der Bundeswehr
Universität München

Bachelor BAU, EIT, LRT

Mathematik 1 - 3

Bachelor EIT

Mathematik 4

Inhaltsverzeichnis

Teil I

Mathematik 1

Kapitel 1

Zahlen und Vektoren

1.1 Mengen und Abbildungen

Der Begriff „Menge“ gehört zu den Grundbegriffen der Mathematik. Eine formale Definition dessen, was Mathematiker unter „Menge“ verstehen, ist sehr problematisch; daher begnügen wir uns mit einer umgangssprachlichen Beschreibung, die von GEORG CANTOR, dem Begründer der Mengenlehre, stammt:

> *Unter einer Menge verstehen wir jede Zusammenfassung von bestimmten, wohlunterschiedenen Objekten unserer Anschauung oder unseres Denkens zu einem Ganzen.*

Wir werden Mengen mit Großbuchstaben bezeichnen. Die Objekte der Menge A werden Elemente von A genannt. Man schreibt $a \in A$ (bzw. $a \notin A$), wenn das Objekt a ein (bzw. kein) Element von A ist.
Neben der Aufzählung von Elementen bei endlichen Mengen zum Beispiel durch

$$A = \{a_1, a_2, \ldots, a_n\}, \tag{1.1}$$

wobei die Reihenfolge keine Rolle spielt, werden Mengen sehr häufig durch die Angabe einer Eigenschaft $\mathcal{E}$, die genau allen Elementen einer zu betrachtenden Menge zukommt, beschrieben. Zum Beispiel könnte man die Menge $\{-\sqrt{2}, +\sqrt{2}\}$ auch durch $\{x \in \mathbb{R}; x^2 = 2\}$ darstellen. Im allgemeinen werden Mengen, deren Elemente eine spezielle Eigenschaft besitzen, folgendermaßen notiert:

$$\{x; x \text{ erfüllt die Eigenschaft } \mathcal{E}\} \quad \text{bzw.} \quad \{x \in G; x \text{ erfüllt die Eigenschaft } \mathcal{E}\}. \tag{1.2}$$

Wie das obige Beispiel zeigt, kann der Term „x erfüllt die Eigenschaft $\mathcal{E}$“ auch durch eine mathematische Formel ausgedrückt werden. Die leere Menge $\emptyset$ besitzt kein Element.
Eine Menge B heißt Teilmenge von A (in Zeichen: $B \subseteq A$), wenn jedes Element von B auch Element von A ist. Zwei Mengen A, B heißen gleich, in Zeichen: $A = B$, falls sowohl $A \subseteq B$ als auch $B \subseteq A$ gilt. Eine Menge B heißt echte Teilmenge von A (in Zeichen: $B \subset A$), wenn $B \subseteq A$ und $B \neq A$ gilt. Für zwei Mengen A, B sind der Durchschnitt $A \cap B$, die Vereinigung $A \cup B$ und das Komplement $A \setminus B$ definiert durch

$$A \cap B := \{x; x \in A \text{ und } x \in B\}, \tag{1.3}$$

$$A \cup B := \{x; x \in A \text{ oder } x \in B\}, \tag{1.4}$$

$$A \setminus B := \{x; x \in A \text{ und } x \notin B\}. \tag{1.5}$$

Das Zeichen „:=“ heißt „definitionsgemäß gleich“, wobei der Doppelpunkt bei dem zu definierenden Ausdruck steht. Besitzen zwei Mengen A, B kein gemeinsames Element, so heißen sie „disjunkt“, d.h. es gilt $A \cap B = \emptyset$.
Als direktes (oder kartesisches) Produkt von n Mengen $A_1, \ldots, A_n$ bezeichnet man die Menge

$$A_1 \times \ldots \times A_n := \{(a_1, \ldots, a_n); a_i \in A_i \text{ für alle } i = 1, \ldots, n\}. \tag{1.6}$$

Die Elemente $(a_1, \ldots, a_n)$ dieser Menge werden als „geordnete n-Tupel“ bezeichnet. Der Begriff „geordnet“ ist dadurch gerechtfertigt, dass gilt:

$$(a_1, \ldots, a_n) = (a'_1, \ldots, a'_n) \text{ genau dann, wenn } a_i = a'_i \text{ für alle } i = 1, \ldots, n. \tag{1.7}$$

Beispiel(e) 1.1

- Für $A_1 = \{\alpha, \beta\}$ und $A_2 = \{1, 2, 3\}$ ergibt sich:

  $$A_1 \times A_2 = \{(\alpha, 1), (\alpha, 2), (\alpha, 3), (\beta, 1), (\beta, 2), (\beta, 3)\}$$

  und

  $$A_2 \times A_1 = \{(1, \alpha), (2, \alpha), (3, \alpha), (1, \beta), (2, \beta), (3, \beta)\}$$

- Für $A_1 = \{Latein, Physik\}$ und $A_2 = \{1, 2, 3, 4, 5, 6\}$ ergibt sich:

  $$\begin{aligned} A_1 \times A_2 \quad = \quad & \{(Latein, 1), (Latein, 2), (Latein, 3), (Latein, 4), \\ & (Latein, 5), (Latein, 6), (Physik, 1), (Physik, 2), \\ & (Physik, 3), (Physik, 4), (Physik, 5), (Physik, 6)\} \end{aligned}$$

Für den Fall $A_1 = A_2 = \ldots = A_n = A$ schreibt man

$$A^n \quad \text{für} \quad \underbrace{A \times \ldots \times A}_{n\text{-mal}}. \tag{1.8}$$

Seien A und B zwei Mengen. Eine Abbildung (oder Funktion) f von A nach B wird durch

$$f : A \to B, \quad x \mapsto f(x) \tag{1.9}$$

dargestellt und ist eine Vorschrift, die jedem $x \in A$ genau ein Element $f(x) \in B$ zuordnet. Man nennt A den Definitionsbereich, B den Wertebereich und die Elemente von A werden als Argumente der Abbildung f bezeichnet; $f(x)$ heißt das Bild von x unter f (oder der Funktionswert von f an der Stelle x). Die Funktion f heißt injektiv, falls aus $f(x_1) = f(x_2)$ folgt: $x_1 = x_2$ für alle $x_1, x_2 \in A$. Die Injektivität einer Funktion hängt nicht nur von der Abbildungsvorschrift ab, sondern auch vom Definitionsbereich, wie die folgenden Beispiele zeigen:

Beispiel(e) 1.2

- $f_1 : \{1, 2, 3, 4\} \to \{1, 4, 9, 16\}, \quad x \mapsto x^2$.
  Diese Funktion ist injektiv.
- $f_2 : \{-1, 1, 2, 3, 4\} \to \{1, 4, 9, 16\}, \quad x \mapsto x^2$.
  Diese Funktion ist nicht injektiv, denn $f_2(-1) = f_2(1)$.

Für jede Teilmenge $C \subseteq A$ bezeichnet man für eine Abbildung f mit Definitionsbereich A die Menge

$$f(C) := \{f(x); x \in C\} \tag{1.10}$$

als Bild von C unter f. Das Bild $f(A)$ des gesamten Definitionsbereichs von f unter f wird als Wertemenge bezeichnet. Ist für unsere Abbildung

$$f : A \to B, \quad x \mapsto f(x) \tag{1.11}$$

die Wertemenge $f(A)$ gleich der Menge B, so heißt die Funktion f surjektiv.

Beispiel(e) 1.3

- $f_3 : \{1,2,3,4\} \to \{1,4,9,16\}, \quad x \mapsto x^2$.
  Diese Funktion ist surjektiv.
- $f_4 : \{1,2,3,4\} \to \{1,2,\ldots,16\}, \quad x \mapsto x^2$.
  Diese Funktion ist nicht surjektiv, denn $\{1,4,9,16\} \neq \{1,2,\ldots,16\}$.
- $f_5 : \{1,2,3,4\} \to \{1,4,9\}, \quad x \mapsto x^2$.
  Diese Abbildungsvorschrift ist nicht korrekt formuliert, denn B ist eine echte Teilmenge von $f_5(A)$.

Funktionen, die injektiv und surjektiv sind (in unseren Beispielen die Funktion f_1 bzw. f_3) werden als bijektiv bezeichnet. Für eine bijektive Funktion

$$f : A \to B, \quad x \mapsto f(x) \tag{1.12}$$

gibt es somit zu jedem $y \in B$ genau ein $x \in A$ mit: $f(x) = y$. Es existiert also eine weitere Abbildung

$$f^{-1} : B \to A, \quad y \mapsto f^{-1}(y) \tag{1.13}$$

mit: $f^{-1}(f(x)) = x$ für alle $x \in A$ und $f(f^{-1}(y)) = y$ für alle $y \in B$. Man bezeichnet f^{-1} als Umkehrfunktion oder Umkehrabbildung von f. Die auf jeder Menge A definierte Funktion

$$\mathrm{Id}_A : A \to A, \quad x \mapsto x \tag{1.14}$$

heißt Identität auf A. Zwei Abbildungen

$$f : A \to B, \quad x \mapsto f(x) \quad \text{und} \quad g : C \to D, \quad z \mapsto g(z) \tag{1.15}$$

sind genau dann gleich (in Zeichen: $f = g$), falls $A = C$, $B = D$ und $f(x) = g(x)$ für alle $x \in A$. An dieser Definition der Gleichheit für Abbildungen sieht man, dass die Darstellung der Abbildungsvorschrift (etwa mit Hilfe des Zeichens „x“ oder des Zeichens „z“) keine Rolle spielt. Durch Einschränkung des Definitionsbereichs einer Abbildung $f : A \to B, x \mapsto f(x)$ auf eine Teilmenge $A_0 \subseteq A$ erhält man die Restriktion

$$f_{|A_0} : A_0 \to B, x \mapsto f(x). \tag{1.16}$$

1.2 Reelle Zahlen

Die Menge der reellen Zahlen bezeichnen wir mit $\mathbb{R}$. Besondere Teilmengen von $\mathbb{R}$ sind

$$\mathbb{N} := \{1,2,3\ldots\} \quad \text{Menge der natürlichen Zahlen} \tag{1.17}$$

$$\mathbb{N}_0 := \mathbb{N} \cup \{0\} = \{0,1,2,3\ldots\} \tag{1.18}$$

$$\mathbb{Z} := \{\ldots -3,-2,-1,0,1,2,3\ldots\} \quad \text{Menge der ganzen Zahlen} \tag{1.19}$$

$$\mathbb{Q} := \left\{\frac{m}{n};\, m,n \in \mathbb{Z}, n \neq 0\right\} \quad \text{Menge der rationalen Zahlen.} \tag{1.20}$$

Neben den rationalen Zahlen gibt es in $\mathbb{R}$ die irrationalen Zahlen, die durch unendliche, nichtperiodische Dezimalbrüche gekennzeichnet sind. Irrationale Zahlen sind zum Beispiel $\sqrt{2}$ oder π. Neben der Aufteilung der reellen Zahlen in rationale und irrationale Zahlen gibt es auch die Aufteilung in algebraische und transzendente Zahlen.
Die reellen Zahlen sind durch eine Vergleichsrelation $\leq$ vollständig geordnet, wobei für $x, y \in \mathbb{R}$ gilt:

$$x \leq y \quad :\Longleftrightarrow \quad (x-y) \leq 0 \quad \text{bzw.} \tag{1.21}$$

$$x < y \quad :\Longleftrightarrow \quad (x-y) < 0. \tag{1.22}$$

Das Zeichen „$:\Longleftrightarrow$“ bedeutet „definitionsgemäß genau dann, wenn “. Mit Hilfe der Ordnungsstruktur „$\leq$“ betrachten wir nun Intervalle reeller Zahlen:

- abgeschlossenes Intervall mit $a, b \in \mathbb{R}$, $a \leq b$:

$$[a, b] := \{x \in \mathbb{R};\, a \leq x \leq b\} \tag{1.23}$$

- offenes Intervall mit $a, b \in \mathbb{R}$, $a < b$:

$$(a, b) := \{x \in \mathbb{R};\, a < x < b\} \tag{1.24}$$

- halboffene Intervalle mit $a, b \in \mathbb{R}$, $a < b$:

$$(a, b] := \{x \in \mathbb{R};\, a < x \leq b\} \text{ und } [a, b) := \{x \in \mathbb{R};\, a \leq x < b\}. \tag{1.25}$$

Es ist üblich, noch zwei Symbole „$-\infty$“ und „∞“ (minus unendlich, unendlich) einzuführen und $-\infty < \infty$ festzulegen. Man definiert für $a \in \mathbb{R}$:

$$\begin{aligned}
(-\infty, a] &:= \{x \in \mathbb{R};\, x \leq a\} && (1.26)\\
(-\infty, a) &:= \{x \in \mathbb{R};\, x < a\} && (1.27)\\
[a, \infty) &:= \{x \in \mathbb{R};\, a \leq x\} && (1.28)\\
(a, \infty) &:= \{x \in \mathbb{R};\, a < x\}. && (1.29)
\end{aligned}$$

Eine Menge S reeller Zahlen heißt nach oben beschränkt, wenn es eine reelle Zahl b gibt, sodass $S \subseteq (-\infty, b]$. In diesem Fall heißt b eine obere Schranke von S. Analog dazu nennt man eine reelle Zahl a eine untere Schranke von S und S nach unten beschränkt, falls $S \subseteq [a, \infty)$. Ist $S \subset \mathbb{R}$ nach oben beschränkt, so nennt man die kleinste obere Schranke s das Supremum von S (in Zeichen: $s = \sup\{S\}$). Offensichtlich braucht $\sup\{S\}$ nicht zur Menge S zu gehören. Es ist nicht selbstverständlich, dass die Menge der oberen Schranken von $S \subseteq (-\infty, b]$ eine kleinste Zahl enthält. Dies gehört zu den Grundannahmen, die man bei der axiomatischen Einführung der reellen Zahlen fordert. Damit hat auch jede nach unten beschränkte Menge $U \subseteq [a, \infty)$ eine größte untere Schranke (ein Infimum $\inf\{U\}$).

Beispiel(e) 1.4

- $\sup\{[a, b]\} = \sup\{[a, b)\} = b$
- $\sup\{\{x \in \mathbb{Q};\, x^2 < 2\}\} = \sqrt{2}$
- $\inf\{\{1 + \frac{1}{n};\, n \in \mathbb{N}\}\} = 1$.

Der Betrag $|a|$ einer reellen Zahl ist definiert durch

$$|a| := \begin{cases} a & \text{falls} \quad a \geq 0 \\ -a & \text{falls} \quad a < 0 \end{cases}. \tag{1.30}$$

Offensichtlich gilt für $a, b \in \mathbb{R}$:

$$-|a| \quad \leq \quad a \quad \leq \quad |a|, \tag{1.31}$$

$$|-a| \quad = \quad |a|, \tag{1.32}$$

$$|ab| \quad = \quad |a| \cdot |b|, \tag{1.33}$$

$$\left|\frac{a}{b}\right| \quad = \quad \frac{|a|}{|b|}, \quad b \neq 0, \tag{1.34}$$

$$|a| \leq b \quad \Longleftrightarrow \quad -b \leq a \leq b. \tag{1.35}$$

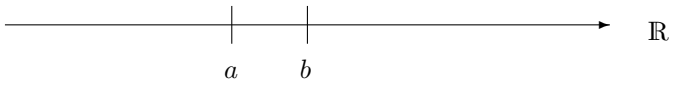
Aus $-|a| \leq a \leq |a|$ und $-|b| \leq b \leq |b|$ folgt:

$$-(|a| + |b|) \leq a + b \leq |a| + |b| \tag{1.36}$$

und damit die Dreiecksungleichung

$$|a + b| \leq |a| + |b|. \tag{1.37}$$

Betrachtet man zwei reelle Zahlen a, b auf der Zahlengeraden,



so ist $|a - b|$ der Abstand der zu a und b gehörenden Punkte auf der Zahlengeraden.

1.3 Die vollständige Induktion

Die vollständige Induktion ist ein Beweisschema, mit dessen Hilfe häufig Aussagen $A(n)$, die für jede ganze Zahl $n \geq n_0$ gelten sollen, bewiesen werden. Ein Beweis mit vollständiger Induktion gliedert sich in zwei Schritte:

1. Induktionsbeginn: Man zeigt, dass $A(n)$ für $n = n_0$ gilt.
2. Schluss von n auf $(n+1)$: Für beliebiges $n \geq n_0$ setzt man die Gültigkeit von $A(n)$ voraus (Induktionsvoraussetzung) und leitet daraus die Gültigkeit von $A(n+1)$ ab.

Können beide Schritte ausgeführt werden, dann gilt $A(n)$ für alle $n \geq n_0$; denn nach dem ersten Schritt gilt $A(n_0)$, nach dem zweiten Schritt $A(n_0+1)$ und ebenso (immer wieder Schluss von n auf $(n+1)$) $A(n_0+2)$, $A(n_0+3)$, ...

Beispiel(e) 1.5

- $1+2+3+\cdots+n=\frac{n(n+1)}{2} \quad (n\in\mathbb{N})$

  Beweis mit vollständiger Induktion:
  Induktionsbeginn $n=1$**:** $\quad 1=\frac{1\cdot 2}{2} \quad \surd$
  Schluss von n **auf** $(n+1)$**:** Sei $n\in\mathbb{N}$ beliebig und die Formel richtig für dieses n. Dann folgt:

$$1+2+3+\cdots+n+(n+1)=\frac{n(n+1)}{2}+(n+1)=\frac{(n+1)(n+2)}{2}$$

  Das ist die zu beweisende Formel für $(n+1)$. Sie gilt somit für alle $n\in\mathbb{N}$.

  q.e.d.

- Für alle reellen Zahlen $h\geq -1$ und alle $n\in\mathbb{N}$ gilt:

$$(1+h)^n\geq 1+n\cdot h \tag{1.38}$$

  Beweis mit vollständiger Induktion:
  Induktionsbeginn $n=1$**:** $(1+h)^1\geq 1+h \quad \surd$
  Schluss von n **auf** $(n+1)$**:** Aus $(1+h)^n\geq 1+n\cdot h$ folgt durch Multiplikation mit $(1+h)\geq 0$:

$$(1+h)^{n+1}\geq (1+n\cdot h)\cdot(1+h)=1+(n+1)\cdot h+nh^2\geq 1+(n+1)\cdot h$$

  q.e.d.

Das Prinzip der vollständigen Induktion kann auch zur Definition einer Größe D_n für alle ganzen Zahlen $n\geq n_0$ verwendet werden. Man spricht in diesem Zusammenhang von einer rekursiven Definition:

1. D_{n_0} wird festgelegt
2. Man sieht D_k für $n_0\leq k\leq n$ als bekannt an und drückt D_{n+1} durch $D_{n_0},\ldots,D_n$ aus.

Beispiel(e) 1.6

- Die Potenzen a^n für eine beliebige Basis $a\in\mathbb{R}$ und Exponenten $n\in\mathbb{N}_0$ werden definiert durch

$$a^0:=1, \quad a^{n+1}:=a\cdot a^n. \tag{1.39}$$

- Die Fakultäten $n!$ (lies: n Fakultät) werden definiert durch

$$0!:=1, (n+1)!:=n!(n+1). \tag{1.40}$$

  So ist etwa 6! = 720 und 9! = 362880

Als Permutation einer Menge A bezeichnet man jede bijektive Abbildung

$$f:A\to A$$

von A auf sich. Ist A endlich mit den n verschiedenen Elementen $\{a_1,\ldots,a_n\}$, so interpretiert man eine Permutation von A als Zuordnung der a_i zu n verschiedenen, mit $1,\ldots,n$ durchnume-

rierten Plätzen, so dass auf jeden Platz genau ein Element kommt. Für jedes $n \in \mathbb{N}$ gilt der Satz:

„Jede n-elementige Menge besitzt genau $n!$ verschiedene Permutationen“.

Der Beweis wird mit vollständiger Induktion geführt (Übung).

Für die Summe und das Produkt der Zahlen $a_m, a_{m+1}, \ldots, a_n \in \mathbb{R}$ schreibt man:

$$a_m + a_{m+1} + \cdots + a_n = \sum_{k=m}^{n} a_k, \qquad a_m \cdot a_{m+1} \cdot \ldots \cdot a_n = \prod_{k=m}^{n} a_k. \tag{1.41}$$

Eine rekursive Definition ersetzt die Punkte:

$$\sum_{k=m}^{m} a_k := a_m, \qquad \sum_{k=m}^{n+1} a_k := \left(\sum_{k=m}^{n} a_k \right) + a_{n+1}. \tag{1.42}$$

Man beachte die Unabhängigkeit vom Summationsindex

$$\sum_{k=m}^{n} a_k = \sum_{i=m}^{n} a_i, \tag{1.43}$$

ferner:

$$\sum_{k=m}^{n} a_k = \sum_{k=m}^{l} a_k + \sum_{k=l+1}^{n} a_k \quad (m \leq l < n). \tag{1.44}$$

Analoges gilt für das Produktzeichen.

1.4 Binomialkoeffizienten

Für zwei ganze Zahlen n, k mit $0 \leq k \leq n$ bezeichnet man die natürliche Zahl

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} \qquad \text{(lies: } n \text{ über } k\text{)} \tag{1.45}$$

als Binomialkoeffizient. Der Binomialkoeffizient gibt die Anzahl der verschiedenen k-elementigen Teilmengen einer Menge mit n Elementen an. Es gibt also genau $\binom{n}{k}$ Möglichkeiten, aus n Objekten genau k auszuwählen.

Beispiel(e) 1.7
Im Lotto „6 aus 49“ gibt es

$$\binom{49}{6} := 13983816 \tag{1.46}$$

verschiedene Möglichkeiten, „6 Richtige“ zu ziehen.

Aus der Definition der Binomialkoeffizienten folgen sofort einige Eigenschaften:

$$\binom{n}{0} = \binom{n}{n} = 1, \quad \binom{n}{1} = \binom{n}{n-1} = n, \quad \binom{n}{2} = \binom{n}{n-2} = \frac{n(n-1)}{2}, \tag{1.47}$$

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \tag{1.48}$$

Damit lassen sich die Binomialkoeffizienten durch Addition berechnen:
Beginnend mit $n = 1$ schreibt man die Zahlen $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$ in eine Zeile, dabei ist $\binom{n}{0} = \binom{n}{n} = 1$ festgelegt. Die Zahlen $\binom{n+1}{0}, \ldots, \binom{n+1}{n+1}$ werden gemäß der Formel

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \tag{1.49}$$

berechnet. Auf diese Weise entsteht das sogenannte Pascal-Dreieck:

$$\begin{array}{ccccccccccc}
\binom{0}{0} & & & & & & & & & & \\
\binom{1}{0} & & \binom{1}{1} & & & & & & & & \\
\binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & & & \\
\binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & & \\
\binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5}
\end{array}
\qquad
\begin{array}{ccccccccccc}
 & & & & 1 & & & & & & \\
 & & & 1 & & 1 & & & & & \\
 & & 1 & & 2 & & 1 & & & & \\
 & 1 & & 3 & & 3 & & 1 & & & \\
1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}
\tag{1.50}$$

Dieses Schema ist nach Blaise Pascal (1623-1662), einem bedeutenden Universalgenie, benannt. Zu seinen herausragenden Leistungen gehören:

- Untersuchungen zum Phänomen des Luftdruckes (Barometerversuche). Die physikalische Größe „Druck“ wird in Pascal [1 Pa] gemessen.
- Konstruktion und Herstellung der ersten funktionstüchtigen Rechenmaschine. Die Programmiersprache PASCAL ist nach Blaise Pascal benannt.
- Planung und Durchführung der ersten Omnibuslinie (Einweihung am 18. März 1662).
- Zusammen mit Fermat Begründung der Stochastik.
- Pascal ist mit Augustinus bis heute der bedeutendste christliche Philosoph.

In Verbindung mit Binomialkoeffizienten ist die folgende „binomische Formel“ von Interesse: Für beliebige reelle Zahlen $a, b \in \mathbb{R}$ und jede ganze Zahl $n \geq 0$ gilt:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k. \tag{1.51}$$

Der Beweis wird mit vollständiger Induktion geführt:

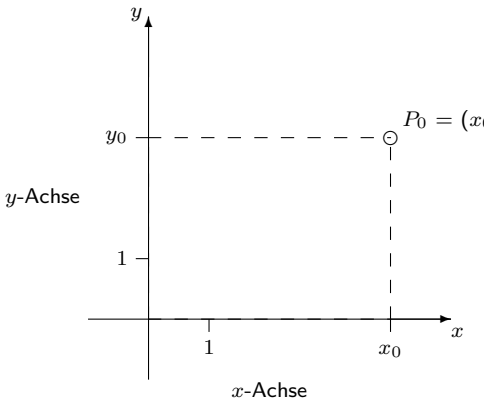
Induktionsbeginn: $n = 0$**:** $(a+b)^0 = 1 = \binom{0}{0} a^0 \cdot b^0$
Schluss von n **auf** $(n+1)$**:**

$$
\begin{aligned}
(a+b)^{n+1} &= (a+b)^n \cdot (a+b) \qquad \overset{\text{Ind.-Voraussetzung}}{\overset{\downarrow}{=}} \\
&= \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k} \cdot b^k \cdot (a+b) = \\
&= \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k+1} \cdot b^k + \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k} \cdot b^{k+1} = \\
&= a^{n+1} + \sum_{k=1}^{n} \binom{n}{k} \cdot a^{n-k+1} \cdot b^k + \sum_{k=0}^{n-1} \binom{n}{k} \cdot a^{n-k} \cdot b^{k+1} + b^{n+1} = \\
&= a^{n+1} + \sum_{k=1}^{n} \binom{n}{k} \cdot a^{n-k+1} \cdot b^k + \\
&\quad + \sum_{k=1}^{n} \binom{n}{k-1} \cdot a^{n-(k-1)} \cdot b^k + b^{n+1} = \\
&= a^{n+1} + \sum_{k=1}^{n} \left\{ \binom{n}{k} + \binom{n}{k-1} \right\} \cdot a^{n-k+1} \cdot b^k + b^{n+1} = \\
&= \binom{n+1}{0} a^{n+1} + \sum_{k=1}^{n} \binom{n+1}{k} \cdot a^{n-k+1} b^k + \binom{n+1}{n+1} b^{n+1} = \\
&= \sum_{k=0}^{n+1} \binom{n+1}{k} \cdot a^{n+1-k} \cdot b^k
\end{aligned}
$$

q.e.d.

1.5 Die Ebene

In einer Ebene E entsteht ein kartesisches Koordinatensystem (benannt nach René Descartes (1596-1650), philosophischer Gegenspieler von Blaise Pascal) durch Vorgabe eines Punktes O und zweier aufeinander senkrechter Zahlengeraden, der x-Achse und der y-Achse, deren Nullpunkt jeweils in O liegt. Dabei muss die y-Achse durch eine positive Drehung (gegen den Uhrzeigersinn) aus der x-Achse hervorgehen. Fällt man für einen beliebigen Punkt P_0 in der Ebene die Lote auf die Achsen, so bestimmen die beiden Fußpunkte die x- bzw. y-Koordinate x_0 bzw. y_0 von P_0 und man schreibt $P_0 = (x_0, y_0)$. Der Punkt $O = (0, 0)$ heißt Nullpunkt oder Ursprung des Koordinatensystems.



(1.52)

Nach Festlegung eines kartesischen Koordinatensystems gibt es zu jedem Zahlenpaar $(x, y) \in \mathbb{R}^2$ genau einen Punkt $P \in E$ mit $P = (x, y)$ und umgekehrt.
Man kann somit Teilmengen von $\mathbb{R}^2$ (etwa Lösungsmengen von Gleichungen bzw. Ungleichungen) als Punktmengen in E veranschaulichen und umgekehrt geometrische Gebilde durch Funktionen, Gleichungen oder Ungleichungen als Teilmengen des $\mathbb{R}^2$ beschreiben.
Sei $I \subseteq \mathbb{R}$ und $f : I \to \mathbb{R}$ eine Funktion. Die Punktmenge

$$G_f := \{(x, y);\ x \in I \ \text{ und } \ y = f(x)\} = \{(x, f(x));\ x \in I\} \tag{1.53}$$

in der mit einem kartesischen Koordinatensystem versehenen Ebene heißt der Graph der Funktion f oder die Kurve $y = f(x)$.

Beispiel(e) 1.8

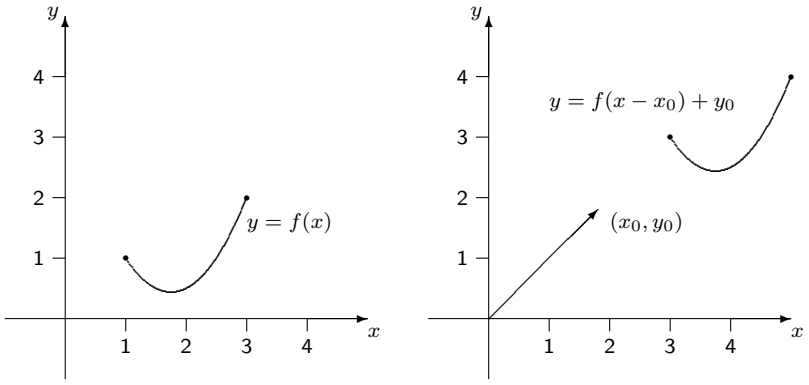
- Für eine gegebene Funktion

  $$f : [a, b] \to \mathbb{R}, \quad x \mapsto f(x)$$

  entsteht der Graph der Funktion

  $$g : [a + x_0, b + x_0] \to \mathbb{R} \quad x \mapsto f(x - x_0) + y_0 \tag{1.54}$$

  mit $(x_0, y_0) \in \mathbb{R}^2$ durch diejenige Parallelverschiebung des Graphen von f, die einen Punkt (x, y) nach $(x + x_0,\ y + y_0)$ verschiebt. Dabei bleibt das Koordinatensystem fest.



(1.55)

- Sei $C \subseteq E$ eine Kurve und - nach Wahl eines kartesischen Koordinatensystems - $X = (x, y)$ ein beliebiger Punkt auf C.
  Wird die zwischen x und y bestehende Abhängigkeit auf die Form

  $$F(x, y) = 0 \tag{1.56}$$

  gebracht, d.h. $C = \{(x, y);\ F(x, y) = 0\}$, dann nennt man „$F(x, y) = 0$“ die Gleichung der Kurve C:

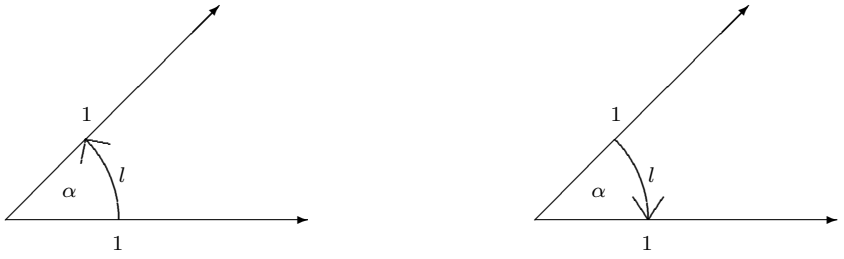
  (i) die Gerade durch die Punkte $A = (a_1, a_2)$ und $B = (b_1, b_2)$ $(A \neq B)$ hat die Gleichung:

  $$(b_2 - a_2)(x - a_1) - (b_1 - a_1)(y - a_2) = 0. \tag{1.57}$$

  (ii) die Kreisfläche (mit Rand) um $A = (a_1, a_2)$ mit dem Radius r wird beschrieben durch die Ungleichung

  $$(x - a_1)^2 + (y - a_2)^2 \leq r^2 \tag{1.58}$$

Ein Winkel α entsteht durch Drehung eines Zeigers um einen gegebenen Punkt der Ebene. Die Länge des zugehörigen Einheitsbogens sei l.



(1.59)

Wir nennen l, bzw. $-l$, das Bogenmaß von α und schreiben $\alpha = l$, bzw. $\alpha = -l$, wenn die Drehung im positiven (gegen Uhrzeigersinn) bzw. im Uhrzeigersinn erfolgt. Ein Winkel von α° besitzt das Bogenmaß

$$x = \frac{\pi}{180^\circ} \cdot \alpha \tag{1.60}$$

mit der Kreiszahl

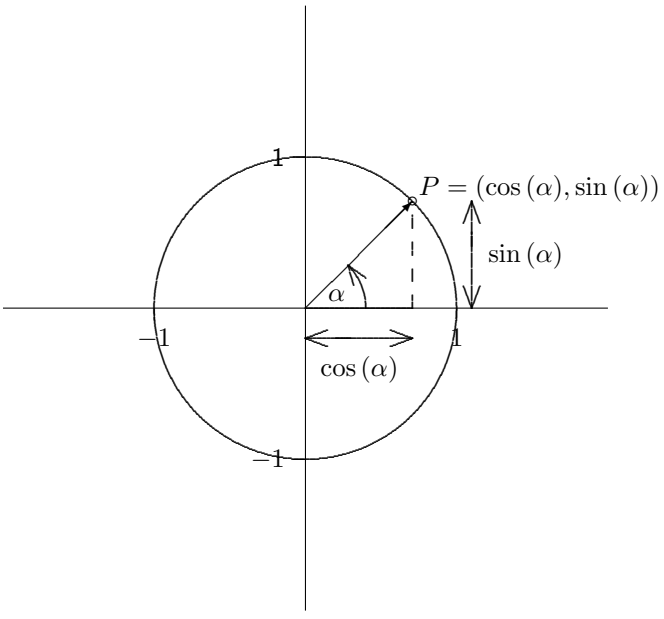
$$\pi = 3.14159265358979\ldots \quad . \tag{1.61}$$

Wird in der mit kartesischen Koordinaten versehenen Ebene der vom Ursprung zum Punkt $(1, 0)$ weisende Zeiger um den Winkel α gedreht (α im Bogenmaß), dann bewegt sich die Spitze auf dem Einheitskreis (um O) bis zu einem Punkte P, dessen Koordinaten mit $P = (\cos(\alpha), \sin(\alpha))$ bezeichnet werden. Die derart für alle $\alpha \in \mathbb{R}$ erklärten Funktionen

$$\cos : \mathbb{R} \to [-1, 1], \quad \alpha \mapsto \cos(\alpha) \tag{1.62}$$

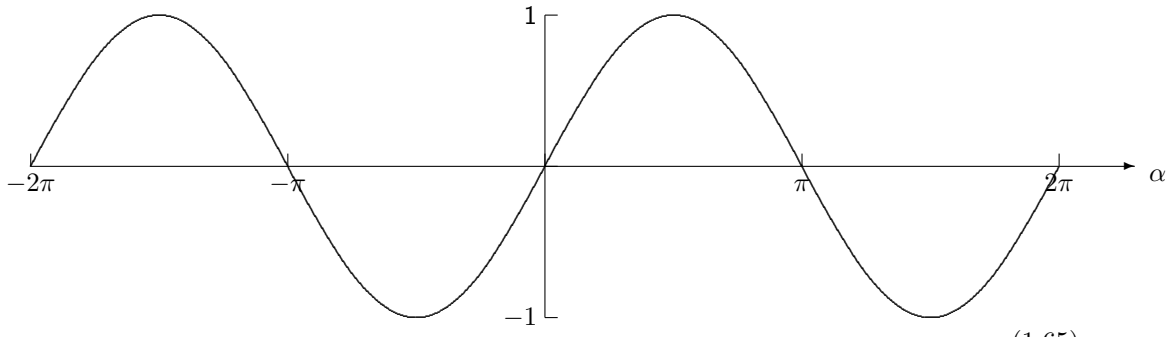
$$\sin : \mathbb{R} \to [-1, 1], \quad \alpha \mapsto \sin(\alpha) \tag{1.63}$$

heißen Cosinus- bzw. Sinusfunktionen. Dabei ist zu beachten, dass mit jeder vollen Umdrehung das Bogenmaß des Winkels um 2π, also dem Umfang des Einheitskreises, vergrößert wird.



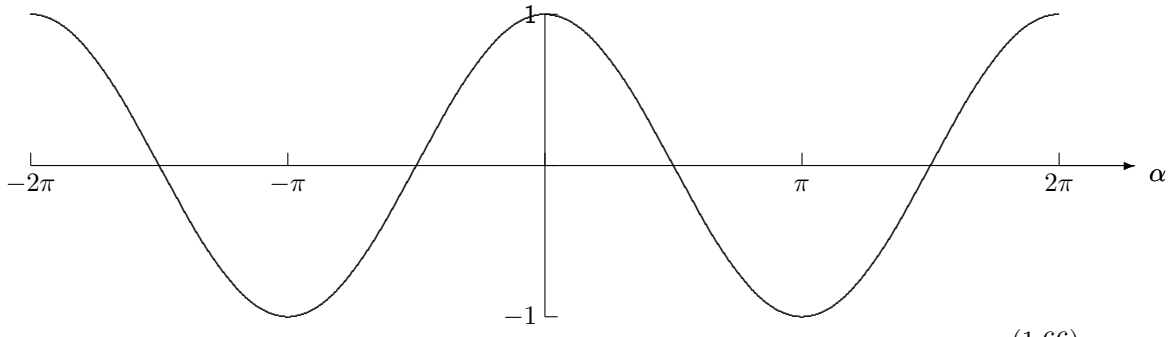
(1.64)

Graph der Sinusfunktion:



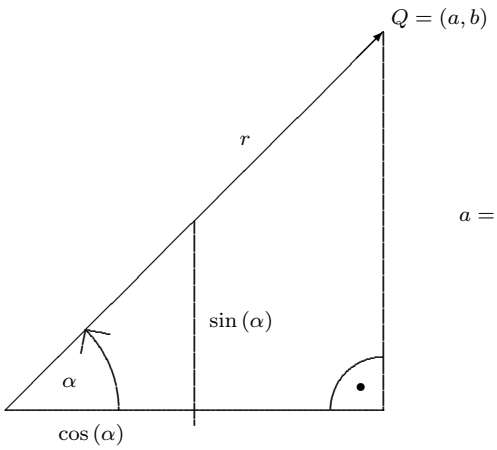
(1.65)

Graph der Cosinusfunktion:



(1.66)

Hat der sich drehende Zeiger die Länge r, so gilt



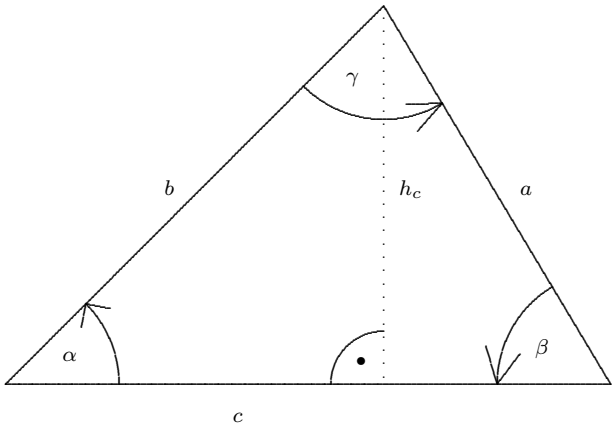
$$a = r \cdot \cos(\alpha), \quad b = r \cdot \sin(\alpha)$$

(1.67)

Somit gilt im rechtwinkligen Dreieck mit der Hypotenuse r, der Ankathete a und der Gegenkathete b:

$$\frac{a}{r} = \cos(\alpha), \quad \frac{b}{r} = \sin(\alpha). \tag{1.68}$$

Für ein beliebiges Dreieck



(1.69)

gilt ferner der Cosinussatz:

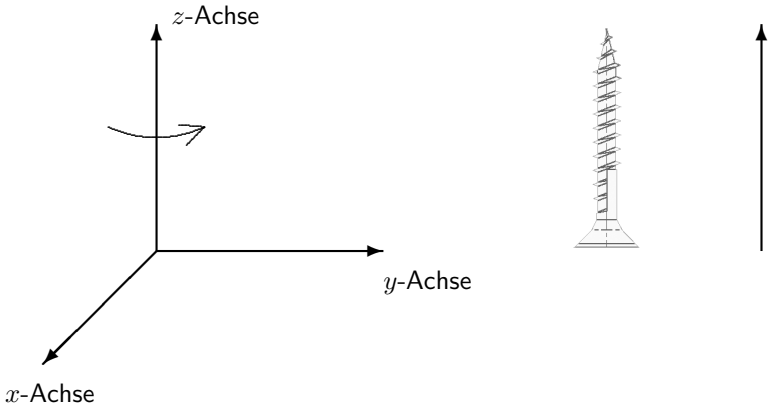
$$a^2 = b^2 + c^2 - 2bc\cos(\alpha) \tag{1.70}$$

und der Sinussatz:

$$\frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c}\ . \tag{1.71}$$

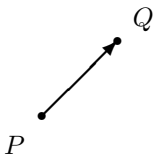
1.6 Vektoren

Kartesische Koordinatensysteme im Raum bestehen aus drei sich in einem Punkt O (Nullpunkt oder Ursprung) rechtwinklig schneidenden Zahlengeraden gleicher Längeneinheit und jeweils mit dem Nullpunkt im Schnittpunkt O. Man bezeichnet sie als x-, y- und z-Achse derart, dass diese Achsen ein Rechtssystem bilden, d.h. die Drehung der positiven x-Achse um $90°$ in die positive y-Achse muss zusammen mit einer Verschiebung in Richtung der positiven z-Achse eine Rechtsschraubung darstellen.

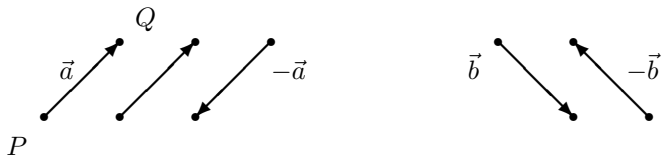


(1.72)

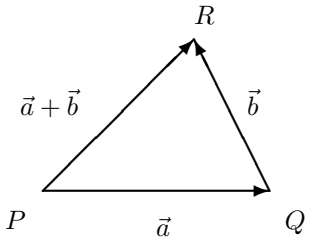
Zu je zwei Punkten P und Q des Raumes gibt es genau eine Parallelverschiebung aller Punkte, die P nach Q abbildet. Diese Verschiebung wird mit $\vec{PQ}$ bezeichnet und heißt „Vektor von P nach Q“ . Das lateinische „vector“ bedeutet „Träger“ : $\vec{PQ}$ trägt P nach Q. Der Vektor $\vec{PQ}$ wird dargestellt durch einen Pfeil von P nach Q.



Wird unter $\vec{PQ}$ ein Punkt R nach S verschoben, dann hat offenbar $\vec{RS}$ dieselbe Wirkung wie $\vec{PQ}$, d.h. $\vec{RS} = \vec{PQ}$. Die Länge des Pfeiles $\vec{PQ}$ ist der Abstand zwischen P und Q. Zwei gleich lange und gleich gerichtete Pfeile im Raum stellen somit denselben Vektor dar.
Anstelle „der Pfeil $\vec{a}$ repräsentiert einen Vektor“ sagt man „$\vec{a}$ ist ein Vektor“ und berücksichtigt, dass $\vec{a}$ im Raum frei parallel verschoben werden kann und nicht an einen festen Anfangspunkt gebunden ist.
Den zu $\vec{a}$ gleich langen, aber entgegengesetzt gerichteten Vektor bezeichnen wir mit $-\vec{a}$; er macht die durch $\vec{a}$ bewirkte Parallelverschiebung rückgängig. Für $\vec{a} = \vec{PQ}$ gilt: $-\vec{a} = \vec{QP}$:



Üblicherweise wird der Nullvektor $\vec{O}$ eingeführt, der eine Verschiebung bezeichnet, bei der sich nichts bewegt; d.h. $\vec{O} = \vec{PP}$.
Führt man zwei Parallelverschiebungen, erst $\vec{a} = \vec{PQ}$, dann $\vec{b} = \vec{QR}$, hintereinander aus, so ergibt sich wieder eine Parallelverschiebung, $\vec{c} = \vec{PR}$

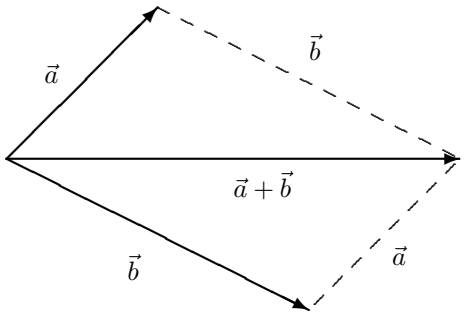


(1.73)

Wir nennen $\vec{c}$ die Summe von $\vec{a}$ und $\vec{b}$ und schreiben dafür

$$\vec{c} = \vec{a} + \vec{b}. \tag{1.74}$$

Haben die Pfeile $\vec{a}$ und $\vec{b}$ den gleichen Anfangspunkt, so gewinnt man die Summe $\vec{a} + \vec{b}$ (geometrisch) nach der Parallelogrammregel:

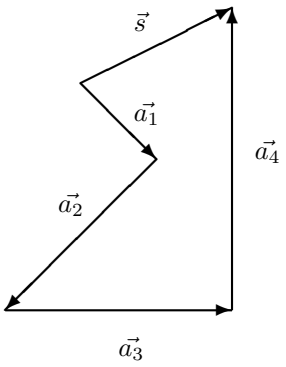


(1.75)

Offenbar gelten für beliebige Vektoren $\vec{a}, \vec{b}, \vec{c}$ die folgenden Rechenregeln:

$$\begin{aligned} \vec{a} + (-\vec{a}) &= \vec{O} \\ \vec{a} + \vec{O} &= \vec{a} \\ \vec{a} + \vec{b} &= \vec{b} + \vec{a} \quad \text{(Kommutativgesetz)} \\ (\vec{a} + \vec{b}) + \vec{c} &= \vec{a} + (\vec{b} + \vec{c}) \quad \text{(Assoziativgesetz)} \end{aligned} \tag{1.76}$$

Die Summe $\vec{a_1} + \vec{a_2} + \ldots + \vec{a_n}$ ist derjenige Vektor $\vec{s}$, der vom Anfangspunkt zum Endpunkt der aus $\vec{a_1}, \vec{a_2}, \ldots, \vec{a_n}$ gebildeten Vektorkette zeigt.



(1.77)

Die Differenz von Vektoren wird erklärt durch

$$\vec{a} - \vec{b} := \vec{a} + (-\vec{b}). \tag{1.78}$$

Zu einer reellen Zahl $\alpha \geq 0$ und einem Vektor $\vec{a}$ bezeichnet $\alpha \cdot \vec{a}$ (oder kürzer: $\alpha\vec{a}$) denjenigen Vektor, der dieselbe Richtung und Orientierung hat wie $\vec{a}$, aber die α-fache Länge. Im Fall $\alpha < 0$ setzt man $\alpha \cdot \vec{a} = -(|\alpha| \cdot \vec{a})$. Sonderfälle dieser Festlegung sind $0\vec{a} = \vec{O}$ und $\alpha\vec{O} = \vec{O}$ für jede reelle Zahl α und jeden Vektor $\vec{a}$. Es gelten die folgenden Rechenregeln ($\alpha, \beta \in \mathbb{R}$):

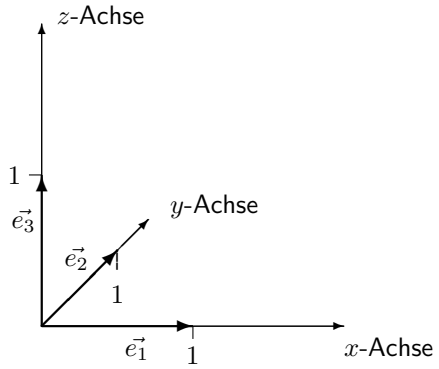
$$\begin{aligned} \alpha(\beta\vec{a}) &= (\alpha\beta)\vec{a} \\ \alpha(\vec{a} + \vec{b}) &= \alpha\vec{a} + \alpha\vec{b} \\ (\alpha + \beta)\vec{a} &= \alpha\vec{a} + \beta\vec{a}. \end{aligned} \tag{1.79}$$

Die Länge eines Vektors $\vec{a}$ (das ist für $\vec{a} = \vec{PQ}$ die Länge der Strecke $[PQ]$) nennt man seinen „Betrag“ und schreibt dafür $|\vec{a}|$ oder $\|\vec{a}\|$.
Offensichtlich gilt:

$$|\alpha\vec{a}| = |\alpha| \cdot |\vec{a}|, \quad \text{insbesondere} \quad |-\vec{a}| = |\vec{a}| \tag{1.80}$$

$$|\vec{a} + \vec{b}| \leq |\vec{a}| + |\vec{b}| \quad \text{(Dreiecksungleichung).} \tag{1.81}$$

Der Nullvektor hat keine positive Länge, d.h. $|\vec{O}| = 0$. Ein Vektor vom Betrag 1 heißt „Einheitsvektor“.
Legt man im Raum ein kartesisches Koordinatensystem fest, so werden neben dem Ursprung O gleichzeitig drei Einheitsvektoren $\vec{e_1}, \vec{e_2}, \vec{e_3}$ in Richtung der positiven x-, y- und z-Achse ausgezeichnet.



(1.82)

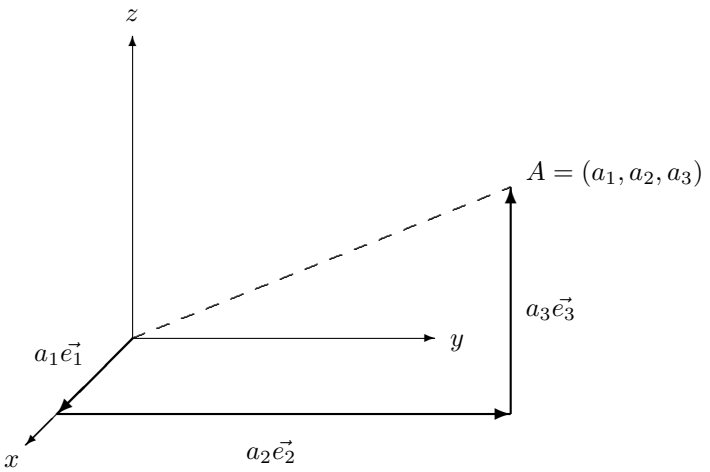
Wir nennen $(\vec{e_1}, \vec{e_2}, \vec{e_3})$ eine kartesische Basis und bezeichnen das Koordinatensystem mit $(O, \vec{e_1}, \vec{e_2}, \vec{e_3})$. Der Vektor $\vec{a} = \vec{OA}$ heißt Ortsvektor des Punktes $A = (a_1, a_2, a_3)$; er ist eindeutig zerlegbar als Summe

$$\vec{a} = a_1\vec{e_1} + a_2\vec{e_2} + a_3\vec{e_3}. \tag{1.83}$$

Abkürzend schreibt man bei festgelegtem Koordinatensystem

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} :\Longleftrightarrow \vec{a} = a_1\vec{e_1} + a_2\vec{e_2} + a_3\vec{e_3} = \vec{OA} \quad \text{mit} \quad A = (a_1, a_2, a_3). \tag{1.84}$$

Man nennt $a_i\vec{e_i}$ die Komponente von $\vec{a}$ in $\vec{e_i}$-Richtung $(i = 1, 2, 3)$ und die Zahlen $a_i \in \mathbb{R}$ die Koordinaten des Vektors $\vec{a}$ bezüglich $(\vec{O}, \vec{e_1}, \vec{e_2}, \vec{e_3})$.



