f^+) - \mathrm{I}_{\mathcal{Z}}(f^+)\right| \leq \left|\mathrm{I}^{\mathcal{Z}}(f) - \mathrm{I}_{\mathcal{Z}}(f)\right|.$$

Mit der üblichen Argumentation sehen wir dann, dass f^+ Riemann-integrierbar ist.

Die erste Ungleichung folgt nun aus *ii)* durch Betrachten von $f \leq |f|$ und $-f \leq |f|$. Die zweite Ungleichung folgt aus *i)* angewendet auf $|f|$.

Section 7.2
Hauptsatz der Differential- und Integralrechnung

Ist $v(t)$ die Geschwindigkeit eines Objektes zur Zeit t so beschreibt

$$s(t) := \int_{t_0}^{t} v(t')\,\mathrm{d}t'$$

die seit dem Zeitpunkt t_0 zurückgelegte Strecke. Die Ableitung s' von s sollte wieder die Geschwindigkeit ergeben. Der folgende, als **Hauptsatz der Differential- und Integralrechnung** bekannte Satz zeigt diesen Sachverhalt in allgemeiner Form.

Theorem 7.2.1

Sei $f : [a, b] \to \mathbb{R}$ stetig. Dann ist die durch

$$F(x) := \int_{a}^{x} f(t)\,\mathrm{d}t\,, \qquad x \in [a, b]$$

definierte Funktion $F : [a, b] \to \mathbb{R}$ stetig und differenzierbar und es gilt

$$F'(x) = f(x)\,, \qquad x \in (a, b).$$

Sei $x \in (a, b)$. Wegen der Sätze 7.1.7 und 7.1.8 gilt dann für $h > 0$ mit $x + h \leq b$

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{h} \int_x^{x+h} f(t)\,\mathrm{d}t \leq \max_{t \in [x, x+h]} f(t) = f(x_h^*)$$

und ebenso

$$\frac{F(x+h) - F(x)}{h} \geq \min_{t \in [x, x+h]} f(t) = f(x_{*,h})\,,$$

wobei $x_h^*, x_{*,h} \in [x, x+h]$ gemäß Satz 5.2.8 gewählt worden sind. Für $h \to 0$ haben wir dann $x_h^*, x_{*,h} \to x$ und die Stetigkeit von f ergibt dann

$$f(x_h^*), f(x_{*,h}) \to f(x)\,.$$

Damit folgt die rechtsseitige Differenzierbarkeit von F mit rechtsseitiger Ableitung $f(x)$. Analog ergibt sich die linksseitige Differenzierbarkeit und somit die behauptete Differenzierbarkeit mit $F' = f$.

Die noch zu beweisende Stetigkeit in a und b kann ebenfalls analog gezeigt werden.

Das folgende Korollar ist als Mittelwertsatz der Integralrechnung bekannt.

Korollar 7.2.2

Sei $f : [a, b] \to \mathbb{R}$ stetig. Dann existiert ein $x_0 \in (a, b)$ mit

$$\int_a^b f(x)\,\mathrm{d}x = (b - a)f(x_0)\,.$$

BEWEIS

Wir betrachten die in Satz 7.2.1 definierte Funktion $F : [a, b] \to \mathbb{R}$. Diese ist stetig und differenzierbar mit $F' = f$, und damit gibt es nach dem Mittelwertsatz der Differentialrechnung, siehe Satz 6.2.3, ein $x_0 \in (a, b)$ mit

$$f(x_0) = F'(x_0) = \frac{F(b) - F(a)}{b - a} = \frac{1}{b - a} \int_a^b f(t)\,\mathrm{d}t\,.$$

Ein einfaches Umstellen ergibt dann die Behauptung.

Der Hauptsatz der Differential- und Integralrechnung motiviert die folgende Definition.

Definition 7.2.3

Sei $f : [a, b] \to \mathbb{R}$ stetig. Dann heißt jede stetige und differenzierbar Funktion $F : [a, b] \to \mathbb{R}$ mit $F' = f$ Stammfunktion von f.

Der Hauptsatz der Differential- und Integralrechnung zeigt, dass es zu jeder stetigen Funktion mindestens eine Stammfunktion gibt und dass diese durch Integration bestimmt werden kann.

Haben wir eine Stammfunktion F zu einer gegebenen Funktion f, so ist für $c \in \mathbb{R}$ ist wegen $(F + c)' = F' = f$ die Funktion $F + c$ wiederum eine Stammfunktion von f. Insbesondere hat jede stetige Funktion unendlich viele Stammfunktionen.

FAST EINDEUTIGKEIT VON STAMMFUNKTIONEN

Der folgende Satz zeigt, dass es keine weiteren Stammfunktionen gibt.

Theorem 7.2.4

Seien $f : [a, b] \to \mathbb{R}$ stetig und F_1 und F_2 Stammfunktionen von f. Dann gibt es ein $c \in \mathbb{R}$ mit $F_1 = F_2 + c$.

BEWEIS

Wir betrachten die Funktion $h := F_1 - F_2$. Diese ist stetig und es gilt $h'(x) = F_1'(x) - F_2'(x) = f(x) - f(x) = 0$ für alle $x \in (a, b)$. Damit gibt es nach Korollar 6.2.5 ein $c \in \mathbb{R}$ mit $h(x) = c$ für alle $x \in (a, b)$. Da h stetig ist gilt dann auch $h(a) = h(b) = c$. Dies ergibt die Behauptung.

Das folgende Korollar zeigt, dass zur Berechnung von Integralen eine Stammfunktion ausreicht.

Korollar 7.2.5

Sei $f : [a, b] \to \mathbb{R}$ stetig und F eine Stammfunktion von f. Dann gilt

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)\,.$$

Beweis.

Sei F_0 die im Hauptsatz 7.2.1 konstruierte Stammfunktion von f. Nach Konstruktion erfüllt F_0 die behauptete Formel. Nach Satz 7.2.4 gibt es dann ein $c \in \mathbb{R}$ mit $F_0 = F + c$ und dies ergibt $F(b) - F(a) = F_0(b) - F_0(a)$. □

Wir bezeichnen mit dem **unbestimmten Integral**

$$\int f(x)\,\mathrm{d}x$$

die Menge aller Stammfunktionen der stetigen Funktion f. Es gilt also (etwas informell geschrieben)

$$\int f(x)\,\mathrm{d}x = F(x) + c$$

mit einer beliebigen Konstanten c genau dann, wenn $F'(x) = f(x)$ für alle x gilt.

Ist schließlich F eine Stammfunktion von f, so schreiben wir auch

$$F\Big|_a^b := F(x)\Big|_a^b := F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x\,.$$

WICHTIGE STAMMFUNKTIONEN

Beispiele zu unbestimmten Integralen ergeben sich aus den schon hergeleiteten Ableitungen. Wir führen diese nachfolgend auf:

$$\int x^n \,\mathrm{d}x = \frac{1}{n+1} x^{n+1} + c\,, \qquad n \in \mathbb{N}_0$$

$$\int x^{-1} \,\mathrm{d}x = \ln|x| + c\,, \qquad x \neq 0$$

$$\int x^{-n-1} \,\mathrm{d}x = -\frac{1}{n} x^{-n} + c\,, \qquad n \in \mathbb{N}_0$$

$$\int \mathrm{e}^x \,\mathrm{d}x = \mathrm{e}^x + c$$

$$\int \cos x \,\mathrm{d}x = \sin x + c$$

$$\int \sin x \,\mathrm{d}x = -\cos x + c$$

$$\int \frac{1}{1+x^2} \,\mathrm{d}x = \arctan x + c$$

$$\int \frac{1}{\sqrt{1-x^2}} \,\mathrm{d}x = \arcsin x + c = -\arccos x + c\,, \qquad |x| < 1\,.$$

Hierbei soll $x \neq 0$ bedeuten, dass 0 nicht im Integrationsbereich liegt.

Sind $f, g : [a, b] \to \mathbb{R}$ stetig differenzierbare Funktionen, so gilt mit der Produktregel der Differentialrechnung

$$(f \cdot g)' = f' \cdot g + f \cdot g' .$$

Durch Umstellen nach $f \cdot g'$ und anschließende Integration erhalten wir damit das folgende Resultat, das als partielle Integration bekannt ist.

Theorem 7.2.6

Seien $f, g : [a, b] \to \mathbb{R}$ stetig differenzierbare Funktionen. Dann ist die Stammfunktion von $f \cdot g'$ durch

$$\int f(x) \cdot g'(x)\,\mathrm{d}x = f \cdot g - \int f'(x) \cdot g(x)\,\mathrm{d}x$$

gegeben und es gilt

$$\begin{aligned}\int_a^b f(x) \cdot g'(x)\,\mathrm{d}x &= f \cdot g \Big|_a^b - \int_a^b f'(x) \cdot g(x)\,\mathrm{d}x \\ &= f(b)g(b) - f(a)g(a) - \int_a^b f'(x) \cdot g(x)\,\mathrm{d}x .\end{aligned}$$

Durch Betrachten von $f(x) := x$ und $g(x) := \exp(x)$ erhalten wir

$$\int x\mathrm{e}^x \,\mathrm{d}x = x\mathrm{e}^x - \int \mathrm{e}^x \,\mathrm{d}x = (x-1)\cdot \mathrm{e}^x \,.$$

Setzen wir $f(x) := \ln x$ und $g(x) := x$ erhalten wir

$$\int \ln x \,\mathrm{d}x = x \ln x - \int \frac{x}{x} \,\mathrm{d}x = x \ln x - x \,.$$

Aus der Kettenregel der Differentialrechnung folgt wiederum der folgende Satz, der als Substitutionsregel bekannt ist.

Theorem 7.2.7

Sei $f : [c, d] \to \mathbb{R}$ stetig und $\varphi : [a, b] \to [c, d]$ stetig differenzierbar. Dann gilt für alle $x \in [a, b]$:

$$\int_a^x f(\varphi(t)) \cdot \varphi'(t)\,\mathrm{d}t = \int_{\varphi(a)}^{\varphi(x)} f(s)\,\mathrm{d}s\,.$$

Ist F eine Stammfunktion von f, so lässt sich die obige Aussage auch als

$$\int f(\varphi(t)) \cdot \varphi'(t)\,\mathrm{d}t = F(\varphi(x))\,.$$

schreiben.

BEWEIS

Sei F eine Stammfunktion von f. Dann gilt mit der Kettenregel

$$(F \circ \varphi)'(t) = F'(\varphi(t)) \cdot \varphi'(t) = f(\varphi(t)) \cdot \varphi'(t)\,.$$

Damit ist $F \circ \varphi$ eine Stammfunktion von $(f \circ \varphi) \cdot \varphi'$ und mit zweimaliger Anwendung des Hauptsatzes der Differential- und Integralrechnung folgt

$$\int_a^x f(\varphi(t)) \cdot \varphi'(t)\,\mathrm{d}t = F \circ \varphi \Big|_a^b = F(\varphi(b)) - F(\varphi(a)) = \int_{\varphi(a)}^{\varphi(x)} f(s)\,\mathrm{d}s$$

für alle $x \in (a, b]$. Für $x = a$ ist nichts zu beweisen.

In den folgenden Beispielen sei F immer eine Stammfunktion von f.

Für $\lambda \neq 0$ und $c \in \mathbb{R}$ gilt mit $\varphi(t) := \lambda t + c$:

$$\int_a^x f(\lambda t + c)\,\mathrm{d}t = \frac{1}{\lambda}\int_a^x f(\lambda t + c)\lambda\,\mathrm{d}t = \frac{1}{\lambda}\int_{\lambda a + c}^{\lambda x + c} f(s)\,\mathrm{d}s$$
$$= \frac{F(\lambda x + c) - F(\lambda a + c)}{\lambda}\,.$$

Durch Betrachten von $\varphi(t) := t^2$ erhalten wir

$$\int t f(t^2)\,\mathrm{d}t = \frac{1}{2}\int f(t^2)2t\,\mathrm{d}t = \frac{1}{2}F(x^2)\,.$$

Für $f(x) := \exp(-x)$ ergibt sich wegen $F(x) = -\exp(-x)$ damit

$$\int t\exp(t^2)\,\mathrm{d}t = -\frac{1}{2}\exp(-x^2)\,.$$

Manchmal ist es aber auch hilfreich, die Substitution für bestimmte, feste f zu betrachten. So gilt beispielsweise für $f(x) := x$:

$$\int_a^x \varphi(t) \cdot \varphi'(t)\,\mathrm{d}t = \int_{\varphi(a)}^{\varphi(x)} s\,\mathrm{d}s = \frac{s^2}{2}\Big|_{\varphi(a)}^{\varphi(x)} = \frac{\varphi^2(x) - \varphi^2(a)}{2}\,.$$

Damit ist $\varphi^2/2$ Stammfunktion von $\varphi \cdot \varphi'$. Diese Formeln lassen sich aber auch mit partieller Integration mit $f := g := \varphi$ herleiten.

Für $f(x) := x^{-1}$ ergibt sich wiederum

$$\begin{aligned}\int_a^x \frac{\varphi'(t)}{\varphi(t)}\,\mathrm{d}t = \int_a^x f(\varphi(t)) \cdot \varphi'(t)\,\mathrm{d}t = \int_{\varphi(a)}^{\varphi(x)} s^{-1}\,\mathrm{d}s = \ln|s|\Big|_{\varphi(a)}^{\varphi(x)} \\ = \ln|\varphi(x)| - \ln|\varphi(a)|\,,\end{aligned}$$

wobei 0 nicht im obigen Integrationsbereich der Funktion $s \mapsto s^{-1}$ liegen darf. Insbesondere ist damit $\ln|\varphi(x)|$ eine Stammfunktion von $\frac{\varphi'}{\varphi}$.

ANWENDUNG: DIFFERENTIALGLEICHUNGEN

Differentialgleichungen erster Ordnung mit trennbaren Veränderlichen können direkt durch Integration gelöst werden. Gesucht ist eine differenzierbare Funktion $f : \mathbb{R} \to \mathbb{R}$, oder auf einem Teilintervall von $\mathbb{R}$, mit der Eigenschaft, dass $f'(x)$ sich als Produkt einer Funktion von $f(x)$ und einer Funktion von x schreiben lässt. Wir betrachten zwei Beispiele.

Das erste Beispiel ist

$$f'(x) = xf(x)\,.$$

Falls $f(x) \neq 0$ gilt, kann man die Gleichung umformen und erhält

$$\frac{f'(x)}{f(x)} = x$$

und damit nach Integration beider Seiten und Anwenden der Substitutionsregel

$$\ln|f(x)| = \int \frac{f'(x)}{f(x)}\,\mathrm{d}x = \int x\,\mathrm{d}x = \frac{1}{2}x^2 + C$$

und damit

$$f(x) = c\mathrm{e}^{x^2/2}$$

mit einer Konstanten $c \neq 0$. Daneben gibt es noch die konstante Lösung $f(x) = 0$, die wir am Anfang ausgeschlossen haben.

Section 7.3
Partialbruch-Zerlegung

In diesem Abschnitt wollen wir rationale Funktionen, d.h. Funktionen der Form

$$f(x) := \frac{p(x)}{q(x)}\,, \tag{7.3.1}$$

wobei p und q Polynome sind, integrieren. Grundlegend für unsere weitere Annahme ist dabei, dass wir q in seine linearen und quadratischen Faktoren wie in Abschnitt 162 beschrieben zerlegen können. Mit anderen Worten sind alle reellen und komplexen Nullstellen von q bekannt. Außerdem nehmen wir ohne Einschränkung an, dass $\mathbf{deg}\, q > 0$ gilt, da im Fall $\mathbf{deg}\, q = 0$ der Nenner eine Konstante ist und das Integrieren damit trivial ist.

Gilt nun $\deg p \geq \deg q$, so gibt es mit Hilfe der Polynomdivision, siehe Satz 2.7.3, Polynome r und s mit

$$p(x) = s(x)q(x) + r(x)\,, \qquad x \in \mathbb{R}\,,$$

sowie $\deg s = \deg p - \deg q$ und $\deg r < \deg q$. Damit lässt sich f schreiben als

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

Die Integration des Polynoms s ist dann eine einfache Übung und es bleibt die Integration der rationalen Funktion

$$g(x) := \frac{r(x)}{q(x)}$$

übrig. Für diese gilt $\deg r < \deg q$ und daher können wir von vornherein $\deg p < \deg q$ in (7.3.1) annehmen.

BEISPIEL

Bevor wir in unseren allgemeinen Betrachtungen weitergehen, wollen wir uns nun zunächst ein Beispiel anschauen. Dazu betrachten wir die Funktion

$$f(x) := \frac{x}{x^2-1} = \frac{x}{(x-1)\cdot(x+1)}, \qquad x \neq \pm 1.$$

Unser Ansatz ist es nun $a, b \in \mathbb{R}$ zu finden mit

$$\frac{x}{(x-1)\cdot(x+1)} = \frac{a}{x-1} + \frac{b}{x+1} = \frac{a(x+1)+b(x-1)}{(x-1)\cdot(x+1)} = \frac{(a+b)x+b-a}{(x-1)\cdot(x+1)}.$$

Damit muss also $x = (a+b)x + b - a$ für alle $x \neq \pm 1$ gelten, und dies impliziert $1 = a + b$ und $0 = a - b$. Lösen dieser beiden Gleichungen ergibt zunächst $a = b$ und damit $1 = 2a$, d.h. $a = b = 1/2$. Die resultierenden Brüche

$$\frac{1}{2}\cdot\frac{1}{x-1} \qquad \text{und} \qquad \frac{1}{2}\cdot\frac{1}{x+1}$$

lassen sich nun leicht integrieren, so ist beispielsweise $\ln|x-1|$ eine Stammfunktion von $(x-1)^{-1}$. Insgesamt erhalten wir auf diese Weise

$$\int \frac{x}{x^2-1}\,\mathrm{d}x = \frac{1}{2}\int \frac{1}{x-1}\,\mathrm{d}x + \frac{1}{2}\int \frac{1}{x+1}\,\mathrm{d}x$$

Es ist leicht zu sehen, dass unser Ansatz im obigen Beispiel auch dann noch funktioniert, wenn das Zähler-Polynom von der allgemeinen Form $p(x) = mx + c$ ist. Wie sieht es aber mit komplizierteren rationalen Funktionen aus, bei Nullstellen mehrfach auftreten können und auch im Komplexen liegen können?

Dazu erinnern wir uns zunächst daran, dass reelle Polynome in endlich viele lineare und quadratische Faktoren zerfallen, siehe Abschnitt 162. Haben nun p und q einen gemeinsamen solchen Faktor h, d.h. es gibt Polynome $\tilde{p}$ und $\tilde{q}$ mit $p(x) = h(x) \cdot \tilde{p}(x)$ und $q(x) = h(x) \cdot \tilde{q}(x)$, so gilt

$$f(x) = \frac{p(x)}{q(x)} = \frac{h(x) \cdot \tilde{p}(x)}{h(x) \cdot \tilde{q}(x)} = \frac{\tilde{p}(x)}{\tilde{q}(x)}.$$

Ferner gilt mit der Polynomdivision aus Satz 2.7.3, dass $\mathbf{deg}\, h + \mathbf{deg}\, \tilde{p} = \mathbf{deg}\, p$ und $\mathbf{deg}\, h + \mathbf{deg}\, \tilde{q} = \mathbf{deg}\, q$. Damit folgt $\mathbf{deg}\, \tilde{p} < \mathbf{deg}\, \tilde{q}$.

Falls $\tilde{p}$ und $\tilde{q}$ weitere gemeinsame Faktoren haben, so lässt sich dieses Kürzen solange wiederholen, bis es keine weiteren gemeinsamen Faktoren gibt. Im folgenden können wir daher zusätzlich annehmen, dass p und q in (7.3.1) in gekürzter Form vorliegen, d.h. keine gemeinsamen Faktoren haben.

Um diese letzte Annahme umzuformulieren, erinnern wir uns daran, dass die linearen Faktoren von q von der Gestalt $(x - \lambda_j)$ sind, wobei λ_j eine reelle Nullstelle von p ist. Ferner sind die quadratischen Faktoren reelle Polynome vom Grad 2, die die Gestalt

$$Q(x) := (x - \lambda_j) \cdot (x - \overline{\lambda}_j) = (x - \alpha_j - \mathrm{i}\beta_j)(x - \alpha_j + \mathrm{i}\beta_j) = (x - \alpha_j)^2 + \beta_j^2$$

haben, wobei $\lambda_j = \alpha_j + \mathrm{i}\beta_j$ und $\overline{\lambda}_j = \alpha_j - \mathrm{i}\beta_j$ zwei komplexe Nullstellen von q sind, siehe auch Lemma 2.7.1.

Da eine analoge Beschreibung auch für p gilt, ist unsere Annahme, dass p und q in gekürzter Form vorliegen, äquivalent zu der Annahme, dass p und q keine gemeinsamen komplexen Nullstellen haben.

Das folgende Lemma zeigt wie man mit Hilfe einer reellen Nullstelle von q die rationale Funktion vereinfachen kann.

Lemma 7.3.1

Seien p und q reelle Polynome mit $\deg p < \deg q$. Ferner sei $\lambda \in \mathbb{R}$ eine Nullstelle von q mit Vielfachheit v, d.h. es gibt ein Polynom r mit $q(x) = (x-\lambda)^v \cdot r(x)$ für alle $x \in \mathbb{R}$ und $r(\lambda) \neq 0$. Wir setzen

$$a := \frac{p(\lambda)}{r(\lambda)}\,.$$

Dann gibt es ein Polynom p_1 mit $\deg p_1 < \deg q - 1$ und

$$\frac{p(x)}{q(x)} = \frac{a}{(x-\lambda)^v} + \frac{p_1(x)}{(x-\lambda)^{v-1} \cdot r(x)} \tag{7.3.2}$$

für alle $x \in \mathbb{R}$ mit $q(x) \neq 0$. Ferner kann dieses Polynom p_1 durch

$$p(x) - ar(x) = (x-\lambda)p_1(x)\,, \qquad x \in \mathbb{R} \tag{7.3.3}$$

bestimmt werden.

Wir definieren $s(x) := p(x) - ar(x)$ für $x \in \mathbb{R}$. Dann gilt $\deg s \leq \max\{\deg p, \deg r\} < \deg q$ und

$$s(\lambda) = p(\lambda) - \frac{p(\lambda)}{r(\lambda)} \cdot r(\lambda) = 0\,.$$

Damit existiert nach Korollar 2.7.4 ein Polynom p_1 mit $\deg p_1 = \deg s - 1 < \deg q - 1$ mit $s(x) = (x - \lambda) \cdot p_1(x)$ für alle $x \in \mathbb{R}$. Mit anderen Worten erfüllt p_1 die Gleichung (7.3.3). Ferner gilt

$$\begin{aligned}\frac{a}{(x-\lambda)^\nu} + \frac{p_1(x)}{(x-\lambda)^{\nu-1} \cdot r(x)} &= \frac{ar(x)}{(x-\lambda)^\nu \cdot r(x)} + \frac{(x-\lambda)\cdot p_1(x)}{(x-\lambda)^\nu \cdot r(x)} \\ &= \frac{ar(x)}{q(x)} + \frac{p(x) - ar(x)}{q(x)} \\ &= \frac{p(x)}{q(x)}\end{aligned}$$

für alle $x \in \mathbb{R}$ mit $q(x) \neq 0$.

Lemma 7.3.2

Seien p und q reelle Polynome mit $\deg p < \deg q$. *Ferner sei $\lambda \in \mathbb{C} \setminus \mathbb{R}$ eine Nullstelle von q mit Vielfachheit w, d.h. es gibt ein reelles Polynom r mit $q(x) = Q^w(x) \cdot r(x)$ für alle $x \in \mathbb{R}$ und $r(\lambda) \neq 0$, wobei wir das reelle quadratische Polynom*

$$Q(x) := (x - \lambda) \cdot (x - \overline{\lambda}), \qquad x \in \mathbb{R}$$

betrachten. Wir setzen

$$\gamma := \frac{p(\lambda)}{r(\lambda)}, \qquad b := \frac{\operatorname{Im} \gamma}{\operatorname{Im} \lambda}, \qquad c := -\frac{\operatorname{Im}(\gamma\overline{\lambda})}{\operatorname{Im} \lambda}$$

Dann gibt es ein reelles Polynom p_2 mit $\deg p_2 < \deg q - 2$ *und*

$$\frac{p(x)}{q(x)} = \frac{bx + c}{Q^w(x)} + \frac{p_2(x)}{Q^{w-1}(x) \cdot r(x)} \tag{7.3.4}$$

für alle $x \in \mathbb{R}$ mit $q(x) \neq 0$. Ferner kann dieses Polynom p_2 durch

$$p(x) - (bx + c)r(x) = Q(x)p_2(x), \qquad x \in \mathbb{R} \tag{7.3.5}$$

bestimmt werden.

BEWEIS 1/2

Wir definieren $s(x) := p(x) - (bx + c)r(x)$ für $x \in \mathbb{R}$. Wegen $\deg r = \deg q - 2$ gilt dann $\deg s \leq \max\{\deg p, 1 + \deg r\} < \deg q$.

Um s durch Q teilen zu können, wollen wir nun zeigen, dass λ und $\overline{\lambda}$ Nullstellen von s sind. Dazu bemerken wir zunächst, dass aus $r(\lambda) \neq 0$ wegen Lemma 2.7.1 auch $r(\overline{\lambda}) \neq 0$ folgt. Außerdem gilt

$$\overline{\gamma} = \frac{\overline{p(\lambda)}}{\overline{r(\lambda)}} = \frac{p(\overline{\lambda})}{r(\overline{\lambda})}\,, \tag{7.3.6}$$

wobei wir im letzten Schritt benutzt haben, dass p und r reelle Polynome sind. Ferner gilt

$$b = \frac{2\mathrm{i} \cdot \operatorname{Im}\gamma}{2\mathrm{i} \cdot \operatorname{Im}\lambda} = \frac{\gamma - \overline{\gamma}}{\lambda - \overline{\lambda}}$$

und

$$c = -\frac{2\mathrm{i} \cdot \operatorname{Im}(\gamma\overline{\lambda})}{2\mathrm{i} \cdot \operatorname{Im}\lambda} = -\frac{\gamma\overline{\lambda} - \overline{\gamma\overline{\lambda}}}{\lambda - \overline{\lambda}} = \frac{\overline{\gamma}\lambda - \gamma\overline{\lambda}}{\lambda - \overline{\lambda}}\,.$$

Damit gilt

$$s(\lambda) = p(\lambda) - (b\lambda + c)r(\lambda) = r(\lambda)\big(\gamma - b\lambda - c\big)$$

und wegen

BEWEIS 2/2

Analog folgt aus (7.3.6)

$$s(\overline{\lambda}) = p(\overline{\lambda}) - (b\overline{\lambda} + c)r(\overline{\lambda}) = r(\overline{\lambda})\big(\overline{\gamma} - b\overline{\lambda} - c\big)$$

und wegen

$$\overline{\gamma} - b\overline{\lambda} - c = \frac{\overline{\gamma}(\lambda - \overline{\lambda})}{\lambda - \overline{\lambda}} - \frac{(\gamma - \overline{\gamma}) \cdot \overline{\lambda}}{\lambda - \overline{\lambda}} - \frac{\overline{\gamma}\lambda - \gamma\overline{\lambda}}{\lambda - \overline{\lambda}} = 0$$

folgt auch $s(\overline{\lambda}) = 0$. Damit gibt es, wie im Abschnitt 162 ausgeführt, ein reelles Polynom p_2 mit $\mathbf{deg}\, p_2 = \mathbf{deg}\, s - 2 < \mathbf{deg}\, q - 2$ und

$$s(x) = (x - \lambda)(x - \overline{\lambda}) \cdot p_2(x) = Q(x)p_2(x)\,, \qquad x \in \mathbb{R}.$$

Mit anderen Worten erfüllt p_2 die Gleichung (7.3.5). Damit folgt aber auch

$$\begin{aligned}\frac{bx + c}{Q^w(x)} + \frac{p_2(x)}{Q^{w-1}(x) \cdot r(x)} &= \frac{(bx + c) \cdot r(x)}{Q^w(x) \cdot r(x)} + \frac{Q(x) \cdot p_2(x)}{Q^w(x) \cdot r(x)} \\ &= \frac{(bx + c) \cdot r(x)}{q(x)} + \frac{p(x) - (bx + c)r(x)}{q(x)} \\ &= \frac{p(x)}{q(x)}\end{aligned}$$

für alle $x \in \mathbb{R}$ mit $q(x) \neq 0$.

Wendet man die Lemmata 7.3.1 und 7.3.2 nacheinander auf alle reellen und komplexen Nullstellen von q an, so ergibt sich die folgende Vereinfachung von (7.3.1).

Theorem 7.3.3

Seien p und q Polynome mit $\deg p < \deg q$. Ferner seien $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ die reellen Nullstellen von q mit Vielfachheiten $v_1, \ldots, v_n$ und $Q_1, \ldots, Q_m$ die quadratischen Faktoren von q mit Vielfachheiten $w_1, \ldots, w_m$. Dann gibt es Konstanten $a_{i,j}, b_{i,j}, c_{i,j} \in \mathbb{R}$ mit

$$\frac{p(x)}{q(x)} = \sum_{i=1}^{n} \sum_{j=1}^{v_i} \frac{a_{i,j}}{(x-\lambda_i)^j} + \sum_{i=1}^{m} \sum_{j=1}^{w_i} \frac{b_{i,j}x + c_{i,j}}{Q_i^j(x)}$$

für alle $x \in \mathbb{R}$ mit $q(x) \neq 0$.

Zunächst betrachten wir nochmal die Funktion

$$f(x) := \frac{x}{x^2-1} = \frac{x}{(x-1)\cdot(x+1)}\,, \qquad x \neq \pm 1.$$

Dann ist $\lambda := 1$ eine einfache Nullstelle des Nenner-Polynoms $q(x) := x^2 - 1$. Setzen wir entsprechend $r(x) := x + 1$, so ist

$$a = \frac{p(1)}{r(1)} = \frac{1}{1+1} = \frac{1}{2}\,,$$

wobei $p(x) := x$ das Zähler-Polynom ist. Um p_1 aus Lemma 7.3.1 zu bestimmen, betrachten wir nun (7.3.3):

$$p(x) - ar(x) = x - \frac{1}{2}\cdot(x+1) = \frac{1}{2}\cdot(x-1) = \frac{1}{2}\cdot(x-\lambda)\,.$$

Damit ist $p_1 = \frac{1}{2}$ und (7.3.2) liefert mit $v = 1$ und $\lambda = 1$ die schon bekannte Zerlegung

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

Ein etwas komplizierteres und vollständiges Beispiel ist durch die Funktion

$$f(x) := \frac{x^5 + x^2 + x + 1}{x^4 - 2x^3 + 2x^2 - 2x + 1}$$

gegeben. Da das Zähler-Polynom noch keinen kleineren Grad als das Nenner-Polynom hat, führen wir zunächst Polynom-Division durch. Dies ergibt

$$f(x) = x + 2 + \frac{2x^3 - x^2 + 4x - 1}{x^4 - 2x^3 + 2x^2 - 2x + 1}\,.$$

Nun zerlegen wir das Nenner-Polynom in seine Faktoren

$$x^4 - 2x^3 + 2x^2 - 2x + 1 = (x-1)^2(x^2+1)\,.$$

Aus unserem allgemeinen Satz 7.3.3 wissen wir dann, dass es $a_1, a_2, b, c \in \mathbb{R}$ gibt mit

$$\frac{2x^3 - x^2 + 4x - 1}{x^4 - 2x^3 + 2x^2 - 2x + 1} = \frac{a_1}{(x-1)^2} + \frac{a_2}{x-1} + \frac{bx + c}{x^2 + 1} \tag{7.3.7}$$

Diese Konstanten können wir zum Beispiel durch iteratives Anwenden von Lemma 7.3.1 bzw. Lemma 7.3.2 bestimmen. Zur Bestimmung von a_1 setzen wir $r(x) = x^2 + 1$. Für die Nullstelle $\lambda = 1$ ergibt dies

$$a_1 = \frac{p(1)}{r(1)} = \frac{2 \cdot 1^3 - 1^2 + 4 \cdot 1 - 1}{1^2 + 1} = \frac{4}{2} = 2$$

und mit (7.3.3) sehen wir

$$\begin{aligned} p(x) - a_1 r(x) = 2x^3 - x^2 + 4x - 1 - 2(x^2 + 1) &= 2x^3 - 3x^2 + 4x - 3 \\ &= (x - 1) \cdot (2x^2 - x + 3)\,, \end{aligned}$$

wobei im letzten Schritt Polynomdivision benutzt wurde. Damit ist $p_1(x) := 2x^2 - x + 3$ und Lemma 7.3.1 ergibt

$$\frac{2x^3 - x^2 + 4x - 1}{x^4 - 2x^3 + 2x^2 - 2x + 1} = \frac{2}{(x-1)^2} + \frac{2x^2 - x + 3}{(x-1)(x^2+1)}\,.$$

Um a_2 zu bestimmen, betrachten wir den zweiten Bruch mit Hilfe von Lemma 7.3.1. Dies ergibt

$$a_2 = \frac{p_1(1)}{r(1)} = \frac{2 \cdot 1^2 - 1 + 3}{2} = \frac{4}{2} = 2\,.$$

Mit (7.3.3) sehen wir ferner

Damit ist $p_2 = -1$ und Lemma 7.3.1 ergibt

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

Insgesamt haben wir damit $a_1 = a_2 = 2$, $b = 0$ und $c = -1$ erhalten, d.h.

$$\frac{2x^3 - x^2 + 4x - 1}{x^4 - 2x^3 + 2x^2 - 2x + 1} = \frac{2}{(x-1)^2} + \frac{2}{x-1} - \frac{1}{x^2+1}$$

für alle $x \neq 1$.

BEISPIEL

Alternativ können wir die Form (7.3.7) auch direkt ausnutzen, um die Konstanten zu bestimmen. Addieren wir nämlich die 3 Brüche auf der rechten Seite und vergleichen den resultierenden Zähler mit p, so ergibt sich

$$\begin{aligned}
&2x^3 - x^2 + 4x - 1 \\
&= a_1 \cdot (x^2 + 1) + a_2 \cdot (x - 1) \cdot (x^2 + 1) + (bx + c) \cdot (x - 1)^2 \\
&= a_1 \cdot (x^2 + 1) + a_2 \cdot (x^3 - x^2 + x - 1) + b(x^3 - 2x^2 + x) + c \cdot (x^2 - 2x + 1) \\
&= (a_2 + b)x^3 + (a_1 - a_2 - 2b + c)x^2 + (a_2 + b - 2c)x + a_1 - a_2 + c\,.
\end{aligned}$$

Das Vergleichen der Koeffizienten ergibt das Gleichungssystem

$$\begin{aligned}
2 &= a_2 + b \\
-1 &= a_1 - a_2 - 2b + c \\
4 &= a_2 + b - 2c \\
-1 &= a_1 - a_2 + c\,.
\end{aligned}$$

Diese muss dann noch gelöst werden, was wir an dieser Stelle überspringen, da es einfach zu sehen ist, dass die obigen Koeffizienten $a_1 = a_2 = 2$, $b = 0$ und $c = -1$ das Gleichungssystem lösen.

Dieser Ansatz kann übrigens etwas vereinfacht werden, wenn man zunächst die reellen Nullstellen einsetzt. In obigen Fall wäre dies $x = 1$, wodurch unser obiger Zählervergleich sich auf

$$4 = 2 - 1 + 4 - 1 = a_1(1^2 + 1) + a_2 \cdot 0 + (bx + c) \cdot 0 = 2a_1$$

reduziert. Damit haben wir $a_1 = 2$ und das Gleichungssystem in 4 Variablen reduziert sich auf eins in 3 Variablen. Analog kann man dann auch die beiden komplexen Nullstellen $\pm\mathrm{i}$ einsetzen, um ein Gleichungssystem in b und c zu bekommen.

Zum Abschluss dieses Abschnittes listen wir die Formeln auf, die uns beim integrieren rationaler Funktionen helfen werden. Statt einer Darstellung als komplexer Partialbrüche bietet sich mitunter eine reelle Form an, wir geben den wichtigsten Fall dazu ebenfalls an. Es gilt für $v > 1$ und $Q > P^2$:

$$\int \frac{1}{x-\lambda}\,\mathrm{d}x = \ln|x-\lambda| + C\,, \qquad x \neq \lambda\,,$$

$$\int \frac{1}{(x-\lambda)^v}\,\mathrm{d}x = \frac{1}{-v+1}\cdot\frac{1}{(x-\lambda)^{v-1}} + C\,, \qquad x \neq \lambda\,,$$

$$\int \frac{1}{x^2+2Px+Q}\,\mathrm{d}x = \frac{1}{\sqrt{Q-P^2}}\arctan\frac{x+P}{\sqrt{Q-P^2}} + C\,,$$

$$\int \frac{x+P}{x^2+2Px+Q}\,\mathrm{d}x = \frac{1}{2}\ln(x^2+2Px+Q) + C\,.$$

Um die letzten beiden Formel anzuwenden, müssen wir dann lediglich noch

$$\frac{bx+c}{x^2+2Px+Q} = b\cdot\frac{x+P}{x^2+2Px+Q} + (c-bP)\cdot\frac{1}{x^2+2Px+Q}$$

beachten. Die Integration von höheren Potenzen von quadratischen Polynomen im Nenner ist ebenfalls möglich. Hier wird zunächst partielle Integration benutzt, um sukzessive die Potenz zu verringern. Wir verzichten auf die recht technischen Details.

Section 7.4
Uneigentliche Integrale

Im folgenden sei I ein Intervall und $f : I \to \mathbb{R}$. Wir wollen nun untersuchen, wann wir für f auch ein Integral über I definieren können. Dazu sagen wir, dass f **lokal Riemann-integrierbar** ist, falls f auf jedem abgeschlossenen Teil-Intervall von I Riemann-integrierbar ist.

Ist nun z.B. $I = [a, b)$ und $f : I \to \mathbb{R}$ lokal Riemann-integrierbar, dann sagen wir, dass f **uneigentlich Riemann-integrierbar** ist, falls der Grenzwert

$$\int_a^b f(x)\,\mathrm{d}x := \lim_{\varepsilon \to 0^+} \int_a^{b-\varepsilon} f(x)\,\mathrm{d}x$$

existiert. In diesem Fall heißt das Integral auf der linken Seite **uneigentliches Riemann-Integral**. Im Fall $I = [a, \infty)$ definieren wir analog

$$\int_a^\infty f(x)\,\mathrm{d}x := \lim_{r \to \infty} \int_a^r f(x)\,\mathrm{d}x\,,$$

falls der Grenzwert existiert, und die obigen Sprechweisen bleiben die gleichen. Halboffene Intervalle der Form $(a, b]$ und $(-\infty, b]$ werden ebenfalls analog betrachtet.

Offene Intervalle $I := (a, b)$ werden dadurch behandelt, dass man ein $c \in I$ fixiert und die uneigentliche Riemann-Integrierbarkeit auf $(a, c]$ und $[c, b)$ fordert. Das uneigentliche Integral ist dann

$$\int_a^b f(x)\,\mathrm{d}x := \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x\,.$$

Dieser Ansatz ist unabhängig von der Wahl von c, wie man sich leicht mit Satz 7.1.7 überlegen kann. Im Fall $I := (-\infty, \infty)$ ist der Ansatz analog.

Es gilt zum Beispiel

$$\int_0^1 \frac{1}{\sqrt{x}}\,\mathrm{d}x = \lim_{\varepsilon \to 0^+} \int_\varepsilon^1 \frac{1}{\sqrt{x}}\,\mathrm{d}x = 2\sqrt{1} - \lim_{\varepsilon \to 0^+} 2\sqrt{\epsilon} = 2$$

und

$$\int_1^\infty \frac{1}{x^2}\,\mathrm{d}x = \lim_{r \to \infty} \int_1^r \frac{1}{x^2}\,\mathrm{d}x = \lim_{r \to \infty} \frac{-1}{r} - \frac{-1}{1} = 1$$

Die Konvergenz uneigentlicher Integrale ist stets nachzuweisen. Oft hilft dazu das folgende Vergleichskriterium, dass vom Charakter sehr ähnlich zum dem Majoranten-Kriterium für Reihen ist.

Theorem 7.4.1

Seien $I := [a, b)$ ein halboffenes Intervall mit $b \in \mathbb{R} \cup \{\infty\}$ und $f, g : I \to \mathbb{R}$ lokal Riemann-integrierbar mit $|f(x)| \leq g(x)$ für alle $x \in I$. Ist dann g uneigentlich Riemann-integrierbar, so ist auch f uneigentlich Riemann-integrierbar und es gilt

$$\left| \int_a^b f(x) \, \mathrm{d}x \right| \leq \int_a^b g(x) \, \mathrm{d}x \, .$$

Eine analoge Aussage gilt auch für die anderen Fälle, in denen uneigentliche Riemann-Integrierbarkeit definiert ist.

Wir betrachten hier nur den Fall $I := [a, b)$ mit $b < \infty$. Sei dazu $(b_n)_{n\in\mathbb{N}} \subset [a, b)$ eine Folge mit $b_n \to b$. Wir setzen

$$\alpha_n := \int_a^{b_n} f(x)\,\mathrm{d}x \qquad \text{und} \qquad \gamma_n := \int_a^{b_n} g(x)\,\mathrm{d}x\,.$$

Dann konvergiert die Folge $(\gamma_n)_{n\in\mathbb{N}}$ nach Voraussetzung und damit ist sie auch eine Cauchy-Folge. Für $m, n \geq 1$ mit $b_m \leq b_n$ gilt ferner

$$|\alpha_n - \alpha_m| = \left|\int_{b_m}^{b_n} f(x)\,\mathrm{d}x\right| \leq \int_{b_m}^{b_n} \big|f(x)\big|\,\mathrm{d}x \leq \int_{b_m}^{b_n} g(x)\,\mathrm{d}x = |\gamma_n - \gamma_m|\,,$$

siehe Satz 7.1.7 und Satz 7.1.8. Damit ist auch $(\alpha_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge und damit konvergent.

Typische Vergleichsfunktionen sind $g(x) := x^\alpha$. Für $\alpha > -1$ gilt hierbei

$$\int_0^1 x^\alpha \, \mathrm{d}x = \frac{1^{\alpha+1}}{\alpha+1} - \lim_{\varepsilon \to 0+} \frac{\varepsilon^{\alpha+1}}{\alpha+1} = \frac{1}{\alpha+1} \tag{7.4.1}$$

und dies sind auch die einzigen α, für die Funktion g auf $(0, 1]$ uneigentlich integrierbar ist. Analog gilt für $\alpha < -1$

$$\int_1^\infty x^\alpha \, \mathrm{d}x = \lim_{r \to \infty} \frac{r^{\alpha+1}}{\alpha+1} - \frac{1^{\alpha+1}}{\alpha+1} = -\frac{1}{\alpha+1} \tag{7.4.2}$$

und dies sind wiederum die einzigen α, für die Funktion g auf $[1, \infty)$ uneigentlich integrierbar ist.

Eine Anwendung des zweiten Falls zeigt sofort, dass das uneigentliche Riemann-Integral

$$\int_1^\infty \frac{\sin x}{x^\alpha} \, \mathrm{d}x$$

für alle $\alpha > 1$ existiert.

Theorem 7.4.2

Sei $f : [0, \infty) \to [0, \infty)$ eine stetige und monoton fallende Funktion. Dann gilt für alle $n \geq 1$:

$$\sum_{k=2}^{n} f(k) \leq \int_{1}^{n} f(x)\,\mathrm{d}x \leq \sum_{k=1}^{n-1} f(k)\,. \tag{7.4.3}$$

Insbesondere ist f uneigentlich Riemann-integrierbar genau dann, wenn die Reihe $\sum_{k=1}^{n} f(k)$ konvergiert.

Satz 7.4.2 kann auch mit dem Majoranten-Kriterium aus Satz 4.2.6, bzw. der Monotonie des Integrals, siehe Satz 7.1.8, verbunden werden.

Betrachtet man die Funktion $f(x) := x^{-1}$, so ergibt sich für die harmonische Reihe

$$\sum_{k=2}^{n} k^{-1} \leq \ln n \leq \sum_{k=1}^{n-1} k^{-1}\,.$$

Die Folge der harmonischen Partialsummen wächst daher so schnell wie der natürliche Logarithmus.

Für $x \in [k, k+1]$ gilt $f(k+1) \leq f(x) \leq f(k)$. Damit folgt

$$f(k+1) \leq \int_k^{k+1} f(x)\,\mathrm{d}x \leq f(k)\,. \tag{7.4.4}$$

Summation über $k = 1, \ldots, n-1$ liefert dann die zwei Ungleichungen. Da alle drei Ausdrücke in diesen zwei Ungleichungen monoton wachsend in n sind, ist die Konvergenz der zugehörigen drei Folgen äquivalent zu ihrer Beschränktheit, siehe die Sätze 4.1.6 und 4.1.16.

BEMERKUNG

Die Ungleichungen (7.4.4) können auch zur Abschätzung von Reihen-Resten dienen. Für $\alpha > 1$ gilt beispielsweise

$$\int_k^{k+1} x^{-\alpha}\,\mathrm{d}x = \frac{x^{-\alpha+1}}{-\alpha+1}\bigg|_k^{k+1} = \frac{k^{-\alpha+1} - (k+1)^{-\alpha+1}}{\alpha - 1}$$

Wird dies für $f(x) := x^{-\alpha}$ mit (7.4.4) verbunden, so ergibt eine anschließende Summation über $k = n, \ldots, \infty$:

$$\sum_{k=n+1}^{\infty} k^{-\alpha} = \sum_{k=n}^{\infty} (k+1)^{-\alpha} \leq \frac{n^{-\alpha+1}}{\alpha - 1} \leq \sum_{k=n}^{\infty} k^{-\alpha}\,.$$

Umstellen ergibt für $n \geq 2$:

$$\frac{n^{-\alpha+1}}{\alpha - 1} \leq \sum_{k=n}^{\infty} k^{-\alpha} \leq \frac{(n-1)^{-\alpha+1}}{\alpha - 1}\,.$$

Man beachte ferner, dass für $n \geq 2$ gilt

$$\frac{(n-1)^{-\alpha+1}}{\alpha - 1} = \left(\frac{n-1}{n}\right)^{-\alpha+1} \cdot \frac{n^{-\alpha+1}}{\alpha - 1} \leq 2^{-\alpha+1} \cdot \frac{n^{-\alpha+1}}{\alpha - 1}\,.$$

Die Tails der Reihe über $k^{-\alpha}$ fallen daher so schnell wie die Folge $n^{-\alpha+1}$.

BEISPIEL: DIRICHLET-INTEGRAL

Als ein weiteres Beispiel für uneigentliche Integrale betrachten wir die Funktion $f(x) := \frac{\sin x}{x}$ für $x > 0$ und $f(0) := 1$. Diese Funktion ist stetig, da

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

Wir wollen nun sogenannte Dirichlet-Integral

$$\int_0^\infty \frac{\sin x}{x}\,\mathrm{d}x \tag{7.4.5}$$

untersuchen. Dazu bemerken wir zunächst, dass für $z_1 < z_2$ partielle Integration

$$\int_{z_1}^{z_2} \frac{\sin x}{x}\,\mathrm{d}x = \frac{1}{x} \cdot (-\cos x)\Big|_{z_1}^{z_2} - \int_{z_1}^{z_2} \frac{-1}{x^{-2}} \cdot (-\cos x)\,\mathrm{d}x\,.$$

ergibt. Damit erhalten wir

$$\begin{aligned}\left|\int_1^r \frac{\sin x}{x}\,\mathrm{d}x\right| &\leq \left|\frac{1}{r} \cdot (-\cos r) - \frac{1}{1} \cdot (-\cos 1)\right| + \int_1^r \left|\frac{\cos x}{x^2}\right| \mathrm{d}x \\ &\leq \frac{1}{r} + 1 + \int_1^r x^{-2}\,\mathrm{d}x\,.\end{aligned}$$

Der letzte Ausdruck ist beschränkt für $r \to \infty$ und damit existiert das uneigentliche Riemann-Integral (7.4.5).

BEISPIEL: DIRICHLET-INTEGRAL

Die Funktion $|f|$ ist jedoch nicht uneigentlich Riemann-integrierbar. So gilt für $k \geq 1$:

$$\int_{2k\cdot\pi}^{(2k+1)\cdot\pi} \left|\frac{\sin x}{x}\right| \mathrm{d}x \geq \frac{1}{(2k+1)\pi} \cdot \int_{2k\cdot\pi}^{(2k+1)\cdot\pi} \sin x \,\mathrm{d}x = \frac{2}{(2k+1)\pi}\,.$$

Damit folgt

$$\int_{2\pi}^{(2n+1)\pi} \left|\frac{\sin x}{x}\right| \mathrm{d}x \geq \sum_{k=1}^{n} \int_{2k\cdot\pi}^{(2k+1)\cdot\pi} \left|\frac{\sin x}{x}\right| \mathrm{d}x \geq \frac{2}{\pi} \sum_{k=1}^{n} \frac{1}{2k+1}$$

und dies zeigt, dass

$$\int_{0}^{\infty} \left|\frac{\sin x}{x}\right| \mathrm{d}x = \infty\,.$$

Die Gamma-Funktion $\Gamma : (0, \infty) \to \mathbb{R}$ ist durch

$$\Gamma(x) := \int_0^\infty \mathrm{e}^{-t} t^{x-1} \, \mathrm{d}t \tag{7.4.6}$$

definiert. Wir wollen hier zeigen, dass das zugrunde liegende, uneigentliche Integral für jedes $x \in (0, \infty)$ existiert. Dazu betrachten wir zunächst

$$\int_0^1 \mathrm{e}^{-t} t^{x-1} \, \mathrm{d}t \leq \int_0^1 t^{x-1} \, \mathrm{d}t \,,$$

und da das rechte, uneigentliche Integral wegen $x - 1 > -1$ existiert, siehe (7.4.1) existiert auch das linke, uneigentliche Integral nach Satz 7.4.1. Ferner gilt $\lim_{t \to \infty} \mathrm{e}^{-t} t^{x+1} = 0$, wie eine $\lfloor x + 1 \rfloor$-malige Anwendung der Regel von l'Hospital zeigt. Damit gibt es ein $t_0 \geq 1$ mit

$$\mathrm{e}^{-t} t^{x-1} \leq t^{-2}$$

für alle $t \geq t_0$. Dies ergibt

$$\int_{t_0}^\infty \mathrm{e}^{-t} t^{x-1} \, \mathrm{d}t \leq \int_{t_0}^\infty t^{-2} \, \mathrm{d}t \,,$$

und da das rechte, uneigentliche Integral wegen $x - 1 > -1$ existiert, siehe (7.4.2) existiert auch das linke, uneigentliche Integral nach Satz 7.4.1. Da die Integration auf dem verbliebenen Intervall $[1, t_0]$ kein Problem darstellt

Mit partieller Integration kann man ferner zeigen, dass

$$\Gamma(x+1) = x \cdot \Gamma(x)\,, \qquad x > 0$$

gilt. Da offensichtlich $\Gamma(x) = 1$ gilt, folgt mit Induktion

$$\Gamma(n+1) = n!$$

für alle $n \in \mathbb{N}$.

Chapter 8: Taylorreihen

Wir hatten die Exponentialfunktion in Kapitel 3 durch

$$\exp(z) := \sum_{k=0}^{\infty} \frac{z^k}{k!}, \qquad z \in \mathbb{C}$$

definiert. Gibt es weitere Funktionen, die durch eine Reihe über gewichtete Potenzen z^k beschrieben werden können?

In Theorem 6.1.4 hatten wir gesehen, dass eine differenzierbare Funktion durch ihre Tangente mit Fehler $o(x - x_0)$ approximiert werden kann. Unter welchen Voraussetzungen können wir eine bessere Approximation finden?

Section 8.1
Taylorpolynome

Da die Tangente an einer Funktion die gleiche Steigung wie die Funktion selber hat, wollen wir zunächst den folgenden Ansatz verfolgen, um die obigen Fragen zu beantworten.

Sei dazu I ein Intervall und $f : I \to \mathbb{R}$ eine n-mal in $x_0 \in I$ differenzierbare Funktion. Wir suchen zunächst ein Polynom $T_n := T_{n,f}$ vom Grad n mit

$$T_{n,f}^{(k)}(x_0) = f^{(k)}(x_0) \tag{8.1.1}$$

für alle $k = 0, \ldots, n$. Mit anderen Worten sollen die ersten n Ableitungen von f und T_n im Punkt x_0 übereinstimmen.

Die Gleichungen (8.1.1) ergeben insgesamt $n+1$ Bedingungen an T_n. Sei nun T_n von der Form

$$T_n(x) = \sum_{j=0}^{n} a_j (x - x_0)^j, \qquad x \in \mathbb{R}.$$

Dann gilt für die erste Ableitung

$$T'(x) = \sum_{j=1}^{n} a_j j (x - x_0)^{j-1}, \qquad x \in \mathbb{R},$$

da für $j = 0$ der Term $a_0(x - x_0)^0$ konstant ist und somit in der Ableitung verschwindet. Analog ergibt sich die zweite Ableitung

$$T''(x) = \sum_{j=2}^{n} a_j j(j-1)(x - x_0)^{j-2}$$

und für die k-te Ableitung die Formel

$$T_n^{(k)}(x) = \sum_{j=k}^{n} a_j \cdot j(j-1)\cdots(j-k+1)(x - x_0)^{j-k}, \qquad x \in \mathbb{R}.$$

Einsetzen von $x = x_0$ überführt dann (8.1.1) in

$$\begin{aligned} f^{(k)}(x_0) &= T_n^{(k)}(x_0) \\ &= \sum_{j=k}^{n} a_j \cdot j(j-1)\cdots(j-k+1)(x_0 - x_0)^{j-k} \\ &= a_k k(k-1)\cdots(k-k+1)(x_0 - x_0)^0 \\ &\qquad + \sum_{j=k+1}^{n} a_j \cdot j(j-1)\cdots(j-k+1)(x_0 - x_0)^{j-k} \\ &= a_k k! \,. \end{aligned}$$

Damit haben wir die Formeln

$$a_k = \frac{f^{(k)}(x_0)}{k!}, \qquad k = 0, \ldots, n \tag{8.1.2}$$

gefunden.

Definition 8.1.1

Sei I ein Intervall und $f : I \to \mathbb{R}$ eine n-mal in $x_0 \in I$ differenzierbare Funktion. Dann heißt

$$T_{n,f,x_0}(x) := T_{n,f}(x) := T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \qquad x \in \mathbb{R}.$$

das **um x_0 entwickelte Taylorpolynom n-ten Grades von f**.

Im Folgenden wollen wir quantitativ untersuchen, inwieweit ein Taylorpolynom T_n die Funktion f approximiert, d.h. wir wollen das **Restglied**

$$R_n(x) := f(x) - T_n(x)$$

abschätzen. Aufgrund unserer Konstruktion, genauer gesagt wegen (8.1.1) im Fall $k = 0$, haben wir immer $R_n(x_0) = 0$. Im Folgenden ist also der Fall $x \neq x_0$ besonders interessant.

Der folgende Satz, der als Satz von Taylor bekannt ist, liefert eine erste Antwort.

Theorem 8.1.2

Sei I ein Intervall, $x_0 \in I$ und $f : I \to \mathbb{R}$ eine $(n-1)$-mal differenzierbare Funktion, die im Punkt x_0 sogar n-mal differenzierbar ist. Dann gilt

$$R_n(x) = o(|x - x_0|^n), \qquad x \to x_0 .$$

BEWEIS 1/2

Nach Konstruktion gilt

$$\begin{aligned} T'_{n,f}(x) = \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{k!} k(x-x_0)^{k-1} &= \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{(k-1)!}(x-x_0)^{k-1} \\ &= \sum_{k=0}^{n-1} \frac{f^{(k+1)}(x_0)}{k!}(x-x_0)^{k} \\ &= T_{n-1,f'}(x)\,, \end{aligned}$$

und diese Formel können wir natürlich weiter iterieren, um $T^{(m)}_{n,f} = T_{n-m,f^{(m)}}$ zu erhalten.

BEWEIS 2/2

Mit der Regel von de L'Hospital für $\frac{0}{0}$, siehe Satz 6.2.12, folgt dann sukzessive

$$\begin{aligned}
\lim_{x\to x_0} \frac{f(x)-T_n(x)}{(x-x_0)^n} &= \lim_{x\to x_0} \frac{f'(x)-T'_{n,f}(x)}{n(x-x_0)^{n-1}} = \lim_{x\to x_0} \frac{f'(x)-T_{n-1,f'}(x)}{n(x-x_0)^{n-1}} \\
&= \lim_{x\to x_0} \frac{f''(x)-T'_{n-1,f'}(x)}{n(n-1)(x-x_0)^{n-2}} = \lim_{x\to x_0} \frac{f''(x)-T_{n-2,f''}(x)}{n(n-1)(x-x_0)^{n-2}} \\
&\vdots \\
&= \frac{1}{n!} \lim_{x\to x_0} \frac{f^{(n-1)}(x)-T_{1,f^{(n-1)}}(x)}{x-x_0} \\
&= \frac{1}{n!} \lim_{x\to x_0} \frac{f^{(n-1)}(x)-f^{(n-1)}(x_0)-f^{(n)}(x_0)\cdot(x-x_0)}{x-x_0} \\
&= \frac{1}{n!} \lim_{x\to x_0} \frac{f^{(n-1)}(x)-f^{(n-1)}(x_0)}{x-x_0} - f^{(n)}(x_0) \\
&= 0\,,
\end{aligned}$$

wobei wir die allgemeine Formel $T_{1,g}(x) = g(x_0) + g'(x_0)\cdot(x-x_0)$ für $g := f^{(n-1)}$ benutzt haben.

BEISPIEL

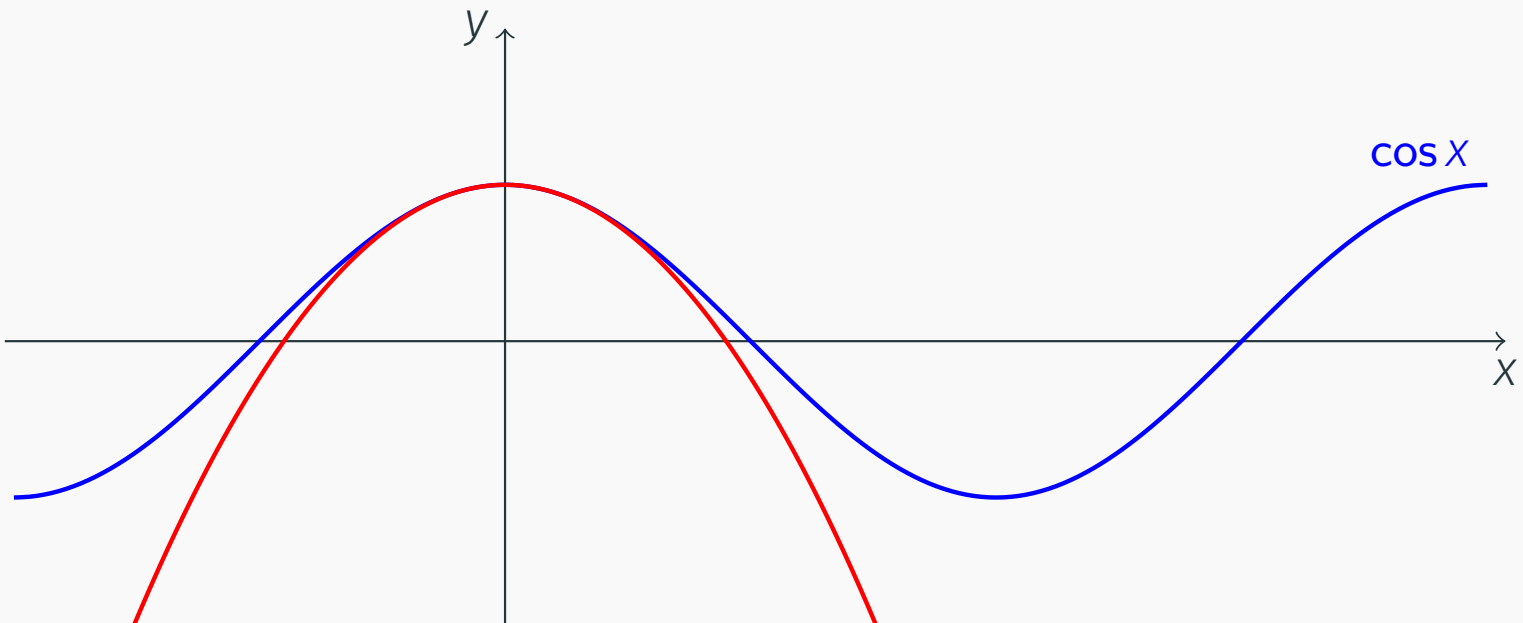
Um den Satz von Taylor an einem Beispiel zu veranschaulichen, betrachten die Funktion $f := \cos$ und $x_0 := 0$. Wegen $f(0) = \cos(0) = 1$ und $f'(0) = -\sin(0) = 0$ und $f''(0) = -\cos(0) = -1$ folgt

$$\cos(x) = T_2(x) + R_2(x) = 1 - \frac{1}{2}x^2 + R_2(x)$$

mit

$$\lim_{x \to 0} \frac{R_2(x)}{x^2} = 0\,.$$

Die Approximation der Kosinusfunktion durch T_2 ist in Abbildung 18 illustriert.



Der folgende Satz ergibt das sogenannte Lagrange Restglied, das eine genauere Abschätzung unter stärkeren Bedingungen liefert.

Theorem 8.1.3

Sei I ein Intervall, $x_0 \in I$ und $f : I \to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion. Dann gilt:

i). Für alle $x \in I$ mit $x \neq x_0$ gibt es ein $\alpha \in (0,1)$, so dass für $\xi := (1-\alpha)x_0 + \alpha x$ gilt

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}.$$

ii). Im Fall $x_0 < x$ gilt

$$|R_n(x)| \leq \sup_{\xi \in [x_0, x]} \frac{f^{(n+1)}(\xi)}{(n+1)!}|x - x_0|^{n+1}$$

und im Fall $x_0 > x$ gilt eine analoge Formel mit $\xi \in [x, x_0]$.

BEWEIS 1/2

Wir betrachten die folgenden Funktionen

$$F(x) := R_n(x) = f(x) - T_n(x)\,,$$
$$G(x) := (x - x_0)^{n+1}$$

für $x \in I$. Es gilt dann nach Konstruktion

$$F(x_0) = F'(x_0) = \cdots = F^{(n)}(x_0) = 0$$

und

$$G(x_0) = G'(x_0) = \cdots = G^{(n)}(x_0) = 0$$

Betrachten wir nun den Fall $x > x_0$. Sukzessives Anwenden des verallgemeinerten Mittelwertsatzes 6.2.9 mit $a := x_0$ und $b := x$ ergibt dann $\xi_i \in (x_0, \xi_{i-1})$ mit $\xi_0 := x$ und

$$\begin{aligned}\frac{F(x)}{G(x)} = \frac{F(x) - F(x_0)}{G(x) - G(x_0)} = \frac{F'(\xi_1)}{G'(\xi_1)} &= \frac{F'(\xi_1) - F'(x_0)}{G'(\xi_1) - G'(x_0)} \\ &= \frac{F''(\xi_2)}{G''(\xi_2)} \\ &= \cdots = \frac{F^{(n+1)}(\xi_{n+1})}{G^{(n+1)}(\xi_{n+1})}\,.\end{aligned}$$

BEWEIS 2/2

Da T_n ein Polynom vom Grad n ist, verschwindet seine $(n+1)$-te Ableitung und daher gilt

$$F^{(n+1)}(\xi_{n+1}) = f^{(n+1)}(\xi_{n+1})\,.$$

Ferner gilt

$$G^{(n+1)}(\xi_{n+1}) = (n+1)!(\xi_{n+1} - x_0)^0 = (n+1)!\,.$$

Zusammengenommen bekommen wir daher

$$R_n(x) = F(x) = G(x) \cdot \frac{F^{(n+1)}(\xi_{n+1})}{G^{(n+1)}(\xi_{n+1})} = (x - x_0)^{n+1} \cdot \frac{f^{(n+1)}(\xi_{n+1})}{(n+1)!}\,.$$

Wegen $\xi_{n+1} \in (x_0, x)$ können wir dann für $\xi := \xi_{n+1}$ ein $\alpha \in (0,1)$ mit $\xi = (1-\alpha)x_0 + \alpha x$ finden. Dies ergibt die Aussage *i)* und die Aussage *ii)* ist eine direkte Konsequenz aus *i)*.

Der Fall $x < x_0$ kann analog bewiesen werden.

ANWENDUNGEN DER RESTGLIEDABSCHÄTZUNGEN

Wir betrachten die Funktion $f := \exp$ und $x_0 := 0$. Wegen $\exp' = \exp$ und $\exp(0) = 1$ folgt $f^{(k)}(0) = 1$ für alle $k \geq 0$. Damit gilt

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{x^k} k! \,, \qquad x \in \mathbb{R}$$

und das Lagrange-Restglied ist

$$R_n(x) = \frac{\exp(\xi)}{(n+1)!} x^{n+1} \,.$$

Nehmen wir nun $x \in [0,1]$ an, so ist $\xi \in [0,1]$ und damit $\exp(\xi) \leq \mathrm{e}$. Dies ergibt

$$\sup_{x \in [0,1]} \left| \exp(x) - T_n(x) \right| = \sup_{x \in [0,1]} \left| R_n(x) \right| \leq \frac{\mathrm{e}}{(n+1)!} \,,$$

d.h. die gleichmäßige Approximation der Exponentialfunktion durch T_2 auf $[0,1]$ kann durch $\frac{\mathrm{e}}{(n+1)!}$ abgeschätzt werden. Für $n = 10$ ergibt sich beispielsweise

$$\sup_{x \in [0,1]} \left| \exp(x) - T_n(x) \right| \leq \frac{\mathrm{e}}{(n+1)!} \leq 6.81 \cdot 10^{-8} \,.$$

BEISPIEL: RELATIVISTISCHE KINETISCHE ENERGIE

Um ein weiteres Beispiel zu betrachten, “erinnern” wir daran, dass die kinetische Energie eines Teilchens mit Ruhemasse m_0 nach der Relativitätstheorie durch

$$E_{\mathrm{rel}} = m_0 c^2 \left(\frac{1}{\sqrt{1-(v/c)^2}} - 1 \right)$$

gegeben ist. Hierbei ist v die Geschwindigkeit des Teilchens und c die Lichtgeschwindigkeit. Für kleine Geschwindigkeiten wollen wir nun E_{rel} mit der kinetischen Energie

$$E_{\mathrm{kin}} = \frac{1}{2} m_0 v^2$$

nach Newton vergleichen. Hierzu definieren wir zunächst

$$f(x) := \frac{1}{\sqrt{1+x}} - 1\,, \qquad\qquad x \in (-1, 0],$$

so dass sich für $x = -(v/c)^2$ die Formel

$$E_{\mathrm{rel}} = m_0 c^2 f\big(-(v/c)^2\big)$$

ergibt. Da wir an $v \to 0$ interessiert sind, entwickeln wir nun f um $x_0 := 0$.

BEISPIEL: RELATIVISTISCHE KINETISCHE ENERGIE

Dazu benötigen wir

$$f'(x) = -\frac{1}{2}(1+x)^{-3/2},$$
$$f''(x) = \frac{3}{4}(1+x)^{-5/2}.$$

Das resultierende Taylorpolynom ist dann

$$T_2(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 = -\frac{x}{2} + \frac{3}{8}x^2,$$

und für $x = -(v/c)^2$ erhalten wir damit

$$\begin{aligned} E_{\mathrm{rel}} = m_0 c^2 f\big(-(v/c)^2\big) &= m_0 c^2 \cdot \left(\frac{v^2}{2c^2} + \frac{3v^4}{8c^4} + \mathrm{o}\Big(\frac{v^6}{c^6}\Big)\right) \\ &= E_{\mathrm{kin}} + m_0 \frac{3v^4}{8c^2} + m_0 \mathrm{o}\Big(\frac{v^6}{c^4}\Big) \end{aligned}$$

für $v \to 0$. Für kleine Geschwindigkeiten ist der Unterschied zwischen den beiden Energien also im Wesentlichen $m_0 \frac{3v^4}{8c^2}$, und selbst dieser Unterschied ist sehr klein.

BEISPIEL: RELATIVISTISCHE KINETISCHE ENERGIE

Wird statt des Taylorpolynoms zweiten Grades nur das ersten Grades betrachtet, so ergibt sich analog

$$E_{\mathrm{rel}} = E_{\mathrm{kin}} + m_0 \mathrm{o}\left(\frac{v^4}{c^2}\right).$$

Da Taylorpolynome ersten Grades affin lineare Funktionen sind, kann daher die Newton'sche Energie E_{kin} als lineare Approximation der relativistischen Energie E_{rel} aufgefasst werden, falls die Geschwindigkeit v klein ist.

Section 8.2
Eigenschaften der gleichmäßigen Konvergenz

VOLLSTÄNDIGKEIT VON $\rfloor([a,b])$

In Abschnitt 324 hatten die Supremumsnorm

$$\|f\|_\infty := \sup_{x\in[a,b]} |f(x)|$$

für stetige Funktionen $f : [a,b] \to \mathbb{R}$, d.h. für $f \in \mathcal{C}([a,b])$ eingeführt. Der folgende Satz zeigt, dass $\mathcal{C}([a,b])$ mit dieser Norm vollständig ist.

Theorem 8.2.1

Sei $(f_n)_{n\in\mathbb{N}} \subset \mathcal{C}([a,b])$ eine Cauchy-Folge bezüglich $\|\cdot\|_\infty$. Dann gibt es ein $f \in \mathcal{C}([a,b])$ mit $f_n \to f$ gleichmäßig.

BEWEIS

Sei $\varepsilon > 0$. Da $(f_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge bezüglich $\|\cdot\|_\infty$ ist, gibt es dann ein $n_0 \geq 1$, so dass

$$\|f_n - f_m\|_\infty \leq \varepsilon$$

für alle $n, m \geq n_0$ gilt. Insbesondere ist dann für jedes $x \in [a, b]$ die Folge $(f_n(x))_{n\geq 1}$ eine Cauchy-Folge, und damit konvergiert sie wegen der Vollständigkeit von $\mathbb{R}$ gegen eine reelle Zahl, die wir mit $f(x)$ bezeichnen. Insgesamt erhalten wir damit eine Funktion $f : [a, b] \to \mathbb{R}$, so dass $f_n \to f$ punktweise.

Sei nun $x \in [a, b]$. Mit dem eben Gesagten gibt es dann ein $n'_x \geq 1$ mit $|f_n(x) - f(x)| \leq \varepsilon$ für alle $n \geq n'_x$. Ohne Einschränkung können wir nun $n'_x \geq n_0$ annehmen. Dann folgt für $n \geq n_0$

$$|f_n(x) - f(x)| \leq |f_n(x) - f_{n'_x}(x)| + |f_{n'_x}(x) - f_n(x)| \leq 2\varepsilon$$

und damit haben wir $\|f_n - f\|_\infty \leq 2\varepsilon$ für $n \geq n_0$. Mit anderen Worten konvergiert $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen f und Satz 5.2.9 zeigt dann, dass f auch stetig ist.

In gewissem Sinne verallgemeinert die folgende Definition die gleichmäßige Konvergenz auf offene Intervalle (a, b).

Definition 8.2.2

Sei $(f_n)_{n\in\mathbb{N}}$ eine Folge von Funktionen $f_n : (a, b) \to \mathbb{R}$ und $f : (a, b) \to \mathbb{R}$. Dann **konvergiert** $(f_n)_{n\in\mathbb{N}}$ **kompakt** gegen f, falls für jedes abgeschlossene Teilintervall $I \subset (a, b)$ gilt:

$$\lim_{n\to\infty} \sup_{x\in I} \left|f_n(x) - f(x)\right| = 0\,.$$

Da es zu jedem $x \in (a, b)$ ein abgeschlossene Teilintervall $I \subset (a, b)$ mit $x \in I$ gibt, muss die Funktion f in der obigen Definition nach Satz 5.2.9 stetig sein.

Theorem 8.2.3

Sei $(f_n)_{n\in\mathbb{N}}$ eine Folge von differenzierbaren Funktionen $f_n : (a,b) \to \mathbb{R}$, für die die Folge $(f_n')_{n\in\mathbb{N}}$ kompakt konvergiert. Gibt es ein $x_0 \in (a,b)$, für das $(f_n(x_0))_{n\geq 1}$ konvergiert, dann konvergiert die Folge $(f_n)_{n\in\mathbb{N}}$ kompakt gegen eine differenzierbare Funktion $f : (a,b) \to \mathbb{R}$ und es gilt

$$f'(x) = \lim_{n\to\infty} f_n'(x)\,, \qquad\qquad x \in (a,b).$$

GEGENBEISPIEL

Bevor wir den Satz 8.2.3 beweisen, wollen wir zunächst darauf hinweisen, dass auf die gleichmäßige Konvergenz der Ableitungen nicht verzichtet werden kann. Hierzu betrachten wir die Funktionen

$$f_n(x) := \frac{\sin(n^2 x)}{n}, \qquad x \in (-1, 1).$$

Für $f := 0$ gilt dann $f_n \to f$ kompakt. Es ist aber auch $f_n'(x) = n \cos(n^2 x)$ und für z.B. $x = 0$ gilt deswegen $f_n'(0) = n \not\to 0 = f'(0)$.

Setzt man $x_m := 1/m$ und

$$a_{n,m} := \frac{f_n(x_m) - f_n(0)}{x_m - 0},$$

so gilt

$$\lim_{m \to \infty} a_{n,m} = f_n'(0) = n$$

und $\lim_{n \to \infty} a_{n,m} = f(x_m) = 0$ und damit

$$\lim_{n \to \infty} \lim_{m \to \infty} a_{n,m} \neq \lim_{m \to \infty} \lim_{n \to \infty} a_{n,m}.$$

BEMERKUNG

Im folgenden Beweis wird zunächst gezeigt, dass die Folge $(f_n)_{n\in\mathbb{N}}$ gegen ein $f \in \mathcal{C}([a,b])$ kompakt konvergiert. Da in unseren späteren Anwendungen diese Konvergenz schon bekannt ist, kann dieser erste Teil des Beweises auch übersprungen werden.

BEWEIS 1/4

Wir zeigen zunächst, dass die Folge $(f_n)_{n\in\mathbb{N}}$ kompakt konvergiert. Sei dazu $[c,d] \subset (a,b)$ und $\varepsilon > 0$. Ohne Einschränkung können wir dabei das Intervall $[c,d]$ so groß wählen, dass $x_0 \in [c,d]$ gilt. Gemäß unseren Annahmen gibt es dann ein $n_0 \geq 1$, so dass

$$|f_n(x_0) - f_m(x_0)| \leq \varepsilon\,,$$
$$|f_n'(\xi) - f_m'(\xi)| \leq \frac{\varepsilon}{d-c}$$

für alle $\xi \in [c,d]$ und $n,m \geq n_0$ gilt. Sei nun $x \in [c,d]$ und $n,m \geq n_0$. Im Fall $x_0 < x$ ergibt dann der Mittelwertsatz 6.2.3 auf $h := f_n - f_m$ angewendet ein $\xi_{n,m} \in [x_0,x] \subset [c,d]$ mit

$$h'(\xi_{n,m}) = \frac{h(x) - h(x_0)}{x - x_0}\,,$$

wobei wir ohne Einschränkung $x_0 < x$ angenommen haben.

Dies ergibt

$$\begin{aligned}
\left|f_n(x) - f_m(x)\right| &= \left|f_n(x) - f_m(x) - f_n(x_0) + f_m(x_0) + f_n(x_0) - f_m(x_0)\right| \\
&\leq \left|h(x) - h(x_0)\right| + \left|f_n(x_0) - f_m(x_0)\right| \\
&\leq |x - x_0| \cdot |h'(\xi_{n,m})| + \varepsilon \\
&= |x - x_0| \cdot \left|f_n'(\xi_{n,m}) - f_m'(\xi_{n,m})\right| + \varepsilon \\
&\leq 2\varepsilon\,,
\end{aligned}$$

wobei wir im letzten Schritt $|x - x_0| \leq |d - c|$ ausgenutzt haben. Eine analoge Rechnung gilt für $x_0 > x$ und im Fall $x = x_0$ gilt die Abschätzung sowieso.

Damit ist die Folge $(f_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge in $\mathcal{C}([c,d])$ und nach Satz 8.2.1 konvergiert sie gleichmäßig gegen ein $f_{c,d} \in \mathcal{C}([c,d])$. Insgesamt haben wir somit ein stetiges $f : (a,b) \to \mathbb{R}$ gefunden mit $f_n \to f$ kompakt.

BEWEIS 3/4

Um die Differenzierbarkeit von f zu zeigen, fixieren wir ein $x_0 \in (a, b)$ und ein Intervall $[c, d] \subset (a, b)$ mit $x_0 \in [c, d]$. Ferner setzen wir

$$g_n(x) := \begin{cases} \frac{f_n(x) - f_n(x_0)}{x - x_0}, & \text{falls } x \neq x_0 \\ f_n'(x_0), & \text{falls } x = x_0 \end{cases}$$

für alle $x \in [c, d]$. Da f_n differenzierbar und damit auch stetig ist, ist die Funktion $g_n : [c, d] \to \mathbb{R}$ stetig.

Wir wollen nun zeigen, dass $(g_n)_{n \in \mathbb{N}}$ eine Cauchy-Folge bezüglich $\|\cdot\|_\infty$ ist. Sei dazu $x \in [c, d]$. Im Fall $x_0 < x$ erhalten wir dann wieder durch Anwendung des Mittelwertsatzes 6.2.3 auf $h := f_n - f_m$ ein $\xi_{n,m} \in [x_0, x]$ mit

$$\begin{aligned} |h'(\xi_{n,m})| &= \left| \frac{h(x) - h(x_0)}{x - x_0} \right| \\ &= \left| \frac{f_n(x) - f_m(x) - f_n(x_0) + f_m(x_1)}{x - x_0} \right| \\ &= |g_n(x) - g_m(x)| . \end{aligned}$$

BEWEIS 4/4

Der Fall $x_0 > x$ ist analog zu behandeln, und im Fall $x = x_0$ gilt die Gleichung für $\xi_{n,m} := x_0$ nach Definition von g. Dies ergibt insgesamt

$$\|g_n - g_m\|_\infty \leq \sup_{\xi \in [c,d]} |h'(\xi)| = \sup_{\xi \in [c,d]} |f_n'(\xi) - f_m'(\xi)| = \|f_n' - f_m'\|_\infty \,.$$

Da $(f_n')_{n\in\mathbb{N}}$ eine Cauchy-Folge bezüglich $\|\cdot\|_\infty$ auf $[c,d]$ ist, ist es auch $(g_n)_{n\in\mathbb{N}}$. Nach Satz 8.2.1 gibt es dann ein $g \in \mathcal{C}([c,d])$ mit $\|g_n - g\|_\infty \to 0$.