(1.85)

Für einen Vektor $\vec{a}$ in allgemeiner Lage, etwa $\vec{a} = \vec{PQ}$ mit $P = (p_1, p_2, p_3)$ und $Q = (q_1, q_2, q_3)$ ergibt sich:

$$\begin{aligned} \vec{PQ} &= \vec{OQ} - \vec{OP} = q_1\vec{e_1} + q_2\vec{e_2} + q_3\vec{e_3} - p_1\vec{e_1} - p_2\vec{e_2} - p_3\vec{e_3} = \\ &= \begin{pmatrix} q_1 - p_1 \\ q_2 - p_2 \\ q_3 - p_3 \end{pmatrix}. \end{aligned} \tag{1.86}$$

Die Summe von Vektoren und die skalaren Vielfachen lassen sich aus der Koordinatendarstellung bezüglich $(O, \vec{e_1}, \vec{e_2}, \vec{e_3})$ sehr einfach berechnen:

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{pmatrix}, \quad \alpha \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \alpha a_1 \\ \alpha a_2 \\ \alpha a_3 \end{pmatrix}. \tag{1.87}$$

Ferner gilt mit dem Satz von Pythagoras:

$$|\vec{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}, \quad \text{falls} \quad \vec{a} = a_1\vec{e_1} + a_2\vec{e_2} + a_3\vec{e_3}. \tag{1.88}$$

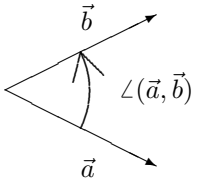
Beispiel(e) 1.9

$$\text{Für } \vec{a} = \begin{pmatrix} 2 \\ 0 \\ -3 \end{pmatrix}, \ \vec{b} = \begin{pmatrix} 0.5 \\ -1 \\ 6 \end{pmatrix} \text{ gilt: } \vec{a} + \vec{b} = \begin{pmatrix} 2.5 \\ -1 \\ 3 \end{pmatrix},$$

$$5\vec{a} = \begin{pmatrix} 10 \\ 0 \\ -15 \end{pmatrix}, 2\vec{a} - 3\vec{b} = \begin{pmatrix} 2.5 \\ 3 \\ -24 \end{pmatrix} \text{ und}$$

$$|\vec{a}| = \sqrt{13}, \ |\vec{b}| = \sqrt{37.25}, \ |\vec{a} + \vec{b}| = \sqrt{16.25}.$$

Trägt man zwei von $\vec{O}$ verschiedene Vektoren $\vec{a}, \vec{b}$ von einem Punkt P aus ab, so bezeichnet man den kleineren der beiden positiv gemessenen Winkel, den die Pfeile $\vec{a}$ und $\vec{b}$ im Scheitel P bilden, als Winkel zwischen $\vec{a}$ und $\vec{b}$ (in Zeichen: $\angle(\vec{a}, \vec{b})$), wobei $0 \leq \angle(\vec{a}, \vec{b}) \leq \pi$.



(1.89)

Offensichtlich gilt:

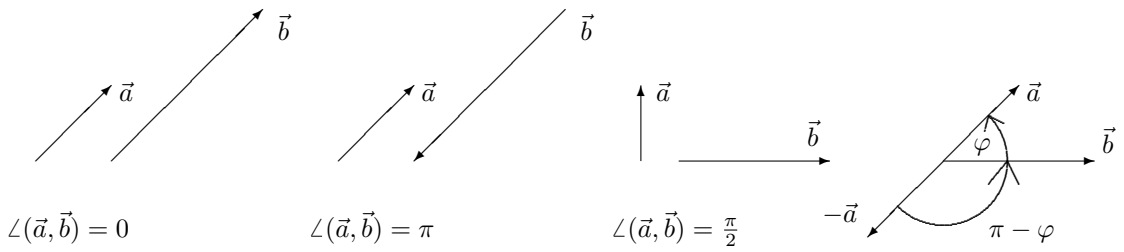
$$\angle(\vec{a}, \vec{b}) = \angle(\vec{b}, \vec{a}) \tag{1.90}$$

$$\angle(\vec{a}, t\vec{a}) = 0, \quad \text{falls } t > 0 \tag{1.91}$$

$$\angle(\vec{a}, t\vec{a}) = \pi, \quad \text{falls } t < 0 \tag{1.92}$$

$$\angle(-\vec{a}, \vec{b}) = \pi - \angle(\vec{a}, \vec{b}) \tag{1.93}$$

Der Vektor $\vec{a}$ heißt orthogonal (senkrecht) zu $\vec{b}$ (in Zeichen: $\vec{a} \perp \vec{b}$), wenn $\angle(\vec{a}, \vec{b}) = \frac{\pi}{2}$. Obwohl zwischen dem Nullvektor $\vec{O}$ und $\vec{a}$ kein Winkel erklärt wird, sagt man dennoch zur Vermeidung von Fallunterscheidungen, dass $\vec{O}$ orthogonal zu jedem beliebigen Vektor $\vec{a}$ ist, also $\vec{O} \perp \vec{a}$.



$\angle(\vec{a},\vec{b}) = 0$ $\quad$ $\angle(\vec{a},\vec{b}) = \pi$ $\quad$ $\angle(\vec{a},\vec{b}) = \frac{\pi}{2}$

Das Skalarprodukt $\vec{a} \cdot \vec{b}$ der Vektoren $\vec{a}$ und $\vec{b}$ ist definiert durch

$$\vec{a} \cdot \vec{b} := \begin{cases} |\vec{a}| \cdot |\vec{b}| \cos\Big(\angle(\vec{a},\vec{b})\Big), & \text{falls} \quad \vec{a} \neq \vec{O} \quad \text{und} \quad \vec{b} \neq \vec{O} \\ 0 & \text{sonst.} \end{cases} \tag{1.94}$$

Das Skalarprodukt wird auch als inneres Produkt bezeichnet und häufig in der Form $\langle \vec{a}, \vec{b} \rangle$ geschrieben.

Rechenregeln für das Skalarprodukt:

$$\begin{array}{rcll} \vec{a} \cdot \vec{b} & = & \vec{b} \cdot \vec{a} & \text{(Kommutativgesetz)} \\ (\alpha\vec{a}) \cdot \vec{b} & = & \vec{a} \cdot (\alpha\vec{b}) & \\ & = & \alpha\,(\vec{a} \cdot \vec{b}) & \text{für alle } \alpha \in \mathbb{R} \\ (\vec{a} + \vec{b}) \cdot \vec{c} & = & \vec{a} \cdot \vec{c} + \vec{b} \cdot \vec{c} & \text{(Distributivgesetz)} \\ \vec{a} \cdot \vec{b} = 0 & \Longleftrightarrow & \vec{a} \perp \vec{b} & \\ \sqrt{\vec{a} \cdot \vec{a}} & = & |\vec{a}|. & \end{array} \tag{1.95}$$

Vektoren $\vec{a}, \vec{b}$ bezüglich einer kartesischen Basis $(\vec{e_1}, \vec{e_2}, \vec{e_3})$ ermöglichen eine einfache Berechnung von $\vec{a} \cdot \vec{b}$ und $\cos\Big(\angle(\vec{a},\vec{b})\Big)$:

Sei $\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$ und $\vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$, so gilt:

$$\begin{aligned} \vec{a} \cdot \vec{b} &= (a_1 \cdot \vec{e_1} + a_2 \cdot \vec{e_2} + a_3 \cdot \vec{e_3}) \cdot (b_1 \cdot \vec{e_1} + b_2 \cdot \vec{e_2} + b_3 \cdot \vec{e_3}) \\ &= a_1 b_1 (\vec{e_1} \cdot \vec{e_1}) + (a_1 b_2 + a_2 b_1)(\vec{e_1} \cdot \vec{e_2}) + \\ &\quad + (a_1 b_3 + a_3 b_1)(\vec{e_1} \cdot \vec{e_3}) + (a_2 b_3 + a_3 b_2)(\vec{e_2} \cdot \vec{e_3}) + \\ &\quad + a_2 b_2 (\vec{e_2} \cdot \vec{e_2}) + a_3 b_3 (\vec{e_3} \cdot \vec{e_3}) \\ &= a_1 b_1 + a_2 b_2 + a_3 b_3. \end{aligned} \tag{1.96}$$

$$\cos\Big(\angle(\vec{a} \cdot \vec{b})\Big) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|} \tag{1.97}$$

$$= \frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{a_1^2 + a_2^2 + a_3^2} \cdot \sqrt{b_1^2 + b_2^2 + b_3^2}}. \tag{1.98}$$

Das Vektorprodukt $\vec{a} \times \vec{b}$ zweier Vektoren $\vec{a}$ und $\vec{b}$ ist der Vektor mit folgenden Eigenschaften:

(i) $\vec{a} \times \vec{b} = \vec{O}$, falls $\vec{a} = \vec{O}$, $\vec{b} = \vec{O}$ oder $\vec{a}$ parallel zu $\vec{b}$.

(ii) In allen anderen Fällen ist $\vec{a} \times \vec{b}$ derjenige Vektor, der

(1) senkrecht auf $\vec{a}$ und $\vec{b}$ steht

(2) mit dem $(\vec{a},\ \vec{b},\ \vec{a} \times \vec{b})$ ein Rechtssystem darstellt und

(3) dessen Betrag gleich dem Flächeninhalt F des von $\vec{a},\ \vec{b}$ aufgespannten Parallelogramms ist.

Für den Betrag des Vektors $\vec{a} \times \vec{b}$ gilt somit

$$|\vec{a} \times \vec{b}| = |\vec{a}| \cdot |\vec{b}| \cdot \sin\Big(\measuredangle(\vec{a}, \vec{b})\Big). \tag{1.99}$$

Rechenregeln für das Vektorprodukt:

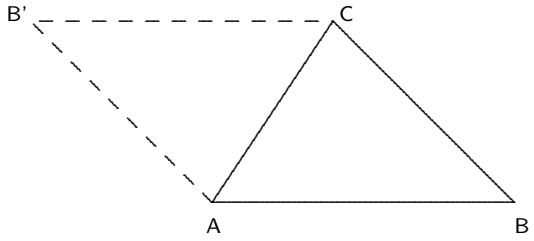
$$\begin{array}{rcl}
\vec{a} \times \vec{a} & = & \vec{O} \\
\vec{a} \times \vec{b} & = & -(\vec{b} \times \vec{a}) \\
\alpha(\vec{a} \times \vec{b}) & = & (\alpha\vec{a}) \times \vec{b} \ = \ \vec{a} \times (\alpha\vec{b}) \quad \text{für alle } \alpha \in \mathbb{R} \\
\vec{a} \times (\vec{b} + \vec{c}) & = & (\vec{a} \times \vec{b}) + (\vec{a} \times \vec{c}) \\
\vec{a} \times \vec{b} = \vec{O} & \Longleftrightarrow & \vec{a} = \vec{O} \quad \text{oder} \quad \vec{b} = \vec{O} \quad \text{oder} \quad \vec{a} \text{ parallel zu } \vec{b} \\
|\vec{a} \times \vec{b}|^2 & = & |\vec{a}|^2 \cdot |\vec{b}|^2 - (\langle \vec{a}, \vec{b} \rangle)^2.
\end{array} \tag{1.100}$$

Sei nun $(\vec{e_1},\ \vec{e_2},\ \vec{e_3})$ eine kartesische Basis und $\vec{a} = a_1\vec{e_1} + a_2\vec{e_2} + a_3\vec{e_3}$, $\vec{b} = b_1\vec{e_1} + b_2\vec{e_2} + b_3\vec{e_3}$, so gilt:

$$\begin{array}{rcl}
\vec{a} \times \vec{b} & = & (a_1 \cdot \vec{e_1} + \ a_2 \cdot \vec{e_2} + \ a_3 \cdot \vec{e_3}) \times (b_1 \cdot \vec{e_1} + \ b_2 \cdot \vec{e_2} + \ b_3 \cdot \vec{e_3}) = \\
& = & a_1b_1 \underbrace{(\vec{e_1} \times \vec{e_1})}_{\vec{O}} + a_1b_2 \underbrace{(\vec{e_1} \times \vec{e_2})}_{\vec{e_3}} + a_1b_3 \underbrace{(\vec{e_1} \times \vec{e_3})}_{-\vec{e_2}} + \\
& & + a_2b_1 \underbrace{(\vec{e_2} \times \vec{e_1})}_{-\vec{e_3}} + a_2b_2 \underbrace{(\vec{e_2} \times \vec{e_2})}_{\vec{O}} + a_2b_3 \underbrace{(\vec{e_2} \times \vec{e_3})}_{\vec{e_1}} + \\
& & + a_3b_1 \underbrace{(\vec{e_3} \times \vec{e_1})}_{\vec{e_2}} + a_3b_2 \underbrace{(\vec{e_3} \times \vec{e_2})}_{-\vec{e_1}} + a_3b_3 \underbrace{(\vec{e_3} \times \vec{e_3})}_{\vec{O}} = \\
& = & \begin{pmatrix} a_2b_3 - a_3b_2 \\ a_3b_1 - a_1b_3 \\ a_1b_2 - a_2b_1 \end{pmatrix}.
\end{array} \tag{1.101}$$

Beispiel(e) 1.10
Der Flächeninhalt eines Dreiecks ABC ist

$$F = \frac{1}{2}\left|\vec{AB} \times \vec{AC}\right|, \tag{1.102}$$



(1.103)

denn die Fläche des Parallelogramms $ABCB'$ ist gerade $\left|\vec{AB} \times \vec{AC}\right|$.

1.7 Komplexe Zahlen

In der mit einem kartesischen (x, y)-Koordinatensystem versehenen Ebene stellen die Punkte der x-Achse die reellen Zahlen dar. Wir gehen dazu über, auch alle anderen Punkte der Ebene als „Zahlen“ aufzufassen. Dazu schreiben wir den Punkt $z = (x,\, y)$ in der Form

$$z = x + iy \tag{1.104}$$

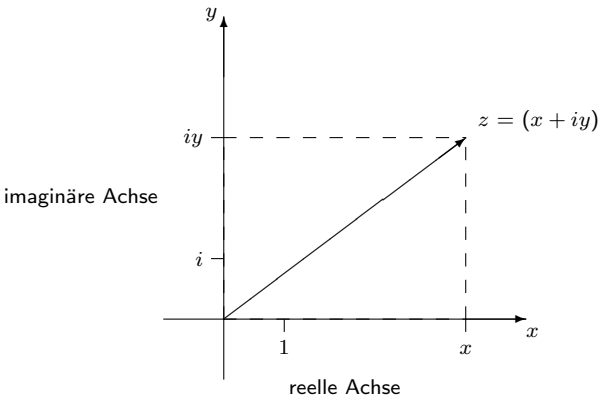
und nennen ihn eine komplexe Zahl mit Realteil

$$\mathrm{Re}(z) := x \tag{1.105}$$

und Imaginärteil

$$\mathrm{Im}(z) := y. \tag{1.106}$$

Die x-Achse heißt reelle Achse, die y-Achse wird imaginäre Achse genannt.



(1.107)

Abkürzend schreibt man

$$x = x + i0 = (x, 0) \tag{1.108}$$

(das sind die reellen Zahlen auf der reellen Achse) und

$$iy = 0 + iy = (0, y) \tag{1.109}$$

(das sind die rein imaginären Zahlen). Die Menge aller komplexen Zahlen wird mit $\mathbb{C}$ bezeichnet:

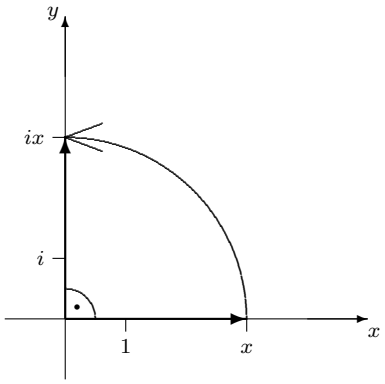
$$\mathbb{C} := \{x + iy; x, y \in \mathbb{R}\}\,. \tag{1.110}$$

Zwei komplexe Zahlen $z = x + iy$ und $w = u + iv$ sind genau dann gleich, wenn $x = u$ und $y = v$ gilt. Es ist üblich, den vom Ursprung O nach z weisenden Zeiger (Ortsvektor) ebenfalls mit z zu bezeichnen. Die Ebene, deren Punkte als komplexe Zahlen aufgefasst werden, heißt komplexe Zahlenebene.
Die Summe und Differenz komplexer Zahlen ist durch

$$(x + iy) + (u + iv) \quad := \quad (x + u) + i(y + v) \tag{1.111}$$

$$(x + iy) - (u + iv) \quad := \quad (x - u) + i(y - v) \tag{1.112}$$

definiert. Sie entspricht der Vektor-Summe bzw. -Differenz der zugehörigen Ortsvektoren. Die Zahl $ix \in \mathbb{C}$ entsteht aus $x \in \mathbb{R}$ durch eine positive Drehung des entsprechenden Ortsvektors um den Winkel $\frac{\pi}{2}$.



(1.113)

Die Multiplikation einer komplexen Zahl z mit der rein imaginären Zahl $i \in \mathbb{C}$ wird nun so erklärt, dass generell der Ortsvektor $i \cdot z$ aus dem Ortsvektor z durch eine positive Drehung um $\frac{\pi}{2}$ hervorgeht.
Somit erhält man

$$i^2 = i \cdot i = -1. \tag{1.114}$$

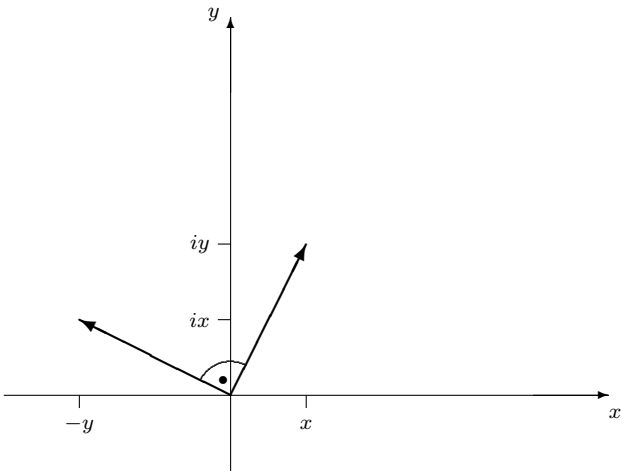
Zusammenfassend wird das Produkt komplexer Zahlen folgendermaßen definiert:

$$(x + iy) \cdot (u + iv) := (xu - yv) + i(xv + yu). \tag{1.115}$$

Diese Definition entspricht dem üblichen „ausmultiplizieren“:

$$(x + iy) \cdot (u + iv) = xu + iyu + ixv - yv = (xu - yv) + i(xv + yu). \tag{1.116}$$

Wegen $i(x+iy) = -y+ix$ stellt $z \mapsto iz$ die gewünschte positive Drehung um den Nullpunkt um $\frac{\pi}{2}$ dar.



(1.117)

Die Potenzen z^n sind wie üblich durch $z^0 := 1$ und $z^{n+1} := z \cdot z^n$ erklärt. Auch in $\mathbb{C}$ gilt die binomische Formel

$$(z + w)^n = \sum_{k=0}^{n} \binom{n}{k} z^k w^{n-k} \qquad (n \in \mathbb{N}_0). \tag{1.118}$$

Beispiel(e) 1.11

$$\begin{aligned}(3 + 0.5i)(2 - 4i) + (-1 + i)^3 &= 6 + i - 12i + 2 + i^3 - 3i^2 + 3i - 1 = \\ &= 10 - 9i.\end{aligned}$$

Die Division ist in $\mathbb{C}$ folgendermaßen erklärt: Man schreibt für $w \neq 0$:

$$q = \frac{z}{w}, \quad \text{falls} \quad w \cdot q = z. \tag{1.119}$$

Für $z = x + iy$ und $w = u + iv \neq 0$ (also $u \neq 0$ oder $v \neq 0$) gilt:

$$\frac{x+iy}{u+iv} = \frac{(x+iy)(u-iv)}{(u+iv)(u-iv)} = \frac{xu+yv}{u^2+v^2} + i\frac{yu-xv}{u^2+v^2}. \tag{1.120}$$

Durch $z^{-n} := \frac{1}{z^n}$ sind für $z \neq 0$ und $n \in \mathbb{N}$ die Potenzen mit negativen Exponenten erklärt. Es ist einfach zu zeigen, dass in $\mathbb{C}$ die üblichen, aus $\mathbb{R}$ bekannten Rechengesetze gelten, etwa

$$\begin{aligned} zw &= wz \\ z(ws) &= (zw)s \\ z(w+s) &= zw + zs \\ z^{m+n} &= z^m \cdot z^n. \end{aligned} \tag{1.121}$$

Auch in $\mathbb{C}$ ist eine Division durch 0 nicht möglich.
Man nennt $\overline{z} = x - iy$ die zu $z = x + iy$ konjugiert komplexe Zahl.

Beispiel(e) 1.12

$$\begin{aligned} \overline{i} &= -i \\ \overline{2} &= 2 \\ \overline{4+2i} &= 4-2i \\ \overline{3-2i} &= 3+2i. \end{aligned}$$

Für den Übergang $z \mapsto \overline{z}$ zur konjugiert komplexen Zahl (das entspricht der Spiegelung an der reellen Achse) gelten die folgenden Rechenregeln:

$$\begin{aligned} \overline{z+w} &= \overline{z} + \overline{w} \\ \overline{z \cdot w} &= \overline{z} \cdot \overline{w} \\ \overline{\left(\frac{z}{w}\right)} &= \frac{\overline{z}}{\overline{w}} \quad (w \neq 0) \\ \overline{(\overline{z})} &= z \\ \mathrm{Re}(z) &= \frac{1}{2}(z + \overline{z}) \\ \mathrm{Im}(z) &= \frac{1}{2i}(z - \overline{z}). \end{aligned} \tag{1.122}$$

Die Länge des Zeigers $z = x + iy$ wird mit $|z|$ bezeichnet und heißt Betrag der komplexen Zahl. Es gilt:

$$\begin{aligned} |z| &= \sqrt{x^2+y^2} = \sqrt{z \cdot \overline{z}}, \quad \text{falls} \quad z = x + iy \\ |z \cdot w| &= |z| \cdot |w| \\ |\overline{z}| &= |z| \\ \left|\frac{z}{w}\right| &= \frac{|z|}{|w|} \quad w \neq 0. \end{aligned} \tag{1.123}$$

Ferner gilt die Dreiecksungleichung

$$|z+w| \leq |z| + |w|, \tag{1.124}$$

die sich leicht mit vollständiger Induktion auf n Summanden verallgemeinern läßt:

$$\left| \sum_{k=1}^{n} z_k \right| \leq \sum_{k=1}^{n} |z_k|. \tag{1.125}$$

Die Gleichung

$$x^2 + 1 = 0 \tag{1.126}$$

hat in $\mathbb{R}$ offenbar keine Lösung, wohl aber in $\mathbb{C}$, nämlich $x_{1,2} = \pm i$ und es gilt:

$$x^2 + 1 = (x+i)(x-i). \tag{1.127}$$

Eine quadratische Gleichung

$$ax^2 + bx + c = 0 \qquad (\text{mit } a,\, b,\, c \in \mathbb{R},\, a \neq 0) \tag{1.128}$$

besitzt in $\mathbb{C}$ die beiden Lösungen

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{1.129}$$

Für $d < 0$ gilt: $\sqrt{d} = i\sqrt{-d}$ ($-d$ ist dann größer 0) und somit:

$$ax^2 + bx + c = a(x - x_1)(x - x_2). \tag{1.130}$$

Für die quadratische Gleichung

$$x^2 + \sqrt{2}x + 1 = 0 \tag{1.131}$$

ergeben sich zum Beispiel die Lösungen

$$x_1 = \frac{\sqrt{2}}{2}(-1+i) \quad \text{und} \quad x_2 = \frac{\sqrt{2}}{2}(-1-i). \tag{1.132}$$

Ist $z_1 \in \mathbb{C} \setminus \mathbb{R}$ die Lösung einer quadratischen Gleichung, so ist nach der obigen Lösungsformel auch $\overline{z_1}$ eine Lösung dieser Gleichung.

Kapitel 2

Lineare Algebra

2.1 Gleichungssysteme und Matrizen

Zur mathematischen Behandlung vieler Probleme der Technik, etwa zur Netzwerkberechnung in der Elektrotechnik oder zur Berechnung von Fachwerken in der Statik, bedient man sich der Matrizenrechnung.

Definition 2.1 (Matrix)
Eine Matrix vom Typ $m \times n$ (oder eine $(m \times n)$-Matrix) ist ein rechteckiges Zahlenschema der Form

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & & \alpha_{2n} \\ \vdots & & & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix}. \tag{2.1}$$

Die Zahlen $\alpha_{ij} \in \mathbb{R}$ (oder $\mathbb{C}$) heißen Komponenten (oder Elemente) der Matrix A.
Man schreibt abkürzend

$$A = (\alpha_{ij})_{m \times n} \tag{2.2}$$

oder nur $A = (\alpha_{ij})$, wenn der Typ feststeht.
Die $(m \times 1)$-Matrizen heißen Spaltenmatrizen oder Spaltenvektoren und haben die Form

$$s = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}. \tag{2.3}$$

Die $(1 \times n)$-Matrizen heißen Zeilenmatrizen oder Zeilenvektoren und haben die Form

$$z = (\alpha_1, \ldots, \alpha_n). \tag{2.4}$$

Bei nur einer Spalte oder Zeile benötigt man keinen zusätzlichen Index. Die Matrix $A = (\alpha_{ij})_{m \times n}$ besteht aus m Zeilenvektoren (mit je n Komponenten)

$$z_i = (\alpha_{i1}, \ldots, \alpha_{in}) \quad (1 \leq i \leq m) \tag{2.5}$$

bzw. aus n Spaltenvektoren (mit je m Komponenten)

$$s_j = \begin{pmatrix} \alpha_{1j} \\ \vdots \\ \alpha_{mj} \end{pmatrix} \quad (1 \leq j \leq n). \tag{2.6}$$

Je nachdem, ob die Zeilen oder die Spalten von A hervorgehoben werden sollen, schreibt man

$$A = \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix} \quad \text{(Zeilendarstellung) bzw. } A = (s_1, \ldots, s_n) \quad \text{(Spaltendarstellung).} \tag{2.7}$$

Die (i,j)-Komponente α_{ij} von A gehört dem i-ten Zeilenvektor z_i und dem j-ten Spaltenvektor s_j an. Man sagt, α_{ij} steht im Schnittpunkt der i-ten Zeile mit der j-ten Spalte.
Zwei Matrizen $A = (a_{ij})$, $B = (b_{ij})$ heißen gleich, in Zeichen: $A = B$, wenn sie vom gleichen Typ $m \times n$ sind und wenn außerdem $\alpha_{ij} = \beta_{ij}$ gilt für alle i, j mit $1 \leq i \leq m$ und $1 \leq j \leq n$.

Beispiel(e) 2.2
Die Matrix $A = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 0 & 4 \end{pmatrix}$ ist vom Typ 2×3, sie hat Zeilenvektoren

$$\begin{aligned} z_1 &= (1,3,5) \\ z_2 &= (2,0,4) \end{aligned}$$

und Spaltenvektoren

$$s_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \quad s_3 = \begin{pmatrix} 5 \\ 4 \end{pmatrix}.$$

Es gilt: $A \neq \begin{pmatrix} 1 & 3 & 5 & 0 \\ 2 & 0 & 4 & 0 \end{pmatrix}$.

Die Menge aller $(m \times n)$-Matrizen mit Komponenten aus $\mathbb{R}$ bezeichnen wir mit $\mathbb{R}^{m \times n}$. Mit $\mathbb{R}^n$ bezeichnen wir üblicherweise die Menge aller Spaltenvektoren $\mathbb{R}^{n \times 1}$. In speziellen Fällen (insbesondere bei Abbildungsvorschriften) wird auch die Menge aller Zeilenvektoren $\mathbb{R}^{1 \times n}$ mit $\mathbb{R}^n$ bezeichnet; dies ist aber die Ausnahme und wird immer explizit kenntlich gemacht. Die Einführung verschiedener Symbole würde mehr zur Verwirrung beitragen, als Klarheit stiften.

Die folgende Definition zeigt, dass man mit Matrizen rechnen kann.

Definition 2.3 (Addition, Subtraktion und Skalar-Multiplikation)
Für $A = (\alpha_{ij})$ und $B = (\beta_{ij})$ aus $\mathbb{R}^{m\times n}$ und jede Zahl $\lambda \in \mathbb{R}$ ist $A + B$, $A - B$ und $\lambda \cdot A$ folgendermaßen definiert:

$$A+B = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{m1} & \dots & \alpha_{mn} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \dots & \beta_{1n} \\ \vdots & & \vdots \\ \beta_{m1} & \dots & \beta_{mn} \end{pmatrix} \tag{2.8}$$

$$:= \begin{pmatrix} \alpha_{11}+\beta_{11} & \dots & \alpha_{1n}+\beta_{1n} \\ \vdots & & \vdots \\ \alpha_{m1}+\beta_{m1} & \dots & \alpha_{mn}+\beta_{mn} \end{pmatrix}, \tag{2.9}$$

$$A-B = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{m1} & \dots & \alpha_{mn} \end{pmatrix} - \begin{pmatrix} \beta_{11} & \dots & \beta_{1n} \\ \vdots & & \vdots \\ \beta_{m1} & \dots & \beta_{mn} \end{pmatrix} \tag{2.10}$$

$$:= \begin{pmatrix} \alpha_{11}-\beta_{11} & \dots & \alpha_{1n}-\beta_{1n} \\ \vdots & & \vdots \\ \alpha_{m1}-\beta_{m1} & \dots & \alpha_{mn}-\beta_{mn} \end{pmatrix}, \tag{2.11}$$

$$\lambda \cdot A = \lambda \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{m1} & \dots & \alpha_{mn} \end{pmatrix} := \begin{pmatrix} \lambda\cdot\alpha_{11} & \dots & \lambda\cdot\alpha_{1n} \\ \vdots & & \vdots \\ \lambda\cdot\alpha_{m1} & \dots & \lambda\cdot\alpha_{mn} \end{pmatrix}. \tag{2.12}$$

Mit dieser Definition läßt sich jeder Spaltenvektor $a = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$ folgendermaßen darstellen:

$$a = \begin{pmatrix} \alpha_1 \\ \\ \vdots \\ \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \alpha_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \begin{pmatrix} 0 \\ \vdots \\ \\ 0 \\ \alpha_n \end{pmatrix} = \tag{2.13}$$

$$= \alpha_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \alpha_n \begin{pmatrix} 0 \\ \vdots \\ \\ 0 \\ 1 \end{pmatrix} \tag{2.14}$$

Somit läßt sich jeder Spaltenvektor $a \in \mathbb{R}^n$ eindeutig in der Form

$$a = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_n e_n \tag{2.15}$$

mit Vektoren

$$e_1 := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \; e_2 := \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \; \ldots, \; e_n := \begin{pmatrix} 0 \\ \vdots \\ \\ 0 \\ 1 \end{pmatrix} \tag{2.16}$$

darstellen.

Das System $(e_1, \ldots, e_n)$ der Vektoren $e_i \in \mathbb{R}^n$ heißt natürliche Basis des $\mathbb{R}^n$.
Die Matrix

$$O := \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{m \times n}, \tag{2.17}$$

deren sämtliche Komponenten Null sind, heißt Nullmatrix (vom Typ $m \times n$).
Die Nullmatrix $O \in \mathbb{R}^{n \times 1}$ (bzw. $O \in \mathbb{R}^{1 \times n}$) wird auch Nullvektor genannt.
Mit $-A := (-1) \cdot A$ gelten für Matrizen die folgenden Rechenregeln:

a) $A + B = B + A$	für alle	$A, B \in \mathbb{R}^{m \times n}$
b) $(A + B) + C = A + (B + C)$	für alle	$A, B, C \in \mathbb{R}^{m \times n}$
c) $A + O = A$	für alle	$A \in \mathbb{R}^{m \times n}$
d) $A + (-A) = O$	für alle	$A \in \mathbb{R}^{m \times n}$
e) $(\lambda\mu)A = \lambda(\mu A)$	für alle	$\lambda, \mu \in \mathbb{R}$, $A \in \mathbb{R}^{m \times n}$
f) $1 \cdot A = A$	für alle	$A \in \mathbb{R}^{m \times n}$
g) $(\lambda + \mu)A = \lambda A + \mu A$	für alle	$\lambda, \mu \in \mathbb{R}$, $A \in \mathbb{R}^{m \times n}$
h) $\lambda(A + B) = \lambda A + \lambda B$	für alle	$\lambda \in \mathbb{R}$, $A, B \in \mathbb{R}^{m \times n}$.

(2.18)

Im Folgenden betrachten wir lineare Gleichungssysteme und ihre Darstellung durch Matrizen. Ein lineares Gleichungssystem mit m linearen Gleichungen für n Unbekannte $x_1, \ldots, x_n$ hat die Form

$$\begin{array}{ccccccccc} \alpha_{11}x_1 & + & \alpha_{12}x_2 & + & \ldots & + & \alpha_{1n}x_n & = & \beta_1 \\ \alpha_{21}x_2 & + & \alpha_{22}x_2 & + & \ldots & + & \alpha_{2n}x_n & = & \beta_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ \alpha_{m1}x_2 & + & \alpha_{m2}x_2 & + & \ldots & + & \alpha_{mn}x_n & = & \beta_m \end{array} \tag{2.19}$$

mit den Koeffizienten $\alpha_{ij} \in \mathbb{R}$ und den Absolutgliedern $\beta_i \in \mathbb{R}$.
Kommt eine Unbekannte in einer Gleichung nicht vor, dann hat sie dort den Koeffizient 0.
Für das obige Gleichungssystem schreibt man

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \vdots & \vdots & & \vdots \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \tag{2.20}$$

oder kurz

$$Ax = b \tag{2.21}$$

mit der Koeffizientenmatrix $A = (\alpha_{ij}) \in \mathbb{R}^{m \times n}$, dem Spaltenvektor $x \in \mathbb{R}^n$ bestehend aus den unbekannten Komponenten $x_1, \ldots, x_n$ und dem Spaltenvektor $b \in \mathbb{R}^m$ bestehend aus den Absolutgliedern.
In der i-ten Zeile der Koeffizientenmatrix stehen die Koeffizienten der i-ten Gleichung, die j-te Spalte von A „gehört“ zur Unbekannten x_j (dort stehen die Koeffizienten von x_j) $\quad (1 \leq i \leq m,$ $1 \leq j \leq n)$.
Ein lineares Gleichungssystem heißt homogen, falls $b = 0$ (d.h. $\beta_1 = \beta_2 = \ldots = \beta_m = 0$), ansonsten inhomogen.
Ein Spaltenvektor $\xi \in \mathbb{R}^n$ mit den Komponenten $\xi_1, \ldots, \xi_n \in \mathbb{R}$ heißt eine Lösung des obigen linearen Gleichungssystems, falls für $x_i = \xi_i$, $i = 1, \ldots, n$, die m Gleichungen tatsächlich erfüllt sind. Nicht jedes lineare Gleichungssystem ist lösbar. Es können die drei folgenden Fälle auftreten:

A) Das lineare Gleichungssystem besitzt keine Lösung, z.B.:

$$\begin{array}{ccccc} 3x_1 & + & 2x_2 & = & 1 \\ 3x_1 & + & 2x_2 & = & 5 \end{array} \tag{2.22}$$

B) Das lineare Gleichungssystem besitzt genau eine Lösung, z.B.:

$$\begin{array}{rcrcl} 3x_1 & + & 2x_2 & = & 1 \\ 3x_1 & + & x_2 & = & 5 \end{array} \quad \text{(also: } x_1 = 3,\ x_2 = -4\text{)} \tag{2.23}$$

C) Das lineare Gleichungssystem besitzt unendlich viele Lösungen, z.B.:

$$\begin{array}{rcrcl} 3x_1 & + & 2x_2 & = & 1 \\ 6x_1 & + & 4x_2 & = & 2 \end{array} \tag{2.24}$$

Für jede Zahl $\lambda \in \mathbb{R}$ ist $x_1 = \frac{1}{3}(1 - 2\lambda)$, $x_2 = \lambda$ eine Lösung.

Zwei lineare Gleichungssysteme $Ax = b$ und $Bx = c$ (mit nicht notwendig gleicher Anzahl von Gleichungen, also $A \in \mathbb{R}^{m_1 \times n}$, $B \in \mathbb{R}^{m_2 \times n}$) heißen äquivalent, wenn sie dieselbe Lösungsmenge besitzen.
Ein nach Carl Friedrich Gauss $(1777 - 1855)$ benanntes Lösungsverfahren für lineare Gleichungssysteme basiert auf der Beobachtung, dass bei folgenden Umformungen ein lineares Gleichungssystem in ein dazu äquivalentes Gleichungssystem übergeht:

(1) Vertauschung zweier Gleichungen

(2) Multiplikation einer Gleichung mit einer Zahl $\alpha \neq 0$

(3) Addition (bzw. Subtraktion) des Vielfachen einer Gleichung zu (bzw. von) einer anderen Gleichung.

Diese Umformungen verändern zwar das lineare Gleichungssystem selbst, aber nicht die entsprechende Lösungsmenge; sie werden übersichtlicher, wenn sie an der sogenannten erweiterten Koeffizientenmatrix

$$(A|b) := \left(\begin{array}{ccc|c} \alpha_{11} & \cdots & \alpha_{1n} & \beta_1 \\ \vdots & & \vdots & \vdots \\ \alpha_{m1} & \cdots & \alpha_{mn} & \beta_n \end{array}\right) \tag{2.25}$$

ausgeführt werden. Man erweitert A um die von den anderen Koeffizienten durch einen senkrechten Strich getrennte Spalte der Absolutglieder. Die den Gleichungsumformungen $(1) - (3)$ entsprechenden Veränderungen der Matrix $(A|b)$ heißen elementare Zeilenumformungen, das sind:

(1) Vertauschen zweier Zeilen

(2) Multiplikation einer Zeile mit einer Zahl $\alpha \neq 0$

(3) Addition (bzw. Subtraktion) des α-fachen einer Zeile zu einer anderen.

Es gilt somit:

> Entsteht $(B|c)$ aus $(A|b)$ durch endlich viele elementare Zeilenumformungen, dann sind $Ax = b$ und $Bx = c$ äquivalent.

Das Gauß-Verfahren zur Lösung eines linearen Gleichungssystems $Ax = b$ besteht nun aus drei Teilen:

(a) der Vorwärtselimination an der erweiterten Matrix $(A|b)$

(b) einer Lösbarkeitsentscheidung (dieser Schritt entfällt bei homogenen linearen Gleichungssystemen, also bei $b = 0$)

(c) der Rückwärtssubstitution.

(a): Vorwärtselimination: Man bringt durch eventuelle Zeilenvertauschung eine Zahl ungleich Null an die erste Stelle der ersten Spalte und annuliert die darunter stehenden Zahlen durch Subtraktion eines passenden Vielfachen der neuen ersten Zeile von der zweiten, dritten usw., d.h. ist $\alpha_{11} \neq 0$ (eventuell nach einer Zeilenvertauschung), dann subtrahiert man das $\frac{\alpha_{21}}{\alpha_{11}}$-fache der ersten Zeile von der zweiten, das $\frac{\alpha_{31}}{\alpha_{11}}$-fache der ersten Zeile von der dritten usw.. Auf diese Weise entsteht aus $(A|b)$ eine Matrix der Form $(B|c)$ mit:

$$(B|c) = \left(\begin{array}{cccccc|c} \diamond & \star & \star & \star & \cdots & \star & \\ & & & & & & \\ & & & A_1 & & & c \\ & & & & & & \end{array}\right) \tag{2.26}$$

mit einer $(m-1) \times n$ Matrix A_1, deren erste Spalte aus lauter Nullen besteht. An der $\diamond$ -Stelle steht eine Zahl ungleich Null. Hat man zu Beginn keine Zeile gefunden, sodass durch Zeilenvertauschung an der ersten Stelle der ersten Spalte ein Zahl ungleich Null steht, so kommt die x_1-Variable in dem linearen Gleichungssystem nicht vor und muss auch nicht berücksichtigt werden. Im Fall $A_1 = 0$ (Nullmatrix) ist die Vorwärtselimination beendet. Andernfalls wiederholt man dasselbe Vorgehen an der ersten von Null verschiedenen Spalte von A_1 (die erste Zeile von B bleibt unverändert). Diese Vorgehensweise wird so lange wiederholt, bis man zu einer Matrix $(M|d)$ gelangt, die eine Zeilenstufenform besitzt:

$$(M|d) = \left(\begin{array}{ccccccccc|c} \diamond & \star & \star & \ldots\ldots & \star & \ldots & \star & | & \alpha_1 \\ 0 & 0 & \diamond & \star & \ldots & & \star & | & \vdots \\ \vdots & \vdots & 0 & \diamond & \star & \ldots\ldots & \star & | & \vdots \\ & & \vdots & 0 & \diamond & \star \ldots & \star & | & \alpha_r \\ & & & \vdots & 0 & 0 \ldots & 0 & | & \alpha_{r+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & | & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 \ldots & 0 & | & \alpha_m \end{array}\right). \tag{2.27}$$

Kennzeichen der Zeilenstufenform:

- In jeder Zeile stehen links von $\diamond$ nur Nullen
- Liest man von oben nach unten, so rückt $\diamond$ pro Zeile um mindestens eine Stelle nach rechts.

Das lineare Gleichungssystem $Mx = d$ ist äquivalent zu $Ax = b$.

Beispiel(e) 2.4

$$\begin{array}{rcrcrcl} 2x_1 & & & - & x_3 & = & 1 \\ 2x_1 & + & 4x_2 & - & x_3 & = & 1 \\ -x_1 & + & 8x_2 & + & 3x_3 & = & 2 \end{array} \tag{2.28}$$

$$\left(\begin{array}{ccc|c} 2 & 0 & -1 & 1 \\ 2 & 4 & -1 & 1 \\ -1 & 8 & 3 & 2 \end{array}\right) \xrightarrow[III. + \frac{1}{2} \cdot I.Zeile]{II. - I.Zeile} \left(\begin{array}{ccc|c} 2 & 0 & -1 & 1 \\ 0 & 4 & 0 & 0 \\ 0 & 8 & 2.5 & 2.5 \end{array}\right) \tag{2.29}$$

$$\xrightarrow{III. - 2 \cdot II.Zeile} \left(\begin{array}{ccc|c} 2 & 0 & -1 & 1 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 2.5 & 2.5 \end{array}\right) \tag{2.30}$$

(b): Die Lösbarkeitsentscheidung (entfällt bei homogenen Gleichungssystemen, da dort $x = 0$ immer eine Lösung ist):
Ist nach Schritt (a) mit dem Ergebnis

$$(M|d) = \left(\begin{array}{cccccccc|c} \diamond & \star & \star & \ldots\ldots & \star & \ldots & \star & | & \alpha_1 \\ 0 & 0 & \diamond & \star & \ldots & & \star & | & \vdots \\ \vdots & \vdots & 0 & \diamond & \star & \ldots\ldots & \star & | & \vdots \\ & & \vdots & 0 & \diamond & \star \ldots & \star & | & \alpha_r \\ & & & \vdots & 0 & 0 \ldots & 0 & | & \alpha_{r+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & | & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 \ldots & 0 & | & \alpha_m \end{array}\right) \tag{2.31}$$

eine der Zahlen $\alpha_{r+1}, \ldots, \alpha_m$ von Null verschieden, dann ist $Mx = d$ und damit $Ax = b$ nicht lösbar. Denn wäre zum Beispiel $\alpha_{r+1} \neq 0$, so würde die $(r+1)$-te Zeile des linearen Gleichungssystems $Mx = d$ lauten: $0 = \alpha_{r+1}$.
Für den Fall $\alpha_{r+1} = \ldots = \alpha_m = 0$ ist die Rückwärtssubstitution durchführbar.

(c): Rückwärtssubstitution: Die in (2.31) zu den Spalten ohne $\diamond$ - Stelle gehörenden Unbekannten sind die freien Variablen, die der Reihe nach gleich $\lambda_1, \ldots, \lambda_{n-r}$ gesetzt werden. Im (2.31) zugeordneten linearen Gleichungssystem bringt man die freien Variablen auf die rechte Seite, ersetzt sie durch den ihnen zugewiesenen „Wert“ λ_i und berechnet der Reihe nach von unten nach oben die zu $\diamond$ -Stellen gehörenden Variablen (in Abhängigkeit von $\lambda_1, \lambda_2, \ldots, \lambda_{n-r}$).

Beispiel(e) 2.5
Wir betrachten das lineare Gleichungssystem

$$\begin{array}{rcrcrcrcr} & & x_2 & - & x_3 & - & x_4 & = & 7 \\ x_1 & - & 4x_2 & + & 2x_3 & & & = & 0 \\ 2x_1 & - & 3x_2 & - & x_3 & - & 5x_4 & = & 35 \\ 3x_1 & - & 7x_2 & + & x_3 & - & 5x_4 & = & 35 \end{array} \tag{2.32}$$

zu (a):

$$(A|b) \quad = \quad \left(\begin{array}{cccc|c} 0 & 1 & -1 & -1 & 7 \\ 1 & -4 & 2 & 0 & 0 \\ 2 & -3 & -1 & -5 & 35 \\ 3 & -7 & 1 & -5 & 35 \end{array}\right) \tag{2.33}$$

$$\overset{\substack{I.\ und\ II.\\ Zeile\ vertauschen}}{\longrightarrow} \quad \left(\begin{array}{cccc|c} 1 & -4 & 2 & 0 & 0 \\ 0 & 1 & -1 & -1 & 7 \\ 2 & -3 & -1 & -5 & 35 \\ 3 & -7 & 1 & -5 & 35 \end{array}\right) \tag{2.34}$$

$$\overset{\substack{III.\ -\ 2\cdot I.\ Zeile\\ IV.\ -\ 3\cdot I.\ Zeile}}{\longrightarrow} \quad \left(\begin{array}{cccc|c} 1 & -4 & 2 & 0 & 0 \\ 0 & 1 & -1 & -1 & 7 \\ 0 & 5 & -5 & -5 & 35 \\ 0 & 5 & -5 & -5 & 35 \end{array}\right) \tag{2.35}$$

$$\overset{\substack{III.\ -\ 5\cdot II.\ Zeile\\ IV.\ -\ 5\cdot II.\ Zeile}}{\longrightarrow} \quad \left(\begin{array}{cccc|c} 1 & -4 & 2 & 0 & 0 \\ 0 & 1 & -1 & -1 & 7 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{2.36}$$

$$= \quad (M|d) \tag{2.37}$$

zu (b): $\alpha_3 = \alpha_4 = 0 \quad \Rightarrow$ lösbar.
zu (c): Die Unbekannten x_3 und x_4 sind frei wählbar, also $x_3 = \lambda_1$, $x_4 = \lambda_2$. Einsetzen in die 2. Zeile und Auflösen nach x_2 ergibt: $x_2 = 7 + \lambda_1 + \lambda_2$. Einsetzen in die erste Zeile und Auflösen nach x_1 ergibt:
$x_1 = 4(7 + \lambda_1 + \lambda_2) - 2 \cdot \lambda_1 = 28 + 2\lambda_1 + 4\lambda_2$.
Die allgemeine Lösung lautet:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 28 \\ 7 \\ 0 \\ 0 \end{pmatrix} + \lambda_1 \begin{pmatrix} 2 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 4 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \qquad \lambda_1,\, \lambda_2 \in \mathbb{R}. \tag{2.38}$$

2.2 Matrizenmultiplikation

Im Folgenden legen wir das Produkt ab eines Zeilenvektors $a \in \mathbb{R}^n$ und eines Spaltenvektors $b \in \mathbb{R}^n$ mit $a = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ und $b = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$ fest:

$$ab := \alpha_1\beta_1 + \alpha_2\beta_2 + \ldots + \alpha_n\beta_n = \sum_{i=1}^{n} \alpha_i\beta_i. \tag{2.39}$$

Somit ergibt das Produkt „Zeile mal Spalte“ eine Zahl.
Für Zeilenvektoren $a, a_1, a_2 \in \mathbb{R}^n$ und Spaltenvektoren $b, b_1, b_2 \in \mathbb{R}^n$ gelten die folgenden Rechenregeln:

$$\begin{aligned} (a_1 + a_2)b &= a_1b + a_2b \\ a(b_1 + b_2) &= ab_1 + ab_2 \\ \alpha(ab) &= (\alpha a)b = a(\alpha b) \qquad \text{für alle } \alpha \in \mathbb{R}. \end{aligned} \tag{2.40}$$

Basierend auf dieser Festlegung definieren wir nun die Multiplikation zweier Matrizen.