Für $x \neq x_0$ gilt nun

$$g(x) = \lim_{n\to\infty} g_n(x) = \lim_{n\to\infty} \frac{f_n(x) - f_n(x_0)}{x - x_0} = \frac{f(x) - f(x_0)}{x - x_0}$$

Da g stetig ist, ergibt dies

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x\to x_0} g(x) = g(x_0)$$

Damit ist f in x_0 differenzierbar mit $f'(x_0) = g(x_0)$. Schließlich gilt

$$\lim_{n\to\infty} f_n'(x_0) \lim_{n\to\infty} g_n'(x_0) = g(x_0) = f'(x_0)\,,$$

wobei wir im ersten Schritt die Definition von $g_n'(x_0)$ ausgenutzt haben.

Section 8.3
Potenzreihen

Zu einer hinreichend differenzierbaren Funktion f hatten wir die Taylorpolynome durch

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \qquad x \in \mathbb{R}$$

definiert. Für $n \to \infty$ ergibt dies die formalen Taylorreihen

$$T_\infty(x) := \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

Im folgenden wollen wir ihr Konvergenzverhalten untersuchen. Dazu betrachten wir Potenzreihen

$$\sum_{k=0}^{\infty} a_k \cdot (x - x_0)^k, \tag{8.3.1}$$

wobei $(a_k)_{k\in\mathbb{N}}$ eine reelle Folge ist. Die folgenden Untersuchungen können aber leicht auf den komplexen Fall übertragen werden.

Der folgende Satz zeigt, dass Potenzreihen in vielen Fällen kompakt konvergieren.

Theorem 8.3.1

Sei $(a_k)_{k\in\mathbb{N}}$ eine reelle Folge und $x_0 \in \mathbb{R}$. Dann gibt es ein eindeutiges $r := r_{x_0} \in [0, \infty]$, so dass die Potenzreihe (8.3.1) *für alle $x \in U(x_0, r)$ absolut konvergiert und für alle $x \notin B(x_0, r)$ divergiert. Ferner gilt*

$$r = \frac{1}{\limsup_{k\to\infty} |a_k|^{1/k}} \,. \tag{8.3.2}$$

Schließlich konvergiert die Potenzreihe (8.3.1) *gleichmäßig auf $B(x_0, \delta)$ für alle $0 < \delta < r$.*

Der Wert r heißt Konvergenzradius und die Formel (8.3.2) wird als Formel von Cauchy-Hadamard bezeichnet.

BEWEIS 1/3

Wir setzen

$$r := \sup\left\{ |w| : \sum_{k=0}^{\infty} a_k w^k \text{ konvergiert} \right\}.$$

Für $x \notin B(x_0, r)$, d.h. $|x - x_0| > r$, konvergiert dann die Potenzreihe (8.3.1) nicht.

Ist $r = 0$, so ist $U(x_0, r) \neq \emptyset$ und die Konvergenz der Potenzreihe ist nicht zu zeigen. Im Fall $r > 0$ wählen wir ein $0 < \delta < r$. Dann gibt es ein w mit $\delta < |w| < r$, so dass

$$\sum_{k=0}^{\infty} a_k w^k$$

konvergiert. Insbesondere ist dann die Folge $(a_k w^k)_{k \geq 0}$ beschränkt, d.h. es gibt ein $M > 0$ mit $|a_k w^k| \leq M$ für alle $k \geq 0$. Für $x \in B(x_0, \delta)$ folgt

$$\left| a_k (x - x_0)^k \right| = |a_k w^k| \cdot \left| \frac{x - x_0}{w} \right|^k \leq M \cdot \left| \frac{\delta}{w} \right|^k.$$

Wegen $q := \frac{\delta}{|w|} < 1$ folgt

$$\sum_{k=m}^{n} |a_k (x - x_0)^k| \leq \sum_{k=m}^{n} M q^k \leq M \cdot \sum_{k=m}^{\infty} q^k = M \cdot \frac{q^m}{1 - q} \to 0$$

für $m \to \infty$.

BEWEIS 2/3

Damit konvergiert die Potenzreihe für das gewählte x absolut und nach Konstruktion gilt dies für dann für alle $x \in U(x_0, r)$. Eine analoge Rechnung liefert zudem

$$\begin{aligned}\sup_{x\in B(x_0,\delta)}\left|\sum_{k=0}^{n} a_k(x-x_0)^k - \sum_{k=0}^{\infty} a_k(x-x_0)^k\right| &\leq \sup_{x\in B(x_0,\delta)} \sum_{k=n+1}^{\infty} |a_k(x-x_0)^k| \\ &\leq M \cdot \frac{q^{n+1}}{1-q}\end{aligned}$$

und damit konvergiert die Potenzreihe auf $B(x_0, \delta)$ auch gleichmäßig.

Um (8.3.2) zu beweisen, definieren wir

$$R := \frac{1}{\limsup_{k\to\infty} |a_k|^{1/k}} .$$

Nach dem Wurzelkriterium, siehe Satz 4.2.9, konvergiert die Potenzreihe (8.3.1) absolut, falls

$$R^{-1}|x-x_0| = \limsup_{k\to\infty} |a_k|^{1/k}|x-x_0| = \limsup_{k\to\infty} |a_k(x-x_0)^k|^{1/k} < 1 .$$

Dies liefert $|x - x_0| < R$ und damit $r \geq R$.

BEWEIS 3/3

Wäre nun $r > R$, d.h. $\limsup_{k\to\infty} |a_k|^{1/k} > 1/r$, so gäbe es ein $\varepsilon > 0$ und eine Teilfolge $(k_n)_{n\in\mathbb{N}}$ mit $|a_{k_n}|^{1/k_n} > 1/r + \varepsilon$ für alle $n \geq 1$. Für

$$x := (1/r + \varepsilon)^{-1} + x_0 = \frac{r}{r + r\varepsilon} + x_0$$

gilt dann

$$|a_{k_n}||x - x_0|^{k_n} = |a_{k_n}| \cdot \left(\frac{r}{r + r\varepsilon}\right)^{k_n} \geq (1/r + \varepsilon)^{k_n} \cdot (1/r + \varepsilon)^{-k_n} = 1\,.$$

Damit ist $(a_k(x - x_0)^k)_{k\geq 1}$ keine Nullfolge, und die Potenzreihe konvergiert nicht für x. Wegen $|x - x_0| = \frac{r}{r+r\varepsilon} < r$ steht dies aber zum Widerspruch zu der Definition von r.

Leitet man alle Summanden der Potenzreihe (8.3.1) ab, so ergibt sich

$$\sum_{k=1}^{\infty} a_k \cdot k \cdot (x - x_0)^{k-1} = \sum_{k=0}^{\infty} a_{k+1} \cdot (k+1)(x - x_0)^k .$$

Insbesondere ist die resultierende Reihe wieder eine Potenzreihe. Der folgende Satz zeigt nun, dass diese “abgeleitete” Potenzreihe den gleichen Konvergenzradius wie die ursprüngliche Potenzreihe hat.

Theorem 8.3.2

Sei $(a_k)_{k\in\mathbb{N}}$ eine reelle Folge und $x_0 \in \mathbb{R}$. Ferner sei r der Konvergenzradius der zugehörigen Potenzreihe (8.3.1)*. Dann ist r auch der Konvergenzradius der Potenzreihe*

$$\sum_{k=1}^{\infty} a_k \cdot k(x - x_0)^{k-1} . \tag{8.3.3}$$

Mit der 2. Regel von L'Hospital, siehe Satz 6.2.13, gilt

$$\lim_{k\to\infty} \ln(1+k)^{1/k} = \lim_{k\to\infty} \frac{\ln(1+k)}{k} = \lim_{k\to\infty} \frac{1}{1+k} = 0$$

und damit folgt $|1+k|^{1/k} \to 1$. Für $b_k := a_{k+1}\cdot(k+1)$ und $c_k := |a_{k+1}|^{\frac{1}{k+1}}$ ergibt diese Konvergenz dann

$$\begin{aligned}\limsup_{k\to\infty} |b_k|^{1/k} = \limsup_{k\to\infty} |a_{k+1}|^{1/k}\cdot|1+k|^{1/k} &= \limsup_{k\to\infty} |a_{k+1}|^{\frac{1}{k+1}\cdot\frac{k+1}{k}} \\ &= \limsup_{k\to\infty} c_k^{\frac{k+1}{k}} \\ &= \limsup_{k\to\infty} \exp\Big(\frac{k+1}{k}\cdot\ln c_k\Big) \\ &= \exp(\ln r^{-1}) = r^{-1},\end{aligned}$$

wobei wir Cauchy-Hadamard Formel (8.3.2) ausgenutzt haben. Damit ist der Konvergenzradius der Potenzreihe (8.3.3) gleich r nach Satz 8.3.1.

Mit diesen Vorbereitungen sind wir nun in der Lage, Potenzreihen ableiten zu können.

Korollar 8.3.3

Sei $(a_k)_{k\in\mathbb{N}}$ eine reelle Folge und $x_0 \in \mathbb{R}$. Ferner sei r der Konvergenzradius der zugehörigen Potenzreihe (8.3.1)*. Dann ist die durch*

$$f(x) := \sum_{k=0}^{\infty} a_k \cdot (x - x_0)^k, \qquad x \in U(x_0, r) \tag{8.3.4}$$

definierte Funktion unendlich oft differenzierbar und es gilt

$$f'(x) = \sum_{k=1}^{\infty} a_k \cdot k(x - x_0)^{k-1}, \qquad x \in U(x_0, r).$$

Für $n \geq 1$ setzen wir

$$f_n(x) := \sum_{k=0}^{n} a_k \cdot (x - x_0)^k, \qquad x \in U(x_0, r).$$

Nach Satz 8.3.1 konvergiert dann die Folge $(f_n)_{n\in\mathbb{N}}$ kompakt gegen f. Ferner zeigt Satz 8.3.2, dass der Konvergenzradius der formal abgeleiteten Potenzreihe (8.3.3) ebenfalls r ist. Nach Satz 8.3.1 konvergiert damit auch die Folge der Ableitungen $(f_n')_{n\in\mathbb{N}}$ kompakt. Satz 8.2.3 zeigt dann, dass f differenzierbar ist und

$$f'(x) = \lim_{n\to\infty} f_n'(x) = \sum_{k=1}^{\infty} a_k \cdot k(x - x_0)^{k-1}$$

gilt. Anwenden der bisher bewiesenen Aussage auf f' zeigt, dass auch f' differenzierbar ist, d.h. f zweimal differenzierbar ist. Durch wiederholtes Anwenden sehen wir schließlich, dass f unendlich oft differenzierbar ist.

Das folgende Korollar zeigt, dass Potenzreihen die Taylorreihen ihrer zugehörigen Funktion sind. Ferner zeigt es, dass die Koeffizienten-Folge $(a_k)_{k\in\mathbb{N}}$ einer Potenzreihe eindeutig durch die zugehörige Funktion bestimmt ist.

Korollar 8.3.4

Sei $(a_k)_{k\in\mathbb{N}}$ eine reelle Folge und $x_0 \in \mathbb{R}$. Ferner sei r der Konvergenzradius der zugehörigen Potenzreihe (8.3.1) *und $f : U(x_0, r) \to \mathbb{R}$ die zugehörige Funktion* (8.3.4)*. Dann gilt*

$$a_k = \frac{f^{(k)}(x_0)}{k!}, \qquad k \geq 0.$$

Beweis.

Für $k = 0$ folgt dies aus $f(x_0) = a_0(x_0 - x_0)^0 = a_0$. Für $k = 1$ folgt dies analog aus $f'(x_0) = a_1 \cdot 1 \cdot (x_0 - x_0)^{1-1} = a_1$. Für $k \geq 2$ ist analog $f^{(k)}(x_0)$ zu betrachten. □

BEMERKUNG

Jedes Polynom ist offensichtlich die von einer Potenzreihe um $x_0 = 0$ mit Konvergenzradius $r = \infty$ dargestellte Funktion. Haben wir daher zwei Polynome p und q und $a < b$ mit $p(x) = q(x)$ für alle $x \in (a, b)$, so zeigt das obige Korollar, dass $\deg p = \deg q$ ist und die Koeffizienten von p und q übereinstimmen. Diese Aussage hätten wir allerdings auch schon im Anschluss von Korollar 6.2.5 beweisen können. Ferner zeigt diese Aussage, dass es in Kapitel 7 unnötig war, potentiell verschiedene Darstellungen von Polynomen zu betrachten.

NICHT JEDES f KANN DURCH TAYLORREIHE DARGESTELLT WERDEN

Korollar 8.3.4 zeigt, dass Potenzreihen die Taylorreihen ihrer dargestellten Funktion f sind. In diesem Fall konvergiert die Taylorreihe naturgemäß gegen f.

Diese Beobachtung wirft die Frage auf, ob wir auch mit einer Funktion f beginnen können und diese durch ihre Taylorreihe um x_0 approximieren können. Offensichtlich muss dazu f in x_0 unendlich oft differenzierbar sein, da wir ansonsten die Taylorkoeffizienten nicht bestimmen können. Dies reicht aber nicht, wie das folgende Beispiel zeigt:

NICHT JEDES f KANN DURCH TAYLORREIHE DARGESTELLT WERDEN

Dazu betrachten wir die Funktion $f : \mathbb{R} \to \mathbb{R}$, die durch

$$f(x) := \begin{cases} 0\,, & \text{falls } x = 0 \\ \exp(-1/x^2)\,, & \text{falls } x \neq 0 \end{cases}$$

definiert ist. Nachrechnen ergibt $f^{(k)}(0) = 0$ für alle $k \geq 0$. Damit ist die um 0 entwickelte Taylorreihe

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = 0\,, \qquad x \in \mathbb{R}.$$

Offensichtlich hat diese Potenzreihe den Konvergenzradius $r = \infty$ und es gilt trotzdem $T_\infty(x) \neq f(x)$ für alle $x \neq 0$.

Funktionen, die (lokal) durch ihre Taylorreihe darstellbar sind, gehören zu den “schönsten” Funktionen und werden in der Funktionentheorie behandelt. Sie haben viele weitere Eigenschaften, die sie von anderen Funktionen abheben.

Die geometrische Reihe

$$\sum_{k=0}^{\infty} x^k$$

ist eine Potenzreihe um $x_0 = 0$ zu der Folge $a_k = 1$ für alle $k \geq 0$. Nach der Cauchy-Hadamard-Formel (8.3.2) ist ihr Konvergenzradius $r = 1$. Außerdem wissen wir, dass ihre dargestellte Funktion f durch

$$f(x) = \frac{1}{1-x}, \qquad x \in (-1, 1)$$

gegeben ist. Ableiten ergibt nach Korollar 8.3.3

$$\frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} k x^{k-1}, \qquad x \in (-1, 1).$$

Die Exponentialfunktion

$$\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!}, \qquad x \in \mathbb{R}$$

ist durch eine Potenzreihe um $x_0 = 0$ zu der Folge $a_k = 1/k!$ für alle $k \geq 0$ definiert. Wegen $\sqrt[k]{k!} \to \infty$ hat sie Konvergenzradius $r = \infty$ und Ableiten nach Korollar 8.3.3 ergibt die wohlbekannte Formel

$$\exp'(x) = \sum_{k=1}^{\infty} \frac{k}{k!} \cdot x^{k-1} = \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \cdot x^{k-1} = \sum_{k=0}^{\infty} \frac{x^k}{k!} = \exp(x)$$

für alle $x \in \mathbb{R}$.

BEISPIELE: WINKELFUNKTIONEN

Es gilt $\sin 0 = 0$, $\sin' 0 = 1$, $\sin'' 0 = 0$, $\sin''' 0 = -1$, usw.. Insgesamt verschwinden damit alle geraden Ableitungen von $\sin$ in 0, d.h.

$$\sin^{(2k)}(0) = 0\,, \qquad\qquad k \geq 0\,,$$

während die ungeraden Ableitungen alternieren, genauer

$$\sin^{(2k+1)}(0) = (-1)^k\,, \qquad\qquad k \geq 0\,,$$

Dies ergibt die Taylorreihe

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}\,,$$

die wiederum den Konvergenzradius $r = \infty$ hat. Analog sehen wir, dass

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}\,,$$

die Taylorreihe der Kosinusfunktion ist.

BEISPIELE: WINKELFUNKTIONEN

Definieren wir die Sinus- und Kosinusfunktion nicht geometrisch, sondern analytisch durch

$$\sin x := \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

$$\cos x := \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}$$

für alle $x \in \mathbb{R}$, so ergibt Ableiten nach Korollar 8.3.3

$$\sin' x = \sum_{k=0}^{\infty} (-1)^k (2k+1) \frac{x^{2k}}{(2k+1)!} = \cos x\,.$$

Analog finden wir $\cos x = -\sin x$.

Viele andere wichtige Funktionen können durch ihre Taylorreihe dargestellt werden, z.B.:

$$\ln(1+x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} x^{k+1},$$

$$\arctan x := \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}$$

wobei wir in beiden Fällen um $x_0 = 0$ entwickelt haben und den Konvergenzradius $r = 1$ haben. Mit anderen Worten gelten die beiden Formeln für alle $x \in (-1, 1)$.

Chapter 9: Lineare Abbildungen

In Definition 3.1.1 hatten wir schon Vektorräume über $\mathbb{R}$ bzw. über $\mathbb{C}$ kennengelernt. Die einfachsten Beispiele solcher Vektorräume waren dabei die Räume $\mathbb{R}^d$ und $\mathbb{C}^d$. Es gibt aber auch deutlich anspruchsvollere Vektorräume, wie z.B. den Raum $\mathcal{C}([a,b])$ aller stetigen Funktionen $f : [a,b] \to \mathbb{R}$.

In diesem Kapitel wollen wir Vektorräume und ausgezeichnete Abbildungen zwischen diesen Räumen genauer untersuchen.

Section 9.1
Basen

Jeder Punkt im Raum $\mathbb{R}^2$ kann eindeutig durch die beiden
Standard-Einheitsvektoren

$$e_1 := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad e_2 := \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

beschrieben werden. Dies ist genau die Grundidee des kartesischen Koordinatensystems. Analog kann jeder Punkt im $\mathbb{R}^3$ durch die drei Standard-Einheitsvektoren

$$e_1 := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad e_2 := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad e_3 := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

beschrieben werden. In diesem Sinne kann man 2 bzw. 3 als die Dimension von $\mathbb{R}^2$ bzw. $\mathbb{R}^3$ ansehen. Diese Idee wollen wir im folgenden auf allgemeinere Vektorräume anwenden.

Im folgenden steht $\mathbb{K}$ immer für $\mathbb{R}$ oder $\mathbb{C}$, um die beiden Fälle gleichzeitig behandeln zu können.

Definition 9.1.1

Seien $(V, +, \cdot)$ ein $\mathbb{K}$-Vektorraum und $W \subset V$. Dann heißt W Unterraum von V, falls $(W, +, \cdot)$ ein $\mathbb{K}$-Vektorraum ist, wobei $+$ und $\cdot$ jeweils auf W eingeschränkt worden sind.

Das folgende Lemma stellt eine einfache Charakterisierung von Unterräumen zur Verfügung.

Lemma 9.1.2

Seien $(V, +, \cdot)$ ein $\mathbb{K}$-Vektorraum und $W \subset V$. Dann ist W genau dann ein Unterraum von V, wenn die folgenden drei Bedingungen erfüllt sind.

i). $0 \in W$

ii). *W ist **abgeschlossen bzgl. der Addition**, d.h. für alle $v, w \in W$ gilt $v + w \in W$.*

iii). *W ist **abgeschlossen bzgl. der Skalarmultiplikation**, d.h. für alle $v \in W$ und $\lambda \in \mathbb{K}$ gilt $\lambda v \in W$.*

In diesem Fall ist 0 auch das neutrale Element von $+$ in W und die inversen Elemente der Addition in W stimmen mit denen von V überein.

BEWEIS

Sei zunächst W ein Unterraum. Nach Definition ist dann $(W, +, \cdot)$ ein $\mathbb{K}$-Vektorraum ist, wobei $+$ und $\cdot$ jeweils auf W eingeschränkt worden sind. Mit anderen Worten bildet die Einschränkung $+_{|W \times W} : W \times W \to V$ sogar nach W ab. Dies zeigt *ii)*. Die Bedingung *iii)* folgt analog. Schließlich kann W nichtleer sein, da es immer ein neutrales Element 0_W enthalten muss. Mit *iii)* folgt $0_V = 0 \cdot 0_W \in W$, wobei $0_V := 0$ das neutrale Element in V bezeichnet. Ferner gilt auch $0 \cdot 0_W = 0_W$ und damit haben wir $0_V = 0_W$

Seien nun umgekehrt *i)* bis *iii)* erfüllt. Nach *ii)* gilt dann $+_{|W \times W} : W \times W \to W$ und *iii)* zeigt $\cdot_{|\mathbb{K} \times W} : \mathbb{K} \times W \to W$. Offensichtlich ist $+$ auch auf W assoziativ und kommutativ und da $0 \in W$ ist, hat $+$ auch ein neutrales Element in W. Schließlich gilt $-v = (-1) \cdot v \in W$ für alle $v \in W$ nach *iii)* und damit hat auch jedes Element in W ein inverses bzgl. $+$. Die übrigen 4 Identitäten gelten in V und damit auch in W.

Im $\mathbb{R}^2$ ist eine Gerade $g := \{x_0 + tv : t \in \mathbb{R}\}$ mit Stützvektor x_0 und Richtungsvektor v genau dann ein Unterraum von $\mathbb{R}^2$ falls $0 \in g$.

Um dies zu sehen, nehmen wir zunächst an, dass g ein Unterraum ist. Dann zeigt Lemma 9.1.2 sofort $0 \in g$. Gilt umgekehrt $0 \in g$, so muss es ein $t^* \in \mathbb{R}$ geben mit $0 = x_0 + t^*v$. Dies zeigt $x_0 = -t^*v$ und damit hat g die Gestalt

$$g := \{-t^*v + tv : t \in \mathbb{R}\} = \{(t - t^*)v : t \in \mathbb{R}\} = \{sv : s \in \mathbb{R}\}.$$

Sind nun $u_1, u_2 \in g$, so gibt es $s_1, s_2 \in \mathbb{R}$ mit $u_1 = s_1v$ und $u_2 = s_2v$. Damit folgt $u_1 + u_2 = (s_1 + s_2)v \in g$. Für $\lambda \in \mathbb{R}$ gilt ferner $\lambda(s_1v) = (\lambda s_1)v \in g$. Damit zeigt Lemma 9.1.2, dass g ein Unterraum ist.

Um weitere Beispiele von Unterräumen zu betrachten, fixieren wir eine nichtleere Menge X und einen $\mathbb{K}$-Vektorraum V. Wir schreiben

$$\mathrm{Abb}(X, V) := \{f : X \to V\}$$

für die Menge aller Abbildungen von X nach V. Dann ist $\mathrm{Abb}(X, V)$ ein Vektorraum, wenn wir die Addition und Skalarmultiplikation wie üblich punktweise definieren, d.h.

$$\begin{aligned}(f + g)(x) &:= f(x) + g(x)\,,\\ (\lambda f)(x) &:= \lambda f(x)\end{aligned}$$

für alle $f, g \in \mathrm{Abb}(X, V)$, $\lambda \in \mathbb{K}$ und $x \in X$. Die konstante Nullfunktion ist dabei z.B. das neutrale Element der Addition. Mit Hilfe diese Raums können wir nun weitere Vektorräume als Unterräume von z.B. $\mathrm{Abb}(X, \mathbb{R})$ identifizieren, wobei die Argumente auch für $\mathbb{C}$ gültig sind.

Ein erstes Beispiel ist der Raum $\mathcal{C}([a,b])$. In der Tat ist die konstante Nullfunktion stetig und damit in $\mathcal{C}([a,b])$. Ferner ist die Summe zweier stetiger Funktionen wieder stetig, und damit ist die von $\mathrm{Abb}([a,b],\mathbb{R})$ stammende punktweise Addition in $\mathcal{C}([a,b])$ abgeschlossen. Das gleiche gilt für die Skalarmultiplikation und damit ist nach Lemma 9.1.2 der Raum $\mathcal{C}([a,b])$ ein Unterraum von $\mathrm{Abb}([a,b],\mathbb{R})$.

Ein weiteres Beispiel ist der Raum aller reellen Polynome auf einen Intervall I, d.h.

$$\mathrm{Pol}(I) := \left\{ p : I \to \mathbb{R} \middle| \; p \text{ ist Polynom } \right\}.$$

Wie im ersten Beispiel sieht man dann wieder mit Lemma 9.1.2, dass $\mathrm{Pol}(I)$ ein Unterraum von $\mathrm{Abb}(I, \mathbb{R})$ ist. Betrachtet man ferner die Menge

$$\mathrm{Pol}(I, n) := \left\{ p : I \to \mathbb{R} \middle| \; p \text{ ist Polynom mit } \mathbf{deg}\, p \leq n \right\},$$

so zeigt (2.7.2), dass für $p, q \in \mathrm{Pol}(I, n)$ auch $\mathbf{deg}(p + q) \leq n$, d.h. $p + q \in \mathrm{Pol}(I, n)$ gilt. Analog zeigt (2.7.3) mit $q = 1$, dass auch die Skalarmultiplikation auf $\mathrm{Pol}(I, n)$ abgeschlossen ist. Schließlich ist $0 \in \mathrm{Pol}(I, n)$ offensichtlich. Damit ist $\mathrm{Pol}(I, n)$ sowohl ein Unterraum von $\mathrm{Pol}(I,)$ als auch von $\mathrm{Abb}(I, \mathbb{R})$.

Letzteres ist keine Überraschung, denn ist W ein Unterraum von V und U ein Unterraum von W, so kann man mit Lemma 9.1.2 leicht zeigen, dass dann auch U ein Unterraum von V ist.

Im folgenden sei V ein $\mathbb{K}$-Vektorraum für $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Für $v_1, \ldots, v_m \in V$ und $\lambda_1, \ldots, \lambda_m \in \mathbb{K}$ heißt

$$v := \sum_{i=1}^{m} \lambda_i v_i$$

Linearkombination von $v_1, \ldots, v_m \in V$. Ferner heißt die Menge aller solcher Linearkombinationen, d.h.

$$\mathbf{span}\{v_1, \ldots, v_m\} := \left\{ \sum_{i=1}^{m} \lambda_i v_i : \lambda_1, \ldots, \lambda_m \in \mathbb{K} \right\}$$

der von $v_1, \ldots, v_m \in V$ aufgespannte Raum. Mit Lemma 9.1.2 ist es leicht zu überprüfen, dass $W := \mathbf{span}\{v_1, \ldots, v_m\}$ ein Unterraum von V ist. In diesem Fall heißt $v_1, \ldots, v_m$ ein Erzeugendensystem von W.

Um die Begriffe zu illustrieren, betrachten wir die Vektoren

$$v_1 := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad v_2 := \begin{pmatrix} -2 \\ 0 \end{pmatrix}.$$

Dann gilt $\mathbf{span}\{v_1, v_2\} = \{\lambda v_1 : \lambda \in \mathbb{R}\}$. Hierbei ist die Inklusion "$\supset$" wegen $\lambda v_1 = \lambda v_1 + 0 v_2$ leicht zu sehen. Haben wir umgekehrt $\lambda_1, \lambda_2 \in \mathbb{R}$, so folgt

$$\lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} -2 \\ 0 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} - 2\lambda_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = (\lambda_1 - 2\lambda_2) v_1 ,$$

und damit haben wir auch "$\subset$".

Betrachten wir stattdessen die Vektoren

$$w_1 := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad w_2 := \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

so gilt $\mathbf{span}\{w_1, w_2\} = \mathbb{R}^2$. Hierbei ist die Inklusion "$\subset$" offensichtlich. Haben wir umgekehrt ein $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$, so müssen wir $\lambda_1 \in \mathbb{R}$ und $\lambda_2 \in \mathbb{R}$ finden, die

$$\lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

erfüllen. Dies liefert die Gleichungen $\lambda_1 + \lambda_2 = x_1$ und $\lambda_2 = x_2$, was wiederum $\lambda_1 = x_1 - x_2$ ergibt.

Sind schließlich $e_1, \ldots, e_d$ die Standard-Einheitsvektoren von $\mathbb{R}^d$, so gilt

$$\mathbf{span}\{e_1, \ldots, e_d\} = \mathbb{R}^d,$$

siehe auch (3.1.1).

Um die Dimension sinnvoll definieren zu können, müssen wir Erzeugendensysteme betrachten, die nicht zu groß sind. Die folgende Definition führt ein entsprechende Konzept ein.

Definition 9.1.3

Seien V ein $\mathbb{K}$-Vektorraum und $v_1, \dots, v_m \in V$. Dann heißen diese Vektoren **linear unabhängig**, falls für alle $\lambda_1, \dots, \lambda_m \in \mathbb{K}$ mit

$$\sum_{i=1}^{m} \lambda_i v_i = 0$$

gilt $\lambda_1 = \cdots = \lambda_m = 0$. Ansonsten heißen die Vektoren **linear abhängig**.

Die obigen Vektoren v_1, v_2 sind linear abhängig, denn es gilt zum Beispiel $2v_1 + v_2 = 0$. Die Vektoren w_1, w_2 sind dagegen linear unabhängig, denn für $x := 0 \in \mathbb{R}^2$ hatte unsere obige Überlegung ja schon $\lambda_2 = x_2 = 0$ und $\lambda_1 = x_1 - x_2 = 0$ gezeigt. Schließlich sind die Vektoren $e_1, \dots, e_d$ in $\mathbb{R}^d$ ebenfalls linear unabhängig, wie man sich leicht klarmachen kann.

Lemma 9.1.4

Seien V ein $\mathbb{K}$-Vektorraum und $v_1, \ldots, v_m \in V$. Dann sind $v_1, \ldots, v_m \in V$ linear unabhängig genau dann, wenn es zu jedem $v \in \mathbf{span}\{v_1, \ldots, v_m\}$ eindeutige Koeffizienten $\lambda_1, \ldots, \lambda_m \in \mathbb{K}$ gibt mit

$$v = \sum_{i=1}^{m} \lambda_i v_i \,. \tag{9.1.1}$$

BEWEIS

Wir nehmen zunächst an, dass $v_1, \ldots, v_m \in V$ linear unabhängig sind. Ferner seien $\lambda_1, \ldots, \lambda_m \in \mathbb{K}$ mit (9.1.1) und $\alpha_1, \ldots, \alpha_m \in \mathbb{K}$ mit

$$v = \sum_{i=1}^{m} \alpha_i v_i$$

gegeben. Dies ergibt

$$0 = \sum_{i=1}^{m} \lambda_i v_i - \sum_{i=1}^{m} \alpha_i v_i = \sum_{i=1}^{m} (\lambda_i - \alpha_i) v_i \,.$$

Aus der linearen Unabhängigkeit folgt dann $\lambda_1 - \alpha_1 = \cdots = \lambda_m - \alpha_m = 0$ und damit auch $\lambda_1 = \alpha_1$, ..., $\lambda_m = \alpha_m$.

Gilt umgekehrt die eindeutige Darstellung in (9.1.1), so folgt sofort die lineare Unabhängigkeit durch Betrachtung von $v := 0$.

Die folgende Definition betrachtet linear unabhängige Erzeugendensysteme.

Definition 9.1.5

Seien V ein $\mathbb{K}$-Vektorraum und $v_1, \ldots, v_m \in V$. Dann ist $v_1, \ldots, v_m$ eine **Basis** von V, falls die Vektoren $v_1, \ldots, v_m$ linear unabhängig sind und $V = \mathbf{span}\{v_1, \ldots, v_m\}$ gilt.

Die obigen Vektoren w_1, w_2 sind eine Basis von $\mathbb{R}^2$ und das gleiche gilt für die beiden Standardeinheitsvektoren $e_1, e_2 \in \mathbb{R}^2$. Räume können also insbesondere mehrere Basen haben.

Ist $v_1, \ldots, v_m$ eine Basis von V, so hat Lemma 9.1.4 gezeigt, dass jedes Element $v \in V$ eine eindeutige Darstellung als Linearkombination von $v_1, \ldots, v_m$ hat.

Das folgende Lemma zeigt, dass wir aus einem linear abhängigen Erzeugendensystem ein Element hinausnehmen können, ohne den aufgespannten Raum zu ändern.

Lemma 9.1.6

Seien V ein $\mathbb{K}$-Vektorraum und $v_1, \ldots, v_m \in V$ linear abhängig mit $v_1 \neq 0$. Dann gibt es ein $j \in \{2, \ldots, m\}$ mit $v_j \in \mathbf{span}\{v_1, \ldots, v_{j-1}\}$ und

$$\mathbf{span}\{v_1, \ldots, v_{j-1}, v_{j+1}, v_m\} = \mathbf{span}\{v_1, \ldots, v_m\}\,.$$

Da $v_1, \dots, v_m$ linear abhängig sind, gibt es Koeffizienten $\lambda_1, \dots, \lambda_m \in \mathbb{K}$, die nicht alle gleich Null sind, mit

$$\lambda_1 v_1 + \cdots + \lambda_m v_m = 0\,.$$

Sei j der größte Index mit $\lambda_j \neq 0$. Dann ist $\lambda_{j+1} = \cdots = \lambda_m = 0$ und es folgt

$$\lambda_1 v_1 + \cdots + \lambda_j v_j = 0\,. \tag{9.1.2}$$

Wäre $j = 1$, so würde dies $\lambda_1 v_1 = 0$ bedeuten, was wegen $\lambda_1 \neq 0$ und $v_1 \neq 0$ unmöglich ist. Damit folgt $j \geq 2$. Aus (9.1.2) folgt dann

$$v_j = -\frac{\lambda_1}{\lambda_j} v_1 - \cdots - \frac{\lambda_{j-1}}{\lambda_j} v_{j-1}\,, \tag{9.1.3}$$

d.h. wir haben $v_j \in \mathbf{span}\{v_1, \dots, v_{j-1}\}$ gezeigt. Sei nun $v \in \mathbf{span}\{v_1, \dots, v_m\}$. Dann gibt es $\alpha_1, \dots, \alpha_m \in \mathbb{K}$ mit

$$v = \alpha_1 v_1 + \cdots + \alpha_m v_m\,.$$

Ersetzen wir nun den Vektor v_j auf der rechten Seite durch (9.1.3), so sehen wir, dass

$$v \in \mathbf{span}\{v_1, \dots, v_{j-1}, v_{j+1}, v_m\}\,.$$

Damit haben wir "$\subset$" gezeigt. Die andere Inklusion ist trivial.

Hat man linear abhängige Vektoren $v_1, \ldots, v_m \in V$ mit $v_1 \neq 0$, so zeigt Lemma 9.1.6, dass wir solange Vektoren wegnehmen können, ohne den aufgespannten Raum zu ändern, bis wir eine linear unabhängige Familie erhalten haben. Diese Familie ist dann eine Basis von $\mathbf{span}\{v_1, \ldots, v_m\}$.

Das folgende Lemma untersucht, wie sich die Erweiterung von Erzeugendensystemen auf die lineare Unabhängigkeit auswirkt.

Lemma 9.1.7

Seien V ein $\mathbb{K}$-Vektorraum, $v_1, \ldots, v_m \in V$ und $W := \mathbf{span}\{v_1, \ldots, v_m\}$. Dann gelten die folgenden Aussagen für $v \in V$:

i). Ist $v \in W$, dann sind $w, v_1, \ldots, v_m$ linear abhängig.

ii). Ist $v \notin W$ und sind $v_1, \ldots, v_m \in V$ linear unabhängig, dann sind auch $w, v_1, \ldots, v_m$ linear unabhängig.

i). Es gibt $\lambda_i \in \mathbb{K}$ mit $v = \lambda_1 v_1 + \cdots + \lambda_m v_m$. Dies impliziert

$$0 = (-1)v + \lambda_1 v_1 + \cdots + \lambda_m v_m \,,$$

und da $-1 \neq 0$ folgt die lineare Abhängigkeit.

ii). Sei $0 = \lambda v + \lambda_1 v_1 + \cdots + \lambda_m v_m$. Dann muss $\lambda = 0$ sein, denn ansonsten hätten wir

$$v = -\frac{\lambda_1}{\lambda} v_1 - \cdots - \frac{\lambda_m}{\lambda} v_m \,,$$

was $v \notin W$ widerspricht. Aus $\lambda = 0$ folgt nun $0 = \lambda_1 v_1 + \cdots + \lambda_m v_m$ und die lineare Unabhängigkeit der $v_1, \ldots, v_m \in V$ sichert deswegen
$\lambda_1 = \cdots = \lambda_m = 0$.

BASEN SIND MAXIMALE, LINEAR UNABHÄNGIGE EZS

Der folgende Satz zeigt, dass linear unabhängige Familien nur eine maximale Größe haben können, falls es ein endliches Erzeugendensystem gibt.

Theorem 9.1.8

Seien V ein $\mathbb{K}$-Vektorraum und $v_1, \ldots, v_m \in V$ ein Erzeugendensystem von V und $w_1, \ldots, w_n \in V$ linear unabhängig. Dann gilt $n \leq m$.

BEWEIS 1/2

Da $w_1, \ldots, w_n$ linear unabhängig sind, gilt $w_1 \neq 0$, ..., $w_n \neq 0$.

Aus $\mathbf{span}\{v_1, \ldots v_m\} = V$ folgt nun mit Lemma 9.1.7, dass $w_1, v_1, \ldots, v_m$ linear abhängig sind. Wegen $w_1 \neq 0$ finden wir mit Lemma 9.1.6 ein j, so dass für $I_1 := \{1, \ldots, m\} \setminus \{j\}$ gilt:

$$V = \mathbf{span}\Big(\{w_1\} \cup \{v_i : i \in I_1\}\Big).$$

Dieser Schritt wird nun iteriert. Vor dem $(j+1)$-ersten Schritt haben wir dabei bereits eine Menge $I_j := \{i_1, \ldots i_{m-j}\} \subset \{1, \ldots, m\}$ mit $|I_j| = m - j$ und

$$V = \mathbf{span}\Big(\{w_1, \ldots, w_j\} \cup \{v_i : i \in I_j\}\Big)$$

BEWEIS 2/2

Nach Lemma 9.1.7 sind dann $w_1, \ldots, w_{j+1}, v_{i_1}, \ldots v_{i_{m-j}}$ linear abhängig. Nach Lemma 9.1.6 gibt es dann einen Vektor u in dieser Liste, der in dem aufgespannten Raum der vorher aufgezählten Vektoren liegt. Wäre u gleich einer der Vektoren $w_1, \ldots, w_{j+1}$, d.h. $u = w_i$ für $i \leq j+1$ so würde dies

$$w_i \in \mathbf{span}\{w_1, \ldots w_{i-1}\} \subset \mathbf{span}\{w_k : k \neq i\}$$

bedeuten. Dies widerspricht nach Lemma 9.1.7 der linearen Unabhängigkeit von $w_1, \ldots, w_n$. Damit muss $u = v_{i_k}$ für ein $i_k \in I_j$ sein. Wir setzen $I_{j+1} := I_j \setminus \{i_k\}$ und erhalten $|I_{j+1}| = m - (j+1)$ und

$$V = \mathbf{span}\Big(\{w_1, \ldots, w_{j+1}\} \cup \{v_i : i \in I_{j+1}\}\Big).$$

Diesen Prozess können wir solange fortsetzen bis n Vektoren in $v_1, \ldots, v_m$ durch die $w_1, \ldots, w_n$ ersetzt worden sind. Dies wäre jedoch im Fall $m < n$ nicht möglich.

Korollar 9.1.9

Seien V ein $\mathbb{K}$-Vektorraum und $v_1, \ldots, v_m$ eine Basis von V. Dann hat jede andere Basis von V genau m Elemente.

Beweis.

Sei $w_1, \ldots, w_n$ eine weitere Basis von V. Da $v_1, \ldots, v_m$ ein Erzeugendensystem von V ist und $w_1, \ldots, w_n$ linear unabhängig sind erhalten wir $n \leq m$ mit Satz 9.1.8.

Vertauschen der Basen in dem obigen Argument ergibt $m \leq n$. □

Mit Hilfe des vorherigen Korollars können wir nun die Dimension eines Vektorraums definieren.

Definition 9.1.10

Seien V ein $\mathbb{K}$-Vektorraum und $v_1, \ldots, v_m$ eine Basis von V. Dann heißt

$$\dim V := m$$

die Dimension von V. Gibt es keine endliche Basis von V, so setzen wir $\dim V := \infty$ und im Fall $V = \{0\}$ setzen wir $\dim V = 0$.

Es gilt $\dim \mathbb{R}^d = d$, da $e_1, \ldots, e_d$ eine Basis sind. Ferner gilt für den Raum der Polynome vom Grad kleiner gleich n über einem Intervall I mit nichtleerem Inneren

$$\dim \mathrm{Pol}(I, n) = n + 1\,,$$

da die Monome $1, x, x^2, \ldots, x^n$ ein Erzeugendensystem von $\mathrm{Pol}(I, n)$ sind. Ferner hatten wir nach Korollar 8.3.4 bemerkt, dass die Koeffizienten eines Polynoms eindeutig sind, und damit sind die Monome auch linear unabhängig.

Der folgende Satz sichert die Existenz von Basen, falls es endliche Erzeugendensysteme gibt.

Theorem 9.1.11

Seien V ein $\mathbb{K}$-Vektorraum, $v_1, \ldots, v_m$ ein Erzeugendensystem von V und $W \neq \{0\}$ ein Unterraum von V. Dann hat W eine Basis und es gilt $\dim W \leq m$.

Betrachtet man den Fall $W = V$ im obigen Lemma, so sieht man, dass jeder nicht-triviale $\mathbb{K}$-Vektorraum V, d.h. $V \neq \{0\}$, der ein endliches Erzeugendensystem hat, auch eine Basis hat.

BEWEIS

Wegen $W \neq \{0\}$ gibt es ein $w_1 \in W$ mit $w_1 \neq 0$. Damit ist w_1 linear unabhängig. Falls $W \neq W_1 := \mathbf{span}\{w_1\}$ gilt, gibt es ein weiteres $w_2 \in W$ mit $w_2 \notin W_1$. Nach Lemma 9.1.7 sind dann w_1, w_2 linear unabhängig. Falls $W \neq W_2 := \mathbf{span}\{w_1, w_2\}$ gilt, gibt es ein weiteres $w_3 \in W$ mit $w_3 \notin W_2$ und nach Lemma 9.1.7 sind dann w_1, w_2, w_3 linear unabhängig.

Sei nun $j \geq 1$, so dass dieser Prozess j mal durchgeführt worden ist, d.h. wir haben linear unabhängige Vektoren $w_1, \ldots, w_j \in W$ gefunden. Wäre $j > m$, so würde dies Satz 9.1.8 widersprechen. Damit gibt es eine maximale Anzahl $n \leq m$ von Durchführungen des Prozesses, was bedeutet, dass wir nach n Durchführungen $W = \mathbf{span}\{w_1, \ldots, w_n\}$ haben. Damit ist $w_1, \ldots, w_n$ ein linear unabhängiges Erzeugendensystem von W, d.h. eine Basis. Wegen $\mathbf{dim}\, W = n$ folgt dann auch die Ungleichung.

Das folgende Korollar zeigt, dass Unterräume nie eine größere Dimension als ihre umgebenden Räume haben können.

Korollar 9.1.12

Seien V ein $\mathbb{K}$-Vektorraum und W ein Unterraum von V. Dann gilt $\mathbf{dim}\, W \leq \mathbf{dim}\, V$.

Beweis.

Ist $\mathbf{dim}\, V = \infty$ oder $\mathbf{dim}\, W = 0$, so ist nichts zu zeigen. Im Fall $\mathbf{dim}\, V < \infty$ und $\mathbf{dim}\, W > 0$ hat V ein endliches Erzeugendensystem der Größe $\mathbf{dim}\, V$ und aus $\mathbf{dim}\, W > 0$ folgt $W \neq \{0\}$. Dann folgt die Behauptung aus Satz 9.1.11. □

Das folgende Korollar untersucht die Situation $\mathbf{dim}\, V = \mathbf{dim}\, W$ im Fall endlicher Dimension.

Korollar 9.1.13

Seien V ein endlich-dimensionaler $\mathbb{K}$-Vektorraum und W ein Unterraum von V. Dann gilt $\mathbf{dim}\, W = \mathbf{dim}\, V$ genau dann, wenn $V = W$.

BEWEIS

Ist $V = W$, so gilt offensichtlich $\mathbf{dim}\, V = \mathbf{dim}\, W$.

Gilt andersherum $\mathbf{dim}\, V = \mathbf{dim}\, W$, so ist $W \subset V$ nach Voraussetzung erfüllt. Wäre nun $W \neq V$, so gäbe es daher ein $v \in V \setminus W$. Nach Satz 9.1.11 hat W eine Basis $w_1, \ldots, w_m$. Lemma 9.1.7 zeigt dann, dass $v, w_1, \ldots, w_m$ linear unabhängig sind und damit folgt $\mathbf{dim}\, V \geq m + 1 = \mathbf{dim}\, W + 1$, was ein Widerspruch zu der Annahme $\mathbf{dim}\, V = \mathbf{dim}\, W$ ist.

Ist I ein Intervall mit nichtleerem Inneren, so hatten wir schon $\dim \mathrm{Pol}(I, n) = n + 1$ gesehen. Da $\mathrm{Pol}(I, n)$ auch Unterraum von $\mathrm{Pol}(I)$, folgt $\dim \mathrm{Pol}(I) \geq n + 1$ für alle $n \geq 1$. Dies zeigt

$$\dim \mathrm{Pol}(I) = \infty .$$

Ferner ist für $I = [a, b]$ der Raum $\mathrm{Pol}(I)$ ein Unterraum von $\mathcal{C}([a, b])$ und damit folgt auch

$$\dim \mathcal{C}([a, b]) = \infty .$$

Dieses Beispiel zeigt auch, dass Korollar 9.1.13 im Falle unendlich-dimensionaler Vektorräume falsch ist.

MAXIMALE, LINEAR UNABHÄNGIGE FAMILIEN SIND BASEN

Das folgende Korollar zeigt, dass linear unabhängige Familien oder Erzeugendensysteme der Größe $\dim V$ Basen sind.

Korollar 9.1.14

Seien V ein $\mathbb{K}$-Vektorraum mit $n = \dim V \in \mathbb{N}$ und $v_1, \ldots, v_n \in V$ linear unabhängig oder ein Erzeugendensystem von V. Dann ist $v_1, \ldots, v_n$ eine Basis von V.

BEWEIS

Wir nehmen zunächst an, dass $v_1, \ldots, v_n \in V$ linear unabhängig sind. Wegen $\mathbf{dim}\, V = n$ gibt es ein Erzeugendensystem $w_1, \ldots, w_n$ von V. Wir nehmen nun an, dass $v_1, \ldots, v_n$ keine Basis von V ist, d.h.

$$\mathbf{span}\{v_1, \ldots, v_n\} \neq V\,.$$

Dann gibt es ein $w \in V$ mit $w \notin \mathbf{span}\{v_1, \ldots, v_n\}$. Nach Lemma 9.1.7 sind dann $v_1, \ldots, v_n, w$ linear unabhängig, d.h. wir haben $n+1$ linear unabhängige Elemente in V. Dies widerspricht Satz 9.1.8.

Seien nun $v_1, \ldots, v_n$ ein Erzeugendensystem von V. Wären sie keine Basis, so wären sie linear abhängig. Da $\mathbf{dim}\, V \geq 1$ kann nicht $v_1 = v_2 = \cdots = v_n = 0$ gelten, d.h. es gibt ein $v_i \neq 0$. Ohne Einschränkung können wir $v_1 \neq 0$ annehmen. Nach Lemma 9.1.6 gibt es dann ein $j \geq 2$ mit

$$\mathbf{span}\{v_1, \ldots, v_{j-1}, v_{j+1}, v_n\} = \mathbf{span}\{v_1, \ldots, v_n\} = V\,.$$

Damit haben wir ein Erzeugendensystem der Größe $n-1$, was nach Satz 9.1.8 und $\mathbf{dim}\, V = n$ unmöglich ist.

EXISTENZ VON BASEN IM ALLGEMEINEN FALL

Man kann Erzeugendensysteme und lineare Unabhängigkeit auf unendliche Familien ausweiten, indem im ersten Fall alle Linearkombinationen über alle endlichen Teilfamilien betrachtet werden und im zweiten Fall verlangt wird, dass alle endlichen Teilfamilien linear unabhängig sind. Nach dieser Verallgemeinerung kann man auch unendlich große Basen als linear unabhängige Erzeugendensysteme definieren. In einem weiteren Schritt kann man dann zeigen, dass jeder Vektorraum eine (potenziell unendlich große) Basis hat. Solche Basen wollen wir hier aber nicht weiter betrachten.

Section 9.2
Lineare Abbildungen

In diesem Kapitel wollen wir Abbildungen zwischen Vektorräumen untersuchen, deren Abbildungsverhalten sich mit der Addition und Skalarmultiplikation vertragen.

Wir beginnen mit der folgenden grundlegenden Definition.

Definition 9.2.1

Seien V und W zwei $\mathbb{K}$-Vektorräume und $T : V \to W$. Dann heißt T **linear**, falls die folgenden zwei Eigenschaften erfüllt sind:

i). **Additivität:** Für alle $v, w \in V$ gilt $T(v + w) = T(v) + T(w)$.

ii). **Homogenität:** Für alle $v \in V$ und $\lambda \in \mathbb{K}$ gilt $T(\lambda v) = \lambda T(v)$.

In diesem Fall heißt T auch **(Vektorraum)-Homomorphismus** und wir schreiben $Tv := T(v)$. Ist T bijektiv und linear, so sprechen wir auch von einem **(Vektorraum)-Isomorphismus**. Wir schreiben ferner

$$\mathcal{L}(V, W) := \left\{ T : V \to W \,\middle|\, T \text{ linear} \right\}.$$

Sind V, W beliebige Vektorräume, so ist $\mathbf{id}_V : V \to V$ linear und die konstante Nullabbildung $v \mapsto 0$ von V nach W ist ebenfalls linear.

Bevor wir weitere Beispiele linearer Abbildungen betrachten, wollen wir den Raum $\mathcal{L}(V, W)$ etwas näher untersuchen.

Lemma 9.2.2

Seien V, W zwei $\mathbb{K}$-Vektorräume. Dann ist $\mathcal{L}(V, W)$ ein Unterraum von $\mathrm{Abb}(V, W)$.

Beweis.

Offensichtlich ist die konstante Nullabbildung linear, d.h. $0 \in \mathcal{L}(V, W)$. Sind $S, T \in \mathcal{L}(V, W)$, so gilt für $v, w \in V$

$$\begin{aligned}(S+T)(v+w) = S(v+w) + T(v+w) &= Sv + Sw + Tv + Tw \\ &= (S+T)v + (S+T)w\end{aligned}$$

damit ist $S + T$ additiv. Analog kann man die Homogenität von $S + T$ zeigen. Mit anderen Worten ist $S + T$ linear, d.h. $S + T \in \mathcal{L}(V, W)$. Für $\lambda \in \mathbb{K}$ kann man analog $\lambda T \in \mathcal{L}(V, W)$ zeigen. Lemma 9.1.2 gibt dann die Behauptung. □

Die Additivität und Homogenität sorgen dafür, dass Operationen im Definitionsbereich von T kompatibel zu denen im Bildbereich sind. Ähnliche Verträglichkeiten hatten wir auch schon früher kennengelernt, z.B.:

i). Monotone Abbildungen vertragen sich mit $\leq$ und streng monotone Abbildungen mit $<$.

ii). Stetige Funktionen vertragen sich mit Grenzwertbildung, genauer gesagt mit den zugehörigen Topologien.

iii). Die Exponentialfunktion $\exp : (\mathbb{R}, +) \to ((0, \infty), \cdot)$ verträgt sich mit den beiden Gruppenoperationen $+$ und $\cdot$ im Definitions- bzw. Bildbereich.

Für $V = W = \mathbb{K}$ ist eine Abbildung $T : \mathbb{K} \to \mathbb{K}$ genau dann linear, wenn es ein $a \in \mathbb{K}$ gibt mit $T(x) = ax$ für alle $x \in \mathbb{K}$. Dieses a lässt sich durch $a = T(1)$ bestimmen.

Bezeichnen $\mathcal{C}(I)$ und $\mathcal{C}^1(I)$ die Mengen aller stetigen bzw. stetig differenzierbaren Abbildungen $f : I \to \mathbb{R}$, wobei I ein Intervall ist, so ist der Ableitungsoperator

$$\begin{aligned} \mathcal{C}^1([a,b]) &\to \mathcal{C}((a,b)) \\ f &\mapsto f' \end{aligned}$$

linear. Umgekehrt ist auch die Abbildung

$$\begin{aligned} \mathcal{C}([a,b]) &\to \mathcal{C}^1((a,b)) \\ f &\mapsto F\,, \end{aligned}$$

wobei F die Stammfunktion von f bezeichnet, linear.

Ist I ein Intervall und $x \in I$, so ist auch der Auswertungsoperator

$$\begin{aligned} \mathcal{C}(I) &\to \mathbb{R} \\ f &\mapsto f(x) \end{aligned}$$

linear.

Bezeichnet $c := \{(x_n)_{n\in\mathbb{N}} \subset \mathbb{R} : \ (x_n)_{n\in\mathbb{N}} \text{ konvergiert } \}$ den Raum aller konvergenten, reellen Folgen, so ist die Abbildung

$$\begin{aligned} c &\to \mathbb{R} \\ (a_n)_{n\in\mathbb{N}} &\mapsto \lim_{n\to\infty} a_n \end{aligned}$$

linear.

Abgesehen vom ersten Beispiel, haben alle anderen Beispiele mindestens einen unendlich-dimensionalen Vektorraum. Wir wollen uns aber auf endlich-dimensionale Vektorräume konzentrieren. Dazu wollen wir als nächstes den Fall $T : \mathbb{K} \to \mathbb{K}$ auf $T : \mathbb{K}^n \to \mathbb{K}^n$ verallgemeinern.

Seien dazu $m, n \geq 1$. Eine $m \times n$-**Matrix** A besteht aus $m \cdot n$ Einträgen $a_{ij} \in \mathbb{K}$ mit $i = 1, \ldots, m$ und $j = 1, \ldots, n$, die üblicherweise so geschrieben werden:

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix}.$$

Die Menge aller $m \times n$-Matrizen wird mit $\mathcal{M}_{\mathbb{K}}(m, n)$ bezeichnet. Ist $\mathbb{K}$ aus dem Zusammenhang bekannt, wird es dabei meistens weggelassen. Man beachte, dass wir für $I := \{1, \ldots, m\} \times \{1, \ldots, n\}$ nichts anderes als $\mathcal{M}_{\mathbb{K}}(m, n) = \mathrm{Abb}(I, \mathbb{K})$ haben.

Für Matrizen der gleichen Größe $m \times n$ kann man eine Addition durch komponentenweise Addition definieren, d.h.

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} := \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{pmatrix}.$$

Analog kann man eine Skalarmultiplikation komponentenweise definieren, d.h.

$$\lambda \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} := \begin{pmatrix} \lambda a_{11} & \dots & \lambda a_{1n} \\ \vdots & & \vdots \\ \lambda a_{m1} & \dots & \lambda a_{mn} \end{pmatrix}.$$

Da diese Operationen genau denen auf $\mathrm{Abb}(I, \mathbb{K})$ entsprechen, sieht man leicht, dass $\mathcal{M}_{\mathbb{K}}(m, n)$ zusammen mit diesen beiden Operationen ein $\mathbb{K}$-Vektorraum ist. Analog kann dies auch direkt verifiziert werden.

MATRIZEN ALS LINEARE OPERATOREN

Ist A eine $m \times n$-Matrix mit Einträgen $a_{ij} \in \mathbb{K}$ und $x \in \mathbb{K}^n$ ein Vektor mit Einträgen $x_1, \ldots, x_n$, so können wir ferner eine Multiplikation durch

$$Ax = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} := \begin{pmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} x_j \end{pmatrix}$$

definieren. Damit ist insbesondere $Ax \in \mathbb{K}^m$. Leichtes Nachrechnen zeigt zudem, dass

$$A(x+y) = Ax + Ay \qquad \text{und} \qquad A(\lambda x) = \lambda Ax$$

für alle $x, y \in \mathbb{K}^n$ und $\lambda \in \mathbb{K}$ gilt. Mit anderen Worten ist die Abbildung

$$\begin{aligned} \mathrm{L}_A : \mathbb{K}^n &\to \mathbb{K}^m \\ x &\mapsto Ax \end{aligned}$$

linear, d.h. jede $m \times n$-Matrix definiert eine lineare Abbildung $\mathbb{K}^n \to \mathbb{K}^m$ vermöge der obigen Multiplikation.

Wenn wir ein lineares Gleichungssystem $Ax = b$ lösen wollen, so suchen wir also nach dem Urbild

$$\mathsf{L}_A^{-1}(\{b\})\,.$$

Sind A und B zwei $m \times n$-Matrizen, so lässt sich ebenfalls

$$(A + B)x = Ax + Bx \qquad \text{und} \qquad \lambda(Ax) = (\lambda A)x$$

für alle $x \in \mathbb{K}^n$ und $\lambda \in \mathbb{K}$ schnell nachrechnen. Die erste Gleichung zeigt dann $\mathsf{L}_{A+B}x = \mathsf{L}_A x + \mathsf{L}_B x$, d.h.

$$\mathsf{L}_{A+B} = \mathsf{L}_A + \mathsf{L}_B\,.$$

Analog zeigt die zweite Gleichung $\mathsf{L}_{\lambda A} = \lambda \mathsf{L}_A$. Mit anderen Worten haben wir das folgende Lemma gezeigt:

Lemma 9.2.3

Für alle $n, m \geq 1$ ist die Abbildung

$$\mathcal{M}_{\mathbb{K}}(m,n) \to \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$$
$$A \mapsto \mathsf{L}_A$$

linear.

Später in Satz 9.2.10 werden wir sehen, dass diese Abbildung sogar ein Isomorphismus ist, d.h. die linearen Abbildungen $\mathbb{K}^n \to \mathbb{K}^m$ lassen sich eindeutig durch Matrizen beschreiben. Im Fall $m = n = 1$ hatten wir dies am Anfang des Unterkapitels schon gesehen.

Der folgende Satz stellt eine weitere wichtige lineare Abbildung vor. Sie dient dazu, Vektorräume V mit $\dim V = d$ mit $\mathbb{K}^d$ zu identifizieren.

Theorem 9.2.4

Seien V ein $\mathbb{K}$-Vektorraum mit $\dim V = d < \infty$ und $\mathcal{B} := (v_1, \ldots, v_d)$ eine Basis von V. Dann ist die ***Koordinatenabbildung***

$$C_{\mathcal{B}} : V \to \mathbb{K}^d$$

$$\sum_{i=1}^{d} x_i v_i \mapsto \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

ein Isomorphismus.

Die Koordinatenabbildung zu einer gegebenen Basis $\mathcal{B}$ ordnet also jedem Vektor v die eindeutigen Koordinaten $x_1, \ldots, x_d$ von v bezüglich dieser Basis zu.

BEWEIS

Wir bemerken zunächst, dass die Abbildung $C_{\mathcal{B}}$ wohldefiniert ist, da nach Lemma 9.1.4 jedes $v \in V$ eine eindeutige Darstellung $v = \sum_{i=1}^{d} x_i v_i$ besitzt. Die Surjektivität ist offensichtlich. Für die Injektivität betrachten wir zwei $v, w \in V$ mit $Tv = Tw = x$. Dann gilt

$$v = \sum_{i=1}^{d} x_i v_i \qquad \text{und} \qquad w = \sum_{i=1}^{d} x_i v_i \,. \tag{9.2.1}$$

Dies ergibt $v = w$. Um die Linearität zu überprüfen, wählen wir $v, w \in V$. Dann gibt es eindeutige $x, y \in \mathbb{K}^d$, so dass (9.2.1) gilt. Wegen

$$v + w = \sum_{i=1}^{d} x_i v_i + \sum_{i=1}^{d} x_i v_i = \sum_{i=1}^{d} (x_i + y_i) v_i$$

folgt dann

$$C_{\mathcal{B}}(v + w) = x + y = C_{\mathcal{B}} v + C_{\mathcal{B}} w \,.$$

Analog kann $C_{\mathcal{B}}(\lambda v) = \lambda C_{\mathcal{B}} v$ für $\lambda \in \mathbb{K}$ nachgewiesen werden.

BEMERKUNGEN

Wir hatten schon gesehen, dass die Monome $1, x, \ldots, x^n$ eine Basis in $\mathrm{Pol}(\mathbb{R}, n)$ bilden. Die zugehörige Koordinatenabbildung ordnet dann jedem Polynom p vom Grad $\deg p \leq n$ seine Koeffizienten zu.

Isomorphismen machen es möglich, Vektorräume miteinander zu identifizieren. Hierbei wird aber nur die Vektorraumstruktur identifiziert. Weitere Strukturen können dagegen verloren gehen. So besteht beispielsweise $\mathrm{Pol}(\mathbb{R}, n)$ aus Funktionen, die punktweise, d.h. für $x \in \mathbb{R}$ ausgewertet werden können. Dies geht bei der obigen Identifizierung von $\mathrm{Pol}(\mathbb{R}, n)$ mit $\mathbb{R}^{n+1}$ verloren, wenn man nur noch im $\mathbb{R}^{n+1}$ arbeitet.

Im folgenden wollen wir einige strukturelle Eigenschaften linear Abbildungen kennenlernen. Wir beginnen mit dem folgenden einfachen Lemma.

Lemma 9.2.5

Seien U, V, W drei $\mathbb{K}$-Vektorräume und $T : U \to V$ und $S : V \to W$ linear. Dann ist auch $S \circ T : U \to W$ linear.

Seien $u_1, u_2 \in U$. Dann gilt $(S \circ T)(u_1 + u_2) = S(T(u_1 + u_2)) = S(Tu_1 + T_u 2) = S(Tu_1) + S(Tu_2) = (S \circ T)(u_1) + (S \circ T)(u_2)$. Die Homogenität kann analog bewiesen werden.

Im Folgenden wird die Komposition linearer Abbildungen auch multiplikativ dargestellt, d.h. wir schreiben

$$ST := S \circ T$$

für die Komposition zweier linearer Abbildungen $T : U \to V$ und $S : V \to W$. Im Fall $S = T$ schreiben wir entsprechend auch $S^2 := S \circ S$, wobei dies natürlich $S : V \to V$ voraussetzt.

Für $\mathbb{K}$-Vektorräume V und W und $T : V \to W$ linear schreiben wir nun

$$\ker T := \{v \in V : Tv = 0\}$$

für den Kern von T. Analog ist

$$\operatorname{ran} T := T(V) = \{Tv : v \in V\}$$

das Bild von T. Das folgende Lemma zeigt, dass beide Mengen Vektorräume sind.

Lemma 9.2.6

Seien V, W zwei $\mathbb{K}$-Vektorräume und $T : V \to W$ linear. Dann ist $\ker T$ ein Unterraum von V und $\operatorname{ran} T$ ein Unterraum von W.

BEWEIS

Für $v \in V$ gilt $T(0) = T(0 \cdot v) = 0Tv = 0$, d.h. $0 \in \mathbf{ker}\, T$ und $0 \in \mathbf{ran}\, T$.

Seien nun $v, w \in \mathbf{ker}\, T$ und $\lambda \in \mathbb{K}$. Dann gilt $T(v + w) = Tv + Tw = 0 + 0 = 0$ und $T(\lambda v) = \lambda TV = \lambda 0 = 0$. Damit ist $\mathbf{ker}\, T$ nach Lemma 9.1.2 ein Unterraum von V.

Seien nun $x, y \in \mathbf{ran}\, T$. Dann gibt es $v, w \in V$ mit $Tv = x$ und $Tw = y$. Dies ergibt $x + y = Tv + Tw = T(v + w)$ und damit ist auch $x + y \in \mathbf{ran}\, T$. Analog gilt $\lambda x = \lambda Tv = T(\lambda x)$ für $\lambda \in \mathbb{K}$, und damit ist auch $\lambda x \in \mathbf{ran}\, T$. Lemma 9.1.2 zeigt dann dann $\mathbf{ran}\, T$ ein Unterraum von W ist.

Das folgende Lemma vereinfacht die Prüfung auf Injektivität für lineare Abbildungen.

Lemma 9.2.7

Seien V, W zwei $\mathbb{K}$-Vektorräume und $T : V \to W$ linear. Dann ist T injektiv genau dann, wenn $Tv \neq 0$ für alle $v \in V \setminus \{0\}$, d.h.

$$\ker T = \{0\}$$

Sind in diesem Fall $v_1, \ldots, v_n \in V$ linear unabhängig, so sind auch $Tv_1, \ldots, Tv_n$ linear unabhängig und insbesondere gilt dann $\dim V \leq \dim W$.

BEWEIS 1/2

Ist T injektiv, so ist jedes Urbild $T^{-1}(w)$ einelementig und damit auch $\ker T = T^{-1}(0)$. Da $0 \in \ker T$ nach Lemma 9.2.6, folgt $\ker T = \{0\}$.

Seien umgekehrt $v, w \in V$ mit $Tv = Tw$. Dann gilt $T(v - w) = 0$ und wegen $\ker T = \{0\}$ folgt dann auch $v - w = 0$. Dies zeigt $v = w$.

Seien nun $v_1, \ldots, v_n$ linear unabhängig. Für $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$ mit

$$\sum_{i=1}^{n} \lambda_i T v_i = 0$$

gilt dann

$$0 = \sum_{i=1}^{n} \lambda_i T v_i = T\left(\sum_{i=1}^{n} \lambda_i v_i\right)$$

und wegen der Injektivität dann auch $\sum_{i=1}^{n} \lambda_i v_i = 0$. Da $v_1, \ldots, v_n$ linear unabhängig sind, folgt $\lambda_1 = \cdots = \lambda_n$. Damit sind auch $Tv_1, \ldots, Tv_n$ linear unabhängig.

BEWEIS 2/2

Ist V endlich-dimensional und $v_1, \ldots, v_n$ eine Basis von V, so haben wir schon gesehen, dass $Tv_1, \ldots, Tv_n \in W$ linear unabhängig sind. Ist nun $w_1, \ldots, w_m$ eine Basis von W, so gilt $n \leq m$ nach Satz 9.1.8.

Ist V unendlich-dimensional, so finden wir für jedes $n \geq 1$ eine linear unabhängige Familie $v_1, \ldots, v_n$ und damit auch eine linear unabhängige Familie $Tv_1, \ldots, Tv_n$ in W. Dies zeigt $\mathbf{dim}\, W = \infty$.

Der folgende Satz zeigt, dass die Umkehrabbildung eines Isomorphismus ein Isomorphismus ist.

Theorem 9.2.8

Seien V, W zwei $\mathbb{K}$-Vektorräume und $T : V \to W$ linear. Ist dann T bijektiv, so ist auch T^{-1} linear und es gilt $\dim V = \dim W$.

Um die Additivität von T^{-1} zu zeigen, seien $w_1, w_2 \in W$. Wir setzen $v_1 := T^{-1}w_1$ und $v_2 := T^{-1}w_2$. Dann gilt $T(v_1 + v_2) = Tv_1 + Tv_2 = w_1 + w_2$ und damit $v_1 + v_2 = T^{-1}(w_1 + w_2)$. Dies ergibt

$$T^{-1}(w_1 + w_2) = v_1 + v_2 = T^{-1}w_1 + T^{-1}w_2 \,.$$

Die Homogenität von T^{-1} kann analog gezeigt werden. Da T injektiv ist, gilt $\mathbf{dim}\, V \leq \mathbf{dim}\, W$ nach Lemma 9.2.7 Ferner ist $T^{-1} : W \to V$ bijektiv ist, siehe auch Abschnitt 47, und damit zeigt Lemma 9.2.7 auch $\mathbf{dim}\, W \leq \mathbf{dim}\, V$.

Das folgende Korollar zeigt, dass endlich dimensionale Vektorräume schon isomorph zueinander sind, falls sie die gleiche Dimension haben.

Korollar 9.2.9

Seien V und W endlich-dimensionale $\mathbb{K}$-Vektorräume. Dann gibt es einen Isomorphismus $T : V \to W$ genau dann, wenn $\dim V = \dim W$.

Beweis.

Seien $\mathcal{A}$ und $\mathcal{B}$ Basen von V und W und $n := \dim V = \dim W$. Dann sind die Koordinatenabbildungen $C_{\mathcal{A}} : V \to \mathbb{K}^n$ und $C_{\mathcal{B}} : W \to \mathbb{K}^n$ Isomorphismen nach Satz 9.2.4. Nach Satz 9.2.8 ist dann auch $C_{\mathcal{B}}^{-1} : \mathbb{K}^n \to W$ ein Isomorphismus. Nach Lemma 9.2.5 ist dann $C_{\mathcal{B}}^{-1} \circ C_{\mathcal{A}} : V \to W$ linear und da beide Koordinatenabbildungen bijektiv sind, ist ihre Komposition ebenfalls bijektiv. □

Der folgende Satz zeigt, dass der Raum $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$ kanonisch isomorph zu $\mathcal{M}_{\mathbb{K}}(m, n)$ ist.

Theorem 9.2.10

Für alle $n, m \geq 1$ ist die Abbildung

$$\begin{aligned} \mathcal{M}_{\mathbb{K}}(m, n) &\to \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m) \\ A &\mapsto \mathsf{L}_A \end{aligned}$$

ein Isomorphismus und seine Umkehrabbildung

$$\begin{aligned} \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m) &\to \mathcal{M}_{\mathbb{K}}(m, n) \\ S &\mapsto \mathsf{M}(S) \end{aligned}$$

ist ebenfalls ein Isomorphismus.

Mit Hilfe des obigen Isomorhismus werden wir im folgenden Begriffe für lineare Abbildungen auf Matrizen übertragen und umgekehrt. Für $A \in \mathcal{M}(m, n)$ schreiben wir z.B.

$$\ker A := \ker \mathsf{L}_A \, .$$

BEWEIS 1/2

In Lemma 9.2.3 hatten wir schon gezeigt, dass die Abbildung linear ist.

Um zu zeigen, dass die Abbildung injektiv ist, wählen wir ein $A \in \mathcal{M}_{\mathbb{K}}(m,n)$ mit $\mathrm{L}_A = 0$, d.h. $\mathrm{L}_A x = 0$ für alle $x \in \mathbb{K}^n$. Gäbe es Indices i^*, j^* mit $a_{i^*j^*} \neq 0$, dann folgt für den j^*-ten Standard-Einheitsvektor $e_{j^*} \in \mathbb{K}^n$ mit den Komponenten $e_{1j^*}, \ldots, e_{nj^*}$, dass

$$0 = \mathrm{L}_A e_{j^*} = A e_{j^*} = \begin{pmatrix} \sum_{j=1}^n a_{1j} e_{jj^*} \\ \vdots \\ \sum_{j=1}^n a_{mj} e_{jj^*} \end{pmatrix} = \begin{pmatrix} a_{1j^*} \\ \vdots \\ e_{mj^*} \end{pmatrix} \neq 0\,.$$

Damit ist die Annahme falsch, und es folgt $A = 0$. Nach Lemma 9.2.7 ist die Abbildung $A \mapsto \mathrm{L}_A$ dann injektiv.

BEWEIS 2/2

Um die Surjektivität zu überprüfen, fixieren wir ein lineares $S : \mathbb{K}^n \to \mathbb{K}^m$. Für $j = 1, \ldots, n$ setzen wir

$$a_j := \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} := Se_j \tag{9.2.2}$$

und damit

$$A := (a_1, \ldots a_n) = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} . \tag{9.2.3}$$

Für $x \in \mathbb{K}^n$ gilt dann

$$Sx = S\left(\sum_{j=1}^{n} x_j e_j\right) = \sum_{j=1}^{n} x_j Se_j = \sum_{j=1}^{n} x_j \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} x_j \end{pmatrix} = Ax .$$

Dies ergibt $S = \mathrm{L}_A$ und damit haben wir die Surjektivität gezeigt. Die Aussagen über die Umkehrabbildung folgen direkt aus Satz 9.2.8.

Das folgende Korollar bestimmt die Dimension des Raums $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$.

Korollar 9.2.11

Für alle $n, m \geq 1$ gilt $\dim \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m) = mn$.

Beweis.

Der Raum $\mathcal{M}_{\mathbb{K}}(m, n)$ ist kanonisch isomorph zu $\mathbb{K}^{mn}$, so dass wir $\dim \mathcal{M}_{\mathbb{K}}(m, n) = mn$ haben. Nun folgt die Behauptung aus Satz 9.2.10 und Satz 9.2.8. □

Das folgende Korollar beschreibt, wie die darstellende Matrix einer linearen Abbildung $S : \mathbb{K}^n \to \mathbb{K}^m$ gefunden werden kann. Es basiert direkt auf der Konstruktion (9.2.2) und (9.2.3) im Beweis von Satz 9.2.10.

Korollar 9.2.12

Sei $S : \mathbb{K}^n \to \mathbb{K}^m$ und $A \in \mathcal{M}_{\mathbb{K}}(m, n)$ die darstellende Matrix von S, d.h. $A := \mathrm{M}(S)$. Dann ist die j-Spalte von A das Bild von e_j unter S, d.h.

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} = \left(Se_1, \dots, Se_n\right).$$

DREH-MATRIX

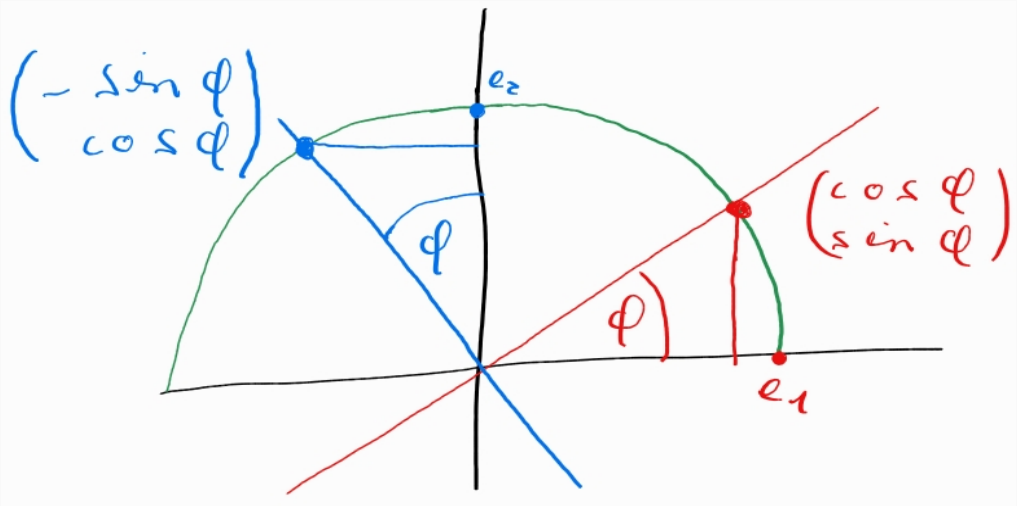
Um ein Beispiel zu betrachten, wollen wir die Drehung S_φ im $\mathbb{R}^2$ um einen Winkel φ gegen den Uhrzeigersinn beschreiben. Nach Korollar 9.2.12 müssen wir dazu nur die Bilder von e_1 und e_2 unter S_φ bestimmen. Es gilt hierbei

$$S_\varphi e_1 = \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} \qquad \text{und} \qquad S_\varphi e_2 = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \end{pmatrix},$$

siehe auch Abbildung 19. Für

$$D_\varphi := \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \tag{9.2.4}$$

haben wir damit $S_\varphi = L_{D_\varphi}$ nach Korollar 9.2.12.



Theorem 9.2.13

Seien V und W endlich-dimensionale $\mathbb{K}$-Vektorräume mit Basen $\mathcal{A} := (v_1, \ldots, v_n)$ und $\mathcal{B} := (w_1, \ldots, w_m)$. Dann gibt es zu jeder linearen Abbildung genau eine Matrix $A \in \mathcal{M}_{\mathbb{K}}(m, n)$ mit

$$S = C_{\mathcal{B}}^{-1} \circ L_A \circ C_{\mathcal{A}} \,. \tag{9.2.5}$$

Ferner gilt:

i). Für alle $j = 1, \ldots, n$ und $i = 1, \ldots, m$ gilt

$$C_{\mathcal{A}} v_j = e_j \qquad \text{und} \qquad C_{\mathcal{B}}^{-1} e_i = w_i \,.$$

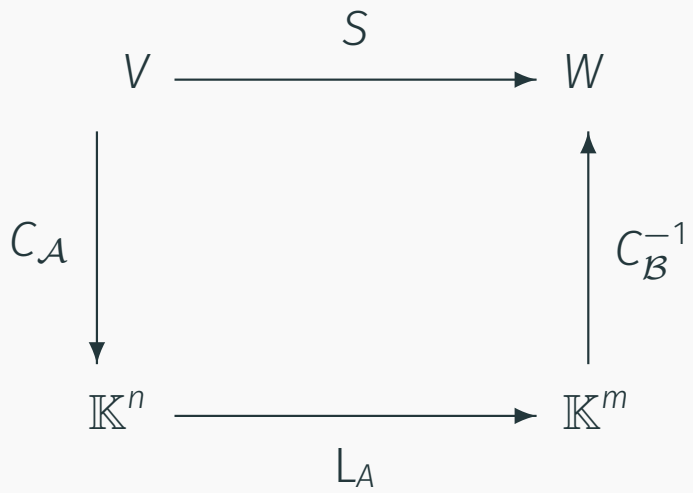
ii). Für alle $j = 1, \ldots, n$ gilt $Sv_j = \sum_{i=1}^{m} a_{ij} w_i$.

iii). Für $v = \sum_{j=1}^{n} x_j v_j \in V$ und $w = \sum_{i=1}^{m} y_i w_i$ mit $Sv = w$ gilt

$$A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \,.$$

BEMERKUNG

Die Gleichung (9.2.5) kann man durch das folgende kommutative Diagramm verdeutlichen



Die Kommutativität bedeutet dabei, dass Pfade entlang der Pfeile zu dem gleichen Ergebnis führen.

BEWEIS 1/2

Die Bedingung $S = C_{\mathcal{B}}^{-1} \circ \mathsf{L}_A \circ C_{\mathcal{A}}$ ist äquivalent zu

$$C_{\mathcal{B}} \circ S \circ C_{\mathcal{A}}^{-1} = \mathsf{L}_A \,.$$

Die Existenz und Eindeutigkeit von A folgt dann aus Satz 9.2.10.

i). Dies folgt direkt aus der Definition der Koordinatenabbildung.

ii). Nachrechnen ergibt

$$\begin{aligned} Sv_j = C_{\mathcal{B}}^{-1} \circ L_A \circ C_{\mathcal{A}} v_j = C_{\mathcal{B}}^{-1} \circ L_A e_j &= C_{\mathcal{B}}^{-1}(Ae_j) \\ &= C_{\mathcal{B}}^{-1} \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} \\ &= C_{\mathcal{B}}^{-1} \left(\sum_{i=1}^{m} a_{ij} e_i \right) \\ &= \sum_{i=1}^{m} a_{ij} C_{\mathcal{B}}^{-1} e_i \\ &= \sum_{i=1}^{m} a_{ij} w_i \,. \end{aligned}$$

iii). Es ist $x = C_{\mathcal{A}} v$ und $y = C_{\mathcal{B}} w$. Dann gilt

$$Ax = L_A C_{\mathcal{A}} v = C_{\mathcal{B}} S v = C_{\mathcal{B}} w = y \,,$$

was die letzte Behauptung zeigt.

Wenden wir $\mathrm{M}(\cdot)$ auf $\mathrm{L}_A = C_{\mathcal{B}} \circ S \circ C_{\mathcal{A}}^{-1}$ an, so ergibt sich

$$A = \mathrm{M}(\mathrm{L}_A) = \mathrm{M}(C_{\mathcal{B}} \circ S \circ C_{\mathcal{A}}^{-1}) .$$

Dies zeigt, dass die darstellende Matrix A von den Basen $\mathcal{A}$ und $\mathcal{B}$ abhängt. Wir sprechen daher auch von der darstellenden Matrix bezüglich der Basen $\mathcal{A}$ und $\mathcal{B}$ und schreiben

$$\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(S) := \mathrm{M}(C_{\mathcal{B}} \circ S \circ C_{\mathcal{A}}^{-1}) .$$

Sind $\mathcal{E}_n$ und $\mathcal{E}_m$ die Standard-Basen von $\mathbb{K}^n$ bzw. $\mathbb{K}^m$, so haben wir für lineare $S : \mathbb{K}^n \to \mathbb{K}^m$:

$$\mathrm{M}(S) = \mathrm{M}_{\mathcal{E}_m}^{\mathcal{E}_n}(S) ,$$

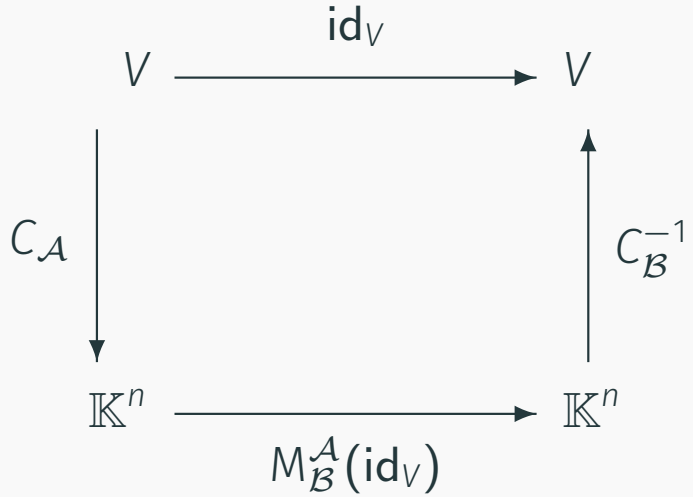
da in diesem Fall $C_{\mathcal{A}} = \mathbf{id}_{\mathbb{K}^n}$ und $C_{\mathcal{B}} = \mathbf{id}_{\mathbb{K}^m}$ gilt.

Der Teil *ii)* stellt eine Möglichkeit zur Berechnung von A zur Verfügung. In der Tat zeigt

$$Sv_j = \sum_{i=1}^{m} a_{ij} w_i ,$$

dass wir zu jedem v_j die Koeffizienten von Sv_j in der Basisdarstellung bzgl. $\mathcal{B}$ berechnen müssen. Dies ergibt dann die j-te Spalte von A.

Betrachten wir Satz 9.2.13 für $V = W$ und $S := \mathrm{id}_V$ und mit ansonsten gleichen Bezeichnungen, so ergibt sich das Diagramm



Haben wir nun ein $v \in V$ mit Basisdarstellungen $v = \sum_{j=1}^{n} x_j v_j$ und $v = \sum_{i=1}^{m} y_i w_i$, so zeigt Teil *iii)*

$$\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(\mathrm{id}_V)x = y\,,$$

d.h. $\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(\mathrm{id}_V)$ beschreibt den Wechsel in den Darstellungen bei dem Wechsel von der Basis $\mathcal{A}$ zu der Basis $\mathcal{B}$. Wir sprechen daher von einer Basiswechselmatrix und schreiben

$$\mathrm{T}_{\mathcal{B}}^{\mathcal{A}} := \mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(\mathrm{id}_V)\,.$$

Ist speziell $V := \mathbb{K}^n$ mit einer beliebigen Basis $\mathcal{A} := (v_1, \ldots, v_n)$, so gilt

$$\mathrm{id}_{\mathbb{K}^n} v_j = v_j = \sum_{i=1}^{n} v_{ij} e_i \,, \tag{9.2.6}$$

wobei $v_{1j}, \ldots, v_{nj}$ die Komponenten von v_j sind, d.h.

$$v_j = \begin{pmatrix} v_{1j} \\ \vdots \\ v_{nj} \end{pmatrix} .$$

Ist $\mathcal{E}_n := (e_1, \ldots, e_n)$ die Standardbasis von $\mathbb{K}^n$, so zeigt Teil *ii)* von Satz 9.2.13 in Verbindung mit (9.2.6) daher

$$\mathrm{T}^{\mathcal{A}}_{\mathcal{E}_n} = \mathrm{M}^{\mathcal{A}}_{\mathcal{E}_n}(\mathrm{id}_{\mathbb{K}^n}) = (v_1, \ldots v_n) \,.$$

Mit anderen Worten ist die Basiswechselmatrix in diesem Fall besonders einfach zu bestimmen. Allgemeinere Basiswechselmatrizen werden wir behandeln, sowie wir mehr über Matrizen wissen.

Section 9.3
Matrizenkalkül

Bis jetzt haben wir gesehen, dass

$$\mathcal{M}_{\mathbb{K}}(m,n) \to \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$$
$$A \mapsto \mathsf{L}_A$$

ein Vektorraum-Isomorphismus ist. Auf den Räumen $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$ haben wir aber eine weitere Struktur, denn die Komposition linearer Abbildungen ist wieder linear. Genauer gesagt haben wir für $R \in \mathcal{L}(\mathbb{K}^l, \mathbb{K}^n)$ und $S \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$:

$$S \circ R \in \mathcal{L}(\mathbb{K}^l, \mathbb{K}^m)\,.$$

Wie kann dieses Verhalten durch Matrizen beschrieben werden?

Um eine Antwort auf diese Frage zu finden, seien

$$R \in \mathcal{L}(\mathbb{K}^l, \mathbb{K}^n) \qquad \text{und} \qquad S \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)\,,$$

sowie $A \in \mathcal{M}_{\mathbb{K}}(n, l)$ und $B \in \mathcal{M}_{\mathbb{K}}(m, n)$ mit $\mathsf{L}_A = R$ und $\mathsf{L}_B = S$. Ferner sei $C \in \mathcal{M}_{\mathbb{K}}(m, l)$ mit $\mathsf{L}_C = S \circ R$.

Aus Korollar 9.2.12 wissen wir dann, dass

$$A = \begin{pmatrix} a_{11} & \dots & a_{1l} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nl} \end{pmatrix} = \left(Re_1, \dots, Re_l\right),$$

sowie

$$B = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} = \left(Se_1, \dots, Se_n\right)$$

und

$$C = \begin{pmatrix} c_{11} & \dots & c_{1l} \\ \vdots & & \vdots \\ c_{m1} & \dots & c_{ml} \end{pmatrix} = \left((S \circ R)e_1, \dots, (S \circ R)e_l\right).$$

Für $e_j \in \mathbb{K}^l$ und die j-te Spalte von C folgt dann

$$\begin{pmatrix} c_{1j} \\ \vdots \\ c_{mj} \end{pmatrix} = (S \circ R)e_j = S(Re_j) = S \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix} = S\left(\sum_{k=1}^{n} a_{kj} e_k\right)$$

$$= \sum_{k=1}^{n} a_{kj} S e_k$$

$$= \sum_{k=1}^{n} a_{kj} \begin{pmatrix} b_{1k} \\ \vdots \\ b_{mk} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{k=1}^{n} b_{1k} a_{kj} \\ \vdots \\ \sum_{k=1}^{n} b_{mk} a_{kj} \end{pmatrix} .$$

Für $i = 1, \ldots, m$ und $j = 1, \ldots, l$ gilt daher

$$c_{ij} = \sum_{k=1}^{n} b_{ik} a_{kj} \, . \tag{9.3.1}$$

Unsere Vorüberlegungen führen nun zu der folgenden Definition.

Definition 9.3.1

Für $l, n, m \geq 1$ seien $A \in \mathcal{M}_{\mathbb{K}}(n, l)$ und $B \in \mathcal{M}_{\mathbb{K}}(m, n)$. Dann ist das Matrixprodukt oder auch die Matrizenmultiplikation

$$C := BA \in \mathcal{M}_{\mathbb{K}}(m, l)$$

die durch (9.3.1) definierte Matrix C.

Unsere Vorüberlegungen liefern sofort den folgenden Satz, der zeigt, dass $A \mapsto \mathsf{L}_A$ die Matrizenmultiplikation in die Komposition von Abbildungen überführt.

Theorem 9.3.2

Für $l, n, m \geq 1$ seien $A \in \mathcal{M}_{\mathbb{K}}(n, l)$ und $B \in \mathcal{M}_{\mathbb{K}}(m, n)$. Dann gilt

$$\mathsf{L}_B \circ \mathsf{L}_A = \mathsf{L}_{BA} \,.$$

Sind $R : \mathbb{K}^l \to \mathbb{K}^n$ und $S : \mathbb{K}^n \to \mathbb{K}^m$ linear und setzen wir $A := \mathsf{M}(R)$ und $B := \mathsf{M}(S)$, d.h. $R = \mathsf{L}_A$ und $S = \mathsf{L}_B$, so gilt mit dem obigen Satz

$$\mathsf{M}(S \circ R) = \mathsf{M}(\mathsf{L}_B \circ \mathsf{L}_A) = \mathsf{M}(\mathsf{L}_{BA}) = BA = \mathsf{M}(S)\mathsf{M}(R) \,. \tag{9.3.2}$$

Damit verträgt sich auch die Umkehrabbildung $\mathsf{M}(\cdot)$ von $A \mapsto \mathsf{L}_A$ mit der Komposition von Abbildungen.

Die Matrizenmultiplikation erfüllt die folgenden Rechenregeln, wobei in jedem Fall angenommen wird, dass die beteiligten Matrizen A, B, C kompatible Größen haben:

$$(A + B)C = AC + BC \tag{9.3.3}$$
$$A(B + C) = AB + AC \tag{9.3.4}$$
$$A(BC) = (AB)C \tag{9.3.5}$$
$$A(\lambda B) = \lambda(AB)\,. \tag{9.3.6}$$

Um beispielsweise (9.3.3) zu zeigen, betrachten wir

$$\begin{aligned} \mathsf{L}_{(A+B)C} = \mathsf{L}_{A+B} \circ \mathsf{L}_C = (\mathsf{L}_A + \mathsf{L}_B) \circ \mathsf{L}_C = \mathsf{L}_A \circ \mathsf{L}_C + \mathsf{L}_B \circ \mathsf{L}_C = \mathsf{L}_{AB} + \mathsf{L}_{BC} \\ = \mathsf{L}_{AC+BC}\,, \end{aligned}$$

wobei der dritte Schritt auf der punktweisen Definition der Addition von Abbildungen basiert. Nach Satz 9.2.10 können wir dann $(A + B)C = AC + BC$ schließen. Die anderen Gleichungen können analog bewiesen werden.

KEINE KOMMUTATIVITÄT

Ähnlich wie die Komposition von Abbildungen ist die Matrizenmultiplikation *nicht* kommutativ, d.h. es gibt Matrizen A und B, die kompatible Größen haben und

$$AB \neq BA$$

erfüllen. Man beachte hierbei, dass das für nicht-quadratische Matrizen A oder B schon die Kompatibilität der Größen verletzt ist. Ein Beispiel quadratischer, nicht kommutativer Matrizen ist

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$$

und

$$\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} .$$

Für eine $m \times n$-Matrix

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$$

entsteht die **transponierte Matrix** A^T durch Vertauschen der Indizes i und j, d.h.

$$A^T := \begin{pmatrix} a_{11} & \dots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \dots & a_{mn} \end{pmatrix} .$$

Für $A \in \mathcal{M}(m,n)$ gilt daher $A^T \in \mathcal{M}(n,m)$. Fasst man Spaltenvektoren $x \in \mathbb{K}^m$ als $m \times 1$-Matrizen auf, so ist x^T ein Zeilenvektor.

Der folgende Satz liefert ein paar elementare Rechenregeln für die Transposition. Da der Beweis trivial ist, wird er weggelassen.

Theorem 9.3.3

Für alle $m, n \geq 1$ ist die Transposition

$$\begin{aligned} \mathcal{M}(m,n) &\to \mathcal{M}(n,m) \\ A &\mapsto A^{\mathsf{T}} \end{aligned}$$

linear und es gilt $(A^{\mathsf{T}})^{\mathsf{T}} = A$ für alle $A \in \mathcal{M}(m,n)$.

Der folgende Satz zeigt, dass sich die Transposition in gewisser Weise auch mit der Matrizenmultiplikation verträgt. Der Beweis ist ebenfalls elementar und wird daher übersprungen.

Theorem 9.3.4

Für alle $A \in \mathcal{M}_{\mathbb{K}}(n, l)$ und $B \in \mathcal{M}_{\mathbb{K}}(m, n)$ gilt

$$(AB)^{\mathsf{T}} = B^{\mathsf{T}} A^{\mathsf{T}} .$$

Sind $v, w \in \mathbb{K}^n$, so kann das Skalarprodukt von v und w durch

$$\langle v, w \rangle = \overline{v}^{\mathsf{T}} w$$

beschrieben werden, wobei wir im Fall $\mathbb{K} = \mathbb{R}$ dies sofort sehen. Für den komplexen Fall verweisen wir auf Kapitel 1. An dieser Stelle bemerken wir nur die Formel

$$\langle v, w \rangle = \overline{\langle w, v \rangle}\,, \qquad\qquad v, w \in \mathbb{C}^n.$$

Ferner ist das dyadische Produkt, oder auch Tensorprodukt, durch

$$v \otimes w := v w^{\mathsf{T}} \in \mathbb{K}^{n \times n}$$

gegeben. Es gilt damit $(v \otimes w)_{ij} = v_i w_j$. Es ist bilinear, d.h. für v fest ist $w \mapsto v \otimes w$ linear und für w fest ist $v \mapsto v \otimes w$ linear. Wir hatten schon in Abschnitt 181 gesehen, dass im Reellen auch das Skalarprodukt bilinear ist.

Das folgende Lemma zeigt, wie sich die Transposition von reellen Matrizen mit dem Skalarprodukt verhält.

Lemma 9.3.5

Für alle $m, n \geq 1$, alle Matrizen $A \in \mathcal{M}_{\mathbb{K}}(m,n)$ und alle $x \in \mathbb{K}^n$ und $y \in \mathbb{K}^m$ gilt

$$\langle y, Ax \rangle = \langle \overline{A}^{\mathsf{T}} y, x \rangle \, .$$

Beweis.

Einfaches Nachrechnen ergibt

$$\begin{aligned} \langle y, Ax \rangle = \overline{y}^{\mathsf{T}} Ax = \overline{y}^{\mathsf{T}} \big((Ax)^{\mathsf{T}} \big)^{\mathsf{T}} = \big((Ax)^{\mathsf{T}} \overline{y} \big)^{\mathsf{T}} = \big(x^{\mathsf{T}} A^{\mathsf{T}} \overline{y} \big)^{\mathsf{T}} &= \big(\langle \overline{x}, A^{\mathsf{T}} \overline{y} \rangle \big)^{\mathsf{T}} \\ &= \langle \overline{x}, A^{\mathsf{T}} \overline{y} \rangle \\ &= \overline{\langle x, \overline{A}^{\mathsf{T}} y \rangle} \\ &= \langle \overline{A}^{\mathsf{T}} y, x \rangle \, , \end{aligned}$$

wobei wir ausgenutzt haben, dass $\langle \overline{x}, A^{\mathsf{T}} \overline{y} \rangle$ eine 1×1-Matrix ist und damit gleich ihrer Transposition ist. □

Das Lemma 9.3.5 motiviert die folgende Definition.

Definition 9.3.6

Seien $m, n \geq 1$ und $S : \mathbb{K}^n \to \mathbb{K}^m$ linear. Dann heißt

$$S^* := L_{\overline{M(S)}^T} : \mathbb{K}^m \to \mathbb{K}^n$$

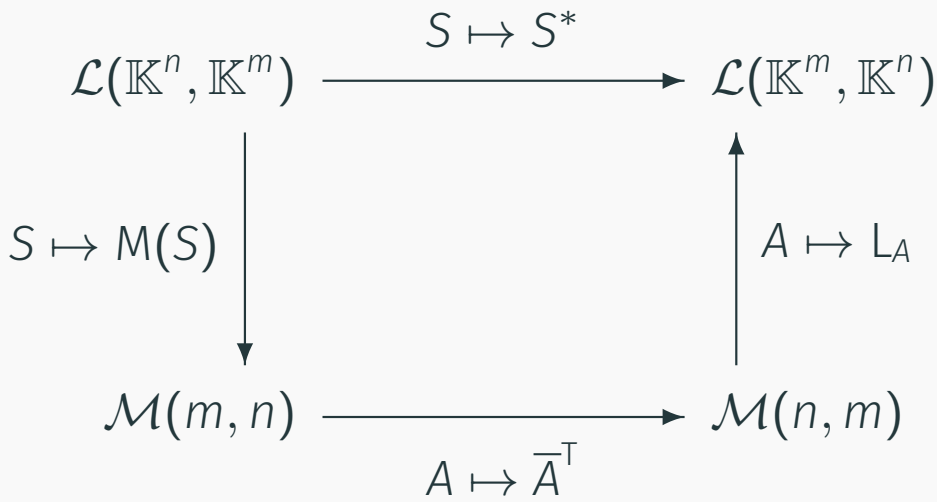
der adjungierte Operator von S.

ADJUNGIERTE ABBILDUNG: BEMERKUNGEN

Ist $A \in \mathcal{M}(m,n)$ die darstellende Matrix von S, d.h. $\mathsf{L}_A = S$, dann wird S^* von $\overline{A}^{\mathsf{T}}$ dargestellt, d.h.

$$S^* = \mathsf{L}_{\overline{A}^{\mathsf{T}}} \,. \tag{9.3.7}$$

Dies entspricht $(\mathsf{L}_A)^* = \mathsf{L}_{\overline{A}^{\mathsf{T}}}$. Die Konstruktion kann auch an dem Diagramm



verdeutlicht werden. Mit Theorem 9.3.3 ist dann sofort ersichtlich, dass die Abbildung

$$\begin{aligned} \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) &\to \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n) \\ S &\mapsto S^* \end{aligned}$$

linear ist.

ADJUNGIERTE ABBILDUNG: BEMERKUNGEN

Im Komplexen ist die Abbildung weiterhin additiv, es gilt aber nur noch $(\lambda S)^* = \overline{\lambda} S^*$.

Ferner gilt in beiden Fällen $(S^*)^* = S$ und aus Theorem 9.3.4 folgt auch

$$(S \circ R)^* = R^* \circ S^*$$

für alle linearen Abbildungen $R : \mathbb{K}^l \to \mathbb{K}^n$ und $S : \mathbb{K}^n \to \mathbb{K}^m$.

Wegen $\mathsf{L}_A x = Ax$ und Lemma 9.3.5 haben wir schließlich für $S = \mathsf{L}_A$ und $S^* = \mathsf{L}_{\overline{A}^\mathsf{T}}$ die Gleichung

$$\langle S^* y, x \rangle = \langle \mathsf{L}_{\overline{A}^\mathsf{T}} y, x \rangle = \langle \overline{A}^\mathsf{T} y, x \rangle = \langle y, Ax \rangle = \langle y, \mathsf{L}_A x \rangle = \langle y, Sx \rangle \tag{9.3.8}$$

für alle $x \in \mathbb{K}^n$ und $y \in \mathbb{K}^m$. Das folgende Resultat zeigt, dass diese Gleichung die Abbildung S^* eindeutig bestimmt.

Theorem 9.3.7

Sei $S : \mathbb{K}^n \to \mathbb{K}^m$ linear. Dann gibt es genau eine lineare Abbildung $T : \mathbb{K}^m \to \mathbb{K}^n$ mit

$$\langle Ty, x\rangle = \langle y, Sx\rangle$$

für alle $x \in \mathbb{K}^n$ und $y \in \mathbb{K}^m$. Diese Abbildung ist $T := S^*$.

Beweis.

Die Gleichung (9.3.8) zeigt, dass $T := S^*$ die Gleichung erfüllt.
Um die Eindeutigkeit zu zeigen, nehmen wir an, dass $T \in \mathcal{L}(\mathbb{K}^m, \mathbb{K}^n)$ die Gleichung erfüllt. Wir setzen $A := \mathsf{M}(S)$ und $B := \mathsf{M}(T)$ und bezeichnen die Einträge mit a_{ij}, bzw. b_{ji} für $i = 1, \dots, m$ und $j = 1, \dots, n$. Ist $e_i \in \mathbb{K}^m$ der i-te Einheitsvektor in $\mathbb{K}^m$ und $e_j \in \mathbb{K}^n$ der j-te Einheitsvektor in $\mathbb{K}^n$, so gilt

$$b_{ji} = \langle e_j, Te_i\rangle = \overline{\langle Te_i, e_j\rangle} = \overline{\langle e_i, Se_j\rangle} = \overline{a}_{ij}\,.$$

Dies zeigt $B = \overline{A}^{\mathsf{T}}$ und damit auch $T = \mathsf{L}_B = \mathsf{L}_{\overline{A}^{\mathsf{T}}} = S^*$. □

Eine quadratische Matrix $D \in \mathcal{M}(n,n)$ mit den Einträgen d_{ij} heißt **Diagonalmatrix**, falls $d_{ij} = 0$ für alle $i \neq j$ gilt. Wir schreiben

$$D = \begin{pmatrix} d_{11} & & 0 \\ & \ddots & \\ 0 & & d_{nn} \end{pmatrix}.$$

Diagonalmatrizen werden also eindeutig durch den Vektor $d = (d_{11}, \ldots, d_{nn})^{\mathsf{T}}$ ihrer **Diagonaleinträge** beschrieben. Wir vereinfachen in der Regel die Notation zu $d = (d_1, \ldots, d_n)^{\mathsf{T}}$.

Eine Diagonalmatrix $E_n \in \mathcal{M}(n,n)$, deren Diagonaleinträge alle gleich 1 sind, heißt n-te **Einheitsmatrix** und wir schreiben

$$E_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

Mit (9.3.1) ist es leicht nachzurechnen, dass

$$AE_n = E_nA = A \tag{9.3.9}$$

für alle $A \in \mathcal{M}(n,n)$ gilt. Die Einheitsmatrix E_n ist also ein neutrales Element

Eine quadratische Matrix $A \in \mathcal{M}(n,n)$ heißt **invertierbar**, falls es eine Matrix $A^{-1} \in \mathcal{M}(n,n)$ gibt mit

$$A^{-1}A = AA^{-1} = E_n\,. \tag{9.3.10}$$

Aus (9.3.9) folgt sofort, dass E_n invertierbar ist für $E_n^{-1} := E_n$. Sind ferner $A, B \in \mathcal{M}(n,n)$ invertierbar, so ist auch AB invertierbar, denn es gilt für $C := B^{-1}A^{-1}$

$$CAB = B^{-1}A^{-1}AB = B^{-1}E_nB = B^{-1}B = E_n$$

und analog auch $ABC = E_n$. Schreiben wir

$$\mathcal{GL}(n) := \left\{A \in \mathcal{M}(n,n) :\ A \text{ ist invertierbar } \right\},$$

so erhalten wir daher eine Gruppe, die aber im Fall $n > 1$ nicht kommutativ ist. Insbesondere ist die durch (9.3.10) gegebene Matrix A^{-1} eindeutig, siehe Abschnitt 87. Im folgenden nennen A^{-1} die **inverse Matrix** von $A \in \mathcal{GL}(n)$ und $\mathcal{GL}(n)$ die **allgemeine lineare Gruppe vom Grad** n. Wie in allen Gruppen haben wir u.a.

$$(AB)^{-1} = B^{-1}A^{-1}\,, \qquad\qquad A, B \in \mathcal{GL}(n).$$

LINEARE GRUPPE IST KEIN UNTERRAUM

Schließlich bemerken wir, dass die Nullmatrix $0 \in \mathcal{M}(n,n)$ nicht invertierbar ist, und damit gilt insbesondere

$$\mathcal{GL}(n) \neq \mathcal{M}(n,n)\,.$$

Aus dem gleichen Grund ist $\mathcal{GL}(n)$ auch kein Unterraum von $\mathcal{M}(n,n)$.

Der folgende Satz stellt einen Zusammenhang zwischen invertierbaren Matrizen und Isomorphismen her.

Theorem 9.3.8

Sei $n \geq 1$. Dann gilt $\mathrm{L}_{E_n} = \mathrm{id}_{\mathbb{K}^n}$ und $A \in \mathcal{M}(n,n)$ ist invertierbar, genau dann wenn L_A bijektiv ist. In diesem Fall gilt

$$\mathrm{L}_{A^{-1}} = \left(\mathrm{L}_A\right)^{-1}. \tag{9.3.11}$$

Haben wir eine lineare Abbildung $S : \mathbb{K}^n \to \mathbb{K}^n$, so zeigt der obige Satz durch Betrachtung von $A := \mathrm{M}(S)$, d.h. $\mathrm{L}_A = S$, dass S genau dann bijektiv ist, wenn A invertierbar ist. In diesem Fall gilt ferner mit der Formel (9.3.11):

$$\mathrm{M}(S^{-1}) = \mathrm{M}(\left(\mathrm{L}_A\right)^{-1}) = \mathrm{M}(\mathrm{L}_{A^{-1}}) = A^{-1} = (\mathrm{M}(S))^{-1}, \tag{9.3.12}$$

d.h. die darstellende Matrix der inversen Abbildung S^{-1} ist die inverse Matrix der darstellenden Matrix der Abbildung S.

BEWEIS

Die Formel $\mathsf{L}_{E_n} = \mathbf{id}_{\mathbb{K}^n}$ folgt sofort aus $\mathsf{L}_{E_n} x = E_n x = x$ für alle $x \in \mathbb{K}^n$.

Sei nun A invertierbar. Dann gilt

$$\mathbf{id}_{\mathbb{K}^n} = \mathsf{L}_{E_n} = \mathsf{L}_{AA^{-1}} = \mathsf{L}_A \circ \mathsf{L}_{A^{-1}}$$

und analog auch $\mathbf{id}_{\mathbb{K}^n} = \mathsf{L}_{A^{-1}} \circ \mathsf{L}_A$. Damit ist L_A bijektiv und $\mathsf{L}_{A^{-1}}$ ist die Umkehrabbildung von L_A.

Sei nun L_A bijektiv und B die darstellende Matrix von $(\mathsf{L}_A)^{-1}$, d.h. $\mathsf{L}_B = (\mathsf{L}_A)^{-1}$. Dann folgt

$$\mathbf{id}_{\mathbb{K}^n} = \mathsf{L}_B \circ \mathsf{L}_A = \mathsf{L}_{BA}$$

und damit $BA = E_n$. Analog sehen wir $AB = E_n$, d.h. B erfüllt (9.3.10). Dies zeigt die Invertierbarkeit von A und die Eindeutigkeit der inversen Matrix ergibt $B = A^{-1}$.

Im folgenden wollen wir die Basiswechselmatrizen, die wir am Ende von Kapitel 2 kennengelernt haben, genauer untersuchen. Wir fangen dazu mit dem folgenden Lemma an:

Lemma 9.3.9

Sei $\mathcal{A} := (v_1, \ldots, v_n)$ eine Basis von $\mathbb{K}^n$. Dann ist die $n \times n$-Matrix $(v_1, \ldots, v_n)$ invertierbar und für

$$A := (v_1, \ldots, v_n)^{-1}$$

und die Koordinatenabbildung $C_{\mathcal{A}}$ gilt

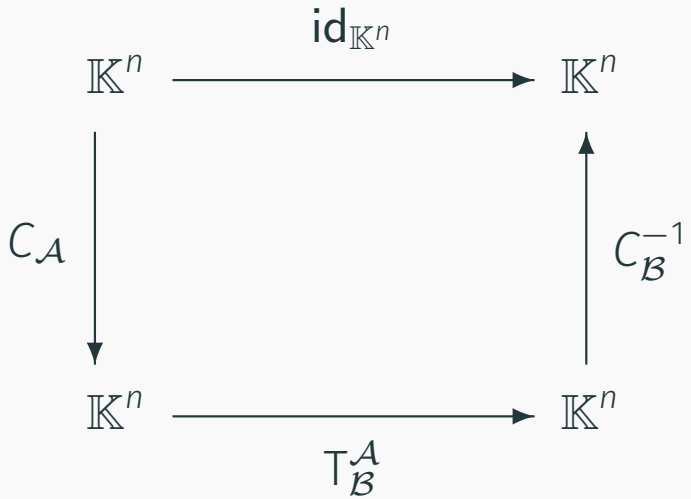
$$\mathrm{M}(C_{\mathcal{A}}) = A \qquad \text{und} \qquad \mathrm{M}(C_{\mathcal{A}}^{-1}) = A^{-1}.$$

Wegen $C_{\mathcal{A}}^{-1} e_i = v_i$ für alle $i = 1, \ldots, n$ gilt dann

$$\mathrm{M}(C_{\mathcal{A}}^{-1}) = (v_1, \ldots, v_n)$$

nach Korollar 9.2.12. Nun sind Koordinatenabbildungen bijektiv, siehe Satz 9.2.4, und damit zeigt Satz 9.3.8, dass die $n \times n$-Matrix $(v_1, \ldots, v_n)$ invertierbar ist. Wegen $A^{-1} = (v_1, \ldots, v_n)$ folgt dann die zweite Gleichung. Die erste Gleichung folgt schließlich aus der zweiten und (9.3.12).

Am Ende von Kapitel 2 hatten wir schon gesehen, dass für $V := \mathbb{K}^n$ mit zwei Basen $\mathcal{A} := (v_1, \ldots, v_n)$ und $\mathcal{B} = (w_1, \ldots, w_n)$ die Basiswechselmatrix $T_{\mathcal{B}}^{\mathcal{A}}$ durch



beschrieben ist, d.h. es galt

$$T_{\mathcal{B}}^{\mathcal{A}} := M_{\mathcal{B}}^{\mathcal{A}}(\mathrm{id}_{\mathbb{K}^n}) = M(C_{\mathcal{B}} \circ \mathrm{id}_{\mathbb{K}^n} \circ C_{\mathcal{A}}^{-1}) = M(C_{\mathcal{B}} \circ C_{\mathcal{A}}^{-1}).$$

Setzen wir nun

$$A := (v_1, \ldots, v_n)^{-1} \qquad \text{und} \qquad B := (w_1, \ldots, w_n)^{-1}, \tag{9.3.13}$$

so zeigen (9.3.2) und Lemma 9.3.9, dass

$$\mathsf{T}_{\mathcal{B}}^{\mathcal{A}} = \mathsf{M}(C_{\mathcal{B}} \circ C_{\mathcal{A}}^{-1}) = \mathsf{M}(C_{\mathcal{B}})\mathsf{M}(C_{\mathcal{A}}^{-1}) = BA^{-1} \tag{9.3.14}$$

gilt. Man beachte hierbei, dass wir $A^{-1} = (v_1, \ldots, v_n)$ haben, d.h., diese Matrix ist ohne zusätzliche Mühe sofort erhältlich. Dagegen müssen wir zur Berechnung von B eine Matrix invertieren.

Im Fall der Standard-Basis $\mathcal{E}_n = (e_1, \ldots, e_n)$ haben wir $(e_1, \ldots, e_n) = E_n$ und damit auch $E_n^{-1} = E_n$. Mit (9.3.14) ergibt dies dann

$$\mathrm{T}_{\mathcal{E}_n}^{\mathcal{A}} = (v_1, \ldots, v_n) \qquad \text{und} \qquad \mathrm{T}_{\mathcal{B}}^{\mathcal{E}_n} = B = (w_1, \ldots, w_n)^{-1}.$$

Die erste Formel haben wir schon am Ende von Kapitel 2 kennengelernt. Insbesondere sind diese Basiswechsel transitiv im Sinne von

$$\mathrm{T}_{\mathcal{B}}^{\mathcal{E}_n} \mathrm{T}_{\mathcal{E}_n}^{\mathcal{A}} = \mathrm{T}_{\mathcal{B}}^{\mathcal{A}}.$$

Mit anderen Worten ist ein iterierter Basiswechsel $\mathcal{A} \to \mathcal{E}_n \to \mathcal{B}$ das gleiche wie ein direkter Basiswechsel $\mathcal{A} \to \mathcal{B}$. Man kann sich leicht überlegen, dass dies auch noch gilt, wenn wir $\mathcal{E}_n$ in der obigen Formel durch eine beliebige andere Basis $\mathcal{C}$ ersetzen. Ferner zeigen unsere obigen Formeln noch

$$\mathrm{T}_{\mathcal{A}}^{\mathcal{B}} = AB^{-1} = (BA^{-1})^{-1} = (\mathrm{T}_{\mathcal{B}}^{\mathcal{A}})^{-1}, \tag{9.3.15}$$

d.h. ein Basiswechsel $\mathcal{B} \to \mathcal{A}$ lässt sich durch Invertieren der Basiswechselmatrix $\mathcal{A} \to \mathcal{B}$ beschreiben.

Haben wir nun zwei endlich-dimensionale $\mathbb{K}$-Vektorräume V und W mit Basen $\mathcal{A}$ und $\mathcal{B}$, und eine lineare Abbildung $S : V \to W$, so wissen wir aus Theorem 9.2.13 schon, dass wir S mit Hilfe der Koordinatenabbildungen $C_{\mathcal{A}} : V \to \mathbb{K}^n$ und $C_{\mathcal{B}} : W \to \mathbb{K}^m$ durch eine Matrix A darstellen können, wobei $n := \mathbf{dim}\, V$ und $m := \mathbf{dim}\, W$ ist. Genauer gesagt ist die darstellende Matrix

$$\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(S) = \mathrm{M}(C_{\mathcal{B}} \circ S \circ C_{\mathcal{A}}^{-1})\,.$$

Haben wir nun weitere Basen $\mathcal{A}'$ und $\mathcal{B}'$ von V und W, so lässt sich S analog mit der Matrix

$$\mathrm{M}_{\mathcal{B}'}^{\mathcal{A}'}(S) = \mathrm{M}(C_{\mathcal{B}'} \circ S \circ C_{\mathcal{A}'}^{-1})$$

darstellen. Um nun $\mathrm{M}_{\mathcal{B}'}^{\mathcal{A}'}(S)$ aus $\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(S)$ zu berechnen, betrachten wir die Abbildung 20 und

$$\begin{aligned}\mathrm{T}_{\mathcal{A}'}^{\mathcal{A}} &:= \mathrm{M}(C_{\mathcal{A}'} \circ C_{\mathcal{A}}^{-1}) = A'A^{-1}\,,\\ \mathrm{T}_{\mathcal{B}'}^{\mathcal{B}} &:= \mathrm{M}(C_{\mathcal{B}'} \circ C_{\mathcal{B}}^{-1}) = B'B^{-1}\,,\end{aligned}$$

wobei die Matrizen $A, A' \in \mathcal{M}_{\mathbb{K}}(n,n)$ und $B, B' \in \mathcal{M}_{\mathbb{K}}(m,m)$ wie in (9.3.13) konstruiert werden.

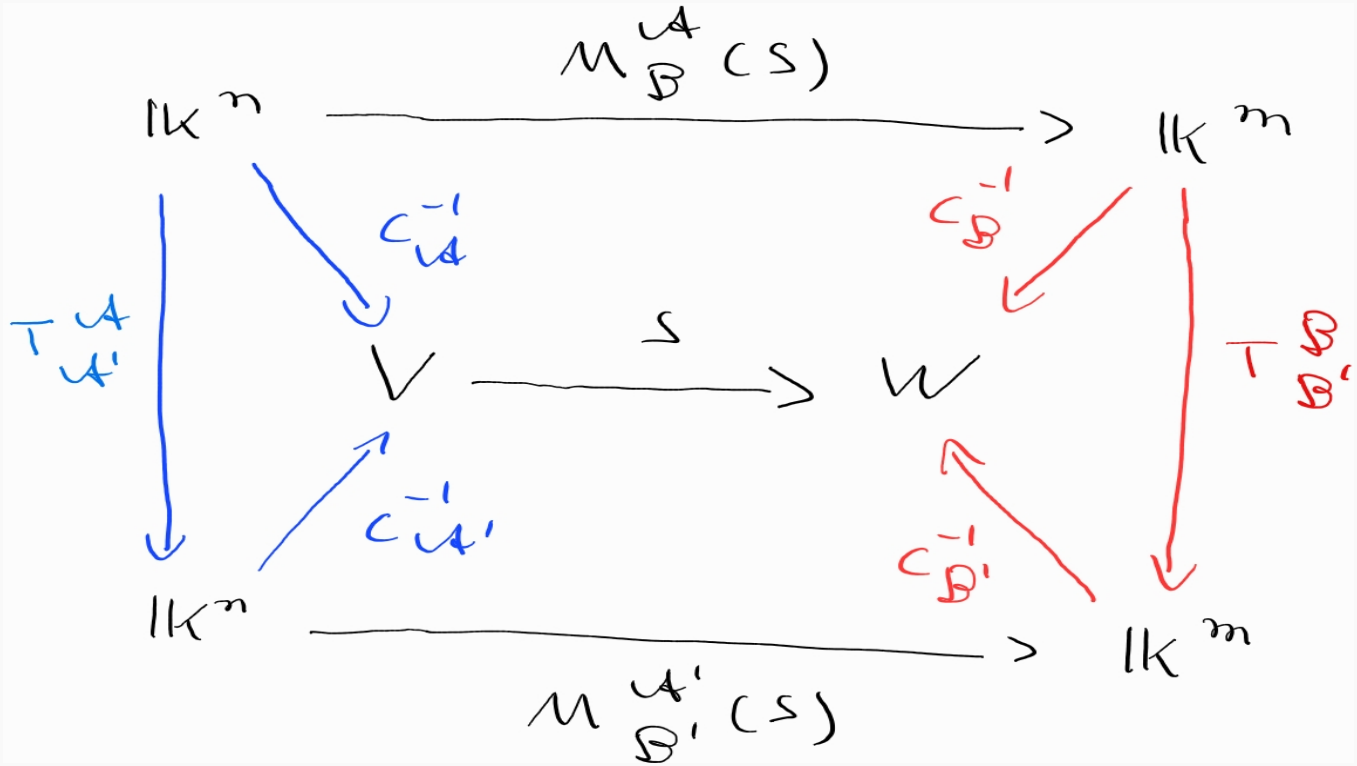
Da das Diagramm in Abbildung 20 kommutiert, haben wir dann

$$T^{\mathcal{B}}_{\mathcal{B}'} M^{\mathcal{A}}_{\mathcal{B}}(S) = M^{\mathcal{A}'}_{\mathcal{B}'}(S) T^{\mathcal{A}}_{\mathcal{A}'} ,$$

oder mit anderen Worten

$$M^{\mathcal{A}'}_{\mathcal{B}'}(S) = T^{\mathcal{B}}_{\mathcal{B}'} M^{\mathcal{A}}_{\mathcal{B}}(S)(T^{\mathcal{A}}_{\mathcal{A}'})^{-1} = T^{\mathcal{B}}_{\mathcal{B}'} M^{\mathcal{A}}_{\mathcal{B}}(S) T^{\mathcal{A}'}_{\mathcal{A}} , \tag{9.3.16}$$

wobei die erste Gleichung, die Gleichung ist, die in der Literatur in der Regel zu finden ist, und die zweite Gleichung (9.3.15) ausgenutzt hat.



Section 9.4
Der Gauss-Algorithmus

Lineare Gleichungssysteme treten in einer Vielzahl von Anwendungen auf. In diesem Unterkapitel wollen wir einen einen systematischen Ansatz zur Lösung solcher Systeme kennenlernen. Dabei werden wir auch eine Möglichkeit kennenlernen, Matrizen zu invertieren.

Um mit einem Beispiel zu beginnen, nehmen wir an, dass wir ein Rechteck mit Umfang 40cm haben und wissen, dass die Seite x_1 um 2cm länger ist, als die Seite x_2. Dies führt zu den beiden Gleichungen

$$\begin{aligned} 2x_1 + 2x_2 &= 40 \\ x_1 &= x_2 + 2\,, \end{aligned}$$

was nach Umstellen zu

$$\begin{aligned} 2x_1 + 2x_2 &= 40 \\ x_1 - x_2 &= 2 \end{aligned}$$

wird. Setzen wir

$$A := \begin{pmatrix} 2 & 2 \\ 1 & -1 \end{pmatrix} \qquad \text{und} \qquad b := \begin{pmatrix} 40 \\ 2 \end{pmatrix}$$

so suchen wir dann nach der Lösungsmenge $\{x \in \mathbb{R}^2 : Ax = b\}$.

Ist A invertierbar, so ist L_A invertierbar nach Satz 9.3.8, und es gibt daher genau ein $x \in \mathbb{R}^2$ mit $Ax = b$. Ist ferner A^{-1} bekannt, so können wir durch Multiplizieren mit A^{-1} von links sofort

$$x = E_2 x = A^{-1}Ax = A^{-1}b$$

bekommen. In der Regel ist A aber nicht quadratisch, und selbst wenn es quadratisch ist, wissen wir weder ob A invertierbar ist noch wie A^{-1} aussieht.

Ein **lineares Gleichungssystem (LGS)** mit m Gleichungen und n Unbekannten ist ein Gleichungssystem der Form

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ \vdots \qquad\qquad &= \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \,, \end{aligned}$$

wobei

$$A := \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \qquad \text{und} \qquad b := \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

bekannt sind und Lösungsvektoren $x \in \mathbb{K}^n$ gesucht sind. Im Folgenden heißt A **Koeffizientenmatrix** des linearen Gleichungssystems und wir schreiben das Gleichungssystem kurz als

$$Ax = b \,.$$

Das Gleichungssystem heißt **homogen**, falls $b = 0$ ist, ansonsten heißt es **inhomogen**. Die Lösungsmenge eines homogen Gleichungssystems ist nach Definition gleich dem Kern der zugehörigen linearen Abbildung

$$\ker A := \ker L_A = \{x \in \mathbb{K}^n : Ax = 0\}.$$

Insbesondere ist 0 immer eine Lösung eines homogenen LGS.

Ist M eine $m \times n$-Matrix mit Einträgen a_{ij}, d.h.

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$$

so heißen

$$\begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} \quad \text{und} \quad (a_{i1}, \dots, a_{in})$$

der j-te **Spaltenvektor** bzw. der i-te **Zeilenvektor** von A.

Ablesen der Lösung in bestimmten Fällen

Ist D eine Diagonalmatrix und sind alle Komponenten der Diagonalen $d = (d_{11}, \ldots, d_{nn})^\mathsf{T}$ ungleich 0, so ist D^{-1} die Diagonalmatrix zu der Diagonalen $(d_{11}^{-1}, \ldots, d_{nn}^{-1})^\mathsf{T}$. Damit ist

$$x := D^{-1}b = \begin{pmatrix} d_{11}^{-1}b_1 \\ \vdots \\ d_{nn}^{-1}b_n \end{pmatrix}$$

Lösung des zugehörigen LGS $Dx = b$.

Ablesen der Lösung in bestimmten Fällen

Ist allgemeiner $A \in \mathcal{M}(n,n)$ von der Form

$$A = \begin{pmatrix} 1 & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix},$$

wobei alle Einträge unterhalb der Diagonalen gleich 0 sind und alle Einträge auf der Diagonalen gleich 1 sind, so lässt sich die Lösung von $Ax = b$ iterativ durch

$$\begin{aligned} x_n &= b_n \\ x_{n-1} + a_{n-1,n} x_n &= b_{n-1}\,, \qquad \text{d.h.} \qquad x_{n-1} = b_{n-1} - a_{n-1,n} b_n \end{aligned}$$

usw. bestimmen. Im Folgenden wollen wir daher allgemeine $A \in \mathcal{M}(m,n)$ auf eine ähnliche Form bringen. Hierbei werden wir systematisch Zeilen multiplizieren und addieren.

Im Folgenden nehmen wir an, dass wir eine Matrix $A \in \mathcal{M}(m,n)$ und einen Vektor $b \in \mathbb{K}^m$ haben. Wir schreiben

$$(A,b) := \begin{pmatrix} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} & b_m \end{pmatrix},$$

d.h. (A,b) ist eine $m \times (n+1)$-Matrix. Der Gauss-Algorithmus besteht nun aus den folgenden Schritten:

(a) Suche Zeile der Matrix A mit $a_{i1} \neq 0$. Falls es eine solche gibt, vertausche die 1. und die i-Zeile von A und springe zu (c).

(b) Suche eine Spalte $j \in \{2, \dots, n\}$, für die es ein i mit $a_{ij} \neq 0$ gibt. Falls es eine solche gibt, vertausche die 1. und die j-te Spalte von A und die Variablen x_1 und x_j und springe zurück zu (a). Ansonsten wird der Algorithmus beendet.

(c) Durch (a) bzw. (b) ist $a_{11} \neq 0$ garantiert. Addiere nun das $-\frac{a_{j1}}{a_{11}}$-fache der 1. Zeile zu den übrigen Zeilen für $i = 2, \ldots, m$ und teile danach die 1. Zeile durch a_{11}. Das Ergebnis ist

$$\begin{pmatrix} 1 & \tilde{a}_{12} & \ldots & \tilde{a}_{1n} & \tilde{b}_1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & \tilde{a}_{m2} & \ldots & \tilde{a}_{mn} & \tilde{b}_m \end{pmatrix},$$

wobei der erste Spaltenvektor gleich $e_1 \in \mathbb{K}^m$ ist.

(d) Falls $m \geq 2$ und $n \geq 2$, gehe zu (a) mit der neuen $(m-1) \times n$ Matrix

$$(\tilde{A}, \tilde{b}) := \begin{pmatrix} \tilde{a}_{22} & \ldots & \tilde{a}_{2n} & \tilde{b}_2 \\ \vdots & & \vdots & \vdots \\ \tilde{a}_{m2} & \ldots & \tilde{a}_{mn} & \tilde{b}_m \end{pmatrix}.$$

Ansonsten wird der Algorithmus beendet.

Hierbei ist zu beachten, dass (a) und (c) die Lösungsmenge des LGS nicht verändern. Der Schritt (b) bedeutet nur, dass wir die Suche von der ersten Zeile auf alle weiteren ausdehnen. Streng genommen kann dieser Schritt weggelassen werden, wenn wir stattdessen die Schritte (c) und (d) entsprechend modifiziert ausführen. Die Beschreibung des Algorithmus und seiner Ausgabe wird aber dadurch unnötig verkompliziert. Der Schritt (d) bedeutet, dass wir im Folgenden die 1. Zeile nicht weiter bearbeiten.

DER GAUSS-ALGORITHMUS: ERGEBNIS

Nach dem Ausführen des Gauss-Algorithmus haben wir ein $0 \leq r \leq m, n$ und eine Matrix

$$A' = \begin{pmatrix} a'_{11} & \dots & a'_{1r} & a'_{1,r+1} & \dots & a'_{1n} & b'_1 \\ \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \dots & a'_{rr} & a'_{r,r+1} & \dots & a'_{rn} & b'_r \\ 0 & & 0 & 0 & & 0 & b_{r+1} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & & 0 & 0 & & 0 & b_m \end{pmatrix},$$

was dem Gleichungssystem

$$\begin{aligned} a'_{11}x'_1 + a'_{12}x'_2 + \dots + a'_{1n}x'_n &= b'_1 \\ &\vdots \\ a'_{rr}x'_r + \dots + a'_{rn}x'_n &= b'_r \\ 0 &= b'_{r+1} \\ &\vdots \\ 0 &= b'_m \end{aligned}$$

mit durch (c) umnummerierten Variablen $x'_1 \dots, x'_n$ entspricht. Ferner garantiert der Algorithmus $a'_{11} \neq 0$, ..., $a'_{rr} \neq 0$.

DER GAUSS-ALGORITHMUS: ERGEBNIS

Im Folgenden wollen wir dieses neue LGS analysieren. Dazu betrachten wir die folgenden Fälle:

(1) Es gibt ein $i \in \{r+1, \ldots, m\}$ mit $b'_i \neq 0$, so steht dies im Widerspruch zu der impliziten Annahme, dass es eine Lösung x und damit auch eine Lösung x' gibt. Damit hat dann das lineare Gleichungssystem $Ax = b$ keine Lösung. Man beachte hierbei, dass im Falle $b = 0$ unsere Algorithmus auch $b' = 0$ ergibt, d.h. der Fall (1) kann für homogene LSG nicht auftreten.

(2) Es gilt $b'_{r+1} = \cdots = b'_m = 0$ und $r = n$. Dann ist $m \geq n$ und die r-te Gleichung im obigen Gleichungssystem ist

$$a'_{nn} x'_n = b'_n$$

Dann ist $x'_n = \frac{b'_n}{a'_{nn}}$, und die Komponenten $x'_{n-1}, \ldots, x'_1$ lassen sich durch sukzessives Einsetzen und Umformen berechnen. Durch Umkehren der Umnummerierung von x nach x' gibt es dann genau eine Lösung von $Ax = b$.

(3) Es gilt $b'_{r+1} = \cdots = b'_m = 0$ und $r < n$. Dann ist die r-te Gleichung im obigen Gleichungssystem

$$a'_{rr}x'_r + \cdots + a'_{rn}x'_n = b'_r\,.$$

Wir können dann die Variablen $x'_{r+1}, \ldots, x'_n$ frei wählen, und die Gleichung nach x'_r auflösen. Die Komponenten $x'_{r-1}, \ldots, x'_1$ lassen sich dann wieder durch sukzessives Einsetzen und Umformen berechnen und die Umkehrung der Umnummerierung ergibt dann die nicht ein-elementige Lösungsmenge von $Ax = b$. Im Fall $r = n - 1$ gibt es dann z.B. eine freie Variable x'_n, so dass die Lösungsmenge eine Gerade im $\mathbb{K}^n$ ist.

Man beachte, dass im Fall $m < n$, d.h. wir haben weniger Gleichungen als Unbekannte, das homogene System $Ax = 0$ zu $b' = 0$ führt. Wegen $r \leq m < n$ sind wir dann automatisch im Fall (3) und daher hat in diesem Fall das System $Ax = 0$ immer eine Lösung $x \neq 0$, d.h. es gilt

$$\dim \ker A \geq 1\,.$$

DER GAUSS-ALGORITHMUS: BEISPIELE

Um ein Beispiel durchzurechnen, betrachten wir das folgende Beispiel

$$\begin{aligned} x_1 + x_2 + x_3 &= 6 \\ x_1 + 2x_2 - x_3 &= 2 \\ 2x_1 + 3x_2 - 3x_3 &= -1\,. \end{aligned}$$

Dann sind

$$A := \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -1 \\ 2 & 3 & -3 \end{pmatrix} \qquad \text{und} \qquad \begin{pmatrix} 6 \\ 2 \\ -1 \end{pmatrix}\,.$$

DER GAUSS-ALGORITHMUS: BEISPIELE

Der Gauss-Algorithmus ergibt dann

1	1	1	6	
1	2	-1	2	
2	3	-3	-1	
1	1	1	6	(a) findet $a_{11} = 1 \neq 0$
0	1	-2	-4	(c) addiert das $-\frac{a_{21}}{a_{11}} = -1$ fache der 1. Zeile
0	1	-5	-13	(c) addiert das $-\frac{a_{31}}{a_{11}} = -2$ fache der 1. Zeile
1	1	1	6	(d) ignoriert 1. Zeile
0	1	-2	-4	(a) findet $\tilde{a}_{22} = 1 \neq 0$
0	0	-3	-9	(c) addiert das $-\frac{\tilde{a}_{32}}{\tilde{a}_{22}} = -1$ fache der 2. Zeile

Danach endet der Algorithmus, da $\tilde{A} = (-3)$ in (d) nur noch eine 1×1-Matrix ist. Wir haben dann $r = n$ und da es keine $b'_{r+1}, \ldots, b'_m$ gibt, sind wir im Fall (2). Da wir nie in (b) umgeordnet haben, gilt ferner $x' = x$. Die 3. Zeile ergibt nun $-3x_3 = -9$, d.h. $x_3 = 3$. Einsetzen von x_3 in die 2. Zeile ergibt

$$-4 = b'_2 = 1 \cdot x_2 - 2x_3 = x_2 - 6\,,$$

d.h. $x_2 = 2$. Einsetzen von x_2 und x_3 in die 1. Zeile ergibt

$$6 = b'_1 = x_1 + x_2 + x_3 = x_1 + 5\,,$$

d.h. $x_1 = 1$. Damit ist

$$x := \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

die eindeutige Lösung des obigen Gleichungssystems.

DER GAUSS-ALGORITHMUS: BEISPIELE

Als weitere Beispiele betrachten wir die Gleichungssysteme

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ x_1 - x_2 - x_3 &= 6 \\ 2x_1 + 7x_2 + 7x_3 &= 1 \end{aligned}$$

und

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ x_1 - x_2 - x_3 &= 5 \\ 2x_1 + 7x_2 + 7x_3 &= 1 \end{aligned}$$

Offensichtlich haben beide Gleichungssysteme die gleiche Koeffizientenmatrix A, aber verschiedene Vektoren

$$b_{(1)} := \begin{pmatrix} 3 \\ 6 \\ 1 \end{pmatrix} \qquad \text{und} \qquad b_{(2)} := \begin{pmatrix} 3 \\ 5 \\ 1 \end{pmatrix}$$

auf der rechten Seite. Wir lösen nun beide Systeme gleichzeitig, indem wir beide Vektoren $b_{(1)}$ und $b_{(2)}$ rechts an die Koeffizientenmatrix anfügen.

DER GAUSS-ALGORITHMUS: BEISPIELE

1	1	1	3	3	
1	-1	-1	6	5	
2	7	7	1	1	
1	1	1	3	3	(a) findet $a_{11} = 1 \neq 0$
0	-2	-2	3	2	(c) addiert das $-\frac{a_{21}}{a_{11}} = -1$ fache der 1. Zeile
0	5	5	-5	-5	(c) addiert das $-\frac{a_{31}}{a_{11}} = -2$ fache der 1. Zeile
1	1	1	3	3	(d) ignoriert 1. Zeile
0	-2	-2	3	2	(a) findet $a_{22} = -2 \neq 0$
0	0	0	2.5	0	(c) addiert das $-\frac{a_{32}}{a_{22}} = -2.5$ fache der 2. Zeile

Hierbei haben wie in den Anmerkungen auf der rechten Seite die Bezeichnung $\tilde{A}$ für die in (d) generierte Teilmatrix zur Vereinfachung durch A ersetzt. In beiden Fällen haben wir damit $r = 2 < 3 = m = n$ und damit sind wir jeweils entweder im Fall (1) oder (3). Außerdem haben wir nie Spalten in (b) vertauscht und damit haben wir $x' = x$.

Für das erste Gleichungssystem haben wir $b'_{(1),3} = 2.5 \neq 0$ und damit sind wir im Fall (1), d.h. es gibt keine Lösung des Systems $Ax = b_{(1)}$.

Für das zweite Gleichungssystem haben wir $b'_{(2),3} = 0$ und damit sind wir im Fall (3). Wir wählen die Variable $x_3 := t$ für $t \in \mathbb{K}$ frei und lösen die 2. Zeile nach x_2 auf

$$-2x_2 - 2t = 2 \qquad \Longrightarrow \qquad x_2 = -(2 + 2t)/2 = -1 - t$$

Nun setzen wir x_2 und x_3 in die 1. Zeile ein und lösen nach x_1, d.h.

$$x_1 + (-1 - t) + t = 3 \qquad \Longrightarrow \qquad x_1 = 4\,.$$

Damit ist

$$\left\{ \begin{pmatrix} 4 \\ -1 - t \\ t \end{pmatrix} : t \in \mathbb{R} \right\} = \left\{ \begin{pmatrix} 4 \\ -1 \\ 0 \end{pmatrix} + t \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} : t \in \mathbb{R} \right\}$$

die Lösungsmenge des Gleichungssystems $Ax = b_{(2)}$. Offensichtlich ist dies eine Gerade.

Mit dem Gauss-Algorithmus können auch Matrizen invertiert werden. Sei dazu $A \in \mathcal{M}(n,n)$ eine Matrix, zu der wir überprüfen wollen, ob sie invertierbar ist und, falls dies der Fall ist, wollen wir auch A^{-1} berechnen. Wir nehmen nun an, dass A invertierbar ist und dass $x_{(j)}$ die j-te Spalte von A^{-1} ist. Aus

$$AA^{-1} = E_n$$

folgt dann

$$Ax_{(j)} = e_j \, , j = 1, \ldots, n. \tag{9.4.1}$$

Wir haben also n lineare Gleichungssysteme mit jeweils der gleichen Koeffizientenmatrix A aber verschiedenen rechten Seiten. Wie im obigen Beispiel können wir diese Systeme simultan lösen.

MATRIZENINVERTIERUNG: BEISPIEL

Als Beispiel betrachten wir dazu die Matrix

$$A := \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{pmatrix}.$$

Der Gauss-Algorithmus liefert dann

1	1	1	1	0	0	
1	2	3	0	1	0	
1	3	6	0	0	1	
1	1	1	1	0	0	(a) findet $a_{11} = 1 \neq 0$
0	1	2	-1	1	0	(c) addiert das $-\frac{a_{21}}{a_{11}} = -1$ fache der 1. Zeile
0	2	5	-1	0	1	(c) addiert das $-\frac{a_{31}}{a_{11}} = -2$ fache der 1. Zeile
1	1	1	1	0	0	(d) ignoriert 1. Zeile
0	1	2	-1	1	0	(a) findet $a_{22} = 1 \neq 0$
0	0	1	1	-2	1	(c) addiert das $-\frac{a_{32}}{a_{22}} = -2$ fache der 2. Zeile

MATRIZENINVERTIERUNG: BEISPIEL

An dieser Stelle sehen wir, dass wir für alle drei Gleichungssysteme $r = n = m$ haben und somit im Fall (2) sind. Damit gibt es eindeutige $x_{(1)}, \ldots, x_{(n)}$, die die n-Gleichungssysteme (9.4.1) lösen. Insbesondere ist unsere Matrix A invertierbar. Wären wir stattdessen im Fall (1) oder (3) gelandet, wäre dagegen die Matrix nicht invertierbar gewesen.

MATRIZENINVERTIERUNG: BEISPIEL

Um A^{-1} zu berechnen, müssen wir jetzt nur noch ihre Spaltenvektoren $x_{(1)}, \ldots, x_{(n)}$ bestimmen. Dies kann im Prinzip wie oben geschehen. Es kann aber auch simultan ausgeführt werden, wobei das Ziel ist, die linke Seite im obigen Schema auf die Gestalt E_3 zu bringen:

1	1	1	1	0	0	
0	1	2	-1	1	0	
0	0	1	1	-2	1	
1	1	0	0	2	-1	addiere das $-\frac{a_{13}}{a_{33}} = -1$ fache der 3. Zeile
0	1	0	-3	5	-2	ddiere das $-\frac{a_{23}}{a_{33}} = -2$ fache der 3. Zeile
0	0	1	1	-2	1	teile durch $a_{33} = 1$
1	0	0	3	-3	1	addiere das $-\frac{a_{12}}{a_{22}} = -1$ fache der 2. Zeile
0	1	0	-3	5	-2	Zeile ist fertig
0	0	1	1	-2	1	Zeile ist fertig

MATRIZENINVERTIERUNG: BEISPIEL

Dann kann die Lösung abgelesen werden. So ist beispielsweise die Lösung des ersten Gleichungssystems $x_{(1)} = (3, -3, 1)^T$. Da $x_{(1)}, \ldots, x_{(n)}$ die Spaltenvektoren von A^{-1} sind, haben wir also insgesamt

$$A^{-1} = \begin{pmatrix} 3 & -3 & 1 \\ -3 & 5 & -2 \\ 1 & -2 & 1 \end{pmatrix} .$$

In Computern werden Zahlen entweder als Ganzzahlen oder als Fließkommazahlen dargestellt, wobei für numerische Berechnungen fast nur die Letzteren in Frage kommen. Diese haben die folgende Gestalt:

Vorzeichenbit	Exponent	Mantisse

Das Vorzeichenbit signalisiert dabei das Vorzeichen ± 1. Die dargestellte Zahl x hat dann den Wert

$$x = \text{Vorzeichen} \cdot \text{Mantisse} \cdot 2^{\text{Exponent}}$$

Ein typischer Wert ist dabei eine Darstellung in 32 Bit (**single precision** oder **float**), d.h. 4 Byte, wobei die Mantisse 23 Bit und der Exponent 8 Bit belegen. Die Genauigkeit ist dann 7 bis 8 Stellen im Dezimalsystem. Zum Beispiel ist

$$1234.56789$$

dann nicht exakt darstellbar sondern nur als

$$1234.56787109375\,,$$

siehe `https://www.h-schmidt.net/FloatConverter/IEEE754.html`, d.h. es tritt ein Fehler in der neunten Stelle auf.

Eine andere typische Darstellung ist 64 Bit (double precision oder double), d.h. 8 Byte, wobei die Mantisse 52 Bit und der Exponent 11 Bit belegen. Die Genauigkeit ist dann ca. 16 Stellen im Dezimalsystem.

In jedem Fall gibt es Zahlen, die nicht dargestellt werden können. Neben der Ungenauigkeit, die schon oben beschrieben wurde, betrifft dies insbesondere zu große Zahlen, zu kleine Zahlen und Zahlen, die zu nahe an 0 liegen.

Arithmetische Operationen werden in diesen Darstellungen in vielen Fällen mit Fehlern ausgeführt, die manchmal vernachlässigbar sind und manchmal nicht.

Wir wollen dies am Beispiel des Gauss-Algorithmus etwas genauer erläutern. Dazu nehmen wir der Einfachheit halber an, dass wir ein System haben, dass nur 3 Stellen Genauigkeit bei seinen Fließkommazahlen erlaubt, für dass aber $10^{-4} \neq 0$ gilt. Wir betrachten nun das folgende Gleichungssystem

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Ausführen des Gauss-Algorithmus ergibt dann als Ergebnis

$$\begin{array}{rr|r} 10^{-4} & 1 & 1 \\ 0 & 1 - 10^4 & 2 - 10^4 \end{array}$$

Dies ergibt die exakten Lösungen

$$x_2 = \frac{b_2'}{a_{22}'} = \frac{2 - 10^4}{1 - 10^4} = \frac{-9998}{-9999} \approx 1$$

und wegen $10^{-4} x_1 + x_2 = 1$ auch

$$x_1 = 10^4 (1 - x_2) = 10^4 \left(1 - \frac{9998}{9999}\right) = \frac{10^4}{9999} \approx 1.$$

Schauen wir uns aber an, wie die Matrix (A', b') auf unserem System gespeichert ist, so haben wir wegen der Genauigkeit von 3 Stellen $1 - 10^4 = -9999 \equiv -10000$ und $2 - 10^4 = -9998 \equiv -10000$. Damit ergibt der Gauss-Algorithmus auf unserem System

$$\begin{array}{rr|r} 10^{-4} & 1 & 1 \\ 0 & -10^4 & -10^4 \end{array}$$

Dies ergibt $\hat{x}_2 = 1$, was eine sinnvolle Approximation des wahren Wertes $x_2 = \frac{9998}{9999}$ ist. Für $\hat{x}_1$ ergibt sich aber

$$10^{-4}\hat{x}_1 + \hat{x}_2 = 1 \qquad \Longrightarrow \qquad \hat{x}_1 = 0\,.$$

Dies ist keine sinnvolle Approximation von $x_1 = \frac{10000}{9999} \approx 1$.

UMSETZUNG AM COMPUTER: MATRIXINVERTIERUNG

Man kann zeigen, dass der Gauss-Algorithmus $\mathcal{O}(n^3)$ Operationen benötigt, um ein $n \times n$-Gleichungssystem zu lösen oder eine $n \times n$-Matrix zu invertieren. Ohne zusätzliche Annahmen an die Koeffizientenmatrix ist dies nicht zu beschleunigen. Ferner braucht er $\mathcal{O}(n^2)$ Speicher. Speichert man beispielsweise alle Zahlen im double precision ab, so werden $8 \cdot n^2$ Bytes zur Speicherung einer $n \times n$ benötigt. Ignoriert man Speicher für das Betriebssystem und mögliche Anwendungen, so können auf einem 64GB-System, d.h. auf einem System mit 68.719.476.736 Bytes, daher maximal 8.589.934.592 Zahlen im double precision abgespeichert werden, was $n \leq 92681$ bedeutet. Dies ist nicht soviel wie man denkt!

Section 9.5
Der Rang einer Matrix

Wir betrachten die 2×3-Matrix

$$A := \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Dann sind die beiden Zeilenvektoren $(0, 1, 1)$ und $(1, 0, 1)$ von A linear unabhängig und damit spannen sie einen zwei-dimensionalen Unterraum vom $\mathbb{R}^3$ auf. Ferner sind die ersten beiden Spaltenvektoren von A gleich e_2 und e_1. Damit spannen die drei Spaltenvektoren

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \text{und} \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

den $\mathbb{R}^2$ auf, also auch einen zwei-dimensionalen Raum. Ist dies Zufall?

Um diese Frage zu beantworten, definieren wir zunächst die beiden Größen.

Definition 9.5.1

Sei $A \in \mathcal{M}_{\mathbb{K}}(m,n)$ mit Zeilenvektoren $a_1, \ldots, a_m \in \mathbb{K}^n$ und Spaltenvektoren $a^{(1)}, \ldots, a^{(n)} \in \mathbb{K}^m$. Dann setzen wir

$$\begin{aligned} \textbf{zeilenrang}\, A &:= \textbf{dim}\, \textbf{span}\{a_1, \ldots, a_m\}\,, \\ \textbf{spaltenrang}\, A &:= \textbf{dim}\, \textbf{span}\{a^{(1)}, \ldots, a^{(n)}\}\,. \end{aligned}$$

Außerdem wollen wir noch den Rang einer linearen Abbildung definieren.

Definition 9.5.2

Seien V und W zwei $\mathbb{K}$-Vektorräume und $S : V \to W$ linear. Dann heißt

$$\mathbf{rang}\, S := \mathbf{dim}\, S(V) := \mathbf{dim}\{Sv : v \in V\}$$

der Rang von S.

Das folgende Lemma stellt einen ersten Zusammenhang zwischen den verschiedenen Größen her.

Lemma 9.5.3

Sei $A \in \mathcal{M}_{\mathbb{K}}(m,n)$. Dann gilt

$$\begin{aligned} \textsf{spaltenrang}\, A &= \textsf{rang}\, \mathsf{L}_A\,, \\ \textsf{zeilenrang}\, A &= \textsf{rang}(\mathsf{L}_A)^*\,. \end{aligned}$$

Mit Korollar 9.2.12 und $S := \mathsf{L}_A$ gilt

$$\begin{aligned}\textbf{spaltenrang}\, A = \textbf{dim}\,\textbf{span}\{Se_1, \ldots, Se_n\} &= \textbf{dim}\left\{\sum_{j=1}^{n} \lambda_j Se_j : \lambda_j \in \mathbb{K}\right\} \\ &= \textbf{dim}\left\{S\left(\sum_{j=1}^{n} \lambda_j e_j\right) : \lambda_j \in \mathbb{K}\right\} \\ &= \textbf{dim}\, S(\mathbb{K}^n) \\ &= \textbf{rang}\, S\,.\end{aligned}$$

Dies zeigt die erste Gleichung. Die zweite Gleichung folgt dann aus

$$\textbf{zeilenrang}\, A = \textbf{spaltenrang}\, \overline{A}^{\mathsf{T}} = \textbf{rang}\, \mathsf{L}_{\overline{A}^{\mathsf{T}}} = \textbf{rang}(\mathsf{L}_A)^*\,,$$

wobei wir im letzten Schritt die Definition von adjungierten Operatoren verwendet haben. Ferner haben wir im ersten Schritt im Fall $\mathbb{K} = \mathbb{C}$ die Tatsache benutzt, dass zu jeder Basis $v_1, \ldots, v_n \in \mathbb{C}^n$ von $\mathbb{C}^n$ auch $\overline{v}_1, \ldots, \overline{v}_n \in \mathbb{C}^n$ eine Basis von $\mathbb{C}$ ist. Letzteres lässt sich mit der Identität

$$\sum_{i=1}^{n} \lambda_i \overline{v}_i = \overline{\sum_{i=1}^{n} \overline{\lambda}_i v_i}$$

leicht zeigen.

Der folgende Satz, der als **Dimensionsformel** bekannt ist, beschreibt, wie der Kern und das Bild einer linearen Abbildung miteinander zusammenhängen.

Theorem 9.5.4

Seien V und W zwei $\mathbb{K}$-Vektorräume mit $\dim V < \infty$ und sei $S : V \to W$ linear. Dann gilt

$$\dim V = \dim \ker S + \operatorname{rang} S = \dim \ker S + \dim \operatorname{ran} S\,.$$

Ist ferner $u_1, \ldots, u_k \in V$ eine Basis von $\ker S$ und $w_1, \ldots, w_r \in W$ eine Basis von $\operatorname{ran} S$, dann ist für jede Wahl von $v_i \in V$ mit $Sv_i = w_i$ für alle $i = 1, \ldots, r$ die Familie

$$\mathcal{A} := (u_1, \ldots, u_k, v_1, \ldots, v_r)$$

eine Basis von V.

BEWEIS 1/2

Wir zeigen zunächst $V = \operatorname{span} \mathcal{A}$. Hierbei ist zu beachten dass die Inklusion $\operatorname{span} \mathcal{A} \subset V$ trivial ist. Um die andere Inklusion zu zeigen, wählen wir uns ein $v \in V$ und setzen $w := Sv$.

Dann gibt es $\lambda_1, \ldots, \lambda_r \in \mathbb{K}$ mit $w = \lambda_1 w_1 + \cdots + \lambda_r w_r$. Nach Wahl der v_i folgt

$$Sv = w = \sum_{i=1}^{r} \lambda_i w_i = \sum_{i=1}^{r} \lambda_i S v_i = S\left(\sum_{i=1}^{r} \lambda_i v_i \right)$$

Damit ist

$$v - \sum_{i=1}^{r} \lambda_i v_i \in \ker S$$

und folglich gibt es $\alpha_1, \ldots, \alpha_k \in \mathbb{K}$ mit

$$v - \sum_{i=1}^{r} \lambda_i v_i = \sum_{j=1}^{k} \alpha_j u_k \,.$$

Umstellen nach v zeigt dann $v \in \operatorname{span} \mathcal{A}$.

BEWEIS 2/2

Um die lineare Unabhängigkeit von $\mathcal{A}$ zu zeigen, wählen wir uns $\alpha_1, \dots, \alpha_k \in \mathbb{K}$ und $\lambda_1, \dots, \lambda_r \in \mathbb{K}$ mit

$$0 = \sum_{j=1}^{k} \alpha_j u_k + \sum_{i=1}^{r} \lambda_i v_i \,. \tag{9.5.1}$$

Dies ergibt

$$0 = S0 = S\left(\sum_{j=1}^{k} \alpha_j u_k\right) + S\left(\sum_{i=1}^{r} \lambda_i v_i\right) = 0 + \sum_{i=1}^{r} \lambda_i S v_i = \sum_{i=1}^{r} \lambda_i w_i$$

und damit folgt $\lambda_1 = \cdots = \lambda_r = 0$. Setzen wir dies in (9.5.1) ein, so erhalten wir

$$0 = \sum_{j=1}^{k} \alpha_j u_k$$

und dies impliziert $\alpha_1 = \cdots = \alpha_k = 0$.

Die Formel folgt dann aus $\mathbf{dim\,ker}\, S = k$ und $\mathbf{dim\,ran}\, S = r$ und der Tatsache, dass $\mathcal{A}$ eine Basis von V ist.

Die Dimensionsformel hat die folgende bemerkenswerte Konsequenz, die zeigt, dass im Falle gleicher Dimension der beteiligten Räume, Bijektivität das gleiche ist wie Injektivität oder Surjektivität.

Korollar 9.5.5

Seien V und W zwei $\mathbb{K}$-Vektorräume mit $\dim V = \dim W < \infty$ und sei $S : V \to W$ linear. Dann sind die folgenden Aussagen äquivalent:

i). S ist bijektiv.

ii). S ist injektiv.

iii). S ist surjektiv.

BEWEIS

Die Implikationen *i)* $\Rightarrow$ *ii)* und *i)* $\Rightarrow$ *iii)* sind trivial.

ii) $\Rightarrow$ *i)*. Ist S injektiv, so gilt $\mathbf{dim}\,\mathbf{ker}\,S = 0$ und damit folgt mit der Dimensionsformel

$$\mathbf{dim}\,W = \mathbf{dim}\,V = \mathbf{rang}\,S\,.$$

Da $\mathbf{rang}\,S$ ein Unterraum von W ist, folgt $\mathbf{ran}\,S = W$ nach Korollar 9.1.13. Damit ist S auch surjektiv.

iii) $\Rightarrow$ *i)*. Ist S surjektiv, haben wir $\mathbf{ran}\,S = W$ und damit $\mathbf{rang}\,S = \mathbf{dim}\,W$. Die Dimensionsformel zeigt dann

$$\mathbf{dim}\,W = \mathbf{dim}\,V = \mathbf{dim}\,\mathbf{ker}\,S + \mathbf{rang}\,S = \mathbf{dim}\,\mathbf{ker}\,S + \mathbf{dim}\,W$$

und damit $\mathbf{dim}\,\mathbf{ker}\,S = 0$. Dies ergibt die Injektivität von S.

Die Aussage des Korollars 9.5.5 ist im Fall $\dim V = \dim W = \infty$ falsch. Als Beispiel betrachten wir dazu zunächst die lineare Abbildung

$$\begin{aligned} I : \mathcal{C}([0,1]) &\to \mathcal{C}([0,1]) \\ f &\mapsto \left(x \mapsto If(x) := \int_0^x f(t)\,\mathrm{d}t \right). \end{aligned}$$

Diese ist injektiv, denn wenn $If(x) = 0$ für alle $x \in [0,1]$ gilt, so folgt mit dem Hauptsatz Differential- und Integralrechnung, siehe Satz 7.2.1, $f(x) = (If)'(x) = 0$ für alle $x \in [0,1]$. Nach Lemma 9.2.7 ist dann die Abbildung I injektiv. Sie ist aber nicht surjektiv, da jedes If nach Satz 7.2.1 differenzierbar ist, es aber stetige, nicht differenzierbare Funktionen gibt.

BEMERKUNG

Ein weiteres Beispiel liefert die Abbildung

$$D : \mathrm{Pol}(\mathbb{R}) \to \mathrm{Pol}(\mathbb{R})$$
$$f \mapsto f' .$$

Diese ist surjektiv, da zu gegebenen $g \in \mathrm{Pol}(\mathbb{R})$ mit $g(x) = \sum_{k=0}^{n} a_k x^k$ wir für

$$f(x) := \sum_{k=0}^{n} \frac{a_k}{k+1} x^{k+1}, \qquad x \in \mathbb{R},$$

die Gleichung $Df = g$ haben. Sie ist aber nicht injektiv, da $D\mathbf{1}_{\mathbb{R}} = 0$ gilt.

Das folgende Korollar zeigt, dass invertierbare Matrizen genau die Matrizen sind, bei denen wir vollen Spalten- und Zeilenrang haben.

Korollar 9.5.6

Sei $n \geq 1$ und $A \in \mathcal{M}_{\mathbb{K}}(n,n)$. Dann sind äquivalent:

i). *A ist invertierbar.*

ii). *$\overline{A}^{\mathsf{T}}$ ist invertierbar.*

iii). $\textbf{spaltenrang}\, A = n$.

iv). $\textbf{zeilenrang}\, A = n$.

In diesem Fall gilt $(\overline{A}^{\mathsf{T}})^{-1} = \overline{(A^{-1})}^{\mathsf{T}}$.

i) $\Rightarrow$ *ii)*. Sei A invertierbar. Dann gilt für $B := \overline{(A^{-1})}^{\mathsf{T}}$:

$$\overline{A}^{\mathsf{T}} B = \overline{A}^{\mathsf{T}} \overline{(A^{-1})}^{\mathsf{T}} = \overline{(A^{-1}A)}^{\mathsf{T}} = \overline{E}_n^{\mathsf{T}} = E_n$$

und analog auch $B\overline{A}^{\mathsf{T}} = E_n$. Dies zeigt, dass $\overline{A}^{\mathsf{T}}$ invertierbar ist.

ii) $\Rightarrow$ *i)*. Wende die Implikation *i)* $\Rightarrow$ *ii)* auf $\overline{A}^{\mathsf{T}}$ an und benutze dabei $\overline{(\overline{A}^{\mathsf{T}})}^{\mathsf{T}} = A$.

i) $\Leftrightarrow$ *iii)*. Die Matrix A ist nach Satz 9.3.8 genau dann invertierbar, wenn L_A bijektiv ist. Dies ist nach Korollar 9.5.5 äquivalent zur Surjektivität von L_A. Da $\mathbf{ran}\,\mathsf{L}_A$ ein Unterraum von $\mathbb{K}^n$ ist, ist nach Korollar 9.1.13 die Surjektivität von L_A wiederum äquivalent zu

$$\mathbf{spaltenrang}\, A = \mathbf{rang}\, \mathsf{L}_A = \mathbf{dim}\, \mathbf{ran}\, \mathsf{L}_A = \mathbf{dim}\, \mathbb{K}^n = n\,,$$

wobei wir in der ersten Gleichheit Lemma 9.5.3 ausgenutzt haben.

ii) $\Leftrightarrow$ *iv)*. Wende die Äquivalenz *i)* $\Leftrightarrow$ *iii)* auf $\overline{A}^{\mathsf{T}}$ an und benutze die Identität $\mathbf{spaltenrang}\, \overline{A}^{\mathsf{T}} = \mathbf{zeilenrang}\, A$.

Das folgende Korollar zeigt, dass wir für jede lineare Abbildung Basen finden, für die die darstellende Matrix besonders einfach sind.

Korollar 9.5.7

Seien V und W endlich-dimensionale $\mathbb{K}$-Vektorräume und $S : V \to W$ linear. Dann gibt es Basen $\mathcal{A}$ und $\mathcal{B}$ von V und W, so dass für $r := \operatorname{rang} S$ gilt

$$\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(S) = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix},$$

wobei die Nullen Teilmatrizen darstellen, deren Einträge alle gleich 0 sind.