Definition 2.6 (Matrixmultiplikation)
Seien $A = (\alpha_{ij})_{m\times n}$ und $B = (\beta_{ij})_{n\times r}$ zwei Matrizen mit der Zeilendarstellung

$$A = \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix}, \qquad z_i \in \mathbb{R}^n,\ i = 1, \ldots, m, \tag{2.41}$$

für A und der Spaltendarstellung

$$B = (s_1, \ldots, s_r), \qquad s_i \in \mathbb{R}^n,\ i = 1, \ldots, r, \tag{2.42}$$

für B, so ist das Produkt AB definiert durch

$$AB := \begin{pmatrix} z_1s_1 & \ldots & z_1s_r \\ z_2s_1 & \ldots & z_2s_r \\ \vdots & & \vdots \\ z_ms_1 & \ldots & z_ms_r \end{pmatrix}. \tag{2.43}$$

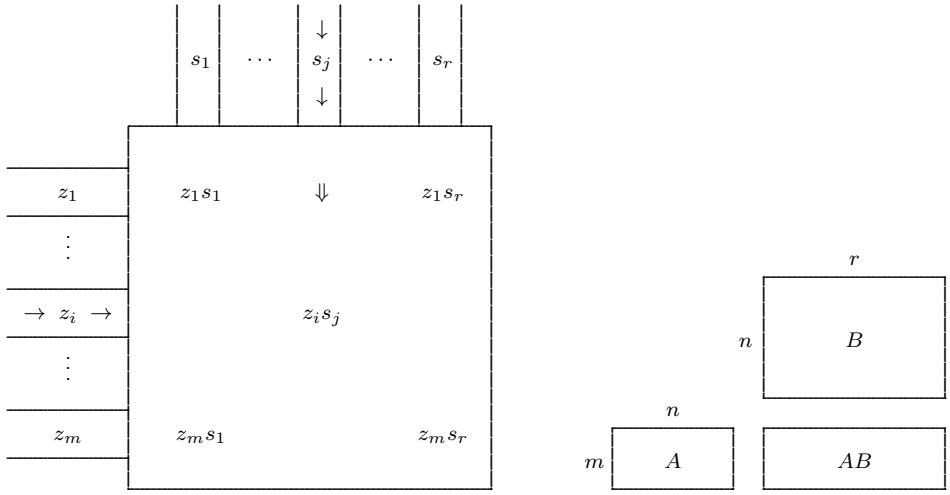
Das Matrixprodukt AB ist nur erklärt für $A \in \mathbb{R}^{m\times n}$ und $B \in \mathbb{R}^{n\times r}$, d.h. die Spaltenzahl von A muss gleich der Zeilenzahl von B sein.
Für $A = (\alpha_{ij})_{m\times n}$, $B = (\beta_{ij})_{n\times r}$ und $AB = (\gamma_{ij})_{m\times r}$ ergibt sich

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik} \cdot \beta_{kj}. \tag{2.44}$$

Beispiel(e) 2.7

$$\begin{pmatrix} 4 & 3 & 0 & 1 & 2 \\ 2 & 1 & 4 & 0 & 1 \\ 0 & 0 & 4 & 1 & 0 \\ 2 & 0 & 1 & 0 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 4 & 2 \\ 3 & 0 & 1 \\ 1 & 1 & 3 \\ 0 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 5 & 21 & 17 \\ 14 & 8 & 10 \\ 13 & 1 & 7 \\ 5 & 4 & 17 \end{pmatrix}. \tag{2.45}$$

Die Matrixmultiplikation läßt sich in folgendem Schema zusammenfassen:



(2.46)

Für das Produkt Ax von $A = (\alpha_{ij})_{m\times n}$ und $x \in \mathbb{R}^n$ erhalten wir

$$Ax = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{m1} & \dots & \alpha_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \alpha_{11}x_1 & + & \alpha_{12}x_2 & + & \dots & + & \alpha_{1n}x_n \\ \vdots & & \vdots & & & & \vdots \\ \alpha_{m1}x_1 & + & \alpha_{m2}x_2 & + & \dots & + & \alpha_{mn}x_n \end{pmatrix}. \quad (2.47)$$

In der bisher nur als Abkürzung verstandenen Schreibweise $Ax = b$ für lineare Gleichungssysteme ergibt sich nun durch die obige Gleichung für Ax die richtige Deutung. Unter Verwendung der Matrix

$$E_n := \begin{pmatrix} 1 & 0 & \dots\dots\dots & 0 \\ 0 & 1 & \ddots & 0 \\ \vdots & 0 & \ddots & \vdots \\ & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}, \quad (2.48)$$

die als $n \times n$-Einheitsmatrix bezeichnet wird, erhalten wir für alle Matrizen A, A_1, $A_2 \in \mathbb{R}^{m\times n}$, B, B_1, $B_2 \in \mathbb{R}^{n\times r}$ und $C \in \mathbb{R}^{r\times s}$:

a) $(A_1 + A_2) \cdot B = A_1B + A_2B$

b) $A(B_1 + B_2) = AB_1 + AB_2$

c) $\alpha(AB) = (\alpha A)B = A(\alpha B)$ für alle $\alpha \in \mathbb{R}$ (2.49)

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

e) $E_mA = AE_n = A$.

Falls $r = m$, gilt im allgemeinen:

$$AB \neq BA. \quad (2.50)$$

Wegen (d) kann man bei Matrixprodukten bestehend aus mehreren Faktoren auf Klammerung verzichten: z.B. $(AB)(CD) = A(BC)D = ABCD$.
Für (quadratische) $(n \times n)$-Matrizen lassen sich Potenzen A^k, $k \in \mathbb{N}_0$, induktiv definieren:

$$A^0 := E_n,\ A^{k+1} = A^k A,\ \text{d.h. } A^k = \underbrace{A \cdot A \cdot \ldots \cdot A}_{k-mal}. \tag{2.51}$$

Offensichtlich gilt: $A^k \cdot A^l = A^{k+l}$, $(A^k)^l = A^{k \cdot l}$, $\quad k,\ l \in \mathbb{N}_0$.

Jeder $(m \times n)$-Matrix $A = (\alpha_{ij})$ kann man eine $(n \times m)$-Matrix $A^\top$ (die transponierte Matrix von A) zuordnen, deren i-te Zeile aus den Elementen der i-ten Spalte von A besteht.

Beispiel(e) 2.8

$$\begin{aligned}
(1,2,3,4)^\top &= \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \\
\begin{pmatrix} 1 \\ 0 \\ \alpha \end{pmatrix}^\top &= (1,0,\alpha) \\
\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 0 & \alpha \\ \beta & x & 2 & 4 \end{pmatrix}^\top &= \begin{pmatrix} 1 & 0 & \beta \\ 2 & 1 & x \\ 3 & 0 & 2 \\ 4 & \alpha & 4 \end{pmatrix}.
\end{aligned}$$

Es gelten folgende Rechenregeln:

$$\begin{array}{llcll}
\text{(a)} & (A+B)^\top & = & A^\top + B^\top & \text{für alle } A, B \in \mathbb{R}^{m \times n} \\
\text{(b)} & (\alpha A)^\top & = & \alpha \cdot A^\top & \text{für alle } A \in \mathbb{R}^{m \times n},\ \alpha \in \mathbb{R} \\
\text{(c)} & (A^\top)^\top & = & A & \text{für alle } A \in \mathbb{R}^{m \times n} \\
\text{(d)} & (AB)^\top & = & B^\top A^\top & \text{für alle } A \in \mathbb{R}^{m \times n},\ B \in \mathbb{R}^{n \times r}.
\end{array} \tag{2.52}$$

Zwei wichtige Klassen von Matrizen werden durch die folgende Definition festgelegt:

Definition 2.9 ((schief-)symmetrische Matrizen)
Eine $(n \times n)$-Matrix A heißt symmetrisch, falls $A = A^\top$ gilt,
sie heißt schiefsymmetrisch, falls $A = -A^\top$.

Für eine symmetrische Matrix $A = (\alpha_{ij})_{n \times n}$ gilt $\alpha_{ij} = \alpha_{ji}$ für alle $i,\ j \in \{1,\ \ldots,\ n\}$.
Für eine schiefsymmetrische Matrix $A = (\alpha_{ij})_{n \times n}$ gilt

$$\alpha_{ij} = -\alpha_{ji} \quad \text{für alle} \quad i, j \in \{1, \ldots, n\}; \tag{2.53}$$

insbesondere gilt dann $\alpha_{ii} = 0$, $i = 1, \ldots, n$.
Eine wichtige Menge von Matrizen ist folgendermaßen definiert:

Definition 2.10 (invertierbare Matrizen)
Eine $(n \times n)$-Matrix A heißt invertierbar, falls es eine $(n \times n)$-Matrix B gibt, sodass $AB = BA = E_n$ gilt. In diesem Fall ist die Matrix B eindeutig bestimmt und wird im allgemeinen mit A^{-1} bezeichnet. A^{-1} heißt die Inverse oder inverse Matrix von A.

Die obige Definition ist formal nicht ganz korrekt, da nicht nur ein Name vergeben wird, sondern auch eine zu beweisende Behauptung (die Eindeutigkeit von A^{-1}) aufgestellt wird. Daher haben wir den folgenden Satz zu beweisen:

Satz 2.11 (Invertierbarkeit und Eindeutigkeit)
Wenn es zu einer Matrix $A \in \mathbb{R}^{n\times n}$ zwei Matrizen B, $C \in \mathbb{R}^{n\times n}$ gibt mit $BA = AC = E_n$, dann ist A invertierbar und $B = C = A^{-1}$.

Beweis:
$B = BE_n = B(AC) = (BA)C = E_nC = C$,
also ist $BA = AB = E_n$ und somit $B = A^{-1} = C$

q.e.d.

Beispiel(e) 2.12

- E_n ist invertierbar und es gilt $E_n^{-1} = E_n$
- Für $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{R}^{2\times 2}$ mit $ad - bc \neq 0$ und $B = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ gilt $AB = BA = E_2$, also sind A, B invertierbar, $A^{-1} = B$ und $B^{-1} = A$.

Mit dem folgenden Satz fassen wir Eigenschaften invertierbarer Matrizen zusammen.

Satz 2.13 (Eigenschaften invertierbarer Matrizen)

(i) Die Inverse einer invertierbaren $(n \times n)$-Matrix ist invertierbar und es gilt:

$$(A^{-1})^{-1} = A. \tag{2.54}$$

(ii) Das Produkt AB zweier invertierbarer $(n \times n)$-Matrizen ist invertierbar und es gilt: $(AB)^{-1} = B^{-1}A^{-1}$.

(iii) Die Transponierte $A^\top$ einer $(n\times n)$-Matrix ist genau dann invertierbar, wenn A invertierbar ist, und es gilt: $(A^\top)^{-1} = (A^{-1})^\top$.

Beweis:

(i) Da $AA^{-1} = A^{-1}A = E_n$, ist A die Inverse zu A^{-1}.

(ii) $AB \cdot B^{-1}A^{-1} = A(BB^{-1})A^{-1} = AE_nA^{-1} = AA^{-1} = E_n = B^{-1}A^{-1}AB$. Daher ist $B^{-1}A^{-1}$ die Inverse zu AB.

(iii) Da $E_n^\top = E_n$, folgt aus $AA^{-1} = E_n$: $E_n = E_n^\top = (AA^{-1})^\top = (A^{-1})^\top A^\top$.
Somit ist $(A^{-1})^\top$ die Inverse zu $A^\top$, also $(A^{-1})^\top = (A^\top)^{-1}$

q.e.d.

Für endlich viele invertierbare $n \times n$-Matrizen $A_1, \ldots, A_k$ folgt somit:

$$(A_1 A_2 \cdots A_k)^{-1} = A_k^{-1} \cdot \ldots \cdot A_1^{-1}. \tag{2.55}$$

In einer quadratischen $(n \times n)$-Matrix $A = (\alpha_{ij})$ nennt man die Zahlen α_{ii}, $1 \leq i \leq n$, Diagonalelemente. Unterhalb der Diagonalen stehen α_{ij} mit $i > j$, oberhalb die Zahlen α_{ij} mit $i < j$. Man nennt A eine untere bzw. obere Dreiecksmatrix, wenn sie höchstens auf und unterhalb (bzw. oberhalb) der Diagonalen von Null verschiedene Elemente hat.

$$A = \begin{pmatrix} \boxed{\alpha_{11}} & \alpha_{12} & \cdots & & \alpha_{1n} \\ \alpha_{21} & \boxed{\alpha_{22}} & \ldots & & \vdots \\ \vdots & & & & \\ \vdots & & & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & & \boxed{\alpha_{nn}} \end{pmatrix} \quad \text{die Diagonale von } A \tag{2.56}$$

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & & \alpha_{1n} \\ 0 & \alpha_{22} & \ldots & & \vdots \\ \vdots & & & & \\ \vdots & & & \ddots & \vdots \\ 0 & \ldots\ldots\ldots & & 0 & \alpha_{nn} \end{pmatrix} \quad \text{eine obere Dreiecksmatrix} \tag{2.57}$$

$$A = \begin{pmatrix} \alpha_{11} & 0 & \cdots & & 0 \\ \alpha_{21} & \alpha_{22} & \ldots & & \vdots \\ \vdots & & & & \vdots \\ \vdots & & & \ddots & 0 \\ \alpha_{n1} & \ldots\ldots\ldots\ldots & & & \alpha_{nn} \end{pmatrix} \quad \text{eine untere Dreiecksmatrix.} \tag{2.58}$$

Die Transponierte einer oberen Dreiecksmatrix ist eine untere Dreiecksmatrix und umgekehrt. Eine obere (bzw. untere) Dreiecksmatrix ist genau dann invertierbar, wenn alle Diagonalelemente von Null verschieden sind. Für eine $(n \times n)$-Matrix sind folgende Aussagen äquivalent (d.h., ist eine Aussage wahr, so gilt das auch für alle anderen):

a) A ist invertierbar

b) Es gibt eine $(n \times n)$-Matrix B mit $AB = E_n$

c) Es gibt eine $(n \times n)$-Matrix C mit $CA = E_n$

d) $Ax = 0 \Rightarrow x = 0$.

2.3 Vektorräume

In der Mathematik spielen Menge eine große Rolle, deren Elemente man addieren und mit einem Zahlenfaktor multiplizieren kann. Mit der Menge der Vektoren in der Ebene (insbesondere den komplexen Zahlen), der Menge der Vektoren im Raum und der Menge der $(m \times n)$-Matrizen haben wir bereits drei Beispiele kennengelernt. Ein weiteres Beispiel ist etwa die Menge aller auf einem Intervall I definierten Funktionen $f : I \mapsto \mathbb{R}$. Man beobachtet nun, dass für die entsprechenden Rechenoperationen (Addition von Elementen einer Menge und Multiplikation mit einem Zahlenfaktor) unabhängig von der Beschaffenheit der Elemente, die die zugrundegelegte Menge bilden, dieselben Rechengesetze gelten. Zur einheitlichen Herleitung der sich daraus ergebenden

Konsequenzen wird der Begriff des $\mathbb{R}$-Vektorraumes eingeführt.

Definition 2.14 ($\mathbb{R}$-Vektorraum)
Eine nichtleere Menge V, in der man zu je zwei Elementen a, $b \in V$ eine Summe $a + b \in V$, zu jedem $a \in V$ und zu jedem $\lambda \in \mathbb{R}$ das λ-fache $\lambda a \in V$ bilden kann, heißt ein $\mathbb{R}$-Vektorraum (oder Vektorraum über $\mathbb{R}$, bzw. linearer Raum über $\mathbb{R}$), wenn folgende acht Rechengesetze (die Vektorraum-Axiome) erfüllt sind:

(V.1) Die Addition ist kommutativ, d.h. für alle a, $b \in V$ gilt

$$a + b = b + a$$

(V.2) Die Addition ist assoziativ, d.h. für alle a, b, $c \in V$ gilt

$$(a + b) + c = a + (b + c)$$

(V.3) Es gibt ein Element $o \in V$, Nullelement genannt, mit

$$a + o = a \quad \text{für alle} \quad a \in V.$$

(V.4) Zu jedem $a \in V$ gibt es genau ein mit $-a$ bezeichnetes Element in V mit $a + (-a) = o$.

(V.5) $1 \cdot a = a$ für alle a (Zahlenfaktor $1 \in \mathbb{R}$).

(V.6) $\lambda(\mu a) = (\lambda\mu)a$ für alle λ, $\mu \in \mathbb{R}$, $a \in V$.

(V.7) $\lambda(a + b) = \lambda a + \lambda b$ für alle $\lambda \in \mathbb{R}$, a, $b \in V$.

(V.8) $(\lambda + \mu)a = \lambda a + \mu a$ für alle λ, $\mu \in \mathbb{R}$, $a \in V$.

Die Elemente eines Vektorraumes nennt man Vektoren; statt $a + (-b)$ schreibt man $a - b$ (Differenz).

Die Axiome (V.1)-(V.8) garantieren, dass man mit Summen, Differenzen und Vielfachen wie gewohnt rechnen darf. Wegen (V.2) kann man in endlichen Summen $a_1 + a_2 + \ldots + a_n$ auf Klammern verzichten und aus den Distributivgesetzen (V.7), (V.8) folgt für alle $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$, $a_1, \ldots, a_n \in V$:

$$(\lambda_1 + \lambda_2 + \ldots + \lambda_k)(a_1 + \ldots + a_n) = \lambda_1 a_1 + \ldots + \lambda_k a_n; \tag{2.59}$$

außerdem gelten die Vorzeichenregeln:

$$\begin{aligned} -a &= (-1)a \\ -(-a) &= a \\ -(a + b) &= (-a) + (-b) = -a - b \quad \text{usw.} \end{aligned} \tag{2.60}$$

Das Nullelement wird immer mit o bezeichnet; bei Verwechslungsgefahr mit o_V.
Auf das überraschende Axiom (V.5) werden wir später eingehen.

Beispiel(e) 2.15

- Ein Vergleich mit den Rechenregeln für $(m \times n)$-Matrizen zeigt, dass die Menge aller $(m \times n)$-Matrizen zusammen mit der komponentenweisen Addition und skalaren Multiplikation ein $\mathbb{R}$-Vektorraum ist. Insbesondere ist die Menge $\mathbb{R}^n$ für jedes $n \in \mathbb{N}$ ein $\mathbb{R}$-Vektorraum.
- Für jedes Intervall $I \subseteq \mathbb{R}$ ist die Menge aller Funktionen

$$f : I \to \mathbb{R} \tag{2.61}$$

durch

$$f + g : I \to \mathbb{R}, \quad x \mapsto f(x) + g(x) \tag{2.62}$$

und

$$\lambda \cdot f : I \to \mathbb{R}, \quad x \mapsto \lambda \cdot f(x), \quad \lambda \in \mathbb{R} \tag{2.63}$$

ein $\mathbb{R}$-Vektorraum.

Häufig erfüllen bereits Teilmengen von $\mathbb{R}$-Vektorräumen die Vektorraumaxiome.

Definition 2.16 (Unterraum, linearer Teilraum)
Eine nichtleere Teilmenge $U \subseteq V$ eines $\mathbb{R}$-Vektorraumes V heißt Unterraum oder linearer Teilraum von V, wenn gilt:

(U.1) $u, v \in U \Rightarrow u + v \in U$

(U.2) $u \in U, \lambda \in \mathbb{R} \Rightarrow \lambda \cdot u \in U$.

Ein Unterraum ist bezüglich der in der Obermenge V definierten Addition und skalaren Multiplikation selbst ein $\mathbb{R}$-Vektorraum, denn mit $a \in U$ liegt nach (U.2) auch $-a = (-1)a$ in U und nach (U.1) auch $a + (-a) = o$.
Alle anderen Vektorraumaxiome sind selbstverständlich auch für U erfüllt.

Beispiel(e) 2.17

- Jeder $\mathbb{R}$-Vektorraum V besitzt den „trivialen“ Unterraum $U = \{o\}$.
- Zu jedem Element v eines $\mathbb{R}$-Vektorraumes V ist

$$\mathbb{R}_v := \{\alpha v; \alpha \in \mathbb{R}\} \tag{2.64}$$

ein Unterraum von V. Für $V = \mathbb{R}^3$ besteht $\mathbb{R}_v$ aus allen zu v parallelen Vektoren (einschließlich dem Nullvektor).

- Im Gegensatz zu

$$V_1 := \left\{ \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} \in \mathbb{R}^4; 2\alpha_1 + 3\alpha_2 + \alpha_4 = 1 \right\} \tag{2.65}$$

ist

$$U := \left\{ \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} \in \mathbb{R}^4; 2\alpha_1 + 3\alpha_2 + 4\alpha_4 = 0 \right\} \tag{2.66}$$

ein Unterraum des $\mathbb{R}^4$.

- Für jede $(m \times n)$-Matrix $A \in \mathbb{R}^{m\times n}$ ist

$$\text{Kern}(A) := \{x \in \mathbb{R}^n;\ Ax = o\}, \quad o \in \mathbb{R}^m \tag{2.67}$$

ein Unterraum des $\mathbb{R}^n$, denn $o \in \text{Kern}(A)$ und für u, $v \in \text{Kern}(A)$, $\lambda \in \mathbb{R}$ gilt:

$$A(u + v) = Au + Av = o + o = o \tag{2.68}$$

und

$$A(\lambda u) = \lambda(Au) = \lambda \cdot o = o. \tag{2.69}$$

Zwei der zentralen Begriffe der Theorie abstrakter Vektorräume werden in der folgenden Definition festgelegt.

Definition 2.18 (Linearkombination, lineare Hülle)
Sei V ein $\mathbb{R}$-Vektorraum. Jede aus endlich vielen Vektoren $v_1, \ldots, v_k \in V$ gebildete Summe der Form

$$\sum_{i=1}^{k} \alpha_i v_i = a_1 v_1 + \ldots + a_k v_k \tag{2.70}$$

mit den Koeffizienten $\alpha_i \in \mathbb{R}$ heißt eine Linearkombination der v_i.
Eine solche Linearkombination wird trivial genannt, wenn sämtliche α_i gleich Null sind. Die Menge

$$\text{Lin}(\{v_1, \ldots, v_k\}) := \left\{ \sum_{i=1}^{k} \alpha_i v_i;\ \alpha_i \in \mathbb{R}, i = 1, \ldots, k \right\} \tag{2.71}$$

heißt lineare Hülle der v_i.

Ohne Mühe ist einzusehen, dass die lineare Hülle $\text{Lin}(\{v_1, \ldots, v_k\})$ ein Unterraum von V ist.

Die folgende Definition stellt einen Zusammenhang zwischen Unterräumen und linearen Hüllen her.

Definition 2.19 (Erzeugendensystem)
Ein Unterraum U eines $\mathbb{R}$-Vektorraumes V wird von den Vektoren $v_1, \ldots, v_k \in V$ erzeugt, oder $\{v_1, \ldots, v_k\}$ ist ein Erzeugendensystem von U, wenn

$$U = \mathsf{Lin}(\{v_1, \ldots, v_k\}). \tag{2.72}$$

Beispiel(e) 2.20
Die Vektoren $e_1^\top = (1,0,0,0)$ $e_2^\top = (0,1,0,0)$ $e_3^\top = (0,0,1,0)$ und $e_4^\top = (0,0,0,1)$ erzeugen den $\mathbb{R}^4$.

Wegen

$$v_i = 0 \cdot v_1 + \ldots + 0 \cdot v_{i-1} + 1 \cdot v_i + 0 \cdot v_{i+1} + \ldots + 0 \cdot v_k \tag{2.73}$$

(vgl. Axiom (V.5)) gilt

$$v_i \in \mathsf{Lin}(\{v_1, \ldots, v_k\}), \quad i = 1, \ldots, n. \tag{2.74}$$

Es stellt sich nun die Frage, unter welchen Voraussetzungen zur Erzeugung von $U = \mathsf{Lin}(\{v_1, \ldots, v_k\})$ tatsächlich sämtliche Vektoren benötigt werden.

Definition 2.21 (linear abhängig, linear unabhängig)
Endlich viele Vektoren $v_1, \ldots, v_k$ eines $\mathbb{R}$-Vektorraumes V heißen linear abhängig, wenn es Zahlen $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}$ gibt, die nicht sämtlich gleich Null sind, sodass gilt:

$$\alpha_1 v_1 + \ldots + \alpha_k v_k = o. \tag{2.75}$$

Die Vektoren heißen linear unabhängig, wenn sie nicht linear abhängig sind; d.h.

$$\alpha_1 v_1 + \ldots + \alpha_k v_k = o \;\Rightarrow\; \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0. \tag{2.76}$$

Um Vektoren $v_1, \ldots, v_k \in V$ auf lineare Unabhängigkeit zu überprüfen, ist die Gleichung

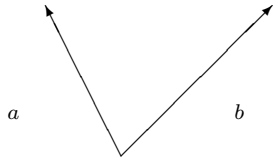
$$x_1 v_1 + \ldots + x_k v_k = o \tag{2.77}$$

zu betrachten. Laut Definition 2.21 gilt:

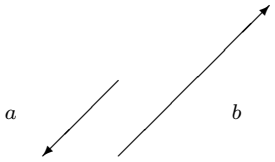
(a) $v_1, \ldots, v_k$ linear abhängig: obige Gleichung hat unendlich viele Lösungen

(b) $v_1, \ldots, v_k$ linear unabhängig: obige Gleichung hat genau eine Lösung, nämlich $x_1 = x_2 = x_3 = \ldots = 0$.

Offensichtlich sind Vektoren $v_1, \ldots, v_k \in V$ genau dann linear abhängig, wenn einer von ihnen als Linearkombination der anderen darstellbar ist.



linear unabhängig



linear abhängig

(2.78)

Beispiel(e) 2.22

- $v_1 = \begin{pmatrix}1\\2\\3\end{pmatrix}$, $v_2 = \begin{pmatrix}0\\2\\1\end{pmatrix}$, $v_3 = \begin{pmatrix}1\\4\\4\end{pmatrix}$ sind linear abhängig, denn es gilt $v_1+v_2-v_3 = \begin{pmatrix}0\\0\\0\end{pmatrix}$ bzw. $v_3 = v_1 + v_2$.
  Anschaulich bedeutet dies, dass v_3 parallel zu der von v_1, v_2 aufgespannten Ebene liegt.
- Die Vektoren $e_1, e_2, \ldots, e_n$ der natürlichen Basis des $\mathbb{R}^n$ sind linear unabhängig, denn das lineare Gleichungssystem

$$x_1 \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} + x_2 \begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix} + \ldots + x_n \begin{pmatrix}0\\\vdots\\0\\1\end{pmatrix} = o \tag{2.79}$$

  hat als einzige Lösung $x_1 = x_2 = \ldots = x_n = 0$.

Betrachtet man eine Matrix in Zeilenstufenform

$$\begin{pmatrix} \diamond & \star & \star & \ldots & & \\ & \diamond & \star & \star & \ldots & \\ & & \diamond & \star & \star & \ldots \\ \mathbf{0} & & & \diamond & \star & \star & \ldots \end{pmatrix} \tag{2.80}$$

so sind die von Null verschiedenen Zeilen linear unabhängig. Analog dazu sind die von Null verschiedenen Spalten einer Matrix in Spaltenstufenform

$$\begin{pmatrix} \diamond & & & \\ \star & \diamond & & \mathbf{0} \\ \star & \star & \diamond & \\ \star & \vdots & \star & \\ \vdots & \vdots & \star & \diamond \\ \star & \star & \star & \star \end{pmatrix} \tag{2.81}$$

linear unabhängig.
Fasst man die zu untersuchenden Vektoren zu einer Matrix zusammen, so ergibt sich der folgende Zusammenhang.

Satz 2.23 (Lineare Unabhängigkeit und Invertierbarkeit)
Für eine $(n \times n)$-Matrix A sind die folgenden Aussagen äquivalent:

a) A ist invertierbar

b) Die Spalten von A sind linear unabhängig

c) Die Zeilen von A sind linear unabhängig

In der folgenden Definition zeichnen wir spezielle Erzeugendensysteme aus.

Definition 2.24 (Basis)
Eine Menge $\{v_1, \ldots, v_n\}$ von Vektoren eines $\mathbb{R}$-Vektorraumes V heißt eine Basis von V, wenn gilt:

(B.1) Die Vektoren $v_1, \ldots, v_n$ sind linear unabhängig

(B.2) $V = \mathsf{Lin}(\{v_1, \ldots, v_n\})$.

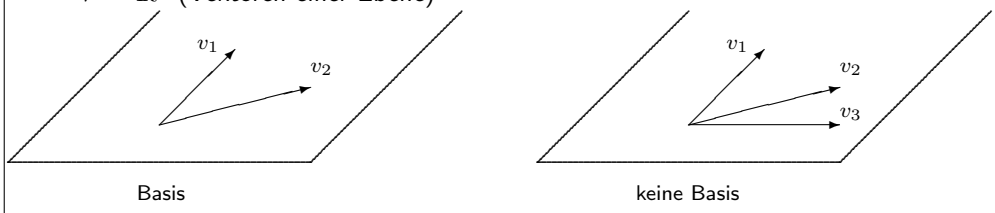
Die Bedeutung einer Basis $\{v_1, \ldots, v_n\}$ eines $\mathbb{R}$-Vektorraumes V ergibt sich aus der Eigenschaft, dass es zu jedem Vektor $v \in V$ genau eine Menge $\{\alpha_1, \ldots, \alpha_n\}$ reeller Zahlen gibt mit

$$v = \sum_{i=1}^{n} \alpha_i v_i. \tag{2.82}$$

Ferner sind m Vektoren aus V immer linear abhängig, falls $m > n$.

Beispiel(e) 2.25

- $V = \mathbb{R}^2$ (Vektoren einer Ebene)



Basis keine Basis (2.83)

- Die Vektoren $e_1, e_2, \ldots, e_n \in \mathbb{R}^n$ bilden eine Basis des $\mathbb{R}^n$
- Die Zeilen (bzw. Spalten) einer invertierbaren $(n \times n)$-Matrix bilden eine Basis des $\mathbb{R}^n$.

In der folgenden Definition betrachten wir besondere $\mathbb{R}$-Vektorräume.

Definition 2.26 (endlichdimensionale $\mathbb{R}$-Vektorräume)
Ein $\mathbb{R}$-Vektorraum V heißt endlichdimensional, falls es endlich viele Vektoren $w_1, \ldots, w_r \in V$ gibt mit

$$V = \mathsf{Lin}(\{w_1, \ldots, w_r\}). \tag{2.84}$$

Die wichtigste Eigenschaft endlichdimensionaler $\mathbb{R}$-Vektorräume $V \neq \{o\}$ besteht in der Tatsache, dass entweder linear unabhängige Vektoren $w_1, \ldots, w_k \in V$ eine Basis von V bilden oder dass diese Vektoren durch Hinzunahme weiterer Vektoren $u_1, \ldots, u_l$ zu einer Basis $\{w_1, \ldots, w_k, u_1, \ldots, u_l\}$ von V ergänzt werden können (Basisergänzungssatz).

Satz und Definition 2.27 (Dimension, Basislänge)
Sei $V \neq \{o\}$ ein endlichdimensionaler $\mathbb{R}$-Vektorraum, so besitzt V eine Basis $\{v_1, \ldots, v_n\}$. Ist $\{w_1, \ldots, w_m\}$ ebenfalls eine Basis von V, so gilt $m = n$. Die gemeinsame Basislänge n aller Basen von V heißt die Dimension von V und wird mit $\mathsf{Dim}(V)$ bezeichnet. Üblicherweise legt man

$$\mathsf{Dim}(\{o\}) = 0 \tag{2.85}$$

fest.

Beispiel(e) 2.28

- $\mathsf{Dim}(\mathbb{R}^n) = n$
- $\mathsf{Dim}\left(\left\{\begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3; 3x + y + z = 0\right\}\right) = 2.$

Ist r die Maximalzahl linear unabhängiger Vektoren aus $v_1, \ldots, v_k$, so gilt:

$$r = \mathsf{Dim}(\mathsf{Lin}(\{v_1, \ldots, v_k\})). \tag{2.86}$$

In einem $\mathbb{R}$-Vektorraum V der Dimension n bilden je n linear unabhängige Vektoren eine Basis von V. Jeder Unterraum U eines endlichdimensionalen $\mathbb{R}$-Vektorraumes V ist ebenfalls endlichdimensional und im Fall $U \neq V$ gilt:

$$\mathsf{Dim}(U) < \mathsf{Dim}(V). \tag{2.87}$$

2.4 Elementarmatrizen

Für eine gegebene $(m \times n)$-Matrix A mit Zeilen $z_1, \ldots, z_m \in \mathbb{R}^n$ und Spalten $s_1, \ldots, s_n \in \mathbb{R}^m$ bezeichnet man mit

$$\mathsf{Lin}(\{z_1, \ldots, z_m\}) = \left\{\sum_{i=1}^{m} \lambda_i z_i;\ \lambda_i \in \mathbb{R}\right\} \tag{2.88}$$

bzw.

$$\mathsf{Lin}(\{s_1, \ldots, s_n\}) = \left\{\sum_{i=1}^{n} \lambda_i s_i;\ \lambda_i \in \mathbb{R}\right\} \tag{2.89}$$

den Zeilen- bzw. Spaltenraum von A. Offensichtlich gilt

$$\mathsf{Lin}(\{z_1, \ldots, z_m\}) = \left\{y^\top A;\ y \in \mathbb{R}^m\right\} \tag{2.90}$$

und

$$\mathsf{Lin}(\{s_1, \ldots, s_n\}) = \left\{Ax;\ x \in \mathbb{R}^n\right\}. \tag{2.91}$$

Wir untersuchen nun die Frage, wann sich durch eine Matrixmultiplikation der Zeilen- bzw. der Spaltenraum nicht ändert.

Satz 2.29 (Zeilenraum und Spaltenraum bei Matrixmultiplikation)
Für alle $A \in \mathbb{R}^{m\times n}$ und alle invertierbaren Matrizen $P \in \mathbb{R}^{m\times m}$ und $Q \in \mathbb{R}^{n\times n}$ gilt:

$$PA \text{ hat denselben Zeilenraum wie } A \tag{2.92}$$

$$AQ \text{ hat denselben Spaltenraum wie } A. \tag{2.93}$$

Beweis:

Wegen $y^\top A = \overbrace{(y^\top P^{-1})}^{z^\top}(PA)$ liegt jedes Element des Zeilenraums von A im Zeilenraum von PA und umgekehrt. Analog mit $(AQ)(Q^{-1}x) = Ax$ für den Spaltenraum.

q.e.d.

Der Zeilenraum einer Matrix $A \in \mathbb{R}^{m\times n}$ kann analog zum Vorgehen bei linearen Gleichungssystemen durch die folgenden elementaren Zeilenumformungen untersucht werden:

(1) Vertauschung zweier Zeilen

(2) Multiplikation einer Zeile mit einem Faktor $\alpha \neq 0$

(3) Addition des Vielfachen einer Zeile zu einer anderen.

Hinter jeder dieser drei elementaren Zeilenumformungen verbirgt sich eine Multiplikation von Matrizen.

zu (1) Die Vertauschung der i-ten mit der j-ten Zeile von A ist gegeben durch $E \cdot A$, wobei

$$E = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & \mathbf{0} & \\ & & 1 & & & & & & & \\ & & & 0 & 0 & \dots & 0 & 1 & & \\ & & & 0 & 1 & & & 0 & & \\ & & & \vdots & & \ddots & & \vdots & & \\ & & & 0 & & & 1 & 0 & & \\ & & & 1 & 0 & \dots & 0 & 0 & & \\ & & & & & & & & 1 & \\ & \mathbf{0} & & & & & & & \ddots & \\ & & & & & & & & & 1 \end{pmatrix} \in \mathbb{R}^{m\times m} \tag{2.94}$$

mit $\alpha_{ij} = \alpha_{ji} = 1$ und $\alpha_{ii} = \alpha_{jj} = 0$. E ist invertierbar und wird Elementarmatrix vom Typ (1) genannt.

zu (2) Die Multiplikation der i-ten Zeile von A mit einem Faktor $\alpha \neq 0$ ist gegeben durch $\overline{E} \cdot A$, wobei die invertierbare Matrix

$$\overline{E} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & \mathbf{0} & \\ & & 1 & & & & \\ & & & \alpha & & & \\ & & & & 1 & & \\ & \mathbf{0} & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \in \mathbb{R}^{m\times m} \tag{2.95}$$

mit $\alpha_{ii} = \alpha$ Elementarmatrix vom Typ (2) genannt wird.

zu (3) Die Addition des α-fachen der i-ten Zeile von A zur j-ten Zeile von A ist gegeben durch $\hat{E} \cdot A$, wobei die invertierbare Matrix

$$\hat{E} = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & \mathbf{0} & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & \alpha & & & & 1 & & \\ & & & & & & & & 1 & \\ & & \mathbf{0} & & & & & & \ddots & \\ & & & & & & & & & 1 \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{2.96}$$

mit $\alpha_{ji} = \alpha$ Elementarmatrix vom Typ (3) genannt wird.

Wird nun eine Matrix $(A|b) \in \mathbb{R}^{m \times (n+1)}$ nach dem Gaußverfahren durch die Vorwärtselimination mit elementaren Zeilenumformungen umgewandelt in eine Matrix $(M|d)$ mit

$$(M|d) = \left(\begin{array}{ccccccccc|c} \diamond & \star & \dots & \star & & & & & & \star \\ & \diamond & \star & \dots & \star & & \mathbf{0} & & & \vdots \\ & & \diamond & \star & \dots & \star & & & & \\ & & & & \diamond & \star & \dots & \star & & \\ & & \mathbf{0} & & & & & & & \vdots \\ & & & & & & & & & \star \end{array}\right), \tag{2.97}$$

so bilden die von Null verschiedenen Zeilen von M eine Basis des Zeilenraumes von M und damit auch des Zeilenraumes von A, denn es gilt

$$M = P \cdot A, \tag{2.98}$$

wobei P eine invertierbare Matrix ist, die ihrerseits aus dem Produkt von Elementarmatrizen des Typs (1), (2) oder (3) besteht. Diese Elementarmatrizen repräsentieren gerade die benötigten elementaren Zeilenumformungen der Vorwärtselimination.
Analog dazu können Elementarmatrizen für elementare Spaltenumformungen definiert werden und somit der Matrix $A \in \mathbb{R}^{m \times n}$ durch

$$\hat{M} = AQ \tag{2.99}$$

mit

$$\hat{M} = \begin{pmatrix} \diamond & & & \\ \star & \diamond & \mathbf{0} & \\ \vdots & \star & \diamond & \\ & \vdots & \star & \\ \vdots & \vdots & \vdots & \mathbf{0} \\ \star & \star & \star & \end{pmatrix} \tag{2.100}$$

eine Matrix $\hat{M}$ mit demselben Spaltenraum wie A zugeordnet werden.
Die von Null verschiedenen Spalten von $\hat{M}$ bilden eine Basis des Spaltenraumes von A.

Definition 2.30 (Rang einer Matrix)
Sei $A \in \mathbb{R}^{m\times n}$ eine $(m \times n)$-Matrix, so wird die Dimension des Zeilenraumes als Rang von A bezeichnet.

Der folgende Satz fasst wichtige Ergebnisse über Matrizen zusammen; dabei sei an den Begriff $\text{Kern}(A) := \{x \in \mathbb{R}^n; Ax = o\}$ für eine $(m \times n)$-Matrix A erinnert.

Satz 2.31 (Eigenschaften von $(m \times n)$-Matrizen)
Für jede $(m \times n)$-Matrix A gilt:

a) Die Dimension des Spaltenraumes von A ist gleich der Dimension des Zeilenraumes von A $(= \text{Rang}(A))$.

b) $\text{Rang}(A) + \text{Dim}(\text{Kern}(A)) = n$.

c) Es gibt invertierbare Matrizen $P \in \mathbb{R}^{m\times m}$ und $Q \in \mathbb{R}^{n\times n}$ mit

$$PAQ = \begin{pmatrix} 1 & & & \\ & \ddots & & \mathbf{0} \\ & & 1 & \\ & \mathbf{0} & & \mathbf{0} \end{pmatrix} \quad \text{mit } \text{Rang}(A) = r. \tag{2.101}$$

d) $\text{Rang}(A^\top) = \text{Rang}(A)$.

e) $\text{Rang}(PAQ) = \text{Rang}(A)$ für alle invertierbaren Matrizen $P \in \mathbb{R}^{m\times m}$ und $Q \in \mathbb{R}^{n\times n}$.

Nach Satz 2.31 ändert sich der Rang einer Matrix durch elementare Zeilen- und Spaltenumformungen nicht.
Mit elementaren Zeilenumformungen kann man die Inverse einer invertierbaren Matrix $A \in \mathbb{R}^{n\times n}$ berechnen. Man beginnt mit der erweiterten Matrix

$$\left(A \;\middle|\; \begin{pmatrix} 1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & 1 \end{pmatrix} \right) \tag{2.102}$$

und führt solange elementare Zeilenumformungen durch, bis man die Form

$$\left(\begin{pmatrix} 1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & 1 \end{pmatrix} \;\middle|\; A^{-1} \right) \tag{2.103}$$

erreicht hat.

Beispiel(e) 2.32
Sei

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} : \tag{2.104}$$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 0 & 1 \end{array}\right) \overset{II. - 2\cdot I.}{\underset{III - I}{\longrightarrow}} \tag{2.105}$$

$$\overset{II. - 2\cdot I.}{\underset{III - I}{\longrightarrow}} \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -2 & 1 & 0 \\ 0 & -2 & -1 & -1 & 0 & 1 \end{array}\right) \overset{III. - \frac{2}{3}\cdot II.}{\longrightarrow} \tag{2.106}$$

$$\overset{III. - \frac{2}{3}\cdot II.}{\longrightarrow} \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -2 & 1 & 0 \\ 0 & 0 & 3 & 1/3 & -2/3 & 1 \end{array}\right) \overset{I. - III.}{\underset{II. + 2\cdot III}{\longrightarrow}} \tag{2.107}$$

$$\overset{I. - III.}{\underset{II. + 2\cdot III}{\longrightarrow}} \left(\begin{array}{ccc|ccc} 1 & 2 & 0 & 2/3 & 2/3 & -1 \\ 0 & -3 & 0 & -4/3 & -1/3 & 2 \\ 0 & 0 & 3 & 1/3 & -2/3 & 1 \end{array}\right) \overset{-\frac{1}{3}\cdot II.}{\underset{\frac{1}{3}\cdot III.}{\longrightarrow}} \tag{2.108}$$

$$\overset{-\frac{1}{3}\cdot II.}{\underset{\frac{1}{3}\cdot III.}{\longrightarrow}} \left(\begin{array}{ccc|ccc} 1 & 2 & 0 & 2/3 & 2/3 & -1 \\ 0 & 1 & 0 & 4/9 & 1/9 & -2/3 \\ 0 & 0 & 1 & 1/9 & -2/9 & 1/3 \end{array}\right) \overset{I. - 2\cdot II.}{\longrightarrow} \tag{2.109}$$

$$\overset{I. - 2\cdot II.}{\longrightarrow} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -2/9 & 4/9 & 1/3 \\ 0 & 1 & 0 & 4/9 & 1/9 & -2/3 \\ 0 & 0 & 1 & 1/9 & -2/9 & 1/3 \end{array}\right) \tag{2.110}$$

Mit elementaren Spaltenumformungen kann man eine Basis des Kerns einer Matrix $A \in \mathbb{R}^{m\times n}$ berechnen.

Man beginnt mit der erweiterten Matrix $\left(\begin{array}{c} A \\ \hline \begin{pmatrix} 1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & 1 \end{pmatrix} \end{array}\right)$ und berechnet die Spaltenstufenform

$$\left(\begin{array}{ccccc|c} \diamond & & & & & \\ \star & \diamond & & & & \\ \vdots & \star & \diamond & & \mathbf{0} & \\ \vdots & \vdots & \vdots & & & \\ \star & \star & \star & & & \\ \hline & & & & & \\ & \star & & & v_1, \ldots, v_r & \end{array}\right) \tag{2.111}$$

Die Vektoren $v_1, \ldots, v_r$ bilden dann eine Basis von Kern(A).

Beispiel(e) 2.33
Sei

$$A = \begin{pmatrix} 3 & 6 & 9 & -3 \\ 0 & 2 & 6 & 10 \\ 15 & 34 & 58 & 3 \\ -9 & -10 & -3 & 49+\alpha \end{pmatrix}, \quad \alpha \in \mathbb{R} \tag{2.112}$$

$$\begin{pmatrix} 3 & 6 & 9 & -3 \\ 0 & 2 & 6 & 10 \\ 15 & 34 & 58 & 3 \\ -9 & -10 & -3 & 49+\alpha \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 6 & 10 \\ 15 & 4 & 13 & 18 \\ -9 & 8 & 24 & 40+\alpha \\ 1 & -2 & -3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \longrightarrow \dots \tag{2.113}$$

$$\longrightarrow \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 15 & 4 & 1 & 0 \\ -9 & 8 & 0 & \alpha \\ 1 & -2 & 3 & 17 \\ 0 & 1 & -3 & -11 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.114}$$

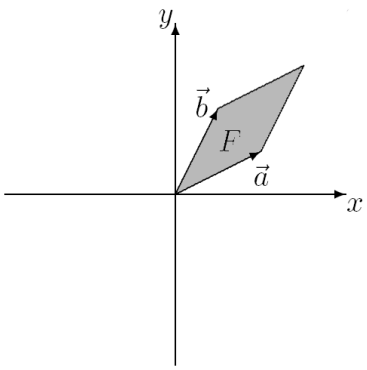
Fall 1: $\alpha \neq 0 \Longrightarrow \mathsf{Rang}(A) = 4 \Longrightarrow \mathsf{Kern}(A) = \{o\}$

Fall 2: $\alpha = 0 \Longrightarrow \mathsf{Rang}(A) = 3 \Longrightarrow \mathsf{Kern}(A) = \left\{ t \cdot \begin{pmatrix} 17 \\ -11 \\ 2 \\ 1 \end{pmatrix}; t \in \mathbb{R} \right\}$

2.5 Determinanten

Determinanten wurden von G. W. Leibniz bereits 1678 eingeführt. Sie dienen unter anderem zur Beschreibung elementargeometrischer Lösungen und zur speziellen Darstellung von Lösungen linearer Gleichungssysteme. Betrachtet man zum Beispiel zwei Vektoren $\vec{a} = \binom{a_1}{a_2}$, $\vec{b} = \binom{b_1}{b_2}$, so ist die Fläche des in der (x, y)-Ebene durch $\vec{a}$, $\vec{b}$ aufgespannten Parallelogrammes gegeben durch

$$F = \left| \begin{pmatrix} a_1 \\ a_2 \\ 0 \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ 0 \end{pmatrix} \right| = |a_1 b_2 - b_1 a_2|. \tag{2.115}$$



(2.116)

Fasst man die Vektoren $\vec{a}$, $\vec{b}$ zu einer (2×2)-Matrix $A = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}$ zusammen, so heißt die Zahl

$$\det(A) := a_1 b_2 - b_1 a_2 \tag{2.117}$$

Determinante der (2×2)-Matrix A. Es ist

$$\det(A) = \pm F = 0 \tag{2.118}$$

genau dann, wenn $\vec{a}$, $\vec{b}$ parallel sind (oder einer der beiden der Nullvektor ist). Damit ist gezeigt:

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \text{ linear unabhängig } \Longleftrightarrow \det\left(\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}\right) \neq 0 \Longleftrightarrow A \text{ invertierbar.} \tag{2.119}$$

Beispiel(e) 2.34

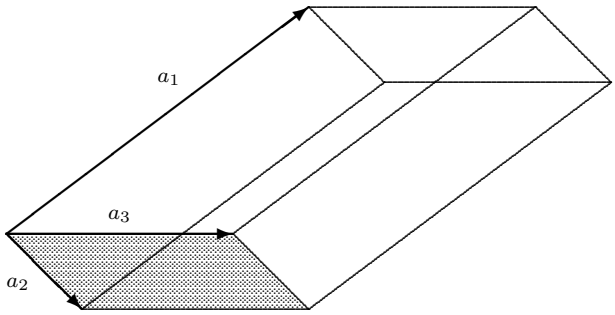
$$\det\left(\begin{pmatrix} 1 & 3 \\ 2 & 5 \end{pmatrix}\right) = 1 \cdot 5 - 2 \cdot 3 = -1 \neq 0. \tag{2.120}$$

Die Matrix ist invertierbar:

$$\begin{pmatrix} 1 & 3 \\ 2 & 5 \end{pmatrix}^{-1} = (-1) \begin{pmatrix} 5 & -3 \\ -2 & 1 \end{pmatrix}. \tag{2.121}$$

Für einen Spat mit den bezüglich einer kartesischen Basis dargestellten Kanten

$$a_1 = \begin{pmatrix} \alpha_{11} \\ \alpha_{21} \\ \alpha_{31} \end{pmatrix}, \; a_2 = \begin{pmatrix} \alpha_{12} \\ \alpha_{22} \\ \alpha_{32} \end{pmatrix}, \; a_3 = \begin{pmatrix} \alpha_{13} \\ \alpha_{23} \\ \alpha_{33} \end{pmatrix} \tag{2.122}$$



(2.123)

kann das Volumen berechnet werden. Es ergibt sich:

$$V = |\alpha_{11}(\alpha_{22}\alpha_{33} - \alpha_{32}\alpha_{23}) - \alpha_{21}(\alpha_{12}\alpha_{33} - \alpha_{32}\alpha_{13}) + \alpha_{31}(\alpha_{12}\alpha_{23} - \alpha_{22}\alpha_{13})| \,. \tag{2.124}$$

Fasst man die Vektoren $\vec{a}_1$, $\vec{a}_2$, $\vec{a}_3$ zu einer (3×3)-Matrix $A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}$ zusammen, so heißt die Zahl

$$\mathsf{det}(A) = \alpha_{11}(\alpha_{22}\alpha_{33} - \alpha_{32}\alpha_{23}) - \alpha_{21}(\alpha_{12}\alpha_{33} - \alpha_{32}\alpha_{13}) + \alpha_{31}(\alpha_{12}\alpha_{23} - \alpha_{22}\alpha_{13}) \tag{2.125}$$

Determinante der (3×3)-Matrix A. Es gilt:

$$\mathsf{Rang}(A) < 3 \iff \mathsf{det}(A) = 0 \tag{2.126}$$

$$\mathsf{Rang}(A) = 3 \iff A \text{ invertierbar} \iff \mathsf{det}(A) \neq 0. \tag{2.127}$$

Ferner gilt:

$$\mathsf{det}(A) = \alpha_{11}\mathsf{det}\left(\begin{pmatrix} \alpha_{22} & \alpha_{23} \\ \alpha_{32} & \alpha_{33} \end{pmatrix}\right) - \alpha_{21}\mathsf{det}\left(\begin{pmatrix} \alpha_{12} & \alpha_{13} \\ \alpha_{32} & \alpha_{33} \end{pmatrix}\right) + \alpha_{31}\mathsf{det}\left(\begin{pmatrix} \alpha_{12} & \alpha_{13} \\ \alpha_{22} & \alpha_{23} \end{pmatrix}\right). \tag{2.128}$$

Beispiel(e) 2.35

$$A = \begin{pmatrix} 0 & -2 & 3 \\ -2 & 1 & -2 \\ 3 & 6 & 5 \end{pmatrix} \tag{2.129}$$

$$\mathsf{det}(A) = 0 \cdot (5 + 12) + 2(-10 - 18) + 3(4 - 3) = -53. \tag{2.130}$$

Analog zur Definition der Determinante für (3×3)-Matrizen lassen sich Determinanten für $(n\times n)$-Matrizen rekursiv definieren:
Für $n = 1$, d.h. $A = (\alpha_{11})$, ist $\mathsf{det}(A) = \alpha_{11}$.
Für $n \geq 2$ ist (Entwicklung von $\mathsf{det}(A)$ nach der ersten Spalte):

$$\mathsf{det}(A) := \alpha_{11}\mathsf{det}(A_{11}) - \alpha_{21}\mathsf{det}(A_{21}) + \alpha_{31}\mathsf{det}(A_{31}) \pm \ldots + (-1)^{n+1}\alpha_{n1}\mathsf{det}(A_{n1}), \tag{2.131}$$

wobei A_{i1} die $((n-1)\times(n-1))$-Matrix bezeichnet, die aus A durch Entfernen der ersten Spalte und der i-ten Zeile entsteht.

Beispiel(e) 2.36

$$A = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 3 & -2 & 1 & 0 \\ 0 & 6 & 3 & -2 \\ 2 & 4 & 3 & 1 \end{pmatrix} \tag{2.132}$$

$$A_{11} = \begin{pmatrix} -2 & 1 & 0 \\ 6 & 3 & -2 \\ 4 & 3 & 1 \end{pmatrix}, \quad A_{21} = \begin{pmatrix} 2 & 0 & 1 \\ 6 & 3 & -2 \\ 4 & 3 & 1 \end{pmatrix}, \tag{2.133}$$

$$A_{31} = \begin{pmatrix} 2 & 0 & 1 \\ -2 & 1 & 0 \\ 4 & 3 & 1 \end{pmatrix}, \quad A_{41} = \begin{pmatrix} 2 & 0 & 1 \\ -2 & 1 & 0 \\ 6 & 3 & -2 \end{pmatrix}, \tag{2.134}$$

$$\det(A) = 1 \cdot |A_{11}| - 3 \cdot |A_{21}| - 2 \cdot |A_{41}| = -72. \tag{2.135}$$

Für obere Dreiecksmatrizen läßt sich die Determinante einfach berechnen.

Satz 2.37 (Determinante einer oberen Dreiecksmatrix)
Für eine obere Dreiecksmatrix gilt

$$\det\left(\begin{pmatrix} \alpha_{11} & & \star \\ & \ddots & \\ \mathbf{0} & & \alpha_{nn} \end{pmatrix}\right) = \alpha_{11} \cdot \ldots \cdot \alpha_{nn} = \prod_{i=1}^{n} \alpha_{ii}. \tag{2.136}$$

Der Beweis besteht aus einfachem Nachrechnen. Wichtig sind die folgenden Rechenregeln für Determinanten.

Satz 2.38 (Rechenregeln für Determinanten)

Sei A eine $(n \times n)$-Matrix mit der Zeilendarstellung $A = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$ und der Spaltendarstellung $A = (s_1, \ldots, s_n)$, so gilt:

(i) Besteht die i-te Zeile (bzw. Spalte) von A aus einer Summe, also $z_i = a + b$ (bzw. $s_i = c + d$), $i \in \{1, \ldots, n\}$, so gilt:

$$\det \begin{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ a+b \\ \vdots \\ z_n \end{pmatrix} \end{pmatrix} = \det \begin{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ a \\ \vdots \\ z_n \end{pmatrix} \end{pmatrix} + \det \begin{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ b \\ \vdots \\ z_n \end{pmatrix} \end{pmatrix} \tag{2.137}$$

bzw.

$$\begin{aligned} \det((s_1, \ldots, c+d, \ldots, s_n)) &= \det((s_1, \ldots, c, \ldots, s_n)) + \\ &\quad + \det((s_1, \ldots, d, \ldots, s_n)). \end{aligned} \tag{2.138}$$

(ii) Sei nun $B = E \cdot A$, wobei E eine Elementarmatrix vom Typ (1), (2) oder (3) ist, so gilt:

$$\begin{aligned} &\det(B) = -\det(A), \text{ falls } E \text{ vom Typ } (1) \text{ ist;} \\ &\det(B) = \alpha \det(A), \text{ falls } E \text{ vom Typ } (2) \text{ ist;} \\ &\det(B) = \det(A), \quad \text{falls } E \text{ vom Typ } (3) \text{ ist;} \end{aligned}$$

(iii) $\det(A^\top) = \det(A)$

(iv) Sei C eine weitere $(n \times n)$-Matrix, so gilt

$$\det(AC) = \det(A) \cdot \det(C) \; (= \det(C) \cdot \det(A) = \det(C \cdot A)) \tag{2.139}$$

(v) A ist invertierbar $\iff \det(A) \neq 0$.

Da die Determinante rekursiv definiert ist, wird im Beweis des obigen Satzes häufig die vollständige Induktion verwendet.
Sei $A = (a_1, \ldots, a_n)$ eine $(n \times n)$-Matrix in Spaltendarstellung, so entsteht

$$\widetilde{A} = (a_j, a_1, a_2, \ldots, a_{j-1}, a_{j+1}, \ldots, a_n) \tag{2.140}$$

durch $(j-1)$ sukzessive Vertauschungen benachbarter Spalten. Also gilt

$$\det(\widetilde{A}) = (-1)^{j-1} \cdot \det(A). \tag{2.141}$$

Berechnet man nun $\det(\widetilde{A})$ wie bisher, so erhält man eine neue Formel für $\det(A)$:

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} \alpha_{ij} \det(A_{ij}), \tag{2.142}$$

wobei die $((n-1)\times(n-1))$-Matrix A_{ij} durch Streichen der i-ten Zeile und der j-ten Spalte aus der Matrix A entsteht.
Mit Hilfe von Determinanten läßt sich die Lösung $x \in \mathbb{R}^n$ eines linearen Gleichungssystems $Ax = b$ gegeben durch eine invertierbare Matrix $A \in \mathbb{R}^{n\times n}$ darstellen (Cramer-Regel):

$$x_i = \frac{\det((a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n))}{\det(A)}, \quad i = 1, \ldots, n, \tag{2.143}$$

wobei $A = (a_1, \ldots, a_n)$ die Spaltendarstellung der Matrix A ist (die i-te Spalte von A wird durch b ersetzt).

2.6 Lineare Abbildungen und Eigenwerte

Seien V, W $\mathbb{R}$-Vektorräume. Mit einer Abbildung $f : V \to W$ wird jedem Vektor $v \in V$ ein eindeutiger Vektor $w \in W$ zugeordnet.

Definition 2.39 (Lineare Abbildung)
Seien V, W $\mathbb{R}$-Vektorräume und $f : V \to W$ eine Abbildung, dann heißt f linear, falls gilt:

(L.1) f ist homogen; d.h. $f(\alpha v) = \alpha \cdot f(v)$ für alle $\alpha \in \mathbb{R}$, $v \in V$.

(L.2) f ist additiv; d.h. $f(u + v) = f(u) + f(v)$ für alle u, $v \in V$.

Eine lineare Abbildung $f : V \to W$ wird auch als linearer Operator, lineare Transformation oder Vektorraumhomomorphismus bezeichnet.

Beispiel(e) 2.40
- Die Nullabbildung $N : V \to V$, $v \mapsto o$ für alle $v \in V$.
- Die Projektionen $p_i : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto x_i$.
- $l : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto Ax$ mit einer festen Matrix $A \in \mathbb{R}^{m\times n}$.

Keine linearen Abbildungen sind zum Beispiel:
- $f : \mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \mapsto (x^2, x + y)$.
- $t_a : V \to V$, $v \mapsto v + a$, $a \in V$, $a \neq o$ (Parallelverschiebung).

Sind zwei lineare Abbildungen f, $g : V \to W$ gegeben, so ist auch die Summe

$$f + g : V \to W, \quad v \mapsto f(v) + g(v) \tag{2.144}$$

und das α-fache

$$\alpha f : V \to W, \quad v \mapsto \alpha \cdot f(v), \quad \alpha \in \mathbb{R} \tag{2.145}$$

lineare Abbildungen.
Mit dieser Addition und skalaren Multiplikation ist die Menge aller linearen Abbildungen von V nach W

$$\mathrm{Hom}(V, W) = \{f : V \to W;\ f \text{ linear}\} \tag{2.146}$$

selbst wieder ein $\mathbb{R}$-Vektorraum.
Sind drei $\mathbb{R}$-Vektorräume U, V, W und zwei lineare Abbildungen

$$g : U \to V \quad \text{und} \quad f : V \to W \tag{2.147}$$

gegeben, so ist auch die Hintereinanderausführung („Komposition“)

$$f \circ g : U \to W, \quad u \mapsto f(g(u)) \tag{2.148}$$

eine lineare Abbildung.
Ist V ein endlichdimensionaler $\mathbb{R}$-Vektorraum und $B = \{v_1, \ldots, v_n\}$ eine Basis von V, dann läßt sich jeder Vektor $v \in V$ eindeutig in der Form

$$v = \alpha_1 v_1 + \ldots + \alpha_n v_n \tag{2.149}$$

mit $\alpha_i \in \mathbb{R}$ darstellen. Man nennt

$$v_B := \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \in \mathbb{R}^n \quad \left(\text{mit } v = \sum_{i=1}^{n} \alpha_i v_i \right) \tag{2.150}$$

den Koordinatenvektor von $v \in V$ bezüglich B. Die lineare Abbildung

$$k_B : V \to \mathbb{R}^n, \quad v \mapsto v_B \tag{2.151}$$

heißt Koordinatenabbildung bezüglich B.
Mit dieser Abbildung überträgt man Probleme aus V in den $\mathbb{R}^n$ (die Grundidee dieser Vorgehensweise stammt von R. Descartes).
Man kann die Abbildung k_B auch umkehren:

$$k_B^{-1} : \mathbb{R}^n \to V, \quad (\alpha_1, \ldots, \alpha_n) \mapsto \sum_{i=1}^{n} \alpha_i v_i. \tag{2.152}$$

Im Folgenden untersuchen wir lineare Abbildungen $f : \mathbb{R}^n \to \mathbb{R}^n$.

Satz 2.41 (Darstellung linearer Abbildungen)
Sei $f : \mathbb{R}^n \to \mathbb{R}^n$ eine lineare Abbildung und $F = (f(e_1), \ldots, f(e_n))$ diejenige $(n \times n)$-Matrix in Spaltendarstellung, deren Spalten aus den Bildern der Einheitsvektoren $e_1, \ldots, e_n$ (natürliche Basis des $\mathbb{R}^n$) besteht, dann gilt:

$$f(x) = Fx \quad \text{für alle } x \in \mathbb{R}^n \tag{2.153}$$

Man nennt F die Abbildungsmatrix von f bezüglich der natürlichen Basis $e_1, \ldots, e_n$.

Im Folgenden untersuchen wir die Begriffe „Länge, Winkel und Orthogonalität“ im $\mathbb{R}^n$.

Definition 2.42 (Skalarprodukt, Betrag, Länge)

Seien $x,\, y \in \mathbb{R}^n$, $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.

Die Zahl

$$x^\top y = \sum_{i=1}^{n} x_i y_i \tag{2.154}$$

heißt Skalarprodukt der Vektoren x, y und

$$|x| := \sqrt{x^\top x} = \sqrt{x_1^2 + \ldots + x_n^2} \tag{2.155}$$

heißt Betrag oder Länge des Vektors x.

Beispiel(e) 2.43

$$x = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 4 \end{pmatrix}, \quad y = \begin{pmatrix} -1 \\ 3 \\ 5 \\ -3 \end{pmatrix} \quad \Longrightarrow \quad x^\top y = -9 \quad |x| = \sqrt{21} \quad |y| = \sqrt{44}. \tag{2.156}$$

Ein Vektor x heißt normiert oder Einheitsvektor, falls $|x| = 1$.
Zu jedem Vektor $x \neq o$ ist $\frac{1}{|x|} \cdot x$ normiert.
Wie im $\mathbb{R}^2$ und $\mathbb{R}^3$ lassen sich die folgenden Rechenregeln für alle $x,\, y,\, z \in \mathbb{R}^n$ und alle $\alpha \in \mathbb{R}$ nachweisen:

(a) $x^\top y = y^\top x$

(b) $\alpha(x^\top y) = (\alpha x)^\top y = x^\top(\alpha y)$

(c) $x^\top(y + z) = x^\top y + x^\top z$

(d) $x^\top x > 0$ für alle $x \neq o$

(e) $|x| = 0 \Longleftrightarrow x = o$

(f) $|\alpha x| = |\alpha| \cdot |x|$

(g) $|x^\top y| \leq |x| \cdot |y|$ (Cauchy-Schwarzsche Ungleichung)

(h) $|x + y| \leq |x| + |y|$ (Dreiecksungleichung)

(2.157)

Wegen (g) gilt für $x \neq o$, $y \neq o$ stets $\frac{|x^\top y|}{|x| \cdot |y|} \leq 1$ und somit

$$-1 \leq \frac{x^\top y}{|x| \cdot |y|} \leq 1. \tag{2.158}$$

Deshalb gibt es genau einen Winkel φ mit $0 \leq \varphi \leq \pi$ und

$$\cos(\varphi) = \frac{x^\top y}{|x| \cdot |y|} \tag{2.159}$$

Wir legen daher fest:

Definition 2.44 (Winkel zwischen Vektoren, orthogonale Vektoren)
Sei $x,\ y \in \mathbb{R}^n$ und $x,\ y \neq o$, so nennt man $\angle(x,y) := \varphi = \cos^{-1}\left(\frac{x^\top y}{|x|\cdot|y|}\right)$ (auch als $\arccos\left(\frac{x^\top y}{|x|\cdot|y|}\right)$ geschrieben) den Winkel zwischen x und y, wobei $\cos^{-1} : [-1,1] \to [0,\pi]$ die Umkehrfunktion der auf $[0,\pi]$ eingschränkten Cosinus-Funktion bezeichnet.
Ferner nennt man $x,\ y \in \mathbb{R}^n$ orthogonal, falls $x^\top y = 0$.
Der Nullvektor ist orthogonal zu allen Vektoren im $\mathbb{R}^n$.

Der wichtige Begriff der Orthogonalität wird nun auf lineare Abbildungen, Matrizen und Basen erweitert.

Definition 2.45 (Orthogonalität von linearen Abbildungen, Matrizen und Basen)

(L.1) Sei $f : \mathbb{R}^n \to \mathbb{R}^n$ eine lineare Abbildung, dann heißt f orthogonal, falls sie das Skalarprodukt invariant läßt, d.h. wenn für alle $x,\ y \in \mathbb{R}^n$ gilt:

$$f(x)^\top f(y) = x^\top y. \tag{2.160}$$

(L.2) Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt orthogonal, falls

$$A^\top A = E \quad (\text{also } A^\top = A^{-1}). \tag{2.161}$$

(L.3) Eine Basis $\{b_1,\ldots,b_n\}$ heißt orthogonal, falls die Vektoren b_i paarweise orthogonal sind; d.h. wenn für $i \neq j$ gilt

$$b_i^\top b_j = 0. \tag{2.162}$$

Die Basis B heißt orthonormal, wenn sie orthogonal ist und alle Basisvektoren normiert sind.

Nach der Definition orthogonaler Vektoren ist der Winkel zwischen diesen Vektoren gleich $\frac{\pi}{2}$, d.h. diese Vektoren stehen senkrecht aufeinander. Der Begriff „orthogonal“ kommt aus dem Altgriechischen „`orthos`“ gerade, richtig und „`gony`“ das Knie. Der „rechte“ Winkel ($= \frac{\pi}{2}$) zwischen Ober- und Unterschenkel ist für das Knie die „richtige“ Sitzposition.

Beispiel(e) 2.46

- Die natürliche Basis $\{e_1,\ldots,e_n\}$ ist eine Orthonormalbasis des $\mathbb{R}^n$

Den Zusammenhang zwischen den Orthonormalitätsbegriffen klärt der folgende

Satz 2.47 (Zusammenhang zwischen den Orthogonalitätsbegriffen)
Für eine Matrix $A \in \mathbb{R}^{n\times n}$ sind äquivalent:

(L.1) A ist orthogonal

(L.2) $(Ax)^\top Ay = x^\top y$ für alle $x,\, y \in \mathbb{R}^n$

(L.3) Die Spalten von A bilden eine orthonormale Basis des $\mathbb{R}^n$.
Insbesondere ist eine lineare Abbildung $f : \mathbb{R}^n \to \mathbb{R}^n$ genau dann orthogonal, wenn die Abbildungsmatrix $F = (f(e_1), \ldots, f(e_n))$ orthogonal ist.

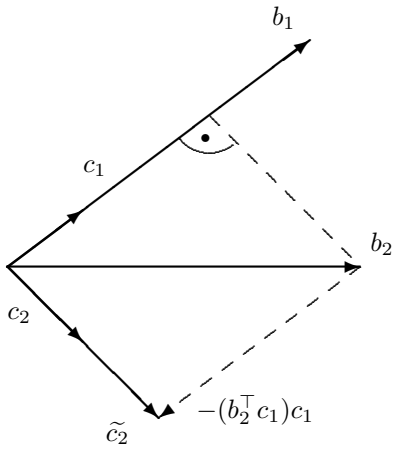
Seien $\{b_1, \ldots, b_r\}$ linear unabhängige Vektoren des $\mathbb{R}^n$, so ermöglicht das nun folgende Schmidtsche Orthonormalisierungsverfahren die Berechnung von Vektoren $\{c_1, \ldots, c_k\}$ mit

$$c_i^\top c_j = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{sonst} \end{cases} \tag{2.163}$$

und

$$\mathsf{Lin}\,(\{b_1, \ldots, b_k\}) = \mathsf{Lin}\,(\{c_1, \ldots, c_k\}) : \tag{2.164}$$

$$c_1 := \frac{b_1}{|b_1|}, \qquad \widetilde{c_i} := b_i - \sum_{m=1}^{i-1} (b_i^\top c_m) c_m, \qquad c_i = \frac{1}{|\widetilde{c_i}|}\,\widetilde{c_i}, \qquad i = 2, \ldots, k. \tag{2.165}$$



(2.166)

Faßt man eine Basis $\{b_1, \ldots, b_n\}$ des $\mathbb{R}^n$ zu einer $(n \times n)$-Matrix $B = (b_1, \ldots, b_n)$ in Spaltendarstellung zusammen, so erhält man bei der Anwendung des Schmidtschen Orthonormalisierungsverfahrens auf $\{b_1, \ldots, b_n\}$ eine Orthonormalbasis $\{q_1, \ldots, q_n\}$ des $\mathbb{R}^n$, welche zu einer Matrix $Q = (q_1, \ldots, q_n)$ in Spaltendarstellung führt.
Zwischen den Matrizen B und Q besteht nun der folgende Zusammenhang:

$$Q = B\widetilde{R}, \tag{2.167}$$

wobei $\widetilde{R}$ eine obere Dreiecksmatrix darstellt, da für die Berechnung der i-ten Spalte von Q nur die ersten i Spalten von B benötigt werden.
Da die Matrix $\widetilde{R}$ invertierbar ist, kann man nach B auflösen:

$$B = Q \cdot \widetilde{R}^{-1}. \tag{2.168}$$

Da nun $R := \widetilde{R}^{-1}$ erneut eine obere Dreiecksmatrix ist, wird die Gleichung

$$B = QR \tag{2.169}$$

als QR-Zerlegung einer invertierbaren Matrix $B \in \mathbb{R}^{n\times n}$ mit einer orthogonalen Matrix Q und einer oberen Dreiecksmatrix R bezeichnet.
Die QR-Zerlegung wird häufig bei numerischen Verfahren angewendet.
Sei $\{b_1, \ldots, b_n\}$ eine Basis des $\mathbb{R}^n$, die wir zu einer invertierbaren $(n \times n)$-Matrix $B = (b_1, \ldots, b_n)$ in Spaltendarstellung zusammenfassen. Die eindeutig bestimmten Koeffizienten $x'_1, \ldots, x'_n \in \mathbb{R}$ der Zerlegung

$$x = \sum_{i=1}^{n} x'_i b_i \tag{2.170}$$

eines Vektors $x \in \mathbb{R}^n$ heißen Koordinaten des Vektors $x \in \mathbb{R}^n$ bezüglich der Basis B. Man nennt

$$x_B := \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix} = x' \quad \text{mit } x = \sum_{i=1}^{n} x'_i b_i \tag{2.171}$$

den Koordinatenvektor von $x \in \mathbb{R}^n$ bezüglich B. Die Abbildung $x \mapsto x_B$ ist linear und es gilt:

$$x = Bx_B \quad (\text{Basiswechsel von } \{b_1, \ldots, b_n\} \text{ zu } \{e_1, \ldots, e_n\})\,. \tag{2.172}$$

$$x_B = B^{-1}x \quad (\text{Basiswechsel von } \{e_1, \ldots, e_n\} \text{ zu } \{b_1, \ldots, b_n\}) \tag{2.173}$$

Jeder $(n \times n)$-Matrix A läßt sich die lineare Abbildung $f : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto Ax$ zuordnen; dabei ist A die Abbildungsmatrix von f bezüglich der natürlichen Basis. Nun betrachten wir die Abbildungsmatrix von f bezüglich einer anderen Basis des $\mathbb{R}^n$.

Definition 2.48 (Abbildungsmatrix bezüglich B)
Seien $f : \mathbb{R}^n \to \mathbb{R}^n$ eine lineare Abbildung und $B = (b_1, \ldots, b_n) \in \mathbb{R}^{n\times n}$ eine invertierbare Matrix in Spaltendarstellung. Die Matrix

$$C := (f(b_1)_B, \ldots, f(b_n)_B), \tag{2.174}$$

in deren Spalten die Bilder der Basisvektoren $b_1, \ldots, b_n$ bezüglich der Basis B stehen, heißt Abbildungsmatrix von f bezüglich B.

Mit Definition 2.48 können wir nun die Änderung der Abbildungsmatrix einer linearen Abbildung bei Basiswechsel beschreiben.

Satz 2.49 (Änderung der Abbildungsmatrix bei Basiswechsel)
Seien $A \in \mathbb{R}^{n\times n}$ und $B = (b_1, \ldots, b_n)$ eine invertierbare $(n \times n)$-Matrix in Spaltendarstellung. Ein Vektor $y = Ax$ lautet in Koordinaten bezüglich der Basis $\{b_1, \ldots, b_n\}$:

$$y_B = (Ax)_B = B^{-1}Ax = B^{-1}ABx_B. \tag{2.175}$$

Für die Abbildungsmatrix C von $f : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto Ax$ bezüglich B gilt daher:

$$C = B^{-1}AB. \tag{2.176}$$

Zu einer linearen Abbildung $f : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto Ax$ gibt es somit je nach Wahl der Basis neben der Matrix A noch verschiedene andere Abbildungsmatrizen, die die lineare Abbildung f repräsentieren.

Definition 2.50 (ähnliche Matrizen)
Zwei $(n \times n)$-Matrizen A, C heißen ähnlich, falls es eine invertierbare $(n \times n)$-Matrix B gibt mit $C = B^{-1}AB$

Ähnliche Matrizen repräsentieren dieselbe lineare Abbildung aber unter Verwendung verschiedener Basen. Es stellt sich im Folgenden die Frage, wie eine Basis des $\mathbb{R}^n$ gewählt werden muss, damit die zugehörige Abbildungsmatrix einer gegebenen linearen Abbildung möglichst einfach wird. Diese Fragestellung wird als algebraisches Eigenwertproblem bezeichnet.
Die Behandlung des algebraischen Eigenwertproblems ist eng verbunden mit der Nullstellenbestimmung von Polynomen, also von Funktionen

$$f : \mathbb{C} \to \mathbb{C}, \quad z \mapsto a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0 \tag{2.177}$$

mit $a_i \in \mathbb{C}$, $i = 1, \ldots, n$.
Daher lassen wir im Folgenden auch komplexe Zahlen als Skalare und als Matrixelemente zu. Die bisherigen Untersuchungen über Vektoren und Matrizen ändern sich dadurch nicht.

Definition 2.51 (Eigenwert, Eigenvektor)
Eine Zahl $\lambda \in \mathbb{C}$ heißt Eigenwert einer Matrix $A \in \mathbb{C}^{n \times n}$, falls es wenigstens einen Spaltenvektor $b \in \mathbb{C}^n$, $b \neq o$, gibt mit

$$Ab = \lambda b. \tag{2.178}$$

Jeder Vektor $b \neq o$, der diese Gleichung erfüllt, heißt Eigenvektor von A zum Eigenwert λ.

Beispiel(e) 2.52

Die Matrix $A = \begin{pmatrix} 0 & -1 & 0 \\ -1 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ hat den Eigenwert $\lambda = -2$ mit dem Eigenvektor $b = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$, denn es gilt:

$$Ab = -2 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$$

Zur Berechnung der Eigenwerte einer $(n \times n)$-Matrix betrachtet man das charakteristische Polynom χ_A:

$$\chi_A : \mathbb{C} \to \mathbb{C}, \quad \lambda \mapsto \det(A - \lambda \cdot E_n) \tag{2.179}$$

Diese Determinate muss berechnet werden.

Definition 2.53 (Eigenwerte und charakteristisches Polynom)
Es ist $\lambda \in \mathbb{C}$ genau dann ein Eigenwert der $(n \times n)$-Matrix A, falls λ eine Nullstelle des charakteristischen Polynoms χ_A ist, d.h. falls

$$\det(A - \lambda E_n) = 0. \tag{2.180}$$

Die Berechnung der Eigenwerte erfolgt daher in zwei Schritten:

(1) Berechne die Nullstellen $\lambda_1, \ldots, \lambda_r$ der charakteristischen Gleichung

$$\chi_A(\lambda) = 0. \tag{2.181}$$

Die Vielfachheit k_i der Nullstelle λ_i heißt algebraische Vielfachheit des Eigenwerts λ_i.

(2) Berechne zu jedem Eigenwert λ_i, $i = 1, \ldots, r$, den Lösungsraum des homogenen linearen Gleichungssystems $(A - \lambda_i E_n)x = o$.
Jede Lösung $b \neq o$ ist ein Eigenvektor zu λ_i.

$$V(\lambda_i) := \{b \in \mathbb{C}^n;\ (A - \lambda_i E_n)b = o\} = \mathsf{Kern}(A - \lambda_i E_n) \tag{2.182}$$

heißt Eigenraum zum Eigenwert λ_i. Man nennt $\mathsf{Dim}(V(\lambda_i))$ die geometrische Vielfachheit des Eigenwertes λ_i.
Algebraische und geometrische Vielfachheit stimmen im allgemeinen nicht überein!

Beispiel(e) 2.54

- $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$,
  $\det(A - \lambda E_n) = \det \begin{pmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{pmatrix} = (1-\lambda)^2 - 4 = \chi_A(\lambda)$
  $\chi_A(\lambda) = 0 \Longleftrightarrow (1-\lambda)^2 = 4 \Longleftrightarrow \lambda_1 = -1, \quad \lambda_2 = +3.$
  Eigenvektoren zu $\lambda_1 = -1$

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} x = -x \Longleftrightarrow \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} x = 0 \Longleftrightarrow x = \mu \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \mu \in \mathbb{R},$$

$$V(\lambda_1) = \left\{ \mu \begin{pmatrix} -1 \\ 1 \end{pmatrix};\ \mu \in \mathbb{R} \right\}.$$

  Eigenvektoren zu $\lambda_2 = +3$

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} x = 3x \Longleftrightarrow \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix} x = 0 \Longleftrightarrow x = \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu \in \mathbb{R},$$

$$V(\lambda_2) = \left\{ \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix};\ \mu \in \mathbb{R} \right\},$$

  Algebraische Vielfachheit gleich geometrische Vielfachheit.

- $A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$, $\lambda_1 = \lambda_2 = 3, V(\lambda_1) = \mathbb{R}^2$.

- $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, $\lambda_1 = \lambda_2 = 2, V(\lambda_1) = \left\{ \mu \begin{pmatrix} 1 \\ 0 \end{pmatrix};\ \mu \in \mathbb{R} \right\}$.
  Algebraische Vielfachheit gleich 2, geometrische Vielfachheit gleich 1.

- $A = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix}$,
  $\lambda_1 = \cos(\varphi) + i\sin(\varphi), \quad \lambda_2 = \cos(\varphi) - i\sin(\varphi),$
  also im allgemeinen keine reellen Eigenwerte.
  $V(\lambda_1) = \left\{ \mu \begin{pmatrix} 1 \\ -i \end{pmatrix}, \quad \mu \in \mathbb{C} \right\}, \quad V(\lambda_2) = \left\{ \mu \begin{pmatrix} 1 \\ i \end{pmatrix}, \quad \mu \in \mathbb{C} \right\}.$

Im folgenden Satz stellen wir nützliche Beobachtungen zusammen:

Satz 2.55 (Eigenschaften von Eigenwerten und Eigenvektoren)
Sei $A \in \mathbb{C}^{n\times n}$:

(a) A und $A^\top$ besitzen dieselben Eigenwerte, jedoch im allgemeinen verschiedene Eigenräume.

(b) Ähnliche Matrizen A und $B^{-1}AB$ haben dasselbe charakteristische Polynom und deshalb dieselben Eigenwerte. b ist genau dann Eigenvektor von A zum Eigenwert λ, falls $B^{-1}b$ Eigenvektor von $B^{-1}AB$ zum Eigenwert λ ist.

(c) A ist genau dann invertierbar, falls alle Eigenwerte ungleich Null sind.
Ist λ ein Eigenwert von A mit Eigenvektor b (A sei invertierbar), so ist $\frac{1}{\lambda}$ Eigenwert von A^{-1} mit Eigenvektor b.

(d) Eigenvektoren $b_1, \ldots, b_r$ zu paarweise verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_r$ sind linear unabhängig.

(e) Besitzt A n linear unabhängige Eigenvektoren $b_1, \ldots, b_n$ zu den nicht notwendig verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_n$, so gilt mit $B := (b_1, \ldots, b_n)$:

$$B^{-1}AB = \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix}. \tag{2.183}$$

Sei nun $f : \mathbb{R}^n \to \mathbb{R}^n$ eine lineare Abbildung mit der Abbildungsmatrix A bezüglich der natürlichen Basis. Besitzt A n linear unabhängige Eigenvektoren $\{b_1, \ldots, b_n\}$ zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$, so kann f auch durch die Abbildungsmatrix

$$C = \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix} \tag{2.184}$$

bezüglich der Basis $\{b_1, \ldots, b_n\}$ dargestellt werden. Die Matrix C läßt die Eigenschaften von f im allgemeinen besser erkennen als die Matrix A.

Teil II

Mathematik 2

Kapitel 3

Analysis einer reellen Veränderlichen

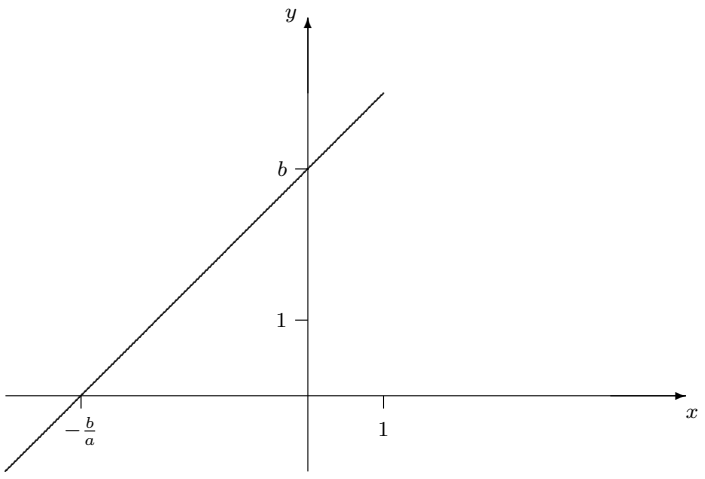
3.1 Funktionen, Grenzwerte, Stetigkeit

Zunächst betrachten wir reelle Funktionen einer reellen Veränderlichen

$$f : \mathbb{D} \to \mathbb{R}, \quad x \mapsto f(x) \tag{3.1}$$

mit der Variablen $x \in \mathbb{D} \subseteq \mathbb{R}$. Jede Funktion dieser Art wird durch einen Graph (die Kurve) $y = f(x)$ veranschaulicht, der aus denjenigen Punkten (x, y) einer mit kartesischen (x, y)-Koordinaten versehenen Ebene besteht, für die $x \in \mathbb{D}$ und $y = f(x)$ gilt.

Affin-lineare Funktionen
$f : \mathbb{R} \to \mathbb{R}$, $x \mapsto ax + b$ mit Konstanten $a, b \in \mathbb{R}$. Die Funktion heißt affin-lineare Funktion. Der Graph dieser Funktion stellt eine Gerade durch die Punkte $(0, b)$ und $(-\frac{b}{a}, 0)$ dar, falls $a \neq 0$. Im Fall $a = 0$ ist f eine konstante Funktion $x \mapsto b$. Ihr Graph stellt die zur x-Achse parallele Gerade $y = b$ dar.

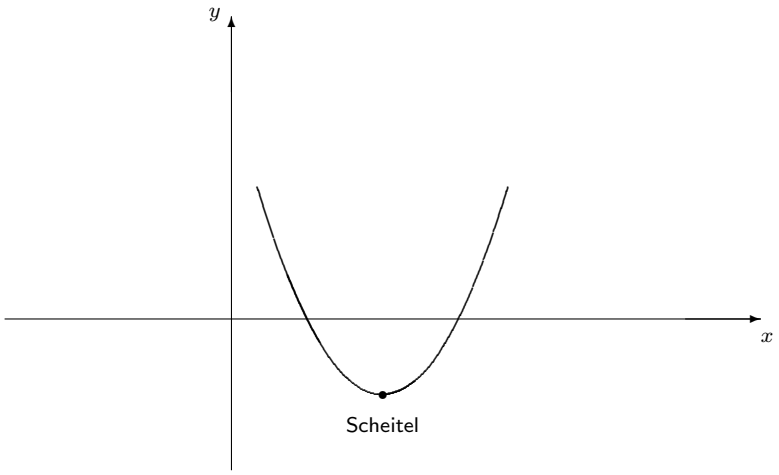


(3.2)

Quadratische Funktionen
$f : \mathbb{R} \to \mathbb{R},\ x \mapsto ax^2 + bx + c$ mit Konstanten $a, b, c \in \mathbb{R}$. Der Graph einer quadratischen Funktion stellt eine Parabel dar, deren Scheitel mittels quadratischer Ergänzung bestimmt wird $(a \neq 0)$:

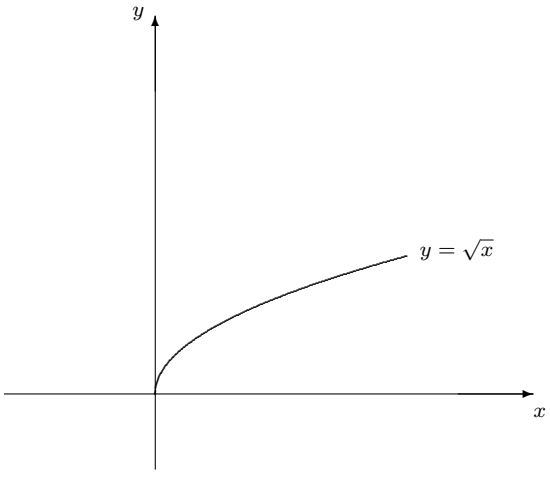
$$\begin{aligned} y &= ax^2 + bx + c \\ &= a\left(x^2 + \frac{b}{a}x\right) + c \\ &= a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2}\right) + c \\ &= a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c \end{aligned}$$

Es handelt sich somit um eine verschobene Parabel $y = ax^2$, deren Scheitel im Punkt $\left(-\frac{b}{2a}, c - \frac{b^2}{4a}\right)$ liegt.



(3.3)

Die Wurzelfunktion
$f : \mathbb{R}_0^+ \to \mathbb{R}_0^+,\ x \mapsto \sqrt{x}$



(3.4)

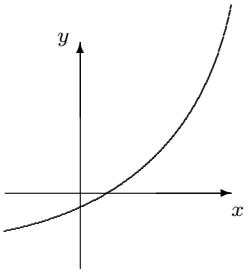
wobei $\mathbb{R}_0^+ := \{x \in \mathbb{R};\ x \geq 0\}$.

Ein für die genauere Analyse von Funktionen wichtiger Begriff ist die Monotonie.

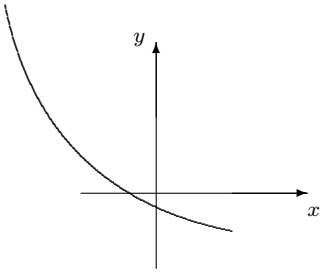
Definition 3.1 (Monotonie)
Eine Funktion $f : \mathbb{D} \to \mathbb{R}$, $x \mapsto f(x)$ heißt

(a) monoton wachsend (bzw. monoton fallend), falls für alle x_1, $x_2 \in \mathbb{D}$ mit $x_2 > x_1$ gilt: $f(x_2) \geq f(x_1)$ (bzw. $f(x_2) \leq f(x_1)$)

(b) streng monoton wachsend (bzw. streng monoton fallend), falls für alle x_1, $x_2 \in \mathbb{D}$ mit $x_2 > x_1$ gilt: $f(x_2) > f(x_1)$ (bzw. $f(x_2) < f(x_1)$).



streng monoton wachsend



streng monoton fallend (3.5)

Zu Funktionen $f : \mathbb{D} \to \mathbb{R}$, $g : \mathbb{D} \to \mathbb{R}$ mit gleichem Definitionsbereich definiert man die Summe, Differenz und das Produkt folgendermaßen:

$$f + g : \mathbb{D} \to \mathbb{R}, \qquad x \mapsto f(x) + g(x) \tag{3.6}$$

$$f - g : \mathbb{D} \to \mathbb{R}, \qquad x \mapsto f(x) - g(x) \tag{3.7}$$

$$f \cdot g : \mathbb{D} \to \mathbb{R}, \qquad x \mapsto f(x) \cdot g(x) \tag{3.8}$$

Der Quotient $\frac{f}{g} : \mathbb{D} \to \mathbb{R}$, $x \mapsto \frac{f(x)}{g(x)}$ ist nur erklärt, falls $g(x) \neq 0$ für alle $x \in \mathbb{D}$.
Mit dem Produkt sind auch die Potenzen $f^n : \mathbb{D} \to \mathbb{R}$, $x \mapsto (f(x))^n$, $n \in \mathbb{N}$, definiert. Das α-fache, $\alpha \in \mathbb{R}$, einer Funktion $f : \mathbb{D} \to \mathbb{R}$ ist definiert durch $\alpha f : \mathbb{D} \to \mathbb{R}$, $x \mapsto \alpha \cdot f(x)$.
Zu Funktionen $f : \mathbb{D} \to \mathbb{R}$ und $g : I \to \mathbb{R}$ mit $g(I) \subseteq \mathbb{D}$, wobei $g(I) := \{g(x) \in \mathbb{R}; x \in I\}$, kann man die Komposition $h = f \circ g$ definiert durch $h : I \to \mathbb{R}$, $x \mapsto f\,(g(x))$ bilden.
Eine Funktion f heißt Polynom vom Grad n, wenn es reelle Zahlen a_0, $a_1, \ldots, a_n$, $a_n \neq 0$, gibt mit

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n = \sum_{i=0}^{n} a_i x^i. \tag{3.9}$$

Die a_k heißen Koeffizienten des Polynoms. Das Nullpolynom $x \mapsto 0$ hat keinen Grad. Das Rechnen mit Polynomen in der Summendarstellung ist sehr übersichtlich

$$\sum_{k=0}^{n} a_k x^k \pm \sum_{k=0}^{n} b_k x^k = \sum_{k=0}^{n} (a_k \pm b_k) x^k. \tag{3.10}$$

Zwei Polynome $x \mapsto \sum\limits_{k=0}^{n} a_k x^k$ und $x \mapsto \sum\limits_{k=0}^{n} b_k x^k$ sind genau dann gleich, wenn $a_k = b_k$ für alle $k \in \{0, \ldots, n\}$ gilt.

Wird ein Funktionswert eines Polynoms $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto \sum\limits_{k=0}^{n} a_k x^k$ aus der Summendarstellung berechnet, so benötigt man dafür $2n - 1$ Multiplikationen und n Additionen. Berechnet man die Werte jedoch aus der Horner-Darstellung

$$x \mapsto (\ldots((a_n x + a_{n-1}) \cdot x + a_{n-2}) \cdot x + \ldots + a_1)\, x + a_0 \tag{3.11}$$

- von innen nach außen - so wird der Aufwand auf n Multiplikationen und n Additionen reduziert.
Als Nullstelle einer Funktion $f : \mathbb{D} \to \mathbb{R}$ bezeichnet man jede Lösung der Gleichung $f(x) = 0$ in $\mathbb{D}$. In praktischen Fällen betrachtet man oft die Nullstellen von Polynomen.
Eine Zahl $b \in \mathbb{R}$ ist genau dann Nullstelle eines Polynoms f, wenn es ein Polynom $h : \mathbb{R} \to \mathbb{R}$ gibt mit $f(x) = (x - b) \cdot h(x)$ für alle $x \in \mathbb{R}$.
Der Faktor $(x - b)$ kann auch mehrfach vorkommen, nämlich dann, wenn auch $h(b) = 0$ und somit $h(x) = (x - b) \cdot h_1(x)$ gilt. Man nennt b l-fache Nullstelle von f und l die Vielfachheit von b, wenn es ein Polynom $g : \mathbb{R} \to \mathbb{R}$ gibt mit

$$f(x) = (x - b)^l \cdot g(x) \quad \text{und} \quad g(b) \neq 0. \tag{3.12}$$

Jede weitere Nullstelle $c \neq b$ von f ist auch Nullstelle von g.

Beispiel(e) 3.2

$$x^5 - 5x^4 + 14x^3 - 22x^2 + 17x - 5 = (x - 1)^3(x^2 - 2x + 5); \tag{3.13}$$

$x = 1$ ist dreifache Nullstelle des entsprechenden Polynoms.

Sind nun $b_1, \ldots, b_r$ die verschiedenen reellen Nullstellen des Polynoms

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \sum_{k=0}^{n} a_k x^k \tag{3.14}$$

mit Grad $f \geq 1$, der jeweiligen Vielfachheit $l_1, \ldots, l_r$, so gibt es ein Polynom $q : \mathbb{R} \to \mathbb{R}$ vom Grad $(n - l_1 - l_2 - \ldots - l_r)$ mit

$$f(x) = (x - b_1)^{l_1} \cdot \ldots \cdot (x - b_r)^{l_r} \cdot q(x) \quad \text{für alle } x \in \mathbb{R}. \tag{3.15}$$

Somit hat jedes Polynom vom Grad $n \geq 1$ höchstens n Nullstellen.
Der Quotient

$$\frac{p}{q} : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \frac{\sum_{k=0}^{n} a_k x^k}{\sum_{k=0}^{m} b_k x^k} \quad (\text{mit } a_n \neq 0,\ b_m \neq 0) \tag{3.16}$$

heißt rationale Funktion. Der Definitionsbereich $\mathbb{D}$ von $\frac{p}{q}$ darf keine Nullstelle des Nennerpolynoms q enthalten. Ist der Grad des Zählerpolynoms p größer oder gleich dem Grad des Nennerpolynoms q (also $n \geq m$), so erhält man mit Hilfe des Polynoms $p_1 : \mathbb{R} \to \mathbb{R}$, $x \mapsto p(x) - \frac{a_n}{b_m} x^{n-m} \cdot q(x)$ durch Einsetzen die folgende Darstellung:

$$\frac{p}{q} : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \frac{a_n}{b_m} x^{n-m} + \frac{p_1(x)}{q(x)}, \tag{3.17}$$

wobei p_1 entweder die Nullfunktion ist oder ein Polynom mit kleinerem Grad als dem Grad von p. Sei nun $p_1 = c_s x^s + \ldots + c_1 x^1 + c_0$ mit $c_s \neq 0$ und $s \geq$ Grad q, so können wir die obige Vorgehensweise für die rationale Funktion $\frac{p_1}{q}$ wiederholen und erhalten

$$\frac{p}{q} : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \frac{a_n}{b_m} x^{n-m} + \frac{c_s}{b_m} x^{s-m} + \frac{p_2(x)}{q(x)} \tag{3.18}$$

mit $p_2 : \mathbb{R} \to \mathbb{R}$, $x \mapsto p_1(x) - \frac{c_s}{b_m} x^{s-m} q(x)$. Ist nun erneut Grad $p_2 \geq$ Grad q, so wiederholt man die Vorgehensweise für $\frac{p_2}{q}$, ansonsten ist das Verfahren beendet und man erhält die Existenz zweier Polynome g, $r : \mathbb{R} \to \mathbb{R}$ mit

$$\frac{p}{q} : \mathbb{D} \to \mathbb{R}, \quad x \mapsto g(x) + \frac{r(x)}{q(x)} \quad \text{mit Grad } r < \text{Grad } q. \tag{3.19}$$

Die eben skizzierte Vorgehensweise wird als Polynomdivision mit Rest bezeichnet.

Beispiel(e) 3.3

$$(x^{12}+x^6+x+1):(2x^4+3) = \underbrace{\frac{1}{2}x^8}_{\frac{a^n}{b^m}x^{n-m}} -\frac{3}{4}x^4+\frac{1}{2}x^2+\frac{9}{8}+\frac{-\frac{3}{2}x^2+x-\frac{19}{8}}{2x^4+3}$$

$$\begin{array}{ll} - & \left(x^{12}+\frac{3}{2}x^8\right) \\ & \hline -\frac{3}{2}x^8+x^6+x+1 \quad (=p_1(x)) \\ - & \left(-\frac{3}{2}x^8-\frac{9}{4}x^4\right) \\ & \hline x^6+\frac{9}{4}x^4+x+1 \quad (=p_2(x)) \\ - & \left(x^6+\frac{3}{2}x^2\right) \\ & \hline \frac{9}{4}x^4-\frac{3}{2}x^2+x+1 \quad (=p_3(x)) \\ - & \left(\frac{9}{4}x^4+\frac{27}{8}\right) \\ & \hline -\frac{3}{2}x^2+x-\frac{19}{8} \quad (=r(x)) \end{array}$$

Mit Hilfe der Polynomdivision kann bei der bereits betrachteten Zerlegung

$$f:\mathbb{R}\to\mathbb{R}, \quad x\mapsto (x-b)^l\cdot g(x) \quad \text{mit } g(b)\neq 0 \tag{3.20}$$

das Polynom g bei gegebenem Polynom f, bei gegebener Nullstelle b und gegebener Vielfachheit l berechnet werden.

Beispiel(e) 3.4

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto x^4 + 2x^3 - 3x^2 - 4x + 4, \quad b = 1, \quad l = 2$$

$$(x^4 + 2x^3 - 3x^2 - 4x + 4) : \underbrace{(x^2 - 2x + 1)}_{(x-1)^2} = x^2 + 4x + 4$$

$$\begin{array}{ll} & (x^4 + 2x^3 - 3x^2 - 4x + 4) \\ - & (x^4 - 2x^3 + x^2) \\ & - - - - - - - - - \\ & 4x^3 - 4x^2 - 4x + 4 \\ - & (4x^3 - 8x^2 + 4x) \\ & - - - - - - - - - \\ & 4x^2 - 8x + 4 \\ - & (4x^2 - 8x + 4) \\ & - - - - - - - - - \\ & - \end{array}$$

Also:

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto (x-1)^2 \underbrace{(x^2 + 4x + 4)}_{g(x)}.$$

Ebenso wie für die rationalen Zahlen ist es auch für rationale Funktionen wichtig, den gemeinsamen Teiler des Zählers und des Nenners zu kürzen, wobei ein Polynom $d : \mathbb{R} \to \mathbb{R}$ Teiler eines Polynoms $p : \mathbb{R} \to \mathbb{R}$ heißt, falls es ein Polynom $p_0 : \mathbb{R} \to \mathbb{R}$ gibt mit $p = d \cdot p_0$.
Die Berechnung eventuell vorhandener gemeinsamer Polynomteiler von Polynomen p und q basiert auf der folgenden Beobachtung:
Falls Grad $p \geq$ Grad q und $\frac{p}{q} = h + \frac{r}{q}$ mit Polynomen h,r und Grad r kleiner als Grad q, dann ist das Polynom d genau dann gemeinsamer Teiler von p und q, falls d gemeinsamer Teiler von r und q ist. Denn aus $p = d \cdot p_0$ und $q = d \cdot q_0$ folgt $r = p - h \cdot q = d(p_0 - h \cdot q_0)$.
Umgekehrt: Aus $q = d \cdot q_0$ und $r = d \cdot r_0$ folgt

$$p = h \cdot q + r = d(h \cdot q_0 + r_0). \tag{3.21}$$

Mittels Polynomdivision wird also die Bestimmung der gemeinsamen Teiler von p und q auf die Bestimmung der gemeinsamen Teiler der Polynome kleineren Grades q und r reduziert. Dieser Reduktionsschritt muss gegebenfalls mehrfach wiederholt werden (Euklidischer Algorithmus für Polynome)

Beispiel(e) 3.5
Wir suchen den gemeinsamen Teiler von

$$p : \mathbb{R} \to \mathbb{R}, \quad x \mapsto x^4 - 2x^3 - 2x^2 - 2x - 3$$

und

$$q : \mathbb{R} \to \mathbb{R}, \quad x \mapsto x^4 - 3x^3 - 7x^2 + 15x + 18 :$$

$p : q$:
$(x^4 - 2x^3 - 2x^2 - 2x - 3) : (x^4 - 3x^3 - 7x^2 + 15x + 18) = 1 + \frac{x^3+5x^2-17x-21}{q(x)}$
- $(x^4 - 3x^3 - 7x^2 + 15x + 18)$

$x^3 + 5x^2 - 17x - 21 \qquad = r(x)$

$q : r$:
$(x^4 - 3x^3 - 7x^2 + 15x + 18) : (x^3 + 5x^2 - 17x - 21) = x - 8 + \frac{50x^2-100x-150}{r(x)}$
- $(x^4 + 5x^3 - 17x^2 - 21x)$

$-8x^3 + 10x^2 + 36x + 18$
- $(-8x^3 - 40x^2 + 136x + 168)$

$50x^2 - 100x - 150 \qquad = r_1(x)$

$r : r_1$:
$(x^3 + 5x^2 - 17x - 21) : (50x^2 - 100x - 150) = \frac{1}{50}x + \frac{7}{50}$
- $(x^3 - 2x^2 - 3x)$

$7x^2 - 14x - 21$
- $(7x^2 - 14x - 21)$

$-$

Somit ist $r_1 : \mathbb{R} \to \mathbb{R}$, $x \mapsto 50x^2 - 100x - 150$ der dem Grad nach größte gemeinsame Teiler von r und r_1 und somit von r und q und damit auch von p und q. Es gilt:

$$p : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{50}(x^2 + 1)(50x^2 - 100x - 150) \tag{3.22}$$

und

$$q : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{50}(x + 2)(x - 3)(50x^2 - 100x - 150). \tag{3.23}$$

Polynome werden häufig zur Interpolation von Punkten benötigt. Ausgehend von Messpunkten $(x_i, y_i) \in \mathbb{R}^2$, $i = 0, \ldots, n$ mit $x_i \neq x_j$ für $i \neq j$ sucht man eine Funktion $f : \mathbb{R} \to \mathbb{R}$ mit $f(x_i) = y_i$ für alle $i = 0, \ldots, n$. Mit Hilfe dieser Funktion kann dann auch einem Argument $\hat{x} \neq x_i$, $i = 0, \ldots, n$ ein Wert $\hat{y} = f(\hat{x})$ zugeordnet werden.
Da es zu $n + 1$ Stützpunkten $(x_0, y_0), \ldots, (x_n, y_n)$ genau ein Polynom $p_n : \mathbb{R} \to \mathbb{R}$ vom Grad kleiner oder gleich n mit $p(x_i) = y_i$ gibt, verwendet man häufig dieses Polynom für die gesuchte Funktion f. Betrachtet man den Ansatz

$$p_n : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \sum_{k=0}^{n} a_k x^k, \tag{3.24}$$

so können die gesuchten Koeffizienten $a_0, \dots a_n$ durch das lineare Gleichungssystem in $a = (a_0, \dots, a_n)$:

$$\begin{aligned} \sum_{k=0}^{n} a_k x_0^k &= y_0 \\ &\vdots \\ \sum_{k=0}^{n} a_k x_n^k &= y_n \end{aligned} \tag{3.25}$$

beziehungsweise

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \cdot a = \begin{pmatrix} y_0 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} \tag{3.26}$$

ausgerechnet werden. Wird ein zusätzlicher Stützpunkt (x_{n+1}, y_{n+1}) hinzugefügt, so muss für die Berechnung von p_{n+1} komplett neu gerechnet werden.