Sei $w_1, \dots, w_r$ eine Basis von $\mathbf{ran}\, S$. Wir ergänzen diese zu einer Basis

$$\mathcal{B} := (w_1, \dots, w_r, w_{r+1}, \dots, w_m)$$

von W. Jetzt wählen wir eine Basis $u_1, \dots, u_k \in V$ von $\mathbf{ker}\, S$, sowie $v_i \in V$ mit $Sv_i = w_i$ für alle $i = 1, \dots, r$. Nach Satz 9.5.4 ist dann

$$\mathcal{A} := (v_1, \dots, v_r, u_1, \dots, u_k)$$

eine Basis von V. Wegen $Sv_i = w_i$ für alle $i = 1, \dots, r$ zeigt dann Teil *ii)* von Satz 9.2.13

$$\begin{pmatrix} a_{11} & \dots & a_{1r} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mr} \end{pmatrix} = \begin{pmatrix} E_r \\ 0 \end{pmatrix}$$

für die ersten r Spalten von $\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(S)$. Da $Su_i = 0$ für alle $i = 1, \dots, k$, zeigt Teil *ii)* von Satz 9.2.13 auch

$$\begin{pmatrix} a_{1,r+1} & \dots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,r+1} & \dots & a_{m,n} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

für die verbleibenden Spalten von $\mathrm{M}_{\mathcal{B}}^{\mathcal{A}}(S)$, wobei wir $n := r + k$ gesetzt haben.

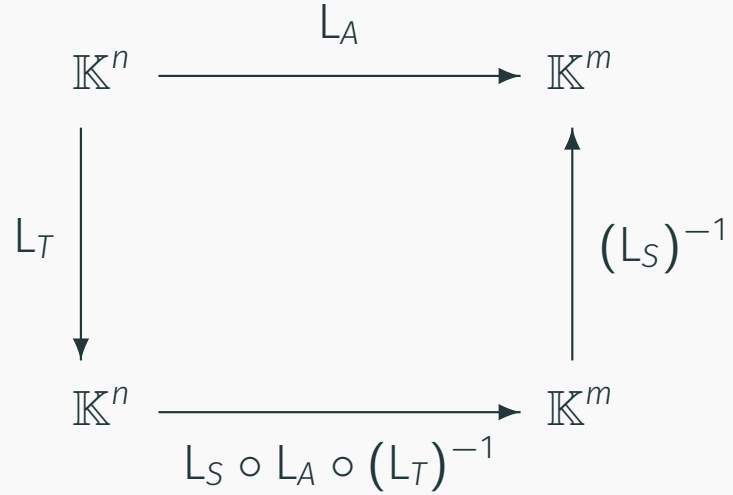
Das folgende Lemma zeigt, dass sich Spalten- und Zeilenrang nicht ändern, wenn mit invertierbaren Matrizen multipliziert wird.

Lemma 9.5.8

Sei A eine $m \times n$-Matrix, sowie $S \in \mathcal{M}(m, m)$ und $T \in \mathcal{M}(n, n)$ invertierbare Matrizen. Dann gilt

$$\begin{aligned} \textbf{spaltenrang}(SAT^{-1}) &= \textbf{spaltenrang}\, A\,, \\ \textbf{zeilenrang}(SAT^{-1}) &= \textbf{zeilenrang}\, A\,. \end{aligned}$$

Wir betrachten das kommutative Diagramm



wobei L_T und $(\mathsf{L}_S)^{-1}$ Isomorphismen sind. Es gilt dann $\mathsf{L}_S \circ \mathsf{L}_A \circ (\mathsf{L}_T)^{-1} = \mathsf{L}_{SAT^{-1}}$ und da Isomorphismen Dimensionen nicht ändern gilt

$$\textbf{rang}\, \mathsf{L}_A = \textbf{dim ran}\, \mathsf{L}_A = \textbf{dim ran}\, \mathsf{L}_{SAT^{-1}} = \textbf{rang}\, \mathsf{L}_{SAT^{-1}}\,.$$

Nun folgt die erste Gleichung aus Lemma 9.5.3. Die zweite Formel kann analog durch Transposition gezeigt werden.

Der folgende Satz zeigt, dass der Spaltenrang und der Zeilenrang einer beliebigen Matrix gleich sind.

Theorem 9.5.9

Sei A eine $m \times n$-Matrix und $r := \textbf{rang}\, \mathrm{L}_A$. Dann gibt es invertierbare Matrizen $S \in \mathcal{M}(m, m)$ und $T \in \mathcal{M}(n, n)$ mit

$$SAT^{-1} = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}$$

und es gilt $\textbf{spaltenrang}\, A = \textbf{zeilenrang}\, A = r$.

BEWEIS

Für $R := \mathsf{L}_A : \mathbb{K}^n \to \mathbb{K}^m$ gibt es nach Korollar 9.5.7 Basen $\mathcal{A}$ und $\mathcal{B}$ von $\mathbb{K}^n$ und $\mathbb{K}^m$, so dass für $r := \textbf{rang}\, R$ gilt

$$\mathsf{M}_{\mathcal{B}}^{\mathcal{A}}(R) = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Offensichtlich gilt

$$\textbf{spaltenrang}\, \mathsf{M}_{\mathcal{B}}^{\mathcal{A}}(R) = r = \textbf{zeilenrang}\, \mathsf{M}_{\mathcal{B}}^{\mathcal{A}}(R)\,. \tag{9.5.2}$$

Ferner haben wir nach Konstruktion $A = \mathsf{M}(R) = \mathsf{M}_{\mathcal{E}_m}^{\mathcal{E}_n}(R)$. Mit der Basiswechsel-Formel (9.3.16) folgt dann

$$A = \mathsf{M}_{\mathcal{E}_m}^{\mathcal{E}_n}(R) = \mathsf{T}_{\mathcal{E}_n}^{\mathcal{B}} \mathsf{M}_{\mathcal{B}}^{\mathcal{A}}(R) \mathsf{T}_{\mathcal{A}}^{\mathcal{E}_n}$$

und setzt man dann

$$S := (\mathsf{T}_{\mathcal{E}_n}^{\mathcal{B}})^{-1} = \mathsf{T}_{\mathcal{B}}^{\mathcal{E}_n} \qquad \text{und} \qquad T := \mathsf{T}_{\mathcal{A}}^{\mathcal{E}_n}\,,$$

so erhält man die erste Behauptung. Die zweite Behauptung folgt aus (9.5.2) und Lemma 9.5.8.

Das folgende Lemma charakterisiert eindeutig lösbare lineare Gleichungssysteme.

Korollar 9.5.10

Für jedes $A \in \mathcal{M}_{\mathbb{K}}(n,n)$ sind die folgenden Aussagen äquivalent:

i). *A ist invertierbar.*

ii). $\operatorname{rang} A = n$

iii). *$Ax = b$ ist für jedes b lösbar.*

iv). *$Ax = b$ ist für jedes b eindeutig lösbar.*

v). *$Ax = 0$ ist eindeutig lösbar.*

i) $\Leftrightarrow$ *ii)*. Dies hatten wir schon in Korollar 9.5.6 gesehen.

Um die restlichen Äquivalenzen zu zeigen, beachten wir zunächst, dass *ii)* zu der Surjektivität von L_A äquivalent ist. Ferner ist *iii)* zu der Bijektivität von L_A äquivalent und *iv)* ist nach Lemma 9.2.7 zu der Injektivität von L_A äquivalent. Die Äquivalenz von *ii)*, *iii)* und *iv)* folgt daher aus Korollar 9.5.5. Die Äquivalenz von *i)* und *iv)* hatten wir in Satz 9.3.8 gezeigt.

Section 9.6
Determinanten

Definition 9.6.1

Sei V ein $\mathbb{K}$-Vektorraum und $n \geq 1$. Dann heißt eine Abbildung $\Delta : V^n \to \mathbb{K}$ **multilinear** oder auch ***n*-linear**, falls

$$\Delta(v_1, \dots v_{i-1}, v_i + w, v_{i+1}, \dots, v_n) = \Delta(v_1, \dots v_{i-1}, v_i, v_{i+1}, \dots, v_n) + \Delta(v_1, \dots v_{i-1}, w, v_{i+1}, \dots, v_n)$$

und

$$\Delta(v_1, \dots v_{i-1}, \lambda v_i, v_{i+1}, \dots, v_n) = \lambda\Delta(v_1, \dots v_{i-1}, v_i, v_{i+1}, \dots, v_n)$$

für alle $i \in \{1, \dots, n\}$, $v_j, w \in V$ und $\lambda \in \mathbb{K}$ gilt.

Offensichtlich sind im Fall $n = 1$ die multilinearen Abbildungen gerade die linearen Abbildungen $V \to \mathbb{K}$. Im Fall $n = 2$ sprechen wir auch von **bilinearen** Abbildungen und im Fall $n = 3$ von **trilinearen** Abbildungen.

Das Skalarprodukt auf $\mathbb{R}^n \times \mathbb{R}^n$, das wir in Abschnitt 181 eingeführt hatten, ist bilinear.

Für $x \in \mathbb{R}^n$ betrachten wir die Abbildung

$$\begin{aligned} x' : \mathbb{R}^n &\to \mathbb{R} \\ y &\mapsto \langle x, y \rangle . \end{aligned}$$

Diese ist linear und damit multilinear mit $n = 1$. Wir sprechen auch von einem Tensor erster Stufe.

Haben wir eine Matrix $A \in \mathcal{M}(n, n)$, so ist die Abbildung

$$\begin{aligned} \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R} \\ (x, y) &\mapsto x^{\mathsf{T}} A y = \langle x, Ay \rangle \end{aligned}$$

bilinear, wobei wir

$$\langle x, Ay \rangle = \sum_{i,j=1}^{n} a_{ij} x_i y_j$$

haben. Wir sprechen auch von einem Tensor zweiter Stufe.

Haben wir eine Familie $(b_{ijk}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$, so ist

$$\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

$$(x, y, z) \mapsto \sum_{i,j,k=1}^{n} b_{ijk} x_i y_j z_k$$

trilinear und wir sprechen von einem Tensor dritter Stufe.

Der folgende Satz, den wir an dieser Stelle nicht beweisen werden, zeigt die Existenz einer ausgezeichneten n-linearen Abbildung $\mathbb{K}^n \times \cdots \times \mathbb{K}^n \to \mathbb{K}$.

Theorem 9.6.2

Für alle $n \geq 1$ gibt es genau eine n-lineare Abbildung
$\det : \mathbb{K}^n \times \cdots \times \mathbb{K}^n \to \mathbb{K}$, *für die gilt:*

$$\det(e_1, \ldots, e_n) = 1\,, \tag{9.6.1}$$

$$\det(a_1, \ldots, a_i, \ldots, a_j, \ldots, a_n) = -\det(a_1, \ldots, a_j, \ldots, a_i, \ldots, a_n) \tag{9.6.2}$$

für alle $a_1, \ldots, a_n \in \mathbb{K}^n$ und $i, j \in \{1, \ldots, n\}$.

Im folgenden heißt die in Satz 9.6.2 betrachtete Abbildung $\det : \mathbb{K}^n \times \cdots \times \mathbb{K}^n \to \mathbb{K}$ **Determinante**. Ist $A \in \mathcal{M}(n,n)$ und sind $a_1, \ldots, a_n$ die Spaltenvektoren von A, so schreiben wir

$$\det A := \det(a_1, \ldots, a_n)$$

und sprechen von der **Determinante von** A. Wir haben damit eine Abbildung

$$\begin{aligned} \det : \mathcal{M}(n,n) &\to \mathbb{K} \\ A &\mapsto \det A \end{aligned}$$

die sich multilinear auf den Spaltenvektoren verhält. Die Bedingung (9.6.1) zeigt dann

$$\det E_n = 1$$

und die Bedingung (9.6.2) besagt, dass sich bei Vertauschen von zwei Spaltenvektoren das Vorzeichen der Determinante ändert.

Ist $\mathbf{vol} : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}$ die Funktion, die zu jeder Wahl von $a_1, \ldots, a_n \in \mathbb{R}^n$ das Volumen des Parallelotops

$$\left\{ \sum_{i=1}^{n} \lambda_i a_i : (\lambda_1, \ldots, \lambda_n) \in [0,1]^n \right\}$$

zuordnet, so kann man

$$\mathbf{vol}(A) = |\det A| \tag{9.6.3}$$

zeigen, wobei A die $n \times n$-Matrix mit Spaltenvektoren $a_1, \ldots, a_n$ ist. Hierbei ist

$$\mathbf{vol}(E_n) = 1 = |\det E_n|$$

offensichtlich, und entsteht A' aus A durch Vertauschen von zwei Spaltenvektoren, so gilt auch

$$\mathbf{vol}(A) = \mathbf{vol}(A'),$$

was zu $|\det A'| = |-\det A| = |\det A|$ passt.

Ferner ist die Funktion **vol** homogen in jeder Komponente, d.h.

$$\mathbf{vol}(a_1, \dots a_{i-1}, \lambda a_i, a_{i+1}, \dots, a_n) = |\lambda| \, \mathbf{vol}(a_1, \dots a_{i-1}, a_i, a_{i+1}, \dots, a_n)$$

und dies passt zu

$$|\det(a_1, \dots a_{i-1}, \lambda a_i, a_{i+1}, \dots, a_n)| = |\lambda| \cdot |\det(a_1, \dots a_{i-1}, a_i, a_{i+1}, \dots, a_n)| \, .$$

Die Additivität der Volumenfunktion in jeder Komponente ist leider nicht so einfach zu erläutern.

Im folgenden haben wir immer eine $n \times n$-Matrix A mit Spaltenvektoren $a_1, \ldots, a_n$.

Haben wir dann $a_i = a_j$ für ein Paar (i, j) mit $i \neq j$, so folgt

$$\det A = 0\,. \tag{9.6.4}$$

Um dies zu sehen, betrachten wir nur den Fall $i = 1$ und $j = 2$. Dann gilt durch Vertauschen der ersten und zweiten Spalte

$$\det A = \det(a_1, a_2, a_3, \ldots, a_n) = -\det(a_2, a_1, a_3, \ldots, a_n) = -\det A\,,$$

wobei wir im letzten Schritt $a_1 = a_2$ ausgenutzt haben. Dies zeigt $2 \det A = 0$ und damit die Behauptung (9.6.4).

Gilt $a_i = 0$ für ein $i = 1, \ldots, n$, so gilt ebenfalls $\det A = 0$. Im Fall $i = 1$ sehen wir dies durch

$$\begin{aligned} \det(0, a_2, a_3, \ldots, a_n) &= \det(0 \cdot 0, a_2, a_3, \ldots, a_n) \\ &= 0 \cdot \det(0, a_2, a_3, \ldots, a_n) \\ &= 0 \end{aligned} \tag{9.6.5}$$

und der Fall $i > 1$ ist analog zu zeigen.

Kombiniert man diese Ergebnisse, so erhalten wir ferner

$$\begin{aligned}&\det(a_1, \lambda a_1 + a_2, a_3, \ldots, a_n)\\ &= \lambda \det(a_1, a_1, a_3, \ldots, a_n) + \det(a_1, a_2, a_3, \ldots, a_n)\\ &= \det(a_1, a_2, a_3, \ldots, a_n)\,,\end{aligned} \tag{9.6.6}$$

wobei wir im letzten Schritt (9.6.4) ausgenutzt haben. Natürlich gelten entsprechende Formeln auch für beliebige andere Spalten i und j mit $i \neq j$. Hierbei bemerken wir, dass diese Operation dem Schritt (c) im Gauss-Algorithmus entspricht, wenn man den Algorithmus auf A^{T} anwendet, um die obige Spaltenoperation in eine Zeilenoperation umzuwandeln.

Ist D eine Diagonalmatrix mit Diagonaleinträgen $d = (d_1, \ldots, d_n)^{\mathrm{T}}$, dann folgt aus der n-fach angewendeten Homogenität und der Normierung (9.6.1)

$$\det D = d_1 \cdot \ldots \cdot d_n \cdot \det E_n = d_1 \cdot \ldots \cdot d_n\,. \tag{9.6.7}$$

Das folgende Lemma verallgemeinert unsere obigen Beobachtungen.

Lemma 9.6.3

Ist $A \in \mathcal{M}(n, n)$ eine Matrix, die nicht invertierbar ist. Dann gilt

$$\det A = 0 .$$

BEWEIS

Wenn A nicht invertierbar ist, gilt **spaltenrang** $A < n$ nach Korollar 9.5.6. Damit sind die Spaltenvektoren $a_1, \ldots, a_n$ von A linear abhängig. Ohne Einschränkung nehmen wir an, dass es $\lambda_2, \ldots, \lambda_n \in \mathbb{K}$ gibt mit

$$a_1 = \sum_{j=2}^{n} \lambda_j a_j \, .$$

Ist $a_1 = 0$, so folgt dann die Behauptung aus (9.6.5). Ist $a_1 \neq 0$, so gibt es ein $j \geq 2$ mit $\lambda_j \neq 0$. Ohne Einschränkung nehmen wir $j = 2$ an. Dann erhalten wir

$$\begin{aligned}
\det(a_1, a_2, \ldots, a_n) &= \lambda_2^{-1} \det(a_1, \lambda_2 a_2, \ldots, a_n) \\
&= \lambda_2^{-1} \det(a_1, \lambda_2 a_2 + \lambda_3 a_3, \ldots, a_n) \\
&\ldots \\
&= \lambda_2^{-1} \det\left(a_1, \sum_{j=2}^{n} \lambda_j a_j, \ldots, a_n\right) \\
&= \lambda_2^{-1} \det(a_1, a_1, \ldots, a_n) \\
&= 0 \, ,
\end{aligned}$$

wobei wir sukzessive (9.6.6) angewendet haben und schließlich (9.6.4) ausgenutzt haben.

Der folgende Satz, den wir an dieser Stelle nicht vollständig beweisen wollen, liefert ein paar weitere wichtige Rechenregeln für Determinanten.

Theorem 9.6.4

Seien $A, B \in \mathcal{M}(n, n)$. Dann gilt:

$$\det A^{\mathsf{T}} = \det A\,, \tag{9.6.8}$$

$$\det(AB) = \det A \cdot \det B \tag{9.6.9}$$

$$\det(AB) = \det(BA)\,. \tag{9.6.10}$$

Ferner ist A genau dann invertierbar, wenn $\det A \neq 0$ gilt und in diesem Fall haben wir

$$\det A^{-1} = \frac{1}{\det A}\,. \tag{9.6.11}$$

Hierbei beachten wir, dass wir für nicht invertierbare Matrizen schon $\det A = 0$ wissen. Ist umgekehrt A invertierbar, so gilt mit (9.6.9)

$$1 = \det E_n = \det AA^{-1} = \det A \cdot \det A^{-1}$$

und daher folgt (9.6.11). Ferner ist (9.6.10) eine Folgerung aus (9.6.9), denn es gilt

$$\det(AB) = \det A \cdot \det B = \det B \cdot \det A = \det(BA)\,.$$

Ist B invertierbar, so kann (9.6.9) recht einfach durch Betrachten der Funktion

$$A \mapsto \frac{\det(AB)}{\det B}$$

gezeigt werden: In der Tat reicht es zu überprüfen, dass diese Funktion alle Eigenschaften der Determinantenfunktion aus Satz (9.6.2) erfüllt. Mit der Eindeutigkeit aus Satz (9.6.2) folgt dann

$$\det A = \frac{\det(AB)}{\det B}\,.$$

Ist B nicht invertierbar, so ist auch AB nicht invertierbar und es folgt (9.6.2) mit Lemma 9.6.3.

Betrachten wir nochmal die Volumenfunktion, so hatten wir schon (9.6.3), d.h.

$$\mathbf{vol}(A) = |\det A|$$

gesehen. Sind die Spaltenvektoren in A linear abhängig, so ist das Parallelotops nicht “voll-dimensional” und wir bekommen

$$\mathbf{vol}(A) = |\det A| = 0\,.$$

Ist ferner $B \in \mathcal{M}(n, n)$ invertierbar, so zeigt (9.6.9)

$$\mathbf{vol}(BA) = |\det B|\,\mathbf{vol}(A)\,.$$

Man beachte hierbei, dass wir bei $\mathbf{vol}(A)$ das von $a_1, \ldots, a_n$ aufgespannte Parallelotop betrachten und bei $\mathbf{vol}(BA)$ das von $Ba_1, \ldots, Ba_n$ aufgespannte Parallelotop.

In (9.6.7) hatten wir schon gesehen, dass Determinanten von Diagonalmatrizen einfach durch das Produkt der Diagonaleinträge zu berechnen sind. Dies gilt auch dann noch, wenn wir eine **obere Dreiecksmatrix**

$$A := \begin{pmatrix} a_{11} & & a_{1n} \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}$$

haben, d.h. eine Matrix, bei der alle Einträge unterhalb der Diagonalen gleich 0 sind. Ist nämlich ein Diagonaleintrag $a_{ii} = 0$, so ist $Ax = 0$ nicht eindeutig lösbar, wie wir beim Gauss-Algorithmus gesehen haben und Korollar 9.5.6 zeigt dann, dass A nicht invertierbar ist, d.h. $\det A = 0$ nach Lemma 9.6.3. Sind jedoch alle Diagonaleinträge $a_{ii} \neq 0$, so können wir die Matrix A durch Operationen der Form (9.6.6) auf Diagonalform bringen, ohne die Diagonale und die Determinante zu ändern. In diesem Zusammenhang bemerken wir, dass der Gauss-Algorithmus in seinem Schritt (c) genau Operationen der Form (9.6.6) durchführt. Ferner ändern die potentiellen Zeilen- oder Spaltenvertauschungen in den Schritten (a) und (b) nur das Vorzeichen, wobei die Aussage für die Zeilenvertauschungen aus $\det A = \det A^{\mathsf{T}}$ folgt.

Um diese Beobachtungen an Beispielen zu verdeutlichen, schreiben wir im folgenden $|A| := \det A$. Dann gilt z.B.

$$\begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix} = - \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = -1 \cdot 1 = -1 \,,$$

da wir einen Zeilentausch vorgenommen haben, und

$$\begin{vmatrix} 1 & 1 & 1 \\ 1 & 2 & -1 \\ 2 & 3 & -3 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & -3 \end{vmatrix} = 1 \cdot 1 \cdot (-3) = -3 \,,$$

wobei wir für die Berechnungen auf das erste Beispiel im Abschnitt 649 verweisen.

Daneben gibt es auch allgemeine Berechnungsformeln. So gilt zum Beispiel

$$\det A = \sum_{\sigma \in S_n} \operatorname{sign} \sigma \cdot a_{1,\sigma(1)} \cdot \dots \cdot a_{n,\sigma(n)} \,, \tag{9.6.12}$$

wobei die **symmetrische Gruppe** $S_n := \operatorname{Perm}(\{1, \dots, n\})$ die Menge alle Permutationen der n-elementigen Menge $\{1, \dots, n\}$ ist,

$$\operatorname{sign} \sigma := (-1)^{|\operatorname{inv}(\sigma)|} \qquad \text{und}$$

das **Vorzeichen einer Permutation** $\sigma \in S_n$ bezeichnet und

$$\operatorname{inv}(\sigma) := \{(i,j) : i < j \text{ und } \sigma(i) > \sigma(j)\}$$

die Menge der **Fehlstände** der Permutation σ ist. Leider ist die Formel (9.6.12) für größere n nicht effizient auswertbar, da wir schon im Satz 2.2.4 gesehen haben, dass $|S_n| = n!$.

Ist A_{ij} die Matrix, die aus der Matrix A durch Streichen der i-ten Zeile und j-ten Spalte entsteht, so besagt der **Laplacesche Entwicklungssatz**, dass

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} \tag{9.6.13}$$

für alle $j = 1, \ldots, n$ und

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} \tag{9.6.14}$$

für alle $i = 1, \ldots, n$ gilt. Im Fall (9.6.13) sprechen wir von der **Entwicklung nach der j-ten Spalte** und im Fall (9.6.14) sprechen wir von der **Entwicklung nach der i-ten Zeile**.

Im Fall von 2×2-Matrizen gilt daher mit der Entwicklung nach der ersten Zeile

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1} \det A_{1,1} - a_{1,2} \det A_{1,2} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}\,.$$

Im Fall von 3×3-Matrizen gilt daher mit der Entwicklung nach der ersten Zeile

$$\begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} = a_{1,1} \begin{vmatrix} a_{2,2} & a_{2,3} \\ a_{3,2} & a_{3,3} \end{vmatrix} - a_{1,2} \begin{vmatrix} a_{2,1} & a_{2,3} \\ a_{3,1} & a_{3,3} \end{vmatrix} + a_{1,3} \begin{vmatrix} a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} \end{vmatrix}$$

$$= a_{1,1}a_{2,2}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2}$$

$$- a_{1,3}a_{2,2}a_{3,1} - a_{1,1}a_{2,3}a_{3,2} - a_{1,2}a_{2,1}a_{3,3}$$

Diese Berechnungsvorschrift ist in Abbildung 21 graphisch veranschaulicht.

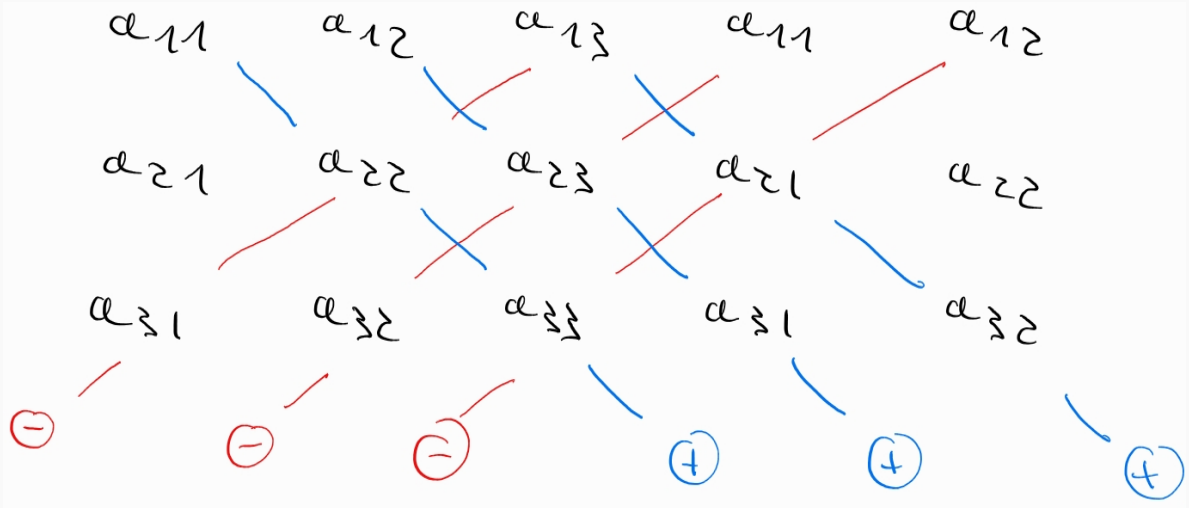


Abbildung: Berechnungshilfe für Determinanten von 3×3-Matrizen. Die ersten beiden Spalten von A werden nochmal rechts an A "angehängt". Die blauen Diagonalen werden multipliziert und bekommen ein positives Vorzeichen und die roten Diagonalen werden ebenfalls multipliziert, bekommen aber ein negatives Vorzeichen. Alle Produkte werden dann addiert.

CHAPTER 10: HILBERTRÄUME

Section 10.1
Grundlagen

Im Folgenden betrachten wir, sofern nicht anders erwähnt, nur den Fall $\mathbb{K} = \mathbb{R}$. Wir beginnen mit der Definition von Skalarprodukten, die unsere Betrachtungen von Abschnitt 181 von $\mathbb{R}^n$ auf $\mathbb{R}$-Vektorräume verallgemeinert.

Definition 10.1.1

Sei V ein $\mathbb{R}$-Vektorraum. Dann heißt eine Abbildung

$$\langle \cdot\,, \cdot \rangle : V \times V \to \mathbb{R}$$

Skalarprodukt, falls die folgenden Bedingungen erfüllt sind:

i). **Bilinearität:** Die Abbildung ist bilinear.

ii). **Positive Definitheit:** Für alle $v \in V$ gilt $\langle v, v \rangle \geq 0$ und

$$\langle v, v \rangle = 0 \qquad \Longleftrightarrow \qquad v = 0\,.$$

iii). **Symmetrie:** Für alle $v, w \in V$ gilt $\langle v, w \rangle = \langle w, v \rangle$.

Wir hatten schon in Abschnitt 181 gesehen, dass

$$\langle x, y\rangle := \sum_{i=1}^{d} x_i y_i \,, \qquad\qquad x, y \in \mathbb{R}^d \tag{10.1.1}$$

ein Skalarprodukt auf $\mathbb{R}^d$ definiert.

Auf dem Raum $\mathcal{C}([a, b])$ können wir ein Skalarprodukt durch

$$\langle f, g\rangle := \int_a^b f(x)g(x)\,\mathrm{d}x, \qquad\qquad f, g \in \mathcal{C}([a, b]) \tag{10.1.2}$$

definieren. Das Überprüfen der Eigenschaften eines Skalarprodukts ist in diesem Fall ebenfalls einfach und wird daher übersprungen.

Im Fall eines $\mathbb{C}$-Vektorraums V ist die Definition eines Skalarprodukts $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ etwas komplizierter. Genauer gesagt wird die Bilinearität durch die **Sesquilinearität** ersetzt, bei der

$$\langle \alpha v + \beta v', w \rangle = \overline{\alpha} \langle v, w \rangle + \overline{\beta} \langle v', w \rangle \,,$$
$$\langle v, \alpha w + \beta w \rangle = \alpha \langle v, w \rangle + \beta \langle v, w' \rangle$$

für alle $v, v', w, w' \in V$ und $\alpha, \beta \in \mathbb{C}$ ersetzt. Mit anderen Worten ist das Skalarprodukt in der ersten Komponente nur noch **semi-linear** und in der zweiten Komponente ist es weiterhin linear. Ferner muss die Symmetrie durch

$$\langle v, w \rangle = \overline{\langle v, w \rangle} \,, \qquad\qquad v, w \in V$$

ersetzt werden. Wir sagen dann auch, dass das Skalarprodukt **hermitisch** ist.

Auf $\mathbb{C}^d$ ist z.B.

$$\langle x, y \rangle := \sum_{i=1}^{d} \overline{x_i} y_i \,, \qquad\qquad x, y \in \mathbb{C}^d \tag{10.1.3}$$

ein Skalarprodukt.

Achtung: In der Literatur, insbesondere in der mathematischen, wird häufig alternativ gefordert, dass komplexe Skalarprodukte in der ersten Komponente linear und in der zweiten Komponente semi-linear sind. In diesem Fall müssen dann stattdessen in (10.1.3) die Terme $x_i \overline{y_i}$ aufsummiert werden.

Lemma 10.1.2

Seien V ein $\mathbb{R}$-Vektorraum und $\langle \cdot\,, \cdot \rangle$ ein Skalarprodukt auf V. Dann gelten die folgenden Aussagen:

i). **Norm:** *Die Abbildung*

$$\begin{aligned} \|\cdot\| : V &\to [0, \infty) \\ v &\mapsto \|v\| := \sqrt{\langle v, v \rangle} \end{aligned}$$

ist eine Norm auf V.

ii). **Cauchy-Schwarz-Ungleichung:** *Für alle $v, w \in V$ gilt*

$$\left|\langle v, w \rangle\right| \leq \|v\| \cdot \|w\| .$$

iii). **Polarisation:** *Für alle $v, w \in V$ gilt:*

$$4\langle v, w \rangle = \|v + w\|^2 - \|v - w\|^2 .$$

iv). **Parallelogrammgleichung:** *Für alle $v, w \in V$ gilt:*

$$\|v + w\|^2 + \|v - w\|^2 = 2\|v\|^2 + 2\|w\|^2 .$$

Die Polarisation besagt, dass wir aus der Norm das Skalarprodukt rekonstruieren können. Man kann zeigen, dass eine Norm genau dann durch ein Skalarprodukt definiert ist, wenn sie die Parallelogrammgleichung erfüllt.

Ist V ein $\mathbb{C}$-Vektorraum mit Skalarprodukt, so kann durch $\|v\| := \sqrt{\langle v, v\rangle}$ ebenfalls eine Norm auf V definiert werden, die die Cauchy-Schwarz'sche Ungleichung und die Parallelogrammgleichung erfüllt. Die Polarisations-Formel muss jedoch durch

$$4\langle v, w\rangle = \|v + w\|^2 - \|v - w\|^2 + \mathrm{i}\|v + \mathrm{i}w\|^2 - \mathrm{i}\|v - \mathrm{i}w\|^2$$

ersetzt werden.

BEWEIS 1/2

ii). Ist $w = 0$, so ist nicht zu zeigen. Ansonsten setzen wir

$$\lambda := -\frac{\langle v, w\rangle}{\langle w, w\rangle}\,.$$

Dann folgt

$$\begin{aligned} 0 \leq \langle v + \lambda w, v + \lambda w\rangle &= \langle v, v\rangle + 2\lambda\langle v, w\rangle + \lambda^2\langle w, w\rangle \\ &= \langle v, v\rangle - 2\frac{(\langle v, w\rangle)^2}{\langle w, w\rangle} + \frac{(\langle v, w\rangle)^2}{\langle w, w\rangle} \\ &= \langle v, v\rangle - \frac{(\langle v, w\rangle)^2}{\langle w, w\rangle}\,. \end{aligned}$$

Umformen ergibt dann die Behauptung.

BEWEIS 2/2

i). Die Definitheit und die Homogenität folgt direkt aus den Eigenschaften des Skalarprodukts. Um die Dreiecksungleichung zu zeigen, wählen wir uns $v, w \in V$. Dann gilt mit *ii)*:

$$\begin{aligned}\|v+w\|^2 = \|v\|^2 + 2\langle v, w\rangle + \|w\|^2 &\leq \|v\|^2 + 2\|v\| \cdot \|w\| + \|w\|^2 \\ &= \left(\|v\| + \|w\|\right)^2.\end{aligned}$$

Ziehen der Wurzel ergibt dann die Behauptung.

iii) und iv). Kann einfach nachgerechnet werden.

Definition 10.1.3

Ein $\mathbb{K}$-Vektorraum V mit Skalarprodukt und zugehöriger Norm heißt **Prä-Hilbertraum**. Ist V vollständig bezüglich dieser Norm, so heißt V **Hilbertraum**.

Die Räume $\mathbb{K}^n$ mit dem Skalarprodukt (10.1.1) bzw. (10.1.3) sind vollständig und damit Hilberträume. Die zugehörige Normen werden **euklidische Normen** genannt. Man kann ferner zeigen, dass alle endlich dimensionalen Prä-Hilberträume vollständig und damit sogar Hilberträume sind.

Der Raum $C([a, b])$ mit dem durch das Riemann-Integral definierte Skalarprodukt (10.1.2) ist *nicht* vollständig und damit auch kein Hilbertraum. Um ihn vollständig zu machen, muss man den Integralbegriff ausweiten.

Lemma 10.1.4

Jeder Teilraum W eines Prä-Hilbertraums V wird zu einem Prä-Hilbertraum, wenn wir das Skalarprodukt von V auf W einschränken.

Das folgende Lemma zeigt, dass Skalarprodukte bzgl. der von ihnen definierten Normen stetig sind.

Lemma 10.1.5

Sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Ist $\| \cdot \|$ die vom Skalarprodukt definierte Norm, dann gilt für alle $v, w \in V$ und $(v_n)_{n \in \mathbb{N}} \subset V$ und $(w_n)_{n \in \mathbb{N}} \subset V$ mit $v_n \to v$ und $w_n \to w$:

$$\langle v_n, w_n \rangle \to \langle v, w \rangle .$$

Mit der Cauchy-Schwarz-Ungleichung folgt

$$\begin{aligned}\left|\langle v_n, w_n\rangle - \langle v, w\rangle\right| &= \left|\langle v_n, w_n\rangle - \langle v_n, w\rangle + \langle v_n, w\rangle - \langle v, w\rangle\right| \\ &\leq \left|\langle v_n, w_n - w\rangle\right| + \left|\langle v_n - v, w\rangle\right| \\ &\leq \|v_n\| \cdot \|w_n - w\| + \|v_n - v\| \cdot \|w\| \,.\end{aligned}$$

Ferner hatten wir in Abschnitt 213 gesehen, dass aus der Konvergenz von $(v_n)_{n\in\mathbb{N}}$ die Beschränktheit der Folge folgt und dass dies wiederum die Existenz eines $B \geq 0$ mit $\|v_n\| \leq B$ für alle $n \geq 1$ impliziert. Setzen wir dieses in die obige Rechnung ein, so folgt die Behauptung für $n \to 0$.

Wir hatten schon in Kapitel 3 gesehen, dass Orthogonalität eine wichtige Rolle spielt. Dieses Konzept kann auf Prä-Hilberträume erweitert werden.

Definition 10.1.6

Sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Dann heißen $v, w \in V$ orthogonal, falls

$$\langle v, w \rangle = 0$$

gilt. In diesem Fall schreiben wir auch $v \perp w$.

Im Reellen gilt wegen $\langle v, w \rangle = \langle w, v \rangle$ offensichtlich $v \perp w \Leftrightarrow w \perp v$ und wegen $\langle v, w \rangle = \overline{\langle w, v \rangle}$ ist dies auch im Komplexen der Fall.

Sind v und w orthogonal, und $\lambda \in \mathbb{K}$, so sind auch λv und w orthogonal. Schließlich gilt $0 \perp w$ für alle $w \in V$ und aus der positiven Definitheit folgt

$$v \perp v \qquad \Longleftrightarrow \qquad v = 0\,.$$

Das folgende Resultat zeigt, dass der Satz von Pythagoras in allen Prä-Hilberträumen gilt.

Theorem 10.1.7

Sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Dann gilt für alle $v, w \in V$ mit $v \perp w$:

$$\|v + w\|^2 = \|v\|^2 + \|w\|^2 .$$

Beweis.

Es gilt

$$\begin{aligned} \|v + w\|^2 = \langle v + w, v + w \rangle &= \langle v, v \rangle + \langle v, w \rangle + \langle w, v \rangle + \langle w, w \rangle \\ &= \langle v, v \rangle + \langle w, w \rangle \\ &= \|v\|^2 + \|w\|^2 \end{aligned}$$

wobei wir $v \perp w \Leftrightarrow w \perp v$ ausgenutzt haben. □

Ist nun V ein Prä-Hilbertraum, $v \in V$ und $W \subset V$, so schreiben wir

$$v \perp W \qquad :\Longleftrightarrow \qquad v \perp w \text{ für alle } w \in W\,.$$

Ist ferner W ein Unterraum von V, so heißt

$$W^{\perp} := \{v \in V : v \perp W\}$$

das orthogonale Komplement von W in V. Einfaches Nachrechnen ergibt die Formeln

$$\{0\}^{\perp} = V\,,$$
$$V^{\perp} = \{0\}\,,$$

sowie das folgende Lemma.

Lemma 10.1.8

Sei V ein Prä-Hilbertraum und $W \subset V$ ein Teilraum. Dann ist $W^{\perp}$ ein Teilraum von V und es gilt $W \cap W^{\perp} = \{0\}$.

Section 10.2
Orthonormalbasen

Die Standardbasis $e_1, \ldots, e_n$ von $\mathbb{K}^n$ besteht aus paarweise orthogonalen Vektoren. Dieses wird in der folgenden Definition verallgemeinert.

Definition 10.2.1

Sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Dann heißt eine endliche Familie $v_1, \ldots, v_n \in V$ Orthogonalsystem, falls für alle $i, j = 1, \ldots, n$ gilt

$$v_i \perp v_j \quad \Longleftrightarrow \quad i \neq j \,.$$

Gilt zusätzlich $\|v_i\| = 1$ für alle $i = 1, \ldots, n$, so sprechen wir von einem Orthonormalsystem (ONS).

Die obigen beiden Definitionen lassen sich sofort auf beliebige Familien $(v_i)_{i \in I}$ ausdehnen. Dieses ist aber nur in unendlich-dimensionalen Räumen interessant, wie wir in Lemma 10.2.2 sehen werden.

BEMERKUNGEN

Ist $v_1, \ldots, v_n \in V$ ein Orthogonalsystem, so gilt $v_i \not\perp v_i$ für alle $i = 1, \ldots, n$. Damit haben wir $\|v_i\|^2 = \langle v_i, v_i \rangle > 0$ für alle i, und durch Normierung

$$w_i := \frac{v_i}{\|v_i\|}, \qquad i = 1, \ldots, n$$

erhalten wir ein Orthonormalsystem $w_1, \ldots, w_n$.

Ist $v_1, \ldots, v_n \in V$ ein Orthonormalsystem, so gilt für alle i, j:

$$\langle v_i, v_j \rangle = \delta_{ij},$$

wobei das Kronecker-Symbol δ_{ij} durch $\delta_{ij} := 0$ falls $i \neq j$ und $\delta_{ij} := 1$ falls $i = j$ definiert ist.

Offensichtlich ist die Standardbasis $e_1, \ldots, e_n$ von $\mathbb{K}^n$ ein Orthonormalsystem. Betrachten wir den Prä-Hilbertraum $\mathcal{C}([0, 2\pi])$ und setzen wir

$$f_k(x) := \cos(kx) \qquad k \geq 0$$
$$f_k(x) := \sin(-kx) \qquad k < 0$$

für alle $k \in \mathbb{Z}$ und $x \in [0, 2\pi]$, so ist jede endliche Familie $f_{k_1}, \ldots, f_{k_n}$ mit $k_i \neq k_j$ für $i \neq j$ ein Orthogonalsystem.

Das folgende Lemma zeigt, dass Orthogonalsysteme automatisch linear unabhängig sind.

Lemma 10.2.2

Sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$ und $v_1, \ldots, v_n \in V$ ein Orthogonalsystem. Dann sind $v_1, \ldots, v_n$ linear unabhängig.

BEWEIS

Seien $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$ mit $\lambda_1 v_1 + \cdots + \lambda_n v_n = 0$. Dann gilt für $j \in \{1, \ldots, n\}$:

$$0 = \left\langle \sum_{i=1}^{n} \lambda_i v_i, v_j \right\rangle = \sum_{i=1}^{n} \overline{\lambda_i} \langle v_i, v_j \rangle = \overline{\lambda_j} \langle v_j, v_j \rangle \,.$$

Aus $\langle v_i, v_j \rangle > 0$ folgt dann $\overline{\lambda_j} = 0$, d.h. $\lambda_j = 0$. Damit haben wir $\lambda_1 = \cdots = \lambda_n = 0$ gezeigt.

Lemma 10.2.2 legt die folgende Definition, nahe.

Definition 10.2.3

Sei V ein endlich-dimensionaler $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Dann heißt $v_1, \ldots, v_n \in V$ **Orthonormalbasis (ONB)** von V, falls $v_1, \ldots, v_n$ ein Orthonormalsystem und eine Basis von V ist.

Offensichtlich ist die Standardbasis $e_1, \ldots, e_n$ eine Orthonormalbasis von $\mathbb{K}^n$.

In unendlich-dimensionalen Hilberträumen werden Orthonormalbasen als maximale Orthonormalsysteme definiert.

Ist $\mathcal{B} := (v_1, \ldots, v_n)$ eine Basis von $\mathbb{K}^n$ und $v \in \mathbb{K}^n$, so ergibt die Koordinatenabbildung $C_{\mathcal{B}} : \mathbb{K}^n \to \mathbb{K}^n$ nach ihrer Definition die Koeffizienten $\lambda := (\lambda_1, \ldots, \lambda_n)^\mathsf{T} := C_{\mathcal{B}} v$ in der Basis-Darstellung

$$v = \sum_{i=1}^{n} \lambda_i v_i \,.$$

Ist $B := (v_1, \ldots, v_n)^{-1}$ so gilt nach Lemma 9.3.9 ferner $\mathsf{M}(C_{\mathcal{B}}) = B$, d.h. wir haben

$$\lambda = Bv \,.$$

Damit benötigt die Koeffizienten-Berechnung eine Matrix-Invertierung. Das folgende Lemma zeigt, dass dies für Orthonormalbasen vereinfacht werden kann.

Lemma 10.2.4

Sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$ und $v_1, \ldots, v_n$ eine Orthonormalbasis von V. Dann gilt für alle $v \in V$:

$$v = \sum_{i=1}^{n} \langle v_i, v \rangle v_i \,. \tag{10.2.1}$$

Für die Standardbasis $e_1, \ldots, e_n$ von $\mathbb{K}^n$ ist die Aussage offensichtlich und im reellen Fall wurde dies auch schon in (3.1.3) bemerkt.

Die Koeffizienten $\langle v_i, v \rangle$ heißen Fourier-Koeffizienten von v bezüglich der Basis $v_1, \ldots, v_n$.

In unendlich-dimensionalen Hilberträumen mit abzählbarer Orthonormalbasis gilt eine analoge Formel, bei der (10.2.1) durch eine Reihe ersetzt wird. Es kann dann gezeigt werden, dass diese Reihe unbedingt konvergiert. Im Falle überabzählbar großer Orthonormalbasen muss dabei zusätzlich gefordert werden, dass nur abzählbar viele Fourier-Koeffizienten ungleich null sind.

BEWEIS

Sei $v \in V$. Da $v_1, \ldots, v_n$ eine Basis von V ist, gibt es dann eindeutige $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$ mit

$$v = \sum_{i=1}^{n} \lambda_i v_i \, .$$

Für $j \in \{1, \ldots, n\}$ gilt dann

$$\langle v_j, v \rangle = \left\langle v_j, \sum_{i=1}^{n} \lambda_i v_i \right\rangle = \sum_{i=1}^{n} \lambda_i \langle v_j, v_i \rangle = \lambda_j \langle v_j, v_j \rangle = \lambda_j \, .$$

Einsetzen in die obige Basis-Darstellung ergibt die Behauptung.

Lemma 10.2.4 deutet an, dass Orthonormalbasen sehr praktisch sein können. Im folgenden wollen wir uns daher der Existenz von Orthonormalbasen widmen, wobei wir ein konstruktives Verfahren, das sogenannte Gram-Schmidtsches Orthonormalisierungsverfahren kennenlernen werden.

Hierzu sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$ und $w_1, \ldots, w_m \in V$ eine linear unabhängige Familie. Wir suchen dann eine Orthonormalbasis $v_1, \ldots, v_m$ von

$$W_m := \mathbf{span}\{w_1, \ldots, w_m\}\,.$$

Diese Basis konstruieren wir iterativ, indem wir Orthonormalbasen von

$$W_j := \mathbf{span}\{w_1, \ldots, w_j\}$$

für $j = 1, \ldots, m$ konstruieren und dabei beachten, dass die angenommene lineare Unabhängigkeit $w_j \notin W_{j-1}$ für alle $j = 2, \ldots, m$ garantiert, siehe Lemma 9.1.7.

Der Konstruktionsanfang für $j = 1$ wird durch

$$v_1 := \frac{w_1}{\|w_1\|} \tag{10.2.2}$$

realisiert, wobei wir daran erinnern, dass aus der linearen Unabhängigkeit der $w_1, \ldots, w_m$ schon $w_i \neq 0$ folgt. Offensichtlich ist dann v_1 eine Orthonormalbasis von W_1.

Wenn wir schon eine Orthonormalbasis $v_1, \ldots, v_{j-1}$ von W_{j-1} für ein $j \geq 2$ konstruiert haben, setzen wir zunächst

$$u_j := w_j - \sum_{i=1}^{j-1} \langle v_i, w_j \rangle v_i \,. \tag{10.2.3}$$

Da $w_j \notin W_{j-1} = \mathbf{span}\{v_1, \ldots, v_{j-1}\}$ gilt, folgt $u_j \notin W_{j-1}$ und damit ist dann die Familie $v_1, \ldots, v_{j-1}, u_j$ linear unabhängig nach Lemma 9.1.7. Insbesondere gilt daher $u_j \neq 0$ und wir können

$$v_j := \frac{u_j}{\|u_j\|} \tag{10.2.4}$$

setzen.

Nach Konstruktion gilt $u_j \in W_j$ und damit auch $v_j \in W_j$. Wir fixieren nun ein $k = 1, \ldots, j-1$. Für $i = 1, \ldots, j-1$ gilt dann $\langle v_k, v_i \rangle = \delta_{ki}$, und dies ergibt:

$$\begin{aligned}\langle v_k, u_j \rangle = \langle v_k, w_j \rangle - \left\langle v_k, \sum_{i=1}^{j-1} \langle v_i, w_j \rangle v_i \right\rangle &= \langle v_k, w_j \rangle - \sum_{i=1}^{j-1} \langle v_i, w_j \rangle \langle v_k, v_i \rangle \\ &= \langle v_k, w_j \rangle - \langle v_k, w_j \rangle \\ &= 0\,.\end{aligned}$$

Dies zeigt auch $v_k \perp v_j$ und damit ist $v_1, \ldots, v_j$ ein Orthonormalsystem in W_j. Nach Lemma 10.2.2 ist dieses linear unabhängig und aus $\dim W_j = j$ und Korollar 9.1.14 folgt dann, dass $v_1, \ldots, v_j$ eine Basis von W_j ist.

Das Gram-Schmidt'sches Orthonormalisierungsverfahren liefert insbesondere die Existenz von Orthonormalbasen, was wir im folgenden Satz nochmal festhalten:

Theorem 10.2.5

Sei V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$ und $0 < \dim V < \infty$. Dann gibt es eine Orthonormalbasis von V.

Gram-Schmidt an einem Beispiel

Um ein Beispiel für das Orthonormalisierungsverfahren zu sehen, betrachten wir die linear unabhängigen Vektoren

$$w_1 := \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \qquad \text{und} \qquad w_2 := \begin{pmatrix} 5 \\ 0 \\ -3 \end{pmatrix} \tag{10.2.5}$$

und die aufgespannte Ebene $\mathcal{E} := \mathbf{span}\{w_1, w_2\}$. Gemäß (10.2.2) setzen wir nun

$$v_1 := \frac{w_1}{\|w_1\|} = \frac{1}{\sqrt{13}} \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} . \tag{10.2.6}$$

Nach (10.2.3) definieren wir weiter

$$u_2 := w_2 - \langle v_1, w_2\rangle v_1 = w_2 - \frac{1}{\|w_1\|^2}\langle w_1, w_2\rangle w_1 = \begin{pmatrix} 5 \\ 0 \\ -3 \end{pmatrix} - \frac{10}{13}\begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}$$

$$= \frac{1}{13}\begin{pmatrix} 65 \\ 0 \\ -39 \end{pmatrix} - \frac{1}{13}\begin{pmatrix} 20 \\ 30 \\ 0 \end{pmatrix}$$

$$= \frac{1}{13}\begin{pmatrix} 45 \\ -30 \\ -39 \end{pmatrix}.$$

Ferner gilt

$$\|u\|_2 = \frac{\sqrt{4446}}{13} = 3\frac{\sqrt{494}}{13}$$

und damit erhalten wir gemäß (10.2.4)

$$v_2 := \frac{u_2}{\|u_2\|} = \frac{1}{\sqrt{494}}\begin{pmatrix} 15 \\ -10 \\ -13 \end{pmatrix}. \tag{10.2.7}$$

GRAM-SCHMIDT BEISPIEL

Um die Orthonormalbasis v_1, v_2 von $\mathcal{E}$ zu einer Orthonormalbasis von $\mathbb{R}^3$ zu ergänzen, benötigen wir zunächst ein $w_3 \notin \mathcal{E}$, da dann w_1, w_2, w_3 nach Lemma 9.1.7 unabhängig sind und wegen $\dim \mathbb{R}^3 = 3$ Korollar 9.1.14 dann zeigt, dass diese drei Vektoren auch eine Basis von $\mathbb{R}^3$ bilden. Wir setzen nun beispielsweise $w_3 := e_1$. Dann ist $w_3 \notin \mathcal{E}$, denn sonst gäbe es $s, t \in \mathbb{R}$ mit

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = w_3 = sw_1 + tw_2 = \begin{pmatrix} 2s + 5t \\ 3s \\ -3t \end{pmatrix}$$

was wegen der letzten beiden Komponenten zu $s = t = 0$ führt, was wiederum $e_1 = 0$ impliziert.

GRAM-SCHMIDT BEISPIEL

Nach Gram-Schmidt, siehe (10.2.3), setzen wir nun

$$
\begin{aligned}
u_3 := w_3 - \langle v_1, w_3\rangle v_1 - \langle v_2, w_3\rangle v_2 &= e_1 - \langle v_1, e_1\rangle v_1 - \langle v_2, e_1\rangle v_2 \\
&= e_1 - \frac{2}{\sqrt{13}} v_1 - \frac{15}{\sqrt{494}} v_2 \\
&= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{13} \begin{pmatrix} 4 \\ 6 \\ 0 \end{pmatrix} - \frac{15}{494} \begin{pmatrix} 15 \\ -10 \\ -13 \end{pmatrix} \\
&= \frac{1}{494} \begin{pmatrix} 494 \\ 0 \\ 0 \end{pmatrix} - \frac{38}{494} \begin{pmatrix} 4 \\ 6 \\ 0 \end{pmatrix} - \frac{15}{494} \begin{pmatrix} 15 \\ -10 \\ -13 \end{pmatrix} \\
&= \frac{1}{494} \begin{pmatrix} 117 \\ -78 \\ 195 \end{pmatrix} \\
&= \frac{1}{38} \begin{pmatrix} 9 \\ -6 \\ 15 \end{pmatrix} .
\end{aligned}
$$

Ferner gilt

$$\|u\|_3 = \frac{\sqrt{9^2 + 6^2 + 15^2}}{494} = \frac{\sqrt{342}}{494} = \frac{\sqrt{9 \cdot 38}}{494} = \frac{3\sqrt{38}}{494}$$

und damit haben wir nach (10.2.4)

$$v_3 = \frac{u_2}{\|u_2\|} = \frac{1}{3\sqrt{38}} \begin{pmatrix} 9 \\ -6 \\ 15 \end{pmatrix} = \frac{1}{\sqrt{38}} \begin{pmatrix} 3 \\ -2 \\ 5 \end{pmatrix} . \tag{10.2.8}$$

Insgesamt ist dann v_1, v_2, v_3 eine Orthonormalbasis von $\mathbb{R}^3$. Natürlich hätten wir v_3 auch durch Normierung von z.B. $v_1 \times v_2$ bestimmen können. Dieses wäre aber ein Zugang gewesen, der in höheren Dimensionen nicht mehr möglich ist, während das Verfahren nach Gram-Schmidt in jeder Dimension funktioniert.

Section 10.3
Orthogonale Abbildungen

Im folgenden seien V ein $\mathbb{K}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$ und W ein Unterraum von V mit $m := \dim W < \infty$. Wir suchen nun eine Abbildung

$$P : V \to V$$

$$v \mapsto Pv$$

die jedem Punkt v den Lotpunkt Pv von v in W zuordnet, d.h. es soll gelten $Pv \in W$ und $(v - Pv) \perp w$ für alle $w \in W$, siehe auch Abbildung 22.

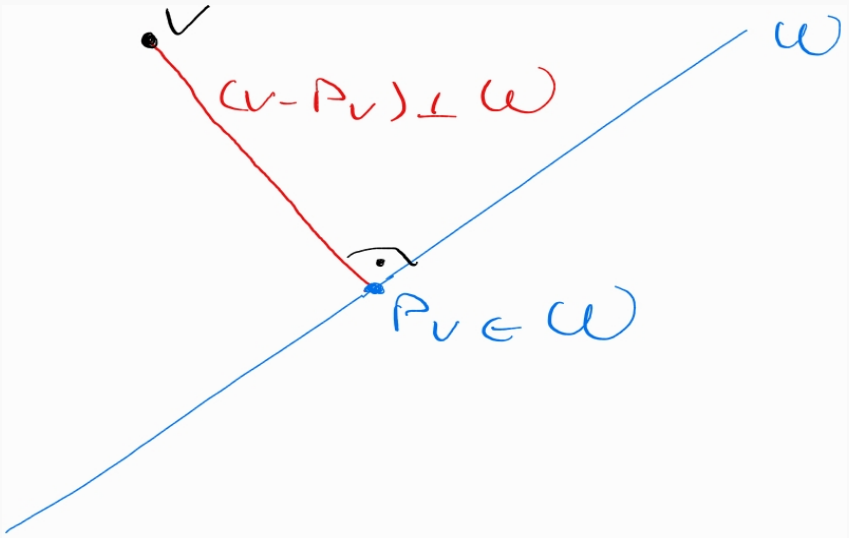


Abbildung: Der Vektor v wird auf den Lotpunkt $Pv \in W$ abgebildet. Die Differenz $v - Pv$ ist dabei senkrecht zu dem Unterraum W.

Um eine solche Abbildung zu konstruieren, wählen wir nach Satz 10.2.5 eine Orthonormalbasis $v_1, \dots, v_m$ von W. Da wir $Pv \in W$ haben möchten, gibt es dann zu jedem $v \in V$ Koeffizienten $\lambda_1, \dots, \lambda_m \in \mathbb{K}$ mit

$$Pv = \sum_{i=1}^{m} \lambda_i v_i \, .$$

Ferner wollen wir $(v - Pv) \perp v_j$ für alle $j = 1, \dots, m$ haben und dies führt zu

$$\begin{aligned} 0 = \langle v_j, v - Pv \rangle &= \langle v_j, v \rangle - \left\langle v_j, \sum_{i=1}^{m} \lambda_i v_i \right\rangle \\ &= \langle v_j, v \rangle - \sum_{i=1}^{m} \lambda_i \langle v_j, v_i \rangle \\ &= \langle v_j, v \rangle - \lambda_j \, . \end{aligned}$$

Mit anderen Worten haben wir $\lambda_j = \langle v_j, v \rangle$ und dies ergibt

$$Pv = \sum_{i=1}^{m} \langle v_i, v \rangle v_i \, , \qquad v \in V. \tag{10.3.1}$$

Der folgende Satz zeigt insbesondere, dass die Wahl der Orthonormalbasis $v_1, \ldots, v_m$ von W das Verhalten von (10.3.1) nicht beeinflusst.

Theorem 10.3.1

Seien V ein Prä-Hilbertraum und W ein Unterraum von V mit $\dim W < \infty$. *Dann gibt es genau eine Abbildung $P : V \to V$ mit*

$$Pv \in W \qquad \text{und} \qquad (v - Pv) \perp W \tag{10.3.2}$$

für alle $v \in V$. Ferner ist diese Abbildung P, die **Orthogonalprojektion** *auf W genannt wird, durch* (10.3.1) *beschrieben und linear.*

BEWEIS 1/2

Wir setzen $m := \dim W$. Im Fall $m = 0$ ist $Pv = 0$ die einzige Wahl und das Überprüfen der weiteren Eigenschaften ist in diesem Fall trivial, so dass wir im Folgenden $m \geq 1$ annehmen.

Für die Existenz von P wählen wir eine Orthonormalbasis $v_1, \ldots, v_m$ von W und betrachten die durch (10.3.1) gegebene Abbildung. Nach Konstruktion haben wir dann $Pv \in W$ für alle $v \in V$ und $\langle v - Pv, v_j \rangle = 0$ für alle $j = 1, \ldots, m$. Sei nun $w \in W$. Dann gibt es $\alpha_1, \ldots, \alpha_m \in \mathbb{K}$ mit $w = \alpha_1 v_1 + \cdots + \alpha_m v_m$. Dies ergibt

$$\langle v - Pv, w \rangle = \left\langle v - Pv, \sum_{j=1}^{m} \alpha_j v_j \right\rangle = \sum_{j=1}^{m} \alpha_j \langle v - Pv, v_j \rangle = 0 \,.$$

BEWEIS 2/2

Um die Eindeutigkeit zu zeigen, nehmen wir an, dass wir eine weitere Abbildung $Q : V \to V$ mit den Eigenschaften (10.3.2) haben. Für $v \in V$ und $w := Pv - Qv \in W$ gilt dann

$$\begin{aligned} \|Pv - Qv\|^2 &= \langle Pv - Qv, w \rangle \\ &= \langle Pv - v, w \rangle + \langle v - Qv, w \rangle \\ &= 0\,, \end{aligned}$$

d.h. $Pv = Qv$. Dies zeigt die Eindeutigkeit.

Die Linearität von P folgt aus (10.3.1) und der Linearität des Skalarprodukts in der zweiten Komponente.

Zu einem gegebenen $v \in V$ sollte der Lotpunkt Pv intuitiv der Punkt $w \in W$ sein, der den kleinsten Abstand zu v hat. Der folgende Satz bestätigt diese Intuition.

Theorem 10.3.2

Seien V ein Prä-Hilbertraum, W ein Unterraum von V mit $\dim W < \infty$ und P die Orthogonalprojektion auf W. Dann gilt für alle $v \in V$ und alle $w \in W$ mit $w \neq Pv$:

$$\|v - Pv\| < \|v - w\| .$$

BEWEIS

Wir fixieren ein $v \in V$ und ein $w \in W$ mit $w \neq Pv$. Dann gilt $v - Pv \in W^{\perp}$ und $Pv - w \in W$. Dies ergibt $(v - Pv) \perp (Pv - w)$ und der Satz von Pythagoras liefert

$$\|v - w\|^2 = \|v - Pv + Pv - w\|^2 = \|v - Pv\|^2 + \|Pv - w\|^2 .$$

Wegen $w \neq Pv$ haben wir nun $\|Pv - w\|^2 > 0$ und Einsetzen dieser Abschätzung in die obige Gleichung ergibt die Behauptung.

BEISPIEL

Sei $\mathcal{E} \subset \mathbb{R}^3$ die Ebene, die durch die in (10.2.5) definierten Vektoren w_1, w_2 aufgespannt wird. Wir haben in (10.2.6) und (10.2.7) schon gesehen, dass

$$v_1 := \frac{1}{\sqrt{13}} \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \qquad \text{und} \qquad v_2 := \frac{1}{\sqrt{494}} \begin{pmatrix} 15 \\ -10 \\ -13 \end{pmatrix}$$

eine Orthonormalbasis von $\mathcal{E}$ ist. Wollen wir nun die Orthogonalprojektion $P : \mathbb{R}^3 \to \mathbb{R}^3$ auf $\mathcal{E}$ berechnen, ergibt sich nach (10.3.1)

$$\begin{aligned} Px &= \langle v_1, x \rangle v_1 + \langle v_2, x \rangle v_2 \\ &= \frac{1}{13} \left\langle x, \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right\rangle \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} + \frac{1}{494} \left\langle x, \begin{pmatrix} 15 \\ -10 \\ -13 \end{pmatrix} \right\rangle \begin{pmatrix} 15 \\ -10 \\ -13 \end{pmatrix} \end{aligned} \tag{10.3.3}$$

für alle $x \in \mathbb{R}^3$. Nach Satz 10.3.2 ist $\|x - Px\|$ der (ungerichtete) Abstand von x zu $\mathcal{E}$. Dieser kann direkt durch Einsetzen berechnet werden. Etwas später werden wir jedoch eine elegantere Methode kennenlernen.

Da die Orthogonalprojektion $P : \mathbb{R}^3 \to \mathbb{R}^3$ linear ist, gibt es auch eine darstellende Matrix $\mathrm{M}(P) \in \mathcal{M}(3,3)$ bezüglich der Standardbasis von $\mathbb{R}^3$. Der folgende Satz ergibt eine allgemeine Berechnungsformel für darstellende Matrizen von Orthogonalprojektionen $\mathbb{K}^n \to \mathbb{K}^n$.

Theorem 10.3.3

Sei $W \subset \mathbb{K}^n$ ein Unterraum und $P : \mathbb{K}^n \to \mathbb{K}^n$ die Orthogonalprojektion auf W. Dann gilt für jede Orthonormalbasis $v_1, \ldots, v_m$ von W:

$$\mathrm{M}(P) = \sum_{k=1}^{m} v_k \overline{v}_k^{\mathsf{T}} = S\overline{S}^{\mathsf{T}},$$

wobei $S := (v_1, \ldots, v_m) \in \mathcal{M}(n,m)$.

Man beachte, dass $v_k \overline{v}_k^{\mathsf{T}}$ und $S\overline{S}^{\mathsf{T}}$ jeweils $n \times n$-Matrizen sind.

Sei $e_1, \dots, e_n$ die Standardbasis von $\mathbb{K}^n$ und $v_{1k}, \dots, v_{nk}$ die Komponenten von v_k, d.h. $v_k = (v_{1k}, \dots, v_{nk})^\mathsf{T}$. Dann gilt

$$Pe_j = \sum_{k=1}^{m} \langle v_k, e_j \rangle v_k = \sum_{k=1}^{m} \overline{v}_{jk} v_k = \sum_{k=1}^{m} v_k \overline{v}_{jk} \,,$$

d.h. wir haben

$$Pe_j = \begin{pmatrix} \sum_{k=1}^{m} v_{k1} \overline{v}_{jk} \\ \vdots \\ \sum_{k=1}^{m} v_{kn} \overline{v}_{jk} \end{pmatrix} .$$

Nach Korollar 9.2.12 ist dies die j-te Spalte der darstellenden Matrix. Da wir ferner

$$v_k \overline{v}_k^\mathsf{T} = \begin{pmatrix} v_{k1} \overline{v}_{1k} & \dots & v_{k1} \overline{v}_{nk} \\ \vdots & & \vdots \\ v_{kn} \overline{v}_{1k} & \dots & v_{kn} \overline{v}_{nk} \end{pmatrix}$$

haben, folgt dann die erste Gleichung. Die zweite Gleichung ist nur eine andere Schreibweise für die erste Gleichung.

Um die darstellende Matrix der Projektion (10.3.3) zu berechnen, müssen wir ihre Einträge

$$v_{1i}v_{j1} + v_{2i}v_{j2}$$

für $i, j = 1, 2, 3$ bestimmen. Für $i = j = 1$ ist dies z.B.

$$v_{11}v_{11} + v_{21}v_{12} = \frac{2 \cdot 2}{13} + \frac{15 \cdot 15}{494} = \frac{4 \cdot 38 + 225}{494} = \frac{377}{494} = \frac{29}{38}.$$

Die anderen Einträge werden analog berechnet und das Endergebnis ist

$$\mathrm{M}(P) = \frac{1}{38}\begin{pmatrix} 29 & 6 & -15 \\ 6 & 34 & 10 \\ -15 & 10 & 13 \end{pmatrix}. \tag{10.3.4}$$

Man beachte, dass diese Matrix symmetrisch im Sinne von $A^T = A$ ist und damit gilt auch $P^* = P$ nach der Definition von adjungierten Operatoren. Alternativ hätten wir auch

$$\mathrm{M}(P) = \begin{pmatrix} \frac{2}{\sqrt{13}} & \frac{15}{\sqrt{494}} \\ \frac{3}{\sqrt{13}} & \frac{-1}{\sqrt{494}} \\ 0 & \frac{-13}{\sqrt{494}} \end{pmatrix} \cdot \begin{pmatrix} \frac{2}{\sqrt{13}} & \frac{3}{\sqrt{13}} & 0 \\ \frac{15}{\sqrt{494}} & \frac{-1}{\sqrt{494}} & \frac{-13}{\sqrt{494}} \end{pmatrix}$$

benutzen können, was effektiv auf die exakt gleichen Rechnungen hinausläuft.

Der folgende Satz zeigt in Verbindung mit Satz 9.3.7, dass die Symmetrie in (10.3.4) kein Zufall ist. Außerdem stellt er weitere Eigenschaften von Orthogonalprojektionen zur Verfügung.

Theorem 10.3.4

Seien V ein Prä-Hilbertraum, W ein Unterraum von V mit $\dim W < \infty$ und P die Orthogonalprojektion auf W. Dann gilt:

$$
\begin{aligned}
Pw &= w\,, && w \in W,\\
\ker P &= W^{\perp}\,,\\
P^2 &= P\,,\\
\langle Pv, w\rangle &= \langle v, Pw\rangle\,, && v, w \in V.
\end{aligned}
$$

BEWEIS 1/2

Sei $w \in W$ und $v := w$. Dann gilt $w \in W$ und $(v - w) \perp W$. Damit erfüllt w die Eigenschaften (10.3.2) zu v und die Eindeutigkeit von P ergibt $Pw = Pv = w$.

Um $\ker P = W^\perp$ zu zeigen, wählen wir zunächst ein $v \in \ker P$, d.h. $Pv = 0$. Für $w \in W$ gilt dann $\langle v, w\rangle = \langle v - Pv, w\rangle = 0$, und damit $v \in W^\perp$. Ist umgekehrt $v \in W^\perp$, so gilt $0 \in W$ und $(v - 0) \perp w$ für alle $w \in W$. Damit muss wegen der Eindeutigkeit von P schon $Pv = 0$ gelten.

Um $P^2 = P$ zu überprüfen, wählen wir ein $v \in V$ und setzen $w := Pv$. Dann haben wir $w \in W$ und es folgt $P(Pv) = Pw = w = Pv$.

BEWEIS 2/2

Seien schließlich $v, w \in V$. Dann gilt mit der Darstellung (10.3.1)

$$\begin{aligned}\langle Pv, w\rangle = \left\langle \sum_{i=1}^{m} \langle v_i, v\rangle v_i, w \right\rangle &= \sum_{i=1}^{m} \overline{\langle v_i, v\rangle}\langle v_i, w\rangle \\ &= \overline{\sum_{i=1}^{m} \overline{\langle v_i, w\rangle}\langle v_i, v\rangle} \\ &= \overline{\left\langle \sum_{i=1}^{m} \langle v_i, w\rangle v_i, v \right\rangle} \\ &= \overline{\langle Pw, v\rangle}\,.\end{aligned}$$

Wegen $\overline{\langle Pw, v\rangle} = \langle v, Pw\rangle$ folgt dann die Behauptung.

Seien V ein endlich-dimensionaler Hilbertraum, W ein Unterraum von V und $P : V \to V$ die orthogonale Projektion auf W. Wir setzen

$$Q := \mathrm{id}_V - P\,. \tag{10.3.5}$$

Dann gilt $\langle Qv, w\rangle = \langle v - Pv, w\rangle = 0$ für alle $v \in V$ und $w \in W$, d.h. wir haben

$$Qv \in W^\perp\,, \qquad\qquad v \in V. \tag{10.3.6}$$

Für $w' \in W^\perp$ gilt ferner

$$\langle w', v - Qv\rangle = \langle w', v - v + Pv\rangle = \langle w', Pv\rangle = 0\,,$$

wobei wir im letzten Schritt sowohl $Pv \in W$ als auch die Definition von $W^\perp$ ausgenutzt haben. Dies zeigt

$$(v - Qv) \perp W^\perp\,, \qquad\qquad v \in V$$

und damit ist Q die orthogonale Projektion auf $W^\perp$. Dies führt zu dem folgenden Satz.

Theorem 10.3.5

Sei V ein endlich-dimensionaler Hilbertraum und W ein Unterraum von V. Dann gibt es zu jedem $v \in V$ eindeutige $w \in W$ und $w^\perp \in W^\perp$ mit

$$v = w + w^\perp .$$

Diese sind durch $w = Pv$ und $w^\perp = v - Pv$ eindeutig bestimmt.

BEWEIS

Wir setzten $w := Pv \in W$ und $w^\perp := v - Pv = Qv$, wobei $Q := \mathbf{id}_V - P$ ist. Dann haben wir $w^\perp \in W^\perp$ nach (10.3.6) und die Zerlegung ist offensichtlich.

Nehmen wir nun zu gegebenen $v \in V$ ein $w \in W$ und ein $w^\perp \in W^\perp$ mit $v = w + w^\perp$, so gilt

$$v - w = w^\perp \in W^\perp .$$

Da wir auch $w \in W$ haben, muss $w = Pv$ nach Satz 10.3.1 gelten. Dies wiederum zeigt $w^\perp = v - w = v - Pv = Qv$.

Interpretation Gram-Schmidt-Verfahren

Seien V ein endlich-dimensionaler Hilbertraum, $w_1, \dots, w_m \in V$ linear unabhängig und $v_1, \dots, v_m$ die Vektoren, die durch das Gram-Schmidt'sche Orthonormalisierungsverfahren konstruiert worden sind. Dann haben wir schon gesehen, dass $v_1, \dots, v_j$ eine Orthonormalbasis von $W_j := \mathbf{span}\{w_1, \dots, w_j\}$ ist. Der entscheidende Schritt in diesem Verfahren war dabei (10.2.3), d.h.

$$u_j := w_j - \sum_{i=1}^{j-1} \langle v_i, w_j \rangle v_i = w_j - P_{W_{j-1}} w_j = P_{W_{j-1}^\perp} w_j \,,$$

wobei $P_{W_{j-1}}$ die Orthogonalprojektion auf W_{j-1} ist und $P_{W_{j-1}^\perp}$ die auf $W_{j-1}^\perp$.

Erweitern wir nun die Vektoren $v_1, \dots, v_m$ zu einer Orthonormalbasis $v_1, \dots, v_n$ von V, und setzen $W := W_m$, so sind $v_{m+1}, \dots, v_n$ eine Orthonormalbasis von $W^\perp$: Nach Lemma 10.2.2 sind diese Vektoren zunächst linear unabhängig. Gäbe es noch ein $w \in W^\perp$ mit

$$w \notin \mathbf{span}\{v_{m+1}, \dots, v_n\},$$

so würden wir mit Gram-Schmidt ein $v_{n+1} \in W^\perp$ finden, so dass $v_{m+1}, \dots, v_{n+1}$ ein Orthonormalsystem in $W^\perp$ ist. Damit wäre auch $v_1, \dots, v_{n+1}$ ein Orthonormalsystem in V, und damit wiederum linear unabhängig. Wegen $\mathbf{dim}\, V = n$ erhalten wir dann mit Satz 9.1.8 einen Widerspruch.

Damit lässt sich die Projektion $Q := P_{W^\perp}$ auf $W^\perp$ durch

$$Qv = \sum_{j=m+1}^{n} \langle v_j, v \rangle v_j\,, \qquad v \in V$$

ausdrücken. Ist zudem $V = \mathbb{K}^n$, so lässt sich die Projektion nach Satz 10.3.3 durch die Matrix

$$\mathrm{M}(Q) = \sum_{k=m+1}^{n} v_k \overline{v}_k^{\mathsf{T}} \tag{10.3.7}$$

darstellen.

ABSTANDSBERECHNUNG: BEISPIEL

Betrachten wir nun beispielhaft wieder die von den Vektoren

$$v_1 := \frac{1}{\sqrt{13}} \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \qquad \text{und} \qquad v_1 := \frac{1}{\sqrt{494}} \begin{pmatrix} 15 \\ -10 \\ -13 \end{pmatrix}$$

aufgespannte Ebene $\mathcal{E} \subset \mathbb{R}^3$, so hatten wir schon die Orthogonalprojektion $P : \mathbb{R}^3 \to \mathbb{R}^3$ auf $\mathcal{E}$ in (10.3.3) bestimmt und in (10.3.4) haben wir auch ihre darstellende Matrix berechnet. Ferner haben wir in Satz 10.3.2 gesehen, das $\|x - Px\|$ der (ungerichtete) Abstand von x zu $\mathcal{E}$ ist. Offensichtlich ist $x - Px = Qx$, d.h. der Abstand lässt sich mit Hilfe der Orthogonalprojektion auf $W^\perp$ berechnen. Um diese Projektion zu bestimmen, müssen wir nach unseren obigen Überlegungen das Orthonormalsystem v_1, v_2 zu einer Orthonormalbasis v_1, v_2, v_3 mit Hilfe von Gram-Schmidt ergänzen. In unserem Beispiel haben wir dies bereits in (10.2.8) getan, d.h. wir haben

$$v_3 = \frac{1}{\sqrt{38}} \begin{pmatrix} 3 \\ -2 \\ 5 \end{pmatrix} .$$

Nach (10.3.1) haben wir damit

$$Qx = \langle v_3, x\rangle v_3 = \frac{1}{38}\left\langle x, \begin{pmatrix} 3 \\ -2 \\ 5 \end{pmatrix}\right\rangle \begin{pmatrix} 3 \\ -2 \\ 5 \end{pmatrix} \tag{10.3.8}$$

für alle $x \in \mathbb{R}^3$. Ferner wird diese Projektion nach Satz 10.3.3 durch die Matrix

$$\mathrm{M}(Q) = v_3 v_3^{\mathsf{T}} = \frac{1}{38}\begin{pmatrix} 9 & -6 & 15 \\ -6 & 4 & -10 \\ 15 & -10 & 25 \end{pmatrix}$$

dargestellt. Mit Hilfe der darstellenden Matrix von P, siehe (10.3.4), können wir $\mathrm{M}(Q)$ alternativ auch durch

$$\begin{aligned} \mathrm{M}(Q) = \mathrm{M}(\mathrm{id}_{\mathbb{R}^3} - P) &= \mathrm{M}(\mathrm{id}_{\mathbb{R}^3}) - \mathrm{M}(P) \\ &= \frac{1}{38}\begin{pmatrix} 38 & 0 & 0 \\ 0 & 38 & 0 \\ 0 & 0 & 38 \end{pmatrix} - \frac{1}{38}\begin{pmatrix} 29 & 6 & -15 \\ 6 & 34 & 10 \\ -15 & 10 & 13 \end{pmatrix} \\ &= \frac{1}{38}\begin{pmatrix} 9 & -6 & 15 \\ -6 & 4 & -10 \\ 15 & -10 & 25 \end{pmatrix} \end{aligned}$$

ABSTANDSBERECHNUNG: BEISPIEL

Ignoriert man den Vorfaktor $\frac{1}{38}$, so benötigt die Berechnung von $Qx = M(Q)x$ genau 9 Multiplikationen und 6 Additionen, während die Berechnung durch (10.3.8) lediglich 6 Multiplikationen und 2 Additionen benötigt. In diesem Fall "lohnt" sich die Matrixdarstellung also nicht. Vergleicht man dies mit der Berechnung von P, so sind die Kosten der Matrixmultiplikation natürlich gleich denen für Q, während die Berechnung von Px durch (10.3.3) jetzt 12 Multiplikationen und 5 Additionen benötigt.

Diese Betrachtungen ignorieren allerdings die Kosten der Matrixerstellung, so dass sie vor allem in Situationen interessant sind, in denen die Projektionen sehr häufig berechnet werden müssen.

Lineare Abbildungen, die Längen oder Winkel erhalten, haben eine besonders schöne Struktur. Diese werden jetzt untersucht.

Definition 10.3.6

Sei $A \in \mathcal{M}_{\mathbb{K}}(n, n)$ eine Matrix, die

$$\langle Ax, Ay \rangle = \langle x, y \rangle \,, \qquad x, y \in \mathbb{K}^n$$

erfüllt. Dann heißt A im Fall $\mathbb{K} = \mathbb{R}$ orthogonal und im Fall $\mathbb{K} = \mathbb{C}$ unitär.

PRODUKT ORTHOGONALER MATRIZEN IST ORTHOGONAL

Das folgende einfache Lemma zeigt, dass Produkte orthogonaler Matrizen orthogonal sind.

Lemma 10.3.7

Sind $A, B \in \mathcal{M}_{\mathbb{K}}(n, n)$ orthogonal/unitär, so ist auch AB orthogonal/unitär.

Beweis.

Für $x, y \in \mathbb{K}^n$ gilt $\langle ABx, ABy\rangle = \langle Bx, By\rangle = \langle x, y\rangle$. □

Der folgende Satz charakterisiert orthogonale bzw. unitäre Matrizen.