Das Polynom p_n kann einfacher berechnet werden, wenn man es in der Form

$$p_n : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \alpha_0 + \alpha_1(x - x_0) + \alpha_2(x - x_0)(x - x_1) + \dots + \alpha_n(x - x_0) \cdot \dots \cdot (x - x_{n-1}) \tag{3.27}$$

darstellt. Aus der Forderung $p_n(x_i) = y_i$, $i = 0, \dots, n$, ergibt sich für die Unbekannten $\alpha_0, \alpha_1, \dots, \alpha_n$:

$$\begin{aligned} y_0 &= \alpha_0 \\ y_1 &= \alpha_0 + \alpha_1(x_1 - x_0) \\ &\vdots \\ y_n &= \alpha_0 + \alpha_1(x_n - x_0) + \dots + \alpha_n(x_n - x_0) \cdot \dots \cdot (x_n - x_{n-1}) \end{aligned} \tag{3.28}$$

beziehungsweise

$$\begin{pmatrix} 1 & 0 & 0 & & \\ 1 & (x_1 - x_0) & 0 & & \\ 1 & (x_2 - x_0) & (x_2 - x_0)(x_2 - x_1) & & \mathbf{0} \\ & & & \ddots & \\ 1 & (x_n - x_0) & \dots & \dots & (x_n - x_0) \cdot \dots \cdot (x_n - x_{n-1}) \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \vdots \\ \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ \\ \vdots \\ y_n \end{pmatrix}. \tag{3.29}$$

Dieses Gleichungssystem kann nun „von oben nach unten“ gelöst werden.
Die Idee dieser Vorgehensweise stammt von Isaac Newton (1642-1727) und hat den großen Vorteil, dass problemlos weitere Stützpunkte (x_{n+1}, y_{n+1}), …, (x_{n+m}, y_{n+m}) in das obige Schema integriert werden können, ohne die Dreiecksform des linearen Gleichungssystems zu verlieren.

Eine der wichtigsten Vorgehensweisen der Analysis ist die Bildung von Grenzwerten. Dabei betrachtet man zunächst spezielle Funktionen $f : \mathbb{N}_0 \to \mathbb{R}$, die auch als Zahlenfolgen bezeichnet werden. Zahlenfolgen werden oft durch die Notation der Funktionswerte $a_0 = f(0)$, $a_1 = f(1)$, $a_2 = f(2)$ … in der Form $\{a_n\}, n \in \mathbb{N}_0$, notiert. Die Zahlen a_n heißen Glieder der Zahlenfolge. Das erste Glied einer Zahlenfolge muss nicht immer a_0 sein. Durch Umbenennung, etwa $b_0 := a_5$, $b_1 := a_6$, $b_n := a_{n+5}$ erreicht man, dass auch a_5, a_6, a_7, … eine Zahlenfolge im obigen Sinne darstellt. Man bezeichnet sie mit $\{a_n\}, n \geq 5$, (allgemein: $\{a_n\}, n \geq n_0$).

Beispiel(e) 3.6

- $a_n = a_0 + n \cdot d$, $d \in \mathbb{R}$ fest, $n \in \mathbb{N}_0$ (arithmetische Folge)
- $a_n = a_0 \cdot q^n$, $q \neq 0$ fest, $n \in \mathbb{N}_0$ (geometrische Folge)
- $a_{n+1} = \frac{1}{2}(a_n + \frac{2}{a_n})$, $n \geq 1$, $a_0 := 2$ (rekursiv definierte Zahlenfolge).

Eine Zahlenfolge heißt beschränkt, falls es Konstanten K_1 und K_2 gibt mit

$$K_1 \leq a_n \leq K_2 \quad \text{für alle } n \in \mathbb{N}_0. \tag{3.30}$$

Die Konstanten K_1 bzw. K_2 heißen untere bzw. obere Schranken der Zahlenfolge.
Für die Analyse von Zahlenfolgen sind die beiden folgenden Begriffe wichtig:

Definition 3.7 (Häufungspunkt und Grenzwert einer Zahlenfolge)
Sei $\{a_n\}, n \in \mathbb{N}_0$, eine Zahlenfolge. Eine Zahl $h \in \mathbb{R}$ heißt Häufungspunkt der Zahlenfolge $\{a_n\}$, falls für jedes offene Intervall (α, β) mit $h \in (\alpha, \beta)$ unendlich viele Folgeglieder der Zahlenfolge in (α, β) liegen.
Eine Zahl $a \in \mathbb{R}$ heißt Grenzwert der Zahlenfolge $\{a_n\}$, falls es zu jedem reellen $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$ gibt mit

$$|a_n - a| < \varepsilon \quad \text{für alle } n \geq n_0. \tag{3.31}$$

Offensichtlich ist jeder Grenzwert a einer Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, ein Häufungspunkt. Die Umkehrung gilt nicht, wie das folgende Beispiel zeigt:

$$a_n = \begin{cases} \frac{1}{n} & \text{für } n \geq 1 \text{ und } n \text{ gerade} \\ 1 & \text{für } n = 0 \\ 2 - \frac{1}{n} & \text{für } n \geq 1 \text{ und } n \text{ ungerade} \end{cases}. \tag{3.32}$$

Diese Zahlenfolge besitzt die beiden Häufungspunkte $h_1 = 0$ und $h_2 = 2$, aber keinen Grenzwert.
Besitzt eine Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, einen Grenzwert a, so schreibt man dafür

$$\lim_{n \to \infty} a_n = a \quad \text{oder} \quad a_n \to a \quad \text{für} \quad n \to \infty. \tag{3.33}$$

und man sagt, die Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, konvergiert gegen a.
Anschaulich bedeutet die Existenz des Grenzwertes a einer Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, dass zu jedem $\varepsilon > 0$ nur endlich viele Folgenglieder außerhalb des Intervalls $(a - \varepsilon, a + \varepsilon)$ liegen dürfen.
Eine gegen Null konvergierende (man sagt auch konvergente) Zahlenfolge heißt Nullfolge. Nicht konvergente Zahlenfolgen heißen divergent. Konvergiert eine Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, gegen a, so ist die Zahlenfolge $\{a_n - a\}, n \in \mathbb{N}_0$, eine Nullfolge.
Eine Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, heißt „bestimmt divergent gegen ∞ (bzw. $-\infty$)“, falls es zu jedem $K \in \mathbb{R}$ ein $n_0 \in \mathbb{N}$ gibt mit:

$$a_n > K \quad (\text{bzw. } a_n < K) \quad \text{für alle } n \geq n_0. \tag{3.34}$$

Man schreibt dafür

$$\lim_{n \to \infty} a_n = \infty \quad (\text{bzw. } \lim_{n \to \infty} a_n = -\infty). \tag{3.35}$$

Beispiel(e) 3.8

- Die Zahlenfolge $a_n = (-1)^n$, $n \in \mathbb{N}_0$, ist divergent.
- Die Zahlenfolge $a_n = n$, $n \in \mathbb{N}_0$, ist bestimmt divergent gegen ∞
- Die Zahlenfolge $a_n = -n$, $n \in \mathbb{N}_0$, ist bestimmt divergent gegen $-\infty$

Offensichtlich sind Zahlenfolgen, die mindestens zwei Häufungspunkte haben, unbestimmt divergent.

Beispiel(e) 3.9

- Sei $a_n = \sum_{k=0}^{n} q^k$, $q \neq 0$ (geometrische Reihe), so ist $\{a_n\}$ konvergent, falls $|q| < 1$ mit dem Grenzwert $a = \frac{1}{1-q}$, unbestimmt divergent für $q \leq -1$ und bestimmt divergent gegen ∞ für $q \geq 1$. Ist $q \neq 1$, so folgt aus der Gleichung

$$(1-q) \cdot a_n = (1-q) \sum_{k=0}^{n} q^k = \sum_{k=0}^{n} q^k - \sum_{k=0}^{n} q^{k+1} = 1 - q^{n+1} \quad \text{für alle } n \in \mathbb{N}_0\text{:}$$

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

Für $|q| < 1$ erhalten wir den Grenzwert $a = \frac{1}{1-q}$, denn für jedes $\varepsilon > 0$ gilt:

$$\begin{aligned} |a_n - a| < \varepsilon &\iff \left| \frac{1-q^{n+1}}{1-q} - \frac{1}{1-q} \right| < \varepsilon \\ &\iff \left| \frac{q^{n+1}}{1-q} \right| < \varepsilon \\ &\iff |q|^{n+1} < (1-q)\varepsilon \end{aligned}$$

Da $|q| < 1$, ist $|q|^{n+1} < (1-q)\varepsilon$ ab einem festen n_0 immer erfüllt.
Für $q > 1$ ist $\frac{1-q^{n+1}}{1-q} = \frac{q^{n+1}-1}{q-1} \to \infty$. Für $q = 1$ ist $a_n = n+1$. Somit ist für $q \geq 1$ die Zahlenfolge $\{a_n\}$ bestimmt divergent gegen ∞. Für $q \leq -1$ gilt

$$a_n = \frac{1-q^{n+1}}{1-q} = \begin{cases} \frac{1-|q|^{n+1}}{1-q} \leq 0 & \text{für } n \text{ ungerade} \\ \\ \frac{1+|q|^{n+1}}{1-q} \geq 1 & \text{für } n \text{ gerade.} \end{cases}$$

Daher ist $\{a_n\}$ für $q \leq -1$ unbestimmt divergent.

- Sei $a_n = \frac{n+2}{n+1}$, $n \in \mathbb{N}_0$, so konvergiert $\{a_n\}$ gegen $a = 1$, denn für jedes $\varepsilon > 0$ gilt:

$$\begin{aligned} |a_n - a| < \varepsilon &\iff \left| \frac{n+2}{n+1} - 1 \right| < \varepsilon \\ &\iff \left| \frac{n+1}{n+1} + \frac{1}{n+1} - 1 \right| < \varepsilon \\ &\iff \frac{1}{n+1} < \varepsilon \\ &\iff n > \frac{1}{\varepsilon} - 1 \end{aligned}$$

Im Folgenden betrachten wir Rechenregeln für Grenzwerte und Konvergenzkriterien. Aus gegebenen Zahlenfolgen $\{a_n\}, n \in \mathbb{N}_0$, und $\{b_n\}, n \in \mathbb{N}_0$, werden durch Addition, Subtraktion, Multiplikation und Division neue Zahlenfolgen $c_n := a_n + b_n$, $d_n := a_n - b_n$, $e_n := a_n \cdot b_n$, $f_n := \frac{a_n}{b_n}$ (falls $b_n \neq 0$ für alle $n \in \mathbb{N}_0$) gewonnen. Ferner können neue Zahlenfolgen dadurch konstruiert werden, dass mit einer gegebenen Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, und Indizes $n_0 < n_1 < n_2 < n_3 < \ldots$ die Zahlenfolge $\{a_{n_k}\}, k \in \mathbb{N}_0$, betrachtet wird. Man nennt $\{a_{n_k}\}, k \in \mathbb{N}_0$, Teilfolge von $\{a_n\}, n \in \mathbb{N}_0$.

Satz 3.10 (Grenzwertregeln)
Sind $\{a_n\}, n \in \mathbb{N}_0$, und $\{b_n\}, n \in \mathbb{N}_0$, konvergente Folgen mit $\lim\limits_{n\to\infty} a_n = a$ und $\lim\limits_{n\to\infty} b_n = b$, dann gilt:

(a) $\lim\limits_{n\to\infty} (a_n \pm b_n) = a \pm b$,

(b) $\lim\limits_{n\to\infty} (a_n \cdot b_n) = ab$, insbesondere $\lim\limits_{n\to\infty} (c \cdot a_n) = c \cdot a$ für alle $c \in \mathbb{R}$,

(c) ist $a \neq 0$, dann gibt es ein $n_1 \in \mathbb{N}$ mit $a_n \neq 0$ für alle $n > n_1$ und für die Zahlenfolgen $\{a_n\}, n \geq n_1$, und $\{b_n\}, n \geq n_1$, gilt:

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

(d) $\lim\limits_{n\to\infty} |a_n| = |a|$,

(e) $\lim\limits_{n\to\infty} \sqrt{a_n} = \sqrt{a}$ falls alle $a_n \geq 0$.

Beispiel(e) 3.11

$$\lim_{n\to\infty} \sqrt{\frac{6n^4 + 3n^2 + 2}{7n^4 + 12n^3 + 6}} = \lim_{n\to\infty} \sqrt{\frac{6 + \frac{3}{n^2} + \frac{2}{n^4}}{7 + \frac{12}{n} + \frac{6}{n^4}}} = \sqrt{\frac{6}{7}} \tag{3.36}$$

Aus der Existenz von $\lim\limits_{n\to\infty} (a_n \pm b_n)$ (bzw. $\lim\limits_{n\to\infty} |a_n|$) folgt im allgemeinen nicht die Konvergenz der Folgen $\{a_n\}$ oder $\{b_n\}$ (bzw. $\{a_n\}$). Die Limesbildung erhält schwache Ungleichungen, aber keine strikten Ungleichungen: Seien $\{a_n\}, n \in \mathbb{N}_0$, und $\{b_n\}, n \in \mathbb{N}_0$ konvergente Zahlenfolgen mit $a_n \leq b_n$ für alle $n \geq n_1 \in \mathbb{N}$, so gilt $\lim\limits_{n\to\infty} a_n \leq \lim\limits_{n\to\infty} b_n$. Für strikte Ungleichungen gilt dies nicht, wie das Beispiel $a_n = 0$, $b_n = \frac{1}{1+n}$, $n \in \mathbb{N}_0$, zeigt, denn $a_n < b_n$ für alle $n \in \mathbb{N}_0$, aber $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} b_n$.
Da Zahlenfolgen spezielle Funktionen sind, heißt eine Zahlenfolge $\{a_n\}, n \in \mathbb{N}_0$, genau dann monoton wachsend (bzw. monoton fallend), falls $a_{n+1} \geq a_n$ (bzw. $a_{n+1} \leq a_n$) für alle $n \in \mathbb{N}_0$.

Satz 3.12 (Monotoniekriterium)
Jede monoton wachsende oder monoton fallende beschränkte Zahlenfolge ist konvergent.

Beispiel(e) 3.13

- Die Zahlenfolge $a_n := \sum_{k=0}^{n} \frac{1}{k!},\ n \geq 0$, ist monoton wachsend. Da

$$0 < \frac{1}{k!} = \frac{1}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot k} \leq \frac{1}{1 \cdot 2 \cdot 2 \cdot \ldots \cdot 2} = \frac{1}{2^{k-1}}, \tag{3.37}$$

ergibt sich für $n \geq 2$ die Abschätzung

$$2 \leq 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{1}{n!} \leq 1 + 1 + \frac{1}{2} + \ldots + \frac{1}{2^{n-1}} \leq 1 + \lim_{n\to\infty} \sum_{k=0}^{n} \left(\frac{1}{2}\right)^k = 3. \tag{3.38}$$

Somit ist $\{a_n\}, n \in \mathbb{N}_0$, auch beschränkt und damit konvergent.
Durch den Grenzwert dieser Folge ist die Eulersche Zahl e definiert:

$$e := \sum_{k=0}^{\infty} \frac{1}{k!} := \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!}, \qquad e = 2.7182818284 \ldots . \tag{3.39}$$

Die Eulersche Zahl e kann auch durch den Grenzwert der Folge

$$c_n := \left(1 + \frac{1}{n}\right)^n, n \geq 1, \tag{3.40}$$

dargestellt werden.

- Die rekursiv definierte Zahlenfolge $a_0 := 1\ a_{n+1} = \frac{6(1+a_n)}{7+a_n}$ ist beschränkt $(0 < a_n < 2)$ und monoton wachsend (vollständige Induktion).
Der noch unbekannte Grenzwert sei a. Aus $a_n \to a$ und $a_{n+1} \to a$ und den Rechenregeln für Grenzwerte folgt:

$$a = \frac{6(1+a)}{7+a} \tag{3.41}$$

d.h., $a = 2$ oder $a = -3$. Da $a_n > 0$ für alle $n \in \mathbb{N}_0$, folgt $a = 2$.

Wie die Zahl e kann man nun weitere Zahlen über den Grenzwert von Zahlenfolgen definieren. Dazu betrachten wir für jedes $x \in \mathbb{R}$ die Zahl e^x definiert durch

$$e^x := \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n \tag{3.42}$$

Die Existenz dieses Grenzwertes für jedes x kann durch geeignete Fallunterscheidungen gezeigt werden. Der Zusammenhang zwischen e^x und der x-ten Potenz von e wird später geklärt.
Die Funktion $\exp : \mathbb{R} \to \mathbb{R}$, $x \mapsto e^x \left(:= \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n\right)$ heißt Exponentialfunktion.

Nach dem Grenzwert von Zahlenfolgen betrachten wir im Folgenden den Grenzwert reellwertiger Funktionen:

Definition 3.14 ((rechtsseitiger, linksseitiger) Grenzwert von Funktionen)
Sei I ein Intervall und $a \in I$. Sei ferner $f : I \setminus \{a\} \to \mathbb{R}$ oder $f : I \to \mathbb{R}$ eine gegebene Funktion, so hat f für x gegen a den rechtsseitigen Grenzwert (bzw. linksseitigen Grenzwert) c, in Zeichen

$$\lim_{x \to a+} f(x) = c \tag{3.43}$$

bzw.

$$\lim_{x \to a-} f(x) = c, \tag{3.44}$$

falls für jede Zahlenfolge $\{x_n\}, n \in \mathbb{N}_0$, aus I mit $x_n \to a$ und $x_n > a$ für alle $n \in \mathbb{N}_0$ (bzw. $x_n \to a$ und $x_n < a$ für alle $n \in \mathbb{N}_0$) die Zahlenfolge $\{f(x_n)\}, n \in \mathbb{N}_0$, gegen c konvergiert. Die Funktion f hat für x gegen a den Grenzwert c, in Zeichen $\lim\limits_{x \to a} f(x) = c$, wenn

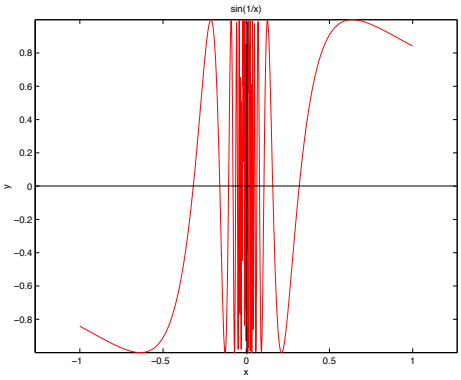
$$\lim_{x \to a+} f(x) = \lim_{x \to a-} f(x) = c. \tag{3.45}$$

Nur linksseitige bzw. rechtsseitige Grenzwerte lassen sich definieren, wenn man in der Definition $a = +\infty$ bzw. $a = -\infty$ mit entsprechender Funktion f zuläßt. In diesen Fällen schreibt man $\lim\limits_{x \to \infty} f(x) = c$ bzw. $\lim\limits_{x \to -\infty} f(x) = c$.

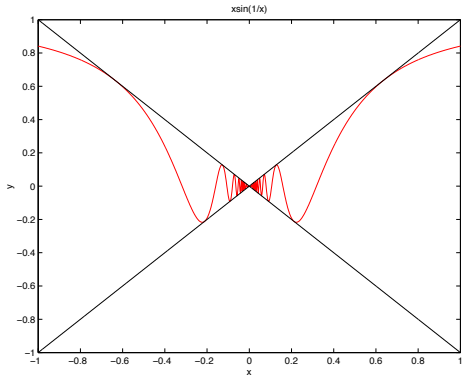
Der Begriff „bestimmt divergent“ wird gemäß obiger Definition für Funktionen anstelle von Zahlenfolgen entsprechend übernommen.

Beispiel(e) 3.15

- Die Funktion $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $x \mapsto \sin\left(\frac{1}{x}\right)$ ist für alle $x \in \mathbb{R} \setminus \{0\}$ erklärt. Sie hat weder einen linksseitigen noch einen rechtsseitigen Grenzwert für $x \to 0$. Dagegen besitzt die Funktion $g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $x \mapsto x \cdot \sin\left(\frac{1}{x}\right)$ den Grenzwert 0 für $x \to 0$.



$x \mapsto \sin\left(\frac{1}{x}\right)$



$x \mapsto x \sin\left(\frac{1}{x}\right)$

- $\lim\limits_{x \to 0+} \frac{1}{x} = \infty$, $\lim\limits_{x \to 0-} \frac{1}{x} = -\infty$
- $f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \begin{cases} x^2 & x > 0 \\ 1 & x = 0 \\ x^2 - 1 & x < 0 \end{cases}$

  $\lim\limits_{x \to 0+} f(x) = 0, \quad \lim\limits_{x \to 0-} f(x) = -1, \quad f(0) = 1$

Grenzwertregeln für Zahlenfolgen lassen sich auch auf Funktionen übertragen:

Satz 3.16 (Grenzwertregeln für Funktionen)
Seien f, g Funktionen gemäß Definition 3.14. Aus $\lim\limits_{x \to a} f(x) = c$ und $\lim\limits_{x \to a} g(x) = d$, $c, d \in \mathbb{R}$, folgt:

(i) $\lim\limits_{x \to a} (f(x) \pm g(x)) = c \pm d$,

(ii) $\lim\limits_{x \to a} (f(x) \cdot g(x)) = c \cdot d$,

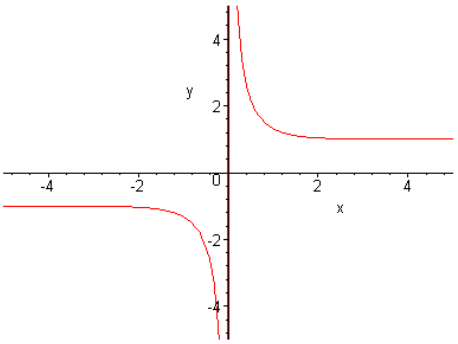
(iii) $\lim\limits_{x \to a} \frac{f(x)}{g(x)} = \frac{c}{d}$, falls $d \neq 0$.

Diese Regeln gelten auch für $a = \pm\infty$, aber nur für $c, d \in \mathbb{R}$.

Beispiel(e) 3.17

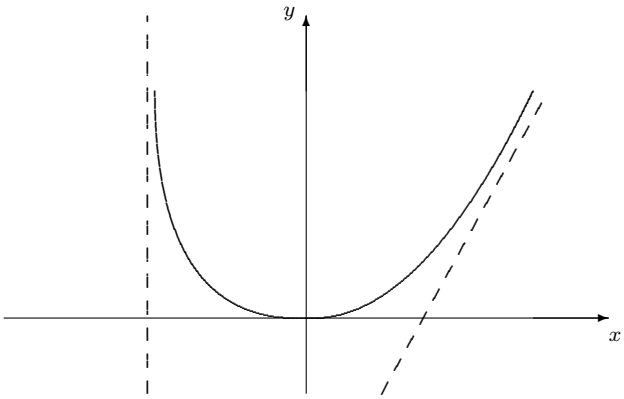
- $\lim\limits_{x\to 2} \frac{x^3+3x+5}{x^2-2x+1} = \frac{8+6+5}{4-4+1} = 19$
- $\lim\limits_{x\to 1} \frac{x^n-1}{x-1} = \lim\limits_{x\to 1}(x^{n-1}+x^{n-2}+\ldots+x+1) = n$
- $\lim\limits_{x\to 0} \frac{\sqrt{x+1}-1}{x} = \lim\limits_{x\to 0} \frac{(\sqrt{x+1}-1)(\sqrt{x+1}+1)}{x(\sqrt{x+1}+1)} = \lim\limits_{x\to 0} \frac{x+1-1}{x\left(\sqrt{x+1}+1\right)} = \frac{1}{2}$

Man nennt die Gerade $x = a$ eine vertikale Asymptote der Kurve $y = f(x)$ bei gegebenem f, wenn $\lim\limits_{x\to a-} f(x) = \pm\infty$ oder $\lim\limits_{x\to a+} f(x) = \pm\infty$. Die Gerade $y = c$ heißt horizontale Asymptote der Kurve $y = f(x)$ bei gegebenem f, falls $\lim\limits_{x\to\infty} f(x) = c$ oder $\lim\limits_{x\to-\infty} f(x) = c$.



(3.46)

Als schräge Asymptote der Kurve $y = f(x)$ bezeichnet man die Gerade $y = px + q$, $p \neq 0$, falls $(f(x) - p \cdot x - q) \to 0$ für $x \to \infty$ oder $x \to -\infty$.



(3.47)

Nun zeichnen wir Funktionen aus, bei denen an einer Stelle der rechtsseitige Grenzwert, der linksseitige Grenzwert und der Funktionswert übereinstimmen.

Definition 3.18 (Stetigkeit)
Sei I ein Intervall und $f : I \to \mathbb{R}$ eine gegebene Funktion, so heißt f in $x_0 \in I$ genau dann stetig, wenn

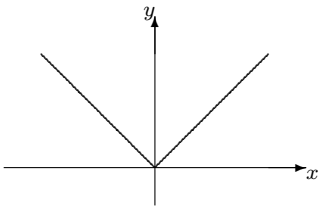
$$\lim_{x\to x_0} f(x) = f(x_0). \tag{3.48}$$

Ist x_0 Randpunkt von I, so ist $\lim\limits_{x\to x_0}$ als einseitiger Grenzwert ($\lim\limits_{x\to x_0+} f(x)$ oder $\lim\limits_{x\to x_0-} f(x)$) zu verstehen. Die Funktion f heißt auf I stetig, falls sie in allen $x_0 \in I$ stetig ist.

Die Stetigkeit von $f : I \to \mathbb{R}$ auf I bedeutet anschaulich, dass der Graph $y = f(x)$ über I eine zusammenhängende Linie (ohne Lücken und Sprünge, aber durchaus mit Spitzen) darstellt.

Beispiel(e) 3.19

- $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto |x|$ ist stetig



(3.49)

- $f : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}$, $x \mapsto \frac{1}{x}$ ist stetig

Aufgrund der Grenzwertregeln sind für stetige Funktionen $f, g : I \to \mathbb{R}$ auch $f \pm g$, $f \cdot g$ stetig. Ferner ist $\frac{f}{g}$ in allen $x \in I$ mit $g(x) \neq 0$ stetig. Seien I, J Intervalle und $f : I \to \mathbb{R}$ sowie $g : J \to \mathbb{R}$ mit $g(J) \subseteq I$ stetig, so ist auch die Funktion $f \circ g : J \to \mathbb{R}$, $x \mapsto f(g(x))$ stetig.
Offensichtlich sind alle Polynome auf $\mathbb{R}$ und alle rationalen Funktionen auf den Teilintervallen des entsprechenden Definitionsbereichs stetig.
Der folgende Satz fasst wichtige Eigenschaften zusammen:

Satz 3.20 (Eigenschaften auf einem abgeschlossenen Intervall stetiger Funktionen)
Sei $f : [a, b] \to \mathbb{R}$ eine auf $[a, b]$ stetige Funktion, so gilt:

a) Schrankensatz: Es gibt ein $K \in \mathbb{R}$ mit $|f(x)| < K$ für alle $x \in [a, b]$ (man sagt, f ist auf $[a, b]$ beschränkt).

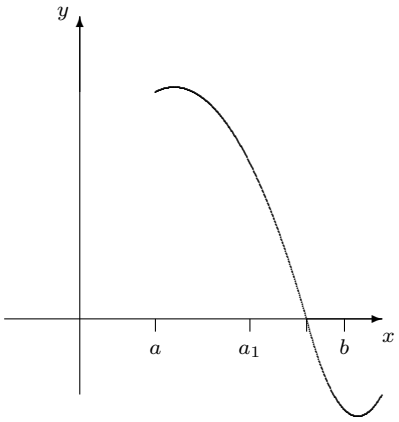
b) Satz von Minimum und Maximum: Es gibt Werte x_0, $x_1 \in [a, b]$ mit

$$f(x_0) \leq f(x) \leq f(x_1) \tag{3.50}$$

für alle $x \in [a, b]$.

c) Nullstellenexistenz: Ist $f(a) \cdot f(b) < 0$, so gibt es ein $\overline{x} \in (a, b)$ mit $f(\overline{x}) = 0$.

Ist $f(a) \cdot f(b) < 0$, so kann eine Stelle $\overline{x}$ mit $f(\overline{x}) = 0$ durch das Bisektionsverfahren berechnet werden:
Wähle $a_1 = \frac{a+b}{2}$. Ist $f(a_1) = 0$, so ist man fertig. Ist $f(a_1) \cdot f(b) < 0$, so ersetze a durch a_1 und wiederhole die Vorgehensweise mit dem Intervall $[a_1, b]$. Ist $f(a) \cdot f(a_1) < 0$, so ersetze b durch a_1 und wiederhole die Vorgehensweise mit $[a, a_1]$.



(3.51)

3.2 Differentiation

Seien $I \subseteq \mathbb{R}$ ein offenes Intervall und $f : I \to \mathbb{R}$ eine Funktion, so heißt f im Punkt $x_0 \in I$ differenzierbar, wenn der Differenzenquotient $\frac{f(x)-f(x_0)}{x-x_0}$ für $x \to x_0$ einen endlichen Grenzwert besitzt.
Dieser Grenzwert wird mit $f'(x_0)$ bezeichnet. Mit $h := x - x_0$ gilt also:

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{3.52}$$

Ist f in jedem Punkt $x_0 \in I$ differenzierbar, so sagt man, f ist auf I differenzierbar. In diesem Fall ist

$$f' : I \to \mathbb{R}, \quad x \mapsto f'(x) \tag{3.53}$$

eine Funktion, die man die Ableitung von f nennt. Den Übergang von f zu f' nennt man differenzieren oder ableiten. Für $f'(x)$ schreibt man auch $\frac{df}{dx}(x)$ oder $\frac{d}{dx}f(x)$.

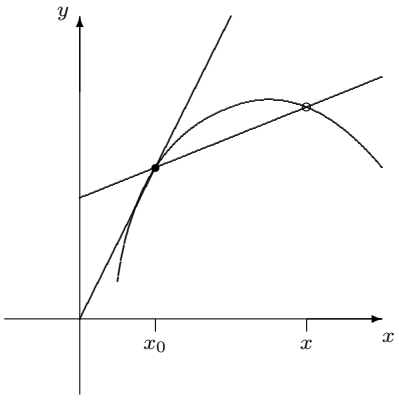
Beispiel(e) 3.21

- $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto c$, $f' : \mathbb{R} \to \mathbb{R}$, $x \mapsto 0$
- $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto ax + b$, $f' : \mathbb{R} \to \mathbb{R}$, $x \mapsto a$, denn

$$\lim_{x \to x_0} \frac{ax + b - (ax_0 + b)}{x - x_0} = \lim_{x \to x_0} \frac{a(x - x_0)}{x - x_0} = a$$

- $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^n$, $n \in \mathbb{N}$, $f' : \mathbb{R} \to \mathbb{R}$, $x \mapsto n \cdot x^{n-1}$
- $f : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}$, $x \mapsto \sqrt{x}$, $f' : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}$, $x \mapsto \frac{1}{2\sqrt{x}}$, denn $\lim_{x \to x_0} \frac{\sqrt{x}-\sqrt{x_0}}{x-x_0} = \lim_{x \to x_0} \frac{\sqrt{x}-\sqrt{x_0}}{(\sqrt{x}-\sqrt{x_0})(\sqrt{x}+\sqrt{x_0})} = \lim_{x \to x_0} \frac{1}{(\sqrt{x}+\sqrt{x_0})} = \frac{1}{2}\frac{1}{\sqrt{x_0}}$.

Betrachtet man bezüglich kartesischer (x, y)-Koordinaten die Kurve $g = f(x)$, so ist $\frac{f(x)-f(x_0)}{x-x_0}$ die Steigung der Geraden durch die Punkte $(x_0, f(x_0))$ und $(x; f(x))$



(3.54)

Der Grenzwert $f'(x_0) = \lim\limits_{x \to x_0} \frac{f(x)-f(x_0)}{x-x_0}$ gibt dann die Steigung der Kurventangente in $(x_0, f(x_0))$ an. Ist also f in x_0 differenzierbar, so ist

$$y = f'(x_0)(x - x_0) + f(x_0) \tag{3.55}$$

die Tangente an den Graphen $y = f(x)$ von f in $(x_0, f(x_0))$.
Da die Wurzelfunktion in $x_0 = 0$ eine senkrechte Tangente hat, existiert die Ableitung in $x_0 = 0$ nicht. Will man eine gegebene differenzierbare Funktion $f : I \to \mathbb{R}$ in der Nähe eines Punktes $x_0 \in I$ durch eine Gerade $g : \mathbb{R} \to \mathbb{R}$, $x \mapsto m \cdot (x - x_0) + f(x_0)$ (damit ist $g(x_0) = f(x_0)$) approximieren, so ist der Parameter m zu bestimmen. Fordert man, dass der relative Fehler $\frac{f(x)-g(x)}{x-x_0}$ für $x \to x_0$ gegen Null konvergiert, so ergibt sich

$$0 = \lim_{x \to x_0} \frac{f(x) - m(x - x_0) - f(x_0)}{x - x_0} = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} - m = f'(x_0) - m. \tag{3.56}$$

Somit ist $m = f'(x_0)$ zu wählen. Es gilt also in der Nähe von x_0:

$$f(x) \approx f'(x_0)(x - x_0) + f(x_0) \tag{3.57}$$

Der folgende Satz stellt eine Beziehung zwischen Differenzierbarkeit und Stetigkeit her:

Satz 3.22 (Differenzierbarkeit und Stetigkeit)
Seien $I \subseteq \mathbb{R}$ ein offenes Intervall und $f : I \to \mathbb{R}$ in $x_0 \in I$ differenzierbar, dann ist f in x_0 auch stetig.

Die Umkehrung dieses Satzes gilt nicht, wie das Beispiel $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto |x|$ an der Stelle $x = 0$ zeigt.

Mit den folgenden Regeln gelingt es, die Ableitung zusammengesetzter Funktionen aus den Ableitungen ihrer einzelnen „Bausteine“ zu ermitteln.

Satz 3.23 (Ableitung zusammengesetzter Funktionen)
Seien I ein offenes Intervall und $f, g : I \to \mathbb{R}$ Funktionen, die in $x \in I$ differenzierbar sind, dann gilt für alle $x \in I$:

$$(f(x) + g(x))' = f'(x) + g'(x) \tag{3.58}$$
$$(c \cdot f(x))' = c \cdot f'(x) \quad \text{für alle } c \in \mathbb{R} \tag{3.59}$$
$$(f(x) \cdot g(x))' = f'(x) \cdot g(x) + f(x) \cdot g'(x) \text{ (Produktregel)} \tag{3.60}$$
$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - g'(x) \cdot f(x)}{g(x)^2} \text{ falls } g(x) \neq 0 \text{ (Quotientenregel)} \tag{3.61}$$

speziell:

$$\left(\frac{1}{g(x)}\right)' = -\frac{g'(x)}{g(x)^2}. \tag{3.62}$$

Mit der Produktregel läßt sich nun die bereits erwähnte Formel $(x^n)' = nx^{n-1}$ für alle $n \in \mathbb{N}$, $x \in \mathbb{R}$, einfach mit vollständiger Induktion beweisen.
Jede rationale Funktion $f : \mathbb{D} \to \mathbb{R}$, $x \mapsto \frac{p}{q}$ kann nun mit der Differentiation für Polynome und der Quotientenregel differenziert werden. Insbesondere gilt für alle $x \in \mathbb{R} \setminus \{0\}$:

$$\left(\frac{1}{x^n}\right)' = -\frac{n}{x^{n+1}} = -n \cdot x^{-n-1}, \quad n \in \mathbb{N}. \tag{3.63}$$

Somit gilt die Formel $(x^n)' = nx^{n-1}$ auch für $n \in \mathbb{Z}, x \in \mathbb{R} \setminus \{0\}$.
Für die Sinus- und Cosinusfunktion sowie für die Tangens- und Cotangensfunktion

$$\tan : \mathbb{R} \setminus \{x \in \mathbb{R}; x = (2k+1)\frac{\pi}{2},\ k \in \mathbb{Z}\} \to \mathbb{R},\ x \mapsto \frac{\sin(x)}{\cos(x)} \tag{3.64}$$
$$\cot : \mathbb{R} \setminus \{x \in \mathbb{R}; x = k \cdot \pi,\ k \in \mathbb{Z}\} \to \mathbb{R},\ x \mapsto \frac{\cos(x)}{\sin(x)} \tag{3.65}$$

gilt für alle x aus dem entsprechenden Definitionsbereich:

$$(\sin(x))' = \cos(x) \tag{3.66}$$
$$(\cos(x))' = -\sin(x) \tag{3.67}$$
$$(\tan(x))' = \frac{1}{(\cos(x))^2} \tag{3.68}$$
$$(\cot(x))' = -\frac{1}{(\sin(x))^2} \tag{3.69}$$

Für den Beweis verwendet man die folgenden Additionstheoreme:

$$\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y) \tag{3.70}$$
$$\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y) \tag{3.71}$$

Damit lassen sich auch die folgenden wichtigen Gleichungen herleiten:

$$\begin{aligned}
\cos(x-y) &= \cos(x)\cos(y)+\sin(x)\sin(y) && (3.72)\\
\sin(x-y) &= \sin(x)\cos(y)-\cos(x)\sin(y) && (3.73)\\
\sin(x)+\sin(y) &= 2\sin\left(\frac{x+y}{2}\right)\cdot\cos\left(\frac{x-y}{2}\right) && (3.74)\\
\sin(x)-\sin(y) &= 2\sin\left(\frac{x-y}{2}\right)\cdot\cos\left(\frac{x+y}{2}\right) && (3.75)\\
\cos(x)+\cos(y) &= 2\cos\left(\frac{x+y}{2}\right)\cdot\cos\left(\frac{x-y}{2}\right) && (3.76)\\
\cos(2x) &= \cos^2(x)-\sin^2(x)=2\cos^2(x)-1 && (3.77)\\
\sin(2x) &= 2\sin(x)\cdot\cos(x) && (3.78)\\
1+\cos(x) &= 2\cos^2\left(\frac{x}{2}\right) && (3.79)\\
1-\cos(x) &= 2\sin^2\left(\frac{x}{2}\right). && (3.80)
\end{aligned}$$

Die Komposition $x \mapsto f(g(x))$ zweier differenzierbarer Funktionen definiert auf einem offenen Intervall I ist ebenfalls differenzierbar und es gilt für alle $x \in I$:

$$\frac{d}{dx}f(g(x)) = f'(g(x))\cdot g'(x), \tag{3.81}$$

denn mit $g(x) \neq g(x_0)$ gilt:

$$\begin{aligned}
\lim_{x\to x_0}\frac{f(g(x))-f(g(x_0))}{x-x_0} &= \lim_{x\to x_0}\frac{(f(g(x))-f(g(x_0)))\cdot(g(x)-g(x_0))}{(g(x)-g(x_0))\cdot(x-x_0)} = \\
&= \lim_{x\to g(x_0)}\frac{(f(x)-f(g(x_0)))}{x-g(x_0)}\cdot\lim_{x\to x_0}\frac{g(x)-g(x_0)}{x-x_0} \qquad (3.82)\\
&= f'(g(x_0))\cdot g'(x_0).
\end{aligned}$$

Beispiel(e) 3.24

- $h:\mathbb{R}\to\mathbb{R}$, $x\mapsto(x^4+6x+5)^3$,
  $f:\mathbb{R}\to\mathbb{R}$, $x\mapsto x^3$, $g:\mathbb{R}\to\mathbb{R}$, $x\mapsto x^4+6x+5$, $h=f\circ g$.
  $h':\mathbb{R}\to\mathbb{R}$, $x\mapsto 3(x^4+6x+5)^2\cdot(4x^3+6)$
- $k:\mathbb{R}\to\mathbb{R}$, $x\mapsto\sin^2(x^4+2x)$,
  $f:\mathbb{R}\to\mathbb{R}$, $x\mapsto x^2$, $g:\mathbb{R}\to\mathbb{R}$, $x\mapsto\sin(x)$, $h:\mathbb{R}\to\mathbb{R}$, $x\mapsto x^4+2x$,
  $k=f\circ g\circ h$, also $k:\mathbb{R}\to\mathbb{R}$, $x\mapsto= f(g(h(x)))$.
  $k':\mathbb{R}\to\mathbb{R}$, $x\mapsto 2(\sin(x^4+2x))\cdot\cos(x^4+2x)\cdot(4x^3+2)$.

Die Ableitung der Ableitung bezeichnen wir, falls sie existiert, mit f'' und für $\frac{d}{dx}\left(\frac{d}{dx}f(x)\right)$ schreiben wir $\frac{d^2}{dx^2}f(x)$.
Allgemein definieren wir

$$f^{(0)}=f,\quad f^{(1)}=f',\quad f^{(n)}=\left(f^{(n-1)}\right)',\quad n\in\mathbb{N}. \tag{3.83}$$

Man sagt, f ist n-mal (stetig) differenzierbar, wenn die n-te Ableitung von f existiert (existiert und stetig ist).
Die Funktion $f:\mathbb{R}\to\mathbb{R}$, $x\mapsto x\cdot|x|$ ist nur einmal stetig differenzierbar.
Eine insbesondere bezüglich der Differentiation interessante Funktion ist die bereits eingeführte

Exponentialfunktion $\exp : \mathbb{R} \to \mathbb{R}$, $x \mapsto \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$, denn es gilt:

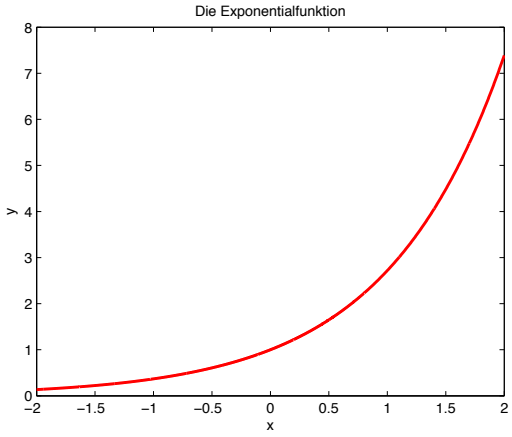
$$\frac{d}{dx}\exp(x) = \exp(x), \quad x \in \mathbb{R}. \tag{3.84}$$

Weitere Eigenschaften der $\exp$-Funktion, die unter Verwendung der unterschiedlichen Darstellungen der $\exp$-Funktion bewiesen werden:

$$\begin{array}{lcll} \exp(1) & = & e & \\ \exp(x) & > & 0 & \text{für alle } x \in \mathbb{R} \\ \exp(x+y) & = & \exp(x)\cdot\exp(y) & \text{für alle } x, y \in \mathbb{R} \\ \exp(-x) & = & \frac{1}{\exp(x)} & \text{für alle } x \in \mathbb{R} \\ \exp(sx) & = & (\exp(x))^s & \text{für alle } x \in \mathbb{R},\ s \in \mathbb{Q}. \end{array}$$

Somit gilt

$$\exp(s) = \exp(s1) = \exp(1)^s = e^s \quad \text{für alle } s \in \mathbb{Q}. \tag{3.85}$$



(3.86)

Aus den speziellen Eigenschaften der Exponentialfunktion folgt, dass jede auf einem offenen Intervall $I \subseteq \mathbb{R}$ differenzierbare Funktion f mit $f' : I \to \mathbb{R}$, $x \mapsto a \cdot f(x)$ von der Form $f : I \to \mathbb{R}$, $x \mapsto c \cdot \exp(ax)$ mit einer Konstanten $c \in \mathbb{R}$ ist.

Beispiel(e) 3.25 (Radioaktiver Zerfall)
Aufgrund von Beobachtungen geht man davon aus, dass für $N(t)$ Atome eines zerfallenden Stoffes die Anzahl ΔN der Zerfälle in einem kleinen Zeitintervall Δt gegeben ist durch

$$\Delta N = -kN\Delta t. \tag{3.87}$$

Unter der Annahme, dass die Funktion N differenzierbar ist, ergibt sich hieraus

$$\frac{d}{dt}N(t) = -k \cdot N(t) \tag{3.88}$$

und somit das Zerfallsgesetz

$$N(t) = N(0)\exp(-kt). \tag{3.89}$$

Alle in der Natur beobachtbaren Wachstums- und Zerfallsprozesse werden mit Hilfe der Exponentialfunktion beschrieben.
Da die Exponentialfunktion eine streng monoton steigende Funktion mit der Wertemenge $\mathbb{R}_0^+ \setminus \{0\}$ ist, existiert eine Umkehrfunktion, die als natürlicher Logarithmus

$$\ln : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}, \quad x \mapsto \ln(x) \tag{3.90}$$

bezeichnet wird. Es gilt somit $\exp(\ln(x)) = x$ für alle $x \in \mathbb{R}_0^+ \setminus \{0\}$ und $\ln(\exp(x)) = x$ für alle $x \in \mathbb{R}$. Ist nun $f : I \to \mathbb{R}$ eine umkehrbare und differenzierbare Funktion, so ist die Umkehrfunktion $f^{-1} : f(I) \to \mathbb{R}$ in allen $x \in f(I)$ mit $f'(f^{-1}(x)) \neq 0$ differenzierbar und es gilt:

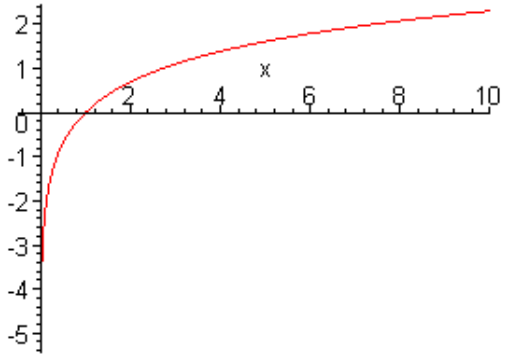
$$\frac{d}{dx} f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}. \tag{3.91}$$

Für den natürlichen Logarithmus erhalten wir somit:

$$\ln'(x) = \frac{1}{\exp(\ln(x))} = \frac{1}{x} \quad \text{für alle } x > 0. \tag{3.92}$$

Ferner gilt:

$$\ln(x \cdot y) = \ln(x) + \ln(y), \quad \ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y), \quad x, y > 0. \tag{3.93}$$



(3.94)

Da $\exp(sx) = (\exp(x))^s$ für alle $x \in \mathbb{R}$ und $s \in \mathbb{Q}$, definieren wir für $x = \ln(a)$, $a > 0$:

$$a^s := \exp(s \ln(a)) \quad \text{für alle} \quad s \in \mathbb{R}. \tag{3.95}$$

(Exponentialfunktion zur Basis a)

Es gilt mit $a, b > 0$, $x, y \in \mathbb{R}$:

$$a^x \cdot a^y = a^{x+y}, \quad (ab)^x = a^x b^x, \quad (a^x)^y = a^{x \cdot y}, \quad \ln(a^x) = x \cdot \ln(a) \tag{3.96}$$

und

$$\frac{d}{dx} a^x = a^x \cdot \ln(a). \tag{3.97}$$

Die Umkehrfunktion zu $f : \mathbb{R} \to \mathbb{R}_0^+ \setminus \{0\}$, $x \mapsto a^x$ lautet für $a \neq 1$: $\log_a : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}$ und ist gegeben durch $x \mapsto \frac{\ln(x)}{\ln(a)}$ (Logarithmus zur Basis a: Wichtige Fälle: $a = 2$ (Informationstheorie) mit der Schreibweise ld für $\log_2$ und $a = 10$ mit der Schreibweise $\log$ für $\log_{10}$ (Nachrichtentechnik)).
Weitere für die Anwendungen wichtige Funktionen sind der
sinus hyperbolicus:

$$\sinh : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{e^x - e^{-x}}{2} \tag{3.98}$$

mit der Umkehrfunktion area sinus hyperbolicus:

$$\operatorname{arsinh} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \ln\left(x + \sqrt{x^2+1}\right), \tag{3.99}$$

cosinus hyperbolicus:

$$\cosh : \mathbb{R} \to [1, \infty), \quad x \mapsto \frac{e^x + e^{-x}}{2} \tag{3.100}$$

mit der Umkehrfunktion area cosinus hyperbolicus des zu $[0, \infty)$ gehörenden Zweiges:

$$\operatorname{arcosh} : [1, \infty) \to [0, \infty), \quad x \mapsto \ln\left(x + \sqrt{x^2-1}\right) \tag{3.101}$$

und der
tangens hyperbolicus:

$$\tanh : \mathbb{R} \to (-1, 1), \quad x \mapsto \frac{\sinh(x)}{\cosh(x)} \tag{3.102}$$

mit der Umkehrfunktion area tangens hyperbolicus:

$$\operatorname{artanh} : (-1, 1) \to \mathbb{R}, \quad x \mapsto \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right). \tag{3.103}$$

Die jeweiligen Ableitungen berechnen sich durch die Kettenregel. Es gilt zum Beispiel:

$$(\operatorname{artanh}(x))' = \left(\frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)\right)' = \frac{1}{2}\left(\frac{1-x}{1+x}\right)\frac{(1-x)+(1+x)}{(1-x)^2} = \frac{1}{1-x^2} \tag{3.104}$$

$$(\sinh(x))' = \frac{e^x + e^{-x}}{2} = \cosh(x) \tag{3.105}$$

$$(\cosh(x))' = \frac{e^x - e^{-x}}{2} = \sinh(x) \tag{3.106}$$

Die trigonometrischen Funktionen $\sin, \cos$ sind nicht über dem ganzen Definitionsbereich umkehrbar. Die Sinusfunktion ist über dem Intervall $-\frac{\pi}{2} \le x \le \frac{\pi}{2}$ streng monoton wachsend und daher dort umkehrbar mit der Umkehrfunktion arcus sinus:

$$\arcsin : [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}] \tag{3.107}$$

Für die Ableitung erhalten wir:

$$(\arcsin(x))' = \frac{1}{\cos(\arcsin(x))} = \frac{1}{\sqrt{1 - \sin^2(\arcsin(x))}} = \frac{1}{\sqrt{1-x^2}}, \quad |x| < 1 \tag{3.108}$$

Die Cosinusfunktion ist im Intervall $[0, \pi]$ streng monoton fallend und daher dort umkehrbar mit der Umkehrfunktion

$$\arccos : [-1, 1] \to [0, \pi]. \tag{3.109}$$

Für die Ableitung gilt mit $|x| < 1$:

$$(\arccos(x))' = -\frac{1}{\sin(\arccos(x))} = -\frac{1}{\sqrt{1 - \cos^2(\arccos(x))}} = -\frac{1}{\sqrt{1-x^2}}. \tag{3.110}$$

Analog erhält man

$$\arctan : \mathbb{R} \to \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \qquad (\arctan(x))' = \frac{1}{1+x^2}, \quad x \in \mathbb{R} \tag{3.111}$$

$$\operatorname{arccot} : \mathbb{R} \to (0, \pi) \qquad (\operatorname{arccot}(x))' = -\frac{1}{1+x^2}, \quad x \in \mathbb{R}. \tag{3.112}$$

Für eine gegebene komplexe Zahl $z = x + iy$ definieren wir nun die Exponentialfunktion folgendermaßen:

$$\exp : \mathbb{C} \to \mathbb{C}, \quad z \mapsto e^z := e^x \cdot (\cos(y) + i \sin(y)). \tag{3.113}$$

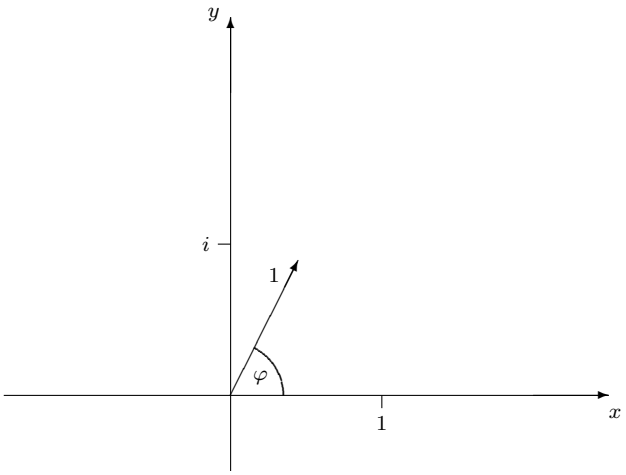
Speziell für $z = iy$ folgt somit:

$$e^{iy} = \cos(y) + i \sin(y), \quad |e^{iy}| = \cos^2(y) + \sin^2(y) = 1. \tag{3.114}$$

Jede komplexe Zahl mit $|z| = 1$ läßt sich also in der Form

$$z = \cos(\varphi) + i \sin(\varphi) \tag{3.115}$$

darstellen. Der Winkel φ kann dabei auch durch $\varphi + 2\pi k$, $k \in \mathbb{Z}$, ersetzt werden.