Theorem 10.3.8

Sei $A \in \mathcal{M}_{\mathbb{K}}(n,n)$. Dann sind die folgenden Aussagen äquivalent:

i). Es gilt $\langle Ax, Ay\rangle = \langle x, y\rangle$ für alle $x, y \in \mathbb{K}^n$.

ii). Es gilt $\|Ax\| = \|x\|$ für alle $x \in \mathbb{K}^n$.

iii). Es gilt $\overline{A}^{\mathsf{T}} A = A\overline{A}^{\mathsf{T}} = E_n$.

iv). Die Zeilen von A sind eine Orthonormalbasis von $\mathbb{K}^n$.

v). Die Spalten von A sind eine Orthonormalbasis von $\mathbb{K}^n$.

Nach *iii)* sind orthogonale, bzw. unitäre Matrizen invertierbar und es gilt

$$A^{-1} = \overline{A}^{\mathsf{T}}. \tag{10.3.9}$$

Für diese Matrizen ist also die Matrixinvertierung besonders einfach.

BEWEIS 1/2

ii) $\Rightarrow$ *i)*. Nachrechnen mit Polarisation.

i) $\Rightarrow$ *ii)*. Folgt sofort mit $x = y$.

i) $\Leftrightarrow$ *iii)*. Die Gleichung $\langle Ax, Ay\rangle = \langle x, y\rangle$ ist äquivalent zu

$$\langle \overline{A}^{\mathsf{T}} Ax, y\rangle = \langle x, y\rangle \,, \tag{10.3.10}$$

und da die komplexe Konjugation und die Transposition selbst-invers sind, ist sie auch äquivalent zu

$$\langle x, \overline{A}^{\mathsf{T}} Ay\rangle = \langle x, y\rangle \tag{10.3.11}$$

Gilt nun *i)*, so erhalten wir durch Betrachten von $x = e_i$ und $y = e_j$ die Gleichung

$$\langle \overline{A}^{\mathsf{T}} Ae_i, e_j\rangle = \langle e_i, \overline{A}^{\mathsf{T}} Ae_j\rangle = \langle e_i, e_j\rangle = \delta_{ij} \,. \tag{10.3.12}$$

Dies zeigt *iii)*. Gilt umgekehrt *iii)*, so erhalten wir (10.3.12). Für $x = \alpha_1 e_1 + \cdots + \alpha_n e_n$ und $y = \beta_1 e_1 + \cdots + \beta_n e_n$ folgen dann (10.3.11) und (10.3.10) mit der (Semi-)-Linearität.

BEWEIS 2/2

iii) $\Leftrightarrow$ *iv)*. Ist a_i ein Zeilenvektor von A, so ist $\overline{a}_i^{\mathsf{T}}$ ein Spaltenvektor von $\overline{A}^{\mathsf{T}}$. Ferner gilt

$$\overline{a}_i^{\mathsf{T}} a_j = \langle a_i, a_j \rangle \,,$$

wobei die linke Seite der Formel den Eintrag b_{ij} von $A\overline{A}^{\mathsf{T}}$ darstellt.

iii) $\Leftrightarrow$ *v)*. Analog.

BEISPIEL: DREHUNGEN

Wir hatten schon in (9.2.4) gesehen, dass eine Drehung gegen den Uhrzeigersinn um den Winkel φ durch die Matrix

$$D_\varphi := \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

beschrieben wird. Es ist leicht nachzurechnen, dass die beiden Spaltenvektoren eine Orthonormalbasis von $\mathbb{R}^2$ sind, wobei aus Kardinalitätsgründen nur überprüft werden muss, dass sie ein Orthonormalsystem bilden. Damit ist D orthogonal. Ferner gilt

$$\det D_\varphi = 1\,.$$

Schließlich ist anschaulich klar, dass der Hintereinanderausführung zweier Drehungen im $\mathbb{R}^2$ wieder eine Drehung ergibt.

BEISPIEL: SPIEGELUNGEN

Analog kann man sich überlegen, dass

$$A_\varphi := \begin{pmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{pmatrix}$$

eine Spiegelung an der Geraden

$$g = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \sin(\varphi/2)x_1 - \cos(\varphi/2)x_2 = 0 \right\}$$

beschreibt. Auch diese Matrix ist wieder orthogonal. Insbesondere haben wir

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \text{und} \qquad A_{\pi/2} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Ferner gilt

$$\det A_\varphi = -1\,.$$

Das folgende Lemma zeigt, dass die Determinanten von orthogonalen Matrizen nur die Werte -1 und 1 annehmen können.

Lemma 10.3.9

Ist A eine orthogonale/unitäre Matrix, so gilt

$$|\det A| = 1\,.$$

Beweis.

Es gilt

$$|\det A|^2 = \overline{\det A} \cdot \det A = \det \overline{A}^{\mathsf{T}} \cdot \det A = \det \overline{A}^{\mathsf{T}} A = \det E_n = 1.$$

Dies ergibt die Behauptung. □

Das vorige Lemma motiviert die folgenden Definitionen.

Definition 10.3.10

Eine orthogonale Matrix $A \in \mathcal{M}_{\mathbb{R}}(n,n)$ heißt Drehung falls $\det A = 1$. Ferner heißt A Spiegelung, falls $\det A = -1$.

Wir haben oben gesehen, dass die Dreh-Matrizen D_φ Drehungen sind und dass die Matrizen A_φ Spiegelungen sind.

Wir nennen die Menge

$$\mathcal{O}(n) := \left\{A \in \mathcal{M}_{\mathbb{R}}(n,n) : A^{\mathsf{T}}A = AA^{\mathsf{T}} = E_n\right\}$$

der orthogonalen Matrizen die orthogonale Gruppe. Ferner heißt die Menge

$$\mathcal{SO}(n) := \left\{A \in \mathcal{O}(n) : \det A = 1\right\}$$

der Drehungen die spezielle orthogonale Gruppe. Beides sind Untergruppen der allgemeinen linearen Gruppe $\mathcal{GL}(n)$. Man beachte, dass die Spiegelungen *keine* Gruppe bilden können, da z.B. die Verknüpfung zweier Spiegelungen eine Drehung ist.

Section 10.4
Eigenwerte

Definition 10.4.1

Seien V ein $\mathbb{K}$-Vektorraum und $S : V \to V$ linear. Dann heißt $\lambda \in \mathbb{K}$ **Eigenwert** von S, falls es ein $v \in V$ mit $v \neq 0$ und

$$Sv = \lambda v$$

gibt. In diesem Fall heißt v **Eigenvektor** zum Eigenwert λ. Die Menge

$$\mathrm{Eig}_S(\lambda) := \{v \in V : Sv = \lambda v\}$$

heißt **Eigenraum** zum Eigenwert λ

Der Eigenraum $\mathrm{Eig}_S(\lambda)$ enthält also alle Eigenvektoren zum Eigenwert λ *und* 0. Man kann leicht zeigen, dass $\mathrm{Eig}_S(\lambda)$ ein Unterraum von V ist.

Im folgenden werden wir wie üblich eine Matrix $A \in \mathcal{M}(n, n)$ mit der linearen Abbildung $\mathsf{L}_A : \mathbb{K}^n \to \mathbb{K}^n$ identifizieren. In diesem Sinne sprechen wir von Eigenwerten λ der Matrix A, wenn diese Eigenwerte von L_A sind. Analoges gilt für Eigenvektoren und Eigenräume.

BEISPIEL

Wir betrachten die Matrix

$$A := \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}.$$

Dann gilt dann

$$Ae_1 = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} = 2e_1,$$

d.h. $\lambda_1 := 2$ ist ein Eigenwert von A und e_1 ein zugehöriger Eigenvektor. Es gilt ferner

$$\mathrm{Eig}_A(\lambda_1) := \{te_1 : t \in \mathbb{R}\}.$$

Analog kann man “ablesen”, dass $\lambda_2 := -1$ ein Eigenwert mit Eigenvektor e_2 von A ist.

Das obige Beispiel lässt sich leicht auf größere Diagonalmatrizen übertragen. Das folgende Lemma hält das Ergebnis fest.

Lemma 10.4.2

Sei $D \in \mathcal{M}(n, n)$ eine Diagonalmatrix mit Diagonaleinträgen $d_1, \ldots, d_n$. Dann ist jedes d_i ein Eigenwert von D mit Eigenvektor e_i.

Lemma 10.4.2 ergibt eine einfache, hinreichende Bedingung für die Eigenwerte einer Diagonalmatrix D. Doch sind dies alle Eigenwerte von D?

Um diese Frage zu beantworten beginnen wir mit dem folgenden Lemma.

Lemma 10.4.3

Sei $A \in \mathcal{M}(n,n)$ und $\lambda \in \mathbb{K}$. Dann sind die folgenden Aussagen äquivalent:

i). λ *ist Eigenwert von* A.

ii). *Die Matrix* $A - \lambda E_n$ *ist nicht invertierbar.*

iii). $\det(A - \lambda E_n) = 0$.

In diesem Fall gilt $\mathrm{Eig}_A(\lambda) = \ker(A - \lambda E_n)$.

BEWEIS

i) $\Leftrightarrow$ *ii)*. Nach Definition ist λ Eigenwert, falls es ein $v \in \mathbb{K}^n \setminus \{0\}$ gibt mit $Av = \lambda v$. Dies ist äquivalent zu $\ker(A - \lambda E_n) \neq \{0\}$. Nach Korollar 9.5.10 ist dies wiederum äquivalent zu der Nicht-Invertierbarkeit von $A - \lambda E_n$.

ii) $\Leftrightarrow$ *iii)*. Dies folgt direkt aus Satz 9.6.4.

Die Formel für $\mathrm{Eig}_A(\lambda)$ ist offensichtlich.

Korollar 10.4.4

Sei $A \in \mathcal{M}(n,n)$ und λ ein Eigenwert von A. Dann ist λ auch ein Eigenwert von A^T. Im Fall $A \in \mathcal{GL}(n)$ gilt zudem $\lambda \neq 0$ und λ^{-1} ist dann ein Eigenwert von A^{-1}.

BEWEIS

Wegen $\det B = \det B^{\mathsf{T}}$ und $E_n^{\mathsf{T}} = E_n$ gilt

$$\det(A^{\mathsf{T}} - \lambda E_n) = \det(A - \lambda E_n)^{\mathsf{T}} = \det(A - \lambda E_n) = 0\,.$$

Damit ist λ Eigenwert von A^{T} nach Lemma 10.4.3.

Wäre $\lambda = 0$, so wäre $\ker A = \ker(A - 0E_n) \neq \{0\}$, was wegen $A \in \mathcal{GL}(n)$ unmöglich ist. Sei nun v ein Eigenvektor zu λ, d.h. $Av = \lambda v$. Dies impliziert

$$v = A^{-1}(\lambda v) = \lambda A^{-1} v$$

und dies impliziert $\lambda^{-1} v = A^{-1} v$.

Wir betrachten jetzt die Matrix

$$A := \begin{pmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Für $\lambda \in \mathbb{R}$ gilt dann

$$\begin{aligned} \det(A - \lambda E_3) &= \left| \begin{pmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \right| = \left| \begin{pmatrix} 3-\lambda & 0 & 0 \\ 0 & 4-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{pmatrix} \right| \\ &= (3-\lambda)(4-\lambda)(1-\lambda). \end{aligned}$$

Damit ist die Funktion $\lambda \mapsto \det(A - \lambda E_3)$ ein Polynom dritten Grades. Insbesondere kann es keine weiteren Eigenwerte als die schon bekannten 3, 4 und 1 geben, da Polynome dritten Grades maximal drei Nullstellen haben.

Ferner gilt $A^{\mathsf{T}} = A$, und damit hat A^{T} natürlich auch die Eigenwerte 3, 4 und 1. Die Matrix A ist ausserdem invertierbar mit

$$A^{-1} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

und eine analoge Überlegung zusammen mit dem obigen Korollar zeigt dass 1/3, 1/4 und 1 die einzigen Eigenwerte von A^{-1} sind.

Wir betrachten jetzt die Dreh-Matrix

$$D_\varphi := \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

Dann gilt

$$\det(D_\varphi - \lambda E_2) = \left| \begin{pmatrix} \cos(\varphi) - \lambda & -\sin\varphi \\ \sin\varphi & \cos(\varphi) - \lambda \end{pmatrix} \right| = \big(\cos(\varphi) - \lambda\big)^2 + \sin^2\varphi\,.$$

Überprüfen zeigt, dass das reelle Polynom $\lambda \mapsto \det(D_\varphi - \lambda E_2)$ vom Grad 2 die beiden komplexen Nullstellen

$$\lambda_{1,2} := \cos\varphi \pm \mathrm{i}\sin\varphi = \mathrm{e}^{\pm \mathrm{i}\varphi}$$

hat. Insbesondere hat die Matrix D_φ in der Regel keine reellen Eigenwerte, sondern nur zwei komplexe Eigenwerte.

Der folgende Satz verallgemeinert die obigen Beobachtungen.

Theorem 10.4.5

Sei $A \in \mathcal{M}_{\mathbb{K}}(n,n)$. Dann ist die Funktion

$$\begin{aligned} p_A : \mathbb{K} &\to \mathbb{K} \\ \lambda &\mapsto \det(A - \lambda E_n) \end{aligned}$$

ein Polynom vom Grad n. Es heißt das ***charakteristische Polynom*** *von A und es gilt für seine Darstellung $p_A(\lambda) = \sum_{i=0}^{n} b_k \lambda^k$:*

$$\begin{aligned} b_n &= (-1)^n \\ b_{n-1} &= (-1)^{n-1} \operatorname{spur} A \\ b_0 &= \det A\,, \end{aligned}$$

wobei $\operatorname{spur} A := \sum_{j=1}^{n} a_{jj}$ die ***Spur*** *von A heißt.*

Einfaches Nachrechnen zeigt, dass für $A, B \in \mathcal{M}(n,n)$ gilt

$$\operatorname{spur}(AB) = \operatorname{spur}(BA)\,. \tag{10.4.1}$$

BEWEIS

Wir erinnern an die allgemeine Berechnungsformel (9.6.12) für Determinanten, d.h.

$$\det B = \sum_{\sigma \in S_n} \operatorname{sign} \sigma \cdot b_{1,\sigma(1)} \cdot \dots \cdot b_{n,\sigma(n)} .$$

Für $B := A - \lambda E_n$, $\sigma \in S_n$ und $i \in \{1, \dots, n\}$ haben wir dann

$$b_{i,\sigma(i)} = \begin{cases} a_{ii} - \lambda & \text{falls } \sigma(i) = i \\ a_{i,\sigma(i)} & \text{sonst.} \end{cases}$$

Damit ist $q_\sigma(\lambda) := b_{1,\sigma(1)} \cdot \dots \cdot b_{n,\sigma(n)}$ ein Polynom vom Grad $|\{i : \sigma(i) = i\}|$. Dies ergibt sofort $\deg p_A \leq n$. Ferner hat nur das Polynom

$$q_{\mathrm{id}}(\lambda) = (a_{11} - \lambda) \cdot \dots \cdot (a_{nn} - \lambda)$$

den Grad n, so dass auch $\deg p_A = n$ folgt. Wegen $\operatorname{sign} \mathrm{id} = 1$ gilt dann auch die Formel für b_n.

Die Formel $b_0 = \det A$ folgt für $\lambda = 0$ und die Formel für b_{n-1} wird hier nicht bewiesen.

Da Polynome vom Grad n maximal n Nullstellen haben können, erhalten wir sofort das folgende Korollar, das die Anzahl von Eigenwerten nach oben begrenzt.

Korollar 10.4.6

Sei $A \in \mathcal{M}(n, n)$. Dann hat A höchstens n verschiedene Eigenwerte.

Nullstellen eines Polynoms können auch mehrfach auftreten. Die folgende Definition berücksichtigt dies.

Definition 10.4.7

Sei $A \in \mathcal{M}_{\mathbb{K}}(n,n)$ und $\lambda \in \mathbb{K}$ ein Eigenwert von A. Dann ist die *algebraische Vielfachheit* $\operatorname{alg}_A(\lambda)$ von λ die Vielfachheit der Nullstelle des charakteristischen Polynoms p_A. Ferner heißt

$$\operatorname{geo}_A(\lambda) := \dim \operatorname{Eig}_A(\lambda)$$

die geometrische Vielfachheit von λ.

Offensichtlich gilt $1 \leq \operatorname{alg}_A(\lambda) \leq n$ und $1 \leq \operatorname{geo}_A(\lambda) \leq n$. Sind ferner $\lambda_1, \ldots, \lambda_m$ alle Eigenwerte von A, so gilt

$$\sum_{i=1}^{m} \operatorname{alg}_A(\lambda_i) \leq n \tag{10.4.2}$$

wobei in $\mathbb{R}$ nicht die Gleichheit gelten muss, wie wir am Beispiel der Drehmatrizen gesehen haben. Im Fall $\mathbb{K} = \mathbb{C}$ gilt jedoch Gleichheit in (10.4.2), da p_A in n Linearfaktoren zerfällt, siehe Satz 2.7.5.

BEISPIEL

Um die beiden Konzepte zu illustrieren, betrachten wir die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} . \tag{10.4.3}$$

Dann gilt

$$p_A(\lambda) = \left| \begin{pmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{pmatrix} \right| = (1-\lambda)^2 .$$

Damit hat hat der Eigenwert $\lambda = 1$ die algebraische Vielfachheit 2. Ferner gilt

$$\mathbf{dim}\, \mathrm{Eig}_A(1) = \mathbf{dim}\, \mathbf{ker}(A - E_n) = \mathbf{dim}\, \mathbf{ker} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 1 ,$$

wobei die letzte Identität z.B. leicht aus der Dimensionsformel aus Satz 9.5.4 und $\mathbf{rang} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 1$ folgt. In diesem Beispiel haben wir also

$$\mathrm{geo}_A(\lambda) < \mathrm{alg}_A(\lambda) .$$

Andererseits haben wir z.B. für $A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ den Eigenwert $\lambda = 0$ mit

$$\mathrm{alg}_A(0) = \mathrm{geo}_A(0) = 2 .$$

Haben wir eine Basis $\mathcal{B}$, die aus Eigenvektoren $v_1, \ldots, v_n \in \mathbb{K}^n$ der Matrix $A \in \mathcal{M}_{\mathbb{K}}(n,n)$ besteht, so finden wir eine besonders einfache Matrix, die die Abbildung $\mathsf{L}_A : \mathbb{K}^n \to \mathbb{K}^n$ darstellt. Wir betrachten dazu die $n \times n$ Matrix

$$S := (v_1, \ldots, v_n)\,.$$

Es gilt $\mathbf{rang}\, S = n$ und damit haben wir $S \in \mathcal{GL}(n)$ nach Korollar 9.5.6. Für $j = 1, \ldots, n$ sei ferner λ_j der Eigenwert zu v_j, d.h. wir haben

$$Av_j = \lambda_j v_j\,, \qquad\qquad j = 1, \ldots, n.$$

Betrachten wir ferner die Diagonalmatrix

$$\Lambda := \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

so gilt

$$AS = \big(Av_1, \ldots, Av_n\big) = \big(\lambda_1 v_1, \ldots, \lambda_n v_n\big) = S\Lambda\,.$$

Mit (9.3.14) und (9.3.16) zeigt dies

$$\Lambda = S^{-1}AS = \mathrm{T}_{\mathcal{B}}^{\mathcal{E}_n} A \mathrm{T}_{\mathcal{E}_n}^{\mathcal{B}} = \mathrm{M}_{\mathcal{B}}^{\mathcal{B}}(\mathrm{L}_A)\,, \tag{10.4.4}$$

d.h. die durch A beschriebene Abbildung L_A hat in der Basis $\mathcal{B}$ die darstellende Matrix Λ.

Diese Beobachtungen inspirieren die folgende Definition.

Definition 10.4.8

Zwei Matrizen $A, B \in \mathcal{M}(n, n)$ heißen **ähnlich**, wenn es ein $S \in \mathcal{GL}(n)$ gibt mit

$$B = S^{-1}AS\,.$$

Die Matrix A heißt **diagonalisierbar**, falls es ein $S \in \mathcal{GL}(n)$ gibt, so dass

$$S^{-1}AS$$

eine Diagonalmatrix ist.

Die Ähnlichkeitsrelation auf $\mathcal{M}(n, n)$ ist, wie man leicht überprüft, eine Äquivalenzrelation. Die diagonalisierbaren Matrizen sind genau die Matrizen, die zu einer Diagonalmatrix ähnlich sind.

Unsere obigen Betrachtungen führen zu der folgenden Charakterisierung diagonalisierbarer Matrizen.

Theorem 10.4.9

Eine Matrix $A \in \mathcal{M}(n, n)$ ist genau dann diagonalisierbar, wenn es eine Basis $\mathcal{B}$ gibt, die aus Eigenvektoren von A besteht.

Gibt es eine Basis aus Eigenwerten, so zeigt (10.4.4), dass A diagonalisierbar ist.

Ist umgekehrt A diagonalisierbar, so gibt es ein $S \in \mathcal{GL}(n)$ und eine Diagonalmatrix $\Lambda \in \mathcal{M}(n,n)$ mit $S^{-1}AS = \Lambda$, d.h.

$$AS = S\Lambda\,.$$

Seien nun $v_1, \ldots, v_n$ die Spaltenvektoren von S, d.h. $S := (v_1, \ldots, v_n)$. Dann gilt

$$\big(Av_1, \ldots Av_n\big) = AS = S\Lambda = \big(\lambda_1 v_1, \ldots, \lambda_n v_n\big)\,.$$

Damit sind $v_1, \ldots, v_n$ Eigenvektoren von A. Da S invertierbar ist, spannen nach Korollar 9.5.6 die Vektoren $v_1, \ldots, v_n$ einen n-dimensionalen Unterraum von $\mathbb{K}^n$ auf und damit sind sie nach Korollar 9.1.14 eine Basis.

Der folgende Satz vergleicht das Eigenwertverhalten ähnlicher Matrizen.

Theorem 10.4.10

Seien $A, B \in \mathcal{M}(n,n)$ und $S \in \mathcal{GL}(n)$ mit $B = S^{-1}AS$. Dann gelten die folgenden Aussagen:

i). *A und B haben die gleichen charakteristischen Polynome.*

ii). *A und B haben die gleichen Eigenwerte $\lambda_1, \ldots, \lambda_m$ und es gilt $\operatorname{alg}_A(\lambda_i) = \operatorname{alg}_B(\lambda_i)$ für alle $i = 1, \ldots, m$.*

iii). *$v \in V$ ist genau dann ein Eigenvektor von A zum Eigenwert λ_i, wenn $w := S^{-1}v$ ein Eigenvektor von B zum Eigenwert λ_i ist.*

iv). *Es gilt $\operatorname{geo}_A(\lambda_i) = \operatorname{geo}_B(\lambda_i)$ für alle $i = 1, \ldots, m$.*

v). *Es gilt $\det A = \det B$ und $\operatorname{spur} A = \operatorname{spur} B$.*

BEWEIS 1/2

i). Es gilt

$$\begin{aligned} p_B(\lambda) = \det(B - \lambda E_n) = \det(S^{-1}AS - \lambda E_n) &= \det(S^{-1}(A - \lambda E_n)S) \\ &= \det S^{-1} \det(A - \lambda E_n) \det S \\ &= \det(S^{-1}S) p_A(\lambda) \\ &= p_A(\lambda)\,. \end{aligned}$$

ii). Folgt sofort aus *i).*

iii). Ist v ein Eigenvektor von A zum Eigenwert λ_i, so gilt

$$Bw = S^{-1}ASS^{-1}v = S^{-1}Av = S^{-1}(\lambda_i v) = \lambda_i S^{-1}v = \lambda_i w\,.$$

Die umgekehrte Implikation folgt aus der Vertauschung der Rollen von A und B mit $S' := S^{-1}$.

BEWEIS 2/2

iv). In *iii)* hatten wir

$$S^{-1}(\mathrm{Eig}_A(\lambda_i)) = \mathrm{Eig}_B(\lambda_i)$$

gesehen. Damit definiert S^{-1} durch Restriktion einen Isomorphismus $\mathrm{Eig}_A(\lambda_i) \to \mathrm{Eig}_B(\lambda_i)$ und Korollar 9.2.8 zeigt dann die Behauptung.

v). Die Aussage für die Determinante folgt aus *i)* mit $\lambda = 0$. Die Aussage für die Spur folgt aus

$$\mathbf{spur}\, B = \mathbf{spur}(S^{-1}AS) = \mathbf{spur}(ASS^{-1}) = \mathbf{spur}\, A$$

wobei wir (10.4.1) für die Matrizen S^{-1} und AS benutzt haben.

Ist $T : V \to V$ eine lineare Abbildung und sind $\mathcal{A}$ und $\mathcal{A}'$ zwei endliche Basen von V, so gilt nach (9.3.16)

$$\mathrm{M}_{\mathcal{A}'}^{\mathcal{A}'}(T) = \mathrm{T}_{\mathcal{A}'}^{\mathcal{A}} \mathrm{M}_{\mathcal{A}}^{\mathcal{A}}(T)(\mathrm{T}_{\mathcal{A}'}^{\mathcal{A}})^{-1},$$

d.h. für $S := (\mathrm{T}_{\mathcal{A}'}^{\mathcal{A}})^{-1} = \mathrm{T}_{\mathcal{A}}^{\mathcal{A}'}$ haben wir

$$\mathrm{M}_{\mathcal{A}'}^{\mathcal{A}'}(T) = S^{-1} \mathrm{M}_{\mathcal{A}}^{\mathcal{A}}(T) S$$

Mit anderen Worten sind verschiedene Darstellungsmatrizen des gleichen Operators T ähnlich zueinander. Satz 10.4.10 ermöglicht es daher, Determinanten, Spuren, charakteristische Polynome und algebraische Vielfachheiten auch für lineare Abbildungen zwischen endlich-dimensionalen Vektorräumen zu definieren.

Eigenwerte, Eigenvektoren, Eigenräume und geometrische Vielfachheiten sind dagegen ohne den Umweg über Matrizen direkt definierbar.

Theorem 10.4.11

Sei $A \in \mathcal{M}(n,n)$. Dann gelten die folgenden Aussagen:

i). Ist λ ein Eigenwert von A, so gilt

$$1 \leq \operatorname{geo}_A(\lambda) \leq \operatorname{alg}_A(\lambda) \leq n\,.$$

ii). Ist A diagonalisierbar, dann gilt

$$\operatorname{geo}_A(\lambda) = \operatorname{alg}_A(\lambda) \tag{10.4.5}$$

für alle Eigenwerte λ von A gilt. Zerfällt p_A zusätzlich in Linearfaktoren, so gilt auch die umgekehrte Implikation.

iii). Sind $v_1, \ldots, v_m$ Eigenvektoren zu paarweise verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_m$, so sind $v_1, \ldots, v_m$ linear unabhängig.

iv). Haben wir paarweise verschiedene Eigenwerte $\lambda_1, \ldots, \lambda_m$ und $\mathcal{B}_j \subset \operatorname{Eig}_A(\lambda_j)$, so dass jedes $\mathcal{B}_j$ aus linear unabhängigen Vektoren besteht, so besteht $\mathcal{B} := \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_m$ aus linear unabhängigen Vektoren.

BEWEIS 1/5

i). Wir fixieren einen Eigenwert λ_0 von A und wählen eine Basis $v_1, \dots, v_m$ von $\mathrm{Eig}_A(\lambda_0)$. Dies ergibt

$$m = \mathrm{geo}_A(\lambda_0)\,.$$

Nun ergänzen wir $v_1, \dots, v_m$ zu einer Basis $\mathcal{B}$ und schreiben $S := \mathsf{L}_A$. Dann ist die darstellende Matrix von S bezüglich der Basis $\mathcal{B}$ von der Block-Gestalt

$$B := \mathsf{M}_{\mathcal{B}}^{\mathcal{B}}(S) = \begin{pmatrix} D_{\lambda_0} & \tilde{A} \\ 0 & A' \end{pmatrix},$$

wobei $D \in \mathcal{M}(m,m)$ eine Diagonalmatrix mit Diagonaleinträgen $(\lambda_0, \dots, \lambda_0)^\mathsf{T} \in \mathbb{K}^m$ ist und $\tilde{A} \in \mathcal{M}(m, n-m)$ und $A' \in \mathcal{M}(n-m, n-m)$ geeignete Matrizen sind. Ähnlich wie für obere Dreiecksmatrizen gilt dann

$$\det B = \det D_\lambda \cdot \det A' = \lambda_0^m \cdot \det A'$$

und wenden wir dies zur Berechnung des charakteristische Polynoms an, so erhalten wir

$$p_B(\lambda) = \left| \begin{pmatrix} D_{\lambda_0 - \lambda} & \tilde{A} \\ 0 & A' - \lambda E_{n-m} \end{pmatrix} \right| = (\lambda_0 - \lambda)^m \cdot \det(A' - \lambda E_{n-m})\,.$$

Damit hat p_B mindestens m-mal die Nullstelle λ_0 und da wir $p_A = p_B$ aus Satz 10.4.10 wissen erhalten wir $m \leq \operatorname{alg}_A(\lambda_0)$. Die restlichen Ungleichungen sind trivial.

iii). Wir zeigen die Behauptung durch Induktion über m. Für den Induktionsanfang $m = 1$ ist hierbei nichts zu zeigen. Sei daher die Behauptung schon für $m - 1$ bekannt. Wir nehmen nun $\alpha_1, \ldots, \alpha_m \in \mathbb{K}$ mit

$$0 = \sum_{j=1}^{m} \alpha_j v_j \,. \tag{10.4.6}$$

Die Multiplikation mit $A - \lambda_m E_n$ ergibt dann

$$0 = \sum_{j=1}^{m} \alpha_j (A - \lambda_m E_n) v_j = \sum_{j=1}^{m} \alpha_j (\lambda_j v_j - \lambda_m v_j) = \sum_{j=1}^{m-1} \alpha_j (\lambda_j - \lambda_m) v_j \,.$$

Nach Induktionsvoraussetzung gilt dann $\alpha_j(\lambda_j - \lambda_m) = 0$ für alle $j = 1, \ldots, m-1$ und wegen $\lambda_j \neq \lambda_m$ für $j \leq m - 1$ folgt dann $\alpha_j = 0$ für alle $j = 1, \ldots, m-1$. Einsetzen dieser α_j in (10.4.6) ergibt $0 = \alpha_m v_m$ und da $v_m \neq 0$ folgt dann auch $\alpha_m = 0$.

iv). Seien $v_{1j}, \dots, v_{k_j,j}$ die Vektoren von $\mathcal{B}_j$. Ferner seien $\alpha_{ij} \in \mathbb{K}$ mit

$$0 = \sum_{j=1}^{m} \sum_{i=1}^{k_j} \alpha_{ij} v_{ij} \, .$$

Wir definieren $v_j := \sum_{i=1}^{k_j} \alpha_{ij} v_{ij} \in \mathrm{Eig}_A(\lambda_j)$. Im Fall $v_j = 0$ zeigt die lineare Unabhängigkeit der $v_{1j}, \dots, v_{k_j,j}$ sofort $\alpha_{1j} = \cdots = \alpha_{k_j,j} = 0$. Sind alle $v_j = 0$, so haben wir also die Behauptung bewiesen. Wir nehmen nun an, dass es Indizes j mit $v_j \neq 0$ gibt. Ohne Einschränkung können wir dann zusätzlich $\{j : v_j \neq 0\} = \{1, \dots, k\}$ für ein $1 \leq k \leq m$ annehmen. Nach unserer Konstruktion sind die Vektoren $v_1, \dots, v_k$ dann Eigenvektoren zu $\lambda_1, \dots \lambda_k$ und wir haben

$$0 = \sum_{j=1}^{m} v_j = \sum_{j=1}^{k} v_j = \sum_{j=1}^{k} \beta_j v_j \, .$$

mit $\beta_j := 1 \neq 0$. Dies widerspricht der linearen Unabhängigkeit von $v_1, \dots, v_k$, die wir aus *iii)* wissen.

BEWEIS 4/5

ii). Im folgenden bezeichnen $\lambda_1, \ldots, \lambda_m$ die paarweise verschiedenen Eigenwerten von A.

Ist A diagonalisierbar, so existiert nach Satz 10.4.9 eine Basis aus Eigenvektoren. Damit gilt $\sum_{i=1}^{m} \operatorname{geo}_A(\lambda_i) = n$. Aus $\sum_{i=1}^{m} \operatorname{alg}_A(\lambda_i) \leq n$ und *i)* folgt dann die Behauptung.

BEWEIS 5/5

Wir nehmen nun umgekehrt an, dass p_A in Linearfaktoren zerfällt und dass (10.4.5) gilt. Dann folgt

$$\sum_{j=1}^{m} \operatorname{alg}_A(\lambda_j) = n\,. \tag{10.4.7}$$

Wählen wir jetzt Basen $\mathcal{B}_j$ von $\operatorname{Eig}_A(\lambda_j)$, dann haben wir nach Voraussetzung zudem

$$|\mathcal{B}_j| = \operatorname{geo}_A(\lambda_j) = \operatorname{alg}_A(\lambda_j) \tag{10.4.8}$$

für alle $j = 1, \ldots, m$. Ferner zeigt *iii)*, dass $\mathcal{B} := \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_m$ aus linear unabhängigen Vektoren besteht. Ausserdem ergibt die Kombination von (10.4.7) und (10.4.8)

$$|\mathcal{B}| = \sum_{j=1}^{m} \operatorname{alg}_A(\lambda_j) = n\,.$$

Damit ist $\mathcal{B}$ eine Basis von $\mathbb{K}^n$, die nach Konstruktion aus Eigenvektoren von S besteht. Satz 10.4.9 zeigt daher, dass A diagonalisierbar ist.

Das folgende Korollar folgt direkt aus Satz 10.4.10 und den Formeln für Spuren und Determinanten von Diagonalmatrizen. Wegen Satz 10.4.11 ist es in der Formulierung egal, welche Vielfachheiten wir betrachten.

Korollar 10.4.12

Sei $A \in \mathcal{M}(n,n)$ diagonalisierbar und $\lambda_1, \ldots, \lambda_n$ alle Eigenwerte von A inklusive ihrer Vielfachheiten. Dann gilt:

$$\operatorname{spur} A = \sum_{j=1}^{n} \lambda_j \,,$$

$$\det A = \prod_{j=1}^{n} \lambda_j \,.$$

BEISPIEL

Als Beispiel betrachten wir nun die Matrix

$$A := \begin{pmatrix} 3 & 2 & 0 \\ -3 & -2 & 0 \\ -3 & -3 & 1 \end{pmatrix}.$$

Durch nachrechnen erhält man dann

$$p_A(\lambda) = \lambda \cdot (1-\lambda)^2, \qquad \lambda \in \mathbb{R},$$

und insbesondere zerfällt p_A in Linearfaktoren. Setzen wir $\lambda_1 := 0$ und $\lambda_2 := 1$, so sehen wir also

$$\operatorname{alg}_A(\lambda_1) = 1 \qquad \text{und} \qquad \operatorname{alg}_A(\lambda_2) = 2\,.$$

Um den Eigenraum $\operatorname{Eig}_A(\lambda_1)$ von λ_1 zu bestimmen, müssen wir das 3×3 Gleichungssystem $Ax = (A - \lambda_1 E_3)x = 0$ lösen. Dies ergibt

$$\operatorname{Eig}_A(\lambda_1) = \mathbb{R} \begin{pmatrix} -2 \\ 3 \\ 3 \end{pmatrix}$$

und insbesondere also $\operatorname{geo}_A(\lambda_1) = \operatorname{alg}_A(\lambda_1)$.

Analog müssen wir zur Bestimmung von $\mathrm{Eig}_A(\lambda_2)$ das 3×3 Gleichungssystem $(A - E_3)x = 0$ lösen. Dies ergibt

$$\mathrm{Eig}_A(\lambda_2) = \mathbb{R} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \mathbb{R} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

wobei die letzten beiden Vektoren linear unabhängig sind. Dies ergibt dann $\mathrm{geo}_A(\lambda_2) = \mathrm{alg}_A(\lambda_2)$. Nach *ii)* von Satz 10.4.11 ist A daher diagonalisierbar und *iv)* von Satz 10.4.11 zeigt, dass

$$\begin{pmatrix} -2 \\ 3 \\ 3 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

eine Basis von Eigenvektoren von A bilden. Setzen wir nun

$$S := \begin{pmatrix} -2 & -1 & 0 \\ 3 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix},$$

so erhalten wir mit (10.4.4)

$$S^{-1}AS = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

BEISPIEL

Betrachten wir noch einmal die Matrix

$$A := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

so hatten wir schon hinter (10.4.3) gesehen, dass $\lambda = 1$ ein Eigenwert von A mit

$$\mathrm{geo}_A(1) = 1 < 2 = \mathrm{alg}_A(1)$$

ist. Daher kann A nach Satz 10.4.11 nicht diagonalisierbar sein.

JORDAN-BLÖCKE

Quadratische Matrizen $A \in \mathcal{M}(n,n)$, mit $a_{ii} = \lambda$ und $a_{i,i+1} = 1$ und $a_{ij} = 0$ sonst heißen **Jordan-Blöcke**. Die obige 2×2-Matrix ist ein Jordan-Block mit $\lambda = 1$ und

$$A := \begin{pmatrix} \lambda & 1 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 \\ 0 & 0 & \lambda & 1 & 0 \\ 0 & 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix}$$

ist ein 5×5-Jordan-Block. Man kann zeigen, dass jede Matrix zu einer Matrix ähnlich ist, die aus Jordan-Blöcken und einer Diagonalmatrix besteht.

Section 10.5
Symmetrische und hermitesche Matrizen

Wir hatten in Kapitel 4 gesehen, dass eine Matrix genau dann diagonalisierbar ist, wenn es eine Basis aus Eigenvektoren gibt. Bis jetzt hatten wir aber keine *einfachen* Kriterien kennengelernt, um zu überprüfen, ob eine Matrix diagonalisierbar ist. Außerdem haben wir gesehen, dass Orthonormalbasen einige Vorteile gegenüber anderen Basen haben. Es wäre also sicher interessant zu wissen, wann wir eine Orthonormalbasis von Eigenvektoren erwarten können. Die folgende Definition wird für diese Fragen eine zentrale Rolle spielen.

Definition 10.5.1

Eine Matrix $A \in \mathcal{M}_{\mathbb{K}}(n, n)$ heißt:

1. symmetrisch, falls $\mathbb{K} = \mathbb{R}$ und $A^{\mathsf{T}} = A$ gilt.
2. hermitesch, falls $\mathbb{K} = \mathbb{C}$ und $\overline{A}^{\mathsf{T}} = A$ gilt.

Ist A eine reelle, symmetrische Matrix, so gilt $A = \overline{A}$ und damit auch $\overline{A}^{\mathsf{T}} = A$. Fasst man reelle symmetrische Matrizen als komplexe Matrizen auf, so sind diese also hermitesch. Im folgenden werden wir uns daher häufig auf hermitesche Matrizen beschränken.

Ist eine Matrix $A \in \mathcal{M}_{\mathbb{K}}(n,n)$ hermitesch, so gilt für die dargestellte Abbildung $S := \mathsf{L}_A$

$$S^* = \mathsf{L}_{\overline{A}^{\mathsf{T}}} = \mathsf{L}_A = S\,,$$

siehe (9.3.7). Haben wir umgekehrt eine selbstadjungierte Abbildung, d.h. eine lineare Abbildung $S : \mathbb{K}^n \to \mathbb{K}^n$ mit $S^* = S$, so zeigt eine analoge Rechnung, dass $A := \mathsf{M}(S)$ hermitesch ist. Hermitesche Matrizen sind also genau die Matrizen, die selbstadjungierte Abbildungen darstellen. Im Fall $\mathbb{K} = \mathbb{C}$ werden solche Abbildungen auch hermitesch genannt.

Nach Lemma 9.3.5 gilt für hermitesche Matrizen A

$$\langle y, Ax\rangle = \langle \overline{A}^{\mathsf{T}} y, x\rangle = \langle Ay, x\rangle\,, \qquad x, y \in \mathbb{K}^n, \tag{10.5.1}$$

und durch Betrachtung von $x = e_i$ und $y = e_j$ für alle $i, j = 1, \ldots, n$ sieht man leicht, dass diese Gleichung hermitesche Matrizen charakterisiert.

Für hermitesche Matrizen $A \in \mathcal{M}(n,n)$ gilt

$$a_{ii} = \overline{a}_{ii}\,, \qquad i = 1, \ldots, n,$$

d.h. ihre Diagonaleinträge sind immer reell.

BESPIEL

Schließlich betrachten wir noch das Beispiel

$$A := \begin{pmatrix} 2 & 1+\mathrm{i} \\ 1-\mathrm{i} & 1 \end{pmatrix}.$$

Dann gilt

$$\overline{A}^{\mathsf{T}} = \begin{pmatrix} \overline{2} & \overline{1-\mathrm{i}} \\ \overline{1+\mathrm{i}} & \overline{1} \end{pmatrix} = \begin{pmatrix} 2 & 1+\mathrm{i} \\ 1-\mathrm{i} & 1 \end{pmatrix} = A\,,$$

d.h. A ist hermitesch.

Der folgende Satz untersucht die Eigenwerte und zugehörigen Eigenräume hermitescher Matrizen.

Theorem 10.5.2

Sei $A \in \mathcal{M}(n,n)$ hermitesch und $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$ die Eigenwerte von A. Dann gelten die folgenden Aussagen:

i). Alle Eigenwerte sind reell, d.h. $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$.

ii). Eigenvektoren zu unterschiedlichen Eigenwerten sind orthogonal zueinander, d.h. für $\lambda_i \neq \lambda_j$ gilt

$$\mathrm{Eig}_A(\lambda_i) \perp \mathrm{Eig}_A(\lambda_j)\,.$$

Komplexe Matrizen haben immer *n* Eigenwerte, wenn man die algebraischen Vielfachheiten berücksichtigt. Fassen wir eine symmetrische Matrix als hermitesche Matrix auf, so hat diese daher *n* komplexe Eigenwerte, die nach Satz 10.5.2 sogar reell sind. *Symmetrische Matrizen haben also immer n reelle Eigenwerte, wenn man die algebraischen Vielfachheiten berücksichtigt.*

i). Sei $v \in \mathbb{C}^n$ ein Eigenvektor zum Eigenwert λ von A. Dann gilt:

$$\lambda\langle v, v\rangle = \langle v, \lambda v\rangle = \langle v, Av\rangle = \langle Av, v\rangle = \langle \lambda v, v\rangle = \overline{\lambda}\langle v, v\rangle \,.$$

Wegen $\langle v, v\rangle > 0$ folgt $\lambda = \overline{\lambda}$, d.h. $\lambda \in \mathbb{R}$.

ii). Seien v ein Eigenvektor zum Eigenwert λ_i und w ein Eigenvektor zum Eigenwert λ_j. Nach *i)* wissen wir $\overline{\lambda_i} = \lambda_i$ und damit folgt

$$\begin{aligned}(\lambda_i - \lambda_j)\langle v, w\rangle = \langle \lambda_i v, w\rangle - \langle v, \lambda_j w\rangle &= \langle Av, w\rangle - \langle v, Aw\rangle \\ &= \langle v, Aw\rangle - \langle v, Aw\rangle \\ &= 0\,,\end{aligned}$$

wobei wir im vorletzten Schritt (10.5.1) ausgenutzt haben. Wegen $\lambda_i \neq \lambda_j$ erhalten wir dann $\langle v, w\rangle = 0$.

ONB AUS EIGENVEKTOREN

Ist $A \in \mathcal{M}(n,n)$ eine hermitesche Matrix mit $\operatorname{geo}_A(\lambda_i) = \operatorname{alg}_A(\lambda_i)$ für alle paarweise verschiedenen Eigenwerte $\lambda_1, \ldots, \lambda_m$ von A, so können wir mit Satz 10.5.2 leicht eine Orthonormalbasis aus Eigenvektoren von A konstruieren. Dazu müssten wir lediglich Orthonormalbasen $\mathcal{B}_i$ für allen Eigenräume $\operatorname{Eig}_A(\lambda_i)$ konstruieren. Nach Satz 10.5.2 ist dann $\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_m$ ein Orthonormalsystem und wegen

$$\sum_{i=1}^{m} |\mathcal{B}_i| = \sum_{i=1}^{m} \operatorname{geo}_A(\lambda_i) = \sum_{i=1}^{m} \operatorname{alg}_A(\lambda_i) = n$$

ist dieses System auch eine Basis.

Der folgende Satz, der als Spektralsatz bekannt ist, zeigt insbesondere, dass diese Konstruktion für alle hermiteschen Matrizen möglich ist.

Theorem 10.5.3

Sei $A \in \mathcal{M}(n,n)$ hermitesch. Dann gilt:

i). Es gibt n reelle Eigenwerte $\lambda_1, \dots, \lambda_n$ von A, wobei algebraische Vielfachheiten berücksichtigt sind.

ii). Es gibt eine Orthonormalbasis $v_1, \dots, v_n$ von $\mathbb{K}^n$, so dass jedes v_i Eigenvektor zum Eigenwert λ_i ist.

iii). Die Matrix $S := (v_1, \dots, v_n)$ ist orthogonal/unitär und A ist diagonalisierbar mit

$$\overline{S}^{\mathsf{T}} A S = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}. \tag{10.5.2}$$

iv). Für alle $x \in \mathbb{K}^n$ gilt

$$Ax = \sum_{i=1}^{n} \lambda_i \langle v_i, x \rangle v_i \,.$$

BEWEIS 1/3

i). Dies haben wir bereits nach Satz 10.5.2 festgestellt.

ii). Wir zeigen die Behauptung durch Induktion über n, wobei für den Induktionsanfang $n = 1$ die Aussage trivial ist.

Für den Induktionsschritt von $n-1$ nach n nehmen wir an, dass *ii)* für $n-1$ wahr ist. Sei nun v_1 ein Eigenvektor zum Eigenwert λ_1. Wir ergänzen diesen zu einer Orthonormalbasis $v_1, w_2, \ldots, w_n$ von $\mathbb{K}^n$. Dann ist $R := (v_1, w_2, \ldots, w_n)$ nach Satz 10.3.8 orthogonal/unitär. Wir setzen $B := \overline{R}^{\mathsf{T}} A R$. Dann ist B wegen

$$\overline{B}^{\mathsf{T}} = \left(R^{\mathsf{T}} \cdot \overline{A} \cdot \overline{R}\right)^{\mathsf{T}} = \overline{R}^{\mathsf{T}} \overline{A}^{\mathsf{T}} \left(R^{\mathsf{T}}\right)^{\mathsf{T}} = \overline{R}^{\mathsf{T}} A R = B$$

hermitesch. Ferner gilt

$$B e_1 = \overline{R}^{\mathsf{T}} A R e_1 = \overline{R}^{\mathsf{T}} A v_1 = \overline{R}^{\mathsf{T}} (\lambda_1 v_1) = \lambda_1 \overline{R}^{\mathsf{T}} v_1 = \lambda_1 e_1 \,,$$

wobei wir $R e_1 = v_1$ und damit auch $\overline{R}^{\mathsf{T}} v_1 = R^{-1} v_1 = e_1$ nach Satz 10.3.8 ausgenutzt haben. Damit ist B von der Blockform

$$B = \begin{pmatrix} \lambda_1 & 0 \\ 0 & A' \end{pmatrix},$$

wobei $A' \in \mathcal{M}(n-1, n-1)$ eine hermitesche Matrix ist.

BEWEIS 2/3

Nach der Induktionsannahme gibt es dann eine Orthonormalbasis $v_2', \dots, v_n'$ von $\mathbb{K}^{n-1}$, die aus Eigenvektoren zu den Eigenwerten $\lambda_2', \dots, \lambda_n'$ von A' besteht. Die Matrix $S_1 := (v_2', \dots, v_n')$ ist dann nach Satz 10.3.8 orthogonal/unitär und es gilt

$$\overline{S}_1^{\mathsf{T}} = S_1^{-1}.$$

Damit zeigt (10.4.4)

$$\overline{S}_1^{\mathsf{T}} A' S_1 = \begin{pmatrix} \lambda_2' & & 0 \\ & \ddots & \\ 0 & & \lambda_n' \end{pmatrix}.$$

Wir setzen nun im Sinne von Blockmatrizen

$$T := \begin{pmatrix} 1 & 0 \\ 0 & S_1 \end{pmatrix}.$$

Nach Konstruktion sind dann die Spaltenvektoren von T eine Orthonormalbasis von $\mathbb{K}^n$. Damit ist T orthogonal/unitär und nach Lemma 10.3.7 ist dann auch $S := RT$ orthogonal/unitär. Sind $v_1, \dots, v_n$ die Spaltenvektoren von S, so bilden diese nach Satz 10.3.8 eine Orthonormalbasis von $\mathbb{K}^n$.

BEWEIS 3/3

Ferner zeigt Nachrechnen

$$\overline{S}^{\mathsf{T}} A S = \overline{RT}^{\mathsf{T}} A R T = \overline{T}^{\mathsf{T}} B T = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2' & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n' \end{pmatrix} =: \Lambda,$$

d.h. wir haben $S\Lambda = AS$. Für $i \geq 2$ zeigt dies wiederum

$$Av_i = ASe_i = S\Lambda e_i = S\lambda_i' e_i = \lambda_i' v_i \,,$$

d.h. λ_i' ist ein Eigenwert von A und v_i ist ein zugehöriger Eigenvektor.

iii). Die Orthogonalität von S folgt direkt aus *ii)* und Satz 10.3.8, der auch $\overline{S}^{\mathsf{T}} = S^{-1}$ zeigt. Nach Satz 10.4.9 und *ii)* ist A diagonalisierbar und (10.4.4) zeigt (10.5.2).

iv). Für $x \in \mathbb{K}^n$ gilt $x = \sum_{i=1}^n \langle v_i, x\rangle v_i$. Eine einfache Rechnung zeigt dann

$$Ax = A\left(\sum_{i=1}^n \langle v_i, x\rangle v_i\right) = \sum_{i=1}^n \langle v_i, x\rangle Av_i = \sum_{i=1}^n \lambda_i \langle v_i, x\rangle v_i \,,$$

da v_i Eigenvektor zum Eigenwert λ_i ist.

BEISPIEL

Im folgenden betrachten wir die symmetrische Matrix

$$A := \begin{pmatrix} 1 & 3 & 0 \\ 3 & -2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Für diese Matrix ist das charakteristische Polynom

$$p_A(\lambda) = -(\lambda - 1)(\lambda + 4)(\lambda - 3), \qquad \lambda \in \mathbb{R},$$

und damit sind die Eigenwerte $\lambda_1 := 1$, $\lambda_2 := -4$ und $\lambda_3 := 3$. Lösen der Gleichungssysteme zugehörigen $(A - \lambda_i E_n)x = 0$ ergibt die normierten Eigenvektoren

$$v_1 := \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \qquad v_2 := \frac{1}{\sqrt{35}} \begin{pmatrix} -3 \\ 5 \\ 1 \end{pmatrix}, \qquad v_3 := \frac{1}{\sqrt{14}} \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix}.$$

Diese bilden eine Orthonormalbasis von Eigenvektoren von A und für

$$S := (v_1, v_2, v_3) = \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{-3}{\sqrt{35}} & \frac{3}{\sqrt{14}} \\ 0 & \frac{5}{\sqrt{35}} & \frac{2}{\sqrt{14}} \\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{35}} & \frac{-1}{\sqrt{14}} \end{pmatrix}$$

gilt

$$S^{\mathsf{T}} A S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Im Folgenden wollen wir uns mit hermiteschen Matrizen, deren Eigenwerte alle positiv sind, näher beschäftigen. Wir beginnen mit der folgenden Definition.

Definition 10.5.4

Sei $A \in \mathcal{M}(n,n)$ eine hermitesche Matrix. Dann heißt A positiv semi-definit, falls für alle $v \in \mathbb{K}^n$ gilt

$$\langle v, Av \rangle \geq 0 \,.$$

Gilt für alle $v \in \mathbb{K}^n$ mit $v \neq 0$ sogar

$$\langle v, Av \rangle > 0 \,,$$

so heißt A positiv definit.

Eine hermitesche Matrix A heißt **negativ (semi)-definit**, falls $-A$ positiv (semi)-definit ist. Matrizen, die weder positiv noch negativ semi-definit sind, heißen **indefinit**. In diesem Fall gibt es also $v, w \in \mathbb{K}^n$ mit

$$\langle v, Av \rangle < 0 \qquad \text{und} \qquad \langle w, Aw \rangle > 0 \,.$$

Die Einheitsmatrix E_n ist positiv definit. Ist allgemeiner D eine $n \times n$-Diagonalmatrix, so ist diese positiv semi-definit, genau dann wenn alle Diagonaleinträge nicht-negativ sind, d.h. $d_i \geq 0$ für alle $i = 1, \ldots, n$. Ferner ist D positiv definit, genau dann wenn $d_i > 0$ für alle $i = 1, \ldots, n$ gilt.

Der folgende Satz verallgemeinert diese Beobachtung mit Hilfe der Eigenwerte.

Theorem 10.5.5

Sei $A \in \mathcal{M}(n, n)$ hermitesch. Dann sind die folgenden Aussagen äquivalent:

i). A ist positiv semi-definit.

ii). Alle Eigenwerte $\lambda_1, \ldots, \lambda_n$ von A sind nicht-negativ, d.h. $\lambda_i \geq 0$ für alle $\lambda = 1, \ldots, n$.

Eine analoge Charakterisierung, bei der alle Eigenwerte $\lambda_i > 0$ erfüllen, gilt für positiv definite Matrizen. Offensichtlich ist eine Matrix genau dann indefinit, falls es Eigenwerte $\lambda_i > 0$ und $\lambda_j < 0$ gibt.

BEWEIS 1/2

i) $\Rightarrow$ *ii)*. Sei v Eigenvektor zum Eigenwert λ. Dann gilt

$$\lambda\langle v, v\rangle = \langle v, \lambda v\rangle = \langle v, Av\rangle \geq 0\,.$$

Aus $v \neq 0$ folgt dann $\lambda \geq 0$.

ii) $\Rightarrow$ *i)*. Sei $v_1, \ldots, v_n$ eine Orthonormalbasis aus Eigenvektoren von A gemäß Satz 10.5.3. Für $v \in \mathbb{K}^n$ gilt dann nach Satz 10.5.3

$$Av = \sum_{i=1}^{n} \langle v_i, v\rangle Av_i = \sum_{i=1}^{n} \lambda_i \langle v_i, v\rangle v_i\,.$$

BEWEIS 2/2

Insgesamt ergibt dies

$$\begin{aligned}\langle v, Av\rangle = \left\langle \sum_{i=1}^{n} \langle v_i, v\rangle v_i, \sum_{j=1}^{n} \lambda_j \langle v_j, v\rangle v_j \right\rangle &= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle \langle v_i, v\rangle v_i, \lambda_j \langle v_j, v\rangle v_j \right\rangle \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\langle v_i, v\rangle} \lambda_j \langle v_j, v\rangle \langle v_i, v_j\rangle \\ &= \sum_{i=1}^{n} \overline{\langle v_i, v\rangle} \lambda_i \langle v_i, v\rangle \\ &= \sum_{i=1}^{n} |\langle v_i, v\rangle|^2 \lambda_i \\ &\geq 0\,,\end{aligned}$$

wobei wir im letzten Teil $\lambda_i \geq 0$ ausgenutzt haben.

Korollar 10.5.6

Sei $A \in \mathcal{M}(n,n)$ eine positiv definite Matrix, $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A und S eine Matrix aus Eigenvektoren gemäß Satz 10.5.3. Dann ist A invertierbar und es gilt

$$A^{-1} = S \begin{pmatrix} \lambda_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda_n^{-1} \end{pmatrix} \overline{S}^{\mathsf{T}}. \tag{10.5.3}$$

Wie nach Satz 10.5.5 bemerkt gilt $\lambda_1, \ldots, \lambda_n > 0$. Damit ist die Diagonalmatrix in (10.5.3) definiert. Ferner gilt mit $\overline{S}^{\mathsf{T}} = S^{-1}$ und (10.5.2)

$$\begin{aligned} AA^{-1} &= S \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \overline{S}^{\mathsf{T}} S \begin{pmatrix} \lambda_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda_n^{-1} \end{pmatrix} \overline{S}^{\mathsf{T}} \\ &= S \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \begin{pmatrix} \lambda_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda_n^{-1} \end{pmatrix} \overline{S}^{\mathsf{T}} \\ &= S\overline{S}^{\mathsf{T}} \\ &= E_n \,. \end{aligned}$$

Die Identität $A^{-1}A = E_n$ kann analog gezeigt werden.

Das folgende Lemma zeigt eine wichtige Methode, positiv semi-definite Matrizen zu erzeugen.

Lemma 10.5.7

Sei $A \in \mathcal{M}(m,n)$. Dann sind die Matrizen

$$B := \overline{A}^{\mathsf{T}} A \in \mathcal{M}(n,n)\,,$$
$$C := A\overline{A}^{\mathsf{T}} \in \mathcal{M}(m,m)$$

positiv semi-definit. Ferner ist $\lambda > 0$ ein Eigenwert von B genau dann, wenn λ ein Eigenwert von C ist. Für $v \in \mathrm{Eig}_B(\lambda)$ gilt in diesem Fall $Av \in \mathrm{Eig}_C(\lambda)$ und analog gilt für $w \in \mathrm{Eig}_C(\lambda)$ auch $\overline{A}^{\mathsf{T}} w \in \mathrm{Eig}_B(\lambda)$. Schließlich gilt

$$\ker B = \ker A \qquad \textit{und} \qquad \ker C = \ker \overline{A}^{\mathsf{T}}\,.$$

BEWEIS

Es gilt $\overline{B}^{\mathsf{T}} = (A^{\mathsf{T}}\overline{A})^{\mathsf{T}} = \overline{A}^{\mathsf{T}}A = B$, d.h. B ist hermitesch. Für $v \in \mathbb{K}^n$ gilt ferner

$$\langle v, Bv\rangle = \langle v, \overline{A}^{\mathsf{T}}Av\rangle = \langle Av, Av\rangle \geq 0\,,$$

wobei wir im zweiten Schritt $\overline{\overline{A}^{\mathsf{T}}}^{\mathsf{T}} = A$ benutzt haben. Die Aussage für C folgt aus der Aussage für B durch Betrachtung von $A' := \overline{A}^{\mathsf{T}}$.

Sei nun $\lambda > 0$ ein Eigenwert von B und $v \in \mathbb{K}^n$ ein zugehöriger Eigenvektor, d.h. $\overline{A}^{\mathsf{T}}Av = \lambda v$. Dann gilt $w := Av \neq 0$, denn sonst hätten wir $\lambda v = \overline{A}^{\mathsf{T}}w = 0$, was $\lambda > 0$ und $v \neq 0$ widerspricht. Ferner haben wir

$$Cw = A\overline{A}^{\mathsf{T}}Av = A\lambda v = \lambda Av = \lambda w\,.$$

Durch Betrachtung von $A' := \overline{A}^{\mathsf{T}}$ erhalten wir wieder die umgekehrten Aussagen.

Sei nun $v \in \ker B$, d.h. $\overline{A}^{\mathsf{T}}Av = 0$. Dies ergibt

$$\langle Av, Av\rangle = \langle v, \overline{A}^{\mathsf{T}}Av\rangle = 0\,,$$

und damit folgt $Av = 0$, d.h. $v \in \ker A$. Ist umgekehrt $v \in \ker A$, so folgt $Bv = \overline{A}^{\mathsf{T}}Av = \overline{A}^{\mathsf{T}}0 = 0$. Die zweite Identität folgt wieder aus der ersten durch Betrachtung von $A' := \overline{A}^{\mathsf{T}}$.

Wir haben im Spektralsatz 10.5.3 gesehen, dass hermitesche Matrizen durch einen Wechsel auf eine geeignete Orthonormalbasis auf Diagonalform gebracht werden können. Außerdem haben wir in Satz 9.5.9 gesehen, dass es zu jeder Matrix $A \in \mathcal{M}(m,n)$ zwei invertierbare Matrizen $S \in \mathcal{M}(m,m)$ und $T \in \mathcal{M}(n,n)$ mit

$$SAT^{-1} = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}$$

gibt, wobei $r = \mathbf{rang}\, A$ gilt und die Matrizen S und T geeignete Basiswechselmatrizen $S := \mathrm{T}^{\mathcal{B}}_{\mathcal{E}_n}$ und $T := \mathrm{T}^{\mathcal{A}}_{\mathcal{E}_n}$ sind. Der folgende Satz, der als Singulärwert-Zerlegung bekannt ist, zeigt, dass wir S und T "im wesentlichen" als orthogonal/unitär wählen können.

Theorem 10.5.8

Sei $A \in \mathcal{M}(m,n)$ und $r := \operatorname{rang} A$. Dann gibt es orthogonale/unitäre Matrizen $U \in \mathcal{M}(m,m)$ und $V \in \mathcal{M}(n,n)$ und eine Matrix $\Sigma \in \mathcal{M}(m,n)$ von der Form

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix},$$

wobei $D \in \mathcal{M}(r,r)$ eine Diagonalmatrix mit positiven Diagonaleinträgen ist, so dass gilt

$$A = U\Sigma\overline{V}^{\mathsf{T}}.$$

Sind ferner $\lambda_1, \ldots, \lambda_r$ die positiven Eigenwerte von $\overline{A}^{\mathsf{T}}A$, so hat D die Diagonaleinträge $\lambda_1^{1/2}, \ldots, \lambda_r^{1/2}$.

Sind $v_1, \ldots, v_n$ die Spaltenvektoren von V und $u_1, \ldots, u_m$ die von U, so bilden beide Familien eine Orthonormalbasis von $\mathbb{K}^n$, bzw. $\mathbb{K}^m$. Wegen $Ve_i = v_i$ haben wir ferner $e_i = V^{-1}v_i = \overline{V}^{\mathsf{T}}v_i$ und die Singulärwert-Zerlegung ergibt für $i = 1, \ldots, r$

$$Av_i = U\Sigma\overline{V}^{\mathsf{T}}v_i = U\Sigma e_i = U\sigma_i e_i = \sigma_i u_i \tag{10.5.4}$$

wobei die Diagonaleinträge $\sigma_i := \lambda_i^{1/2}$ von D die Singulärwerte von A heißen. Modulo der Skalierung mit den Singulärwerten, werden also die ersten r Basiselemente v_i auf die ersten r Basiselemente u_i abgebildet.

BEWEIS 1/4

Wir setzen $B := \overline{A}^{\mathsf{T}} A \in \mathcal{M}(n,n)$. Da diese Matrix nach Lemma 10.5.7 hermitesch ist, gibt es nach dem Spektralsatz 10.5.3 eine orthogonale/unitäre Matrix $V \in \mathcal{M}(n,n)$ und eine Diagonalmatrix $\Lambda \in \mathcal{M}(n,n)$ bestehend aus den Eigenwerten von B mit

$$\overline{V}^{\mathsf{T}} B V = \Lambda \,. \tag{10.5.5}$$

Durch Sortieren der Eigenwerte und zugehörigen Eigenvektoren können wir dabei Λ auf die Gestalt

$$\Lambda = \begin{pmatrix} \Lambda_{\neq 0} & 0 \\ 0 & 0 \end{pmatrix}$$

bringen, wobei die Diagonaleinträge $\lambda_1, \ldots, \lambda_{r'}$ der Diagonalmatrix $\Lambda_{\neq 0} \in \mathcal{M}(r', r')$ alle ungleich 0 sind. Nach der Bemerkung nach Satz 10.5.5 sind diese Diagonaleinträge dann alle positiv und mit Lemma 9.5.8 gilt

$$r = \operatorname{\mathbf{rang}} A = \operatorname{\mathbf{rang}} \Lambda = r' \,.$$

Wir schreiben $D := \Lambda_{\neq 0}^{1/2}$, d.h. $D \in \mathcal{M}(r,r)$ ist die Diagonalmatrix mit den Einträgen $\lambda_1^{1/2}, \ldots, \lambda_r^{1/2}$.

BEWEIS 2/4

Seien nun $v_1, \ldots, v_n$ die Spaltenvektoren von V. Wir betrachten die Matrizen $V_1 := (v_1, \ldots, v_r) \in \mathcal{M}(n, r)$ und $V_2 := (v_{r+1}, \ldots, v_n) \in \mathcal{M}(n, n-r)$. Im Sinne von Blockmatrizen ergibt dies $V = (V_1, V_2)$ und (10.5.5) liest sich als

$$\begin{pmatrix} \overline{V_1}^{\mathsf{T}} \\ \overline{V_2}^{\mathsf{T}} \end{pmatrix} \cdot B \cdot (V_1, V_2) = \begin{pmatrix} D^2 & 0 \\ 0 & 0 \end{pmatrix} . \tag{10.5.6}$$

Wir definieren nun

$$U_1 := AV_1D^{-1} \in \mathcal{M}(m, r) ,$$

wobei $D^{-1} \in \mathcal{M}(r, r)$ die Diagonalmatrix mit den Einträgen $\lambda_1^{-1/2}, \ldots, \lambda_r^{-1/2}$ ist. Es gilt dann

$$\begin{aligned} \overline{U}_1^{\mathsf{T}} U_1 = \left(\overline{A} \cdot \overline{V_1} D^{-1}\right)^{\mathsf{T}} AV_1D^{-1} &= D^{-1}\overline{V_1}^{\mathsf{T}} \cdot \overline{A}^{\mathsf{T}} AV_1D^{-1} \\ &= D^{-1}\overline{V_1}^{\mathsf{T}} \cdot BV_1D^{-1} \\ &= D^{-1}D^2D^{-1} \\ &= E_r , \end{aligned}$$

wobei wir im vorletzten Schritt (10.5.6) benutzt haben.

BEWEIS 3/4

Sind $u_1, \ldots, u_r \in \mathbb{K}^n$ die Spaltenvektoren von U_1, so zeigt diese Rechnung

$$\langle u_i, u_j \rangle = \overline{u}_i^{\mathsf{T}} u_j = \delta_{ij} \,,$$

d.h. $u_1, \ldots, u_r$ bilden ein Orthonormalsystem in $\mathbb{K}^m$. Wir ergänzen dieses durch Vektoren $u_{r+1}, \ldots, u_m$ zu einer Orthonormalbasis von $\mathbb{K}^m$ und setzen

$$U_2 := (u_{r+1}, \ldots, u_m) \in \mathcal{M}(m, m-r) \,.$$

Dann ist $U := (U_1, U_2) \in \mathcal{M}(m,m)$ nach Satz 10.3.8 orthogonal/unitär. Unsere Konstruktion liefert dann

$$U \Sigma \overline{V}^{\mathsf{T}} = (U_1, U_2) \cdot \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \overline{V_1}^{\mathsf{T}} \\ \overline{V_2}^{\mathsf{T}} \end{pmatrix} = (U_1 D, 0) \cdot \begin{pmatrix} \overline{V_1}^{\mathsf{T}} \\ \overline{V_2}^{\mathsf{T}} \end{pmatrix} = U_1 D \overline{V_1}^{\mathsf{T}} \,.$$

Ferner haben wir

$$U_1 D \overline{V_1}^{\mathsf{T}} = A V_1 D^{-1} D \overline{V_1}^{\mathsf{T}} = A V_1 \overline{V_1}^{\mathsf{T}} = A\big(E_n - V_2 \overline{V_2}^{\mathsf{T}}\big) = A - A V_2 \overline{V_2}^{\mathsf{T}} \,,$$

wobei wir im vorletzten Schritt benutzt haben, dass $V_1 \overline{V_1}$ die Orthogonalprojektion auf den von $v_1, \ldots, v_r$ aufgespannten Raum ist, siehe Satz 10.3.3, und diese nach (10.3.5) und (10.3.7) auch durch $E_n - V_2 \overline{V_2}$ berechnet werden kann.

BEWEIS 4/4

Nun ist $v_{r+1}, \ldots, v_n$ nach Konstruktion eine Orthonormalbasis von $\ker B$, und damit zeigt Satz 10.3.3, dass $V_2\overline{V_2}^{\mathsf{T}}$ die Orthogonalprojektion auf $\ker B$ beschreibt. Ferner zeigt Lemma 10.5.7

$$\ker B = \ker \overline{A}^{\mathsf{T}} A = \ker A$$

und dies ergibt $AV_2\overline{V_2}^{\mathsf{T}} = 0$. Kombinieren wir diese Gleichungen erhalten wir $U\Sigma\overline{V}^{\mathsf{T}} = A$.

In der Regel haben Matrizen A keine Inverse, und dies gilt insbesondere für nicht-quadratische Matrizen. Im folgenden wollen wir untersuchen, ob es nicht zumindest Matrizen $A^\dagger$ gibt, die einige Eigenschaften mit Inversen Matrizen teilen. Die folgende Definition präzisiert dies.

Definition 10.5.9

Sei $A \in \mathcal{M}(m,n)$. Dann heißt $X \in \mathcal{M}(n,m)$ eine **Moore-Penrose Pseudo-Inverse** von A, falls die folgenden vier Gleichungen gelten:

$$AXA = A\,, \tag{10.5.7}$$

$$XAX = X\,, \tag{10.5.8}$$

$$\overline{(AX)}^{\mathsf{T}} = AX\,, \tag{10.5.9}$$

$$\overline{(XA)}^{\mathsf{T}} = XA\,. \tag{10.5.10}$$

Ist $m = n$ und $A \in \mathcal{M}(n,n)$ invertierbar, so ist A^{-1} eine Moore-Penrose Pseudo-Inverse von A, wie man durch Nachrechnen sofort überprüfen kann.

EINDEUTIGKEIT DER PSEUDO-INVERSEN

Der folgende Satz zeigt, dass es höchstens eine Moore-Penrose Pseudo-Inverse zu einer gegebenen Matrix geben Kann.

Theorem 10.5.10

Sei $A \in \mathcal{M}(m, n)$. Dann gibt es höchstens eine Moore-Penrose Pseudo-Inverse von A, die wir im folgenden mit $A^\dagger$ bezeichnen.

BEWEIS 1/2

Wir nehmen an, dass X und Y zwei Moore-Penrose Pseudo-Inverse von A sind. Mit (10.5.9) und (10.5.8) und gilt dann

$$X\overline{X}^{\mathsf{T}}\overline{A}^{\mathsf{T}} = X\overline{(AX)}^{\mathsf{T}} = XAX = X$$

und (10.5.9) und (10.5.7) liefern

$$\overline{A}^{\mathsf{T}}AY = \overline{A}^{\mathsf{T}}\overline{(AY)}^{\mathsf{T}} = \overline{(AYA)}^{\mathsf{T}} = \overline{A}^{\mathsf{T}}\,.$$

Mit (10.5.9) und (10.5.8) ergibt dies

$$\begin{aligned} X = X\overline{X}^{\mathsf{T}}\overline{A}^{\mathsf{T}} = X\overline{X}^{\mathsf{T}}\overline{A}^{\mathsf{T}}AY = X\overline{(AX)}^{\mathsf{T}}AY = XAXAY \\ = XAY\,. \end{aligned} \tag{10.5.11}$$

BEWEIS 2/2

Ferner gilt mit (10.5.10) und (10.5.8)

$$\overline{A}^{\mathsf{T}}\overline{Y}^{\mathsf{T}}Y = \overline{(YA)}^{\mathsf{T}}Y = YAY = Y$$

und (10.5.10) mit (10.5.7) liefert

$$XA\overline{A}^{\mathsf{T}} = \overline{(XA)}^{\mathsf{T}}A = \overline{A}^{\mathsf{T}}\overline{X}^{\mathsf{T}}\overline{A}^{\mathsf{T}} = \overline{(AXA)}^{\mathsf{T}} = \overline{A}^{\mathsf{T}}\,.$$

Damit haben wir

$$XAY = XA\overline{A}^{\mathsf{T}}\overline{Y}^{\mathsf{T}}Y = \overline{A}^{\mathsf{T}}\overline{Y}^{\mathsf{T}}Y = \overline{(YA)}^{\mathsf{T}}Y = YAY = Y\,,$$

wobei wir in den letzten beiden Schritten (10.5.10) und (10.5.8) angewendet haben. Führt man die letzte Gleichung mit (10.5.11) zusammen, so ergibt sich $X = Y$.

Pseudo-Inverse von Diagonalmatrizen

Ist $D \in \mathcal{M}(r,r)$ eine Diagonalmatrix deren Diagonaleinträge alle ungleich 0 sind, so hat

$$\Sigma := \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}(m,n)\,,$$

die Moore-Penrose Pseudo-Inverse

$$\Sigma^\dagger := \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}(n,m)\,.$$

Dies lässt sich mit Hilfe der Identitäten

$$\Sigma \cdot \Sigma^\dagger = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}(m,m)$$

und

$$\Sigma^\dagger \cdot \Sigma = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}(n,n)$$

leicht überprüfen. Durch die Singulärwert-Zerlegung erhalten wir dann den folgenden Satz.

Theorem 10.5.11

Für jedes $A \in \mathcal{M}(m,n)$ existiert die Moore-Penrose Pseudo-Inverse $A^\dagger$. Ist ferner

$$A = U\Sigma\overline{V}^\mathsf{T}$$

eine Singulärwert-Zerlegung von A, so gilt

$$A^\dagger = V\Sigma^\dagger\overline{U}^\mathsf{T}\,.$$

Schließlich ist $AA^\dagger : \mathbb{K}^m \to \mathbb{K}^m$ die Orthogonalprojektion auf $\operatorname{ran} A$.

BEWEIS 1/2

Für die Existenz von $A^\dagger$ reicht es zu zeigen, dass $X := V\Sigma^\dagger \overline{U}^\mathsf{T}$ die Eigenschaften (10.5.7) – (10.5.10) erfüllt. Dies geschieht durch elementares Nachrechnen, wobei wir zunächst bemerken, dass

$$AX = U\Sigma\overline{V}^\mathsf{T} V\Sigma^\dagger \overline{U}^\mathsf{T} = U\Sigma\Sigma^\dagger \overline{U}^\mathsf{T} \tag{10.5.12}$$

und

$$XA = V\Sigma^\dagger \overline{U}^\mathsf{T} U\Sigma\overline{V}^\mathsf{T} = V\Sigma^\dagger \Sigma\overline{V}^\mathsf{T}$$

gilt. Dies ergibt zum Beispiel

$$AXA = U\Sigma\Sigma^\dagger \overline{U}^\mathsf{T} U\Sigma\overline{V}^\mathsf{T} = U\Sigma\Sigma^\dagger \Sigma\overline{V}^\mathsf{T} = U\Sigma\overline{V}^\mathsf{T} = A$$

und $XAX = X$ kann analog gezeigt werden. Ferner gilt

$$\begin{aligned}\overline{(AX)}^\mathsf{T} = \overline{X}^\mathsf{T} \cdot \overline{A}^\mathsf{T} = U\overline{\Sigma^\dagger}^\mathsf{T}\overline{V}^\mathsf{T} V\overline{\Sigma}^\mathsf{T}\overline{U}^\mathsf{T} = U\overline{\Sigma^\dagger}^\mathsf{T}\overline{\Sigma}^\mathsf{T}\overline{U}^\mathsf{T} &= U\overline{(\Sigma\Sigma^\dagger)}^\mathsf{T}\overline{U}^\mathsf{T} \\ &= U\Sigma\Sigma^\dagger\overline{U}^\mathsf{T} \\ &= AX\,.\end{aligned}$$

Die Gleichung (10.5.10) kann wieder analog gezeigt werden.

BEWEIS 2/2

Um nun $AA^\dagger$ zu untersuchen, sei $r := \mathbf{rang}\, A$. Mit (10.5.12) und den Identitäten für Diagonalmatrizen gilt dann

$$AA^\dagger = U\Sigma\Sigma^\dagger\overline{U}^\mathsf{T} = U\begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}\overline{U}^\mathsf{T} = U_1\overline{U_1}^\mathsf{T}$$

wobei $U_1 \in \mathcal{M}(m, r)$ aus den ersten r Spalten von U besteht. Da U orthogonal ist, bilden die Spalten von U_1 ein Orthonormalsystem, das wegen (10.5.4) in $\mathbf{ran}\, A$ liegt. Da $\mathbf{dim}\, \mathbf{ran}\, A = \mathbf{rang}\, A = r$, sind die Spalten sogar eine Orthonormalbasis von $\mathbf{ran}\, S$. Satz 10.3.3 zeigt dann, dass $AA^\dagger : \mathbb{K}^m \to \mathbb{K}^m$ die Orthogonalprojektion auf $\mathbf{ran}\, A$ ist.

Die Moore-Penrose Pseudo-Inverse erfüllt ein Reihe von Rechenregeln. So gilt zum Beispiel

$$(A^\dagger)^\dagger = A \qquad \text{und} \qquad (\overline{A}^\mathsf{T})^\dagger = \overline{(A^\dagger)}^\mathsf{T},$$

sowie

$$(\overline{A}^\mathsf{T} A)^\dagger = A^\dagger \overline{A^\dagger}^\mathsf{T} \qquad \text{und} \qquad (\lambda A)^\dagger = \lambda^{-1} A^\dagger$$

für $\lambda \neq 0$. Diesen Identitäten lassen sich ebenfalls durch Verifizieren von (10.5.7) – (10.5.10) beweisen. Sind die Spalten von A linear unabhängig, so ist $\overline{A}^\mathsf{T} A$ invertierbar und es gilt

$$A^\dagger = \left(\overline{A}^\mathsf{T} A\right)^{-1} \overline{A}^\mathsf{T}.$$

Im Fall von unabhängigen Zeilen gilt analog

$$A^\dagger = \overline{A}^\mathsf{T} \left(A \overline{A}^\mathsf{T}\right)^{-1}.$$

Schließlich kann man

$$A^\dagger = \lim_{\delta \searrow 0} \left(\overline{A}^\mathsf{T} A + \delta E_n\right)^{-1} \cdot \overline{A}^\mathsf{T} = \lim_{\delta \searrow 0} \overline{A}^\mathsf{T} \left(A \overline{A}^\mathsf{T} + \delta E_m\right)^{-1}$$

zeigen.

Mit Hilfe der Moore-Penrose Pseudo-Inversen lassen sich die Lösungen von linearen Gleichungssystemen leicht beschreiben. Der folgende Satz zeigt dies.

Theorem 10.5.12

Sei $A \in \mathcal{M}(m,n)$ und $b \in \mathbb{K}^m$. Hat die Gleichung $Ax = b$ eine Lösung, so gilt

$$\{x \in \mathbb{K}^n : Ax = b\} = \left\{A^\dagger b + (E_n - A^\dagger A)y : y \in \mathbb{K}^n\right\}.$$

Hat die Gleichung $Ax = b$ keine Lösung, so gilt für $x^* := A^\dagger b$

$$0 < \|Ax^* - b\| \leq \|Ax - b\|, \qquad x \in \mathbb{K}^n.$$

BEWEIS

Wir nehmen zunächst an, dass es ein $x_0 \in \mathbb{K}^n$ mit $Ax_0 = b$ gibt. Sei nun $y \in \mathbb{K}^n$ und $x := A^\dagger b + (E_n - A^\dagger A)y$. Dann gilt

$$\begin{aligned} Ax = A\big(A^\dagger b + (E_n - A^\dagger A)y\big) = AA^\dagger b + Ay - AA^\dagger Ay &= AA^\dagger Ax_0 \\ &= Ax_0 \\ &= b\,. \end{aligned}$$

Dies zeigt die Inklusion "$\supset$".