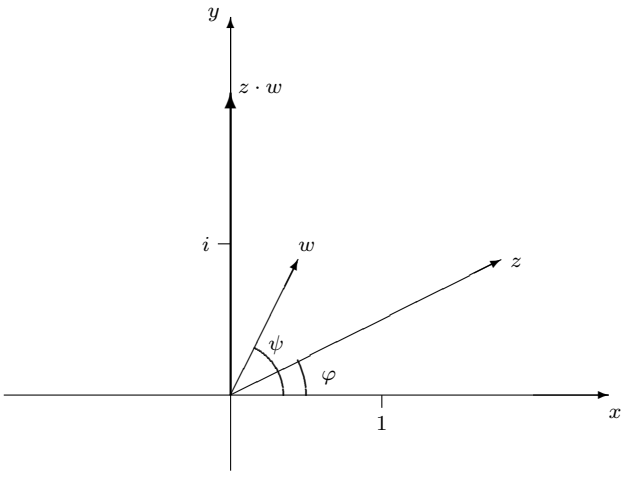
(3.116)

Somit erhalten wir für jede beliebige komplexe Zahl $z = x + iy$ die Darstellung

$$z = |z| \cdot (\cos(\varphi) + i \sin(\varphi)) = |z| \cdot e^{i\varphi}. \tag{3.117}$$

Den kartesischen Koordinaten x und y entsprechen dabei die polaren Koordinaten $|z|$ und φ. Während die Addition $z + w$ und die Subtraktion $z - w$ komplexer Zahlen in kartesischen Koordinaten $z = x + iy$, $w = u + iv$ zu naheliegenden Formeln führte, erhält man nun für die Multiplikation und für die Division komplexer Zahlen $z = |z| \cdot e^{i\varphi}$, $w = |w| \cdot e^{i\psi}$ ($|w| \neq 0$) einfache Formeln:

$$z \cdot w = |z| \cdot e^{i\varphi} \cdot |w| \cdot e^{i\psi} = \underbrace{|z| \cdot |w|}_{=|z \cdot w|} \cdot e^{i(\varphi + \psi)} \tag{3.118}$$



(3.119)

$$\frac{z}{w} = |z| \cdot e^{i\varphi} \cdot \frac{1}{|w|} \cdot e^{-i\psi} = \frac{|z|}{|w|} \cdot e^{i(\varphi-\psi)}, \quad w \neq 0. \tag{3.120}$$

Der Logarithmus naturalis ($\ln$) einer komplexen Zahl $z = |z| \cdot e^{i\varphi} \neq 0$ ist definiert durch

$$\ln(z) := \ln(|z|) + i\varphi. \tag{3.121}$$

Da φ in der Darstellung $z = |z| \cdot e^{i\varphi}$ nicht eindeutig ist, fordert man $-\pi < \varphi \leq \pi$. Der entsprechende Winkel φ wird Hauptargument genannt und mit $\arg(z)$ bezeichnet.
Ein wichtiger Satz im Zusammenhang mit komplexen Zahlen ist der nun folgende:
Fundamentalsatz der Algebra: Seien $n \in \mathbb{N}$ und $a_0, a_1, \ldots, a_n \in \mathbb{C}$ mit $a_n \neq 0$, so gibt es komplexe Zahlen $z_1, z_2, \ldots, z_n$ mit

$$a_n z^n + a_{n-1} z^{n-1} + \ldots + a_2 z^2 + a_1 z + a_0 = a_n(z - z_1) \cdot \ldots \cdot (z - z_n) \tag{3.122}$$

für alle $z \in \mathbb{C}$.

Beispiel(e) 3.26 (n-te Einheitswurzel)
Die Gleichung

$$z^n = 1 \tag{3.123}$$

kann mit dem Ansatz $z = e^{i\varphi}$ aufgelöst werden:

$$z^n = 1 \quad \Longleftrightarrow \quad n\varphi = k2\pi, \quad k = 0, 1, \ldots, (n-1). \tag{3.124}$$

Somit ist

$$z_{k+1} = e^{i2\pi\frac{k}{n}}, \quad k = 0, 1, \ldots, (n-1). \tag{3.125}$$

Den wichtigen Zusammenhang zwischen den trigonometrischen Funktionen $\cos, \sin$ und der komplexen Exponentialfunktion zeigt die folgende Rechnung:

$$\begin{aligned} e^{i(\varphi+\psi)} &= \cos(\varphi+\psi) + i\sin(\varphi+\psi) = e^{i\varphi}e^{i\psi} = \\ &= (\cos(\varphi) + i\sin(\varphi))(\cos(\psi) + i\sin(\psi)) = \\ &= \cos(\varphi)\cos(\psi) - \sin(\varphi)\sin(\psi) + i(\sin(\varphi)\cos(\psi) + \sin(\psi)\cos(\varphi)). \end{aligned} \tag{3.126}$$

Somit reproduzieren sich die Additionstheoreme für die Funktionen $\cos$ und $\sin$:

$$\cos(\varphi+\psi) = \cos(\varphi)\cos(\psi) - \sin(\varphi)\sin(\psi) \tag{3.127}$$

$$\sin(\varphi+\psi) = \sin(\varphi)\cos(\psi) + \sin(\psi)\cos(\varphi). \tag{3.128}$$

Zu den wichtigsten Anwendungen der Differentiation gehört die Berechnung von Maxima und Minima gegebener Funktionen. Eine auf $\mathbb{D} \subseteq \mathbb{R}$ definierte Funktion $f : \mathbb{D} \to \mathbb{R}$ hat in $a \in \mathbb{D}$ ein globales Maximum (Minimum), wenn $f(x) \leq f(a)$ für alle $x \in \mathbb{D}$ gilt ($f(a) \leq f(x)$ für alle $x \in \mathbb{D}$). In diesem Fall heißt a ein globaler Maximierer (Minimierer) von f. Eine Stelle $b \in \mathbb{D}$ heißt lokaler Maximierer (Minimierer) von f, falls es ein $\delta > 0$ gibt mit $f(x) \leq f(b)$ für alle $x \in \mathbb{D} \cap (b-\delta, b+\delta)$ ($f(b) \leq f(x)$ für alle $x \in \mathbb{D} \cap (b-\delta, b+\delta)$). Der Begriff „global“ bezieht sich also immer auf den ganzen Definitionsbereich, der Begriff „lokal“ nur auf die unmittelbare Umgebung von b. Ist x_0 ein Maximierer von f, dann ist x_0 ein Minimierer von $-f$ und umgekehrt. Der folgende Satz beschreibt einen Zusammenhang zwischen der Lage von Extremalstellen (Maximierer und Minimierer) und der Ableitung der gegebenen Funktion f.

Satz 3.27 (Differentiation und Extremalstellen)
Sei $f : I \to \mathbb{R}$ eine auf dem offenen Intervall I differenzierbare Funktion, so gilt:
Ist $x_0 \in I$ (lokale) Extremalstelle von f, so gilt: $f'(x_0) = 0$

Die Bedingung $f'(x_0) = 0$ (waagerechte Tangente des Graphen von f in $(x_0, f(x_0))$) ist zwar notwendig für eine Extremalstelle, aber nicht hinreichend, wie das Beispiel $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^3$ mit $x_0 = 0$ zeigt. Der Satz gibt auch keine Auskunft über Extremalstellen an Intervallenden, an Spitzen oder anderen Stellen, an denen f nicht differenzierbar ist.
Als Kandidaten für Extremalstellen einer Funktion $f : I \to \mathbb{R}$ kommen also in Frage:

a) die Randpunkte von I, falls I nicht offen ist

b) die Punkte aus I, an denen f nicht differenzierbar ist

c) die Punkte aus I, für die $f'(x) = 0$ gilt (stationäre Punkte).

Beispiel(e) 3.28

- Aus einer rechtwinkligen Blechplatte der Seitenlängen 16cm und 10cm soll eine quaderformige, oben offene Wanne mit maximalem Volumen geformt werden.
  Für das Volumen $V(x)$ in Abhängigkeit von der Höhe x der Wanne erhalten wir:
  $V : [0, 5] \to \mathbb{R}$, $x \mapsto (16 - 2x) \cdot (10 - 2x) \cdot x$
  Wegen $V(0) = V(5) = 0$ sind die Randpunkte keine Maximierer. Es gilt:

$$V' : (0, 5) \to \mathbb{R}, \quad x \mapsto 12x^2 - 104x + 160. \tag{3.129}$$

  $V'(x) = 0 \iff 12x^2 - 104x + 160 = 0 \iff x = 2$, da $x \in [0, 5]$.
  Wir erhalten den globalen Maximierer $x_0 = 2$ mit $V(2) = 144\text{cm}^3$.

- In der Informationstheorie betrachtet man Folgen $\{x_i\}$ von Bits $x_i \in \{\pm 1\}$, $i \in \mathbb{N}_0$, mit Wahrscheinlichkeiten $P(x_i = 1) = w$, $P(x_i = -1) = 1 - w$, $w \in [0, 1]$.
  Der Informationsgehalt eines Zeichens x_i wird gemessen durch $E : [0, 1] \to \mathbb{R}$, $w \mapsto -w \cdot \log_2(w) - (1 - w) \cdot \log_2(1 - w)$ für $w \in (0, 1)$, und $E(0) = E(1) = 0$.
  Für welches w wird der Informationsgehalt maximal?

$$E' : (0, 1) \to \mathbb{R}, \quad w \mapsto -\log_2(w) - w \cdot \frac{1}{\ln(2) \cdot w} + \log_2(1 - w) + \frac{1 - w}{\ln(2)(1 - w)}. \tag{3.130}$$

$$E'(w) = 0 \iff \log_2(w) = \log_2(1 - w) \iff w = 1 - w \iff w = \frac{1}{2}. \tag{3.131}$$

  Für $w = \frac{1}{2}$ erhält man den maximalen Informationsgehalt.

Ein wichtiges Hilfsmittel für die Analyse von gegebenen Funktionen bildet der folgende Satz:

Satz 3.29 (Mittelwertsatz)
Ist eine Funktion f auf einem abgeschlossenen Intervall $[a, b]$ definiert und stetig sowie auf dem offenen Intervall (a, b) differenzierbar, dann gibt es mindestens ein $x_0 \in (a, b)$ mit

$$f'(x_0) = \frac{f(b) - f(a)}{b - a}, \quad a \neq b. \tag{3.132}$$

Im Folgenden fassen wir wichtige Anwendungen des Mittelwertsatzes zusammen. Dazu sei $f : I \to \mathbb{R}$ eine auf dem offenen Intervall I differenzierbare Funktion. Es gilt:

(a) $f'(x) > 0$ auf $I \Longrightarrow f$ ist auf I streng monoton wachsend
(b) $f'(x) < 0$ auf $I \Longrightarrow f$ ist auf I streng monoton fallend
(c) $f'(x) \geq 0$ auf $I \Longrightarrow f$ ist auf I monoton wachsend
(d) $f'(x) \leq 0$ auf $I \Longrightarrow f$ ist auf I monoton fallend
(e) $f'(x) = 0$ auf $I \Longrightarrow f$ ist auf I konstant. (3.133)

Beispiel(e) 3.30

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{x}{\sqrt{1+x^2}}, \quad f' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{(\sqrt{1+x^2})^3}, \tag{3.134}$$

somit ist f streng monoton wachsend.

Der Mittelwertsatz kann auch dazu verwendet werden, Kandidaten für Extremalstellen zu klassifizieren.

Satz 3.31 (Extremwert-Test)
Sei $f : (a,b) \to \mathbb{R}$ eine differenzierbare Funktion mit $f'(x_0) = 0$, so hat f in x_0 ein lokales Maximum (Minimum), falls es ein $\varepsilon > 0$ gibt mit $f'(y) > 0$ $(f'(y) < 0)$ für alle $y \in (x_0 - \varepsilon, x_0)$ und $f'(y) < 0$ $(f'(y) > 0)$ für alle $y \in (x_0, x_0 + \varepsilon)$.

Steht für die Funktion f die zweite Ableitung zur Verfügung, so kann der Extremwert-Test auch folgendermaßen durchgeführt werden:

$f'(x_0) = 0$ und $f''(x_0) > 0 \Longrightarrow x_0$ ist lokales Minimum
$f'(x_0) = 0$ und $f''(x_0) < 0 \Longrightarrow x_0$ ist lokales Maximum (3.135)

Die zweite Ableitung bestimmt die Krümmung des Graphen $y = f(x)$.
Ist $f'' > 0$, so ist f konvex (Linkskrümmung), etwa $f : \mathbb{R} \to \mathbb{R}, x \mapsto x^2$.
Ist $f'' < 0$, so ist f konkav (Rechtskrümmung), etwa $f : \mathbb{R} \to \mathbb{R}, x \mapsto -x^2$.
Eine Stelle, an der sich die Krümmung ändert, wird als Wendepunkt bezeichnet. Als Kandidaten für Wendepunkte erhalten wir

a) alle Punkte an denen zwar f, aber nicht f'' definiert ist

b) alle Punkte mit $f''(x_0) = 0$.

Ist eine Funktion $f : I \to \mathbb{R}$ im offenen Intervall I dreimal differenzierbar und gilt für $x_0 \in I$: $f''(x_0) = 0$, $f'''(x_0) \neq 0$, so ist x_0 ein Wendepunkt.
Gilt zudem $f'(x_0) = 0$, so heißt x_0 Sattelpunkt, zum Beispiel $f : \mathbb{R} \to \mathbb{R}, x \mapsto x^3$, $x_0 = 0$.
Der Mittelwertsatz läßt sich folgendermaßen verallgemeinern:

Satz 3.32 (verallgemeinerter Mittelwertsatz)
Seien $f, g : \mathbb{D} \subseteq \mathbb{R} \to \mathbb{R}$ in einem offenen Intervall $(a,b) \subset \mathbb{D}$ differenzierbar, in $[a,b] \subseteq \mathbb{D}$ stetig und $g'(x) \neq 0$ in (a,b), dann gibt es mindestens eine Stelle $\xi \in (a,b)$ mit

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)} \quad (g(b) \neq g(a)). \tag{3.136}$$

Dieser Satz ist Grundlage für die wichtige Regel von L'Hospital:
Sind $f, g : (a, b) \to \mathbb{R}$ differenzierbar mit:

a) $g'(x) \neq 0$, $x \in (a, b)$,

b) $f(x) \to 0$, $g(x) \to 0$ oder $f(x) \to \pm\infty$, $g(x) \to \pm\infty$ für $x \to b-$

c) $\lim\limits_{x \to b-} \frac{f'(x)}{g'(x)} = L$ mit $L \in \mathbb{R} \cup \{\pm\infty\}$,

dann gilt:

$$\lim_{x \to b-} \frac{f(x)}{g(x)} = \lim_{x \to b-} \frac{f'(x)}{g'(x)}. \tag{3.137}$$

Entsprechendes gilt für $x \to a+$ bzw. $x \to \infty$ (dann ist $(a, b) = (a, \infty)$) oder $x \to -\infty$ (dann ist $(a, b) = (-\infty, b)$) .

Beispiel(e) 3.33

- $\lim\limits_{x \to 0} \frac{\sin(x)}{x} = \lim\limits_{x \to 0} \frac{\cos(x)}{1} = 1$
- $\lim\limits_{x \to \infty} x \cdot \ln\left(\frac{x+1}{x-1}\right) = \lim\limits_{x \to \infty} \frac{\ln(x+1) - \ln(x-1)}{\frac{1}{x}} = \lim\limits_{x \to \infty} \frac{\frac{1}{x+1} - \frac{1}{x-1}}{-\frac{1}{x^2}} = \lim\limits_{x \to \infty} \frac{2x^2}{x^2 - 1} = 2$
- $\lim\limits_{x \to 0} \left(\frac{1}{x} - \frac{1}{\sin(x)}\right) = \lim\limits_{x \to 0} \frac{\sin(x) - x}{x \cdot \sin(x)} = \lim\limits_{x \to 0} \frac{\cos(x) - 1}{\sin(x) + x \cdot \cos(x)} = \lim\limits_{x \to 0} \frac{-\sin(x)}{\cos(x) + \cos(x) - x \cdot \sin(x)} = 0$

3.3 Potenzreihen

Die aus einer Zahlenfolge $\{a_k\}, k \in \mathbb{N}_0$, gebildete Summenfolge $\{s_n\}, n \in \mathbb{N}_0$, mit

$$s_n := \sum_{k=0}^{n} a_k \tag{3.138}$$

heißt unendliche Reihe und wird mit $\sum\limits_{k=0}^{\infty} a_k$ bezeichnet. Die Folgenglieder s_i werden als Partialsummen bezeichnet. Die Konvergenz- und Divergenzbegriffe für Zahlenfolgen lassen sich durch die Betrachtung von $\{s_n\}$ auf unendliche Reihen übertragen.

Beispiel(e) 3.34

$\{s_n\} := \sum\limits_{k=0}^{\infty} q^k$ mit $|q| < 1$. Es gilt:

$$\lim_{n \to \infty} s_n = \frac{1}{1 - q}. \tag{3.139}$$

Somit können Konvergenzkriterien für Zahlenfolgen auch für unendliche Reihen verwendet werden. Speziell für unendliche Reihen ist das folgende Leibniz-Kriterium wichtig:

- Sei $\{a_k\}, k \in \mathbb{N}_0$, eine monoton fallende Nullfolge, so konvergiert die unendliche Reihe $\sum\limits_{k=0}^{\infty} (-1)^k a_k$.

Elementare Manipulationen, die bei endlichen Summen erlaubt sind, sind bei unendlichen Reihen nicht uneingeschränkt möglich.

Beispiel(e) 3.35
Die unendliche Reihe $\sum_{k=0}^{\infty} a_k = (1-1) + (1-1) + ..$ mit $a_k = (1-1) = 0$ konvergiert gegen 0. Die unendliche Reihe $\sum_{k=0}^{\infty} (-1)^k = 1 - 1 + 1 - 1 + 1..$, die aus $\sum_{k=0}^{\infty} a_k$ durch Entfernen der Klammern entsteht, ist divergent.

Oft untersucht man, ob eine gegebene unendliche Reihe $\sum_{k=0}^{\infty} a_k$ absolut konvergiert, ob also die unendliche Reihe $\sum_{k=0}^{\infty} |a_k|$ einen Grenzwert besitzt. Dann konvergiert auch $\sum_{k=0}^{\infty} a_k$. Eine direkte Folge des Monotoniekriteriums für Zahlenfolgen ist der folgende

Satz 3.36 (absolute Konvergenz)
Eine unendliche Reihe

$$\sum_{k=0}^{\infty} a_k$$

ist genau dann absolut konvergent, wenn die Zahlenfolge $\{b_n\}, n \in \mathbb{N}_0$, mit $b_n = \sum_{k=0}^{n} |a_k|$ beschränkt ist.

Ein wichtiges Kriterium für die Konvergenz unendlicher Reihen ist das **Quotientenkriterium**: Sei $\sum_{k=0}^{\infty} a_k$ eine unendliche Reihe mit $a_k \neq 0$ für alle $k \geq k_0 \in \mathbb{N}_0$ und sei $\left\{ \left| \frac{a_{k+1}}{a_k} \right| \right\}, k \geq k_0$, konvergent, so gilt:

$$\lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| < 1 \implies \sum_{k=0}^{\infty} a_k \quad \text{ist absolut konvergent} \tag{3.140}$$

$$\lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| > 1 \implies \sum_{k=0}^{\infty} a_k \quad \text{konvergiert nicht.} \tag{3.141}$$

Analog zu Zahlenfolgen kann man auch Folgen $\{f_n\}, n \in \mathbb{N}_0$, von Funktionen $f_i : \mathbb{D} \to \mathbb{R}$, betrachten. Für jedes $x \in \mathbb{D}$ ergibt sich dann eine Zahlenfolge $\{f_n(x)\}, n \in \mathbb{N}_0$. Eine Folge $\{p_n\}, n \in \mathbb{N}_0$,

$$p_n : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \sum_{k=0}^{n} a_k x^k \tag{3.142}$$

heißt Potenzreihe und wird durch $\sum_{k=0}^{\infty} a_k x^k$ notiert. Dabei ist wichtig festzuhalten, dass die Koeffizienten von p_n und p_{n+1} für x^i, $i = 0, \ldots, n$, übereinstimmen.
An einer Potenzreihe interessiert die Menge M aller $x \in \mathbb{R}$, sodass die unendliche Reihe $\{p_n(x)\}$, $n \in \mathbb{N}_0$, konvergiert.
Die Größe

$$R := \begin{cases} \sup\{|x|, x \in M\}, & \text{falls } M \text{ beschränkt ist} \\ \infty & \text{sonst} \end{cases} \tag{3.143}$$

heißt Konvergenzradius.

Beispiel(e) 3.37

- Die Potenzreihe $\sum_{k=0}^{\infty} k!x^k$ hat den Konvergenzradius 0. Sie konvergiert nur für $x = 0$.
- Die Potenzreihe $\sum_{k=0}^{\infty} x^k$ hat den Konvergenzradius 1,
  denn $M = \{x \in \mathbb{R}; |x| < 1\}$

Eine Potenzreihe $\sum_{k=0}^{\infty} a_k x^k$ mit Konvergenzradius $R > 0$ konvergiert absolut für alle $-R < x < R$.
Dies gilt auch für die Potenzreihe $\sum_{k=1}^{\infty} k a_k x^{k-1}$ und damit für

$$\sum_{k=l}^{\infty} k(k-1) \cdot \ldots \cdot (k-(l-1)) a_k \cdot x^{k-l}, \quad l \in \mathbb{N}. \tag{3.144}$$

Ist für eine Potenzreihe $\sum_{k=0}^{\infty} a_k x^k$ $a_k \neq 0$ für $k \geq k_0 \in \mathbb{N}_0$, so läßt sich der Konvergenzradius durch

$$R = \lim_{k\to\infty} \left| \frac{a_k}{a_{k+1}} \right| \tag{3.145}$$

berechnen, falls $\left\{ \left| \frac{a_k}{a_{k+1}} \right| \right\}, k \geq k_0$, konvergiert oder bestimmt divergiert.

Beispiel(e) 3.38

Die Potenzreihe $\sum_{k=0}^{\infty} \frac{k^k}{k!} x^k$ hat den Konvergenzradius

$$R = \lim_{k\to\infty} \left| \frac{a_k}{a_{k+1}} \right| = \lim_{k\to\infty} \left(\frac{k}{k+1} \right)^k = \frac{1}{e}. \tag{3.146}$$

Sei nun $\sum_{k=0}^{\infty} a_k x^k$ eine Potenzreihe mit Konvergenzradius $R > 0$, so konvergieren die Zahlenfolgen $\left\{ \sum_{k=0}^{n} a_k x^k \right\}$ und $\left\{ \sum_{n=l}^{n} k \cdot \ldots \cdot (k-(l-1)) a_k x^{k-l} \right\}$ für alle $x \in (-R, R)$ und für alle $l \in \mathbb{N}$.
Somit existiert eine Funktion $f : (-R, R) \to \mathbb{R}$, $x \mapsto \lim_{n\to\infty} \sum_{k=0}^{n} a_k x^k$ und es gilt (**Differentiation von Potenzreihen**):

$$\begin{aligned}
f' &: (-R, R) \to \mathbb{R}, \qquad x \mapsto \lim_{n\to\infty} \sum_{k=1}^{n} k \cdot a_k x^{k-1} \\
f'' &: (-R, R) \to \mathbb{R}, \qquad x \mapsto \lim_{n\to\infty} \sum_{k=2}^{n} k(k-1) \cdot a_k x^{k-2} \\
f^{(l)} &: (-R, R) \to \mathbb{R}, \qquad x \mapsto \lim_{n\to\infty} \sum_{k=l}^{n} k(k-1) \cdot \ldots \cdot (k-(l-1)) a_k x^{k-l}, \ \ l \in \mathbb{N}.
\end{aligned} \tag{3.147}$$

Beispiel(e) 3.39

- $\exp : \mathbb{R} \to \mathbb{R},\ x \mapsto \lim\limits_{n\to\infty} \sum\limits_{k=0}^{n} \frac{1}{k!} x^k$
- $\sin : \mathbb{R} \to \mathbb{R},\ x \mapsto \lim\limits_{n\to\infty} \sum\limits_{k=0}^{n} \frac{(-1)^k}{(2k+1)!} x^{2k+1}$
- $\cos : \mathbb{R} \to \mathbb{R},\ x \mapsto \lim\limits_{n\to\infty} \sum\limits_{k=0}^{n} \frac{(-1)^k}{(2k)!} x^{2k}$
- $\ln(1+\bullet) : (-1,1) \to \mathbb{R},\ x \mapsto \lim\limits_{n\to\infty} \sum\limits_{k=0}^{n} \frac{(-1)^k}{k+1} x^{k+1} (= \ln(1+x))$

Eine unendliche Reihe der Form

$$\sum_{k=0}^{\infty} a_k (x-a)^k \tag{3.148}$$

heißt Potenzreihe mit dem Zentrum (Entwicklungspunkt) a. Die bisherigen Potenzreihen haben das Zentrum $a = 0$. Für theoretische Überlegungen genügt es, sich auf den Fall $a = 0$ zu beschränken, denn eine Potenzreihe mit Zentrum a geht durch die Substitution $z := x - a$ in die Potenzreihe $\sum\limits_{k=0}^{\infty} a_k z^k$ über. Als Konvergenzradius einer Potenzreihe mit Zentrum a wird der Konvergenzradius der entsprechenden Potenzreihe mit Zentrum 0 betrachtet. Es gilt dann:

$$x \in (a-R, a+R) \implies \sum_{k=0}^{\infty} a_k (x-a)^k \text{ konvergiert} \tag{3.149}$$

$$x \notin [a-R, a+R] \implies \sum_{k=0}^{\infty} a_k (x-a)^k \text{ divergiert.} \tag{3.150}$$

Falls eine Funktion f über $(a-R, a+R)$, $R > 0$, als Potenzreihe

$$f(x) = \lim_{n\to\infty} \sum_{k=0}^{n} a_k (x-a)^k \tag{3.151}$$

mit Zentrum a darstellbar ist, gilt:

$$a_k := \frac{f^{(k)}(a)}{k!} \quad (f^{(k)}\text{: die } k\text{-te Ableitung von } f; \quad f^{(0)} = f) \tag{3.152}$$

Die Potenzreihe $\sum\limits_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x-a)^k$ heißt Taylor-Reihe einer beliebig oft differenzierbaren Funktion $f : \mathbb{D} \to \mathbb{R}$ im Entwicklungspunkt a.
Ist R der Konvergenzradius von $\sum\limits_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x-a)^k$, $(a-R, a+R) \subseteq \mathbb{D}$, und

$$f(x) = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (x-a)^k \tag{3.153}$$

für alle $a - R < x < a + R$, so sagt man: f läßt sich um a als Taylor-Reihe darstellen.

Beispiel(e) 3.40
Es kann passieren, dass die Taylor-Reihe für alle $x \in \mathbb{R}$ konvergiert; aber nur für $x = a$ gleich $f(a)$ ist.
Sei $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto \begin{cases} e^{-\frac{1}{x^2}} & \text{falls } x \neq 0 \\ 0 & x = 0, \end{cases}$
so ist $\lim\limits_{n\to\infty} \sum\limits_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k = 0$ für alle $x \in \mathbb{R}$, aber $f(x) \neq 0$ für $x \neq 0$.

Um festzustellen, wann eine Taylor-Reihe gegen den Funktionswert konvergiert, benötigen wir die Taylor-Formel:
Für jede auf einem offenen Intervall $I \subseteq \mathbb{R}$ $(n+1)$-mal stetig differenzierbare Funktion $f : I \to \mathbb{R}$ und $a, x \in I$ gilt:

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + R_{n+1}(x,a) \text{ mit} \tag{3.154}$$

$$R_{n+1}(x,a) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1} \text{ mit } \xi \text{ zwischen } x \text{ und } a. \tag{3.155}$$

Ist nun f auf dem Intervall I beliebig oft stetig differenzierbar und $a \in I$, dann konvergiert die Taylor-Reihe $\sum\limits_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$ genau dann gegen $f(x)$ für $x \in I$, wenn $\lim\limits_{n\to\infty} R_n(x,a) \to 0$.
Sind die Werte einer Funktion $f : I \to \mathbb{R}$ und ihrer ersten n Ableitungen in einem Punkt $x = a$ aus dem Innern des Intervalls I bekannt, dann wird mit

$$p(x) = f(a) + f'(a)(x-a) + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n \quad \text{(Taylor-Polynom)} \tag{3.156}$$

ein Polynom bestimmt, das die Funktion in einer Umgebung des Punktes $x = a$ gut approximiert. Es gilt:

$$p^{(k)}(a) = f^{(k)}(a) \quad 0 \leq k \leq n \quad (f^{(0)} = f) \tag{3.157}$$

Die Bedeutung der Taylor-Formel besteht darin, dass der Fehler $|f(x) - p(x)|$ durch

$$|f(x) - p(x)| = |R_{n+1}(x,a)| \tag{3.158}$$

darstellbar ist.

Beispiel(e) 3.41

- $e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \frac{e^\xi}{(n+1)!}x^{n+1}$.
  Für $|x| \leq 1$ gilt:
  $$\frac{e^\xi}{(n+1)!}|x|^{n+1} \leq \frac{e}{(n+1)!}|x|^{n+1} \tag{3.159}$$
  Soll der Fehler kleiner oder gleich 10^{-8} sein, so genügt für $x = \frac{1}{10}$ zum Beispiel $n = 5$.
- $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{\sin(\xi)}{8!} \cdot x^8$.

3.4 Integration

Mit der Integration wird das Problem gelöst, aus der Ableitung f' die ursprüngliche Funktion f zu rekonstruieren. Dazu verwendet man den Mittelwertsatz. Zu jeder Zerlegung

$$a = x_0 < x_1 < \ldots < x_n = x \tag{3.160}$$

des Intervalls $[a, x]$, auf dem die Funktion f stetig und in (a, x) differenzierbar ist, gibt es Zwischenpunkte $\xi_i \in [x_{i-1}, x_i]$ mit

$$f(x_i) - f(x_{i-1}) = f'(\xi_i)(x_i - x_{i-1}), \quad i = 1, \ldots, n \tag{3.161}$$

beziehungsweise

$$f(x) - f(x_0) = \sum_{i=1}^{n} f'(\xi_i)(x_i - x_{i-1}). \tag{3.162}$$

Die letzte Gleichung bildet die Grundidee zur Einführung des bestimmten Integrals. Sei f eine auf dem Intervall $[a, b]$ definierte beschränkte Funktion, die an höchstens endlich vielen Stellen nicht stetig ist (derartige Funktionen nennt man stückweise stetig).
Durch Einfügen von $(n-1)$ Teilpunkten

$$a = x_0 < x_1 < \ldots < x_n = b \tag{3.163}$$

wird $[a, b]$ in n Teilintervalle zerlegt.

Für jedes Teilintervall wählt man einen Zwischenpunkt $\xi_i \in [x_{i-1}, x_i]$ und betrachtet die Summe

$$z_n := \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) \tag{3.164}$$

Man bezeichnet z_n nach dem Mathematiker B. RIEMANN (1826-1866) als Riemannsche Summe. Man kann nun zeigen, dass $\lim\limits_{n\to\infty} z_n$ existiert, falls die maximale Intervallbreite der einzelnen Unterteilungen mit $n \to \infty$ gegen Null konvergiert. Außerdem ist dieser Grenzwert $\lim\limits_{n\to\infty} z_n$ unabhängig von der Wahl der Teilpunkte und Zwischenpunkte und wird mit

$$\int_a^b f(x)dx := \lim_{n\to\infty} z_n \tag{3.165}$$

(das bestimmte Integral von f über $[a, b]$) bezeichnet. Die Randpunkte a, b heißen Integrationsgrenzen und $\int\limits_a^b f(x)dx$ ist immer eine Zahl.

Üblicherweise setzt man $\int\limits_a^a f(x)dx = 0$, $\int\limits_b^a f(x)dx = -\int\limits_a^b f(x)dx$ für $a < b$.

Ist $f : [a, b] \to \mathbb{R}$ stetig und $f(x) \geq 0$ für alle $x \in [a, b]$, so ist $\int\limits_a^b f(x)dx$ der Flächeninhalt des von der Kurve $y = f(x)$, der x-Achse und den Geraden $x = a$, $x = b$ begrenzten Flächenstücks.
Verläuft die Kurve $y = f(x)$ ganz unterhalb der x-Achse ($f(x) \leq 0$, $x \in [a, b]$), so kann man den Flächeninhalt I berechnen durch

$$I = \int_a^b (-f(x))dx = -\int_a^b f(x)dx, \quad \text{d.h.} \quad \int_a^b f(x)dx = -I \tag{3.166}$$

Begrenzt $y = f(x)$ Flächenstücke oberhalb und unterhalb der x-Achse, dann ist $\int\limits_a^b f(x)dx$ die Summe der mit Vorzeichen versehenen Flächeninhalte.

Beispiel(e) 3.42

- $\int_a^b c dx = c(b-a)$
- $\int_a^b x dx = \frac{1}{2}(b^2 - a^2)$
- $\int_{-1}^{1} x dx = 0.$

Für auf dem abgeschlossenen Intervall $[a,b]$ stückweise stetige Funktionen $f, g : [a,b] \to \mathbb{R}$ erhalten wir:

$$\int_a^b (\alpha f(x) + \beta g(x))dx = \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx, \quad \alpha, \beta \in \mathbb{R} \tag{3.167}$$

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx \quad (a \le c \le b) \tag{3.168}$$

$$f(x) \le g(x) \text{ für alle } a \le x \le b \Longrightarrow \int_a^b f(x)dx \le \int_a^b g(x)dx. \tag{3.169}$$

Für eine stetige Funktion $f : [a,b] \to \mathbb{R}$ gilt:

$$\left| \int_a^b f(x)dx \right| \le \int_a^b |f(x)|dx \tag{3.170}$$

und aus $m \le f(x) \le M$ für alle $x \in [a,b]$ folgt $m(b-a) \le \int_a^b f(x)dx \le M(b-a)$.

Mit der letzten Doppelungleichung läßt sich der folgende Satz beweisen:

Satz 3.43 (Mittelwertsatz der Integralrechnung)
Sind die Funktionen $f, g : \mathbb{D} \to \mathbb{R}$ mit $[a,b] \subseteq \mathbb{D}$ auf $[a,b]$ stetig und $g(x) \ge 0$ für alle $x \in [a,b]$, dann gibt es mindestens ein $\xi \in [a,b]$ mit

$$\int_a^b f(x) \cdot g(x)dx = f(\xi) \int_a^b g(x)dx. \tag{3.171}$$

Der zu Beginn dieses Kapitels betrachtete Weg, aus der Ableitung f' einer Funktion f die Funktion f zu konstruieren, führt zu einer einfachen Methode, Integrale zu berechnen.

Definition 3.44 (Stammfunktion)
Man nennt eine auf einem offenen Intervall I differenzierbare Funktion F eine Stammfunktion einer Funktion $f : I \to \mathbb{R}$, falls $F'(x) = f(x)$ für alle $x \in I$ gilt.

Offensichtlich sind Stammfunktionen zu f nicht eindeutig, da mit F auch $F + c$, $c \in \mathbb{R}$, eine Stammfunktion von f ist:

$$(F(x) + c)' = F'(x) + 0 = f(x), \quad x \in I. \tag{3.172}$$

Die bereits angedeutete Beziehung zwischen Differentiation und Integration wird durch den folgenden Satz spezifiziert.

Satz 3.45 (Hauptsatz der Differential- und Integralrechnung)
Sei $f : I \to \mathbb{R}$ eine auf dem Intervall I stetige Funktion und $a, b \in I$, $a < b$, dann gilt:

a) Die durch $F_a : I \to \mathbb{R}$, $x \mapsto \int_a^x f(t)dt$ definierte Funktion ist eine Stammfunktion von f, d.h.

$$\frac{d}{dx}\left(\int_a^x f(t)dt\right) = f(x) \quad \text{für alle } x \in (a, b). \tag{3.173}$$

Man nennt F_a eine Integralfunktion. Jede andere Stammfunktion von f hat die Form

$$F : I \to \mathbb{R}, \quad x \mapsto F_a(x) + c, \quad c \in \mathbb{R}. \tag{3.174}$$

b) Sei F eine Stammfunktion von f, so gilt:

$$\int_a^b f(x)dx = F(b) - F(a) =: F(x)|_a^b. \tag{3.175}$$

Der Unterschied zwischen Integral- und Stammfunktion wird im folgenden Beispiel deutlich:
Sei $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x$, so ist die Menge aller Stammfunktionen F gegeben durch $F : \mathbb{R} \to \mathbb{R}$, $x \mapsto \frac{1}{2}x^2 + c$, $c \in \mathbb{R}$. Eine Integralfunktion $F_a : \mathbb{R} \to \mathbb{R}$, $x \mapsto \int_a^x t dt$ für $a \in \mathbb{R}$ ist von der Form $x \mapsto \frac{1}{2}x^2 - \frac{1}{2}a^2$. Somit sind alle Stammfunktionen mit $c \leq 0$ auch Integralfunktionen.

Definition 3.46 (Unbestimmtes Integral)
Die Menge aller Stammfunktionen von f wird mit $\int f(x)dx$ bezeichnet und heißt unbestimmtes Integral.

Beispiel(e) 3.47
mit $I = \mathbb{R}$ ($c \in \mathbb{R}$)

- $\int \cos(x)dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \sin(x) + c$
- $\int x^n dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{x^{n+1}}{n+1} + c$
- $\int \sinh(x)dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \cosh + c$
- $\int \cosh(x)dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \sinh(x) + c$
- $\int_0^{\frac{\pi}{2}} \sin(x)dx = -\cos(x)|_0^{\frac{\pi}{2}} = -\cos(\frac{\pi}{2}) + \cos(0) = 1.$

Häufig wird die Konstante c weggelassen.
Im Folgenden betrachten wir Integrationsregeln, die direkt aus dem Hauptsatz der Differential- und Integralrechnung folgen:
1. Linearität:

$$\int (\alpha f(x) + \beta g(x))dx = \alpha \int f(x)dx + \beta \int g(x)dx. \tag{3.176}$$

Beispiel:

$$\int \sum_{i=0}^{n} a_i x^i dx : \mathbb{R} \to \mathbb{R} \quad x \mapsto \sum_{i=0}^{n} \frac{a_i}{i+1} x^{i+1} + c. \tag{3.177}$$

2. Partielle Integration:

$$\int u'(x) \cdot v(x)dx = u \cdot v - \int u(x) \cdot v'(x)dx. \tag{3.178}$$

Beispiel(e) 3.48

- $\int \underbrace{x}_{v} \cdot \underbrace{e^x}_{u'} dx : \mathbb{R} \to \mathbb{R} \quad x \mapsto x \cdot e^x - (\underbrace{e^x + c}_{\int u(x)\cdot v'(x)dx}) = (x-1)e^x + \bar{c}$
- $\int \underbrace{x}_{v} \cdot \underbrace{\cos(x)}_{u'} dx : \mathbb{R} \to \mathbb{R} \quad x \mapsto x \cdot \sin(x) + (\underbrace{\cos(x) + c}_{\int u(x)\cdot v'(x)dx})$

3. Substitutionsmethode:

$$\int f(g(x)) \cdot g'(x)dx : \mathbb{D} \to \mathbb{R}, \quad x \mapsto F(g(x)) + c, \quad \text{wobei } F \text{ eine Stammfunktion von } f \text{ ist.} \tag{3.179}$$

Für das bestimmte Integral gilt somit:

$$\int_a^b f(g(x))g'(x)dx = F(g(x))|_a^b = F(g(b)) - F(g(a)) = \int_{g(a)}^{g(b)} f(t)dt. \tag{3.180}$$

Gemäß der letzten Gleichungen gibt es zwei mögliche Anwendungen der Substitutionsregel:
1. Version: Berechne $\int f(g(x))g'(x)dx$ durch $F \circ g + c$.

Beispiel(e) 3.49

- $\int \frac{g'(x)}{g(x)} dx = \int f(g(x))g'(x)dx$ mit $f : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}$, $x \mapsto \frac{1}{x}$, $g > 0$,
  also $\int \frac{g'(x)}{g(x)} dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \ln(g(x)) + c.$
- $\int \frac{(\ln(x))^2}{x} dx = \int f(g(x))g'(x)dx$ mit $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$, $g : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}$, $x \mapsto \ln(x)$,
  also: $\int \frac{(\ln(x))^2}{x} dx : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}, \quad x \mapsto \frac{(\ln(x))^3}{3} + c,$
- $\int e^{\sin(x)} \cdot \cos(x)dx : \mathbb{R} \to \mathbb{R}$, $x \mapsto e^{\sin(x)} + c$.

2. Version: Sei $g : \mathbb{D}_1 \to \mathbb{R}$ bijektiv und stetig differenzierbar, $f : \mathbb{D}_2 \to \mathbb{R}$ stückweise stetig und $g(\mathbb{D}_1) \subseteq \mathbb{D}_2$:

$$\int f(x)dx = H \circ g^{-1} + c, \quad \text{wobei } H \text{ Stammfunktion von } (f \circ g) \cdot g'. \tag{3.181}$$

„ersetze x durch $g(t)$, erweitere mit $g'(t)$ und ersetze in den Stammfunktionen t durch $g^{-1}(x)$“.

Beispiel(e) 3.50

- Berechne:

$$\int \frac{e^{3x}}{e^{2x}-1}dx : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}.$$

Wähle: $g = \ln$:

$$\int \frac{t^3}{t^2-1}\frac{1}{t}dt = \int \frac{t^2}{t^2-1}dt =$$

$$= \int \left(1 + \frac{1}{t^2-1}\right) dt : (1,\infty) \to \mathbb{R}, \quad t \overset{\text{Formelsammlung}}{\overset{\downarrow}{\mapsto}} t + \frac{1}{2}\ln\left(\frac{t-1}{t+1}\right) + c,$$

also:

$$\int \frac{e^{3x}}{e^{2x}-1}dx : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}, \quad x \mapsto e^x + \frac{1}{2}\ln\left(\frac{e^x-1}{e^x+1}\right) + c.$$

- Berechne:

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

Wähle $g = \sinh$:

$$\int \frac{1}{\sqrt{\underbrace{\sinh(t)^2+1}_{=\cosh^2(t)}}} \cdot \cosh(t)dt =$$

$$= \int \frac{\cosh(t)}{\cosh(t)}dt : \mathbb{R} \to \mathbb{R}, \quad t \mapsto t + c,$$

also:

$$\int \frac{1}{\sqrt{x^2+1}}dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \operatorname{arsinh}(x) + c.$$

4. Die Integration rationaler Funktionen

Für die Integration rationaler Funktionen ist deren Zerlegung in eine Summe von Partialbrüchen wichtig. Sei $f : \mathbb{D} \to \mathbb{R}$, $x \mapsto \frac{p(x)}{q(x)}$ mit Polynomen p, q ohne gemeinsamen Faktor, Grad $p <$ Grad q. Sei ferner $q : \mathbb{R} \to \mathbb{R}$, $x \mapsto c(x-b_1)^{k_1}(x-b_2)^{k_2} \cdot \ldots \cdot (x-b_r)^{k_r} \cdot q_1(x)^{l_1} \cdot \ldots \cdot q_s(x)^{l_s}$ mit den paarweise verschiedenen reellen Nullstellen b_i der Vielfachheit k_i und verschiedenen quadratischen Polynomen q_j, die in $\mathbb{R}$ keine Nullstelle haben. Dann existieren reelle Zahlen
$A_1^{(1)}, \ldots, A_1^{(k_1)}, \ldots, A_r^{(1)}, \ldots, A_r^{(k_r)}, B_1^{(1)}, \ldots, B_1^{(l_1)}, \ldots, B_s^{(1)}, \ldots, B_s^{(l_s)}$,
$C_1^{(1)}, \ldots, C_1^{(l_1)}, \ldots, C_s^{(1)}, \ldots, C_s^{(l_s)}$ mit:

$$\frac{p}{q} : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \sum_{i=1}^{r}\sum_{j=1}^{k_i} \frac{A_i^{(j)}}{(x-b_i)^j} + \sum_{i=1}^{s}\sum_{j=1}^{l_i} \frac{B_i^{(j)} + C_i^{(j)}x}{(q_i(x))^j}. \tag{3.182}$$

(Partialbruchzerlegung)

Beispiel(e) 3.51

$$f : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \frac{x^2+x+1}{(x-1)^3(x-2)} = \frac{A_1^{(1)}}{(x-1)} + \frac{A_1^{(2)}}{(x-1)^2} + \frac{A_1^{(3)}}{(x-1)^3} + \frac{A_2^{(1)}}{(x-2)}$$

Koeffizientenvergleich:

$$A_1^{(1)}(x-1)^2(x-2) + A_1^{(2)}(x-1)(x-2) + A_1^{(3)}(x-2) + A_2^{(1)}(x-1)^3 = x^2+x+1$$

$$\underbrace{(A_1^{(1)} + A_2^{(1)})}_{=0} x^3 + \underbrace{(-4A_1^{(1)} + A_1^{(2)} - 3A_2^{(1)})}_{1} x^2 +$$

$$+ \underbrace{(5A_1^{(1)} - 3A_1^{(2)} + A_1^{(3)} + 3A_2^{(1)})}_{1} x + \underbrace{(-2A_1^{(1)} + 2A_1^{(2)} - 2A_1^{(3)} - A_2^{(1)})}_{1} =$$

$$= x^2+x+1 \Longrightarrow A_1^{(1)} = -7, \quad A_1^{(2)} = -6, \quad A_1^{(3)} = -3, \quad A_2^{(1)} = 7.$$

Hat man nun eine rationale Funktion f zu integrieren, so schreibt man zunächst f in der Form (Polynomdivision):

$$f : \mathbb{D} \to \mathbb{R}, \quad x \mapsto= g(x) + \frac{p(x)}{q(x)}, \tag{3.183}$$

wobei g, p, q Polynome sind und Grad $p <$ Grad q. Mit der Partialbruchzerlegung für $\frac{p}{q}$ sind somit lediglich Funktionen der Form $x \mapsto \sum_{i=0}^{n} a_i x_i$, $x \mapsto \frac{A}{(x-b)^k}$, $x \mapsto \frac{B+Cx}{q_i(x)^m}$, k, $m \in \mathbb{N}$ zu integrieren, wobei q_i ein quadratisches Polynom ohne reelle Nullstelle ist. Es gilt:

- $$\int \frac{1}{(x-b)} dx : \mathbb{R} \setminus \{b\} \to \mathbb{R}, \quad x \mapsto \ln(|x-b|) + c \tag{3.184}$$

- $$\int \frac{1}{(x-b)^k} dx : \mathbb{R} \setminus \{b\} \to \mathbb{R}, \quad x \mapsto -\frac{1}{k-1} \frac{1}{(x-b)^{k-1}} + c, \quad k > 1 \tag{3.185}$$

- $$\int \frac{1}{x^2+\alpha x+\beta} dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{2}{\sqrt{4\beta-\alpha^2}} \arctan\left(\frac{2x+\alpha}{\sqrt{4\beta-\alpha^2}}\right) + c \tag{3.186}$$

  wobei $\alpha^2 - 4\beta < 0$

- $$\int \frac{1}{(x^2+\alpha x+\beta)^k} dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{2x+\alpha}{(k-1)(4\beta-\alpha^2)(x^2+\alpha x+\beta)^{k-1}} + \frac{2(2k-3)}{(k-1)(4\beta-\alpha^2)} s(x) + c, \tag{3.187}$$

  wobei s eine Stammfunktion von

  $$\mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{(x^2+\alpha x+\beta)^{k-1}}$$

  darstellt ($\alpha^2 - 4\beta < 0$, $k > 1$).

- $$\int \frac{ax+b}{(x^2+\alpha x+\beta)^k}dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -\frac{a}{2(k-1)(x^2+\alpha x+\beta)^{k-1}} + \left(b - \frac{a\alpha}{2}\right) s(x) + c, \tag{3.188}$$

  wobei s eine Stammfunktion von

  $$\mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{(x^2+\alpha x+\beta)^k}$$

  darstellt ($\alpha^2 - 4\beta < 0$, $k > 1$).

- $$\int \frac{ax+b}{(x^2+\alpha x+\beta)}dx : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{a}{2}\ln(|x^2+\alpha x+\beta|) + \left(b - \frac{a\alpha}{2}\right) s(x) + c, \tag{3.189}$$

  wobei s eine Stammfunktion von

  $$\mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{(x^2+\alpha x+\beta)}$$

  darstellt ($\alpha^2 - 4\beta < 0$, $k > 1$).

Falls $\alpha^2 - 4\beta \geq 0$, hat $x \mapsto x^2 + \alpha x + \beta$ reelle Nullstellen.