Ist umgekehrt $x \in \mathbb{K}^n$ mit $Ax = b$. Für $y := x \in \mathbb{K}^n$ gilt dann

$$A^\dagger b + (E_n - A^\dagger A)y = A^\dagger b + x - A^\dagger Ax = A^\dagger b + x - A^\dagger b = x$$

und damit haben wir auch die Inklusion "$\subset$" gezeigt.

Nach Satz 10.5.11 ist $AA^\dagger : \mathbb{K}^m \to \mathbb{K}^m$ die Orthogonalprojektion auf $\mathbf{ran}\, A$ ist. Für $x \in \mathbb{K}^n$ erhalten wir dann

$$\|Ax^* - b\| = \|AA^\dagger - b\| \leq \|Ax - b\|$$

mit Satz 10.3.2.

Section 10.6
Quadratische Formen

Im folgenden wollen wir die Lösungsmengen mehrdimensionaler, quadratischer Polynome untersuchen, wobei wir uns auf den reellen Fall konzentrieren. Wir beginnen mit der folgenden Definition.

Definition 10.6.1

Für $A \in \mathcal{M}(n,n)$, $b \in \mathbb{R}^n$ und $c \in \mathbb{R}$ sei

$$q(x) := \langle x, Ax \rangle + \langle b, x \rangle + c\,, \qquad x \in \mathbb{R}^n.$$

Dann heißt

$$\left\{ x \in \mathbb{R}^n : q(x) = 0 \right\}$$

Quadrik.

Das folgende Lemma zeigt, dass wir ohne Einschränkung der Allgemeinheit immer annehmen können, dass die Matrix A einer Quadrik symmetrisch ist.

Lemma 10.6.2

Sei $A \in \mathcal{M}(n,n)$. Dann ist $B := (A + A^{\mathsf{T}})/2$ symmetrisch und es gilt

$$\langle x, Ax \rangle = \langle x, Bx \rangle \,, \qquad\qquad x \in \mathbb{R}^n.$$

Wegen Lemma 9.3.5, $(A^\mathsf{T})^\mathsf{T} = A$ und der Symmetrie des Skalarprodukts gilt

$$\langle x, Bx\rangle = \frac{1}{2}\langle x, Ax\rangle + \frac{1}{2}\langle x, A^\mathsf{T}x\rangle = \frac{1}{2}\langle x, Ax\rangle + \frac{1}{2}\langle Ax, x\rangle = \langle x, Ax\rangle$$

für alle $x \in \mathbb{R}^n$. Ferner ist B wegen $B^\mathsf{T} = (A + A^\mathsf{T})^\mathsf{T}/2 = (A^\mathsf{T} + A)/2 = B$ symmetrisch.

Wir wollen uns als nächstes einige Beispiele von Quadriken im $\mathbb{R}^2$ anschauen. Wir beginnen mit

$$q(x) := \langle x, E_2 x \rangle - 1 = x_1^2 + x_2^2 - 1\,, \qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\,.$$

Die resultierende Quadrik ist die **Einheitssphäre**:

$$\left\{x \in \mathbb{R}^2 : q(x) = 0\right\} = \left\{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\right\}.$$

Betrachten wir stattdessen die Matrix $A := \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ und das quadratische Polynom

$$q(x) := \langle x, Ax \rangle - 1 = x_1^2 + 2x_2^2 - 1\,, \qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\,,$$

so ist die resultierende Quadrik eine **Ellipse**:

$$\left\{x \in \mathbb{R}^2 : q(x) = 0\right\} = \left\{x \in \mathbb{R}^2 : x_1^2 + 2x_2^2 = 1\right\}.$$

Betrachten wir das quadratische Polynom zu der Matrix $A := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, dem Vektor $b := \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ und der Konstante $c := 0$, d.h.

$$q(x) := \langle x, Ax \rangle + \langle b, x \rangle + c = x_1^2 - x_2\,, \qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\,,$$

so ist die resultierende Quadrik eine **Parabel**:

$$\left\{x \in \mathbb{R}^2 : q(x) = 0\right\} = \left\{x \in \mathbb{R}^2 : x_2 = x_1^2\right\}.$$

Betrachten wir nun die Matrix $A := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ und das quadratische Polynom

$$q(x) := \langle x, Ax \rangle - 1 = x_1^2 - x_2^2 - 1\,, \qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\,,$$

so ist die resultierende Quadrik eine **Hyperbel**:

$$\left\{x \in \mathbb{R}^2 : q(x) = 0\right\} = \left\{x \in \mathbb{R}^2 : x_1^2 - x_2^2 = 1\right\}.$$

Ersetzen wir in dem letzten Beispiel die Konstante -1 durch 0, so ergibt sich das quadratische Polynom

$$q(x) := \langle x, Ax \rangle = x_1^2 - x_2^2\,, \qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\,,$$

mit der zugehörigen Quadrik

$$\{x \in \mathbb{R}^2 : q(x) = 0\} = \{x \in \mathbb{R}^2 : x_1^2 = x_2^2\} = \{x \in \mathbb{R}^2 : |x_1| = |x_2|\}\,,$$

die aus den beiden Diagonalen besteht.

Als nächstes Beispiel betrachten wir die Matrix $A := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ und das quadratische Polynom

$$q(x) := \langle x, Ax \rangle - 1 = x_2^2 - 1\,, \qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\,.$$

Die zugehörige Quadrik ist dann

$$\{x \in \mathbb{R}^2 : q(x) = 0\} = \{x \in \mathbb{R}^2 : x_2^2 = 1\} = \{x \in \mathbb{R}^2 : |x_2| = 1\}\,,$$

die aus zwei waagerechten Geraden besteht.

Quadriken können aber auch leer sein. Ersetzen wir in dem letzten Beispiel die Konstante -1 durch 1, d.h. wir betrachten das quadratische Polynom

$$q(x) := \langle x, Ax \rangle + 1 = x_2^2 + 1\,, \qquad\qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\,,$$

so gilt für die Quadrik

$$\left\{x \in \mathbb{R}^2 : q(x) = 0\right\} = \left\{x \in \mathbb{R}^2 : x_2^2 = -1\right\} = \emptyset\,.$$

Die bisherigen Beispiele hatten immer Diagonalmatrizen A betrachtet, und das resultierende quadratische Polynom sah deswegen sehr einfach aus. Wir können aber auch kompliziertere Polynome betrachten, z.B.

$$q(x) := 7x_1^2 + 6x_2^2 + 5x_3^2 - 4x_1x_2 - 4x_2x_3 + 14x_1 - 8x_2 + 10x_3 + 6$$

für $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$. Setzen wir

$$A := \begin{pmatrix} 7 & -2 & 0 \\ -2 & 6 & -2 \\ 0 & -2 & 5 \end{pmatrix}, \qquad \text{und} \qquad b := \begin{pmatrix} 14 \\ -8 \\ 10 \end{pmatrix}$$

sowie $c := 6$, so ergibt sich

$$q(x) = \langle x, Ax \rangle + \langle b, x \rangle + c\,, \qquad x \in \mathbb{R}^3.$$

Diese Darstellungen von q, sowie die Quadrik von q ist recht unübersichtlich. Im folgenden werden wir daher die sogenannte Hauptachsentransformation beschreiben, die einen Basiswechsel durchführt, um die Darstellungen zu vereinfachen.

HAUPTACHSENTRANSFORMATION: 1. SCHRITT

Im folgenden betrachten wir dazu ein allgemeines quadratisches Polynom q, d.h. wir haben ein $A \in \mathcal{M}(n,n)$, ein $b \in \mathbb{R}^n$ und $c \in \mathbb{R}$, so dass

$$q(x) = \langle x, Ax \rangle + \langle b, x \rangle + c\,, \qquad x \in \mathbb{R}^n$$

gilt. Gemäß Lemma 10.6.2 nehmen wir dabei zusätzlich an, dass A symmetrisch ist.

Der erste Schritt der Hauptachsentransformation besteht aus der Diagonalisierung der Matrix A. Nach Satz 10.5.3 existieren hierzu Eigenwerte $\lambda_1, \ldots, \lambda_n$ von A und eine Orthonormalbasis $\mathcal{B}$ aus zugehörigen Eigenvektoren $v_1, \ldots, v_n$, d.h. $Av_i = \lambda_i v_i$ für alle $i = 1, \ldots, n$. Setzen wir $S := (v_1, \ldots, v_n) \in \mathcal{M}(n,n)$ und

$$\Lambda := \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

so gilt mit (10.4.4)

$$\Lambda = S^{-1}AS = S^{\mathsf{T}}AS\,,$$

wobei wir im letzten Schritt Satz 10.3.8 und (10.3.9) ausgenutzt haben. Für $x \in \mathbb{R}^n$ setzen wir nun $y := S^{-1}x = S^{\mathsf{T}}x$, d.h. $x = Sy$. Dies ergibt

$$\begin{aligned} q(x) = q(Sy) &= \langle Sy, ASy\rangle + \langle b, Sy\rangle + c \\ &= \langle y, S^{\mathsf{T}}ASy\rangle + \langle S^{\mathsf{T}}b, y\rangle + c \\ &= \langle y, \Lambda y\rangle + \langle S^{\mathsf{T}}b, y\rangle + c\,. \end{aligned}$$

Für $\tilde{b} := S^{\mathsf{T}}b$ und

$$q_1(y) := \langle y, \Lambda y\rangle + \langle \tilde{b}, y\rangle + c = \sum_{i=1}^{n} \lambda_i y_i^2 + \sum_{i=1}^{n} \tilde{b}_i y_i + c$$

gilt damit

$$q(x) = q_1(S^{-1}x) = q_1(S^{\mathsf{T}}x)\,, \qquad x \in \mathbb{R}^n, \tag{10.6.1}$$

wobei das neue Polynom q_1 aufgrund der Diagonalform von Λ eine deutlich einfachere Gestalt hat.

Wegen $S^{\mathsf{T}} = S^{-1} = \mathrm{T}_{\mathcal{B}}^{\mathcal{E}_n}$ zeigt dann die Formel (10.6.1), dass durch einem Basiswechsel von $\mathcal{E}_n$ nach $\mathcal{B}$, dass Polynom q durch q_1 ersetzt wird. Nach Lemma 10.3.9 haben wir dabei $\det S^{\mathsf{T}} = \pm 1$. Im Fall $\det S^{\mathsf{T}} = -1$ können wir alternativ aber auch die Orthonormalbasis $\mathcal{B}' := (-v_1, v_2, \ldots, v_n)$ betrachten. Für die resultierende Matrix $S' := (-v_1, v_2, \ldots, v_n)$ gilt dann $\det(S')^{\mathsf{T}} = 1$, so dass wir ohne Einschränkung annehmen können, dass schon unsere Matrix S die Gleichung $\det S^{\mathsf{T}} = 1$ erfüllt. Der Basiswechsel $\mathcal{E}_n$ nach $\mathcal{B}$ entspricht somit einer Drehung.

HAUPTACHSENTRANSFORMATION: 2. SCHRITT

Um q_1 weiter zu vereinfachen, nehmen wir nun an, dass die Eigenwerte so sortiert sind, dass $\lambda_1, \dots, \lambda_m \neq 0$ und $\lambda_{m+1}, \dots, \lambda_n = 0$ gilt. Wir werden nun eine quadratische Ergänzung vornehmen. Für $i = 1, \dots, m$ basiert dies auf der Rechnung

$$\begin{aligned}\lambda_i y_i^2 + \tilde{b}_i y_i = \lambda_i \left(y_i^2 + \frac{\tilde{b}_i y_i}{\lambda_i} \right) &= \lambda_i \left(\left(y_i + \frac{\tilde{b}_i}{2\lambda_i} \right)^2 - \frac{\tilde{b}_i^2}{4\lambda_i^2} \right) \\ &= \lambda_i \left(y_i + \frac{\tilde{b}_i}{2\lambda_i} \right)^2 - \frac{\tilde{b}_i^2}{4\lambda_i} \,.\end{aligned}$$

Für $y \in \mathbb{R}^n$ definieren wir nun $Ty := z$ durch

$$z_i := \begin{cases} y_i + \frac{\tilde{b}_i}{2\lambda_i} & \text{für } i = 1, \dots, m \,, \\ y_i & \text{für } i = m+1, \dots, n \,. \end{cases} \tag{10.6.2}$$

Mit anderen Worten verschieben wir y in den ersten m Koordinaten. Setzen wir ferner

$$\tilde{c} := c - \sum_{i=1}^{m} \frac{\tilde{b}_i^2}{4\lambda_i} \,,$$

und

$$q_2(z) := \sum_{i=1}^{m} \lambda_i z_i^2 + \sum_{i=m+1}^{n} \tilde{b}_i z_i + \tilde{c}\,, \qquad z \in \mathbb{R}^n,$$

so folgt für $y \in \mathbb{R}^n$ und $z := Ty$:

$$\begin{aligned}
q_2(Ty) = q_2(z) &= \sum_{i=1}^{m} \lambda_i z_i^2 + \sum_{i=m+1}^{n} \tilde{b}_i z_i + \tilde{c} \\
&= \sum_{i=1}^{m} \lambda_i \left(y_i + \frac{\tilde{b}_i}{2\lambda_i} \right)^2 + \sum_{i=m+1}^{n} \tilde{b}_i y_i + c - \sum_{i=1}^{m} \frac{\tilde{b}_i^2}{4\lambda_i} \\
&= \sum_{i=1}^{m} \left(\lambda_i y_i^2 + \tilde{b}_i y_i + \frac{\tilde{b}_i^2}{4\lambda_i} \right) + \sum_{i=m+1}^{n} \tilde{b}_i y_i + c - \sum_{i=1}^{m} \frac{\tilde{b}_i^2}{4\lambda_i} \\
&= \sum_{i=1}^{m} \lambda_i y_i^2 + \sum_{i=1}^{n} \tilde{b}_i y_i + c \\
&= \sum_{i=1}^{n} \lambda_i y_i^2 + \sum_{i=1}^{n} \tilde{b}_i y_i + c \\
&= q_1(y)
\end{aligned}$$

wobei wir im vorletzten Schritt $\lambda_{m+1}, \ldots, \lambda_n = 0$ ausgenutzt haben.

HAUPTACHSENTRANSFORMATION: 2. SCHRITT

Damit haben wir also

$$q(x) = q_1(S^\mathsf{T} x) = q_2\big(T(S^\mathsf{T} x)\big)\,, \qquad\qquad x \in \mathbb{R}^n.$$

Durch die *bijektive* Transformation $x \mapsto S^\mathsf{T} x \mapsto T(S^\mathsf{T} x)$ können wir also q durch das vereinfachte Polynom q_2 ausdrücken.

HAUPTACHSENTRANSFORMATION: 3. SCHRITT

Im letzten Schritt vereinfachen wir noch den affin linearen Term $\sum_{i=m+1}^{n} \tilde{b}_i z_i + \tilde{c}$ von q_2. Ist bei diesem $\tilde{b}_{m+1}, \ldots, \tilde{b}_n = 0$, so kann die Summe offensichtlich einfach weggelassen werden, d.h. wir haben

$$q_2(z) = \sum_{i=1}^{m} \lambda_i z_i^2 + \tilde{c}\,, \qquad\qquad z \in \mathbb{R}^n.$$

Ansonsten gibt es ein $j \in \{m+1, \ldots, n\}$ mit $\tilde{b}_j \neq 0$. Wir setzen $\xi := Rz$ mit

$$\xi_i := \begin{cases} z_i & \text{für } i \neq j\,, \\ z_j + \frac{\tilde{c}}{\tilde{b}_j} & \text{für } i = j\,. \end{cases}$$

Damit ist R, wie schon T eine Verschiebung um einen festen Vektor. Für

$$q_3(\xi) := \sum_{i=1}^{m} \lambda_i \xi_i^2 + \sum_{i=m+1}^{n} \tilde{b}_i \xi_i\,, \qquad\qquad \xi \in \mathbb{R}^n,$$

ergibt dies

$$q_3(Rz) = q_3(\xi) = \sum_{i=1}^{m} \lambda_i \xi_i^2 + \sum_{i=m+1}^{n} \tilde{b}_i \xi_i = \sum_{i=1}^{m} \lambda_i z_i^2 + \sum_{i=m+1}^{n} \tilde{b}_i z_i + \tilde{c} = q_2(z)$$

für alle $z \in \mathbb{R}^n$.

HAUPTACHSENTRANSFORMATION: 3. SCHRITT

Kombiniert man diese Identität wieder mit unseren vorherigen Überlegungen, so erhalten wir

$$q(x) = q_2\big(T(S^\mathsf{T}x)\big) = q_3\big((R \circ T)(S^\mathsf{T}x)\big)\,, \qquad x \in \mathbb{R}^n.$$

Hierbei ist $R \circ T$ wieder eine Verschiebung, so dass die bijektive Transformation $R \circ T \circ S^\mathsf{T}$ wieder aus einer Drehung und einer anschließenden Verschiebung besteht.

Ist p das aus der Hauptachsentransformation resultierende Polynom, d.h. $p = q_2$ oder $p = q_3$, und $V : \mathbb{R}^n \to \mathbb{R}^n$ die zugehörige Transformation, d.h. $V = T \circ S^{\mathsf{T}}$ bzw. $V = R \circ T \circ S^{\mathsf{T}}$, dann gilt mit $z := Vx$, d.h. $x = V^{-1}x$

$$\begin{aligned}\{x \in \mathbb{R}^n : q(x) = 0\} &= \{x \in \mathbb{R}^n : p(Vx) = 0\} \\ &= \{V^{-1}z : z \in \mathbb{R}^n \text{ mit } p(z) = 0\} \\ &= V^{-1}\big(\{z \in \mathbb{R}^n : p(z) = 0\}\big)\end{aligned}$$

Mit anderen Worten reicht es aus, die Quadriken zu verstehen, die zu den Polynomen p gehören, die aus einer Hauptachsentransformation resultieren.

Die 2, bzw. 3 Schritte der Hauptachsentransformation, führen nun auf die folgenden 3 Arten von Polynomen, die als **Normalformen** bezeichnet werden:

(A) Der Fall $m = n$, d.h. $\lambda_1, \ldots, \lambda_n \neq 0$. In diesem Fall führen wir nur die ersten beiden Schritte aus und das vereinfachte Polynom ist

$$p(z) = q_2(z) = \sum_{i=1}^{n} \lambda_i z_i^2 + \tilde{c}\,.$$

(B) Der Fall $m < n$ und $\tilde{b}_{m+1} = \cdots = \tilde{b}_n = 0$. In diesem Fall führen wir wieder nur die ersten beiden Schritte aus und das vereinfachte Polynom ist

$$p(z) = q_2(z) = \sum_{i=1}^{m} \lambda_i z_i^2 + \tilde{c}\,.$$

(C) Der Fall $m < n$ und $\tilde{b}_j \neq 0$ für ein $j \in \{m+1, \ldots, n\}$. In diesem Fall führen wir alle 3 Schritte aus und das vereinfachte Polynom ist

$$p(z) = q_3(z) = \sum_{i=1}^{m} \lambda_i z_i^2 + \sum_{i=m+1}^{n} \tilde{b}_i z_i\,.$$

HAUPTACHSENTRANSFORMATION: BEISPIEL

Im Fall $n = 3$ betrachten wir

$$A := \begin{pmatrix} 5 & -2 & -4 \\ -2 & 8 & -2 \\ -4 & -2 & 5 \end{pmatrix}, \qquad \text{und} \qquad b := \begin{pmatrix} 16 \\ 8 \\ -2 \end{pmatrix}$$

sowie $c := 19$. Damit haben wir das Polynom

$$\begin{aligned} q(x) &= \langle x, Ax \rangle + \langle b, x \rangle + c \\ &= 5x_1^2 + 8x_2^2 + 5x_3^2 - 4x_1x_2 - 8x_1x_3 - 4x_2x_3 + 16x_1 + 8x_2 - 2x_3 + 19\,. \end{aligned}$$

Um den ersten Schritt der Hauptachsentransformation durchzuführen, müssen wir zunächst die Eigenwerte und eine zugehörige Orthonormalbasis von Eigenvektoren bestimmen. Dazu berechnen wir zunächst das charakteristische Polynom

$$p_A(\lambda) = \mathbf{det}(A - \lambda E_3) = -\lambda(\lambda - 9)^2\,.$$

Die Eigenwerte sind demnach $\lambda_1 := \lambda_2 := 9$ und $\lambda_3 := 0$ und wir haben $m = 2$.

HAUPTACHSENTRANSFORMATION: BEISPIEL

Um Eigenvektoren von λ_1 und λ_2 zu bestimmen, bemerken wir zunächst, dass $(A - 9E_n)x = 0$ dem Gleichungssystem

$$\begin{pmatrix} -4 & -2 & -4 \\ -2 & -1 & -2 \\ -4 & -2 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

entspricht. Dies führt zu der Gleichung $2x_1 + x_2 + 2x_3 = 0$, für die

$$v_1 := \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \qquad \text{und} \qquad v_2 := \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix}$$

zwei linear unabhängige Lösungen sind.

HAUPTACHSENTRANSFORMATION: BEISPIEL

Um eine Orthonormalbasis von diesem Unterraum zu konstruieren, benutzen wir das Gram-Schmidt-Verfahren. Dies ergibt

$$w_1 := \frac{1}{\|v\|_1} v_1 = \frac{1}{3} v_1$$

und wegen $\langle v_2, w_1 \rangle = 1$ auch

$$u_2 := v_2 - \langle v_2, w_1 \rangle w_1 = \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix} - w_1 = \frac{1}{3} \left(\begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \right) = \frac{2}{3} \begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix} .$$

Wegen $\|u_2\| = 2$ folgt dann

$$w_2 := \frac{u_2}{\|u_2\|} = \frac{1}{3} \begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix} .$$

Analog ergibt das Lösen des Gleichungssystems $Ax = (A - \lambda_3 E_3)x = 0$ zum Beispiel

$$w_3 := \frac{1}{3} \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} ,$$

wobei wir $\|w_3\| = 1$ haben.

Nach Satz 10.5.2 sind dann (w_1, w_2, w_3) eine Orthonormalbasis, die nach unserer Konstruktion aus Eigenvektoren besteht. Setzen wir

$$S := \frac{1}{3}\begin{pmatrix} 1 & -2 & 2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{pmatrix} \qquad \text{und} \qquad \Lambda := \begin{pmatrix} 9 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

sowie

$$\tilde{b} = S^{\mathsf{T}} b = \begin{pmatrix} 24 \\ -12 \\ 0 \end{pmatrix},$$

so ergibt sich $\det S = 1$ und

$$q_1(y) := \langle y, \Lambda y \rangle + \langle \tilde{b}, y \rangle = 9y_1^2 + 9y_2^2 + 24y_1 - 12y_2 + 19$$

für $y \in \mathbb{R}^3$.

Im 2. Schritt der Hauptachsentransformation verschieben wir jetzt gemäß (10.6.2). Dies ergibt

$$z_1 := y_1 + \frac{\tilde{b}_1}{2\lambda_1} = y_1 - \frac{3}{4}\,, \qquad \text{und} \qquad z_2 := y_2 + \frac{\tilde{b}_2}{2\lambda_2} = y_2 - \frac{2}{3}\,,$$

sowie $z_3 := y_3$. Ferner gilt

$$\tilde{c} := c - \frac{\tilde{b}_1^2}{4\lambda_1} - \frac{\tilde{b}_2^2}{4\lambda_2} = 19 - 16 - 4 = -1\,,$$

so dass wir

$$q_2(z) := 9z_1^2 + 9z_2^2 - 1\,, \qquad z \in \mathbb{R}^3$$

erhalten. Da wir $m = 2$ und $\tilde{b}_{m+1} = \tilde{b}_3 = 0$ haben, entfällt der 3. Schritt und unser Polynom $p := q_2$ ist von der Form (B).

HAUPTACHSENTRANSFORMATION: BEISPIEL

Die Quadrik

$$Q := \{x \in \mathbb{R}^3 : q_2(x) = 0\} = \left\{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = 1\right\}$$

ist in jeder Ebene $\mathcal{E}_{x_3} := \{(x_1, x_2, x_3)^\mathsf{T} : x_1, x_2 \in \mathbb{R}\}$ für $x_3 \in \mathbb{R}$ einen Kreis mit Mittelpunkt 0 und Radius 1/3 als Lösungsmenge. Da $\mathcal{E}_{x_3}$ parallel zur (x_1, x_2)-Ebene ist, besteht damit Q aus einem Zylinder, für den die x_3-Achse die Mittelpunkt-Achse bildet, und der den Radius 1/3 hat.

Die Transformation $V^{-1} = (S^\mathsf{T})^{-1} \circ T^{-1} = S \circ T^{-1}$ zeigt dann, dass die Quadrik von unserem ursprünglichen Polynom q durch Verschiebung T^{-1} und Drehung S aus dem Zylinder Q entsteht.

EINIGE NORMALFORMEN

Im Folgenden wollen wir noch einige Normalformen und ihre Quadriken in den Räumen $\mathbb{R}^2$ und $\mathbb{R}^3$ auflisten.

Wir beginnen mit den Fällen für $n = m = 2$, wobei wir die Koordinaten mit x und y statt mit x_1 und x_2 bezeichnen und $a, b \neq 0$ sind:

Ellipse: $$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0\,,$$

Hyperbel: $$\frac{x^2}{a^2} - \frac{y^2}{b^2} - 1 = 0\,,$$

Leere Menge: $$\frac{x^2}{a^2} + \frac{y^2}{b^2} + 1 = 0\,,$$

Punkt 0: $$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 0\,,$$

Geradenpaar: $$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0\,.$$

Hierbei sind die obigen Hyperbeln in der sogenannten 1. Hauptlage. Vertauscht man x und y, so erhält man die 2. Hauptlage.

Für $n = 2$ und $m = 1$ haben wir für $a \neq 0$ und $b \neq 0$:

Parabel:	$x^2 - 2py = 0\,,$
Parallele Geraden:	$x^2 - a^2 = 0\,,$
Leere Menge:	$x^2 + a^2 = 0$

und entsprechendes gilt beim Vertauschen der Rollen von x und y.

Für $n = 2$ und $m = 0$ haben wir für $b_1, b_2, c \in \mathbb{R}$ mit $|b_1| + |b_2| > 0$:

Gerade:	$b_1x + b_2y + c = 0\,,$	
Leere Menge:	$c = 0\,,$	falls $c \neq 0$,
Ganzer Raum:	$c = 0\,,$	falls $c = 0$.

Im Fall $n = 3$ gibt es naturgemäß mehr Fälle als im $\mathbb{R}^2$. Im folgenden verzichten wir hierbei auf mögliche Skalierungen von x, y und z:

Ellipsoid:	$x^2 + y^2 + z^2 - 1 = 0\,,$
einschaliges Hyperbolid:	$x^2 + y^2 - z^2 - 1 = 0\,,$
zweischaliges Hyperbolid:	$x^2 + y^2 - z^2 + 1 = 0\,,$
Kegel:	$x^2 + y^2 - z^2 = 0\,,$
elliptisches Paraboloid:	$x^2 + y^2 - z = 0\,,$
hyperbolisches Paraboloid:	$x^2 - y^2 - z = 0\,.$

SCHNITTE VON QUADRIKEN UND EBENEN

Ist $Q \subset \mathbb{R}^n$ eine Quadrik, und $\mathcal{E} \subset \mathbb{R}^n$ eine Ebene, so kann

$$Q \cap \mathcal{E}$$

wieder als eine Quadrik im $\mathbb{R}^{n-1}$ aufgefasst werden. Um dies zu sehen, lösen wir die Ebenen-Gleichung $\langle w, x \rangle + c = 0$ nach einer Variablen auf, z.B.

$$x_n = \langle \hat{w}, (x_1, \ldots, x_{n-1})^T \rangle + \hat{c}\,,$$

wenn $w_n \neq 0$ gilt. Einsetzen von x_n in das quadratische Polynom q der Quadrik Q ergibt ein neues quadratisches Polynom in den Variablen $x_1, \ldots, x_{n-1}$.

Ein Beispiel für diese Beobachtung sind Kegelschnitte.

APTER 11: MEHRDIMENSIONALE DIFFERENTIALRECHN

EINFÜHRUNG

In diesem Kapitel werden wir Ableitungen für Funktionen $f : \mathbb{R}^n \to \mathbb{R}^m$ untersuchen. Hierbei werden wir einerseits verschiedene Ableitungsbegriffe kennenlernen und andererseits wichtige Eigenschaften differenzierbarer Funktionen betrachten.

Section 11.1
Partielle Ableitungen

Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$ lässt sich durch m Komponentenfunktionen $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ darstellen, d.h.

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}, \qquad x \in \mathbb{R}^n.$$

Wir betrachten daher zunächst den Fall $m = 1$, d.h. Abbildungen

$$\begin{aligned} f : \mathbb{R}^n &\to \mathbb{R} \\ (x_1, \ldots, x_n)^\top &\mapsto f(x_1, \ldots, x_n)\,. \end{aligned}$$

Fixieren wir $n - 1$ Variablen, so erhalten wir eine Funktion $\mathbb{R} \to \mathbb{R}$ für die wir schon einen Ableitungsbegriff haben. Dies führt zu der folgenden Definition.

Definition 11.1.1

Sei $f : \mathbb{R}^n \to \mathbb{R}$ und $j \in \{1, \ldots, n\}$. Dann heißt f in $x \in \mathbb{R}^n$ **partiell in der j-ten Koordinate differenzierbar**, falls

$$\frac{\partial f(x)}{\partial x_j} := \lim_{h \to 0} \frac{f(x + he_j) - f(x)}{h}$$

existiert. In diesem Fall heißt $\frac{\partial f(x)}{\partial x_j}$ die **partielle Ableitung von f in x in Richtung x_j**.

In der Literatur gibt es diverse weitere Schreibweisen für partielle Ableitungen, wie z.B.

$$\frac{\partial f}{\partial x_j}(x) := \frac{\partial}{\partial x_j} f(x) := \partial_{x_j} f(x) := D_j f(x) := \frac{\partial f(x)}{\partial x_j}\,.$$

Analog kann die partielle Ableitung für $f : U \to \mathbb{R}$ mit $U \subset \mathbb{R}^n$ **offen** definiert werden. Hierbei ist eine Menge $U \subset \mathbb{R}^n$ offen, falls es zu jedem $x \in U$ ein $\delta > 0$ mit $U(x, \delta) \subset U$ gibt. Offene Kugeln sind Beispiele offener Mengen und $\mathbb{R}^n$ ist auch offen.

Eine Funktion $f : U \to \mathbb{R}$ mit $U \subset \mathbb{R}^n$ ist **partiell differenzierbar**, falls sie partiell differenzierbar für alle $x \in U$ und $j \in \{1, \ldots, n\}$ ist. Sind in diesem Fall alle partiellen Ableitungen stetig, so heißt f **stetig partiell differenzierbar**.

Die Rechenregeln für Ableitungen aus Kapitel 6 übertragen sich direkt auf partielle Ableitungen. So gilt beispielsweise

$$\frac{\partial(\alpha f + \beta g)}{\partial x_j}(x) = \alpha \frac{\partial f}{\partial x_j}(x) + \beta \frac{\partial g}{\partial x_j}(x)\,,$$

$$\frac{\partial fg}{\partial x_j}(x) = f(x)\frac{\partial g}{\partial x_j}(x) + g(x)\frac{\partial f}{\partial x_j}(x)\,,$$

$$\frac{\partial}{\partial x_j}\left(\frac{f}{g}\right)(x) = \frac{1}{(g(x))^2} \cdot \left(g(x)\frac{\partial f}{\partial x_j}(x) - f(x)\frac{\partial g}{\partial x_j}(x)\right),$$

wobei $f, g : U \to \mathbb{R}$ und $\alpha, \beta \in \mathbb{R}$ sind und im dritten Fall zusätzlich $g(x) \neq 0$ gelten muss.

Ist $f : \mathbb{R}^2 \to \mathbb{R}$ beispielsweise durch $f(x) := \|x\|^2 = x_1^2 + x_2^2$ für $x \in \mathbb{R}^2$ definiert, so gilt

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1 \qquad \text{und} \qquad \frac{\partial f}{\partial x_2}(x_1, x_2) = 2x_2 \,.$$

Definition 11.1.2

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ partiell differenzierbar. Dann heißt

$$\operatorname{grad} f(x) := \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right), \qquad x \in U$$

der Gradient von f in x. Ferner schreiben wir

$$\nabla f(x) := \big(\operatorname{grad} f(x)\big)^{\mathsf{T}} = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix},$$

wobei ∇ als Nabla gesprochen wird.

In der Literatur wird der Gradient häufig auch als Spaltenvektor, und nicht wie oben als Zeilenvektor, aufgefasst. Die Abbildung $f \mapsto \nabla f$ wird auch als Nabla-Operator bezeichnet. Man beachte, dass ∇f eine Abbildung $\nabla f : U \to \mathbb{R}^n$ ist.

GRADIENT UND ∇-OPERATOR: BEMERKUNGEN

Offensichtlich ist $\operatorname{grad} f : U \to \mathbb{R}^n$. Ferner übertragen sich obigen Rechenregeln für partielle Ableitungen direkt auf den Gradienten. So gilt beispielsweise

$$\operatorname{grad}(\alpha f + \beta g) = \alpha \operatorname{grad} f + \beta \operatorname{grad} g\,,$$
$$\operatorname{grad}(fg) = f \operatorname{grad} g + g \operatorname{grad} f\,,$$

wobei wieder $f, g : U \to \mathbb{R}$ und $\alpha, \beta \in \mathbb{R}$ angenommen wurde. Analoge Regeln gelten für ∇.

Ist $f(x, y) := \mathrm{e}^x \cdot \sin y$, so gilt

$$\operatorname{grad} f(x, y) = \left(\mathrm{e}^x \sin y, \mathrm{e}^x \cos y\right).$$

Für die Funktion $r : \mathbb{R}^n \to \mathbb{R}$ mit $r(x) := \|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$ gilt mit der eindimensionalen Kettenregel, siehe Satz 6.1.7,

$$\frac{\partial r}{\partial x_i}(x) = \frac{2x_i}{2\sqrt{x_1^2 + \cdots + x_n^2}} = \frac{x_i}{r(x)}\,, \qquad x \in \mathbb{R}^n \setminus \{0\}.$$

Dies ergibt

$$\nabla r(x) = \frac{1}{\|x\|} x\,, \qquad x \in \mathbb{R}^n \setminus \{0\}.$$

Ist nun $h : (0, \infty) \to \mathbb{R}$ eine differenzierbare Funktion, so ist die Funktion $f := h \circ r : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$, d.h. $f(x) := h(\|x\|)$, wegen der eindimensionalen Kettenregel partiell differenzierbar und es gilt

$$\frac{\partial f}{\partial x_i}(x) = \frac{\partial (h \circ r)}{\partial x_i}(x) = h'(r(x)) \frac{\partial r}{\partial x_i}(x) = \frac{h'(\|x\|) x_i}{\|x\|} .$$

Mit anderen Worten haben wir

$$\nabla h(\|\cdot\|)(x) = \nabla f(x) = \frac{h'(\|x\|)}{\|x\|} \cdot x , \qquad x \in \mathbb{R}^n \setminus \{0\}. \tag{11.1.1}$$

Funktionen der Form $x \mapsto h(\|x\|)$ heißen rotationsinvariant.

Am Anfang des Kapitels hatten wir uns zunächst auf den Fall $m = 1$ reduziert. Mit dem bisher eingeführten können wir jetzt auch den Fall $m > 1$ betrachten.

Definition 11.1.3

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^m$. Ferner seien $f_1, \ldots, f_m$ die Komponentenfunktionen von f, d.h.

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}.$$

Sind dann $f_1, \ldots, f_m$ partiell in x differenzierbar, so heißt f in x partiell differenzierbar und

$$Jf(x) := \begin{pmatrix} \operatorname{grad} f_1(x) \\ \vdots \\ \operatorname{grad} f_m(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \ldots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \ldots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix}$$

ist die Jacobi-Matrix von f im Punkt x.

JACOBI-MATRIX: BEMERKUNGEN

Offensichtlich ist die obige Jacobi-Matrix eine $m \times n$-Matrix und wir haben insgesamt eine Abbildung

$$U \to \mathcal{M}(m, n)$$
$$x \mapsto Jf(x)\,.$$

ABLEITUNGEN HÖHERER ORDNUNG

Ist $f : U \to \mathbb{R}$ partiell differenzierbar, so ist

$$g := \frac{\partial f}{\partial x_j} : U \to \mathbb{R}$$

Damit können wir auch partielle Ableitungen von g betrachten, d.h.

$$\frac{\partial^2 f}{\partial x_i \partial x_j} := \frac{\partial g}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j} f\right),$$

wobei wir annehmen, dass die entsprechende partielle Ableitung existiert. Existieren alle partiellen Ableitungen $\frac{\partial^2 f}{\partial x_i \partial x_j}$ für $i, j = 1, \ldots, n$, so heißt f **zweimal partiell differenzierbar** und sind diese partiellen Ableitungen alle stetig, so sagen wir, dass f **zweimal stetig partiell differenzierbar**

Da $\frac{\partial^2 f}{\partial x_i \partial x_j}$ wiederum eine Abbildung $U \to \mathbb{R}$ ist, können wir dieses, die Existenz vorausgesetzt, weiter iterieren, also zum Beispiel die dritte partielle Ableitung

$$\frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k} := \frac{\partial}{\partial x_i}\left(\frac{\partial^2 f}{\partial x_j \partial x_k}\right)$$

betrachten. Wir sprechen dann von **partiellen Ableitungen höherer Ordnung** und die **k-fache (stetige) partielle Differenzierbarkeit** ist auf naheliegenderweise definiert.

ABLEITUNGEN HÖHERER ORDNUNG: BEISPIEL

Im Allgemeinen ist die Reihenfolge der partiellen Ableitungen wichtig. So gilt für die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$, die durch $f(0) := 0$ und

$$f(x) := x_1 \cdot x_2 \cdot \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}, \qquad x \neq 0$$

definiert ist, zum Beispiel

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(0) \neq \frac{\partial^2 f}{\partial x_2 \partial x_1}(0).$$

Der folgende Satz, der als Satz von Schwarz bekannt ist, zeigt, dass dieses Phänomen für zweimal stetig partiell differenzierbare Funktionen nicht auftritt.

Theorem 11.1.4

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ zweimal stetig partiell differenzierbar. Dann gilt für alle $i, j = 1, \ldots, n$:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Der Beweis dieses Satzes ist nicht schwierig und beruht auf Anwendung des Mittelwertsatzes. Er wird hier aus Zeitgründen nicht durchgeführt.

Haben wir eine k-mal stetig differenzierbare Funktion, so können wir Satz 11.1.4 $(k-1)$-mal anwenden um zu sehen, dass bei allen partiellen Ableitungen der Ordnung kleiner gleich k die Reihenfolge der partiellen Ableitungen keine Rolle spielt.

Ist $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ zweimal partiell differenzierbar, so ist $\nabla f : U \to \mathbb{R}^n$ und wir können die Jacobi-Matrix von ∇f betrachten:

$$Hf(x) := J\nabla f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix}.$$

Die $n \times n$-Matrix $Hf(x)$ heißt die **Hesse-Matrix** von f in Punkt x. Das folgende Korollar, das sofort aus Satz 11.1.4 folgt, zeigt, dass die Hesse-Matrix "typischerweise" symmetrisch ist.

Korollar 11.1.5

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ zweimal stetig partiell differenzierbar. Dann ist $Hf(x)$ für alle $x \in U$ symmetrisch.

In unserem ersten Beispiel schauen wir uns nochmal die Abbildung $f : \mathbb{R}^2 \to \mathbb{R}$ an, die durch $f(x_1, x_2) := x_1^2 + x_2^2$ definiert ist. Es gilt dann

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1 \qquad \text{und} \qquad \frac{\partial f}{\partial x_2}(x_1, x_2) = 2x_2\,,$$

und ein weiteres partielles Ableiten ergibt

$$Hf(x) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2E_2\,, \qquad x \in \mathbb{R}^2.$$

In einem weiteren Beispiel fixieren wir ein $A \in \mathcal{M}(m,n)$ mit Einträgen a_{ij} und betrachten die zugehörige, lineare Abbildung $f : \mathbb{R}^n \to \mathbb{R}^m$. Es gilt somit

$$f(x) = Ax = \begin{pmatrix} \sum_{j=1}^n a_{1j}x_j \\ \vdots \\ \sum_{j=1}^n a_{mj}x_j \end{pmatrix}, \qquad x \in \mathbb{R}^n.$$

Ist nun f_i die i-te Komponentenfunktionen von f, so gilt für $j \in \{1, \ldots, n\}$

$$\frac{\partial f_i}{\partial x_j}(x) = a_{ij}$$

für alle $x \in \mathbb{R}^n$. Dies ergibt die Formel

$$Jf(x) = A\,, \qquad x \in \mathbb{R}^n\,,$$

die den eindimensionalen Fall auf einfache Weise verallgemeinert. Insbesondere ist die Jacobi-Matrix von linearen Abbildungen eine konstante Abbildung.

BEISPIEL: ABLEITUNG QUADRATISCHER FUNKTIONEN

Für unser letztes Beispiel fixieren wir nun eine quadratische Matrix $A \in \mathcal{M}(n,n)$ und betrachten die Funktion $f : \mathbb{R}^n \to \mathbb{R}$, die durch

$$f(x) := \langle x, Ax \rangle , \qquad x \in \mathbb{R}^n$$

definiert ist. Ausgeschrieben ist dies

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j , \qquad x \in \mathbb{R}^n .$$

Für $k = 1 \ldots, n$ erhalten wir dann

$$\frac{\partial f}{\partial x_k}(x) = \sum_{i=1}^{n} a_{ik} x_i + \sum_{j=1}^{n} a_{kj} x_j = \langle Ae_k, x \rangle + \langle A^{\mathsf{T}} e_k, x \rangle = \langle (A + A^{\mathsf{T}}) e_k, x \rangle$$

$$= \langle e_k, (A^{\mathsf{T}} + A) x \rangle .$$

Da $\frac{\partial f}{\partial x_k}(x)$ die k-te Spalte von $\mathbf{grad}\, f(x)$ ist und $\langle e_k, (A^{\mathsf{T}} + A)x \rangle$ die k-te Spalte von $A^{\mathsf{T}} + A = A + A^{\mathsf{T}}$, ergibt dies

$$\mathbf{grad}\, f(x) = (A + A^{\mathsf{T}})x , \qquad x \in \mathbb{R}^n . \tag{11.1.2}$$

Insbesondere ist $x \mapsto \mathbf{grad}\, f(x)$ linear und unser vorheriges Beispiel ergibt

$$Hf(x) = J\nabla f(x) = J(\mathbf{grad}\, f(x))^{\mathsf{T}} = A + A^{\mathsf{T}} , \qquad x \in \mathbb{R}^n .$$

Section 11.2
Mehrdimensionale Ableitungen

Partielle Ableitungen beschreiben ihrer Definition nach das Verhalten einer Funktion in Richtung der Achsen. Leider kann es in andere Richtungen jedoch zu unerwünschtem Verhalten kommen, wie das folgende Beispiel verdeutlicht.

Wir betrachten dazu die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$, die durch

$$f(x,y) := \begin{cases} \frac{xy}{(x^2+y^2)^2}\,, & \text{falls } (x,y)^\mathsf{T} \neq 0 \\ 0\,, & \text{sonst} \end{cases} \tag{11.2.1}$$

definiert ist. Mit der Produktregel für die Zerlegung $x \cdot \frac{y}{(x^2+y^2)^2}$ gilt dann

$$\frac{\partial f}{\partial x}(x,y) = \frac{y}{(x^2+y^2)^2} - \frac{4x^2y}{(x^2+y^2)^3}$$

und analog erhalten wir

$$\frac{\partial f}{\partial y}(x,y) = \frac{x}{(x^2+y^2)^2} - \frac{4xy^2}{(x^2+y^2)^3}$$

für $(x,y)^\mathsf{T} \neq 0$.

Ferner gilt

$$\frac{\partial f}{\partial x}(0,0) = \lim_{h \to 0} \frac{f(h,0) - f(0,0)}{h} = \lim_{h \to 0} \frac{0-0}{h} = 0$$

und analog $\frac{\partial f}{\partial y}(0,0) = 0$. Damit ist f partiell differenzierbar. Es gilt aber auch

$$f(1/n, 1/n) = \frac{n^{-2}}{(n^{-2}+n^{-2})^2} = \frac{n^{-2}}{4n^{-4}} = \frac{n^2}{4} \to \infty$$

für $n \to \infty$. Damit ist f auf der Diagonalen nicht mal beschränkt und nicht stetig.

Um dieses Problem zu lösen, erinnern wir uns an Satz 6.1.4, der zeigte, dass eine Funktion $f : \mathbb{R} \to \mathbb{R}$ in $x_0 \in \mathbb{R}$ genau dann differenzierbar ist, wenn es ein $a \in \mathbb{R}$ und eine Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ mit $\varphi(h) = \mathrm{o}(h)$ für $h \to 0$ gibt, so dass

$$f(x_0 + h) = f(x_0) + a \cdot h + \varphi(h)\,, \qquad h \in \mathbb{R}$$

gilt. In diesem Fall ist ferner $a = f'(x_0)$. Mit anderen Worten lässt sich f in x_0 linear mit Ordnung $\mathrm{o}(h)$ für $x_0 + h \to x_0$ approximieren.

Die folgende Definition benutzt nun diese Charakterisierung, um mehrdimensionale Ableitungen zu definieren.

Definition 11.2.1

Sei $U \subset \mathbb{R}^n$ offen, $f : U \to \mathbb{R}^m$ und $x_0 \in U$. Dann ist f in x_0 Fréchet-differenzierbar, falls es ein $A \in \mathcal{M}(m, n)$ und eine Funktion $r : U \to \mathbb{R}^m$ gibt mit

$$f(x_0 + h) = f(x_0) + Ah + r(h) \tag{11.2.2}$$

für alle $h \in \mathbb{R}^n$ mit $x_0 + h \in U$, sowie $r(0) = 0$ und

$$\lim_{h \to 0} \frac{r(h)}{\|h\|} = 0\,. \tag{11.2.3}$$

Ferner heißt f Fréchet-differenzierbar, wenn es Fréchet-differenzierbar in allen $x_0 \in U$ ist.

Die Fréchet-Differenzierbarkeit wird häufig auch als totale Differenzierbarkeit, oder auch nur als Differenzierbarkeit bezeichnet.

ABLEITUNG IST JACOBI-MATRIX

Der folgende Satz zeigt, dass Fréchet-Differenzierbarkeit stärker als die partielle Differenzierbarkeit ist und dass die Matrix A in (11.2.2) eindeutig ist.

Theorem 11.2.2

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^m$ in x_0 Fréchet-differenzierbar. Dann ist f in x_0 partiell differenzierbar und $f'(x_0) := Jf(x_0)$ ist die einzige Matrix, die (11.2.2) *mit* (11.2.3) *erfüllt.*

Im Falle der Fréchet-Differenzierbarkeit von f in x_0 wird $f'(x_0)$ auch als Fréchet-Ableitung, als totale Ableitung, oder auch nur als Ableitung von f in x_0 bezeichnet.

Sei A eine Matrix $A \in \mathcal{M}(m,n)$, die (11.2.2) mit (11.2.3) erfüllt. Für $i = 1, \ldots, m$ betrachte die Komponentenfunktion $f_i := \langle e_i, f \rangle$. Für $j = 1, \ldots, n$ gilt dann

$$\begin{aligned} \langle e_i, f(x_0 + he_j) \rangle &= \langle e_i, f(x_0) + A(he_j) + r(he_j) \rangle \\ &= f_i(x_0) + \langle e_i, Ae_j \rangle \cdot h + \langle e_i, r(he_j) \rangle \,. \end{aligned}$$

Setzen wir dann $\varphi(h) := \langle e_i, r(he_j) \rangle$, so folgt mit (11.2.3) und der Cauchy-Schwarz-Ungleichung

$$\left| \frac{\varphi(h)}{|h|} \right| = \left| \frac{\langle e_i, r(he_j) \rangle}{\|he_j\|} \right| \leq \left| \frac{\|e_i\| \cdot \|r(he_j)\|}{\|he_j\|} \right| \to 0$$

für $h \to 0$. Satz 6.1.4 zeigt dann, dass f_i in x_0 partiell in der j-ten Koordinate differenzierbar ist mit

$$a_{ij} = \langle e_i, Ae_j \rangle = \frac{\partial f_i}{\partial x_j}(x) \,.$$

Die Definition von $Jf(x_0)$ ergibt dann die Behauptung.

Recht analog lässt sich das folgende Lemma zeigen.

Lemma 11.2.3

Sei $U \subset \mathbb{R}^n$ offen, $x_0 \in U$ und $f : U \to \mathbb{R}^m$ mit Komponentenfunktion $f_1, \ldots, f_m$. Dann ist f in x_0 Fréchet-differenzierbar, genau dann wenn alle $f_1, \ldots, f_n$ in x_0 Fréchet-differenzierbar sind.

BEWEIS 1/1

Ist f in x_0 Fréchet-differenzierbar, so gilt mit $A := f'(x_0)$

$$\begin{aligned} f_i(x_0 + h) &= \langle e_i, f(x_0 + h)\rangle + \langle e_i, Ah\rangle + \langle e_i, r(h)\rangle \\ &= f_i(x_0) + \langle A^\mathsf{T} e_i, h\rangle + \langle e_i, r(h)\rangle \\ &= f_i(x_0) + a_i^\mathsf{T} h + \langle e_i, r(h)\rangle\,, \end{aligned}$$

wobei $a_i := A^\mathsf{T} e_i$ der i-te Spaltenvektor von A^T ist, d.h. a_i^T ist der i-te Zeilenvektor von A. Mit der Cauchy-Schwarz-Ungleichung haben wir ferner

$$\left|\frac{\langle e_i, r(h)\rangle}{\|h\|}\right| \leq \left|\frac{\|e_i\| \cdot \|r(h)\|}{\|h\|}\right| \to 0$$

und damit ist f_i in x_0 Fréchet-differenzierbar.

BEWEIS 2/2

Die umgekehrte Implikation ist ähnlich zu beweisen. Hierzu sei $a_i \in \mathcal{M}(1,n)$ die Ableitung $f_i'(x_0)$. Wir schreiben

$$A := \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \in \mathcal{M}(m,n)\,.$$

Dann gilt

$$\begin{aligned} f(x_0+h) = \sum_{i=1}^{m} \langle e_i, f(x_0+h)\rangle e_i &= \sum_{i=1}^{m} f_i(x_0+h)e_i \\ &= \sum_{i=1}^{m} \Big(f_i(x_0)e_i + (a_i \cdot h)e_i + r_i(h)e_i \Big) \\ &= f(x_0) + Ah + \sum_{i=1}^{m} r_i(h)e_i\,. \end{aligned}$$

Definieren wir nun die letzte Summe als $r(h)$, so erfüllt $r(h)$ die Aussage (11.2.3), da alle $r_i(h)$ die Aussage (11.2.3) erfüllen.

Ist $f : U \to \mathbb{R}^m$ in x_0 Fréchet-differenzierbar, so haben wir

$$\frac{f(x_0 + h) - f(x_0) - f'(x_0)h}{\|h\|} = \frac{r(h)}{\|h\|} \to 0$$

für $h \to 0$. Multiplizieren wir mit $\|h\|$, so erhalten wir

$$f(x_0 + h) - f(x_0) - f'(x_0)h = r(h) \to 0$$

für $h \to 0$. Wegen $f'(x_0)h \to 0$ für $h \to 0$ haben wir dann das folgende Lemma gezeigt, welches Korollar 6.1.5 verallgemeinert.

Lemma 11.2.4

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^m$ in x_0 Fréchet-differenzierbar. Dann ist f in x_0 stetig.

Die in (11.2.1) definierte, partiell differenzierbare Abbildung f ist nicht Fréchet-differenzierbar in 0, da f nicht stetig in 0 ist. Insbesondere gibt es partiell differenzierbare Abbildungen, die nicht Fréchet-differenzierbar sind, d.h. die Umkehrung von Satz 11.2.2 gilt im Allgemeinen nicht.

STETIGE PARTIELLE DIFFERENZIERBARKEIT REICHT

Wenn wir jedoch stetig partiell differenzierbare Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ betrachten, so sind diese auch Fréchet-differenzierbar, wie der folgende Satz zeigt.

Theorem 11.2.5

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ stetig partiell differenzierbar. Dann ist f in allen $x_0 \in U$ Fréchet differenzierbar.

Mit Hilfe von Lemma 11.2.3 lässt sich der Satz 11.2.5 leicht auf den Fall $f : U \to \mathbb{R}^m$ mit $U \subset \mathbb{R}^n$ offen verallgemeinern.

BEWEIS 1/3

Ohne Einschränkung können wir $U = U(x_0, \delta)$ für ein $\delta > 0$ annehmen. Wir fixieren nun ein $h \in \mathbb{R}^n$ mit $\|h\| < \delta$. Für $k = 0, \ldots, n$ definieren wir

$$z^{(k)} := x_0 + \sum_{j=1}^{k} h_j e_j \,.$$

Dies ergibt $z^{(0)} = x_0$ und $z^{(n)} = x_0 + h$. Ferner unterscheiden sich $z^{(k)}$ und $z^{(k-1)}$ nur in der k-ten Koordinate. Ist $h_k \neq 0$, so wenden wir nun den Mittelwertsatz 6.2.3 auf die k-te Koordinate an, um ein $\xi_k \in [0, 1]$ zu erhalten, so dass für

$$y^{(k)} := z^{(k-1)} + \xi_k h_k e_k$$

gilt

$$\frac{f(z^{(k)}) - f(z^{(k-1)})}{h_k} = \frac{\partial f}{\partial x_k}(y^{(k)}) \,.$$

Dies ergibt

$$f(z^{(k)}) - f(z^{(k-1)}) = h_k \frac{\partial f}{\partial x_k}(y^{(k)}) \,. \tag{11.2.4}$$

Im Fall $h_k = 0$ ist dies auch wahr, da wir dann $z^{(k)} = z^{(k-1)}$ haben und deshalb z.B. $y^{(k)} := z^{(k-1)}$ betrachten können. Aufsummieren der Gleichungen (11.2.4) über k führt zu

$$f(x_0 + h) - f(x_0) = \sum_{k=1}^{n} \Big(f(z^{(k)}) - f(z^{(k-1)}) \Big) = \sum_{k=1}^{n} h_k \frac{\partial f}{\partial x_k}(y^{(k)}) \,, \qquad (11.2.5)$$

wobei wir im ersten Schritt ein Teleskop-Summen-Argument benutzt haben. Wir definieren nun

$$a_k := \frac{\partial f}{\partial x_k}(x_0) \,,$$

$$r(h) := \sum_{k=1}^{n} h_k \cdot \Big(\frac{\partial f}{\partial x_k}(y^{(k)}) - a_k \Big) \,.$$

Dies führt (11.2.5) in

$$f(x_0 + h) = f(x_0) + ah + r(h)$$

über, wobei $a := (a_1, \ldots, a_n) \in \mathcal{M}(1, n)$ ist. Mit $y_h := (y^{(1)}, \ldots, y^{(n)})^{\mathsf{T}}$ gilt ferner

$$\left| \frac{r(h)}{\|h\|} \right| = \left| \frac{\big\langle \operatorname{grad} f(y_h) - \operatorname{grad} f(x_0), h \big\rangle}{\|h\|} \right| \leq \| \operatorname{grad} f(y_h) - \operatorname{grad} f(x_0) \| \,.$$

BEWEIS 3/3

Für $h \to 0$ gilt nun $y_h \to x_0$ nach Konstruktion der Komponenten $y^{(k)}$ von y_h. Dies impliziert $\| \mathbf{grad}\, f(y_h) - \mathbf{grad}\, f(x_0) \| \to 0$, da f stetig partiell differenzierbar ist.

Der Beweis des folgenden Satzes ist ähnlich zu dem von Satz 11.2.5 und wird deshalb übersprungen.

Theorem 11.2.6

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ partiell differenzierbar. Sind die partiellen Ableitungen beschränkt, so ist f stetig.

Ist $f : U \to \mathbb{R}^m$ Fréchet-differenzierbar, so ist

$$f' : U \to \mathcal{M}(m, n) = \mathbb{R}^{m \times n} .$$

Ist f' sogar stetig, so heißt f **stetig differenzierbar**. Da f' wiederum in einen $\mathbb{R}^k$ abbildet, können wir auch die Fréchet-Differenzierbarkeit von f' und die zugehörige Abbildung f'' betrachten. Identifizieren wir Matrizen mit linearen Abbildungen, so ergibt sich dabei

$$f'' : U \to \mathcal{M}(m \times n, n) = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^{m \times n}) = \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)) ,$$

wobei der letzte Raum der Menge aller bilinearen Abbildungen $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ vermöge der Identifikation

$$(x, x') \mapsto (Sx)(x') , \qquad x, x' \in \mathbb{R}^n$$

für $S \in \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m))$ entspricht. Die zweite Ableitungen $f''(x_0)$ ist damit eine bilineare Abbildung $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ und weiteres Iterieren ergibt, dass die k-te Ableitung $f^{(k)}(x_0)$ eine k-lineare $\mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}^m$ ist, wobei das k-fache Produkt von $\mathbb{R}^n$ betrachtet wird.

Partielle Ableitungen betrachten das Verhalten einer Funktion beim Verändern des Arguments parallel zu den Achsen des Koordinatensystems. Die folgende Definition verallgemeinert dies auf beliebige Richtungen.

Definition 11.2.7

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$. Für $x_0 \in U$ und $v \in \mathbb{R}^n$ mit $v \neq 0$ heißt

$$D_v f(x_0) := \lim_{h \to 0} \frac{f(x_0 + hv) - f(x_0)}{h}$$

die Richtungs- oder Gâteaux-Ableitung von f in x_0 in Richtung v, falls dieser Grenzwert existiert. Ferner heißt f Gâteaux-differenzierbar in x_0, falls $D_v f(x_0)$ für alle $v \in \mathbb{R}^n \setminus \{0\}$ existiert.

Man beachte, dass mit dieser Definition $D_{e_i} f(x_0) = D_i f(x_0)$ gilt, d.h. die Richtungs-Ableitung verallgemeinert tatsächlich die partiellen Ableitungen.

FRÉCHET IMPLIZIERT GÂTEAUX

Der folgende Satz zeigt, dass die Fréchet-Differenzierbarkeit die Gâteaux-Differenzierbarkeit impliziert.

Theorem 11.2.8

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$. Ist f Fréchet-differenzierbar in $x_0 \in U$, so ist f auch f Gâteaux-differenzierbar in x_0 und es gilt

$$D_v f(x_0) = \langle \nabla f(x_0), v \rangle .$$

BEWEIS

Da f in x_0 Fréchet-differenzierbar ist, gilt für $v \in \mathbb{R}^n$ mit $v \neq 0$ und $A := f'(x_0) = (\nabla f(x_0))^\mathsf{T}$:

$$\begin{aligned} 0 &= \lim_{h \to 0} \frac{f(x_0 + hv) - f(x_0) - A(hv)}{h\|v\|} \\ &= \frac{1}{\|v\|} \lim_{h \to 0} \frac{f(x_0 + hv) - f(x_0) - hAv}{h} \\ &= \frac{1}{\|v\|} \left(D_v f(x_0) - Av \right). \end{aligned}$$

Wegen $Av = (\nabla f(x_0))^\mathsf{T} v = \langle \nabla f(x_0), v \rangle$ folgt dann die Behauptung.

BEISPIEL

Als Beispiel betrachten wir nun die Funktion $f(x,y) := x^2 + y^2$ für $x, y \in \mathbb{R}$. Wir haben dann

$$\nabla f(x,y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}, \qquad \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2.$$

Für $x_0 := 1$ und $y_0 := 0$ haben wir also

$$\nabla f(x_0, y_0) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

Sei nun $v := (1,1)^T$. Mit Satz 11.2.8 haben wir dann

$$D_v f(x_0, y_0) = \left\langle \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\rangle = 2,$$

während für den Richtungsvektor $2v$ gilt

$$D_{2v} f(x_0, y_0) = \left\langle \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\rangle = 4.$$

Damit hängt die Richtungsableitung nicht nur von der Richtung, sondern auch von der Länge des Richtungsvektors ab.

BEISPIEL

Betrachten wir nun nur $v \in \mathbb{R}^2$ mit $\|v\| = 1$, so ist

$$D_v f(x_0, y_0) = \left\langle \begin{pmatrix} 2 \\ 0 \end{pmatrix}, v \right\rangle$$

maximal für $v := \frac{1}{\sqrt{2}} \begin{pmatrix} 2 \\ 0 \end{pmatrix}$. Ferner gilt

$$D_v f(x_0, y_0) = 0$$

für $v \perp \begin{pmatrix} 2 \\ 0 \end{pmatrix}$, d.h. für $v = \pm e_2$.

Der folgende Satz verallgemeinert die Maximalitäts-Beobachtung in dem obigen Beispiel.

Theorem 11.2.9

Sei $U \subset \mathbb{R}^n$ offen, $x_0 \in U$ und $f : U \to \mathbb{R}$ in x_0 Fréchet-differenzierbar mit $\nabla f(x_0) \neq 0$. Dann gilt

$$\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|} = \operatorname*{argmax}_{v \in \mathbb{R}^n : \|v\| = 1} D_v f(x_0)\,.$$

Der Gradient $\mathbf{grad}\, f(x_0)$ ergibt also die Richtung des steilsten Anstiegs von f in x_0. Dies wird häufig zur Maximierung von Funktionen $f : U \to \mathbb{R}$ benutzt.

Für $v \in \mathbb{R}^n$ mit $\|v\| = 1$ gilt

$$D_v f(x_0) = \left\langle \nabla f(x_0), v \right\rangle \leq \|\nabla f(x_0)\| \cdot \|v\| = \|\nabla f(x_0)\| .$$

und für $v^* := \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}$ gilt sogar

$$D_{v^*} f(x_0) = \left\langle \nabla f(x_0), v^* \right\rangle = \|\nabla f(x_0)\| .$$

Zusammen zeigt dies, dass das Maximum von $v \mapsto D_v f(x_0)$ in v^* angenommen wird.

Section 11.3
Differentialoperatoren

Eine Abbildung $f : \mathbb{R}^n \to \mathbb{R}^n$ heißt **Vektorfeld**. Im Fall von $n = 3$ kann dies also zum Beispiel ein Kraftfeld sein, das in jedem Punkt $x \in \mathbb{R}^n$ die Kraft auf ein Objekt in diesem Punkt beschreibt. In diesem Kapitel wollen wir ein paar wichtige Differential-Operatoren für solche Vektorfelder kennenlernen.

Bevor wir damit beginnen, sagen wir, dass ein $f : \mathbb{R}^n \to \mathbb{R}^n$ ein **Gradientenfeld** ist, falls es eine partiell differenzierbare Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}$ gibt mit

$$f(x) = \nabla\varphi(x)\,, \qquad x \in \mathbb{R}^n\,. \tag{11.3.1}$$

Analoge Definition sind für $f : U \to \mathbb{R}^n$ für offene Mengen $U \subset \mathbb{R}^n$ üblich. Gradientenfelder werden häufig auch **konservative Felder** genannt.

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^n$ eine Funktionen, deren Komponentenfunktionen partiell differenzierbar sind. Dann heißt

$$\operatorname{div} f(x) := \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i}(x)\,, x \in U$$

Divergenz von f. Es ist $\operatorname{div} f : U \to \mathbb{R}$. Andere Schreibweisen für die Divergenz sind

$$\nabla^{\mathsf{T}} f(x) := \langle \nabla, f(x) \rangle := \operatorname{div} f(x)\,.$$

Anschaulich gesprochen, beschreibt die Divergenz $\operatorname{div} f(x)$ die Zu- und Abflüsse in dem Punkt x, wobei Quellen positive Divergenz und Senken negative Divergenz haben. Genauer wird dies in dem Gaußschen Integralsatz beschrieben, der in der HM 3 behandelt wird. Ein Vektorfeld $f : U \to \mathbb{R}^n$ heißt divergenz- oder quellenfrei, falls

$$\operatorname{div} f(x) = 0\,, \qquad x \in U.$$

Betrachten wir zum Beispiel die Identität $\mathrm{id}_{\mathbb{R}^n} : \mathbb{R}^n \to \mathbb{R}^n$, so gilt

$$\operatorname{div} \mathrm{id}_{\mathbb{R}^n}(x) = \sum_{i=1}^{n} \frac{\partial x_i}{\partial x_i}(x) = \sum_{i=1}^{n} 1 = n\,. \tag{11.3.2}$$

Damit ist die Identität nicht divergenzfrei. Konstante Funktionen sind jedoch immer divergenzfrei.

Es ist leicht zu überprüfen, dass die Divergenz linear ist, d.h. für $f, g : U \to \mathbb{R}^n$ und $\alpha, \beta \in \mathbb{R}$ gilt

$$\operatorname{div}(\alpha f + \beta g) = \alpha \operatorname{div} f + \beta \operatorname{div} g\,.$$

Das folgende Lemma zeigt die Produktregel der Divergenz.

Lemma 11.3.1

Seien $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^n$ eine Funktionen, deren Komponentenfunktionen partiell differenzierbar sind und $\varphi : U \to \mathbb{R}$ partiell differenzierbar. Dann gilt

$$\operatorname{div}(\varphi f) = \langle \nabla \varphi, f \rangle + \varphi \operatorname{div} f\,.$$

Mit der Produktregel für reelle Ableitungen gilt

$$\begin{aligned}\operatorname{div}(\varphi f)(x) = \sum_{i=1}^{n} \frac{\partial(\varphi f_i)}{\partial x_i}(x) &= \sum_{i=1}^{n} \frac{\partial \varphi}{\partial x_i}(x) \cdot f_i(x) + \sum_{i=1}^{n} \varphi(x) \frac{\partial f_i}{\partial x_i}(x) \\ &= \langle \nabla \varphi(x), f(x) \rangle + \varphi(x) \operatorname{div} f(x)\end{aligned}$$

für alle $x \in U$. Dies zeigt die Behauptung.

BEISPIEL

Um ein Beispiel für das Lemma 11.3.1 zu betrachten, sei $k \in \mathbb{N}$ und $g_k : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}^n$ die durch

$$g_k(x) := \frac{1}{\|x\|^k} \cdot x$$

definierte Funktion. Mit $r(x) := \|x\|$ und $h(t) := t^{-k}$ zeigt dann Lemma 11.3.1 zusammen mit (11.3.2) und (11.1.1)

$$\begin{aligned}
\operatorname{div} g_k(x) = \left\langle \nabla r^{-k}(x), x \right\rangle + \frac{1}{r^k(x)} \operatorname{div} x &= \langle \nabla (h \circ r)(x), x \rangle + \frac{1}{r^k(x)} n \\
&= \langle (h' \circ r)(x) \cdot g_1(x), x \rangle + \frac{n}{\|x\|^k} \\
&= \langle -k\|x\|^{-k-1} \cdot g_1(x), x \rangle + \frac{n}{\|x\|^k} \\
&= -k\|x\|^{-k-2} \langle x, x \rangle + \frac{n}{\|x\|^k} \\
&= -k\|x\|^{-k} + \frac{n}{\|x\|} \\
&= \frac{n-k}{\|x\|^k} \qquad (11.3.3)
\end{aligned}$$

für alle $x \neq 0$. Damit ist g_k genau dann divergenzfrei, wenn $k = n$ gilt.

Sei $U \subset \mathbb{R}^n$ offen. Dann ordnet der Laplace-Operator jeder zweimal partiell differenzierbaren Funktion $\varphi : U \to \mathbb{R}$ die Funktion

$$\Delta\varphi := \sum_{i=1}^{n} \frac{\partial^2 \varphi}{\partial^2 x_i}$$

zu. Es gilt $\Delta\varphi : U \to \mathbb{R}$ und

$$\Delta\varphi = \mathbf{div}(\nabla\varphi)\,, \tag{11.3.4}$$

wobei dies wegen

$$\mathbf{div}(\nabla\varphi)(x) = \sum_{i=1}^{n} \frac{\partial(\nabla\varphi)_i}{\partial x_i}(x) \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \frac{\partial\varphi}{\partial x_i}(x) = \Delta\varphi(x)$$

für alle $x \in U$ gilt.

Ist $\varphi : U \to \mathbb{R}$ eine Funktion mit

$$\Delta\varphi(x) = 0\,, \qquad x \in U, \tag{11.3.5}$$

so zeigt (11.3.4), dass das Gradientenfeld $f := \nabla\varphi : U \to \mathbb{R}^n$ ein divergenzfreies Vektorfeld ist, denn wir haben

$$\operatorname{div} f = \operatorname{div}(\nabla\varphi) = \Delta\varphi = 0\,.$$

Die Gradientenfelder der Lösungen der Laplace-Gleichung oder Potential-Gleichung (11.3.5) sind also divergenzfreie Vektorfelder.

Das folgende Lemma berechnet den Laplace-Operator für rotationsinvariante Funktionen.

Lemma 11.3.2

Sei $h : (0, \infty) \to \mathbb{R}$ zweimal differenzierbar und $f : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ die durch

$$f(x) = h(\|x\|), \qquad x \neq 0 \tag{11.3.6}$$

definierte, rotationsinvariante Funktion. Dann gilt

$$\Delta f(x) = h''(\|x\|) + (n-1) \cdot \frac{h'(\|x\|)}{\|x\|}, \qquad x \neq 0.$$

Für $x \neq 0$ definieren wir $r(x) := \|x\|$ und $g(x) := \|x\|^{-1}x$. Dann haben wir

$$\begin{aligned}
\Delta f = \Delta(h \circ r) = \operatorname{div}(\nabla(h \circ r)) &= \operatorname{div}\big((h' \circ r) \cdot g\big) \\
&= \langle \nabla(h' \circ r), g\rangle + (h' \circ r) \cdot \operatorname{div} g \\
&= \langle (h'' \circ r) \cdot g, g\rangle + (h' \circ r) \cdot \frac{n-1}{r} \\
&= (h'' \circ r)\langle g, g\rangle + (h' \circ r) \cdot \frac{n-1}{r} \\
&= (h'' \circ r) + (h' \circ r) \cdot \frac{n-1}{r}\,,
\end{aligned}$$

wobei wir nacheinander (11.3.4), (11.1.1), Lemma 11.3.1, (11.1.1), (11.3.3) und $\langle g, g\rangle = 1$ angewendet haben.

Mit Hilfe von Lemma 11.3.2 können wir nun die rotations-invarianten Lösungen der Laplace-Gleichung bestimmen.

Theorem 11.3.3

Sei $h : (0, \infty) \to \mathbb{R}$ zweimal differenzierbar und $f : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ die durch (11.3.6) *bestimmte rotations-invariante Funktion. Dann gilt*

$$\Delta f = 0$$

genau dann, wenn es ein $c_1, c_2 \in \mathbb{R}$ gibt, so dass für alle $r > 0$ gilt:

$$h(r) = \begin{cases} c_1 r^{-n+2} + c_2 & \text{falls } n \neq 2, \\ c_1 \ln r + c_2 & \text{falls } n = 2. \end{cases}$$

Im Fall $n = 3$ und $c_2 = 0$ erhalten wir die Lösungen $f(x) = c_1 \|x\|^{-1}$, die z.B. als Potential eines Gravitationsfeldes einer punktförmigen Masse im Ursprung angesehen werden kann. Das zugehörige Gradientenfeld entspricht dann dem Kraftfeld. Alternativ kann auch das Potential eines elektrischen Feldes betrachtet werden.

BEWEIS

Wir nehmen zunächst an, dass $\Delta f = 0$ erfüllt ist. Setzen wir $r(x) := \|x\|$ für $x \in \mathbb{R}^n \setminus \{0\}$, so gilt mit Lemma 11.3.2:

$$0 = \Delta f(x) = h''(r) + (n-1) \cdot \frac{h'(r)}{r}$$

Wir definieren nun $g(r) := h'(r)$ für $r > 0$. Dann erhalten wir

$$g'(r) = -\frac{n-1}{r} \cdot g(r)\,, \qquad\qquad r > 0.$$

Wie im Abschnitt 453 lösen wir diese Differential-Gleichung:

$$\begin{aligned} \ln|g(x)| = \int \frac{g'(x)}{g(x)}\,\mathrm{d}x = -(n-1)\int x^{-1}\,\mathrm{d}x &= -(n-1)\ln x + C \\ &= \ln\left(e^C \cdot x^{-(n-1)}\right). \end{aligned}$$

Dies führt zu

$$g(x) = \pm e^C \cdot x^{-(n-1)} = cx^{-(n-1)}$$

für $c \in \mathbb{R}$. Berechnen der Stammfunktionen von g ergibt dann die Gestalt von h.

Die umgekehrte Implikation kann durch einfaches Nachrechnen überprüft werden.

Ist $U \subset \mathbb{R}^3$ offen und $f : U \to \mathbb{R}^3$ eine Funktion, deren Komponentenfunktionen partiell differenzierbar sind. Dann heißt

$$\operatorname{rot} f := \begin{pmatrix} \frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3} \\ \frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2} \end{pmatrix} : U \to \mathbb{R}^3$$

Rotation von f. Eine andere Schreibweise für die Rotation ist

$$\nabla \times f := \operatorname{rot} f .$$

Ein Vektorfeld $f : U \to \mathbb{R}^3$ heißt rotationsfrei, falls $\operatorname{rot} f(x) = 0$ für alle $x \in U$ gilt.

Die Rotation ist linear, d.h. für $f, g : U \to \mathbb{R}^3$ und $\alpha, \beta \in \mathbb{R}$ gilt

$$\operatorname{rot}(\alpha f + \beta g) = \alpha \operatorname{rot} f + \beta \operatorname{rot} g ,$$

wobei natürlich vorausgesetzt wird, dass $\operatorname{rot} f$ und $\operatorname{rot} g$ definiert sind. Ist ferner $\varphi : U \to \mathbb{R}$ partiell differenzierbar, so gilt die Produktregel

$$\operatorname{rot}(\varphi f)(x) = \nabla\varphi(x) \times f(x) + \varphi(x) \operatorname{rot} f(x) , \qquad x \in U,$$

wobei wir wieder annehmen, dass $\operatorname{rot} f$ definiert ist. Diese Gleichung ist durch einfaches Nachrechnen zu überprüfen.

Das folgende Lemma zeigt, dass Gradientenfelder rotationsfrei sind.

Lemma 11.3.4

Sei $U \subset \mathbb{R}^3$ offen und $\varphi : U \to \mathbb{R}$ zweimal stetig partiell differenzierbar. Dann gilt

$$\mathrm{rot}(\nabla\varphi)(x) = 0\,, \qquad x \in U.$$

Sei $f_i := \frac{\partial \varphi}{\partial x_i}$. Dann gilt

$$\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3} = \frac{\partial^2 \varphi}{\partial x_2 \partial x_3} - \frac{\partial^2 \varphi}{\partial x_3 \partial x_2} = 0$$

nach dem Satz von Schwarz, siehe Satz 11.1.4. Analoge Rechnungen für die übrigen beiden Koordinaten ergeben dann die Behauptung.

BEISPIEL

Sei $\psi : (0, \infty) \to \mathbb{R}$ stetig differenzierbar und $f : \mathbb{R}^3 \setminus \{0\} \to \mathbb{R}^3$ durch

$$f(x) := \psi(\|x\|) \cdot x$$

definiert. Ist dann $t \mapsto t\psi(t)$ auf jedem Intervall $(0, a]$ uneigentlich Riemann-integrierbar, so gilt

$$\mathbf{rot}\, f(x) = 0\,, \qquad x \neq 0. \tag{11.3.7}$$

Um dies zu sehen, definieren wir $h : (0, \infty) \to \mathbb{R}$ durch

$$h(t) := \int_0^t s\psi(s)\,\mathrm{d}s\,, \qquad t > 0.$$

Dies ergibt dann $h'(t) = t\psi(t)$ für alle $t > 0$. Damit lässt sich f als

$$f(x) = \frac{h'(\|x\|)}{\|x\|} x\,, \qquad x \neq 0$$

schreiben. Nach (11.1.1) gilt dann

$$f(x) = \nabla h(\|\cdot\|)(x)$$

und da h zweimal stetig differenzierbar nach Konstruktion ist, zeigt Lemma 11.3.4 dann (11.3.7).

Partielle Differentialgleichungen betrachten Ableitungen, die sowohl vom Raum als auch von der Zeit abhängen können. Im Folgenden interpretieren wir daher $x \in \mathbb{R}^n$ als Raum-Koordinaten und $t \in \mathbb{R}$ als Zeit. Der Laplace-Operator bezieht sich dann immer nur auf die Raum-Koordinaten. Ferner ist u eine Funktion, die von x und t abhängen kann.

Einige wichtige, lineare partielle Differentialgleichungen zweiter Ordnung sind dann die folgenden Typen:

Elliptischer Typ (Laplace-Gleichung)

$$0 = \Delta u \,.$$

Zweimal stetig differenzierbare Lösungen u heißen **harmonische Funktionen**. In Satz 11.3.3 hatten wir schon die rotationsinvarianten Lösungen bestimmt.

Parabolischer Typ (Wärmeleitungs- oder Diffusionsgleichung)

$$\frac{\partial u}{\partial t} = a \cdot \Delta u \,,$$

wobei $a > 0$. Im Fall $n = 1$ ist dies

$$\frac{\partial u}{\partial t}(x,t) = a \cdot \frac{\partial^2 u}{\partial^2 x}(x,t)$$

und Nachrechnen ergibt, dass

$$u(x,t) := \frac{1}{\sqrt{at}} \exp\left(-\frac{x^2}{4at}\right)$$

eine Lösung ist. Analog ist im Fall $n \geq 1$ die Funktion

$$u(x,t) := \frac{1}{(at)^{n/2}} \exp\left(-\frac{\|x\|^2}{4at}\right)$$

eine Lösung.

Hyperbolischer Typ (Wellengleichung)

$$\frac{\partial^2 u}{\partial^2 t} = c^2 \cdot \Delta u \,,$$

für $c \in \mathbb{R}$, oder lang ausgeschrieben

$$\frac{\partial^2 u}{\partial^2 t}(x,t) = c^2 \sum_{i=1}^{n} \frac{\partial^2 u}{\partial^2 x_n}(x,t) \,.$$

Diese partielle Differentialgleichung ist linear, denn sind u_1 und u_2 Lösungen und $\alpha, \beta \in \mathbb{R}$, so ist auch $\alpha u_1 + \beta u_2$ eine Lösung.

Im Fall $n = 1$ ist die Wellengleichung

$$\frac{\partial^2 u}{\partial^2 t}(x,t) = c^2 \cdot \frac{\partial^2 u}{\partial^2 x}(x,t)\,.$$

Sind dann f und g zwei beliebige, zweimal differenzierbare Funktionen $\mathbb{R} \to \mathbb{R}$, so ist die durch

$$u(x,t) := f(x+ct) + g(x-ct) \tag{11.3.8}$$

definierte Funktion eine Lösung, der Wellengleichung. Um dies zu sehen, schreiben wir $\tilde{f}(x,t) := f(x+ct)$. Dann gilt

$$\frac{\partial^2 \tilde{f}}{\partial^2 t}(x,t) = c^2 f''(x+ct) \qquad \text{und} \qquad \frac{\partial^2 \tilde{f}}{\partial^2 x}(x,t) = f''(x+ct)\,.$$

Damit ist $\tilde{f}$ eine Lösung der Wellengleichung und eine analoge Rechnung zeigt, dass auch $(x,t) \mapsto g(x-ct)$ eine Lösung ist. Mit der Linearität folgt dann, dass auch das obige u eine Lösung ist.