Beispiel(e) 3.52

$$\begin{aligned} & \int \frac{3x^5 - 2x^4 + 4x^3 + 4x^2 - 7x + 6}{(x-1)^2(x^2+1)^2}dx = \\ = & \int \left(\frac{1}{x-1} + \frac{2}{(x-1)^2} + \frac{2x+1}{x^2+1} + \frac{4}{(x^2+1)^2}\right) dx : \mathbb{R} \setminus \{1\} \to \mathbb{R} \\ x \quad \mapsto & \quad \ln(|x-1|) - \frac{2}{x-1} + \ln(x^2+1) + \arctan(x) + \frac{2x}{x^2+1} + 2\arctan(x) + c. \end{aligned}$$

Ist eine Funktion $f : [a, b) \to \mathbb{R}$ auf jedem abgeschlossenen Intervall $[a, c]$, $a < c < b$, $b \in \mathbb{R} \cup \{\infty\}$ stückweise stetig, so heißt das Integral

$$\int_a^b f(x)dx := \lim_{c \to b-} \int_a^c f(x)dx, \tag{3.190}$$

bzw.

$$\int_a^\infty f(x)dx := \lim_{c \to \infty} \int_a^c f(x)dx \tag{3.191}$$

uneigentliches Integral. Analog definiert man für $f : (a, b] \to \mathbb{R}$:

$$\int_a^b f(x)dx := \lim_{c \to a+} \int_c^b f(x)dx \tag{3.192}$$

bzw.

$$\int_{-\infty}^b f(x)dx := \lim_{c \to -\infty} \int_c^b f(x)dx. \tag{3.193}$$

Beispiel(e) 3.53

- $\int_1^\infty \frac{1}{x}dx = \lim_{c\to\infty}\int_1^c \frac{1}{x}dx = \lim_{c\to\infty}\ln(c) = \infty$
- $\int_0^1 \frac{1}{x}dx = \lim_{c\to 0+}\int_c^1 \frac{1}{x}dx = \lim_{c\to 0+}(-\ln(c)) = \infty$
- $\int_1^\infty \frac{1}{x^\alpha}dx = \begin{cases}\frac{1}{\alpha-1} & \text{falls } \alpha > 1\\ \infty & \text{falls } \alpha \le 1\end{cases}$
- $\int_0^1 \frac{1}{x^\alpha}dx = \begin{cases}\infty & \text{falls } \alpha \ge 1\\ \frac{1}{1-\alpha} & \text{falls } \alpha < 1\end{cases}$.

Ein an beiden Grenzen uneigentliches Integral ist definiert durch

$$\int_a^b f(x)dx = \lim_{u\to a+}\int_u^c f(x)dx + \lim_{v\to b-}\int_c^v f(x)dx \quad \text{mit } a < c < b. \tag{3.194}$$

Die Grenzwerte auf der rechten Seite sind voneinander unabhängig!
Es gilt also: $\int_{-\infty}^{\infty} x dx$ ist nicht definiert!

Beispiel(e) 3.54

- $$\begin{aligned}\int_{-\infty}^{\infty}\frac{1}{1+x^2}dx &= \lim_{a\to-\infty}\int_a^0\frac{1}{1+x^2}dx + \\ &\quad + \lim_{b\to\infty}\int_0^b\frac{1}{1+x^2}dx = \\ &= \arctan(0) - \lim_{a\to-\infty}\arctan(a) + \lim_{b\to\infty}\arctan(b) - \\ &\quad - \arctan(0) = \frac{\pi}{2} + \frac{\pi}{2} = \pi.\end{aligned}$$

- $\Gamma : (0,\infty) \to \mathbb{R},\ x \mapsto \int_0^\infty e^{-t}t^{x-1}dt$
  Die Gammafunktion Γ ist über ein an beiden Seiten uneigentliches Integral definiert. Es gilt:
  $$\Gamma(x+1) = x\cdot\Gamma(x), \quad \Gamma(n+1) = n! \text{ für alle } n \in \mathbb{N}. \tag{3.195}$$

Anwendung der Integrationstheorie:
Funktionsapproximation:
Die bereits betrachtete Taylorformel für eine auf dem offenen Intervall I $(n+1)$-mal stetig diffe-

renzierbare Funktion

$$f : I \to \mathbb{R}, \quad x \mapsto \sum_{i=0}^{n} \frac{1}{i!} f^{(i)}(a)(x-a)^i + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1} \quad (a \in I) \tag{3.196}$$

kann auch mit Hilfe der Integrationstheorie in der Form

$$f : I \to \mathbb{R}, \quad x \mapsto \sum_{i=0}^{n} \frac{1}{i!} f^{(i)}(a)(x-a)^i + \underbrace{\frac{1}{n!} \int_a^x (x-t)^n f^{(n+1)}(t) dt}_{R_{n+1}(x,a)} \tag{3.197}$$

geschrieben werden. Diese Form des Restgliedes ist häufig praktikabler als die differentielle Form. Weitere Anwendungen der Integrationstheorie folgen in den folgenden Kapiteln.

Kapitel 4

Gewöhnliche Differentialgleichungen

Es sei $y : \mathbb{R}^+ \to \mathbb{R}, \quad t \mapsto y(t)$ eine Funktion mit der Interpretation: $y(t)$ ist die Menge eines zum Zeitpunkt t vorhandenen radioaktiven Materials. Ein radioaktiver Zerfallsprozess kann mathematisch durch folgende Gleichung beschrieben werden:

$$\dot{y}(t) = a \cdot y(t) \qquad \text{mit} \qquad a < 0\,. \tag{4.1}$$

Hier bezeichnen wir mit $\dot{y}$, wie in der Physik anstelle von y' üblich, die Ableitung der zeitabhängigen Funktion y. Gleichung (4.1) sagt aus, dass die Abnahme $\dot{y}(t)$ des Materials zum Zeitpunkt t abhängt von der zu diesem Zeitpunkt noch vorhandenen Restmenge an Material. Gleichung (4.1) stellt also einen funktionalen Zusammenhang her zwischen einer (gesuchten) Funktion und ihren Ableitungen und heißt deshalb **Differentialgleichung**.

In Gleichung (4.1) tritt nur die erste Ableitung von y auf — es handelt sich um eine Differentialgleichung *erster Ordnung*. Die allgemeinste Differentialgleichung erster Ordnung für eine Funktion $y : t \mapsto y(t)$ hat die Form

$$F(t, y(t), \dot{y}(t)) = 0\,, \tag{4.2}$$

wobei F im Fall des Beispiels (4.1) die durch $F(u, v, w) = w - av$ gegebene lineare Funktion ist. Eine Funktion $y : t \mapsto y(t)$ heißt **Lösung** der Differentialgleichung (4.2), wenn y auf einem Intervall $I \subseteq \mathbb{R}$ differenzierbar ist und die Gleichung (4.2) für alle $t \in I$ erfüllt. Hier wie in allen folgenden Abschnitten verstehen wir unter einem Intervall $I \subseteq \mathbb{R}$ immer eine Menge der Form $I = (a, b)$ (offenes Intervall), $I = [a, b]$ (abgeschlossenes Intervall), $I = (a, b]$ oder $I = [a, b)$ (halboffenes Intervall) mit $a < b$. Ist I bei a offen, dann ist auch $a = -\infty$ erlaubt und ist I bei b offen, dann ist auch $b = \infty$ erlaubt. An abgeschlossenen Intervallrändern sind Stetigkeits- und Differenzierbarkeitsforderungen stets einseitig zu verstehen.

Im Beispiel (4.1) kann man sofort nachprüfen, dass die Funktionen

$$y_C : t \mapsto y_C(t) = C \cdot e^{at}\,, \quad C \in \mathbb{R},$$

auf $I = \mathbb{R}$ definierte Lösungen der Differentialgleichung ist — es sind dies auch die einzigen Lösungen, wie wir bald sehen werden. Kennt man zusätzlich die Menge y_0 des zum Zeitpunkt t_0 vorhandenen Materials, kann man die unbestimmte Konstante C aus der Gleichung $y_0 = C \exp(at_0)$ bestimmen und bekommt eine *eindeutige* Lösung der Differentialgleichung. Man spricht von einem **Anfangswertproblem**, wenn neben einer Differentialgleichung noch der Funktionswert vorgegeben wird, den die Lösung an einer Stelle annehmen soll.

Betrachten wir ein weiteres Beispiel, bei dem es nicht mehr um eine skalare, sondern um eine vektorwertige Funktion (eine Kurve) $\mathbf{r} : \mathbb{R}_0^+ \to \mathbb{R}^3$ geht: Eine Aufgabenstellung der Mechanik

besteht darin, für Zeitpunkte $t \geq 0$ die Bahnkurve $\mathbf{r}(t) \in \mathbb{R}^3$ eines Massenpunkts zu bestimmen, wenn die Beschleunigung $\mathbf{a}(t) = \ddot{\mathbf{r}}(t)$ vorgegeben ist. (Hier ist mit $\ddot{\mathbf{r}}$ die komponentenweise zu verstehende 2. Ableitung von $\mathbf{r}$ gemeint). Es ist also eine Funktion $\mathbf{r}$ gesucht, deren 2. Ableitung bekannt ist. In diesem einfachen Fall erhält man $\mathbf{r}$ durch 2-malige Integration der Komponenten von $\mathbf{a}$. Dabei entsteht ein unbestimmter linearer Term der Form $\mathbf{u} + t \cdot \mathbf{v}$ mit $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$, der wiederum durch Zusatzbedingungen festgelegt werden kann, am einfachsten durch die Anfangsbedingungen $\mathbf{u} = \mathbf{r}(0)$ (bekannter Ort des Massenpunkts zum Zeitpunkt 0) und $\mathbf{v}(0) = \dot{\mathbf{r}}(0)$ (bekannte Anfangsgeschwindigkeit des Massenpunkts).

Im allgemeinen wird die Beschleunigung $\mathbf{a}$ nicht explizit gegeben sein, man weiß jedoch, dass sie durch eine Kraft $\mathbf{F}$ verursacht wird, so dass nach dem 2. Newtonschen Bewegungsgesetz (konstante Masse m vorausgesetzt)

$$\mathbf{F} = m\ddot{\mathbf{r}} = m\mathbf{a}$$

gilt. Typischerweise ist die Kraft eine Funktion der Gestalt

$$\mathbf{F} = \mathbf{F}(\mathbf{r}, \dot{\mathbf{r}}, t)$$

(zum Beispiel wirken Gravitationskräfte in Abhängigkeit vom Ort $\mathbf{r}$ des Massenpunkts oder Reibungskräfte in Abhängigkeit von der Geschwindigkeit $\dot{\mathbf{r}}$) und damit lautet die **Bewegungsgleichung** für die Funktion $\mathbf{r}$:

$$m\ddot{\mathbf{r}} = \mathbf{F}(\mathbf{r}, \dot{\mathbf{r}}, t) \ .$$

Die Bewegungsgleichung besteht hier aus einem ganzen System von Differentialgleichungen, stellt also wiederum einen funktionalen Zusammenhang zwischen (den Komponenten) einer (gesuchten) Funktion und deren Ableitungen dar. Bewegungsgleichungen oder allgemeiner Energieerhaltungssätze der Physik sind der wichtigste Grund für das Auftreten von Differentialgleichungen.

Wir geben nun die formale Definition zunächst von skalaren und im Anschluss daran von Systemen von Differentialgleichungen an. Dabei kehren wir wieder zu der gewohnten Schreibweise $y', y'', \ldots$ für erste, zweite und höhere Ableitungen einer Funktion y zurück und nennen die unabhängige Variable wieder x statt t.

4.1 Gewöhnliche Differentialgleichungen n-ter Ordnung

Definition 4.1 (gewöhnliche Differentialgleichungen n-ter Ordnung)
Seien $n \in \mathbb{N}$, $F : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} \subseteq \mathbb{R}^{n+2}$, $f : \mathbb{G} \to \mathbb{R}$, $\mathbb{G} \subseteq \mathbb{R}^{n+1}$.
Eine Bestimmungsgleichung

$$F\left(x, y(x), y'(x), \ldots, y^{(n)}(x)\right) = 0 \quad \text{für alle } x \in I \subseteq \mathbb{R} \text{ offen.} \tag{4.3}$$

für eine gesuchte Funktion $y : I \to \mathbb{R}$, $x \mapsto y(x)$ heißt implizite gewöhnliche Differentialgleichung n-ter Ordnung.
Erlaubt F die Gestalt

$$y^{(n)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right), \tag{4.4}$$

so spricht man von einer expliziten gewöhnlichen Differentialgleichung n-ter Ordnung.

Eine n-mal stetig differenzierbare Funktion $y : I \to \mathbb{R}$ heißt explizite Lösung von (4.3) bzw. (4.4), falls

$$\left(x, y(x), y'(x), \ldots, y^{(n)}(x)\right) \in \mathbb{D} \quad \text{für alle } x \in I \text{ und} \tag{4.5}$$

$$F\left(x, y(x), y'(x), \ldots, y^{(n)}(x)\right) = 0 \quad \text{für alle } x \in I \tag{4.6}$$

beziehungsweise

$$\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right) \in \mathbb{G} \quad \text{für alle } x \in I \text{ und} \tag{4.7}$$

$$y^{(n)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right) = 0 \quad \text{für alle } x \in I. \tag{4.8}$$

Eine implizite Lösung ist eine durch $g : \mathbb{B} \to \mathbb{R}$, $\mathbb{B} \subseteq \mathbb{R}^2$ mit

$$g(x, y) = 0 \quad \text{für alle} \quad x, y \in \mathbb{B} \tag{4.9}$$

gegebene Funktion $y : I \to \mathbb{R}$, die die entsprechende Differentialgleichung löst und die wenigstens auf einem Teilintervall $\widetilde{I} \subseteq I$ explizit in der Form $x \mapsto y(x)$ darstellbar ist.

Beispiel(e) 4.2

- $y'(x) = xy^2(x)$ ist eine explizite gewöhnliche Differentialgleichung erster Ordnung. Eine Lösung für $I = \mathbb{R}$ ist
$$y : \mathbb{R} \to \mathbb{R},\ x \mapsto -\frac{2}{1+x^2}.$$
- $\big(y(x) \cdot y^{(5)}(x)\big)^2 + xy^{(2)}(x) + \ln(y(x)) = 0$
ist eine implizite gewöhnliche Differentialgleichung 5.Ordnung. Eine Lösung wäre z.B. $y : \mathbb{R} \to \mathbb{R}$, $x \mapsto 1$.
- $y(x) \cdot y'(x) + x = 0$ ist eine implizite gewöhnliche Differentialgleichung 1.Ordnung mit einer impliziten Lösung $g(x, y) = 0$, wobei
$$g : \mathbb{R}^2 \to \mathbb{R}, \quad (x, y) \to x^2 + y^2 - c, \quad c \geq 0.$$
- Die implizite gewöhnliche Differentialgleichung erster Ordnung
$$(y'(x))^2 + 1 = 0$$
besitzt keine Lösung.

Wie man bereits am Beispiel
$$y''(x) = 0$$
mit den Lösungsfunktionen $y_{a,b} : \mathbb{R} \to \mathbb{R}$, $x \mapsto ax + b$, $a, b \in \mathbb{R}$ sieht, kann die Lösungsmenge einer gewöhnlichen Differentialgleichung mehrparametrige Kurvenscharen enthalten. Man fasst jede r-parametrige Kurvenschar $y_{c_1,\ldots,c_r} : I \to \mathbb{R}$, $c_i \in I_i \subseteq \mathbb{R}$ als eine Lösung mit r freien Parametern (den sogenannten Integrationskonstanten $c_1, \ldots, c_r$) auf. Jede einzelne Lösung, die keine wählbaren Konstanten enthält, wird als spezielle oder partikuläre Lösung bezeichnet. Gehört sie keiner Lösungsschar an, so nennt man sie singuläre Lösung. Eine parameterabhängige Lösung einer gewöhnlichen Differentialgleichung n-ter Ordnung heißt allgemein, wenn sie n frei wählbare Konstanten enthält. Sie heißt vollständig, wenn dadurch sämtliche Lösungen erfasst werden. Entsprechende Begriffe gelten auch für implizite Lösungen.

Beispiel(e) 4.3

- $y'(x) - 5y(x) = 0$ hat auf $I = \mathbb{R}$ die allgemeine Lösung

$$y_c : \mathbb{R} \to \mathbb{R}, \quad x \mapsto c \cdot e^{5x}, \quad c \in \mathbb{R}.$$

  Diese Lösung ist vollständig.

- $y''(x) + 9y(x) - 9 = 0$ besitzt auf $\mathbb{R}$ die spezielle Lösung

$$y : \mathbb{R} \to \mathbb{R}, \quad x \mapsto 1$$

  und die allgemeine Lösung

$$y_{c_1,c_2} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto 1 + c_1 \sin(3x) + c_2 \cos(3x)$$

  mit $c_1, c_2 \in \mathbb{R}$. Diese ist vollständig.

- $|y'(x)| + |y| = 0$ hat keine allgemeine Lösung, sondern nur die singuläre Lösung

$$y : \mathbb{R} \to \mathbb{R}, \quad x \mapsto 0.$$

- $y'(x)^2 - 4xy'(x) + 4y(x) = 0$ ist eine implizite gewöhnliche Differentialgleichung erster Ordnung mit der allgemeinen Lösung

$$y_c : \mathbb{R} \to \mathbb{R}, \quad x \mapsto 2cx - c^2, \quad c \in \mathbb{R}$$

  und der singulären Lösung $y : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$.

Fügt man einer gewöhnlichen Differentialgleichung zusätzliche Bedingungen hinzu, um die Integrationskonstanten festzulegen, so erhält man als einfachsten Fall Anfangswertprobleme:

$$y^{(n)}(x) = f\left(x, y(x), \ldots, y^{(n-1)}(x)\right) \tag{4.10}$$

mit vorgegebenen Werten $y(x_0) = y_0$, $y'(x_0) = y_1$, ..., $y^{(n-1)}(x_0) = y_{n-1}$, wobei

$$(x_0, y_0, y_1, \ldots, y_{n-1})$$

im Definitionsbereich von f liegen muss.
Die Bezeichnung „Anfangswertproblem“ stammt aus technischen Anwendungen, in denen man Zeitfunktionen $x : [a, b] \to \mathbb{R}$ betrachtet, die durch

$$x^{(n)}(t) = f\left(t, x(t), \dot{x}(t), \ldots, x^{(n-1)}(t)\right) \tag{4.11}$$

mit Anfangszustand $x(t_0) = 0$, $\dot{x}(t_0) = \dot{x}_0$,...,$x^{(n-1)}(t_0) = x_0^{(n-1)}$, $t_0 \in]a, b[$ gegeben sind ($(t_0, x_0, \ldots, x_0^{(n-1)})$ im Definitionsbereich von f).

Beispiel(e) 4.4
Für ein lineares Pendel modelliert durch

$$\ddot{x}(t) + \omega^2 x(t) = 0, \quad x(0) = x_0, \dot{x}(0) = v_0 \tag{4.12}$$

gibt es die eindeutige Lösung:

$$x : \mathbb{R}_0^+ \to \mathbb{R}, \quad t \mapsto x_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t), \quad \omega \neq 0. \tag{4.13}$$

Ein Anfangswertproblem heißt lokal lösbar, falls es ein $\epsilon > 0$ gibt, sodass auf $I = (x_0 - \epsilon, x_0 + \epsilon)$ eine Lösung $y : (x_0 - \epsilon, x_0 + \epsilon) \to \mathbb{R}$ der gewöhnlichen Differentialgleichung

$$y^{(n)}(x) = f\left(x, y(x), \ldots, y^{(n-1)}(x)\right) \tag{4.14}$$

mit den Anfangsbedingungen

$$y(x_0) = y_0,\, y'(x_0) = y_1, \ldots, y^{(n-1)}(x_0) = y_{n-1}, \tag{4.15}$$

existiert ($(x_0, y_0, y_1, \ldots, y_{n-1})$ im Definitionsbereich von f). Diese Lösung heißt lokale Lösung.
Ein Anfangswertproblem heißt „sachgemäß gestellt“, falls

- die Existenz und Eindeutigkeit einer lokalen Lösung
- die stetige Abhängigkeit der lokalen Lösung von den Anfangswerten

gewährleistet ist.
Für sachgemäß gestellte Probleme stellen sich zusätzliche Fragen:

- globale Existenz,
- explizite Bestimmung einer Lösung.

Beispiel(e) 4.5
Jede kubische Funktion $y : \mathbb{R} \to \mathbb{R}$, $x \mapsto (x - c)^3, c \in \mathbb{R}$, genügt der gewöhnlichen Differentialgleichung

$$y'(x) = 3\sqrt[3]{y^2(x)}. \tag{4.16}$$

Da auch $y : \mathbb{R} \to \mathbb{R}$, $x \mapsto 0$ eine singuläre Lösung dieser gewöhnlichen Differentialgleichung ist, hat das Anfangswertproblem

$$y'(x) = 3\sqrt[3]{y^2(x)}, \quad y(0) = 0 \tag{4.17}$$

zumindest zwei Lösungen: $\hat{y} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto x^3$ und $y : \mathbb{R} \to \mathbb{R}, \quad x \mapsto 0$.

Im Folgenden betrachten wir spezielle Klassen gewöhnlicher Differentialgleichungen. Eine gewöhnliche Differentialgleichung der Form

$$y'(x) = h(x)g(y(x)) \tag{4.18}$$

mit stetigen, auf Intervallen $I \subseteq \mathbb{R}$ und $J \subseteq \mathbb{R}$ definierten Funktionen $h : I \to \mathbb{R}$ und $g : J \to \mathbb{R}$ heißt trennbar. Eine Lösung der obigen trennbaren gewöhnlichen Differentialgleichung ist für $\xi \in I, \eta \in J$ und dem Anfangswert $y(\xi) = \eta$ implizit gegeben durch

$$\int_{\eta}^{y(x)} \frac{1}{g(t)} dt = \int_{\xi}^{x} h(t) dt, \tag{4.19}$$

falls $g(t) \neq 0$ für alle $t \in J$.
Ist $g(\eta) = 0$ für ein $\eta \in J$, dann ist die Funktion $y : I \to \mathbb{R}$, $x \mapsto \eta$ eine konstante Lösung des Anfangswertproblems

$$y'(x) = h(x)g(y(x)), \quad y(\xi) = \eta, \quad \xi \in I. \tag{4.20}$$

Beispiel(e) 4.6

- $y'(x) = y^2(x), \quad y(0) = 1. \qquad \int_1^{y(x)} \frac{1}{t^2} dt = \int_0^x dt \;\Leftrightarrow\; -\frac{1}{y(x)} + 1 = x,$
  also: $y : (-\infty, 1) \to \mathbb{R},\ x \mapsto \frac{1}{1-x}$.
- $y'(x) = \frac{1-y(x)}{x}, \quad y(\xi) = \eta,\ \eta \neq 1,\ \xi \neq 0.$

$$\int_\eta^{y(x)} \frac{1}{1-t} dt = \int_\xi^x \frac{1}{t} dt \quad \Leftrightarrow \quad -\ln(|1-y(x)|) + \ln(|1-\eta|) = \ln(|x|) - \ln(|\xi|), \quad (4.21)$$

also: $|1 - y(x)| = \left|\frac{\xi(1-\eta)}{x}\right|$.
Für $\xi > 0,\ |\eta| < 1$ folgt: $y : (0, \infty) \to \mathbb{R},\ x \mapsto 1 - \frac{\xi(1-\eta)}{x}$.

Lineare gewöhnliche Differentialgleichungen erster Ordnung sind von der Form

$$y'(x) + a(x)y(x) = f(x) \tag{4.22}$$

mit Funktionen definiert auf einem Intervall $I \subseteq \mathbb{R}$.
Diese gewöhnliche Differentialgleichung heißt homogen, falls $f \equiv 0$; f wird Störfunktion genannt. Ist die Funktion a auf I stetig und ist $f \equiv 0$, so erhält man durch

$$y : I \to \mathbb{R}, \quad x \mapsto \exp\left(-\int a(x)dx\right), \tag{4.23}$$

eine allgemeine Lösung der gewöhnlichen Differentialgleichung

$$y'(x) + a(x)y(x) = 0 \tag{4.24}$$

(die Konstante c ist in der Menge aller Stammfunktionen $\int a(x)dx$ versteckt). Diese gewöhnliche Differentialgleichung wird als Modell für Wachstumsprozesse gedeutet.

Beispiel(e) 4.7
Für $a(x) \equiv -a$ hat das gewöhnliche Anfangswertproblem

$$y'(x) = a \cdot y(x), \quad y(\xi) = \eta \tag{4.25}$$

die Lösung

$$y : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \eta \cdot e^{a(x-\xi)}. \tag{4.26}$$

Für $a < 0$ handelt es sich um einen Zerfallsprozess, für $a > 0$ um einen Wachstumsprozess.

Für die lineare gewöhnliche Differentialgleichung

$$y'(x) + a(x)y(x) = f(x) \tag{4.27}$$

mit nichtverschwindender Störfunktion f erhält man für stetige Funktionen a, f eine vollständige allgemeine Lösung

$$y : I \to \mathbb{R}, \quad x \mapsto e^{-A(x)} \left(\int_\xi^x e^{A(t)} f(t)dt + c \right), \quad c \in \mathbb{R}, \tag{4.28}$$

wobei $A(x) := \int_{\xi}^{x} a(t)dt$ und $\xi \in I$ beliebig.
Es gilt somit für gewöhnliche Anfangswertprobleme der Form

$$y'(x) + a(x)y(x) = f(x), \quad y(\xi) = \eta : \tag{4.29}$$

- Es existiert eine eindeutige Lösung

$$y : I \to \mathbb{R}, \quad x \mapsto e^{-A(x)} \left(\int_{\xi}^{x} e^{A(t)} f(t)dt + \eta \right), \tag{4.30}$$

  wobei $A(x) = \int_{\xi}^{x} a(t)dt$.

Die vollständige allgemeine Lösung y der inhomogenen gewöhnlichen Differentialgleichung

$$y'(x) + a(x)y(x) = f(x) \tag{4.31}$$

ist gegeben durch

$$y = y_h + y_p, \tag{4.32}$$

wobei y_p eine spezielle (partikuläre) Lösung der obigen gewöhnlichen Differentialgleichung ist (etwa eine Lösung zu fest gewählten Anfangswerten) und y_h die allgemeine Lösung der zugehörigen homogenen gewöhnlichen Differentialgleichung darstellt.

Beispiel(e) 4.8

- Betrachte

$$y'(x) + \frac{1}{x}y(x) = x^3, \quad x > 0 : \tag{4.33}$$

$$\int \frac{1}{t}dt : (0,\infty) \to \mathbb{R},\ x \mapsto \ln(x) + c, \quad y_h : (0,\infty) \to \mathbb{R},\ x \mapsto e^{-\ln(x)-c} = \frac{\hat{c}}{x}. \tag{4.34}$$

Mit $\xi = 1, \eta = 0$ erhält man:

$$y_p : I \to \mathbb{R}, \quad x \mapsto e^{-\ln(x)} \left(0 + \int_1^x e^{\ln(t)} t^3 dt\right) = \frac{x^4}{5} - \frac{1}{5x}. \tag{4.35}$$

Somit ist

$$y : I \to \mathbb{R}, \quad x \mapsto \frac{\tilde{c}}{x} + \frac{x^4}{5} \tag{4.36}$$

die allgemeine Lösung.

- RL-Stromkreis:
Wird in einem Stromkreis eine Spule der Induktivität L und ein Widerstand R in Serie geschaltet an eine Spannungsquelle $U(t)$ angeschlossen, so ergibt sich für die Stromstärke $I(t)$:

$$L \cdot \dot{I}(t) + R \cdot I(t) = U(t). \tag{4.37}$$

$I_h : [0,\infty) \to \mathbb{R},\ t \mapsto c \cdot \exp\left(-\frac{R}{L}t\right)$ (exponentielles Abklingen der Stromstärke nach dem Abschalten der Spannungsquelle).
Allgemeine Lösung:
1. Fall: Gleichspannung $U : [0,\infty) \to \mathbb{R},\ t \mapsto U_0$:

$$I : [0,\infty) \to \mathbb{R},\ t \mapsto \frac{U_0}{R} - \left(\frac{U_0}{R} - I_0\right) \exp\left(-\frac{R}{L}t\right), \tag{4.38}$$

wobei $I_p : [0,\infty) \to \mathbb{R},\ t \mapsto \frac{U_0}{R}$ und $I(0) = I_0$.
2. Fall: Wechselspannung $U : [0,\infty) \to \mathbb{R},\ t \mapsto U_0 \cos(\omega t)$:

$$I : [0,\infty) \to \mathbb{R},\ t \mapsto c \cdot \exp\left(-\frac{R}{L}t\right) + \frac{U_0}{\sqrt{R^2 + \omega^2 L^2}} \cos\left(\omega t - \arctan\left(\frac{\omega L}{R}\right)\right), \tag{4.39}$$

wobei

$$I_p : [0,\infty) \to \mathbb{R},\ t \mapsto \frac{U_0}{\sqrt{R^2 + \omega^2 L^2}} \cos\left(\omega t - \arctan\left(\frac{\omega L}{R}\right)\right). \tag{4.40}$$

In den Anwendungen spielen spezielle gewöhnliche Differentialgleichungen zweiter Ordnung eine wichtige Rolle. Beginnen wir zunächst mit der homogenen linearen Differentialgleichung zweiter Ordnung

$$y''(x) + ay'(x) + by(x) = 0, \quad a, b \in \mathbb{R}. \tag{4.41}$$

Der Ansatz $y : \mathbb{R} \to \mathbb{R},\ x \mapsto e^{\lambda x}$ liefert:

$$\lambda^2 e^{\lambda x} + a\lambda e^{\lambda x} + be^{\lambda x} = 0, \tag{4.42}$$

also:

$$\lambda^2 + a\lambda + b = 0, \quad \lambda_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2}. \tag{4.43}$$

1. Fall: λ_1, λ_2 reell und $\lambda_1 \neq \lambda_2$:
Die vollständige allgemeine Lösung lautet:

$$y : \mathbb{R} \to \mathbb{R},\ x \mapsto c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}. \tag{4.44}$$

2. Fall: λ_1, λ_2 komplex, also $\lambda_1 = \bar{\lambda}_2$:
Die vollständige allgemeine Lösung lautet mit $\lambda_1 = \alpha + \beta i,\ \beta \neq 0$:

$$\begin{aligned} y : \mathbb{R} \to \mathbb{R},\ x &\mapsto c_1 \mathsf{Re}(e^{\lambda_1 x}) + c_2 \mathsf{Im}(e^{\lambda_1 x}) = \\ &= c_1 e^{\alpha x} \cos(\beta x) + c_2 e^{\alpha x} \sin(\beta x) \\ &= c_1 \mathsf{Re}(e^{\lambda_2 x}) - c_2 \mathsf{Im}(e^{\lambda_2 x}). \end{aligned} \tag{4.45}$$

3. Fall: λ_1, λ_2 reell und $\lambda_1 = \lambda_2 = -\frac{a}{2}$:
Ansatz: Wähle $y : \mathbb{R} \to \mathbb{R},\ x \mapsto c(x) e^{\lambda_1 x}$, so folgt:

$$y' : \mathbb{R} \to \mathbb{R},\ x \mapsto c'(x) e^{\lambda_1 x} + c(x) e^{\lambda_1 x} \lambda_1 \tag{4.46}$$

und

$$y'' : \mathbb{R} \to \mathbb{R},\ x \mapsto c''(x) e^{\lambda_1 x} + c'(x) e^{\lambda_1 x} \lambda_1 + c'(x) e^{\lambda_1 x} \lambda_1 + c(x) e^{\lambda_1 x} \lambda_1^2. \tag{4.47}$$

Einsetzen in die gegebene gewöhnliche Differentialgleichung liefert $c''(x) = 0$ und damit:

$$c : \mathbb{R} \to \mathbb{R},\ x \mapsto c_1 + c_2 x, \quad c_1, c_2 \in \mathbb{R} \tag{4.48}$$

sowie

$$y : \mathbb{R} \to \mathbb{R},\ x \mapsto = c_1 e^{-\frac{ax}{2}} + c_2 x e^{-\frac{ax}{2}}. \tag{4.49}$$

Beispiel(e) 4.9

- $y''(x) - 4y'(x) + 4y(x) = 0.$
  Wegen $\lambda_1 = \lambda_2 = 2$ ergibt sich

$$y : \mathbb{R} \to \mathbb{R},\ x \mapsto c_1 e^{2x} + c_2 x e^{2x}. \tag{4.50}$$

- $y''(x) - 6y'(x) + 34y(x) = 0.$
  Wegen $\lambda_1 = 3 + 5i,\ \lambda_2 = 3 - 5i$ ergibt sich

$$y : \mathbb{R} \to \mathbb{R},\ x \mapsto c_1 e^{3x} \cos(5x) + c_2 e^{3x} \sin(5x). \tag{4.51}$$

Für die vollständige allgemeine Lösung der inhomogenen linearen gewöhnlichen Differentialgleichung

$$y''(x) + ay'(x) + by(x) = f(x) \tag{4.52}$$

mit konstanten Koeffizienten $a, b \in \mathbb{R}$ und der Störfunktion $f : I \to \mathbb{R}$, $I \subseteq \mathbb{R}$ ein Intervall, genügt es, die vollständige allgemeine Lösung y_h der homogenen gewöhnlichen Differentialgleichung mit einer partikulären Lösung y_p der inhomogenen gewöhnlichen Differentialgleichung zu kombinieren:

$$y = y_h + y_p. \tag{4.53}$$

Sei nun f stetig und $y_h : I \to \mathbb{R},\ x \mapsto c_1 y_1(x) + c_2 y_2(x)$ mit:

$$y_1(x) y_2'(x) - y_2(x) y_1'(x) \neq 0 \quad \text{für alle } x \in I, \tag{4.54}$$

so erhält man eine partikuläre Lösung der inhomogenen gewöhnlichen Differentialgleichung (etwa zum Anfangswert $y_p(\xi) = 0,\ \xi \in I$):

$$y_p : I \to \mathbb{R},\ x \mapsto -y_1(x) \int_\xi^x \frac{y_2(t) f(t)}{y_1(t) y_2'(t) - y_2(t) y_1'(t)} dt + y_2(x) \int_\xi^x \frac{y_1(t) f(t)}{y_1(t) y_2'(t) - y_2(t) y_1'(t)} dt. \tag{4.55}$$

Diese Formel der partikulären Lösung erhält man, wenn man den Ansatz

$$y_p : I \to \mathbb{R},\, x \mapsto c_1(x)y_1(x) + c_2(x)y_2(x) \tag{4.56}$$

in die inhomogene gewöhnliche Differentialgleichung mit Anfangswert $y_p(\xi) = 0$ einsetzt und die entsprechenden Formeln für $c_1(x)$ und $c_2(x)$ herleitet (Variation der Konstanten).

Beispiel(e) 4.10

$$y''(x) - 6y'(x) + 5y(x) = \ln(x), \quad x > 0. \tag{4.57}$$

$$y_h : (0,\infty) \to \mathbb{R},\, x \mapsto c_1 e^x + c_2 e^{5x}, \tag{4.58}$$

$$\begin{aligned} y_p : (0,\infty) \to \mathbb{R},\, x \quad &\mapsto \quad -e^x \int_\xi^x \frac{e^{5t}\ln(t)}{5e^t e^{5t} - e^{5t}e^t}dt + e^{5x}\int_\xi^x \frac{e^t \ln(t)}{5e^t e^{5t} - e^{5t}e^t}dt = \\ &= \quad -\frac{1}{4}e^x \int_\xi^x e^{-t}\ln(t)dt + \frac{1}{4}e^{5x}\int_\xi^x e^{-5t}\ln(t)dt. \end{aligned} \tag{4.59}$$

Gilt nun **nicht** die Bedingung $y_1(x)y_2'(x) - y_2(x)y_1'(x) \neq 0$, so gibt es verschiedene Möglichkeiten, eine partikuläre Lösung zu berechnen (z. B. Laplace-Transformation).

4.2 Gewöhnliche Differentialgleichungssysteme

In vielen Anwendungen betrachtet man Funktionen $y_1, \ldots, y_n : I \to \mathbb{R}$, deren Ableitungen $y_i'(x)$ neben $y_i(x)$ und x auch von $y_1(x), \ldots, y_{i-1}(x), y_{i+1}(x), \ldots, y_n(x)$ abhängen. Dies führt zu einer vektoriellen gewöhnlichen Differentialgleichung erster Ordnung:

$$(y_1'(x), \ldots, y_n'(x))^\top = \mathbf{f}(x, y_1(x), \ldots, y_n(x)), \tag{4.60}$$

wobei $\mathbf{f} : \mathbb{D} \to \mathbb{R}^n$, $\mathbb{D} \subseteq \mathbb{R}^{n+1}$.

Definition 4.11 (n-dim. gewöhnliches Differentialgleichungssystem)
Seien $n \in \mathbb{N}$, $\mathbf{f} : \mathbb{D} \to \mathbb{R}^n$, $\mathbb{D} \subseteq \mathbb{R}^{n+1}$ und $I \subseteq \mathbb{R}$ ein Intervall, so heißt eine Funktion

$$\mathbf{y} : I \to \mathbb{R}^n, \quad x \mapsto \mathbf{y}(x) \tag{4.61}$$

Lösungskurve des n-dim. gewöhnlichen Differentialgleichungssystems

$$\mathbf{y}'(x) := (y_1'(x), \ldots, y_n'(x))^\top = \mathbf{f}(x, \mathbf{y}(x)) := \mathbf{f}(x, y_1(x), \ldots, y_n(x)), \tag{4.62}$$

falls $(x, \mathbf{y}(x)) \in \mathbb{D}$ und $\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x))$ für alle $x \in I$.
Ist $\mathbf{y}(x_0) = \mathbf{y_0}$, so ist $\mathbf{y}$ eine Lösung des gewöhnlichen Anfangswertproblems

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(x_0) = \mathbf{y_0}. \tag{4.63}$$

Analog zum eindimensionalen Fall heißt $\mathbf{y}$ allgemeine Lösung, falls $\mathbf{y}$ n freie Parameter enthält. Eine allgemeine Lösung heißt vollständig, falls damit alle möglichen Lösungen erfasst werden.

Eine wichtige Klasse von gewöhnlichen Differentialgleichungssystemen erster Ordnung ist durch

eindimensionale gewöhnliche Differentialgleichungen n-ter Ordnung

$$y^{(n)}(x) = a_{n-1}(x)y^{(n-1)}(x) + \ldots + a_1(x)y'(x) + a_0(x)y(x) + b(x) \tag{4.64}$$

mit $a_{n-1}, \ldots, a_0, b : I \to \mathbb{R}$ gegeben, denn es gilt mit

$$\begin{aligned}
\mathbf{f} : \mathbb{D} \quad &\to \quad \mathbb{R}^n, \quad (x, \mathbf{z}) \mapsto \begin{pmatrix} z_2 \\ z_3 \\ \vdots \\ z_n \\ a_{n-1}(x)z_n + \ldots + a_1(x)z_2 + a_0(x)z_1 + b(x) \end{pmatrix} = \\
&= \underbrace{\begin{pmatrix} 0 & 1 & \ldots\ldots\ldots\ldots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & 1 \\ a_0(x) & a_1(x) & \cdots & a_{n-2}(x) & a_{n-1}(x) \end{pmatrix}}_{M(x)} \mathbf{z} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b(x) \end{pmatrix},
\end{aligned} \tag{4.65}$$

$$\mathbf{y}'(x) = M(x) \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(x) \end{pmatrix} \quad \text{und} \quad \mathbf{y} : I \to \mathbb{R}^n,\ x \mapsto \begin{pmatrix} y(x) \\ y'(x) \\ \vdots \\ y^{(n-1)}(x) \end{pmatrix} : \tag{4.66}$$

$$\mathbf{y}'(x) = M(x)\mathbf{y}(x) + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b(x) \end{pmatrix} \tag{4.67}$$

$$\Longleftrightarrow$$

$$y^{(n)}(x) = a_{n-1}(x)y^{(n-1)}(x) + \ldots + a_1(x)y'(x) + a_0(x)y(x) + b(x). \tag{4.68}$$

Beispiel(e) 4.12

- Van der Pol-Differentialgleichung

  $$y''(x) = (\alpha - \beta x^2)y'(x) - y(x) \tag{4.69}$$

  beziehungsweise

  $$\mathbf{y}'(x) = \begin{pmatrix} 0 & 1 \\ -1 & \alpha - \beta x^2 \end{pmatrix} \mathbf{y}(x). \tag{4.70}$$

  Beschreibung der Änderung der Gittervorspannung in Triodenschaltungen.

- Räuber-Beute-Modell:

  $$\begin{aligned} y_1'(x) &= -(\alpha - \beta y_2(x))y_1(x) && (4.71) \\ y_2'(x) &= (\gamma - \delta y_1(x))y_2(x). && (4.72) \end{aligned}$$

  Die Population „Räuber“ ($\hat{=}\, y_1$) lebt von der „Beute“ ($\hat{=}\, y_2$). Die Sterberate der Räuber ohne Beute ($y_2 \equiv 0$) ist α. Die Geburtsrate der Beute ohne Räuber ($y_1 \equiv 0$) ist γ.

Im Folgenden betrachten wir die Existenz und Eindeutigkeit der Lösung gewöhnlicher Differentialgleichungssysteme

Satz 4.13 (Existenz und Eindeutigkeit)
Sei $\mathbf{f} : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n, (x, \mathbf{z}) \mapsto \mathbf{f}(x, \mathbf{z})$, eine Abbildung mit

$$\|\mathbf{f}(x, \mathbf{z}_1) - \mathbf{f}(x, \mathbf{z}_2)\|_2 \leq L\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \quad \text{für } x \in [a,b];\ \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n \tag{4.73}$$

(Lipschitz-Bedingung), die in x stetig ist. Sei ferner $\eta \in \mathbb{R}^n$, $n \in \mathbb{N}$, so hat jedes Anfangswertproblem

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(x_0) = \eta \tag{4.74}$$

mit $x_0 \in (a,b)$ eine eindeutige Lösung $\mathbf{y} : [a,b] \to \mathbb{R}^n$.

Die wichtigsten Systeme gewöhnlicher Differentialgleichungen sind lineare gewöhnliche Differentialgleichungssysteme der Form:

$$\mathbf{y}'(x) = A(x)\mathbf{y}(x) + \tilde{\mathbf{b}}(x) \tag{4.75}$$

mit der Koeffizientenmatrix $A(x) \in \mathbb{R}^{n,n}$ und einer Störfunktion $\tilde{\mathbf{b}}$. Im Folgenden nehmen wir an, dass $A : I \to \mathbb{R}^{n,n}$ und $\tilde{\mathbf{b}} : I \to \mathbb{R}^n$ über einem Intervall $I \subseteq \mathbb{R}$ definiert sind. Für $\tilde{\mathbf{b}} \equiv 0$ auf I spricht man von einem homogenen linearen gewöhnlichen Differentialgleichungssystem.

Beispiel(e) 4.14

$$\mathbf{y}'(x) = \begin{pmatrix} 3 & 2 & x^2 \\ \sin(x) & 0 & 5 \\ x & x^2 & x^3 \end{pmatrix} \mathbf{y}(x) + \begin{pmatrix} \cos(\omega x) \\ 0 \\ \sqrt{x} \end{pmatrix}. \tag{4.76}$$

Für lineare Systeme gewöhnlicher Differentialgleichungen gelten wichtige Eigenschaften:

- Aus

$$\mathbf{y}'(x) = A(x)\mathbf{y}(x) + \tilde{\mathbf{b}}_1(x) \tag{4.77}$$
$$\mathbf{w}'(x) = A(x)\mathbf{w}(x) + \tilde{\mathbf{b}}_2(x) \tag{4.78}$$

  folgt:

$$\mathbf{v} := \alpha\mathbf{y} + \beta\mathbf{w} \tag{4.79}$$

  ist für $\alpha, \beta \in \mathbb{R}$ die Lösung von

$$\mathbf{v}'(x) = A(x)\mathbf{v}(x) + \alpha\tilde{\mathbf{b}}_1(x) + \beta\tilde{\mathbf{b}}_2(x). \tag{4.80}$$

- $\mathbf{y} = \mathbf{y}_h + \mathbf{y}_p$.
- $\{\mathbf{y}(x) : I \to \mathbb{R};\ \mathbf{y}'(x) = A(x)\mathbf{y}(x)\}$ ist ein Vektorraum über $\mathbb{R}$ der Dimension n.
- Hat man n linear unabhängige Lösungen $\mathbf{y}_1, \ldots, \mathbf{y}_n$ des homogenen linearen Differentialgleichungssystems

$$\mathbf{y}'(x) = A(x)\mathbf{y}(x) \tag{4.81}$$

  gefunden, so ist mit $Y : I \to \mathbb{R}^{n,n}$, $x \mapsto (\mathbf{y}_1(x) \ldots \mathbf{y}_n(x)) \in \mathbb{R}^{n,n}$ durch

$$\mathbf{y} : I \to \mathbb{R}^n,\ x \mapsto Y(x) \left(\int_\xi^x Y^{-1}(t)\tilde{\mathbf{b}}(t)dt + c \right), \quad c \in \mathbb{R}^n \tag{4.82}$$

  eine vollständige allgemeine Lösung des inhomogenen gewöhnlichen Differentialgleichungssystems gegeben.

- Ist speziell $A(x) \equiv A$ unabhängig von x, so erhält man mit

$$e^{\bullet A} : I \to \mathbb{R}^{n,n},\ x \mapsto \sum_{k=0}^{\infty} \frac{x^k}{k!} A^k = E + xA + \frac{x^2}{2} A \cdot A + \ldots \tag{4.83}$$

durch

$$\mathbf{y}(x) = e^{xA} c, \quad c \in \mathbb{R}^n \tag{4.84}$$

die vollständige allgemeine Lösung des homogenen Systems linearer gewöhnlicher Differentialgleichungen

$$\mathbf{y}'(x) = A\mathbf{y}(x). \tag{4.85}$$

Für von x abhängige Matrizen $A(x)$ ist die Lösung nur in Spezialfällen möglich.

4.3 Lineare Differentialgleichungssysteme mit konstanten Koeffizienten

Es wird noch einmal der in der Praxis häufig auftretende Fall des linearen DGL-Systems mit konstanten Koeffizienten

$$\mathbf{y}'(x) = A\mathbf{y}(x) + \mathbf{b}(x), \quad A \in \mathbb{R}^{n,n}, \tag{4.86}$$

betrachtet. Anfangswertprobleme zu solchen Systemen kann man günstig mit der Laplace-Transformation berechnen – siehe das spätere Kapitel über Fourier-Analysis.

Die allgemeine Lösung von (4.86) hat die Form $\mathbf{y} = \mathbf{y}_h + \mathbf{y}_p$ mit einer partikulären Lösung $\mathbf{y}_p$ von (4.86) und der allgemeinen Lösung $\mathbf{y}_h$ des homogenen Systems

$$\mathbf{y}' = A\mathbf{y}\ ,$$

die sich theoretisch wie in (4.84) mithilfe der Matrix-Exponentialfunktion angeben lässt. Praktischer ist folgender Zugang: Wenn $\mathbf{v}$ ein Eigenvektor von A zum Eigenwert λ ist, dann ist die Funktion $\mathbf{y}(x) = \exp(\lambda x)\mathbf{v}$ eine Lösung des homogenen Systems, denn

$$\mathbf{y}'(x) = \lambda e^{\lambda x}\mathbf{v} \quad \text{und} \quad A\mathbf{y}(x) = e^{\lambda x} A\mathbf{v} = \lambda e^{\lambda x}\mathbf{v}\ .$$

Als Nullstellen des charakteristischen Polynoms können Eigenwerte und damit auch Eigenvektoren komplex sein. Solange jedoch die Matrix A reellwertig ist, erhält man durch Übergang zum konjugiert Komplexen:

$$A\mathbf{v} = \lambda\mathbf{v} \quad \Longleftrightarrow \quad A\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}\ ,$$

das heißt mit λ und $\mathbf{v}$ ist stets auch $\bar{\lambda}$ ein Eigenwert und $\bar{\mathbf{v}}$ ein Eigenvektor zu A. Mit der komplexen Lösung $\exp(\lambda x)\mathbf{v}$ erhält man die beiden linear unabhängigen reellen Lösungen $\mathrm{Re}(\exp(\lambda x)\mathbf{v})$ und $\mathrm{Im}(\exp(\lambda x)\mathbf{v})$.

Kann man n linear unabhängige Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n$ zu Eigenwerten $\lambda_1, \ldots, \lambda_n$ (die nicht notwendig unterschiedlich sein müssen) finden, so hat man auf diese Art die allgemeine homogene Lösung in der Form

$$\mathbf{y}_h(x) = c_1 e^{\lambda_1 x}\mathbf{v}_1 + \ldots + c_n^{\lambda_n x}\mathbf{v}_n$$

mit beliebigen Konstanten $c_1, \ldots, c_n$ gefunden. In manchen Fällen ist das möglich, so zum Beispiel bei symmetrischen Matrizen, bei denen sogar immer garantiert ist, dass alle Eigenwerte reellwertig sind. Aber auch bei komplexen Eigenwerten lassen sich durch Übergang zu Real- und Imaginärteil immer reelle Lösungen finden.

Beispiel(e) 4.15

$$\mathbf{y}'(x) = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ -1 & 1 & 0 \end{pmatrix} \mathbf{y}(x). \tag{4.87}$$

Hier ergeben sich die Eigenwerte aus

$$\det(A - \lambda E) = -\lambda^3 + 2\lambda - 4 = 0 \quad \Longrightarrow \quad \lambda = -2, \lambda = 1 \pm i$$

und damit die Eigenvektoren

$$\lambda = -2 : (A + 2E)\mathbf{v} = 0 \quad \Longrightarrow \quad \mathbf{v} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

sowie

$$\lambda = 1 + i : (A - (1+i)E)\mathbf{v} = 0 \quad \Longrightarrow \quad \mathbf{v} = \begin{pmatrix} 2 - 2i \\ 2 \\ 1 + i \end{pmatrix}.$$

Drei linear unabhängige komplexe Lösungen sind damit $\mathbf{y}_1(x), \mathbf{y}_2(x)$ und $\mathbf{y}_3(x)$ mit $\mathbf{y}_3(x) = \bar{\mathbf{y}}_2(x)$ und

$$\mathbf{y}_1(x) = e^{-2x} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \quad \mathbf{y}_2(x) = e^{(1+i)x} \begin{pmatrix} 2 - 2i \\ 2 \\ 1 + i \end{pmatrix}.$$

Drei linear unabhängige reelle Lösungen sind $\mathbf{y}_1$, $\tilde{\mathbf{y}}_2 =$ Re $\mathbf{y}_2$ und $\tilde{\mathbf{y}}_3 =$ Im $\mathbf{y}_2$, also

$$\tilde{\mathbf{y}}_2(x) = e^x \left[\cos x \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} - \sin x \begin{pmatrix} -2 \\ 0 \\ 1 \end{pmatrix} \right], \tilde{\mathbf{y}}_3(x) = e^x \left[\sin x \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} + \cos x \begin{pmatrix} -2 \\ 0 \\ 1 \end{pmatrix} \right].$$

Es gibt durchaus Matrizen $A \in \mathbb{R}^{n,n}$, die weniger als n linear unabhängige Eigenvektoren besitzen, etwa

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

die den dreifachen Eigenwert $\lambda = 1$ besitzt, dessen Eigenraum

$$\text{Kern}(A - E) = \left\{ c \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} ; \, c \in \mathbb{R} \right\}$$

jedoch nur die Dimension 1 besitzt. Auch in solchen Fällen ist es jedoch möglich, die allgemeine Lösung von $y' = Ay$ anzugeben. Man benötigt dafür die folgende Verallgemeinerung von Eigenvektoren.