Wenn wir nun Anfangswerte für z.B. $t = 0$ haben, d.h. wir haben Funktionen $u_0, u_1 : \mathbb{R} \to \mathbb{R}$ mit

$$\begin{aligned} u(x,0) &= u_0(x)\,, \\ \frac{\partial u}{\partial t}(x,0) &= u_1(x) \end{aligned}$$

für alle $x \in \mathbb{R}$, so lässt sich eine Lösung finden, die diesen Anfangswerten und der Wellengleichung genügt. Um dies zu sehen, betrachten wir zunächst ein u der Form (11.3.8). Dann haben wir

$$u_1(x) = \frac{\partial u}{\partial t}(x,0) = cf'(x) - cg'(x)\,.$$

Integration ergibt dann

$$\begin{aligned} f(x) - g(x) &= \int_0^x f'(s) - g'(s)\,\mathrm{d}s + f(0) - g(0) \\ &= \frac{1}{c}\int_0^x u_1(s)\,\mathrm{d}s + f(0) - g(0)\,. \end{aligned}$$

Ferner haben wir $f(x) + g(x) = u(x, 0) = u_0(x)$. Addieren dieser beiden Gleichungen ergibt

$$f(x) = \frac{1}{2c} \int_0^x u_1(s)\,\mathrm{d}s + \frac{f(0) - g(0)}{2} + \frac{u_0(x)}{2}$$

und Subtrahieren ergibt

$$\begin{aligned} g(x) &= -\frac{1}{2c} \int_0^x u_1(s)\,\mathrm{d}s - \frac{f(0) - g(0)}{2} + \frac{u_0(x)}{2} \\ &= \frac{1}{2c} \int_x^0 u_1(s)\,\mathrm{d}s - \frac{f(0) - g(0)}{2} + \frac{u_0(x)}{2} \end{aligned}$$

für alle $x \in \mathbb{R}$. Einsetzen in (11.3.8) ergibt dann

$$u(x, t) = \frac{u_0(x + ct) + u_0(x - ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} u_1(s)\,\mathrm{d}s\,.$$

Dies ist auch als D'Alembert-Lösung der Wellengleichung bekannt.

Section 11.4
Kettenregel

Die Kettenregel für eindimensionale Funktionen ist eines der wichtigsten Werkzeuge, um Ableitungen komplizierter Funktionen zu bestimmen. In diesem Kapitel wollen wir sie auf den Fall $g \circ f$ mit

$$f : \mathbb{R}^n \to \mathbb{R}^m \qquad \text{und} \qquad g : \mathbb{R}^m \to \mathbb{R}^k$$

übertragen.

Zunächst brauchen wir ein paar kleine Vorbereitungen. Dazu erinnern wir daran, dass wir in Abschnitt 310 gesehen haben, dass der Einheitsball $B(0,1)$ bezüglich der euklidischen Norm $\|\cdot\|$ folgenkompakt ist. Stetige Funktionen $f : B(0,1) \to \mathbb{R}$ sind daher beschränkt, wie nach Satz 5.2.8 erläutert worden ist.

Ist nun $A \in \mathcal{M}(m,n)$, so ist die zugehörige lineare Abbildung $x \mapsto Ax$ stetig und die Abbildung $y \mapsto \|y\|$ ist ebenfalls stetig. Damit gilt

$$\|A\| := \sup_{x \in B(0,1)} \|Ax\| < \infty. \tag{11.4.1}$$

Das folgende Lemma, dessen Beweis einfach ist und deswegen weggelassen wird, zeigt, dass wir auf diese Weise eine Norm auf $\mathcal{M}(m,n)$ definiert haben.

Lemma 11.4.1

Für alle $m, n \geq 1$ definiert (11.4.1) *eine Norm auf $\mathcal{M}(m, n)$.*

Ist $x \in \mathbb{R}^n$ und $A \in \mathcal{M}(m, n)$, so gilt

$$\|Ax\| \leq \|A\| \cdot \|x\| . \tag{11.4.2}$$

Für $x = 0$ ist hierbei nicht zu zeigen, und für $x \neq 0$ gilt wegen $\|x\|^{-1} x \in B(0, 1)$:

$$\left\| A\left(\|x\|^{-1} x\right) \right\| \leq \|A\| .$$

Multiplizieren mit $\|x\|$ auf beiden Seiten ergibt dann die gewünschte Ungleichung (11.4.2).

Ist nun $B \in \mathcal{M}(k, m)$, so gilt für $x \in B(0, 1)$ und $y := Ax$ wegen einer zweimaligen Anwendung von (11.4.2)

$$\|BAx\| = \|By\| \leq \|B\| \cdot \|y\| = \|B\| \cdot \|Ax\| \leq \|B\| \cdot \|A\| \cdot \|x\| \leq \|A\| \cdot \|B\| .$$

Mit anderen Worten haben wir

$$\|BA\| \leq \|B\| \cdot \|A\| . \tag{11.4.3}$$

Der folgende Satz liefert nun die mehrdimensionale Kettenregel.

Theorem 11.4.2

Seien $U \subset \mathbb{R}^n$ und $V \subset \mathbb{R}^m$ offen, sowie

$$f : U \to \mathbb{R}^m \qquad \text{und} \qquad g : V \to \mathbb{R}^k$$

und $x \in U$ mit $y := f(x) \in V$. Sind dann f in x und g in y Fréchet-differenzierbar, so ist $g \circ f$ in x Fréchet-differenzierbar und es gilt

$$J(g \circ f)(x) = J(g)(f(x)) \cdot J(f)(x)\,.$$

Man beachte, dass $J(g)(y) \in \mathcal{M}(k, m)$ und $J(f)(x) \in \mathcal{M}(m, n)$ gilt. Dies "passt" zu der Matrizenmultiplikation und zu $J(g \circ f)(x) \in \mathcal{M}(k, n)$.

Die mehrdimensionale Kettenregel wird häufig als Merksatz: "Äußere mal innere Ableitung" formuliert. Anders als im eindimensionalen Fall ist hierbei die Reihenfolge im Produkt wichtig!

BEWEIS 1/3

Wegen der angenommenen Fréchet-Differenzierbarkeit haben wir für $\xi \in \mathbb{R}^n$ und $\eta \in \mathbb{R}^m$:

$$f(x+\xi) = f(x) + J(f)(x) \cdot \xi + r_1(\xi)\,,$$
$$g(y+\eta) = g(y) + J(g)(y) \cdot \eta + r_2(\eta)\,,$$

wobei die Rest-Terme $r_1(\xi)$ und $r_2(\eta)$ die Bedingung (11.2.3) erfüllen. Damit haben wir für $\eta := J(f)(x) \cdot \xi + r_1(\xi)$:

$$\begin{aligned} g \circ f(x+\xi) &= g(f(x+\xi)) \\ &= g\big(f(x) + J(f)(x) \cdot \xi + r_1(\xi)\big) \\ &= g(y+\eta) \\ &= g(y) + J(g)(y) \cdot \eta + r_2(\eta) \\ &= g(f(x)) + J(g)(y) \cdot \Big(J(f)(x) \cdot \xi + r_1(\xi)\Big) + r_2(\eta) \\ &= g \circ f(x) + J(g)(y) \cdot J(f)(x) \cdot \xi + r_3(\xi)\,, \end{aligned}$$

wobei wir

$$r_3(\xi) := J(g)(f(x)) \cdot r_1(\xi) + r_2\big(J(f)(x) \cdot \xi + r_1(\xi)\big)$$

gesetzt haben. Es bleibt dann noch $\frac{r_3(\xi)}{\|\xi\|} \to 0$ für $\xi \to 0$ zu zeigen.

BEWEIS 2/3

Dafür bemerken wir zunächst, dass mit (11.4.2) folgt:

$$\frac{\|J(g)(f(x)) \cdot r_1(\xi)\|}{\|\xi\|} \leq \frac{\|J(g)(f(x))\| \cdot \|r_1(\xi)\|}{\|\xi\|} \to 0\,,$$

da r_1 die Bedingung (11.2.3) erfüllt.

Da r_1 die Bedingung (11.2.3) erfüllt, gibt es Konstanten $K > 0$ und $\delta \in (0, 1]$ mit

$$\|r_1(\xi)\| \leq K \cdot \|\xi\| \tag{11.4.4}$$

für alle $\xi \in U(0, \delta)$. Wir setzen nun $\psi(\eta) := r_2(\eta)/\|\eta\|$, d.h. wir haben

$$r_2(\eta) = \|\eta\| \cdot \psi(\eta)$$

und $\psi(\eta) \to 0$ für $\eta \to 0$. Damit erhalten wir

$$r_2\big(J(f)(x) \cdot \xi + r_1(\xi)\big) = \big\|J(f)(x) \cdot \xi + r_1(\xi)\big\| \cdot \big\|\psi\big(J(f)(x) \cdot \xi + r_1(\xi)\big)\big\|\,.$$

BEWEIS 3/3

Nun gilt mit (11.4.2) und (11.4.4) für $\xi \in U(0, \delta)$:

$$\frac{\left\| J(f)(x) \cdot \xi + r_1(\xi) \right\|}{\|\xi\|} \leq \frac{\|J(f)(x)\| \cdot \|\xi\| + \|r_1(\xi)\|}{\|\xi\|} \leq \|J(f)(x)\| + K\,.$$

Ferner gilt für $\eta_\xi := J(f)(x) \cdot \xi + r_1(\xi)$:

$$\|\eta_\xi\| = \|J(f)(x) \cdot \xi + r_1(\xi)\| \leq \|J(f)(x)\| \cdot \|\xi\| + \|r_1(\xi)\| \to 0$$

für $\xi \to 0$. Damit folgt $\|\psi(\eta_\xi)\| \to 0$ für $\xi \to 0$. Insgesamt haben wir damit $\frac{r_3(\xi)}{\|\xi\|} \to 0$ für $\xi \to 0$ gezeigt.

Ist $h := g \circ f$ so wird die Kettenregel häufig auch durch die intuitive Formel

$$\frac{\partial h_k}{\partial x_j} = \sum_{l=1}^{m} \frac{\partial g_k}{\partial y_l} \frac{\partial f_l}{\partial x_j}$$

beschrieben. Diese drückt aus, wie sich die einzelnen Einträge von $J(g \circ f)$ berechnen lassen.

BEISPIEL

Für ein erstes Beispiel betrachten wir die Funktionen $f : \mathbb{R} \to \mathbb{R}^2$ und $g : \mathbb{R}^2 \to \mathbb{R}^2$, die durch

$$f(x) := \begin{pmatrix} 0 \\ x^3 \end{pmatrix} \qquad \text{und} \qquad g(y_1, y_2) := \begin{pmatrix} y_1 \\ y_1^2 + y_2^2 \end{pmatrix}$$

gegeben sind. Wir haben dann

$$J(f)(x) = \begin{pmatrix} 0 \\ 3x^2 \end{pmatrix} \qquad \text{und} \qquad J(g)(y_1, y_2) := \begin{pmatrix} 1 & 0 \\ 2y_1 & 2y_2 \end{pmatrix}.$$

Da für $y := f(x)$ gilt: $y_1 = 0$ und $y_2 = x^3$, folgt

$$J(g)(f(x)) = \begin{pmatrix} 1 & 0 \\ 0 & 2x^3 \end{pmatrix}.$$

Damit haben wir insgesamt

$$J(g \circ f)(x) = J(g)(f(x)) \cdot J(f)(x) = \begin{pmatrix} 1 & 0 \\ 0 & 2x^3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 3x^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 6x^5 \end{pmatrix}.$$

Im Folgenden haben wir $U, V \subset \mathbb{R}^n$ offen und $\Phi : U \to V$ Fréchet-differenzierbar und bijektiv. Aus der Bijektivität folgt dann

$$\Phi^{-1} \circ \Phi(u) = u\,, \qquad\qquad u \in U.$$

Ist dann Φ^{-1} auch Fréchet-differenzierbar, so liefert die Kettenregel

$$E_n = J(\mathrm{id}_U)(u) = J(\Phi^{-1} \circ \Phi)(u) = J(\Phi^{-1})(\Phi(u)) \cdot J(\Phi)(u)$$

und durch Betrachten von $\Phi \circ \Phi^{-1}(v) = v$ erhalten wir mit $u := \Phi^{-1}(v)$ auch

$$\begin{aligned} E_n = J(\mathrm{id}_V)(v) = J(\Phi \circ \Phi^{-1})(v) &= J(\Phi)(\Phi^{-1}(v)) \cdot J(\Phi^{-1})(v) \\ &= J(\Phi)(u) \cdot J(\Phi^{-1})(\Phi(u)) \end{aligned}$$

Für $x := \Phi(u)$ zeigen diese Rechnungen

$$J(\Phi^{-1})(x) = \big(J(\Phi)(u)\big)^{-1}. \tag{11.4.5}$$

Wir nehmen nun an, wir haben eine Fréchet-differenzierbare Funktion $f : U \to \mathbb{R}^m$ und eine Funktion $\tilde{f} : V \to \mathbb{R}^m$ mit

$$f(u) = \tilde{f}(\Phi(u)), \qquad u \in U.$$

Mit anderen Worten ist $\tilde{f}$ die Funktion, die das Verhalten von f ausdrückt, wenn wir einen Koordinatenwechsel $u \mapsto \Phi(u)$ vorgenommen haben. Offensichtlich folgt aus $f = \tilde{f} \circ \Phi$ sofort $\tilde{f} = f \circ \Phi^{-1}$. Wir wollen nun die Ableitungen von f und $\tilde{f}$ vergleichen. Mit $x := \Phi(u)$ folgt mit aus der Kettenregel

$$\begin{aligned} J(f)(u) = J(\tilde{f} \circ \Phi)(u) &= J(\tilde{f})(\Phi(u)) \cdot J(\Phi)(u) \\ &= J(\tilde{f})(x) \cdot J(\Phi)(u). \end{aligned}$$

Mit (11.4.5) folgt dann

$$J(\tilde{f})(x) = J(f)(u) \cdot \big(J(\Phi)(u)\big)^{-1} = J(f)(u) \cdot J(\Phi^{-1})(x)$$

und Ausnutzen von $x := \Phi(u)$ führt zu

$$J(\tilde{f})(\Phi(u)) = J(f)(u) \cdot J(\Phi^{-1})(\Phi(u))\,. \tag{11.4.6}$$

Um dies an einem konkreten Beispiel zu verdeutlichen, betrachten wir $\Phi : (0,\infty) \times (-\pi,\pi) \to \mathbb{R}^2 \setminus \{0\}$ mit

$$\Phi(r,\varphi) := \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix}$$

für $r > 0$ und $\varphi \in (-\pi,\pi)$. Mit anderen Worten überführt Φ Polar-Koordinaten in kartesische Koordinaten. Wir setzen

$$U := (0,\infty) \times (-\pi,\pi)\,,$$
$$V := \Phi(U) = \mathbb{R}^2 \setminus (-\infty,0]e_1\,.$$

Wir haben dann für $u = (r, \varphi)$:

$$J(\Phi)(u) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}$$

und gemäß der Formel

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

und (11.4.5) dann auch

$$J(\Phi^{-1})(\Phi(u)) = \big(J(\Phi)(u)\big)^{-1} = \frac{1}{r\cos^2\varphi + r\sin^2\varphi} \begin{pmatrix} r\cos\varphi & r\sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$$
$$= \begin{pmatrix} \cos\varphi & \sin\varphi \\ -r^{-1}\sin\varphi & r^{-1}\cos\varphi \end{pmatrix}$$

Die Formel (11.4.6) wird dann zu

$$J(\tilde{f})\big((r\cos\varphi, r\sin\varphi)^{\mathsf{T}}\big) = J(f)(r,\varphi) \cdot \begin{pmatrix} \cos\varphi & \sin\varphi \\ -r^{-1}\sin\varphi & r^{-1}\cos\varphi \end{pmatrix}.$$

Section 11.5
Satz von Taylor

MOTIVATION

Der eindimensionale Satz von Taylor, siehe Satz 8.1.2, ermöglichtete es, hinreichend glatte Funktionen lokal durch Polynome zu approximieren. In diesem Kapitel wollen wir dies auf Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ verallgemeinern.

Zunächst benötigen wir hierzu den folgenden, mehrdimensionalen Mittelwertsatz, der auch für sich genommen von Interesse ist.

Theorem 11.5.1

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ Fréchet-differenzierbar. Ferner seien $a, b \in U$ mit

$$[a, b] := \{a + t(b - a) : t \in [0, 1]\} \subset U\,. \tag{11.5.1}$$

Dann gibt es ein $\theta \in (0, 1)$ mit

$$f(b) - f(a) = \langle \nabla f(a + \theta(b - a)), b - a \rangle\,.$$

Die Bedingung (11.5.1) besagt, dass die Strecke zwischen a und b komplett in U liegt. Ist $U \subset \mathbb{R}^n$ **konvex**, d.h. wir haben $[a, b] \subset U$ für alle $a, b \in U$, so ist dies "automatisch" erfüllt. Hierbei ist zu beachten, dass die offenen und abgeschlossenen Kugeln bzgl. einer Norm aufgrund der Dreiecksungleichung und der Homogenität immer konvex sind.

Der Satz 11.5.1 zeigt unter der Bedingung $[a, b] \subset U$, dass es auf dies Strecke $[a, b]$ einen Punkt $\xi := a + \theta(b - a)$ gibt mit

$$f(b) - f(a) = \langle \nabla f(\xi), b - a \rangle .$$

Im Fall $n = 1$ ist dies exakt die Aussage des Mittelwertsatzes 6.2.3.

BEWEIS

Wir betrachten die Funktion $h : [0,1] \to \mathbb{R}$, die durch

$$h(t) := f(a + t(b-a))$$

definiert ist. Da die Abbildung $\varphi : [0,1] \to \mathbb{R}^n$, die durch $\varphi(t) := a + t(b-a)$ definiert ist, Fréchet-differenzierbar ist, zeigt die Kettenregel, siehe Satz 11.4.2, dass h differenzierbar ist. Ferner ist h auch stetig nach Lemma 11.2.4. Der Mittelwertsatzes 6.2.3 liefert daher ein $\theta \in (0,1)$ mit

$$h(1) - h(0) = h'(\theta) \cdot (1-0)\,.$$

Wegen $h(1) = f(b)$ und $h(0) = f(a)$ folgt dann

$$f(b) - f(a) = h'(\theta)\,.$$

Ferner gilt mit der Kettenregel und $h = f \circ \varphi$:

$$h'(\theta) = J(f)(\varphi(\theta)) \cdot J(\varphi)(\theta) = \big(\mathbf{grad}\, f(a + \theta(b-a))\big) \cdot (b-a)\,.$$

Mit $(\nabla f)^{\mathsf{T}} = \mathbf{grad}\, f$ folgt dann die Behauptung.

BEMERKUNG

Für Funktionen $f : U \to \mathbb{R}^m$ mit $m > 1$ ist keine zu Satz 11.5.1 analoge Aussage möglich. Um dies zu illustrieren, betrachten wir die Abbildung $f : \mathbb{R} \to \mathbb{R}^2$, die durch

$$f(t) := \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$

definiert ist. Wir setzen ferner $a := 0$ und $b := 2\pi$. Wir haben dann $f(b) - f(a) = 0$ und

$$J(f)\big(a + \theta(b-a)\big) = J(f)(2\pi\theta) = \begin{pmatrix} -\sin(2\pi\theta) \\ \cos(2\pi\theta) \end{pmatrix}.$$

Dies ergibt

$$J(f)\big(a + \theta(b-a)\big) \cdot (b-a) = 2\pi \begin{pmatrix} -\sin(2\pi\theta) \\ \cos(2\pi\theta) \end{pmatrix}.$$

Da es aber kein $\theta \in \mathbb{R}$ mit $\sin(2\pi\theta) = 0$ und $\cos(2\pi\theta) = 0$ gibt, ist die Gleichung

$$f(b) - f(a) = J(f)\big(a + \theta(b-a)\big) \cdot (b-a)$$

unmöglich.

Im folgenden wollen wir versuchen, eine abgeschwächte Fassung des Mittelwertsatzes im Fall $m > 1$ zu bekommen. Dazu sei $f : [a, b] \to \mathbb{R}^n$ eine stetige Funktion. Ihr Integral wird dann komponentenweise erklärt, d.h.

$$\int_a^b f(t)\,\mathrm{d}t := \begin{pmatrix} \int_a^b f_1(t)\,\mathrm{d}t \\ \vdots \\ \int_a^b f_m(t)\,\mathrm{d}t \end{pmatrix}, \tag{11.5.2}$$

wobei $f_1, \ldots, f_m$ die Komponentenfunktionen von f sind.

Das folgende Lemma zeigt, dass Normen in solche vektorwertigen Integrale reingezogen werden können. Es verallgemeinert damit die Abschätzung (7.1.13) von Satz 7.1.8.

Lemma 11.5.2

Sei $f : [a, b] \to \mathbb{R}^n$ stetig. Dann gilt

$$\left\| \int_a^b f(t)\,\mathrm{d}t \right\| \leq \int_a^b \|f(t)\|\,\mathrm{d}t\,.$$

Wir setzen

$$v := \int_a^b f(t)\,\mathrm{d}t$$

und $K := \|v\|$ und nehmen ohne Einschränkung $K \neq 0$ an. Dann gilt mit der Linearität des Integrals und der Definition des Skalarproduktes

$$\begin{aligned} K^2 = \langle v, v\rangle = \left\langle \int_a^b f(t)\,\mathrm{d}t, v \right\rangle &= \int_a^b \langle f(t), v\rangle\,\mathrm{d}t \\ &\leq \int_a^b \|f(t)\| \cdot \|v\|\,\mathrm{d}t \\ &= K \int_a^b \|f(t)\|\,\mathrm{d}t\,, \end{aligned}$$

wobei wir für die Ungleichung die Cauchy-Schwarz-Ungleichung benutzt haben. Teilen wir nun durch K, so erhalten wir die Behauptung.

Der folgende Satz liefert nun die angekündigte, abgeschwächte Version des Mittelwertsatzes im Fall $m > 1$.

Theorem 11.5.3

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^m$ Fréchet-differenzierbar. Ferner seien $a, b \in U$ mit $[a, b] \subset U$. Dann gilt

$$f(b) - f(a) = \int_0^1 J(f)\big(a + t(b-a)\big) \cdot (b-a)\,\mathrm{d}t$$

und

$$\|f(b) - f(a)\| \leq \int_0^1 \big\|J(f)\big(a + t(b-a)\big)\big\|\,\mathrm{d}t \cdot \|b-a\|\,,$$

wobei das erste Integral als (11.5.2) *zu verstehen ist.*

Seien $f_1, \ldots, f_m$ die Komponentenfunktionen von f. Mit dem Hauptsatz 7.2.1 und der mehrdimensionalen Kettenregel, siehe Satz 11.4.2, gilt dann

$$\begin{aligned} f_i(b) - f_i(a) &= \int_0^1 \frac{\partial f_i(a + t(b-a))}{\partial t} \, \mathrm{d}t \\ &= \int_0^1 \mathbf{grad}\, f_i\big(a + t(b-a)\big) \cdot (b-a) \, \mathrm{d}t \,. \end{aligned}$$

Zusammensetzen über $i = 1, \ldots, m$ ergibt dann die erste Behauptung. Die zweite folgt dann mit Lemma 11.5.2 und der allgemeinen Abschätzung (11.4.2) für Matrix-Normen.

Wir müssen zunächst ein paar neue Notationen einführen. Dazu sei $\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ ein **Multi-Index**. Wir definieren dann

$$|\alpha| := \alpha_1 + \cdots + \alpha_n \,,$$
$$\alpha! := \alpha_1! \cdot \ldots \cdot \alpha_n! \,.$$

Ist ferner $f : \mathbb{R}^n \to \mathbb{R}$ α-mal stetig differenzierbar, so schreiben wir

$$D_i^{\alpha_i} f := \frac{\partial^{\alpha_i} f}{\partial^{\alpha_i} x_i}$$

für die α_i-malige Ableitung in Richtung i und

$$D^\alpha f := D_1^{\alpha_1} \cdots D_n^{\alpha_n} f$$

für die $|\alpha|$-malige partielle Ableitung in die Richtungen $i = 1, \ldots, n$ mit Vielfachheiten $\alpha_1, \ldots, \alpha_n$. Schließlich schreiben wir für $x = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n$:

$$x^\alpha := x_1^{\alpha_1} \cdot \ldots \cdot x_n^{\alpha_n}$$

für ein multidimensionales Monom.

Wir beginnen mit dem folgenden Satz, der höhere Ableitungen in eine Richtung berechnet.

Theorem 11.5.4

Sei $U \subset \mathbb{R}^n$ offen, $k \geq 1$ und $f : U \to \mathbb{R}$ eine k-mal stetig differenzierbare Funktion. Ferner seien $x \in U$ und $v \in \mathbb{R}^n$ mit $[x, x+v] \subset U$. Für

$$\begin{aligned} g : [0,1] &\to \mathbb{R} \\ t &\mapsto f(x+tv) \end{aligned}$$

gilt dann

$$g^{(k)}(t) = \sum_{\alpha \in \mathbb{N}_0^n : |\alpha| = k} \frac{k!}{\alpha!} D^\alpha f(x+tv) \cdot v^\alpha \,, \qquad t \in [0,1].$$

BEWEIS 1/3

Wir zeigen zunächst per Induktion über k, dass

$$g^{(k)}(t) = \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} D_{i_k} \cdots D_{i_1} f(x + tv) \cdot v_{i_1} \cdots v_{i_k} \tag{11.5.3}$$

gilt. Dazu sei wieder $\varphi(t) := x + tv$, d.h., wir haben $g = f \circ \varphi$. Für $k = 1$ ergibt nun die Kettenregel

$$\begin{aligned} g'(t) = J(f)(\varphi(t)) \cdot J(\varphi)(t) &= \operatorname{\mathbf{grad}} f(x + tv) \cdot v \\ &= \sum_{i=1}^{n} D_i f(x + tv) \cdot v_i \,. \end{aligned}$$

Alternativ können wir auch Satz 11.2.8 anwenden.

BEWEIS 2/3

Für den Induktionsschritt nehmen wir nun an, dass die Aussage schon für $k-1$ gezeigt wurde. Es gilt dann

$$\begin{aligned} g^{(k)}(t) &= \frac{\partial g^{(k-1)}}{\partial t}(t) \\ &= \frac{\partial}{\partial t}\left(\sum_{i_1=1}^{n} \cdots \sum_{i_{k-1}=1}^{n} D_{i_{k-1}} \cdots D_{i_1} f(x+tv) \cdot v_{i_1} \cdots v_{i_{k-1}}\right) \\ &= \sum_{j=1}^{n} D_j \left(\sum_{i_1=1}^{n} \cdots \sum_{i_{k-1}=1}^{n} D_{i_{k-1}} \cdots D_{i_1} f(x+tv) \cdot v_{i_1} \cdots v_{i_{k-1}}\right) v_j \,, \end{aligned}$$

wobei wir analog zum Fall $k=1$ die Kettenregel angewendet haben. Setzen wir $i_k := j$ und wenden wir die Linearität von D_j an, so ergibt sich (11.5.3).

BEWEIS 3/3

Die Behauptung folgt nun durch geschicktes Zusammenfassen der Summen. Hierzu sei $i_1, \ldots, i_k \in \{1, \ldots, n\}$. Kommt unter diesen k Zahlen, die Zahl 1 genau α_1-mal vor, die Zahl 2 genau α_2-mal vor, ...und die Zahl n genau α_n-mal vor, so folgt aus dem Satz von Schwarz, siehe Satz 11.1.4:

$$D_{i_k} \cdots D_{i_1} f(x + tv) \cdot v_{i_1} \cdots v_{i_k} = D^\alpha f(x + tv) v_1^{\alpha_1} \cdots v_n^{\alpha_n} \, .$$

Eine solche Kombination kommt genau

$$\frac{k!}{\alpha_1! \cdots \alpha_n!}$$

mal vor, wobei für $n = 2$ dies schon im Satz 2.2.5 gezeigt wurde und die allgemeine Aussage auch elementar bewiesen werden kann. Mit unseren Notationen für v^α und $\alpha!$ folgt dann die Behauptung.

Mit diesen Vorbereitungen können wir nun die mehrdimensionale Version der Taylor-Entwicklung mit Lagrange-Restglied beschreiben.

Theorem 11.5.5

Sei $U \subset \mathbb{R}^n$ offen, $k \geq 1$ und $f : U \to \mathbb{R}$ eine $(k+1)$-mal stetig differenzierbare Funktion. Ferner seien $x \in U$ und $\xi \in \mathbb{R}^n$ mit $[x, x+\xi] \subset U$. Dann existiert ein $\theta \in [0,1]$ mit

$$f(x+\xi) = \sum_{\alpha \in \mathbb{N}_0^n : |\alpha| \leq k} \frac{D^\alpha f(x)}{\alpha!} \xi^\alpha + \sum_{\alpha \in \mathbb{N}_0^n : |\alpha| = k+1} \frac{D^\alpha f(x+\theta\xi)}{\alpha!} \xi^\alpha .$$

Man beachte, dass alle Terme in der ersten Summe mehr-dimensionale Monome vom Grad $|\alpha|$ sind. Insgesamt ist daher die erste Summe ein mehr-dimensionales Polynom vom Grad k.

Wir betrachten wieder die Funktion

$$\begin{aligned} g : [0,1] &\to \mathbb{R} \\ t &\mapsto f(x + t\xi)\,. \end{aligned}$$

Mit dem Lagrange-Restglied, siehe Satz 8.1.3 für $t_0 = 0$ und $t = 1$, haben wir dann ein $\theta \in [t_0, t] = [0,1]$ mit

$$\begin{aligned} f(x+\xi) &= g(1) \\ &= \sum_{m=0}^{k} \frac{g^{(m)}(0)}{m!}(1-0)^m + \frac{g^{(k+1)}(\theta)}{(k+1)!}(1-0)^{k+1} \\ &= \sum_{m=0}^{k} \frac{g^{(m)}(0)}{m!} + \frac{g^{(k+1)}(\theta)}{(k+1)!}\,. \end{aligned}$$

Setzen wir jetzt die Formel aus Satz 11.5.4 ein, so folgt die Behauptung.

Der mehrdimensionale Satz von Taylor folgt nun als Korollar.

Korollar 11.5.6

Sei $U \subset \mathbb{R}^n$ offen, $k \geq 1$ und $f : U \to \mathbb{R}$ eine k-mal stetig differenzierbare Funktion. Ferner seien $x \in U$ und $\delta > 0$ mit $U(x, \delta) \subset U$. Dann gibt es eine Funktion $r : U(x, \delta) \to \mathbb{R}$ mit

$$f(x + \xi) = \sum_{\alpha \in \mathbb{N}_0^n : |\alpha| \leq k} \frac{D^\alpha f(x)}{\alpha!} \xi^\alpha + r(\xi) \tag{11.5.4}$$

für alle $\xi \in U(0, \delta)$, sowie $r(0) = 0$ und

$$\lim_{\xi \to 0} \frac{r(\xi)}{\|\xi\|^k} = 0 \,. \tag{11.5.5}$$

Wir wenden Satz 11.5.5 für $k-1$ an und erhalten ein $\theta_\xi \in [0,1]$, so dass gilt

$$\begin{aligned} f(x+\xi) &= \sum_{\alpha\in\mathbb{N}_0^n:|\alpha|\le k-1} \frac{D^\alpha f(x)}{\alpha!}\xi^\alpha + \sum_{\alpha\in\mathbb{N}_0^n:|\alpha|=k} \frac{D^\alpha f(x+\theta_\xi\xi)}{\alpha!}\xi^\alpha \\ &= \sum_{\alpha\in\mathbb{N}_0^n:|\alpha|\le k} \frac{D^\alpha f(x)}{\alpha!}\xi^\alpha + \sum_{\alpha\in\mathbb{N}_0^n:|\alpha|=k} \frac{D^\alpha f(x+\theta_\xi\xi)-D^\alpha f(x)}{\alpha!}\xi^\alpha . \end{aligned}$$

Wir setzen daher

$$r_\alpha(\xi) := \frac{D^\alpha f(x+\theta_\xi\xi)-D^\alpha f(x)}{\alpha!}$$

für $\alpha\in\mathbb{N}_0^n$ mit $|\alpha|=k$. Da nach Voraussetzung alle $D^\alpha f$ stetig sind, folgt $r_\alpha(\xi)\to 0$ für $\xi\to 0$. Setzen wir nun

$$r(\xi) := \sum_{\alpha\in\mathbb{N}_0^n:|\alpha|=k} r_\alpha(\xi)\cdot\xi^\alpha ,$$

so folgt sofort (11.5.4) und $r(\xi)=0$. Ferner haben wir

$$\frac{\xi^\alpha}{\|\xi\|^k} = \frac{\xi_1^{\alpha_1}\cdots\xi_n^{\alpha_n}}{\|\xi\|^{\alpha_1}\cdots\|\xi\|^{\alpha_n}} \le 1$$

und damit zeigt $r_\alpha(\xi)\to 0$ für $\xi\to 0$ auch schon (11.5.5).

DIE TERME BIS ZUR ORDNUNG 2

Im folgenden wollen wir die einzelnen Summen

$$\sum_{\alpha\in\mathbb{N}_0^n:|\alpha|=k} \frac{D^\alpha f(x)}{\alpha!}\xi^\alpha$$

in (11.5.4) für $k = 0, 1, 2$ genauer betrachten.

Im Fall $k = 0$ ist dies einfach, da es nur ein $\alpha \in \mathbb{N}_0^n$ gibt mit $|\alpha| = 0$, nämlich $\alpha = 0$. Dies ergibt

$$\sum_{\alpha\in\mathbb{N}_0^n:|\alpha|=0} \frac{D^\alpha f(x)}{\alpha!}\xi^\alpha = \frac{D^0 f(x)}{0!}\xi^0 = f(x)\,. \tag{11.5.6}$$

Der Fall $k = 1$ ist ebenfalls noch sehr übersichtlich, da nur die Tupel $\alpha = e_i \in \mathbb{N}_0^d$ für $i = 1, \ldots, n$ die Gleichung $|\alpha| = 1$ erfüllen. Da wir für $\alpha = e_i$ die Identitäten $D^\alpha = D_i$, $e_i! = 1$ und $\xi^{e_i} = \xi_i$ haben, folgt dann

$$\sum_{\alpha\in\mathbb{N}_0^n:|\alpha|=1} \frac{D^\alpha f(x)}{\alpha!}\xi^\alpha = \sum_{i=1}^n D_i f(x)\xi_i = \langle \nabla f(x), \xi\rangle\,. \tag{11.5.7}$$

DIE TERME BIS ZUR ORDNUNG 2

Im Fall $k = 2$ gibt es zwei Sorten von $\alpha \in \mathbb{N}_0^n$ mit $|\alpha| = 2$, nämlich $\alpha = 2e_i$ und $\alpha = e_i + e_j$, wobei wir bei letzteren ohne Einschränkung $i < j$ annehmen können, da dies nicht den Multi-Index α verändert. Im ersten Fall gilt nun

$$D^{2e_i} f = D_i^2 f\,, \qquad (2e_i)! = 2\,, \qquad \xi^{2e_i} = \xi_i^2\,,$$

während wir im zweiten Fall

$$D^{e_i+e_j} f = D_i D_j f\,, \qquad (e_i + e_j)! = 1\,, \qquad \xi^{e_i+e_j} = \xi_i \cdot \xi_j$$

haben. Dies ergibt

$$\begin{aligned}
\sum_{\alpha\in\mathbb{N}_0^n:|\alpha|=2} \frac{D^\alpha f(x)}{\alpha!}\xi^\alpha &= \frac{1}{2}\sum_{i=1}^n D_i^2 f(x)\xi_i^2 + \sum_{i<j} D_i D_j f(x)\xi_i\xi_j \\
&= \frac{1}{2}\sum_{i=1}^n D_i^2 f(x)\xi_i^2 + \frac{1}{2}\sum_{i\neq j} D_i D_j f(x)\xi_i\xi_j \\
&= \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n D_i D_j f(x)\xi_i\xi_j \\
&= \frac{1}{2}\langle \xi, H(f)(x)\cdot\xi\rangle\,.
\end{aligned}$$

Damit haben wir das folgende Korollar gezeigt.

Korollar 11.5.7

Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion. Ferner seien $x \in U$ und $\delta > 0$ mit $U(x, \delta) \subset U$. Dann gibt es eine Funktion $r : U(x, \delta) \to \mathbb{R}$ mit

$$f(x + \xi) = f(x) + \langle \nabla f(x), \xi \rangle + \frac{1}{2} \langle \xi, H(f)(x) \cdot \xi \rangle + r(\xi) \tag{11.5.8}$$

für alle $\xi \in U(0, \delta)$, wobei r sowohl $r(0) = 0$ als auch (11.5.5) für $k = 2$ erfüllt.

BEISPIEL

Wir wollen das letzte Korollar jetzt noch an einem Beispiel illustrieren. Dazu sei $f : \mathbb{R}^3 \to \mathbb{R}$ durch

$$f(x_1, x_2, x_3) := x_1 x_2^2 \sin x_3$$

definiert. Ferner interessieren wir uns für den Punkt $x := (1, 2, 0)^T$, wobei wir zunächst beachten, dass

$$f(1, 2, 0) = 0 \,.$$

Um nun die Entwicklung (11.5.8) zu berechnen, berechnen wir noch alle Ableitungen der Ordnungen 1 und 2. Dies ergibt

Für die Ableitungen der ersten Ordnung ergibt dies

$$\begin{aligned}
\frac{\partial f}{\partial x_1}(1,2,0) &= x_2^2 \sin x_3\big|_{(1,2,0)} = 0\,,\\
\frac{\partial f}{\partial x_2}(1,2,0) &= 2x_1x_2 \sin x_3\big|_{(1,2,0)} = 0\,,\\
\frac{\partial f}{\partial x_3}(1,2,0) &= x_1x_2^2 \cos x_3\big|_{(1,2,0)} = 4\,,
\end{aligned}$$

und damit

$$\big\langle \nabla f(1,2,0), \xi \big\rangle = 4\xi_3\,.$$

Für die Ableitungen der zweiten Ordnung haben wir ferner

$$\begin{aligned}
\frac{\partial^2 f}{\partial^2 x_1}(1,2,0) &= 0\big|_{(1,2,0)} = 0\,,\\
\frac{\partial^2 f}{\partial x_2 \partial x_1}(1,2,0) &= 2x_2 \sin x_3\big|_{(1,2,0)} = 0\,,\\
\frac{\partial^2 f}{\partial x_3 \partial x_1}(1,2,0) &= x_2^2 \cos x_3\big|_{(1,2,0)} = 4\,,\\
\frac{\partial^2 f}{\partial^2 x_2}(1,2,0) &= 2x_1 \sin x_3\big|_{(1,2,0)} = 0\,,\\
\frac{\partial^2 f}{\partial x_3 \partial x_2}(1,2,0) &= 2x_1 x_2 \cos x_3\big|_{(1,2,0)} = 4\,,\\
\frac{\partial^2 f}{\partial^2 x_3}(1,2,0) &= -x_1 x_2^2 \sin x_3\big|_{(1,2,0)} = 0\,.
\end{aligned}$$

Dies ergibt

$$H(f)(1,2,0) = \begin{pmatrix} 0 & 0 & 4 \\ 0 & 0 & 4 \\ 4 & 4 & 0 \end{pmatrix}$$

und damit

$$\left\langle \xi, H(f)(1,2,0) \cdot \xi \right\rangle = \left\langle \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}, \begin{pmatrix} 4\xi_3 \\ 4\xi_3 \\ 4\xi_1 + 4\xi_2 \end{pmatrix} \right\rangle = 8\left(\xi_1\xi_3 + \xi_2\xi_3\right).$$

Insgesamt ergibt dies

$$f\big((1,2,0) + \xi\big) = 4\xi_3 + 4(\xi_1\xi_3 + \xi_2\xi_3) + r(\xi).$$

Section 11.6
Mehrdimensionale Extrema

Für Funktionen $\mathbb{R} \to \mathbb{R}$ können Maxima und Minima über die ersten und zweiten Ableitungen detektiert werden. In diesem Kapitel wollen wir dies auf Funktionen $\mathbb{R}^n \to \mathbb{R}$ ausdehnen. Wir beginnen mit der Definition von (lokalen) Extrema.

Definition 11.6.1

Sei $U \subset \mathbb{R}^n$ offen, $f : U \to \mathbb{R}$ und $x^* \in U$. Dann hat f in x^* ein **lokales Maximum**, falls es ein $\delta > 0$ gibt mit $U(x^*, \delta) \subset U$ und

$$f(x) \leq f(x^*), \qquad x \in U(x^*, \delta).$$

Gilt sogar $f(x) < f(x^*)$ für alle $x \in U(x^*, \delta) \setminus \{0\}$, so sprechen wir von einem **strikten lokalem Maximum**.

Wie üblich hat f ein **(striktes) lokales Minimum** in x^*, falls $-f$ ein (striktes) lokales Maximum hat. Schließlich hat f in x^* ein **(striktes) lokales Extremum**, falls f in x^* ein (striktes) lokales Maximum oder Minimum hat.

Im eindimensionalen Fall verschwindet die Ableitung in einem Extremum. Der folgende Satz zeigt, dass dies auch im mehrdimensionalen Fall gilt.

Theorem 11.6.2

Sei $U \subset \mathbb{R}^n$ offen, $f : U \to \mathbb{R}$ partiell differenzierbar und $x^* \in U$, so dass f in x^* ein lokales Extremum hat. Dann gilt

$$\operatorname{grad} f(x^*) = 0\,.$$

Wir fixieren ein $i \in \{1, \dots, n\}$ und $\delta > 0$ gemäß der Definition von lokalen Extrema. Ferner definieren wir $h : (-\delta, \delta) \to \mathbb{R}$ durch

$$h(t) := f(x^* + te_i)\,, \qquad t \in (-\delta, \delta).$$

Dann hat h in 0 ein Extremum und ist nach Konstruktion auch differenzierbar. Mit dem eindimensionalen Fall, siehe Satz 6.2.2 folgt

$$0 = h'(0) = \frac{\partial f}{\partial x_i}(x^*)\,.$$

Dies zeigt die Behauptung.

Da die Funktion $x \mapsto \mathbf{grad}\, f(x)$ im Allgemeinen nichtlinear ist, ist die Bedingung

$$\mathbf{grad}\, f(x^*) = 0$$

typischerweise ein *nicht-lineares* Gleichungssystem. Dieses hat oft keine explizite Lösung sondern muss numerisch gelöst werden. Hierbei kann n sehr groß sein: Neuronale Netze auf einem Desktop-Computer haben leicht Millionen von Parametern und bei ihrem Training entspricht diese Anzahl gerade n. Ernsthafte neuronale Netze sind um einige Größenordnungen größer, z.B. hat ChatGPT (momentan) ca. 175 Milliarden Parameter!

Wie im eindimensionalen Fall ist $\mathbf{grad}\, f(x^*) = 0$ keine hinreichende Bedingung für ein Extremum. Um ein Beispiel für $n = 2$ zu sehen, betrachten wir die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$, die durch

$$f(x, y) := x^2 - y^2 \tag{11.6.1}$$

definiert ist. Es gilt dann

$$\mathbf{grad}\, f(x, y) = (2x, -2y)$$

und damit folgt $\mathbf{grad}\, f(0, 0) = 0$. Die Funktion f hat aber kein Extremum in 0, da gilt

$$f(x, 0) > f(0, 0) > f(0, y)$$

für alle $x, y \neq 0$.

Im eindimensionalen Fall hat $f : \mathbb{R} \to \mathbb{R}$ ein Extremum in x^*, falls $f'(x^*) = 0$ und $f''(x^*) \neq 0$ gilt, siehe Satz 6.2.8. Dies kann im mehrdimensionalen Fall so nicht gelten, da für die in (11.6.1) definierte Funktion f gilt

$$\operatorname{\mathbf{grad}} f(0) = 0 \qquad \text{und} \qquad H(f)(0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \neq 0$$

und wir gerade gesehen haben, dass f kein Extremum in 0 hat.

Der folgende Satz löst dieses Problem, indem er die Hesse-Matrix genauer betrachtet.

Theorem 11.6.3

Sei $U \subset \mathbb{R}^n$ offen, $f : U \to \mathbb{R}$ zweimal stetig partiell differenzierbar und $x^* \in U$ mit

$$\operatorname{grad} f(x^*) = 0 \,.$$

Dann gelten die folgenden Aussagen:

i). Hat f in x^* ein lokales Minimum, so ist $H(f)(x^*)$ positiv semi-definit.

ii). Ist $H(f)(x^*)$ positiv definit, so hat f in x^* ein striktes lokales Minimum.

Analoge Aussagen gelten für Maxima, indem wir die Funktion $-f$ betrachten. Da wir dann $H(-f)(x^*) = -H(f)(x^*)$ haben, wird die positive (semi)-Definitheit durch die negative (semi)-Definitheit ersetzt werden müssen.

Ist $H(f)(x^*)$ indefinit, so kann insbesondere f kein lokales Extremum in x^* haben.

Die Idee des Beweises beruht einerseits auf Korollar 11.5.7, das zeigt, dass das Verhalten von f in einer kleinen Umgebung von x^* durch die quadratischen Funktion $\xi \mapsto \langle \xi, H(f)(x^*) \cdot \xi \rangle$ dominiert wird und andererseits auf den Eigenschaften von symmetrischen Matrizen, die wir in Kapitel 5 herausgearbeitet hatten.

BEWEIS 1/4

Da $\mathbf{grad}\, f(x^*) = 0$ gilt, gibt es nach Korollar 11.5.7 ein $\delta > 0$ mit

$$f(x) - f(x^*) = \frac{1}{2}\langle x - x^*, A(x - x^*)\rangle + r(x - x^*)\,, \qquad x \in U(x^*, \delta),$$

wobei wir $A := H(f)(x^*)$ und $x := x^* + \xi$ gesetzt haben und $r : U(x^*, \delta) \to \mathbb{R}$ sowohl $r(0) = 0$ als auch

$$\lim_{\xi \to 0} \frac{r(x - x^*)}{\|x - x^*\|^2} = 0$$

erfüllt. Damit gibt es zu jedem $\varepsilon > 0$ ein $\delta_\varepsilon \in (0, \delta)$ mit

$$|r(x - x^*)| \leq \varepsilon \|x - x^*\|^2\,, \qquad x \in U(x^*, \delta_\varepsilon).$$

Nach dem Satz von Schwarz, siehe Satz 11.1.4, ist A symmetrisch. Satz 10.5.3 zeigt dann, dass es eine orthogonale Matrix $S \in \mathcal{M}(n, n)$ gibt mit

$$\Lambda := \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} = S^{\mathsf{T}} A S\,,$$

wobei $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ die n Eigenwerte von A sind.

BEWEIS 2/4

Da $S^{-1} = S^{\mathsf{T}}$ gilt, siehe (10.3.9), folgt $A = S\Lambda S^{\mathsf{T}}$. Wir setzen nun $y := x - x^*$ und $z := S^{\mathsf{T}} y$. Dann folgt

$$\begin{aligned}\left\langle x - x^*, A(x - x^*)\right\rangle = \langle y, S\Lambda S^{\mathsf{T}} y\rangle &= \langle S^{\mathsf{T}} y, \Lambda S^{\mathsf{T}} y\rangle \\ &= \langle z, \Lambda z\rangle \\ &= \sum_{i=1}^{n} \lambda_i z_i^2\end{aligned}$$

und da S^{T} orthogonal ist, zeigt Satz 10.3.8:

$$\|z\| = \|S^{\mathsf{T}} y\| = \|y\| = \|x - x^*\|\,.$$

BEWEIS 3/4

i). Wir nehmen an, es gibt ein $\lambda_i < 0$. Sei v_i ein Eigenvektor zu λ_i, wobei wir ohne Einschränkung $\|v_i\| = 1$ annehmen können. Für $\varepsilon := -\frac{\lambda_i}{3} > 0$ und $t \in (-\delta_\varepsilon, \delta_\varepsilon)$ folgt dann

$$\begin{aligned} f(x^* + tv_i) - f(x^*) = \frac{1}{2}\langle tv_i, A(tv_i)\rangle + r(tv_i) &= \frac{\lambda_i}{2}t^2 + r(tv_i) \\ &\leq \frac{\lambda_i}{2}t^2 + \varepsilon t^2 \\ &= \frac{\lambda_i}{6}t^2 \,. \end{aligned}$$

Durch Betrachten von $t := \delta_\varepsilon/2$ haben wir dann $x := x^* + tv_i \in U(x^*, \delta_\varepsilon) \subset U(x^*, \delta)$ und $f(x) - f(x^*) < 0$.

BEWEIS 4/4

ii). Da A positiv definit ist, gilt $\lambda_1, \ldots, \lambda_n > 0$, siehe Satz 10.5.5 und die Bemerkung danach. Ohne Einschränkung können wir $\lambda_n = \mathbf{min}\, \lambda_i$ annehmen. Für $\varepsilon := \frac{\lambda_n}{3}$ und $x \in U(x^*, \delta_\varepsilon)$ folgt dann

$$\begin{aligned} f(x) - f(x^*) &= \frac{1}{2}\langle x - x^*, A(x - x^*)\rangle + r(x - x^*) \\ &= \frac{1}{2}\sum_{i=1}^{n} \lambda_i z_i^2 + r(x - x^*) \\ &\geq \frac{\lambda_n}{2}\sum_{i=1}^{n} z_i^2 + r(x - x^*) \\ &\geq \frac{\lambda_n}{2}\|x - x^*\|^2 - \varepsilon\|x - x^*\|^2 \\ &= \frac{\lambda_n}{6}\|x - x^*\|^2 . \end{aligned}$$

Im Fall $x \neq x^*$ gilt daher $f(x) > f(x^*)$.

Wir beginnen mit der Funktion $f : \mathbb{R}^2 \to \mathbb{R}$, die durch

$$f(x, y) := x^2 + y^2$$

definiert ist. Es gilt dann

$$\mathbf{grad}\, f(x, y) = (2x, 2y)\,,$$

und damit ist $(x, y) = 0$ der einzige Kandidat für ein Extremum. Wegen

$$H(f)(0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

sehen wir dann, dass f ein striktes Minimum in 0 hat, was natürlich keine Überraschung ist.

Wir betrachten nun die Funktionen $f_\pm : \mathbb{R}^2 \to \mathbb{R}$, die durch

$$f_\pm(x, y) := x^2 \pm y^4$$

definiert sind. Es gilt dann

$$\mathbf{grad}\, f_\pm(x, y) = (2x, \pm 4y^3)\,,$$

und damit ist $(x, y) = 0$ wieder der einzige Kandidat für ein Extremum. Wir haben ferner

$$H(f_\pm)(0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}\,,$$

was eine positiv semi-definite Matrix ist. Mit Satz 11.6.3 können wir daher keine Aussage treffen, ob es ein Extremum in 0 gibt. Und tatsächlich hat f_+ ein striktes Minimum in 0, während f_- kein Extremum in 0 hat.

Als letztes Beispiel sei $f : \mathbb{R}^2 \to \mathbb{R}$, die durch

$$f(x, y) := y^2(x - 1) + x^2(x + 1)$$

definierte Funktion. Es gilt dann

$$\mathbf{grad}\, f(x, y) = \left((y^2 + 3x^2 + 2x, 2y(x - 1)\right).$$

Damit ist das nichtlineare Gleichungssystem

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

zu lösen. Aus der zweiten Gleichung können wir nun $x = 1$ oder $y = 0$ herleiten. Im Fall $x = 1$ ist dann die erste Gleichung $y^2 + 5 = 0$, und diese hat keine Lösung. Im Fall $y = 0$ ist die erste Gleichung dagegen $3x^2 + 2x = 0$ und diese hat die beiden Lösungen $x_1 = 0$ und $x_2 = -2/3$. Unsere Kandidaten für Extrema sind somit $(0, 0)$ und $(-2/3, 0)$. Um eine Entscheidung mit Satz 11.6.3 zu treffen, berechnen wir nun die Hesse-Matrix

$$H(f)(x, y) = \begin{pmatrix} 6x + 2 & 2y \\ 2y & 2(x - 1) \end{pmatrix}.$$

Dies ergibt dann für unsere Kandidaten

$$H(f)(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$
$$H(f)(-2/3,0) = \begin{pmatrix} -2 & 0 \\ 0 & -10/3 \end{pmatrix}.$$

Damit gibt es kein Extremum in 0 und ein striktes lokales Maximum in $(-2/3, 0)$.

Chapter 12: Implizite Funktionen

Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ ist in einer Umgebung von $x \in \mathbb{R}$ invertierbar, falls zum Beispiel f einmal stetig differenzierbar ist und $f'(x) \neq 0$ gilt. Können wir dies auf $f : \mathbb{R}^n \to \mathbb{R}^n$ verallgemeinern?

Sei $F : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^m$ eine "glatte" Funktion mit $F(0, 0) = 0$. Wann gibt es dann eine Funktion g mit $F(x, g(x)) = 0$ in einer Umgebung von 0 in $\mathbb{R}^k$?

Seien $f, g : \mathbb{R}^n \to \mathbb{R}$ Funktionen. Wie können wir f auf der Niveau-Menge $\{x : g(x) = 0\}$ optimieren?

Section 12.1
Der Satz über Implizite Funktionen

Wir betrachten die Abbildung $f : \mathbb{R}^2 \to \mathbb{R}$, die durch

$$f(x, y) := x^2 + y^2 - r^2$$

definiert ist. Dann ist $\{(x, y) : f(x, y) = 0\}$ der Kreis mit Radius r. Dieser kann in Teilen durch

$$y = g(x) = \sqrt{r^2 - x^2}\,, \qquad x \in [-r, r]$$

beschrieben werden. Es gilt dann $f(x, g(x)) = 0$ und wir sagen, dass g implizit durch f gegeben ist. Wir wollen nun solche impliziten Funktionen untersuchen.

Theorem 12.1.1

Seien $a \in \mathbb{R}^k$, $b \in \mathbb{R}^m$ und $U_1 := U(a, r_1)$ und $U_2 := U(b, r_2)$ mit $r_1, r_2 > 0$. Ferner sei

$$F : U_1 \times U_2 \to \mathbb{R}^m$$
$$(x, y) \mapsto F(x, y)$$

in (a, b) Fréchet-differenzierbar, so dass $F(a, b) = 0$ gilt und

$$\frac{\partial F}{\partial y}(a, b) \in \mathcal{M}(m, m)$$

invertierbar ist. Gibt es dann eine stetige Funktion $g : U_1 \to U_2$ mit $g(a) = b$ und

$$F(x, g(x)) = 0\,, \qquad x \in U_1,$$

dann ist g in a Fréchet-differenzierbar und es gilt

$$g'(a) = -\left(\frac{\partial F}{\partial y}(a, b)\right)^{-1} \frac{\partial F}{\partial x}(a, b)\,. \tag{12.1.1}$$

BEWEIS 1/4

Ohne Einschränkung können wir $(a, b) = (0, 0)$ annehmen. Wir definieren

$$A := \frac{\partial F}{\partial x}(0,0) \in \mathcal{M}(m,k)\,,$$
$$B := \frac{\partial F}{\partial y}(0,0) \in \mathcal{M}(m,m)\,.$$

Da F in $(0, 0)$ Fréchet-differenzierbar ist, gilt mit Block-Matrix-Schreibweise

$$F(x,y) = F(0,0) + (A,B) \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \varphi(x,y)$$
$$= Ax + By + \varphi(x,y)\,,$$

wobei die Funktion $\varphi : U_1 \times U_2 \to \mathbb{R}^m$ die Eigenschaften $\varphi(0, 0) = 0$ und

$$\lim_{(x,y)\to 0} \frac{\varphi(x,y)}{\|(x,y)\|} = 0 \tag{12.1.2}$$

erfüllt. Für $y := g(x)$ folgt dann

$$0 = F(x, g(x)) = Ax + Bg(x) + \varphi(x, g(x))\,.$$

Da B invertierbar ist, können wir dies zu

$$\begin{aligned} g(x) &= -B^{-1}Ax - B^{-1}\varphi(x, g(x)) \\ &= g(0) - B^{-1}Ax - B^{-1}\varphi(x, g(x)) \end{aligned} \tag{12.1.3}$$

umschreiben, wobei wir im letzten Schritt die angenommene Identität $g(0) = 0$ benutzt haben. Nach der Definition der Fréchet-Differenzierbarkeit reicht es daher

$$\lim_{x \to 0} \frac{B^{-1}\varphi(x, g(x))}{\|x\|} = 0 \tag{12.1.4}$$

zu zeigen. Dazu definieren wir

$$C_1 := \|B^{-1}A\| \qquad \text{und} \qquad C_2 := \|B^{-1}\|,$$

wobei wir für die Matrix-Normen und ihre Eigenschaften an Abschnitt 949 erinnern. Wegen (12.1.2) gibt es nun $\delta_i \in (0, r_i)$ mit

$$\begin{aligned} \|\varphi(x, y)\| \leq \frac{1}{2C_2} \cdot \|(x, y)\| &\leq \frac{1}{2C_2} \cdot \big(\|(x, 0)\| + \|(0, y)\|\big) \\ &= \frac{1}{2C_2} \cdot \big(\|x\| + \|y\|\big) \end{aligned} \tag{12.1.5}$$

für alle $x \in U(0, \delta_1)$ und $y \in U(0, \delta_2)$.

BEWEIS 3/4

Da g stetig ist, gibt es ferner ein $\delta \in (0, \delta_1)$ mit

$$\|x\| < \delta \quad \Longrightarrow \quad \|g(x)\| < \delta_2 \,.$$

Für $x \in U(0, \delta)$ haben wir also $g(x) \in U(0, \delta_2)$ und damit ergibt (12.1.5):

$$\|\varphi(x, g(x))\| \leq \frac{1}{2C_2} \cdot \big(\|x\| + \|g(x)\|\big) \,.$$

Für $x \in U(0, \delta)$ impliziert (12.1.3) daher

$$\begin{aligned} \|g(x)\| &= \| - B^{-1}Ax\| + \|B^{-1}\varphi(x, g(x)\| \\ &\leq C_1\|x\| + C_2\|\varphi(x, g(x))\| \\ &\leq \Big(C_1 + \frac{1}{2}\Big)\|x\| + \frac{1}{2}\|g(x)\| \end{aligned}$$

und lösen wir dies nach $\|g(x)\|$ auf, so erhalten wir

$$\|g(x)\| \leq \big(2C_1 + 1\big)\|x\|$$

für alle $x \in U(0, \delta)$. Mit der Dreiecksungleichung schließen wir dann auf

$$\|(x, g(x))\| \leq \|x\| + \|g(x)\| \leq 2(C_1 + 1)\|x\| \,.$$

BEWEIS 4/4

Für $x \in U(0, \delta)$ ergibt dies nun

$$\frac{\|B^{-1}\varphi(x, g(x))\|}{\|x\|} \leq C_2 \frac{\|\varphi(x, g(x))\|}{\|x\|} \leq 2(C_1 + 1)C_2 \frac{\|\varphi(x, g(x))\|}{\|(x, g(x))\|},$$

und da wegen der Stetigkeit von g aus $x \to 0$ auch $g(x) \to 0$ folgt, erhalten wir schließlich (12.1.4).

Theorem 12.1.2

Seien $a \in \mathbb{R}^k$, $b \in \mathbb{R}^m$ und $U_1 := U(a, r_1)$ und $U_2 := U(b, r_2)$ mit $r_1, r_2 > 0$. Ferner sei

$$F : U_1 \times U_2 \to \mathbb{R}^m$$
$$(x, y) \mapsto F(x, y)$$

stetig Fréchet-differenzierbar, es gelte $F(a, b) = 0$ und

$$\frac{\partial F}{\partial y}(a, b) \in \mathcal{M}(m, m)$$

sei invertierbar. Dann gibt es $\rho_1 \in (0, r_1)$ und $\rho_2 \in (0, r_2)$, so dass es für

$$V_1 := U(a, \rho_1) \qquad \text{und} \qquad V_2 := U(b, \rho_2)$$

genau eine Funktion $g : V_1 \to V_2$ gibt mit

$$F(x, g(x)) = 0\,, \qquad x \in V_1$$

und diese ist dann auch stetig.

BEWEIS 1/8

Wir nehmen wieder ohne Einschränkung $(a, b) = (0, 0)$ an und definieren

$$B := \frac{\partial F}{\partial y}(0,0) \in \mathcal{M}(m,m)\,.$$

Ferner betrachten wir die Funktion

$$\begin{aligned} G : U_1 \times U_2 &\to \mathbb{R}^m \\ (x,y) &\mapsto y - B^{-1}F(x,y)\,. \end{aligned}$$

Dann ist G stetig Fréchet-differenzierbar und wegen

$$\frac{\partial G}{\partial y}(x,y) = E_m - B^{-1}\frac{\partial F}{\partial y}(x,y)$$

gilt

$$\frac{\partial G}{\partial y}(0,0) = 0$$

und $G(0,0) = 0 - B^{-1}F(0,0) = 0$. Ferner ist $y \mapsto \frac{\partial G}{\partial y}(x,y)$ stetig und damit gibt es $\tilde{r}_i \in (0, r_i)$ mit

$$\left\| \frac{\partial G}{\partial y}(x,y) \right\| \leq \frac{1}{2}\,, \qquad (x,y) \in B(0,\tilde{r}_1) \times B(0,\tilde{r}_2). \tag{12.1.6}$$

Wir setzen $\rho_2 := \tilde{r}_2$. Da $G(0,0) = 0$ und G stetig ist, gibt es dann ein $\rho_1 \in (0, \tilde{r}_1)$ mit

$$\varepsilon := \sup_{x \in B(0,\rho_1)} \|G(x,0)\| < \frac{\rho_2}{2}\,. \tag{12.1.7}$$

Nach diesen Vorbereitungen bemerken wir nun zunächst, dass $F(x,y) = 0$ genau dann gilt wenn $B^{-1}F(x,y) = 0$ gilt und dies ist wiederum äquivalent zu $G(x,y) = y$. Damit haben wir

$$F(x,y) = 0 \qquad \Longleftrightarrow \qquad G(x,y) = y \tag{12.1.8}$$

für alle $(x,y) \in (U_1 \times U_2)$.

Seien nun $x \in V_1$ und $y_1, y_2 \in V_2$ mit $F(x, y_1) = 0 = F(x, y_2)$. Dann zeigt (12.1.8)

$$y_1 - y_2 = G(x, y_1) - G(x, y_2)$$

und mit dem abgeschwächten Mittelwertsatz, siehe Satz 11.5.3, auf $y \mapsto G(x, y)$ angewendet erhalten wir

$$\begin{aligned} \|y_2 - y_1\| &= \|G(x, y_1) - G(x, y_2)\| \\ &\leq \int_0^1 \left\| \frac{\partial G}{\partial y}(x, y_1 + t(y_2 - y_1)) \right\| \mathrm{d}t \cdot \|y_2 - y_1\| \\ &\leq \frac{1}{2} \|y_2 - y_1\| . \end{aligned}$$

wobei wir im letzten Schritt (12.1.6) benutzt haben. Dies zeigt $y_1 = y_2$ und damit die Eindeutigkeit von g.

BEWEIS 4/8

Um die Existenz zu zeigen, definieren wir eine Folge $(g_i)_{i\in\mathbb{N}}$ von Funktionen $g_i : B(0,\rho_1) \to \mathbb{R}^m$ induktiv durch $g_0 := 0$ und

$$g_{i+1} := G(x, g_i(x))\,, \qquad x \in B(0,\rho_1).$$

Da g_0 und G stetig sind, sehen wir mit Induktion schnell, dass alle g_i stetig sind. Mit Hilfe von Induktion wollen wir nun auch

$$\sup_{x\in B(0,\rho_1)} \|g_{i+1}(x) - g_i(x)\| < 2^{-i}\varepsilon \tag{12.1.9}$$

und

$$g_i(x) \in U(0,2\varepsilon)\,, \qquad x \in B(0,\rho_1) \tag{12.1.10}$$

zeigen. Für den Induktionsanfang beachten wir dazu, dass wir wegen $g_0 = 0$ auch $g_1(x) = G(x,0)$ haben und dies zeigt

$$\sup_{x\in B(0,\rho_1)} \|g_1(x) - g_0(x)\| = \sup_{x\in B(0,\rho_1)} \|G(x,0)\| = \varepsilon < \rho_2$$

mit (12.1.7). Ferner ist $g_0(x) = 0 \in U(0,2\varepsilon)$ für $x \in B(0,\rho_1)$ offensichtlich.

Für den Induktionsschritt fixieren wir nun ein $j \geq 0$ und nehmen an, dass (12.1.9) und (12.1.10) für alle $i = 0, \ldots, j$ wahr sind. Für $t \in [0,1]$ und $x \in B(0, \rho_1)$ haben wir dann

$$\begin{aligned} \|g_{j-1}(x) + t(g_j(x) - g_{j-1}(x))\| &= \|(1-t)g_{j-1}(x) + tg_j(x)\| \\ &\leq (1-t)\|g_{j-1}(x)\| + t\|g_j(x)\| \\ &< 2(1-t)\varepsilon + 2t\varepsilon \\ &= 2\varepsilon\,, \end{aligned}$$

wobei wir zweimal (12.1.10) benutzt haben.

Mit Satz 11.5.3 folgt dann

$$
\begin{aligned}
&\|g_{j+1}(x) - g_j(x)\| \\
&= \|G(x, g_j(x)) - G(x, g_{j-1}(x))\| \\
&\leq \int_0^1 \left\| \frac{\partial G}{\partial y}(x, g_{j-1}(x) + t(g_j(x) - g_{j-1}(x))) \right\| \mathrm{d}t \cdot \|g_j(x) - g_{j-1}(x)\| \\
&\leq \frac{1}{2} \cdot \|g_j(x) - g_{j-1}(x)\| \\
&< 2^{-1} 2^{-(j-1)} \varepsilon \\
&= 2^{-j} \varepsilon \,,
\end{aligned}
$$

wobei wir in der zweiten Ungleichung die Abschätzung (12.1.6) benutzt haben. Damit haben wir (12.1.9) gezeigt.

BEWEIS 7/8

Für den Beweis von (12.1.10) wenden wir die schon bewiesene Ungleichung (12.1.9) und ein Teleskopsummen-Argument an:

$$\begin{aligned} \|g_j(x)\| &= \Big\| \sum_{i=0}^{j-1} \big(g_{i+1}(x) - g_i(x)\big) \Big\| \\ &\leq \sum_{i=0}^{j-1} \|g_{i+1}(x) - g_i(x)\| \\ &< \sum_{i=0}^{j-1} 2^{-i}\varepsilon \\ &< 2\varepsilon\,. \end{aligned}$$

Mit (12.1.9) sehen wir nun, dass jede Komponentenfunktionenfolge von $(g_i)_{i\in\mathbb{N}}$ eine Cauchy-Folge bezüglich der Supremumsnorm ist. Damit konvergiert jede Komponentenfunktionenfolge gegen eine stetige Funktion, vgl. Satz 8.2.1. Insgesamt erhalten wir damit eine stetige Funktion $g : B(0,\rho_1) \to \mathbb{R}^m$ mit

$$\lim_{i\to\infty} g_i(x) = g(x)\,, \qquad\qquad x \in B(0,\rho_1).$$

Aus (12.1.10) folgt dann

$$g(x) \in B(0, 2\varepsilon) \subset U(0, \rho_2) = V_2$$

für alle $x \in B(0, \rho_1)$, da wir $2\varepsilon < \rho_2$ haben, siehe (12.1.7). Ferner gilt mit der Stetigkeit von G:

$$g(x) = \lim_{i \to \infty} g_{i+1}(x) = \lim_{i \to \infty} G(x, g_i(x)) = G(x, g(x))$$

für alle $x \in B(0, \rho_1)$. Mit (12.1.8) folgt dann $F(x, g(x)) = 0$ für alle $x \in B(0, \rho_1)$ und damit ist $g_{|V_1} : V_1 \to V_2$ unsere gesuchte Funktion.

BEMERKUNGEN UND BEISPIEL

Um die bisherigen Ergebnisse zu illustrieren, betrachten wir die Funktion $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, die durch

$$F(x, y) := x^2 + y^2 - 1$$

definiert ist. Dann gilt $F(0, 1) = 0$ und wegen $\frac{\partial F}{\partial y}(x, y) = 2y$ haben wir auch $\frac{\partial F}{\partial y}(0, 1) = 2 \in \mathcal{M}(1, 1)$, was offensichtlich eine invertierbare 1×1-Matrix ist. Nach Satz 12.1.2 gibt es dann $\varepsilon, \delta > 0$ und eine stetige Funktion $g : (-\varepsilon, \varepsilon) \to (1 - \delta, 1 + \delta)$ mit

$$F(x, g(x)) = 0\,, \qquad x \in (-\varepsilon, \varepsilon). \tag{12.1.11}$$

Einfaches Nachrechnen zeigt ferner

$$g(x) = \sqrt{1 - x^2}$$

und es ist leicht zu überprüfen, dass das größtmögliche ε gerade $\varepsilon = 1$ ist, wenn wir $\delta := 1$ setzen.

Ferner haben wir auch $F(1,0) = 0$. In diesem Fall ist aber $\frac{\partial F}{\partial y}(1,0) = 0$ und da dies keine invertierbare 1×1-Matrix ist, können wir Satz 12.1.2 nicht anwenden, um eine Funktion g mit (12.1.11) zu erhalten. Das dies kein Artefakt ist, zeigt Abbildung 23. Vertauschen wir jedoch die Rollen von x und y, so können wir mit Satz 12.1.2 eine Funktion $h : (-\delta, \delta) \to (1-\varepsilon, 1+\varepsilon)$ mit

$$F(h(y), y) = 0\,, \qquad\qquad y \in (-\delta, \delta)$$

finden, da wir $\frac{\partial F}{\partial x}(1,0) = 2 \neq 0$ haben. Offensichtlich ist $h(y) = \sqrt{1-y^2}$.

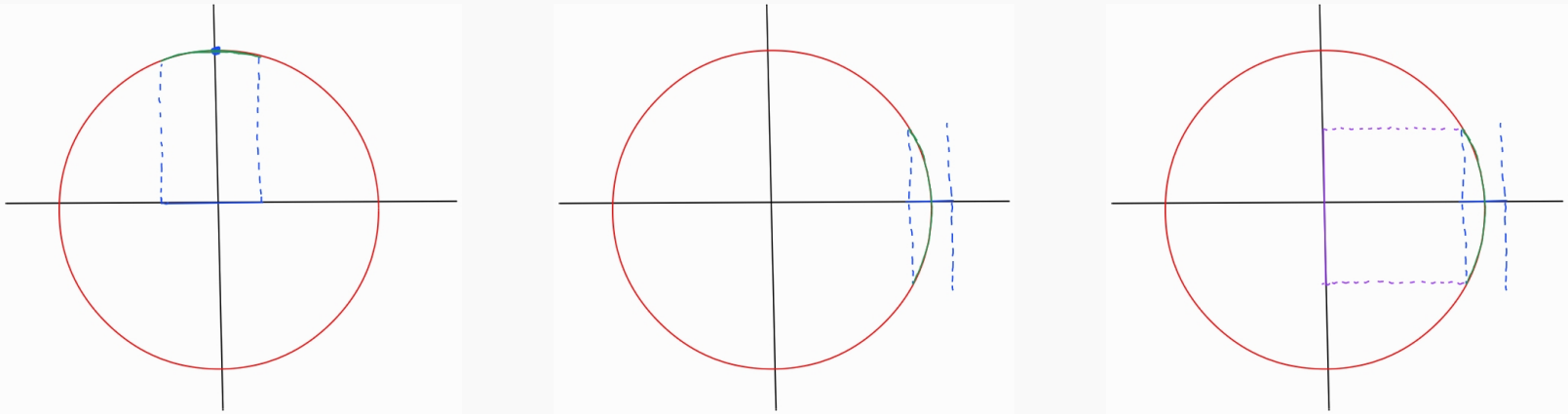


Abbildung: Links: Die rote Menge $\{(x, y) : F(x, y) = 0\}$ für die Funktion $F(x, y) := x^2 + y^2 - 1$. Im Punkt $(0, 1)$ ist $\frac{\partial F}{\partial y}(0, 1) = 2$ und wir können z.B. auf dem blauen Intervall eine eindeutige Funktion g finden mit $F(x, g(x)) = 0$. **Mitte:** Die gleiche Situation, diesmal aber im Punkt $(1, 0)$. Hier gilt $\frac{\partial F}{\partial y}(0, 1) = 2$ und egal, wie wir das blauen Intervall wählen, können wir kein eindeutiges g mit $F(x, g(x)) = 0$ finden, da wir sowohl oberhalb als auch unterhalb der x-Achse ein y mit $F(x, y) = 0$ haben. **Rechts:** Vertauschen der Rollen von x und y macht es möglich, ein lila Intervall auf der y-Achse zu finden, so dass wir eine Funktion h mit $F(h(y), y) = 0$ haben.

Betrachten wir die Situation in Satz 12.1.2, so folgt aus der Invertierbarkeit von

$$\frac{\partial F}{\partial y}(a,b) \in \mathcal{M}(m,m)$$

schon die Invertierbarkeit von $\frac{\partial F}{\partial y}(x,y)$ für hinreichend kleine Umgebungen, d.h. $x \in U(a,\tilde{\rho}_1)$ und $y \in U(b,\tilde{\rho}_2)$. In der Tat ist $\mathbf{det} : \mathcal{M}(m,m) \to \mathbb{R}$ stetig und damit ist auch

$$(x,y) \mapsto \mathbf{det}\Big(\frac{\partial F}{\partial y}(x,y)\Big)$$

stetig.

Aus

$$\det\Big(\frac{\partial F}{\partial y}(a,b)\Big) \neq 0$$

folgt dann die oben behauptete Invertierbarkeit in einer kleinen Umgebung. Da V_1 und V_2 in Satz 12.1.2 nicht weiter spezifiziert sind, können wir also ohne Einschränkung zusätzlich annehmen, dass $\frac{\partial F}{\partial y}(x,y)$ für alle $(x,y) \in V_1 \times V_2$ invertierbar ist. Mit Satz 12.1.1 erhalten wir dann die Differenzierbarkeit von $g : V_1 \to V_2$ mit

$$g'(x) = -\left(\frac{\partial F}{\partial y}(x,y)\right)^{-1} \frac{\partial F}{\partial x}(x,y)\,, \qquad (x,y) \in V_1 \times V_2.$$

Anders als im obigen Beispiel ist die Funktion g im Allgemeinen nicht explizit berechenbar. Der Beweis von 12.1.2 liefert jedoch ein iteratives Verfahren zur Approximation von g. Dazu betrachten wir die Funktion

$$G : U_1 \times U_2 \to \mathbb{R}^m$$
$$(x, y) \mapsto y - \left(\frac{\partial F}{\partial y}(a, b)\right)^{-1} F(x, y)\,.$$

Setzen wir nun $g_0 := 0$ und

$$g_{i+1}(x) := G(x, g_i(x))$$

so zeigte der Beweis von Satz 12.1.2, dass $g_i \to g$ gleichmäßig auf einer hinreichend kleinen Kugel um a. Setzen wir $y_i := g_i(x)$, so lässt sich die obige, induktive Definition auch als

$$y_{i+1} := y_i - \left(\frac{\partial F}{\partial y}(a, b)\right)^{-1} F(x, y_i)$$

schreiben. Dies ist das sogenannte Newton-Verfahren, das nach dem eben gesagten für $x \in V_1$ die Lösung $y \in V_2$ von $F(x, y) = 0$ durch die Folge $(y_i)_{i\in\mathbb{N}}$ approximiert, d.h. $y_i \to y$.

HÖHENLINIEN

Im folgenden sie $U \subset \mathbb{R}^2$ offen und $f : U \to \mathbb{R}$ stetig differenzierbar. Für $\lambda \in \mathbb{R}$ betrachten wir dann die **Niveau-Menge**

$$N_\lambda := \{(x, y) \in U : f(x, y) = \lambda\},$$

die anschaulich gesprochen der Höhenlinie zur Höhe λ auf einer Landkarte entspricht. Wir nehmen nun an, wir haben einen Punkt $(x^*, y^*) \in N_\lambda$ mit

$$\operatorname{grad} f(x^*, y^*) \neq (0, 0),$$

d.h. f hat *kein* Extremum in (x^*, y^*). Wir nehmen nun an, dass $\frac{\partial f}{\partial y}(x^*, y^*) \neq 0$ gilt, der andere Fall lässt sich durch Vertauschen von x und y analog behandeln. Für die durch $F(x, y) := f(x, y) - \lambda$ gegebene Funktion $F : U \to \mathbb{R}$ folgt dann

$$\frac{\partial F}{\partial y}(x^*, y^*) \neq 0 \qquad \text{und} \qquad F(x^*, y^*) = 0.$$

Nach Satz 12.1.2 und unseren vorherigen Bemerkungen existieren dann $\varepsilon, \delta > 0$ und eine stetig differenzierbare Funktion $g : (x^* - \delta, x^* + \delta) \to (y^* - \varepsilon, y^* + \varepsilon)$ mit

$$F(x, g(x)) = 0, \qquad x \in (x^* - \delta, x^* + \delta).$$

Mit anderen Worten haben wir

$$N_\lambda \cap \big((x^* - \delta, x^* + \delta) \times (y^* - \varepsilon, y^* + \varepsilon)\big) = \big\{(x, g(x)) : x \in (x^* - \delta, x^* + \delta)\big\}\,,$$

d.h. die Höhenlinie N_λ zur Höhe λ lässt sich *lokal* durch die Funktion g beschreiben und insbesondere handelt es sich tatsächlich um eine “Linie”.

Parametrisieren wir diese Linie nun durch $\varphi : (-\delta, \delta) \to \mathbb{R}^2$, d.h. $\varphi(t) := (t, g(x^* + t))^\mathsf{T}$ für $t \in (-\delta, \delta)$, so haben wir $f(\varphi(t)) = \lambda$ für alle $t \in (-\delta, \delta)$ und die mehrdimensionale Kettenregel ergibt dann

$$0 = (\mathbf{grad}\, f)(\varphi(t)) \cdot \varphi'(t)\,.$$

Der Vektor $\varphi'(t) \in \mathbb{R}^2$ steht daher nach Satz 11.2.9 senkrecht zu der Richtung des steilsten Anstieg von f im Punkt $(t, \varphi(t))$. Ferner beschreibt nach der Definition der Fréchet-Differenzierbarkeit die Abbildung

$$h \mapsto \varphi(t) + \varphi'(t) \cdot h\,, \qquad\qquad h \in \mathbb{R}$$

die Tangente an $\varphi(t)$, wobei wir $\varphi'(t) \in \mathcal{M}(2, 1)$ benutzen. Insgesamt sind also die Richtung des steilsten Anstiegs und die Richtung der Tangente senkrecht zueinander.

Haben wir eine Funktion $f : \mathbb{R}^n \to \mathbb{R}^n$, so ist diese in der Regel nicht invertierbar. Der folgende Satz, der Satz 6.1.8 verallgemeinert, zeigt aber die lokale Invertierbarkeit für Punkte, in denen die Ableitung von f eine invertierbare Matrix ist.

Theorem 12.1.3

Sei $U \subset \mathbb{R}^n$ offen, $f : U \to \mathbb{R}^n$ stetig differenzierbar und $a \in U$, so dass $Jf(a)$ invertierbar ist. Dann gibt es offene Mengen $V_1 \subset U$ und $V_2 \subset \mathbb{R}^n$ mit $a \in V_1$ und $b := f(a) \in V_2$, so dass

$$f_{|V_1} : V_1 \to V_2$$

bijektiv ist. Ferner ist die resultierende Umkehrabbildung $(f_{|V_1})^{-1}$ stetig differenzierbar und es gilt

$$J(f_{|V_1})^{-1}(b) = \big(Jf(a)\big)^{-1}.$$

BEWEIS 1/2

Wir betrachten die Funktion $F : U \times \mathbb{R}^n \to \mathbb{R}^n$, die durch

$$F(x, y) := f(x) - y$$

definiert ist. Dann gilt $F(a, b) = 0$ und

$$\frac{\partial F}{\partial x}(x, y) = Jf(x)\,.$$

Da die Matrix $Jf(a)$ invertierbar ist, gibt es nach Satz 12.1.2 Radien $\rho_1, \rho_2 > 0$ und genau eine Funktion

$$g : U(b, \rho_2) \to U(a, \rho_1)$$

mit $0 = F(g(y), y) = f(g(y)) - y$ für alle $y \in U(b, \rho_2)$, d.h.

$$f(g(y)) = y\,, \qquad\qquad y \in U(b, \rho_2)\,.$$

Wir setzen $V_2 := U(b, \rho_2)$ und $V_1 := U(a, \rho_1) \cap f^{-1}(V_2)$, wobei $f^{-1}(V_2)$ das Urbild von V_2 unter f bezeichnet. Mit der Stetigkeit von f kann man sich leicht überlegen, dass $f^{-1}(V_2)$ offen ist, und damit ist auch V_1 offen. Ferner ist $a \in f^{-1}(U(f(a), \rho_2)) = f^{-1}(V_2)$, und damit haben wir $a \in V_1$.

BEWEIS 2/2

Um die Surjektivität zu zeigen, wählen wir ein $y \in V_2$. Für $x := g(y) \in U(a, \rho_1)$ folgt dann $f(x) = f(g(y)) = y$. Wegen $y \in V_2$ zeigt dies auch $x \in f^{-1}(V_2)$ und damit haben wir insgesamt $x \in V_1$.

Für die Injektivität wählen wir zwei $x, x' \in V_1$ mit $f(x) = f(x')$. Für $y := f(x)$ haben wir dann

$$F(x, y) = f(x) - y = f(x') - y = F(x', y)$$

und damit ergibt die Eindeutigkeit der Funktion g die Identität $x = g(y) = x'$.

Ferner hatten wir in Abschnitt 1022 erläutert, dass wir durch geeignete Verkleinerung der offenen Mengen V_1 und V_2 die stetige Differenzierbarkeit von g garantieren können. Damit ist $f^{-1} = g$ stetig differenzierbar und die Formel für $J(f_{|V_1})^{-1}(b)$ folgt aus der mehrdimensionalen Kettenregel, sowie $f \circ g = \mathbf{id}_{V_2}$ und $g \circ f = \mathbf{id}_{V_1}$.

Section 12.2
Extremwerte unter Nebenbedingungen

Im folgenden wollen wir reellwertige Funktionen $F(x, y)$ über die (x, y) optimieren, die der Nebenbedingung $g(x, y) = 0$ genügen.

Ist beispielsweise $g(x, y) = y - \varphi(x)$, so haben wir für $g(x, y) = 0$ also $y = \varphi(x)$ und damit wollen wir dann die Funktion

$$x \mapsto F(x, \varphi(x))$$

optimieren. Dieses Beispiel lässt sich verallgemeinern. Haben wir zum Beispiel eine implizit gegebene Funktion φ mit $g(x, \varphi(x)) = 0$ und folgt aus $g(x, y) = 0$ schon $y = \varphi(x)$, so haben wir wieder

$$x \mapsto F(x, \varphi(x))$$

zu optimieren. Der folgende Satz greift diese Idee in seinem Beweis auf.

Theorem 12.2.1

Sei $U \subset \mathbb{R}^n$ offen, $a \in U$ und $g : U \to \mathbb{R}$ stetig differenzierbar mit $\operatorname{grad} g(a) \neq 0$. *Ferner sei*

$$N := \{x \in U : g(x) = 0\}$$

und $f : U \to \mathbb{R}$ sei stetig differenzierbar und es gibt ein $\delta > 0$ mit

$$f(a) \geq f(x)\,, \qquad x \in N \cap U(a, \delta). \tag{12.2.1}$$

Dann existiert ein $\lambda \in \mathbb{R}$ mit

$$\operatorname{grad} f(a) = \lambda \operatorname{grad} g(a)\,. \tag{12.2.2}$$

Die Bedingung (12.2.1) besagt, dass die Funktion $f_{|N} : N \to \mathbb{R}$ ein lokales Maximum in a hat. Das resultierende λ heißt Lagrange-Multiplikator. Die Bedingung (12.2.2) ist eine *notwendige* Bedingung für das Vorhandensein eines lokalen Maximums in N. Ein analoges Resultat für lokale Minima in N kann durch Betrachtung von $-f$ erzielt werden.

BEWEIS 1/4

Da wir $\mathbf{grad}\, g(a) \neq 0$ vorausgesetzt haben, können wir ohne Einschränkung

$$\frac{\partial g}{\partial x_n}(a) \neq 0$$

annehmen. Wir setzen $a' := (a_1, \ldots, a_{n-1})$, d.h. wir haben $(a', a_n) = a$. Nach Satz 12.1.2 und den Erläuterungen in Abschnitt 1022 gibt es dann offene Mengen $V_1 \subset \mathbb{R}^{n-1}$ und $V_2 \subset \mathbb{R}$ mit $a' \in V_1$ und $a_n \in V_2$ mit $V_1 \times V_2 \subset U$, sowie eine eindeutige, stetig differenzierbare Funktion $\varphi : V_1 \to V_2$ mit

$$M \cap (V_1 \times V_2) = \{x \in V_1 \times V_2 : x_n = \varphi(x_1, \ldots, x_n)\}\,.$$

Für $x' \in V_1$ setzen wir nun

$$\psi(x') := \begin{pmatrix} x' \\ \varphi(x') \end{pmatrix},$$

d.h. wir haben eine Funktion $\psi : V_1 \to \mathbb{R}^n$. Offensichtlich is ψ stetig differenzierbar mit

$$\psi'(x') = \begin{pmatrix} E_{n-1} \\ \varphi'(x') \end{pmatrix} \in \mathcal{M}(n, n-1)\,, \qquad x' \in V_1.$$

Ferner sichert uns unsere Konstruktion

$$g(\psi(x')) = g(x', \varphi(x')) = 0$$

für alle $x' \in V_1$. Wegen $\phi(a') = a_n$ haben wir und außerdem auch

$$\psi(a') = \begin{pmatrix} a' \\ \varphi(a') \end{pmatrix} = a$$

Nach der mehrdimensionalen Kettenregel erhalten wir somit

$$0 = (g \circ \psi)'(a') = g'(\psi(a')) \cdot \psi'(a') = \mathbf{grad}\, g(a) \cdot \begin{pmatrix} E_{n-1} \\ \varphi'(a') \end{pmatrix} \in \mathbb{R}^{n-1},$$

und durch komponentenweise Betrachtung ergibt dies

$$0 = \frac{\partial g}{\partial x_i}(a) + \frac{\partial g}{\partial x_n}(a) \cdot \frac{\partial \varphi}{\partial x_i}(a') \tag{12.2.3}$$

für alle $i = 1, \ldots, n-1$. Wir betrachten nun die Funktion $h : V_1 \to \mathbb{R}$, die durch

$$h(x') := f(x', \varphi(x')) = f \circ \psi(x')$$

definiert ist.

BEWEIS 3/4

Da f ein lokales Maximum in $N \cap B(a, \delta)$ besitzt, hat h dann ein lokales Maximum in a' und mit Satz 11.6.2 bekommen wir daher

$$h'(a') = 0\,.$$

Wegen $h = f \circ \psi$ erhalten wir dann analog zu (12.2.3) die Gleichung

$$0 = \frac{\partial f}{\partial x_i}(a) + \frac{\partial f}{\partial x_n}(a) \cdot \frac{\partial \varphi}{\partial x_i}(a')\,. \tag{12.2.4}$$

Wir setzen nun

$$\lambda := \frac{\partial f}{\partial x_n}(a) \cdot \left(\frac{\partial g}{\partial x_n}(a)\right)^{-1}.$$

Aus (12.2.4) und (12.2.3) folgt dann für $i = 1, \ldots, n-1$:

$$\begin{aligned}\frac{\partial f}{\partial x_i}(a) = -\frac{\partial f}{\partial x_n}(a) \cdot \frac{\partial \varphi}{\partial x_i}(a') &= -\frac{\partial f}{\partial x_n}(a) \cdot \left(-\frac{\partial g}{\partial x_i}(a) \cdot \left(\frac{\partial g}{\partial x_n}(a)\right)^{-1}\right) \\ &= \lambda \frac{\partial g}{\partial x_i}(a)\,.\end{aligned}$$

Ferner sichert die Definition von λ auch

$$\frac{\partial f}{\partial x_n}(a) = \frac{\partial f}{\partial x_i}(a) \cdot \left(\frac{\partial g}{\partial x_n}(a)\right)^{-1} \cdot \frac{\partial g}{\partial x_n}(a) = \lambda \frac{\partial g}{\partial x_n}(a)$$

und damit haben wir insgesamt (12.2.2) gezeigt.

Als erstes Beispiel wollen wir die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$, die durch $f(x, y) := xy$ definiert ist, betrachten. Wir wollen dann die Extrema von f auf der Einheitskugel $B := B(0, 1)$ bestimmen. Dazu betrachten wir zunächst das Innere von B, d.h. die offene Kugel $U(0, 1)$. Wegen

$$\mathbf{grad}\, f(x, y) = (y, x)$$

ist dann $(x^*, y^*) := (0, 0)$ nach Satz 11.6.2 der einzige Kandidat für ein Extremum. Es gilt aber

$$H := H(f)(0, 0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

und wegen $\mathbf{det}(H - \lambda E_2) = \lambda^2 - 1$ hat diese Matrix die Eigenwerte $\lambda_1 = 1$ und $\lambda_2 = -1$. Nach Satz 11.6.3 kann es daher kein lokales Extremum in $(0, 0)$ geben und damit gibt es kein lokales Extremum in der offenen Kugel $U(0, 1)$.

Es bleibt die Betrachtung des Randes übrig. Dazu setzen wir $g(x, y) := x^2 + y^2 - 1$ und bemerken, dass

$$N := \{(x, y) \in \mathbb{R}^2 : g(x, y) = 0\}$$

gerade der Rand von $B(0, 1)$ ist. Nach Konstruktion gilt nun

$$\mathbf{grad}\, g(x, y) = (2x, 2y)$$

und falls es ein lokales Extremum (x^*, y^*) von f auf N gibt, muss es nach Satz 12.2.1 ein $\lambda \in \mathbb{R}$ geben mit

$$\mathbf{grad}\, f(x^*, y^*) = \lambda\, \mathbf{grad}\, g(x^*, y^*)\,.$$

Aus den Formeln für die Gradienten erhalten wir dann die Gleichungen

$$y^* = 2\lambda x^*$$
$$x^* = 2\lambda y^* .$$

Einsetzen führt zu $y^* = 4\lambda^2 y^*$ und damit $0 = y^*(1 - 4\lambda^2)$. Ist nun $y^* = 0$, so haben wir auch $x^* = 2\lambda y^* = 0$ und dies widerspricht $g(x^*, y^*) = 0$. Damit führt die Existenz eines lokalen Extremums zu $4\lambda^2 = 1$, d.h. zu $\lambda = \pm 1/2$. Im Fall $\lambda = 1/2$ ergeben unsere Gleichungen dann $x^* = y^*$ und wegen $g(x^*, y^*) = 0$ führt dies zu $x^* = y^* = 1/\sqrt{2}$ bzw. $x^* = y^* = -1/\sqrt{2}$. Es gilt dann $f(x^*, y^*) = 1/2$ und weitere elementare Betrachtungen zeigen, dass dies globale Maxima von f auf $U(0, 1)$ sind.

Analog sehen wir im Fall $\lambda = -1/2$, dass $x^* = -y^* = 1/\sqrt{2}$ bzw. $x^* = -y^* = -1/\sqrt{2}$ globale Minima von f auf der Einheitskugel sind.

Um ein weiteres Beispiel zu betrachten, fixieren wir eine symmetrische Matrix $A \in \mathcal{M}(n,n)$ und betrachten die Funktion $f : \mathbb{R}^n \to \mathbb{R}$, die durch

$$f(x) := \langle x, Ax \rangle$$

definiert ist. Wir wollen nun die Extrema von f auf der Einheitsspähre $S := \{x \in \mathbb{R}^n : \|x\| = 1\}$ bestimmen. Dazu betrachten wie wieder die Funktion $g : \mathbb{R}^n \to \mathbb{R}$, die durch

$$g(x) := \|x\|^2 - 1$$

definiert ist. Offensichtlich ist dann $\{x \in \mathbb{R}^n : g(x) = 0\} = S$. Ferner gilt

$$\mathbf{grad}\, g(x) = 2x$$

und damit ist $\mathbf{grad}\, g(x) \neq 0$ für alle $x \in S$. Ferner hatten wir in (11.1.2) schon

$$\mathbf{grad}\, f(x) = (A + A^{\mathsf{T}})x\,, \qquad\qquad x \in \mathbb{R}^n.$$

gesehen. Da A symmetrisch ist, folgt $\mathbf{grad}\, f(x) = 2Ax$ für alle $x \in \mathbb{R}^n$.

Nun ist f stetig und man kann durch n-maliges Anwenden vom Satz von Bolzano-Weierstraß, siehe Satz 5.2.5, zeigen, dass S folgenkompakt ist. Nach Satz 5.2.8 und der dort angedeuteten Verallgemeinerung auf folgenkompakte Räume gibt es dann ein $x^* \in S$, so dass f ein Maximum in x^* hat. Satz 12.2.1 zeigt dann die Existenz eines $\lambda \in \mathbb{R}$ mit

$$\operatorname{grad} f(x^*) = \lambda \operatorname{grad} g(x^*) \,.$$

Die obigen Formeln für die Gradienten führen dann zu

$$Ax^* = \lambda x^* \,,$$

und wegen $x^* \neq 0$ muss λ einer der Eigenwerte von A sein. Ferner ist x^* ein Eigenvektor zu λ.

Sind nun $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A und $x_1, \ldots, x_n \in S$ zugehörige, normalisierte Eigenvektoren. Dann gilt

$$f(x_i) = \langle x_i, Ax_i \rangle = \lambda_i \langle x_i, x_i \rangle = \lambda_i \, .$$

Damit folgt

$$\max_{x \in S} \langle x, Ax \rangle = \max_{i=1,\ldots,n} \lambda_i \, ,$$

d.h. das Maximum von f ist gleich dem größten Eigenwert von A. Analog kann man zeigen, dass das Minimum von f dem kleinsten Eigenwert von A gleicht.

CHAPTER 13: KURVEN

Section 13.1
Bogenlänge

Definition 13.1.1

Seien $a, b \in \mathbb{R}$ mit $a < b$ und $\gamma : [a, b] \to \mathbb{R}^n$ stetig. Dann heißt γ **Kurve** und $\gamma(a)$ und $\gamma(b)$ heißen **Anfangs-** und **Endpunkt** der Kurve. Ferner heißt die Kurve γ:

i). **geschlossen**, falls $\gamma(a) = \gamma(b)$ gilt.

ii). **stetig differenzierbar**, falls $\gamma_{|(a,b)}$ differenzierbar ist und es eine stetige Funktion $g : [a, b] \to \mathbb{R}^n$ gibt mit $\gamma'(t) = g(t)$ für alle $t \in (a, b)$.

Im Fall $n = 3$ kann eine Kurve als der Weg eines Teilchens durch den Raum von der Zeit $t = a$ bis zur Zeit $t = b$ interpretiert werden. Die Stetigkeit sichert dann, dass das Teilchen nicht “springt”.

Ist $\gamma : [a, b] \to \mathbb{R}^n$ eine stetig differenzierbare Kurve, so ist die stetige Funktion $g : [a, b] \to \mathbb{R}^n$, die $\gamma'(t) = g(t)$ für alle $t \in (a, b)$ erfüllt, eindeutig. Wir setzen daher manchmal $\gamma'(t) := g(t)$ für $t \in \{a, b\}$.

Wir betrachten zunächst die Kurve

$$\begin{aligned} \gamma_1 : [0, 2\pi] &\to \mathbb{R}^2 \\ t &\mapsto (\cos t, \sin t)\,. \end{aligned} \tag{13.1.1}$$

Ihr Bild ist der Einheitskreis und in der Zeit von $t = 0$ bis $t = 2\pi$ "umrundet" die Kurve den Kreis genau einmal.

Betrachten wir nun die Kurve

$$\begin{aligned} \gamma_2 : [0, 2\pi] &\to \mathbb{R}^2 \\ t &\mapsto (\cos(2t), \sin(2t))\,, \end{aligned}$$

so ist diese wieder geschlossenen und ihr Bild ist wieder der Einheitskreis. Die Kurve γ_2 umrundet jedoch in der Zeit von $t = 0$ bis $t = 2\pi$ den Kreis zweimal.

Betrachten wir schließlich die Kurve

$$\begin{aligned} \gamma_3 : [0, 2\pi] &\to \mathbb{R}^2 \\ t &\mapsto (\cos(3t/2), \sin(3t/2))\,, \end{aligned}$$

so ist *nicht* geschlossen. Trotzdem ist ihr Bild wieder der Einheitskreis.

Für $r > 0$ heißt die Kurve

$$\begin{aligned} \gamma_4 : [a, b] &\to \mathbb{R}^2 \\ t &\mapsto r(t - \sin t, 1 - \cos t) \end{aligned} \tag{13.1.2}$$

Zykloide. Sie beschreibt den Weg eines festen Punktes auf einem Kreis mit Radius r, wenn dieser auf der x-Achse abgerollt wird. Im Fall $[a, b] = [0, 2\pi]$ befindet sich hierbei der Mittelpunkt des Kreises zur Zeit t im Punkt (rt, r) und der Anfangspunkt ist $\gamma_4(0) = (0, 0)$. Eine Zykloide ist nicht geschlossen.

Für $r > 0$ und $h \neq 0$ heißt die Kurve

$$\begin{aligned} \gamma_5 : [a, b] &\to \mathbb{R}^3 \\ t &\mapsto (r \cos t, r \sin t, ht) \end{aligned}$$

Schraubenlinie. Bei ihr "bewegt sich das Teilchen" auf einem Kreis mit Radius r, der parallel zu der xy-Ebene liegt, und gleichzeitig entlang der z-Achse. Ist $h > 0$, so ist die Schraubenlinie rechtsgängig, ansonsten linksgängig. Die Schraubenlinie ist nicht geschlossen.

Für $a, b > 0$ heißt die Kurve

$$\begin{aligned} \gamma_6 : [0, b] &\to \mathbb{R}^2 \\ t &\mapsto (at \cos t, at \sin t) \end{aligned}$$

Archimedische Spirale. Bei ihr "dreht sich das Teilchen" gleichmäßig um den Nullpunkt und mit proportionaler Geschwindigkeit bewegt es sich gleichzeitig vom Nullpunkt weg. Dementsprechend ist die Kurve nicht geschlossen.

ADDITION VON KURVEN

Sind $\gamma_1 : [a, b] \to \mathbb{R}^n$ und $\gamma_2 : [b, c] \to \mathbb{R}^n$ zwei Kurven mit mit $\gamma_1(b) = \gamma_2(b)$, so können wir diese beiden Kurven im folgenden Sinne addieren

$$\gamma_1 \oplus \gamma_2 : [a, c] \to \mathbb{R}^n$$
$$t \mapsto \begin{cases} \gamma_1(t), & \text{falls } t \in [a, b]\,, \\ \gamma_2(t), & \text{falls } t \in [b, c]\,. \end{cases}$$

Wegen $\gamma_1(b) = \gamma_2(b)$ bleibt dann die Stetigkeit erhalten, d.h. $\gamma_1 \oplus \gamma_2$ ist wieder eine Kurve. Anschaulich gesprochen werden die beiden Kurven γ_1 und γ_2 bei der Addition “zusammengeklebt”. Hierbei ist die Bedingung $\gamma_1(b) = \gamma_2(b)$ wichtig, und insbesondere folgt aus der Existenz von $\gamma_1 \oplus \gamma_2$ nicht die Existenz von $\gamma_2 \oplus \gamma_1$.

BOGENLÄNGE

Wegen der physikalischen Interpretation einer Kurve, ist es naheliegend nach der Länge von ihr zu fragen.

Leider gibt es Kurven, die keine (endliche) Länge haben. Anekdotisch schreibt man der Küstenlinie Englands diese Eigenschaft zu, da die Küstenlinie mit zunehmend feiner werdenden Maßstab immer länger erscheint. Mathematisch gesehen ist die Koch'sche Schneeflocke eine Kurve mit nicht endlicher Länge. Das gleiche gilt für fast alle Pfade der Brown'schen Bewegung.

Definition 13.1.2

Sei $\gamma : [a, b] \to \mathbb{R}^n$ eine Kurve, $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$ und

$$\mathcal{Z} : \qquad a = t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N = b$$

eine Zerlegung von $[a, b]$. Dann heißt

$$\mathrm{L}_{\mathcal{Z}}(\gamma) := \sum_{i=1}^{N} \|\gamma(t_i) - \gamma(t_{i-1})\|$$

die Länge der durch $\mathcal{Z}$ gegebenen Geradenapproximation von γ.

Eine Illustration von Geradenapproximationen und der zugehörigen Längen findet sich Abbildung 24.

Sind $\mathcal{Z}_1$ und $\mathcal{Z}_2$ Zerlegungen von $[a, b]$ mit $\mathcal{Z}_1 \subset \mathcal{Z}_2$ so zeigt die Dreiecksungleichung

$$\mathrm{L}_{\mathcal{Z}_1}(\gamma) \leq \mathrm{L}_{\mathcal{Z}_2}(\gamma)\,. \tag{13.1.3}$$

Ferner sei daran erinnert, dass wir für zwei beliebige Zerlegungen $\mathcal{Z}_1$ und $\mathcal{Z}_2$ von $[a, b]$ immer $\mathcal{Z}_i \subset \mathcal{Z}_1 \cup \mathcal{Z}_2$ und $\delta(\mathcal{Z}_i) \geq \delta(\mathcal{Z}_1 \cup \mathcal{Z}_2)$ für $i = 1, 2$ haben, siehe Abschnitt 408. Grob gesprochen können wir daher, ähnlich wie bei den Darboux'schen Untersummen, die approximierte Länge durch Verfeinerung der Zerlegung vergrößern.

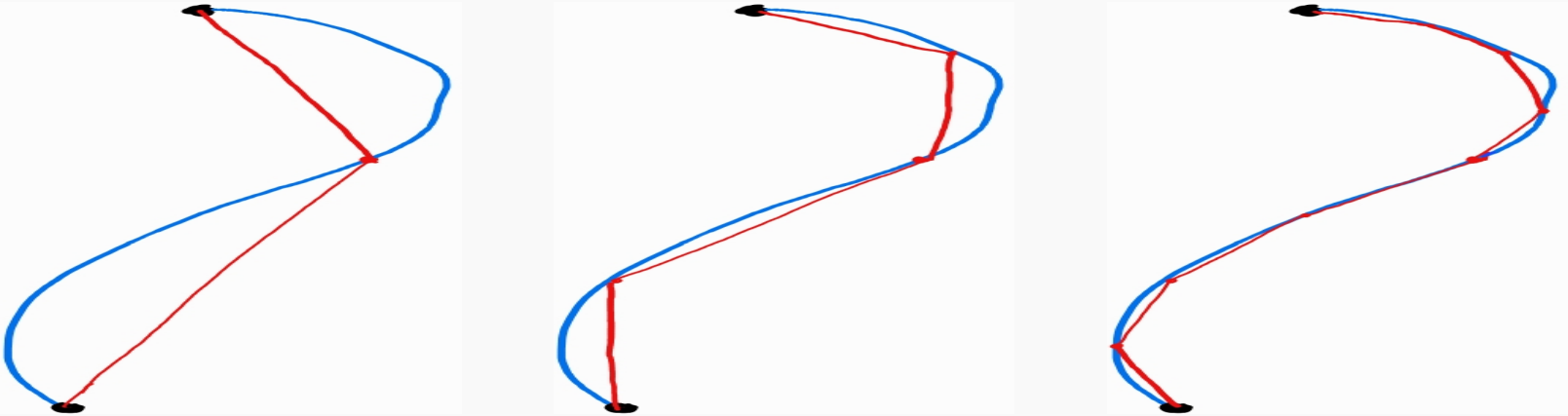


Abbildung: Links: Eine Kurve γ (in blau) und eine Geradenapproximationen mit einer Zerlegung $\mathcal{Z}_1$ der Größe $N = 2$. Die Länge der roten Strecken wird in $\mathrm{L}_{\mathcal{Z}_2}(\gamma)$ summiert. **Mitte:** Die gleiche Kurve und eine Zerlegung $\mathcal{Z}_2$ der Größe $N = 4$ mit $\mathcal{Z}_1 \subset \mathcal{Z}_2$. Aufgrund der Dreiecksungleichung haben wir $\mathrm{L}_{\mathcal{Z}_1}(\gamma) \leq \mathrm{L}_{\mathcal{Z}_2}(\gamma)$, siehe auch (13.1.3). **Rechts:** Eine weitere Zerlegung $\mathcal{Z}_3$ der Größe $N = 8$ mit $\mathcal{Z}_2 \subset \mathcal{Z}_3$. Die zugehörige Geradenapproximation beschreibt die Kurve schon recht gut und es ist zu sehen, dass die einzelnen Streckenabschnitte sich immer mehr den Tangenten annähern.

Aufgrund dieser Beobachtung ist die folgende Definition naheliegend.

Definition 13.1.3

Sei $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$ und $\gamma : [a,b] \to \mathbb{R}^n$ eine Kurve. Dann heißt γ rektifizerbar, falls

$$\mathrm{L}(\gamma) := \sup\left\{ \mathrm{L}_{\mathcal{Z}}(\gamma) : \mathcal{Z} \text{ Zerlegung von } [a,b] \right\} < \infty.$$

In diesem Fall heißt $\mathrm{L}(\gamma)$ die Länge von γ.

Man kann zeigen, dass die Rektifizierbarkeit unabhängig von der gewählten Norm ist. Dies beruht auf der Äquivalenz von Normen auf dem $\mathbb{R}^n$, die besagt, dass es zu je zwei beliebigen Normen $\|\cdot\|_1$ und $\|\cdot\|_2$ auf dem $\mathbb{R}^n$ Konstanten $c_1, c_2 > 0$ gibt mit

$$c_1\|x\|_1 \leq \|x\|_2 \leq c_2\|x\|_1\,, \qquad\qquad x \in \mathbb{R}^n.$$

Die Länge hängt jedoch von der Norm ab.

Sind $x_1, x_2 \in \mathbb{R}^n$ mit $x_1 \neq x_2$ und $\gamma : [a, b] \to \mathbb{R}^n$ eine rektifizierbare Kurve mit Anfangspunkt x_1 und Endpunkt x_2, so gilt

$$\|x_2 - x_1\| \leq \mathrm{L}(\gamma)\,.$$

Dies folgt aus (13.1.3), wenn wir die triviale Partition $\mathcal{Z}_1 : a = t_0, b = t_1$ betrachten, da in diesem Fall $\mathrm{L}_{\mathcal{Z}_1}(\gamma) = \|x_2 - x_1\|$ gilt. Da $\|x_2 - x_1\|$ auch die Länge der Strecke von x_1 nach x_2 ist, sehen wir also, dass diese Strecke immer der kürzeste Weg von x_1 nach x_2 darstellt, wobei diese Beobachtung *unabhängig von der Norm* ist.

Sind $\gamma_1 : [a, b] \to \mathbb{R}^n$ und $\gamma_2 : [b, c] \to \mathbb{R}^n$ zwei rektifizierbare Kurven mit mit $\gamma_1(b) = \gamma_2(b)$, so ist auch $\gamma_1 \oplus \gamma_2$ rektifizierbar und es gilt

$$\mathrm{L}(\gamma_1 \oplus \gamma_2) = \mathrm{L}(\gamma_1) + \mathrm{L}(\gamma_2)\,. \tag{13.1.4}$$

Der Beweis ist vergleichbar zu dem von Satz 7.1.7 und wird daher weggelassen.

Das folgende Lemma, das eine Art Mittelwertsatz-Abschätzung liefert, ist wichtig für die Berechnung von Längen.

Lemma 13.1.4

Sei $\gamma : [a, b] \to \mathbb{R}^n$ eine stetig differenzierbar Kurve. Dann gibt es zu jedem $\varepsilon > 0$ ein $\delta > 0$, so dass für alle $t_1, t_2 \in [a, b]$ mit $0 < |t_1 - t_2| \leq \delta$ gilt

$$\left\| \frac{\gamma(t_2) - \gamma(t_1)}{t_2 - t_1} - \gamma'(t_i) \right\| \leq \varepsilon \,, \qquad i = 1, 2.$$

BEWEIS 1/2

Wir betrachten zunächst den Fall $n = 1$. Sei dazu $\varepsilon > 0$. Da $\gamma' : [a, b] \to \mathbb{R}$ stetig ist, ist es nach Satz 5.2.10 sogar gleichmäßig stetig. Damit existiert ein $\delta > 0$, so dass für alle $s, t \in [a, b]$ mit $|s - t| \leq \delta$ gilt

$$\|\gamma'(t) - \gamma'(s)\| \leq \varepsilon \,. \tag{13.1.5}$$

Wir fixieren nun $t_1, t_2 \in [a, b]$ mit $0 < |t_1 - t_2| \leq \delta$. Ohne Einschränkung nehmen wir ferner $t_1 < t_2$ an. Nach dem eindimensionalen Mittelwertsatz 6.2.3 gibt es dann ein $s \in (t_1, t_2)$ mit

$$\frac{\gamma(t_2) - \gamma(t_1)}{t_2 - t_1} = \gamma'(s) \,.$$

Wegen $|s - t_i| \leq \delta$ liefert (13.1.5) dann

$$\left| \frac{\gamma(t_2) - \gamma(t_1)}{t_2 - t_1} - \gamma'(t_i) \right| = |\gamma'(s) - \gamma'(t_i)| \leq \varepsilon \,.$$

BEWEIS 2/2

Für den Fall $n > 1$ fixieren wir eine Konstante $c > 0$ mit $\|x\| \leq c\|x\|_\infty$ für alle $x \in \mathbb{R}^n$. Das ergibt

$$\left\| \frac{\gamma(t_2) - \gamma(t_1)}{t_2 - t_1} - \gamma'(t_i) \right\| \leq c \sup_{j=1,\dots,n} \left| \frac{\gamma_j(t_2) - \gamma_j(t_1)}{t_2 - t_1} - \gamma_j'(t_i) \right|,$$

wobei γ_j die j-te Komponentenfunktion von γ bezeichnet. Damit impliziert der eindimensionale Fall den Fall $n > 1$.

Mit diesen Vorbereitungen können wir jetzt die Länge einer rektifizierbaren Kurve berechnen,.

Theorem 13.1.5

Sei $\gamma : [a, b] \to \mathbb{R}^n$ eine stetig differenzierbar Kurve. Dann ist γ rektifizierbar und es gilt

$$\mathrm{L}(\gamma) = \int_a^b \|\gamma'(t)\| \, \mathrm{d}t \,. \tag{13.1.6}$$

Ferner gilt die Abschätzung

$$\mathrm{L}(\gamma) \leq (b - a) \cdot \sup_{t \in [a,b]} \|\gamma'(t)\| \,. \tag{13.1.7}$$

Mit Hilfe von (13.1.4) lässt sich (13.1.6) auf stückweise stetig differenzierbare Kurven, d.h. auf Kurven der Form $\gamma := \gamma_1 \oplus \cdots \oplus \gamma_m$ mit stetig differenzierbaren Teilkurven $\gamma_1, \ldots, \gamma_m$ verallgemeinern. Die Tatsache, dass dann $\gamma'(t)$ an den $m - 1$ "Klebestellen" nicht definiert sein muss, hat dabei keinen Einfluss auf das Riemann-Integral in (13.1.6).

BEWEIS 1/3

Da $t \mapsto \|\gamma'(t)\|$ stetig ist, ist die Funktion auch Riemann-integrierbar nach Satz 7.1.4. Sei nun $\varepsilon > 0$. Dann existiert nach Satz 7.1.6 ein $\delta > 0$, so dass für alle Zerlegungen $\mathcal{Z} : t_0, \ldots, t_N$ von $[a, b]$ der Feinheit $\delta(\mathcal{Z}) \leq \delta$ gilt:

$$\left| \int_a^b \|\gamma'(t)\| \, \mathrm{d}t - \sum_{i=1}^{N} \|\gamma'(t_i)\|(t_i - t_{i-1}) \right| \leq \varepsilon.$$

Nach Lemma 13.1.4 existiert ferner ein $\delta^* \in (0, \delta]$ mit

$$\left\| \frac{\gamma(t_i) - \gamma(t_{i-1})}{t_i - t_{i-1}} - \gamma'(t_i) \right\| \leq \frac{\varepsilon}{b - a}$$

für alle Zerlegungen $\mathcal{Z} : t_0, \ldots, t_N$ von $[a, b]$ der Feinheit $\delta(\mathcal{Z}) \leq \delta^*$.

Für eine solche Zerlegung gilt dann

$$\begin{aligned}
\left|\mathrm{L}_{\mathcal{Z}}(\gamma) - \int_a^b \|\gamma'(t)\|\,\mathrm{d}t\right| &\leq \left|\sum_{i=1}^{N} \|\gamma(t_i) - \gamma(t_{i-1})\| - \sum_{i=1}^{N} \|\gamma'(t_i)\|(t_i - t_{i-1})\right| \\
&\qquad + \left|\sum_{i=1}^{N} \|\gamma'(t_i)\|(t_i - t_{i-1}) - \int_a^b \|\gamma'(t)\|\,\mathrm{d}t\right| \\
&\leq \sum_{i=1}^{N} \left|\frac{\|\gamma(t_{i+1}) - \gamma(t_i)\|}{t_i - t_{i-1}} - \|\gamma'(t_i)\|\right| \cdot (t_i - t_{i-1}) + \varepsilon \\
&\leq \sum_{i=1}^{N} \frac{\varepsilon}{b-a} \cdot (t_i - t_{i-1}) + \varepsilon \\
&= 2\varepsilon\,.
\end{aligned}$$

BEWEIS 3/3

Sei schließlich $\mathcal{Z}$ eine beliebige Zerlegung von $[a, b]$ und $N > (b-a)/\delta^*$. Ferner sei $\mathcal{Z}_N$ die zugehörige äquidistante Zerlegung. Dann gilt $\delta(\mathcal{Z}_N) = (b-a)/N \leq \delta^*$ und damit ist $\mathcal{Z}' := \mathcal{Z} \cup \mathcal{Z}_N$ eine Zerlegung mit $\delta(\mathcal{Z}') \leq \delta^*$ und (13.1.3) zusammen mit unserer obigen Abschätzung ergibt

$$\mathrm{L}_{\mathcal{Z}}(\gamma) \leq \mathrm{L}_{\mathcal{Z}'}(\gamma) \leq \int_a^b \|\gamma'(t)\| \,\mathrm{d}t + 2\varepsilon\,.$$

Damit ist γ rektifizerbar mit

$$\mathrm{L}(\gamma) \leq \int_a^b \|\gamma'(t)\| \,\mathrm{d}t\,.$$

Ein weiteres Anwenden unserer obigen Abschätzung ergibt zudem

$$\lim_{N\to\infty} \mathrm{L}_{\mathcal{Z}_N}(\gamma) \to \int_a^b \|\gamma'(t)\| \,\mathrm{d}t\,.$$

Insgesamt haben wir damit (13.1.6) gezeigt und (13.1.7) ist eine direkte Konsequenz aus (13.1.6) und Satz 7.1.8.

Das folgende Korollar ist eine direkte Konsequenz von Satz 13.1.5, so dass sich der Beweis erübrigt.

Korollar 13.1.6

Sei $\gamma : [a, b] \to \mathbb{R}^n$ eine stetig differenzierbar Kurve und $\|\cdot\|$ die euklidische Norm. Dann gilt

$$\mathrm{L}(\gamma) = \int_a^b \sqrt{(\gamma_1'(t))^2 + \cdots + (\gamma_n'(t))^2}\,\mathrm{d}t\,.$$

Mit dem vorherigen Korollar lässt sich nun auch die Länge eines Graphen bestimmen.

Korollar 13.1.7

Sei $f : [a, b] \to \mathbb{R}$ stetig differenzierbar und $\gamma : [a, b] \to \mathbb{R}$ die durch $\gamma(t) := (t, f(t))$ definierte Funktion. Dann gilt

$$\mathrm{L}(\gamma) = \int_a^b \sqrt{1 + (f'(t))^2}\,\mathrm{d}t\,.$$

Man beachte, dass die obige Kurve γ den Graphen von f beschreibt.

Bevor wir noch einige Eigenschaften der Länge einer Kurve diskutieren, wollen wir zunächst ein paar Beispiele betrachten.

Wir beginnen mit der Kurve γ_1, die den Einheitskreis einmal durchläuft, siehe (13.1.1). In diesem Fall gilt

$$\gamma_1'(t) = (-\sin t, \cos t)\,, \qquad\qquad t \in [0, 2\pi]$$

und für die euklidische Norm erhalten wir daher die wenig überraschende Identität

$$\mathrm{L}(\gamma_1) = \int_0^{2\pi} \sqrt{(-\sin t)^2 + (\cos t)^2}\,\mathrm{d}t = \int_0^{2\pi} 1\,\mathrm{d}t = 2\pi\,.$$

Wir betrachten nun die Zykloide $\gamma_4 : [0, 2\pi]$ zum Radius $r > 0$, siehe (13.1.2). Es ist dann

$$\gamma_4'(t) = r(1 - \cos t, \sin t)\,, \qquad t \in [0, 2\pi]\,,$$

und damit gilt für die euklidische Norm:

$$\begin{aligned}\|\gamma_4'(t)\| = \sqrt{r^2(1-\cos t)^2 + r^2 \sin^2 t} &= r\sqrt{1 - 2\cos t + \cos^2 t + \sin^2 t} \\ &= r\sqrt{2 - 2\cos t}\,.\end{aligned}$$

Aus den Additionstheoremen, siehe Satz 2.5.5 erhalten wir zudem

$$1 - \cos t = 1 - \cos^2(t/2) + \sin^2(t/2) = 2\sin^2(t/2)\,,$$

und damit haben wir insgesamt

$$\begin{aligned}\mathrm{L}(\gamma_4) = \int_0^{2\pi} \|\gamma_4'(t)\|\,\mathrm{d}t = 2r\int_0^{2\pi} \sin(t/2)\,\mathrm{d}t &= 4r\int_0^{\pi} \sin(t)\,\mathrm{d}t \\ &= -4r\cos\Big|_0^{\pi} \\ &= 8r\,.\end{aligned}$$

Ist $\gamma : [a, b] \to \mathbb{R}^n$ eine Kurve und $\varphi : [c, d] \to [a, b]$ stetig und bijektiv mit $\varphi(c) = a$ und $\varphi(d) = b$, so ist

$$\gamma \circ \varphi : [c, d] \to \mathbb{R}^n$$

wieder ein Kurve, die die gleichen Anfangs- und Endpunkte hat. Physikalisch kann diese umparametrisierte Kurve $\gamma \circ \varphi$ so interpretiert werden, dass sich zwar die Geschwindigkeit und evtl. die ‘’Flugrichtung” des Teilchen ändert, nicht aber dessen “Flugbahn”. Intuitiv sollte daher die Weglänge von dieser Umparametrisierung unabhängig sein. In diesem Abschnitt wollen wir uns überlegen, dass dies tatsächlich so ist.

Zunächst wollen wir aber zwei wichtige Umparametrisierungen vorstellen. Dazu seien $a < b$ reelle Zahlen. Dann ist die Abbildung

$$\begin{aligned} \varphi_{a,b} : [0,1] &\to [a,b] \\ t &\mapsto (1-t)a + tb \end{aligned}$$

stetig differenzierbar, bijektiv und streng monoton wachsend. Mit Hilfe von $\varphi_{a,b}$ können wir also eine Kurve $\gamma : [a,b] \to \mathbb{R}^n$ zu einer $\gamma \circ \varphi_{a,b} : [0,1] \to \mathbb{R}^n$ umparametrisieren. Sind ferner $c < d$ zwei weitere, reelle Zahlen, so können wir auch die umparametrisierte Kurve

$$\gamma \circ (\varphi_{a,b} \circ \varphi_{c,d}^{-1}) : [c,d] \to \mathbb{R}^n$$

betrachten. Umparametrisierungen erlauben es also insbesondere, den Definitionsbereich einer Kurve “geeignet zu wählen”. Man beachte dabei, dass die Umparametrisierung $\varphi_{a,b} \circ \varphi_{c,d}^{-1}$ weiterhin wachsend ist, und damit wurde die “Durchlaufrichtung” der Kurve nicht geändert.

Betrachtet man dagegen die Rückwärtsparametrisierung

$$\begin{aligned} \varphi_{-1} : [a,b] &\to [a,b] \\ t &\mapsto a+b-t\,, \end{aligned}$$

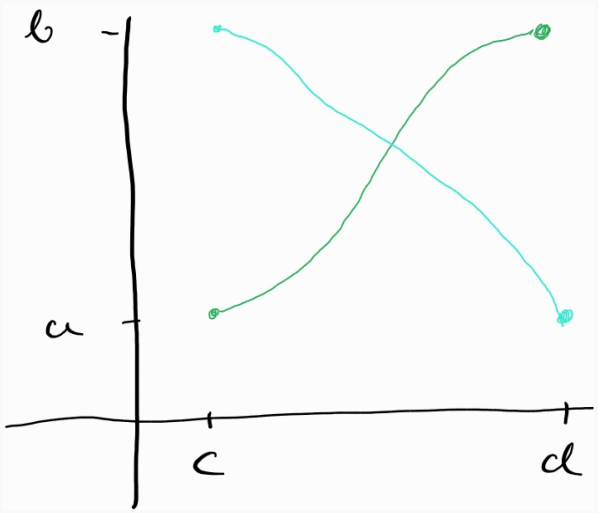
so ist diese wieder stetig differenzierbar und bijektiv, aber auch *fallend*. Ferner haben sich die Anfangs- und Endpunkte durch die Umparametrisierung von γ zu $\gamma \circ \varphi_{-1}$ vertauscht, da zum Beispiel $\gamma \circ \varphi_{-1}(a) = \gamma(b)$ gilt.

Wir beginnen mit dem folgenden Lemma, das zeigt, dass unsere Annahmen an φ schon die strenge Monotonie von φ implizieren. Für eine Illustration des Lemmas und seines Beweises verweisen wir auf die Abbildung 25. Ferner ist das folgende Lemma in gewisser Weise eine Umkehrung des Umkehrsatzes 5.2.7.

Lemma 13.1.8

Sei $\varphi : [c, d] \to [a, b]$ stetig und injektiv. Dann ist φ streng monoton.

MONOTONIE VON UMPARAMETRISIERUNGEN



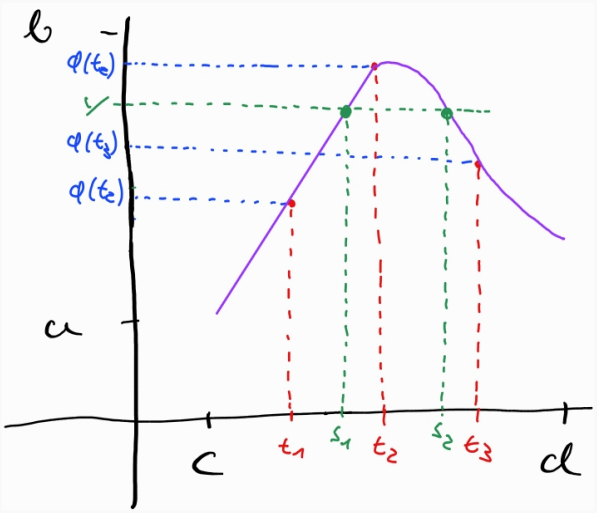


Abbildung: Links: Zwei stetige und bijektive Abbildungen $[c, d] \to [a, b]$, die auch monoton sind **Rechts:** Eine nicht-monotone Abbildung $\varphi : [c, d] \to [a, b]$. Die Punkte $t_1 < t_2 < t_3$ sind so gewählt, dass die Monotonie verletzt ist. Zum Wert y finden wirmit dem Zwischenwertsatz 5.2.2 dann $s_1 \in (t_1, t_2)$ und $s \in (t_2, t_3)$ mit $y = \varphi(s_1) = \varphi(s_2)$, was der Injektivität widerspricht.

Wir nehmen an, dass φ nicht streng monoton ist. Dann existieren $t_1, t_2, t_3 \in [c, d]$ mit $t_1 < t_2 < t_3$ und zum Beispiel $\varphi(t_1) \leq \varphi(t_2)$ und $\varphi(t_3) \leq \varphi(t_2)$.

Da φ injektiv ist, muss dann sogar $\varphi(t_1) < \varphi(t_2)$ und $\varphi(t_3) < \varphi(t_2)$ gelten. Wir setzen

$$y := \frac{\max\{\varphi_(t_1), \varphi(t_3)\} + \varphi(t_2)}{2}.$$

Wegen $\varphi(t_1) < y < \varphi(t_2)$ gibt es nach dem Zwischenwertsatz 5.2.2 dann ein $s_1 \in (t_1, t_2)$ mit $\varphi(s_1) = y$. Ferner gilt auch $\varphi(t_3) < y < \varphi(t_2)$ und damit finden wir analog ein $s_2 \in (t_2, t_3)$ mit $\varphi(s_2) = y$. Wegen $s_1 < t_2 < s_2$ widerspricht dies der Injektivität von φ.

Der andere Fall $\varphi(t_1) \geq \varphi(t_2)$ und $\varphi(t_3) \geq \varphi(t_2)$ ist analog zu behandeln.

Mit diesen Vorbereitungen können wir nun zeigen, dass sich die Länge einer Kurve bei einer Umparametrisierung nicht ändert.

Theorem 13.1.9

Sei $\gamma : [a, b] \to \mathbb{R}^n$ eine stetig differenzierbare Kurve und sei $\varphi : [c, d] \to [a, b]$ stetig differenzierbar und bijektiv. Dann ist $\gamma \circ \varphi$ rektifizierbar und es gilt

$$\mathrm{L}(\gamma \circ \varphi) = \mathrm{L}(\gamma)\,.$$

BEWEIS 1/2

Da $\gamma \circ \varphi : [c, d] \to \mathbb{R}^n$ stetig differenzierbar ist, folgt die Rektifizierbarkeit aus Satz 13.1.5. Ferner zeigt (13.1.6) zusammen mit der Kettenregel

$$\begin{aligned} \mathrm{L}(\gamma \circ \varphi) = \int_c^d \|(\gamma \circ \varphi)'(t)\| \, \mathrm{d}t &= \int_c^d \big\|(\gamma'(\varphi(t)) \cdot \varphi'(t)\big\| \, \mathrm{d}t \\ &= \int_c^d \|(\gamma'(\varphi(t))\| \cdot |\varphi'(t)| \, \mathrm{d}t \, . \end{aligned}$$

Aus Lemma 13.1.8 wissen wir schon, dass φ streng monoton ist. Sei daher φ ohne Einschränkung streng wachsend. Dann gilt $\varphi'(t) \geq 0$, siehe Korollar 6.2.5. Ferner folgt aus der Bijektivität und der Monotonie $\varphi(c) = a$ und $\varphi(d) = b$, denn wäre zum Beispiel $\varphi(c) > a$, so müsste es ein $t \in (c, d]$ mit $\varphi(t) = a$ geben. Dies widerspricht der Annahme, dass φ streng wachsend ist.

BEWEIS 2/2

Mit der Substitutionsregel, siehe Satz 7.2.7, folgt dann

$$\begin{aligned}\int_c^d \|(\gamma'(\varphi(t))\| \cdot |\varphi'(t)|\,\mathrm{d}t = \int_c^d \|(\gamma'(\varphi(t))\| \cdot \varphi'(t)\,\mathrm{d}t &= \int_{\varphi(c)}^{\varphi(d)} \|\gamma'(s)\|\,\mathrm{d}s \\ &= \int_a^b \|\gamma'(s)\|\,\mathrm{d}s \\ &= \mathrm{L}(\gamma)\,.\end{aligned}$$

Kombinieren wir beide Rechnungen erhalten wir die Formel $\mathrm{L}(\gamma \circ \varphi) = \mathrm{L}(\gamma)$. Im Fall einer fallenden Funktion φ ist die Rechnung analog.

Section 13.2
Wegintegrale

MOTIVATION

Ist $\gamma : [a, b] \to \mathbb{R}^n$ eine rektifizierbare Kurve, so wollen wir im Folgenden Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ bzw. $F : \mathbb{R}^n \to \mathbb{R}^n$ entlang dieser Kurve integrieren.

WEGINTEGRALE ERSTER ART

Im folgenden haben wir ein stetig differenzierbare Kurve $\gamma : [a, b] \to \mathbb{R}^n$ und eine stetige Funktion $f : \mathbb{R}^n \to \mathbb{R}$. Dann ist die Funktion

$$t \mapsto f(\gamma(t)) \cdot \|\gamma'(t)\|$$

ebenfalls stetig und damit existiert das Wegintegral erster Art

$$\int_\gamma f(x)\,\mathrm{d}x := \int_a^b f(\gamma(t)) \cdot \|\gamma'(t)\|\,\mathrm{d}t\,.$$

Ist die Kurve γ geschlossen, so schreiben wir auch

$$\oint_\gamma f(x)\,\mathrm{d}x := \int_\gamma f(x)\,\mathrm{d}x\,.$$

Ist $f = 1$, so ist offensichtlich

$$\int_\gamma f(x)\,\mathrm{d}x = \mathrm{L}(\gamma)\,.$$

Das Wegintegral erster Art verallgemeinert damit unsere Definition der Länge einer Kurve.

Zur Definition des Wegintegrals erster Art benötigen wir eigentlich keine Funktion f, die auf dem ganzen $\mathbb{R}^n$ definiert ist. Tatsächlich reicht es, eine stetige Funktion $f : \Gamma \to \mathbb{R}$ zu haben, wobei $\Gamma := \gamma([a, b])$ das Bild der Kurve γ ist.

Man kann das Wegintegral erster Art durch Riemann-Summen approximieren: Haben wir eine Folge $(\mathcal{Z}_m)_{m\in\mathbb{N}}$ von Zerlegungen $\mathcal{Z}_m : t_0^{(m)}, \dots, t_{N^{(m)}}^{(m)}$ von $[a,b]$ mit $\delta(\mathcal{Z}_m) \to 0$, so gilt nach Satz 7.1.6:

$$\int_\gamma f(x)\,\mathrm{d}x = \lim_{m\to\infty} \sum_{k=1}^{N^{(m)}} f\big(\gamma(t_k^{(m)})\big) \cdot \|\gamma'(t_k^{(m)})\| \cdot (t_k^{(m)} - t_{k-1}^{(m)})\,.$$

Wie bei der Definition von $\mathrm{L}(\gamma)$ kann ferner der Term $\|\gamma'(t_k^{(m)})\|$ durch

$$\frac{\|\gamma(t_k^{(m)}) - \gamma(t_{k-1}^{(m)})\|}{t_k^{(m)} - t_{k-1}^{(m)}}$$

ersetzt werden, ohne den Grenzwert zu ändern, d.h. wir haben

$$\int_\gamma f(x)\,\mathrm{d}x = \lim_{m\to\infty} \sum_{k=1}^{N^{(m)}} f\big(\gamma(t_k^{(m)})\big) \cdot \|\gamma(t_k^{(m)}) - \gamma(t_{k-1}^{(m)})\|\,.$$

Der Beweis ist analog zu dem von $\mathrm{L}(\gamma)$ und wird daher weggelassen. Eine Interpretation, die auf dieser Approximation basiert findet sich in Abbildung 26. Diese Approximation ermöglicht es auch, die Annahmen an f und γ zu relaxieren.

Ist die Kurve γ fixiert, so ist das Wegintegral erster Art linear. Dies ist aus der Definition offensichtlich. Ferner kann man analog zum Satz 13.1.9 zeigen, das sich das Wegintegral erster Art nicht ändert, wenn die Kurve umparametrisiert wird, d.h. es gilt

$$\int_{\gamma\circ\varphi} f(x)\,\mathrm{d}x = \int_{\gamma} f(x)\,\mathrm{d}x$$

für alle $\varphi : [c,d] \to [a,b]$, die stetig differenzierbar und bijektiv sind. Haben wir schließlich stetig differenzierbare Kurven $\gamma_1 : [a,b] \to \mathbb{R}^n$ und $\gamma_2 : [b,c] \to \mathbb{R}^n$ mit $\gamma_1(b) = \gamma_2(b)$, so können wir

$$\int_{\gamma_1\oplus\gamma_2} f(x)\,\mathrm{d}x := \int_{\gamma_1} f(x)\,\mathrm{d}x + \int_{\gamma_2} f(x)\,\mathrm{d}x$$

setzen. Wenn man das Wegintegral erster Art stattdessen über obige Approximationen definiert, ergibt sich diese Identität von allein.

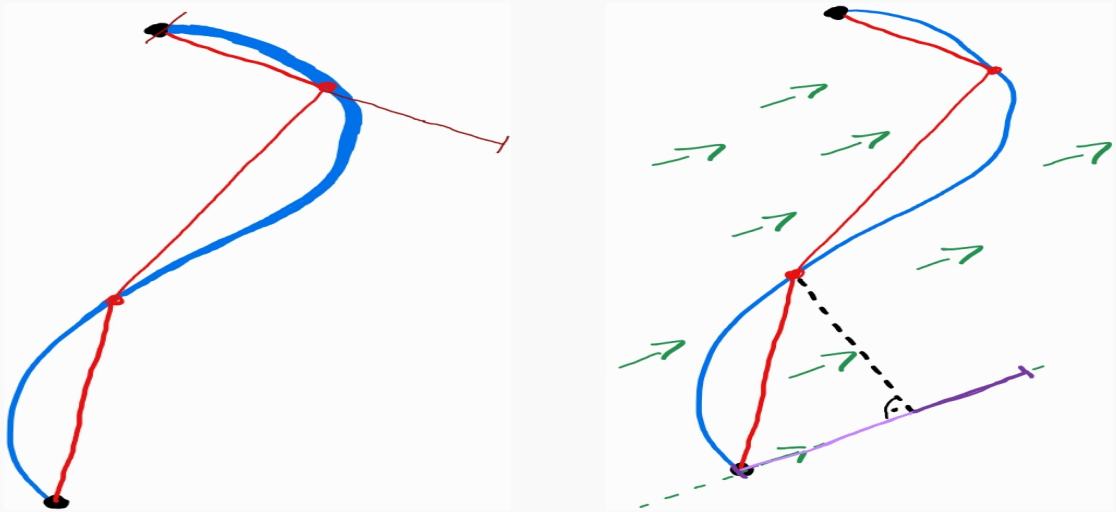


Abbildung: Links: Eine Kurve γ (in blau) und eine stückweise lineare Approximation (in rot). Die Funktion $f : \Gamma \to \mathbb{R}$ wird durch die Dicke der Kurve dargestellt. Der Approximationsterm $f(\gamma(t_N)) \cdot \|\gamma(t_N) - \gamma(t_{N-1})\|$ entspricht der roten Strecke mal der Dicke der Kurve, was durch die Länge der lila Strecke dargestellt wird. Physikalisch gesprochen kann die Funktion f zum Beispiel die Dichte oder den Widerstand eines eindimensionalen Drahtes beschreiben. **Rechts:** Approximation des Wegintegral 2. Art für ein konstantes Vektorfeld, das durch die grünen Pfeile angedeutet wird. Das Skalarprodukt $\langle F(\gamma(t_1)), \gamma(t_1) - \gamma(t_0)\rangle$ entspricht der Länge der verlängerten, lila Strecke, d.h. dem Anteil der roten Strecke in Richtung der grünen Pfeile mal der Länge der grünen Pfeile. Physikalisch gesprochen entspricht das

Im folgenden haben wir ein stetig differenzierbare Kurve $\gamma : [a, b] \to \mathbb{R}^n$ und eine stetige Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$. Dann ist die Funktion

$$t \mapsto \langle F(\gamma(t)), \gamma'(t) \rangle$$

ebenfalls stetig und damit existiert das Wegintegral zweiter Art

$$\int_\gamma \langle F(x),\, \mathrm{d}x \rangle := \int_a^b \langle F(\gamma(t)), \gamma'(t) \rangle \,\mathrm{d}t\,.$$

Ist die Kurve γ geschlossen, so schreiben wir auch

$$\oint_\gamma \langle F(x),\, \mathrm{d}x \rangle := \int_\gamma \langle F(x),\, \mathrm{d}x \rangle\,.$$

BEMERKUNGEN

Zur Definition des Wegintegrals zweiter Art benötigen wir eigentlich keine Funktion F, die auf dem ganzen $\mathbb{R}^n$ definiert ist. Tatsächlich reicht es wieder, eine stetige Funktion $F : \Gamma \to \mathbb{R}^n$ für $\Gamma := \gamma([a, b])$ zu haben. In Anwendungen haben wir häufig eine stetige Funktion $F : U \to \mathbb{R}^n$ mit $\Gamma \subset U$.

Man kann auch das Wegintegral zweiter Art durch Riemann-Summen approximieren: Haben wir eine Folge $(\mathcal{Z}_m)_{m\in\mathbb{N}}$ von Zerlegungen $\mathcal{Z}_m : t_0^{(m)}, \dots, t_{N^{(m)}}^{(m)}$ von $[a,b]$ mit $\delta(\mathcal{Z}_m) \to 0$, so gilt nach Satz 7.1.6:

$$\int_\gamma \langle F(x), \mathrm{d}x\rangle = \lim_{m\to\infty} \sum_{k=1}^{N^{(m)}} \left\langle f\left(\gamma(t_k^{(m)})\right), \gamma'(t_k^{(m)})\right\rangle \cdot (t_k^{(m)} - t_{k-1}^{(m)}).$$

Wieder kann der Term $\gamma'(t_k^{(m)})$ hierbei durch

$$\frac{\gamma(t_k^{(m)}) - \gamma(t_{k-1}^{(m)})}{t_k^{(m)} - t_{k-1}^{(m)}}$$

ersetzt werden, ohne den Grenzwert zu ändern, d.h. wir haben

$$\int_\gamma \langle F(x), \mathrm{d}x\rangle = \lim_{m\to\infty} \sum_{k=1}^{N^{(m)}} \left\langle F\left(\gamma(t_k^{(m)})\right), \gamma(t_k^{(m)}) - \gamma(t_{k-1}^{(m)})\right\rangle.$$

Der Beweis ist wieder ähnlich zu dem von $\mathrm{L}(\gamma)$ und wird daher auch weggelassen. Eine Interpretation, die auf dieser Approximation basiert, findet sich in Abbildung 26. Diese Approximation ermöglicht es wieder, die Annahmen an F und γ zu relaxieren.

Ist die Kurve γ fixiert, so ist das Wegintegral zweiter Art linear. Dies ist aus der Definition offensichtlich. Ferner kann man analog zum Satz 13.1.9 zeigen, das sich das Wegintegral zweiter Art nicht ändert, wenn die Kurve umparametrisiert wird *ohne die Durchlaufrichtung zu ändern*, d.h. es gilt

$$\int_{\gamma\circ\varphi} \langle F(x),\, \mathrm{d}x\rangle = \int_{\gamma} \langle F(x),\, \mathrm{d}x\rangle$$

für alle $\varphi : [c,d] \to [a,b]$, die stetig differenzierbar, *wachsend* und bijektiv sind. Dieser Unterschied zum Wegintegral erster Art liegt an der Tatsache, dass beim Wegintegral zweiter Art die Ableitung φ' durch das Anwenden der Kettenregel nicht als Betrag, d.h. vorzeichenlos, erscheint. Ist stattdessen $\varphi : [c,d] \to [a,b]$ stetig differenzierbar, *fallend* und bijektiv, so gilt entsprechend

$$\int_{\gamma\circ\varphi} \langle F(x),\, \mathrm{d}x\rangle = -\int_{\gamma} \langle F(x),\, \mathrm{d}x\rangle \tag{13.2.1}$$

Diese Eigenschaften stimmen auch mit der Interpretation in Abbildung 26 überein.

Haben wir schließlich stetig differenzierbare Kurven $\gamma_1 : [a, b] \to \mathbb{R}^n$ und $\gamma_2 : [b, c] \to \mathbb{R}^n$ mit $\gamma_1(b) = \gamma_2(b)$, so können wir

$$\int_{\gamma_1 \oplus \gamma_2} \langle F(x),\, \mathrm{d}x \rangle := \int_{\gamma_1} \langle F(x),\, \mathrm{d}x \rangle + \int_{\gamma_2} \langle F(x),\, \mathrm{d}x \rangle \tag{13.2.2}$$

setzen. Wenn man das Wegintegral zweiter Art stattdessen über obige Approximationen definiert, ergibt sich diese Identität von allein.

Section 13.3
Berechnung von Potenzialen

Im folgenden wollen wir uns mit der Frage beschäftigen, unter welchen Bedingungen ein Wegintegral zweiter von der Kurve unabhängig ist, solange die Anfangs- und Endpunkte sich nicht ändern.

Wir beginnen mit der folgenden Definition.

Definition 13.3.1

Sei $U \subset \mathbb{R}^n$ offen und $F : U \to \mathbb{R}^n$ stetig. Dann heißt F **wirbelfrei**, falls für alle geschlossenen, stückweise stetig differenzierbaren Kurven $\gamma : [a, b] \to U$ gilt

$$\oint_\gamma \langle F(x), \, \mathrm{d}x \rangle = 0 \, .$$

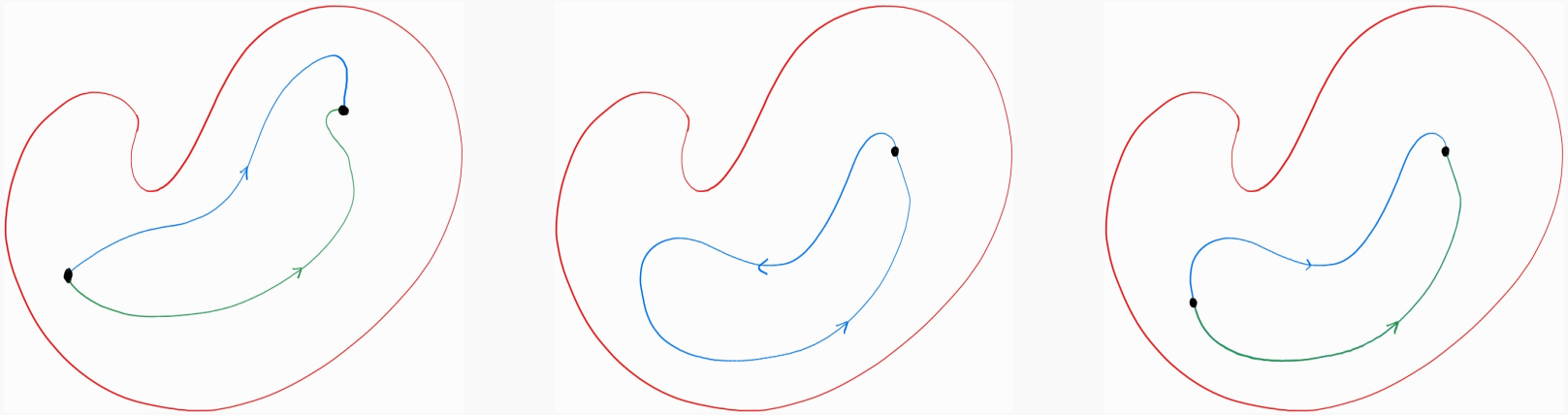


Abbildung: Links: Eine Menge U (in rot) und zwei Punkte (in schwarz) in U, die durch zwei Kurven in U (in grün und blau) miteinander verbunden sind. Die beiden Pfeile deuten an, dass der linke Punkt jeweils der Anfangspunkt ist. Wird nun die Durchlaufrichtung der grünen Kurve geändert und das Ergebnis mit der blauen Kurve verklebt, so erhalten wir eine geschlossene Kurve, deren Anfangs- und Endpunkt der linke Punkt ist. Kurve (in blau) in U verbunden werden können, während die Strecke (in grün) zwischen den beiden Punkten nicht komplett in U liegt. **Mitte:** Ein geschlossene Kurve in der gleichen Menge U, deren Anfangs- und Endpunkt in schwarz gekennzeichnet ist. **Rechts:** Wählen wir einen Punkt der geschlossenen Kurve (links in Schwarz) und betrachten wir die resultierenden Teilkurven (in grün und blau), so sehen wir, dass wir in der Situation der linken Illustration sind,

Um die Definition in Hinblick auf unsere Eingangsfrage zu verstehen, seien $F : U \to \mathbb{R}^n$ stetig und $\gamma_1, \gamma_2 : [a, b] \to U$ zwei stückweise stetig differenzierbare Kurven mit $\gamma_1(a) = \gamma_2(a)$ und $\gamma_1(b) = \gamma_2(b)$. Ferner sei $\gamma_2^- : [b, 2b - a] \to U$ durch

$$\gamma_2^-(t) := \gamma_2(2b - t)\,, \qquad t \in [b, 2b - a]$$

definiert. Dann gilt $\gamma_2^-(b) = \gamma_2(b) = \gamma_1(b)$ und deswegen ist die Kurve $\gamma_1 \oplus \gamma_2^- : [a, 2b - a]$ definiert und nach Voraussetzung auch stückweise stetig differenzierbar. Ferner gilt

$$\gamma_1 \oplus \gamma_2^-(2b - a) = \gamma_2^-(2b - a) = \gamma_2(a) = \gamma_1(a) = \gamma_1 \oplus \gamma_2^-(a)\,,$$

d.h. die Kurve $\gamma_1 \oplus \gamma_2^-$ ist geschlossen. Außerdem ist die Umparametrisierung $t \mapsto 2b - t$ fallend und wegen (13.2.2) und (13.2.1) folgt daher

$$\begin{aligned}\oint_{\gamma_1 \oplus \gamma_2^-} \langle F(x), \mathrm{d}x\rangle &= \oint_{\gamma_1} \langle F(x), \mathrm{d}x\rangle + \oint_{\gamma_2^-} \langle F(x), \mathrm{d}x\rangle \\ &= \oint_{\gamma_1} \langle F(x), \mathrm{d}x\rangle - \oint_{\gamma_2} \langle F(x), \mathrm{d}x\rangle\,.\end{aligned}$$

Ist F wirbelfrei, so erhalten wir also

$$\oint_{\gamma_1} \langle F(x),\, \mathrm{d}x\rangle = \oint_{\gamma_2} \langle F(x),\, \mathrm{d}x\rangle\,. \tag{13.3.1}$$

Umgekehrt lässt sich jede geschlossene, stückweise stetig differenzierbaren Kurve γ in der Form $\gamma = \gamma_1 \oplus \gamma_2^-$ darstellen, und damit sind die wirbelfreien Vektorfelder genau die Vektorfelder, auf die unsere Eingangsfrage eine positive Antwort hat. Eine Illustration dieser Beobachtungen findet sich in Abbildung 27.

Um wirbelfrei Vektorfelder genauer zu untersuchen, benötigen wir noch die folgende Definition, die auch für andere Fragen interessant ist.

Definition 13.3.2

Sei $U \subset \mathbb{R}^n$. Dann heißt die Menge U (Pfad)-zusammenhängend, falls es für alle $x, y \in U$ eine Kurve $\gamma : [a, b] \to U$ gibt mit $\gamma(a) = x$ und $\gamma(b) = y$.

Man beachte, dass die Kurve γ in die Menge U abbilden muss. Ist U konvex, siehe die Diskussion nach dem mehrdimensionalen Mittelwertsatz 11.5.1, so ist U auch zusammenhängend: Um dies zu überprüfen, betrachten wir die Kurve $\gamma : [0, 1] \to \mathbb{R}^n$, die durch

$$\gamma(t) := x + t(y - x)\,, \qquad t \in [0, 1]$$

definiert ist. Die Konvexität sichert dann $\gamma([0, 1]) = [x, y] \subset U$. Illustrationen für zusammenhängende Mengen finden sich in Abbildung 28.

Man kann ferner zeigen, dass wir für offene, zusammenhängende Mengen U für alle $x, y \in U$ immer auch stückweise stetig differenzierbare Kurve $\gamma : [a, b] \to U$ gibt mit $\gamma(a) = x$ und $\gamma(b) = y$ finden. Für konvexe Mengen ist dies offensichtlich.

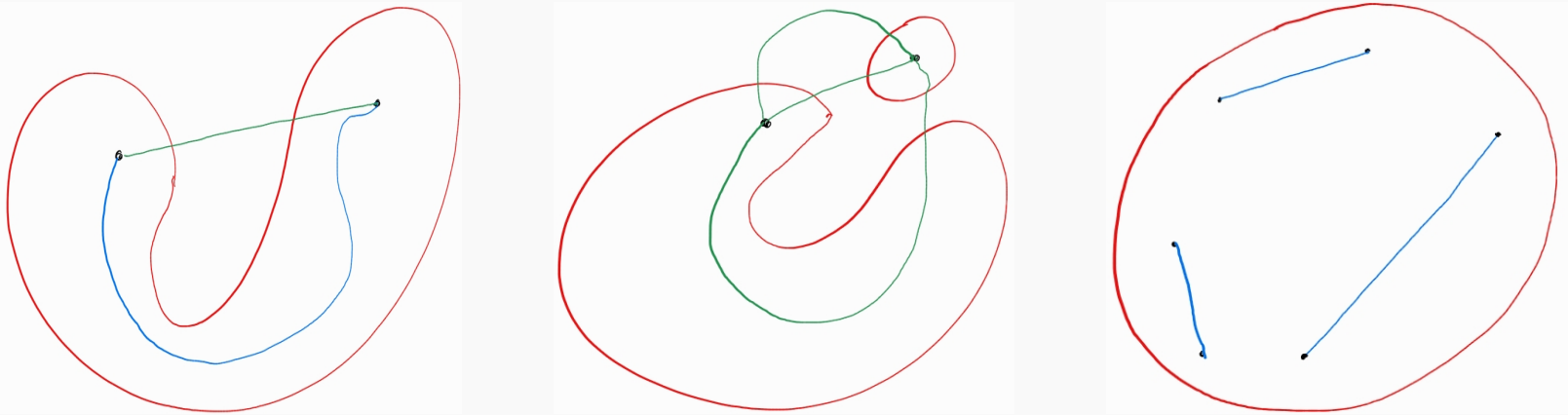


Abbildung: Links: Eine Menge U (in rot), die zusammenhängend aber nicht konvex ist, da z.B. die beiden schwarzen Punkte durch eine Kurve (in blau) in U verbunden werden können, während die Strecke (in grün) zwischen den beiden Punkten nicht komplett in U liegt. **Mitte:** Eine Menge U (in rot), die nicht zusammenhängend ist, da z.B. die beiden schwarzen Punkte nicht durch eine Kurve verbunden werden können, die komplett in U liegt. Die drei grünen Kurven illustrieren dies. **Rechts:** Eine konvexe Menge U (in rot). Für jedes Paar schwarzer Punkte in U liegt die Strecke zwischen ihnen ebenfalls

Im folgenden wollen wir wirbelfreie Vektorfelder charakterisieren. Wir erinnern dazu daran, dass eine Abbildung $F : U \to \mathbb{R}^n$ Gradientenfeld heißt, falls es eine partiell differenzierbare Abbildung $\varphi : U \to \mathbb{R}$ gibt mit

$$F(x) = \nabla\varphi(x)\,, \qquad\qquad x \in U\,,$$

siehe auch (11.3.1). In diesem Fall nennen wir die Abbildung φ ein Potential von F.

Man beachte, dass, wenn φ ein Potential von F ist, auch die Funktion $\varphi + c$ für $c \in \mathbb{R}$ ein Potential von F ist.

Im Fall $n = 3$ gilt für Gradientenfelder F mit zweimal stetig differenzierbaren Potentialen φ die Gleichung

$$\mathbf{rot}\, F = \mathbf{rot}\, \nabla\varphi = 0\,,$$

wie wir im Lemma 11.3.4 schon gesehen haben. Für z.B. $U = \mathbb{R}^3 \setminus \{0\}$ gibt es jedoch rotationsfreie Vektorfelder, die keine Gradientenfelder sind. Grob gesprochen ist hierfür das Loch $\{0\}$ schuld.

Der folgende Satz zeigt, dass stetige Gradientenfelder gerade wirbelfreie Vektorfelder sind. Ferner stellt er einen Zusammenhang zwischen Potentialen und Wegintegralen zweiter Art her.

Theorem 13.3.3

Sei $U \subset \mathbb{R}^n$ offen und zusammenhängend und $F : U \to \mathbb{R}^n$ stetig. Dann gelten die folgenden Aussagen

i). Ist F ein Gradientenfeld mit Potential φ, so ist F wirbelfrei und für alle stückweise stetig differenzierbaren Kurven $\gamma : [a, b] \to U$ gilt

$$\int_\gamma \langle F(x),\, \mathrm{d}x\rangle = \varphi(\gamma(b)) - \varphi(\gamma(a))\,.$$

ii). Ist F wirbelfrei, so ist F ein Gradientenfeld mit Potential

$$\varphi(y) := \int_{\gamma_{y^*,y}} \langle F(x),\, \mathrm{d}x\rangle\,, \qquad\qquad y \in U, \tag{13.3.2}$$

wobei wir ein $y^* \in U$ fixieren und für jedes $y \in U$ eine beliebige, stückweise stetig differenzierbare Kurve $\gamma_{y^*,y} : [a, b] \to U$ mit $\gamma_{y^*,y}(a) = y^*$ und $\gamma_{y^*,y}(b) = y$ nehmen können.

BEWEIS 1/3

i). Es reicht, stetig differenzierbare Kurven zu betrachten. Sei nun $\gamma : [a, b] \to U$ stetig differenzierbar. Dann gilt

$$\begin{aligned}
\int_\gamma \langle F(x),\, \mathrm{d}x\rangle &= \int_\gamma \langle \nabla\varphi(x),\, \mathrm{d}x\rangle \\
&= \int_a^b \big\langle \nabla(\gamma(t)), \gamma'(t)\big\rangle \,\mathrm{d}t \\
&= \int_a^b \sum_{i=1}^n \frac{\partial\varphi}{\partial x_i}(\gamma(t)) \cdot \gamma_i'(t)\,\mathrm{d}t \\
&= \int_a^b (\varphi\circ\gamma)'(t)\,\mathrm{d}t \\
&= \varphi(\gamma(a)) - \varphi(\gamma(b))\,,
\end{aligned}$$

wobei wir im vorletzten Schritt die mehrdimensionale Kettenregel und im letzten Schritt den Hauptsatz der Differential- und Integralrechnung benutzt haben.

BEWEIS 2/3

ii). Wir fixieren ein $y \in U$ und ein $\varepsilon > 0$ mit $B(y, \varepsilon) \subset U$. Damit ist $[y, y + he_i] \subset U$ für alle $h \in (0, \varepsilon]$ und $i = 1, \ldots, n$. Ist nun $\gamma_h : [0, h] \to U$ die Kurve, die durch

$$\gamma_h(t) := y + te_i$$

definiert ist und $\gamma_{y^*,y} : [-1, 0] \to U$ eine stückweise stetig differenzierbare Kurve mit $\gamma_{y^*,y}(-1) = y^*$ und $\gamma_{y^*,y}(0) = y$, so ist $\gamma_{y^*,y} \oplus \gamma_h : [-1, \varepsilon] \to U$ eine stückweise stetig differenzierbare Kurve mit Anfangspunkt y^* und Endpunkt $y + he_i$. Wegen der Wirbelfreiheit und (13.3.1), sowie (13.2.2) folgt dann

$$\begin{aligned}
\varphi(y + he_i) - \varphi(y) &= \int_{\gamma_{y^*,y} \oplus \gamma_h} \langle F(x),\, \mathrm{d}x \rangle - \int_{\gamma_{y^*,y}} \langle F(x),\, \mathrm{d}x \rangle \\
&= \int_{\gamma_h} \langle F(x),\, \mathrm{d}x \rangle \\
&= \int_0^h \big\langle F(\gamma_h(t)), \gamma_h'(t) \big\rangle \,\mathrm{d}t \\
&= \int_0^h \big\langle F(y + te_i), e_i \big\rangle \,\mathrm{d}t \\
&= \int_0^h F_i(y + te_i) \,\mathrm{d}t\,.
\end{aligned}$$

BEWEIS 3/3

Dies ergibt

$$\frac{\varphi(y + he_i) - \varphi(y)}{h} = \frac{1}{h} \int_0^h F_i(y + te_i)\,\mathrm{d}t \to F_i(y)$$

für $h \to 0$, wobei die Konvergenz analog zum Beweis des Hauptsatzes der Differential- und Integralrechnung gezeigt werden kann, d.h. wir benutzen

$$\left| \frac{1}{h} \int_0^h F_i(y + te_i)\,\mathrm{d}t \right| \leq \max_{s \in [0,h]} |F_i(y + te_i)| \to 0\,.$$

Da $i \in \{1, \ldots, n\}$ beliebig war, folgt $\nabla\varphi(y) = F(y)$.

BEISPIEL

Wir betrachten das zentrale Kraftfeld $F : \mathbb{R}^3 \setminus \{0\} \to \mathbb{R}^3$, dass durch

$$F(x) := \frac{x}{\|x\|^3}, \qquad x \neq 0$$

definiert ist. Dann definiert

$$\varphi(x) := -\|x\|^{-1}, x \neq 0$$

ein Potential von F, da für $h(r) := -r^{-1}$ die Formel (11.1.1) die Identität

$$\nabla\varphi(x) = \nabla h(\|\cdot\|)(x) = \frac{h'(\|x\|)}{\|x\|} \cdot x = F(x)$$

für alle $x \neq 0$ ergibt. Damit haben wir nach Satz 13.3.3 für jede stückweise stetig differenzierbare Kurve $\gamma : [a, b] \to \mathbb{R}^3 \setminus \{0\}$:

$$\int_\gamma \langle F(x), \, \mathrm{d}x\rangle = \frac{1}{\|\gamma(b)\|} - \frac{1}{\|\gamma(a)\|}.$$