Definition 4.16 (Hauptvektoren)
Ein Vektor $\mathbf{v} \in \mathbb{C}^n$ heißt **Hauptvektor der Stufe** $\ell \in \mathbb{N}$ zum Eigenwert $\lambda \in \mathbb{C}$ von $A \in \mathbb{R}^{n,n}$, wenn

$$(A - \lambda E)^\ell \mathbf{v} = 0 \quad \text{und} \quad (A - \lambda E)^{\ell - 1} \mathbf{v} \neq 0 \, .$$

Beispiel(e) 4.17

(a): Jeder Eigenvektor $\mathbf{v}$ von A ist ein Hauptvektor der Stufe 1, denn

$$(A-\lambda E)\mathbf{v}=0 \quad \text{und} \quad (A-\lambda E)^0\mathbf{v}=\mathbf{v}\neq 0.$$

(b): Die Matrix

$$A=\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\end{pmatrix}$$

hat den dreifachen Eigenwert $\lambda=1$. Der Vektor e_1 ist Eigenvektor (und damit Hauptvektor der Stufe 1). Wegen

$$(A-E)e_2=e_1 \quad \Longrightarrow (A-E)^2e_2=0$$

ist dann e_2 Hauptvektor der Stufe 2. Ebenso ist e_3 wegen $(A-E)e_3=e_1+e_2$ Hauptvektor der Stufe 3.

(c): Allgemein sind für einen Hauptvektor $\mathbf{v}$ der Stufe ℓ zum Eigenwert λ die Vektoren

$$\mathbf{v}, \quad (A-\lambda E)\mathbf{v}, \quad (A-\lambda E)^2\mathbf{v}, \quad \ldots, \quad (A-\lambda E)^{\ell-1}\mathbf{v}$$

linear unabhängige Hauptvektoren der Stufen $\ell, \ell-1, \ldots, 1$.

Wenn es auch nicht zu jeder Matrix $A\in\mathbb{R}^{n,n}$ n linear unabhängige Eigenvektoren gibt, so gilt dennoch

Satz 4.18 (Hauptvektor-Basis)
Zu jedem k-fachen Eigenwert λ von $A\in\mathbb{R}^{n,n}$ gibt es k linear unabhängige Hauptvektoren. Die Hauptvektoren zu verschiedenen Eigenwerten sind linear unabhängig.

Damit gibt es zu jeder Matrix $A\in\mathbb{R}^{n,n}$ n linear unabhängige Hauptvektoren.

Beispiel(e) 4.19
Die Matrix

$$A = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix}$$

hat Eigenwerte

$$\det(A - \lambda E) = -\lambda^3 + 4\lambda^2 - 5\lambda + 2 = 0 \quad \Longrightarrow \quad \lambda = 2, \lambda = 1 \quad \text{(doppelt)}.$$

Zum ersten Eigenwert ergibt sich der Eigenvektor

$$\lambda = 2 : (A - 2E)\mathbf{v} = 0 \quad \Longrightarrow \quad \mathbf{v}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

und zum zweiten Eigenwert der Eigenvektor

$$\lambda = 1 : (A - E)\mathbf{v} = 0 \quad \Longrightarrow \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Es gibt keinen zu zu $\mathbf{v}_2$ linear unabhängigen Eigenvektor zu A, jedoch bekommt man

$$(A - E)^2\mathbf{v} = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}\mathbf{v} = 0 \quad \Longrightarrow \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

$\mathbf{v}_3$ ist ein Hauptvektor der Stufe 2 und $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ ist eine Basis aus Hauptvektoren.

In Verallgemeinerung des Sachverhalts bei Eigenvektoren gilt nun der

Satz 4.20 (Allgemeine Lösung linearer DGL-Systeme)
Es sei $\mathbf{v}$ ein Hauptvektor der Stufe ℓ zum Eigenwert λ der Matrix $A \in \mathbb{R}^{n,n}$. Dann ist

$$\mathbf{y}_v(x) := e^{\lambda x}\left[\mathbf{v} + x(A - \lambda E)\mathbf{v} + \frac{x^2}{2!}(A - \lambda E)^2\mathbf{v} + \ldots + \frac{x^{\ell-1}}{(\ell-1)!}(A - \lambda E)^{\ell-1}\mathbf{v}\right]$$

eine Lösung des homogenen Differentialgleichungssytems $\mathbf{y}'(x) = A\mathbf{y}(x)$.

Mit einer nach Satz 4.18 existierenden Basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ aus Hauptvektoren von A ist dann

$$\mathbf{y}_h(x) = c_1\mathbf{y}_{v_1}(x) + \ldots + c_n\mathbf{y}_{v_n}(x),$$

die allgemeine Lösung des homogenen Systems.

Beispiel(e) 4.21
Im Fall von Beispiel 4.19 bekommt man die Lösungsbasis $\{\mathbf{y}_{v_1}, \mathbf{y}_{v_2}, \mathbf{y}_{v_3}\}$ mit

$$\mathbf{y}_{v_1}(x) = e^{2x}\begin{pmatrix}0\\1\\1\end{pmatrix}, \mathbf{y}_{v_2}(x) = e^{x}\begin{pmatrix}1\\1\\0\end{pmatrix}, \mathbf{y}_{v_3}(x) = e^{x}\left[\begin{pmatrix}0\\0\\1\end{pmatrix} + x\begin{pmatrix}-1\\-1\\0\end{pmatrix}\right].$$

4.4 Stabilität

In diesem Abschnitt betrachten wir sogenannte **autonome Differentialgleichungssysteme**. Das sind spezielle Systeme der Bauart

$$\mathbf{y}'(x) = \mathbf{f}(\mathbf{y}(x))$$

mit $\mathbf{f}: D \to \mathbb{R}^n$ und $D \subseteq \mathbb{R}^n$. Im Unterschied zum allgemeinen Fall hängt die rechte Seite nicht explizit von x ab. Die Interpretation hiervon wird deutlicher, wenn man die unabhängige Variable nicht mit x, sondern mit t („Zeit“) bezeichnet, also zu

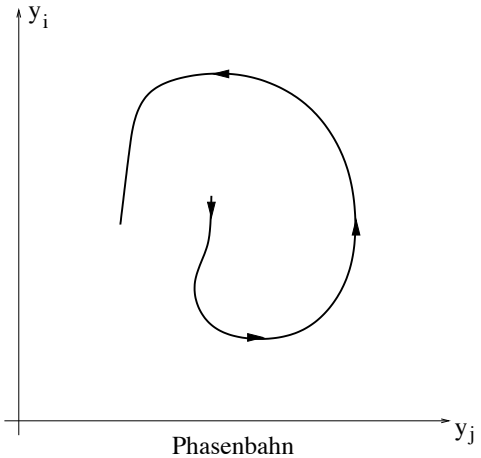
$$\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t)) \tag{4.88}$$

übergeht mit der in der Physik üblichen Konvention, die Ableitung nach der Zeit mit einem Punkt zu kennzeichnen. $\mathbf{y}(t)$ ist der Zustand eines Systems zum Zeitpunkt t. Zum Beispiel könnte man die Flugbahn einer Raumkapsel modellieren, die durch die Komponenten **Tangentialgeschwindigkeit** $v(t)$, **Bahnneigungswinkel** $\gamma(t)$ und **Flughöhe** $h(t)$ gekennzeichnet ist, so dass in diesem Fall

$$\mathbf{y}(t) = \begin{pmatrix}y_1(t)\\y_2(t)\\y_3(t)\end{pmatrix} = \begin{pmatrix}v(t)\\\gamma(t)\\h(t)\end{pmatrix}.$$

Der Zustand des Systems ändert sich mit der Zeit und die spezielle Interpretation des Differentialgleichungssystems (4.88) ist, dass Zustandsänderungen $\dot{\mathbf{y}}(t)$ nur vom augenblicklichen Zustand $\mathbf{y}(t)$ selbst abhängen, aber nicht explizit vom Zeitpunkt, wann dieser Zustand angenommen wird. Dies ist ein in der Steuerungs- und Regelungstechnik wichtiger Fall.

Die Menge der Zustände, die eine Lösung $\mathbf{y}(t)$ von (4.88) ausgehend von einem bestimmten Startzustand $\mathbf{y}(t_0) = \mathbf{y}_0$ annehmen kann, nennt man eine **Phasenbahn**. Interessiert man sich nur für die Zustände selbst, aber nicht dafür, wann sie angenommen werden, dann ist es oft aufschlussreicher, Phasenbahnen zu skizzieren als Lösungskurven. Schematisch sieht das so aus:



Phasenbahn

Mit den Pfeilen wird angedeutet, in welcher Richtung die Phasenbahn durchlaufen wird.

Beispiel(e) 4.22
Wir betrachten das homogene System

$$\dot{\mathbf{y}}(t) = \begin{pmatrix} -4 & 2 \\ -3 & 1 \end{pmatrix} \mathbf{y}(t) \, .$$

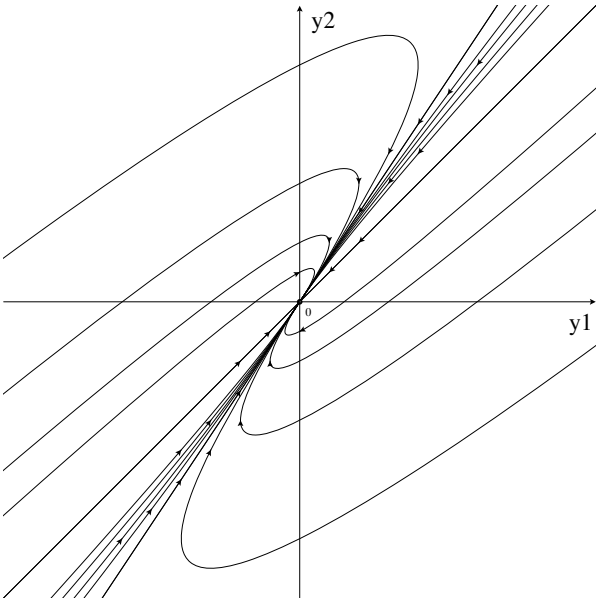
A hat Eigenwerte $\lambda_1 = -2$ und $\lambda_2 = -1$ mit Eigenvektoren

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{und} \quad \mathbf{v}_2 = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \, .$$

Die allgemeine Lösung ist dann

$$\mathbf{y}(t) = c_1 e^{-2t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 e^{-t} \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \quad c_1, c_2 \in \mathbb{R} \, .$$

Wir skizzieren einige Phasenbahnen (zu jeder einzelnen gehört eine spezielle Wahl von c_1 und c_2):



Man beobachtet und erkennt auch an der allgemeinen Lösung, dass hier sämtliche Phasenbahnen in die spezielle Lösung $\mathbf{y}(t) \equiv 0$ einmünden.

Konstante Lösungen wie im letzten Beispiel die Nulllösung $\mathbf{y}(t) = 0$ bekommen einen speziellen Namen.

Definition 4.23 (Stationäre Lösungen)
Eine konstante Funktion $\mathbf{y}(t) = \mathbf{a} \in \mathbb{R}^n$ heißt **stationäre Lösung** oder **Gleichgewichtslösung** (GGL) von (4.88), wenn $\mathbf{f}(\mathbf{a}) = 0$.

GGL entsprechen Zuständen, in denen sich ein System nicht mehr ändert. Oft interessiert man sich für das sogenannte Stabilitätsverhalten eines Systems „in der Nähe“ einer GGL: kann das System „ausbrechen“oder bleibt es für alle Zeiten „stabil“ in der Nähe der GGL (oder – wie im letzten Beispiel – konvergiert es sogar gegen die GGL)? Dazu die folgende

Definition 4.24 (Stabilität)
Eine GGL $\mathbf{a}$ von $\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y})$ heißt

(a): $\mathbf{stabil}$, wenn es für jedes $\varepsilon > 0$ ein $\delta > 0$ gibt, so dass für jede Lösung gilt

$$|\mathbf{y}(0) - \mathbf{a}| < \delta \quad \Longrightarrow \quad |\mathbf{y}(t) - \mathbf{a}| < \varepsilon \quad \text{für alle } t > 0,$$

(b): $\mathbf{asymptotisch\ stabil}$, wenn ein $\delta > 0$ existiert, so dass für jede Lösung gilt

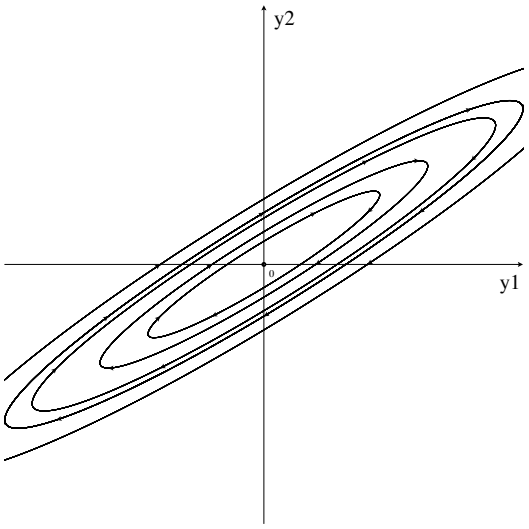
$$|\mathbf{y}(0) - \mathbf{a}| < \delta \quad \Longrightarrow \quad \lim_{t\to\infty} \mathbf{y}(t) = \mathbf{a},$$

(c): $\mathbf{instabil}$, wenn sie nicht stabil ist.

Beispiel(e) 4.25
Das Beispiel (4.22) zeigt eine asymptotisch stabile GGL. Beim homogenen System

$$\dot{\mathbf{y}}(t) = \begin{pmatrix} -3 & 5 \\ -2 & 3 \end{pmatrix} \mathbf{y}(t)$$

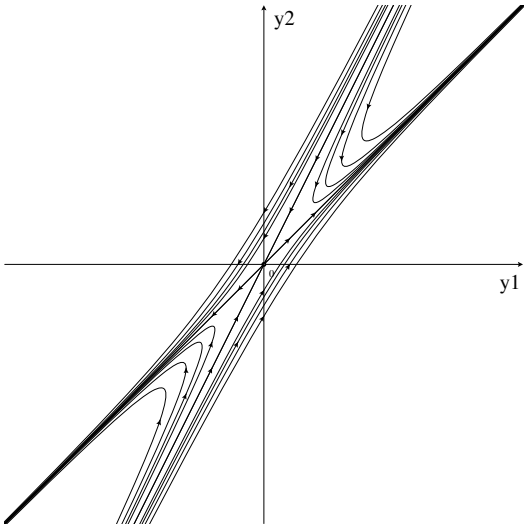
lauten die Eigenwerte von A $\lambda_{1,2} = \pm i$ und wir haben folgende Phasenbahnen



Die GGL $\mathbf{y}(t) = 0$ ist in diesem Fall stabil. Die GGL $\mathbf{y}(t) = 0$ ist instabil im Fall von

$$\dot{\mathbf{y}}(t) = \begin{pmatrix} 4 & -3 \\ 6 & -5 \end{pmatrix} \mathbf{y}(t)$$

in dem A die Eigenwerte $\lambda_1 = 1$ und $\lambda_2 = -2$ hat. Wir skizzieren einige Phasenbahnen:



Entscheidend für die Stabilität in den gezeigten Beispielen sind allein die Eigenwerte der Matrix A. Dass das immer so ist, zeigt der folgende

Satz 4.26 (Stabilitätssatz für lineare Systeme)
Es seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte der Matrix $A \in \mathbb{R}^{n,n}$ des linearen DGL-Systems $\dot{\mathbf{y}} = A\mathbf{y}$.

(a): Genau dann ist eine GGL asymptotisch stabil, wenn $\mathrm{Re}(\lambda_k) < 0$ für alle k.

(b): Wenn $\mathrm{Re}(\lambda_k) > 0$ für ein k, dann ist eine GGL instabil.

(c): Genau dann ist eine GGL stabil, wenn erstens $\mathrm{Re}(\lambda_k) \leq 0$ für alle k und wenn zweitens zu jedem Eigenwert λ_k mit $\mathrm{Re}(\lambda_k) = 0$ nur Eigenvektoren, aber keine Hauptvektoren der Stufe 2 oder höher existieren.

Der **Beweis** dieses Satzes ergibt sich fast unmittelbar aus Satz 4.20.

Bemerkung. Das inhomogene lineare DGL-System mit konstanten Koeffizienten $\dot{\mathbf{y}} = A\mathbf{y} + \mathbf{b}$ hat als GGL die Lösungen $\mathbf{a}$ des linearen Gleichungssystems $A\mathbf{a} + \mathbf{b} = 0$. Wenn eine GGL $\mathbf{a}$ existiert, dann kann man das DGL-System umschreiben in der Form

$$\dot{\mathbf{y}} = A(\mathbf{y} - \mathbf{a}) \quad \Longleftrightarrow \quad \dot{\mathbf{z}} = A\mathbf{z} \quad \text{mit} \quad \mathbf{z} = \mathbf{y} - \mathbf{a}\,.$$

Folglich wird auch in diesem Fall die Art der GGL unabhängig von $\mathbf{b}$ von den Eigenwerten von A bestimmt.

Kapitel 5

Fourier-Analysis

In diesem Kapitel haben wir es häufig mit „komplexwertigen Funktionen mit reellen Argumenten“ zu tun. Dabei handelt es sich um Folgendes. Es sei $I \subseteq \mathbb{R}$ ein reelles Intervall und $f : I \to \mathbb{C}$ eine Funktion mit komplexen Funktionswerten (aber reellen Argumenten). Dann ist für jedes $x \in I$ die Zahl $f(x) \in \mathbb{C}$ zerlegbar in Real- und Imaginärteil:

$$f(x) = u(x) + iv(x) \quad \text{mit} \quad u(x) = \text{ Re } f(x) \quad \text{und} \quad v(x) = \text{ Im } f(x) \, .$$

Die komplexwertige Funktion f definiert also zwei reellwertige Funktion $u : I \to \mathbb{R}$ und $v : I \to \mathbb{R}$ mit $f(x) = u(x) + iv(x)$ für alle $x \in I$. Mit der üblichen Identifikation von $\mathbb{C}$ mit $\mathbb{R}^2$ können wir uns komplexwertige Funktionen mit reellen Argumenten also immer als Kurven $I \to \mathbb{R}^2$ vorstellen:

$$f : I \to \mathbb{C}, \quad f(x) = u(x) + iv(x) = \begin{pmatrix} u(x) \\ v(x) \end{pmatrix} \, .$$

Gemäß dieser Vorstellung übernehmen wir aus der Theorie der reellen, ebenen Kurven die folgenden Definitionen

- f ist stetig in $x_0 \in I$, wenn u und v in x_0 stetig sind
- f ist differenzierbar in $x_0 \in I$, wenn u und v in x_0 differenzierbar sind. Für die Ableitung gilt $f'(x_0) = u'(x_0) + iv'(x_0)$.
- f ist Riemann-integrierbar auf $[a, b] \subseteq I$, wenn u und v es sind. Dann ist $\displaystyle\int_a^b f(x)\, dx = \int_a^b u(x)\, dx + i \int_a^b v(x)\, dx$.

5.1 Fourierreihen

Viele periodische Funktionen können als Überlagerung (Superposition) von Schwingungen dargestellt werden. In diesem Abschnitt wird besprochen, wie das geht.

Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ oder $f : \mathbb{R} \to \mathbb{C}$ wird **periodisch mit Periode** T (oder kurz T-periodisch) genanannt, wenn

$$f(t + T) = f(t) \quad \text{für alle } t \in \mathbb{R} \, . \tag{5.1}$$

Ist eine Funktion $g : [a, a + T) \to \mathbb{C}$ nur auf einem endlichen Intervall $I := [a, a + T)$ definiert, so kann diese Funktion **T-periodisch fortgesetzt** werden. Das bedeutet, dass man eine Funktion $f : \mathbb{R} \to \mathbb{C}$ durch die Festlegung

$$f(t) := g(t - nT)$$

definiert, wobei $n \in \mathbb{Z}$ so gewählt wird, dass $t - nT \in I$. Dies ist stets für genau ein $n \in \mathbb{Z}$ der Fall.

Wir betrachten jetzt eine Klasse $\mathscr{F}$ von Funktionen $f : [0,T] \to \mathbb{C}$, die sich durch folgende Eigenschaften auszeichnen

(a): Es gibt endlich viele „**Sprungstellen**“ $0 = t_0 < t_1 < \ldots < t_p = T$.

(b): Auf jedem offenen Intervall $t_j < t < t_{j+1}$ sind $f(t)$ und $\dot{f}(t) := \dfrac{d}{dt} f(t)$ stetig.

(c): Für $t \to t_j + 0$ und $t \to t_{j+1} - 0$ existieren die Grenzwerte $f(t)$ und $\dot{f}(t)$.

(d): An den Sprungstellen gelte die Konvention

$$f(t_j) := \frac{1}{2}\left(f(t_j - 0) + f(t_j + 0)\right)$$

und

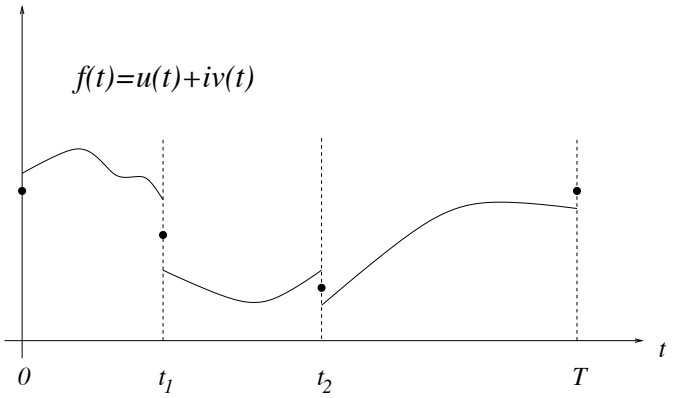
$$f(0) = \frac{1}{2}\left(f(0+) + f(T - 0)\right) \ .$$

(e): Außerhalb von $[0,T]$ werde f periodisch fortgesetzt, falls benötigt:

$$f(t + jT) := f(t) \quad \forall t \in [0,T], \quad \forall j \in \mathbb{Z} \ .$$

Jedes periodische Zeitsignal in der Technik fällt in diese Klasse oder lässt sich durch ein solches f idealisieren.

Das folgende Bild zeigt einen Vertreter der Klasse $\mathscr{F}$.



Für zwei Funktionen $f, g \in \mathscr{F}$ definieren wir

$$\langle f \mid g \rangle := \frac{1}{T} \int_0^T \overline{f(t)} g(t)\, dt \tag{5.2}$$

Jeweils für $f, f_1, f_2, g, g_1, g_2 \in \mathscr{F}$ und $\alpha, \beta \in \mathbb{R}$ gelten die leicht zu überprüfenden Regeln

$$\begin{aligned}
\langle f \mid \alpha g_1 + \beta g_2 \rangle &= \alpha \langle f \mid g_1 \rangle + \beta \langle f \mid g_2 \rangle \\
\langle \alpha f_1 + \beta f_2 \mid g \rangle &= \overline{\alpha} \langle f_1 \mid g \rangle + \overline{\beta} \langle f_2 \mid g \rangle \\
\langle f \mid g \rangle &= \overline{\langle g \mid f \rangle} \\
\langle f \mid f \rangle &\geq 0 \text{ und } = 0 \text{ genau für } f = 0 \ .
\end{aligned}$$

Wegen dieser Eigenschaften nennt man $\langle f \mid g \rangle$ ein Skalarprodukt auf dem Vektorraum $\mathscr{F}$.

Weiterhin benutzen wir die abkürzenden Schreibweisen

$$\left.\begin{array}{rcl} C_n(t) & := & \cos(2\pi nt/T) \\ S_n(t) & := & \sin(2\pi nt/T) \\ E_{\pm n}(t) & := & \exp(\pm 2\pi int/T) = C_n(t) \pm iS_n(t) \end{array}\right\} n = 0, 1, 2, \ldots$$

Wenn t die Zeit in Sekunden misst, dann beschreibt $E_{\pm n}(t)$ eine Kreisschwingung mit einer Frequenz von n/T Hertz: n Umdrehungen in T Sekunden.

Wir haben die folgenden Formeln

- $E_m \cdot E_n = E_{m+n}$
- $\overline{E_n} = E_{-n}$
- $\dfrac{1}{T}\displaystyle\int_0^T E_n(t)\,dt = \begin{cases} 1, & \text{falls } n = 0 \\ 0, & \text{falls } n \neq 0 \end{cases}$
- $\langle E_m \mid E_n \rangle = \dfrac{1}{T}\displaystyle\int_0^T \overline{E_m(t)} E_n(t)\,dt = \delta_{n,m} = \begin{cases} 1, & \text{falls } n = m \\ 0, & \text{falls } n \neq m \end{cases}$

Aufgrund der letzten Eigenschaft nennt man $\{E_n\}_{n=-\infty}^{\infty}$ ein **Orthonormalsystem**.

Definition 5.1 (Trigonometrisches Polynom)
Jede Funktion der Form

$$\gamma_{-n}E_{-n} + \ldots + \gamma_n E_n\,, \qquad \text{alle } \gamma_k \in \mathbb{C}\,,$$

heißt **trigonometrisches Polynom vom Grad** n.

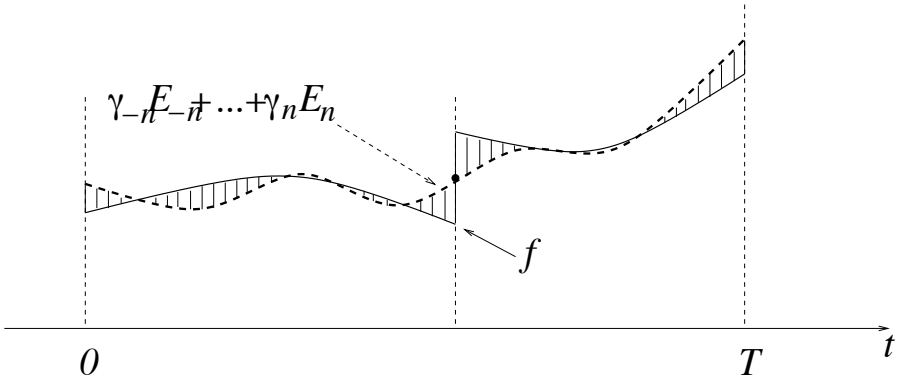
Es soll nun eine Funktion $f \in \mathscr{F}$ durch ein trigonometrisches Polynom vom Grad n „approximiert“ werden, das heißt wir suchen ein trigonometrisches Polynom, das möglichst „dicht“ bei f liegt. Dazu muss geklärt werden, wie denn der Abstand zwischen zwei Funktionen $f, g \in \mathscr{F}$ im allgemeinen und damit der Abstand zwischen f und einem trigonometrischen Polynom im Besonderen gemessen werden soll. Wir führen dazu die weitere Schreibweise

$$\|f\|_2 := \sqrt{\langle f \mid f \rangle} = \left(\frac{1}{T}\int_0^T |f(t)|^2\,dt\right)^{\frac{1}{2}} \tag{5.3}$$

ein und erklären damit den Abstand von zwei Funktionen $f, g \in \mathscr{F}$ als deren „mittlere quadratische Abweichung“:

$$\|f - g\|_2 = \left(\frac{1}{T}\int_0^T |f(t) - g(t)|^2\,dt\right)^{\frac{1}{2}}\,. \tag{5.4}$$

Die folgende Skizze illustriert die mittlere quadratische Abweichung einer Funktion von einem trigonometrischen Polynom.



Mit dieser Festlegung lässt sich präzisieren: gesucht sind die Koeffizienten γ_k eines trigonometrischen Polynoms so, dass

$$\|f - \sum_{k=-n}^{n} \gamma_k E_k\|_2$$

minimal wird.
Behauptung: Die optimale Wahl ist

$$\gamma_k := \langle E_k \mid f \rangle = \frac{1}{T} \int_0^T f(t) e^{-2\pi i k t/T}\, dt\ .$$

Beweis. Wir setzen zur Abkürzung $c_k := \langle E_k \mid f \rangle$ und entsprechend $\overline{c_l} := \langle f \mid E_l \rangle$. Mit den Eigenschaften eines Orthonormalsystems ergibt sich für beliebige $\gamma_k \in \mathbb{C}$

$$\begin{aligned}
\| \sum_{k=-n}^{n} \gamma_k E_k - f\|_2^2 &= \langle \sum_k \gamma_k E_k - f \mid \sum_l \gamma_l E_l - f \rangle \\
&= \sum_k \sum_l \overline{\gamma_k} \gamma_l \delta_{k,l} - \sum_k \overline{\gamma_k} c_k - \sum_l \gamma_l \overline{c_l} + \langle f \mid f \rangle \\
&= \underbrace{\sum_k |\gamma_k - c_k|^2}_{=:\,A} + \underbrace{\langle f \mid f \rangle - \sum_k |c_k|^2}_{=:\,B}\ .
\end{aligned}$$

Hier ist der Anteil B unabhängig von der Wahl der γ_k. Nur Anteil A ist durch die Wahl der γ_k beeinflussbar und wird offenbar am kleinsten, nämlich gleich 0, wenn man $\gamma_k = c_k$ wählt, wie behauptet. q.e.d.

Die eben bewiesene Aussage nochmals als

Satz 5.2 (Bestapproximation mit trigonometrischen Polynomen)
Im Sinn der kleinsten mittleren quadratischen Abweichung ist die beste Approximation einer Funktion $f \in \mathscr{F}$ durch ein trigonometrisches Polynom

$$\gamma_{-n}E_{-n} + \ldots + \gamma_n E_n$$

gegeben durch die Wahl

$$\gamma_k = c_k = \langle E_k \mid f \rangle \ .$$

Die Bestapproximation ist eindeutig: jede andere Approximation ist strikt schlechter.

Weiterhin gilt die **Besselsche Ungleichung**:

$$\sum_{k=-n}^{n} |c_k|^2 \leq \langle f \mid f \rangle \ .$$

Zum Nachweis der Besselschen Ungleichung muss man nur den obigen Beweis zitieren:

$$0 \leq \|f - \sum_{k=-n}^{n} c_k E_k\|_2^2 = \langle f \mid f \rangle - \sum_{k=-n}^{n} |c_k|^2 \tag{5.5}$$

Da die Besselsche Ungleichung unabhängig von n gilt, folgt insbesondere

$$\sum_{k=-\infty}^{\infty} |c_k|^2 \leq \langle f \mid f \rangle$$

und daraus das sogenannte **Riemannsche Lemma**:

$$c_{\pm n} \to 0 \quad \text{für} \quad n \to \infty \ . \tag{5.6}$$

Die Koeffizienten des bestapproximierenden trigonometrischen Polynoms bekommen nun einen offiziellen Namen:

Definition 5.3 (Fourier-Koeffzienten und Fourier-Polynom)
Die **Fourier-Koeffizienten** von $f \in \mathscr{F}$ sind gegeben durch

$$\begin{aligned}
c_k(f) &:= \langle E_k \mid f \rangle &&= \frac{1}{T}\int_0^T f(t)e^{-2\pi ikt/T}\,dt, && k = 0, \pm 1, \pm 2, \ldots \\
a_k(f) &:= \langle C_k \mid f \rangle &&= \frac{1}{T}\int_0^T f(t)\cos(2\pi kt/T)\,dt, && k = 0, 1, 2, \ldots \\
b_k(f) &:= \langle S_k \mid f \rangle &&= \frac{1}{T}\int_0^T f(t)\sin(2\pi kt/T)\,dt, && k = 0, 1, 2, \ldots
\end{aligned}$$

Das **Fourier-Polynom vom Grad** n ist definiert durch

$$T_{n,f}(t) = \sum_{k=-n}^{n} c_k(f)E_k(t) = \sum_{k=-n}^{n} c_k(f)e^{2\pi ikt/T} \ .$$

Unmittelbar aus der Definition der Fourier-Koeffizienten ergibt sich

$$c_k = a_k - ib_k \qquad \text{und} \qquad c_{-k} = a_k + ib_k \ ,$$

was auf

$$2a_k = c_k + c_{-k} \qquad \text{und} \qquad 2b_k = i(c_k - c_{-k}) \tag{5.7}$$

führt und insbesondere auf $a_0 = c_0$. Damit lässt sich das Fourier-Polynom auch direkt mithilfe von $\sin$- und $\cos$-Funktionen schreiben:

$$T_{n,f} = a_0 + \sum_{k=1}^{n} \left(2a_k \cos(2\pi kt/T) + 2b_k \sin(2\pi kt/T)\right) \ . \tag{5.8}$$

Oft werden in der Literatur $2a_k$ mit a_k und $2b_k$ mit b_k bezeichnet.

Beispiel(e) 5.4
Die „Sägezahn-Funktion“ ist definiert durch

$$f(t) = \begin{cases} 0, & \text{für } t = 0 \\ \frac{1}{2} - \frac{t}{2}, & \text{für } 0 < t < 2\pi \end{cases} \ .$$

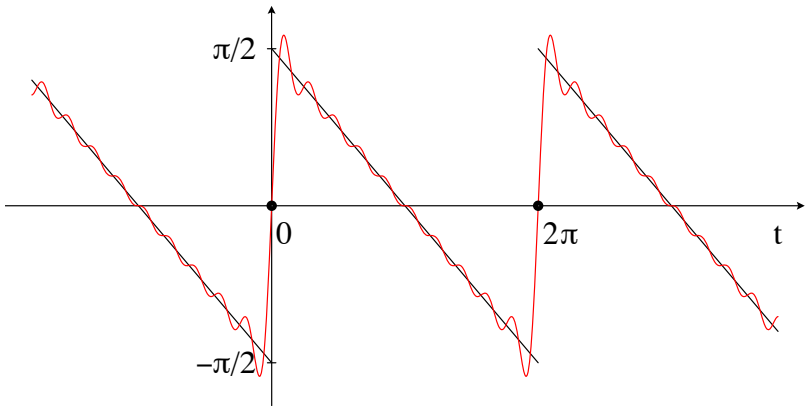
Die Fourier-Koeffizienten ergeben sich zu

$$c_k = \begin{cases} 0, & \text{falls } k = 0 \\ -i/(2k), & \text{falls } k \neq 0 \end{cases} \quad \text{bzw.} \quad \begin{aligned} a_k &= 0, & k &= 0, 1, 2, \ldots \\ b_k &= 1/(2k), & k &= 1, 2, 3, \ldots \end{aligned}$$

Damit haben wir das Fourier-Polynom

$$T_{n,f}(t) = \sum_{k=1}^{n} \frac{\sin(kt)}{k} \ .$$

Wir skizzieren die Sägezahnfunktion f (in schwarz) und $T_{10,f}$ (in rot):



Beispiel(e) 5.5
Für die Funktion

$$f(t) = t \cdot (1-t), \qquad 0 \le t \le 1$$

ergeben sich die folgenden Fourier-Koeffizienten

$$c_k = \begin{cases} 1/6, & \text{falls } k = 0 \\ -1/(2\pi^2 k^2), & \text{falls } k \neq 0 \end{cases} \quad \text{bzw.} \quad \begin{array}{lll} a_0 = 1/6, & a_k & = & -1/(2\pi^2 k^2) \\ & b_k & = & 0 \end{array}$$

Das n-te Fourier-Polynom ist

$$T_{n,f}(t) = \frac{1}{6} - \sum_{k=1}^{n} \frac{\cos(2\pi k t)}{\pi^2 k^2} \ .$$

Der Approximationsfehler an der Stelle $t = 0$ ist

$$T_{n,f}(0) - f(0) = \frac{1}{6} - \frac{1}{\pi^2}\left(\frac{1}{1^2} + \ldots + \frac{1}{n^2}\right) \ , \tag{5.9}$$

In den beiden vorangegangenen Beispielen bestätigt sich das Riemannsche Lemma: die Fourier-Koeffizienten $c_{\pm k}$ gehen für $k \to \infty$ gegen Null. Allerdings tun sie das nicht immer gleich schnell, sondern im zweiten Beispiel schneller als im ersten. Das liegt daran, dass die Funktion f im Beispiel 5.5 im Gegensatz zu der in Beispiel 5.4 stetig ist. Man kann Folgendes zeigen.

Es sei $f \in \mathscr{F}$ mit den zusätzlichen Eigenschaften

- f sei $(n-1)$-mal stetig differenzierbar ($n = 1$ bedeutet: f ist stetig, $n = 0$ bedeutet, dass f nicht einmal stetig ist)
- die n-te und die $(n+1)$-te Ableitung existieren und sind stückweise stetig

dann gilt

$$c_k(f) = \mathscr{O}\left(\frac{1}{k^{n+1}}\right) \ , \tag{5.10}$$

das heißt die Fourier-Koeffizienten fallen so schnell gegen Null ab wie die Folge $1/k^{n+1}$.

Die folgenden Regeln helfen bei der Berechnung von Fourierkoeffizienten.

Satz 5.6 (Berechnung von Fourier-Koeffizienten)
Es seien $f, g \in \mathscr{F}$ und $\alpha, \beta \in \mathbb{R}$. Dann gelten

(a): **Linearität**. Es ist $c_k(\alpha f + \beta g) = \alpha c_k(f) + \beta c_k(g)$.

(b): **Konjugation**. Es ist $c_k(\overline{f}) = \overline{c_{-k}(f)}$.

(c): **Streckung**. Die Funktion $h(t) := f(ct)$ hat die Periode T/c. Man erhält

$$c_k(h) = \frac{c}{T} \int_0^{T/c} f(ct) e^{-2\pi i k c t/T}\, dt = \frac{1}{T} \int_0^T f(\tau) e^{-2\pi i k \tau/T}\, d\tau = c_k(f)\ .$$

Eine Umskalierung der unabhängigen Variablen ändert also die Fourierkoeffizienten nicht.

(d): **Zeitverschiebung**. Die Funktion $h(t) := f(t-a)$ entspricht einer Verschiebung von f um $+a$ auf der t-Achse. Dann

$$c_k(h) = \frac{1}{T} \int_0^T f(t-a) e^{-2\pi i k t/T}\, dt = e^{-2\pi i k a/T} \cdot c_k(f)\ .$$

Eine Zeit-Translation entspricht also einer Rotation der Fourier-Koeffizienten. Diesen Effekt beobachtet man häufig in der digitalen Signalverarbeitung.

(e): **Symmetrien**.

- Wenn f reellwertig ist, dann gilt

$$c_{-k} = \overline{c_k} \quad \Longrightarrow \quad a_k, b_k \in \mathbb{R}\ .$$

- Wenn f „gerade“, d.h. $f(T-t) = f(t)$ bzw. $f(-t) = f(t)$, dann $c_{-k} = c_k$ und $b_k = 0$.
- Wenn f „ungerade“, d.h. $f(T-t) = -f(t)$ bzw. $f(-t) = -f(t)$, dann $c_{-k} = -c_k$ und $a_k = 0$.
- „Bei halber Periode“:

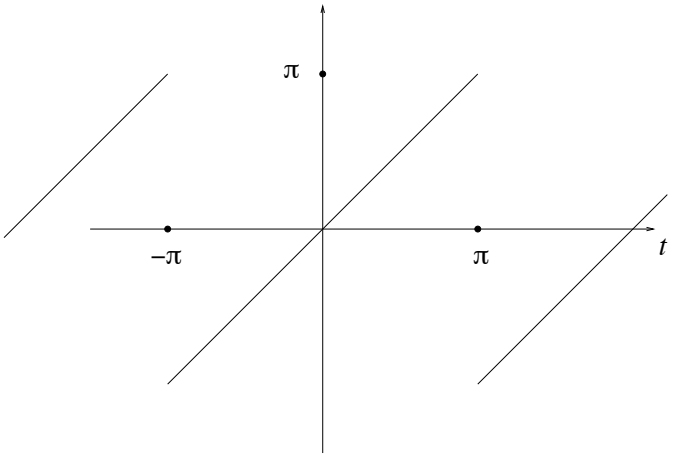
$$f(\frac{T}{2} + t) = f(t) \quad \Longrightarrow \quad c_{2k+1} = 0$$

$$f(\frac{T}{2} + t) = -f(t) \quad \Longrightarrow \quad c_{2k} = 0$$

Beispiel(e) 5.7
In Beispiel 5.4 hatten wir die Fourier-Koeffizienten einer „Sägezahn-Funktion“ f berechnet. Betrachten wir nun die daraus abgeleitete Funktion

$$h(t) = -2f(t+\pi) \ ,$$

die in der folgenden Skizze dargestellt ist.



Mit den Regeln des letzten Satzes ergibt sich wegen $\exp(-\pi ik) = (-1)^k$, dass $c_k(h) = 2(-1)^{k+1}c_k(f)$, also

$$T_{n,h} = 2\left(\frac{\sin(t)}{1} - \frac{\sin(2t)}{2} + \frac{\sin(3t)}{3} - \frac{\sin(4t)}{4} \pm \ldots + (-1)^{n+1}\frac{\sin(nt)}{n}\right) \ .$$

Bisher haben wir nur Fourier-Polynome betrachtet. Wird die Approximation immer besser, wenn man den Polynomgrad n immer höher treibt? In den praktisch relevanten Fällen ist das tatsächlich so, wie im folgenden Satz ausgesagt.

Satz 5.8 (Punktweise Konvergenz der Fourierreihe)
Es sei $f \in \mathscr{F}$ mit Fourier-Polynom

$$T_{n,f}(t) = \sum_{k=-n}^{n} c_k(f) E_k(t) = \sum_{k=-n}^{n} c_k(f) e^{2\pi i k t/T} \ .$$

Setzt man

$$T_{\infty,f}(t) := \lim_{n\to\infty} T_{n,f}(t) = \sum_{k=-\infty}^{\infty} c_k(f) e^{2\pi i k t/T}$$

so konvergiert diese Reihe und es gilt

$$f(t) = T_{\infty,f}(t) \quad \text{für alle } t \in \mathbb{R}.$$

Darüber hinaus gilt die **Parsevalsche Identität**:

$$\sum_{k=-\infty}^{\infty} |c_k(f)|^2 = \frac{1}{T} \int_0^T |f(t)|^2 \, dt \ . \tag{5.11}$$

Beispiel(e) 5.9
In Beispiel 5.4 hatten wir für die Funktion

$$f(t) = t \cdot (1-t), \qquad 0 \le t \le 1,$$

das n-te Fourier-Polynom berechnet:

$$T_{n,f}(t) = \frac{1}{6} - \sum_{k=1}^{n} \frac{\cos(2\pi k t)}{\pi^2 k^2} \ .$$

Mit Satz 5.8 ergibt sich für alle $t \in \mathbb{R}$ die Identität

$$t(1-t) = \frac{1}{6} - \sum_{k=1}^{n} \frac{\cos(2\pi k t)}{\pi^2 k^2} \ .$$

insbesondere also für $t = 0$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \ .$$

Es sollen noch zwei Anwendungen von Fourier-Reihen skizziert werden.

Mittels Fourier-Koeffizienten besteht die Möglichkeit, eine Funktion (ein Signal) durch eine diskrete Menge von Zahlen (eben die Fourier-Koeffizienten) zu beschreiben. Praktikabel für den Rechnereinsatz ist allerdings nur eine approximative Beschreibung von Funktionen durch Fourier-Polynome, denen *endlich viele* Fourier-Koeffizienten entsprechen. Bei festem Polynomgrad wird die Approximation desto besser, je schneller die Fourier-Koeffizienten gegen Null abfallen. Dabei kann nun (5.10) berücksichtigt werden.

Beispiel(e) 5.10
Setzt man die Funktion

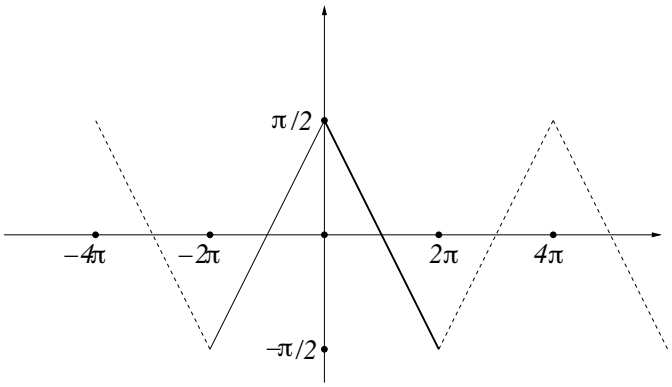
$$f(t) = \frac{1}{2} - \frac{t}{2}, \quad 0 < t < 2\pi,$$

wie in Beispiel 5.4 2π-periodisch fort, dann ergibt sich das das n-te Fourier-Polynom

$$T_{n,f}(t) = \sum_{k=1}^{n} \frac{\sin(kt)}{k} \, .$$

Die Fourier-Koeffizienten fallen gemäß (5.10) nur langsam gegen Null, weil die fortgesetzte Funktion unstetig ist.

Wählt man hingegen erst eine gerade Fortsetzung auf das Intervall $[-2\pi, 2\pi]$ und anschließend eine 4π-periodische Fortsetzung, so erhält man eine stetige Funktion, deren Fourier-Koeffizienten wesentlich schneller gegen Null abfallen.



Diese Technik findet beispielsweise in der JPEG-Bildkompression Anwendung.

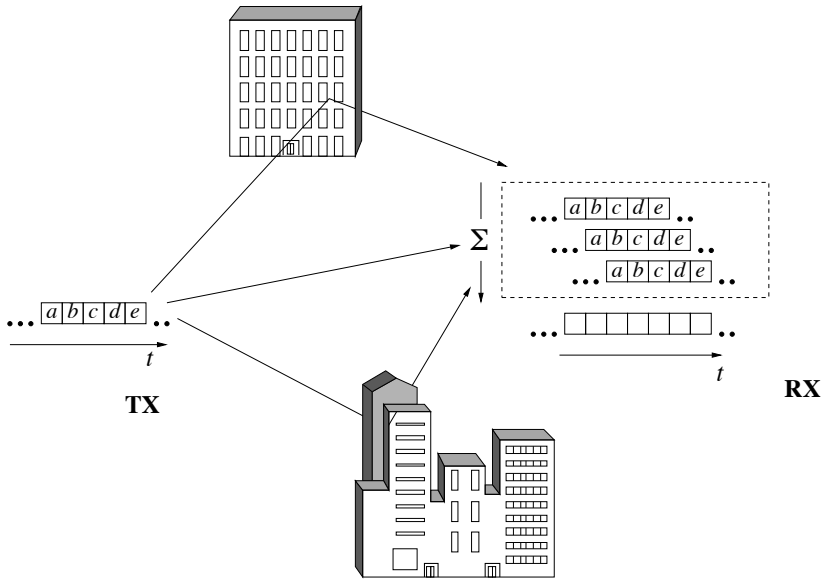
Eine in der Signalverarbeitung häufig auftretende Operation ist die folgende Verknüpfung zweier Funktionen.

Definition 5.11 (Faltung)
Die Faltung (engl. **convolution**) zweier Funktionen $f, g \in \mathscr{F}$ ist definiert als

$$(f * g)(t) := \frac{1}{T} \int_0^T f(\tau) g(t - \tau) \, d\tau, \quad t \in \mathbb{R} \, .$$

$f * g$ ist also der Name einer neuen Funktion und $(f * g)(t)$ ist deren Wert an der Stelle t.

Die Faltung wird zum Beispiel benötigt, um die durch Mehrwegeausbreitung eines Signals entstehende Bildung von Interferenzen zu modellieren. Ein diskrete (näherungsweise) Beschreibung dieses Effekts wird durch die folgende Skizze veranschaulicht.



Die Beschreibung im analogen Modell ist

$$z(t) = \frac{1}{T}\int_0^T M(\tau)y(t-\tau)\,d\tau\ , \tag{5.12}$$

wobei

- $y(t)$ das zu übertragende Signal zum Zeitpunkt t ist,
- $y(t-\tau)$ das zu übertragende Signal zum früheren Zeitpunkt $t-\tau$ ist,
- $M(\tau)$ ein (in aller Regel schnell abklingender) Gewichtskoeffizient („der Kanal“) ist und
- $z = M * y$ das empfangene, durch Interferenzen gestörte Signal ist.

Bemerkung. In dieser Anwendung ist M in aller Regel keine periodische Funktion. Andererseits benötigt man M nur auf dem Intervall $[0,T]$ und kann sich diese Funktion dann ohne Schaden als periodisch fortgesetzt denken.

In der skizzierten Situation hat ein Empfänger die Aufgabe, aus einem empfangenen, gemäß (5.12) verzerrten Signal z das ursprünglich gesendete Signal y zurück zu gewinnen. Dies nennt man „Entzerrung“.

Da mit $f, g \in \mathscr{F}$ auch $f * g \in \mathscr{F}$ kann man die Fourier-Koeffizienten von $f * g$ berechnen. Dabei ergibt sich

Satz 5.12 (Fourier-Koeffizienten von $f * g$)
Für alle $k \in \mathbb{Z}$ ist

$$c_k(f * g) \quad = \quad c_k(f) \cdot c_k(g)\ .$$

Beweis durch Nachrechnen:

$$
\begin{aligned}
c_k(f*g) &= \frac{1}{T}\int_0^T (f*g)(t)e^{-2\pi ikt/T}\,dt = \frac{1}{T^2}\int_0^T \left(\int_0^T f(\tau)g(t-\tau)\,d\tau\right) e^{-2\pi ikt/T}\,dt \\
&= \frac{1}{T^2}\int_0^T f(\tau)\left(\int_0^T g(t-\tau)e^{-2\pi ik(t-\tau+\tau)/T}\,dt\right)\,d\tau \\
&= \frac{1}{T^2}\int_0^T f(\tau)e^{-2\pi ik\tau/T}\left(\int_0^T g(t-\tau)e^{-2\pi ik(t-\tau)/T}\,dt\right)\,d\tau \\
&\qquad [\text{ Substitution } t' = t-\tau] \\
&= \frac{1}{T}\int_0^T f(\tau)e^{-2\pi ik\tau/T}\underbrace{\left(\frac{1}{T}\int_{-\tau}^{T-\tau} g(t')e^{-2\pi ikt'/T}\,dt'\right)}_{=\frac{1}{T}\int_0^T \cdots}\,d\tau \\
&= \frac{1}{T}\int_0^T f(\tau)e^{-2\pi ik\tau/T}\,d\tau\cdot c_k(g) = c_k(f)\cdot c_k(g)\,. \qquad \text{q.e.d}
\end{aligned}
$$

Mit Hilfe des Satzes 5.12 kann das Problem der Entzerrung im Prinzip gelöst werden. Dabei wird vorausgesetzt, dass der Empfänger die Funktion M kennt oder schätzen kann („Kanalschätzung“).

Auflösung von (5.12) nach y.

Diese Aufgabe lässt sich in drei Schritten lösen:

(a): **Fourier-Analyse:** $M(t) \Longrightarrow \{c_k(M)\}$, $z(t) \Longrightarrow \{c_k(z)\}$.

(b): Berechne $c_k(y) = c_k(z)/c_k(M)$.

(c): **Fourier-Synthese:** $\{c_k(y)\} \Longrightarrow y(t)$

Offenbar funktioniert der vorgeschlagene Lösungsweg — in der Signalverarbeitung nennt man ihn den **zero forcing equalizer** — nicht, wenn Fourier-Koeffizienten $c_k(M) = 0$ auftreten.

5.2 Fourier-Transformation

Für eine T-periodische, Riemann-integrierbare Funktionen $f : \mathbb{R} \to \mathbb{C}$ sind die Fourierkoeffizienten $c_k(f)$ definiert gemäß:

$$
\boxed{c_k(f) = \frac{1}{T}\int_0^T f(t)e^{-2\pi ikt/T}\,dt\,.} \tag{5.13}
$$

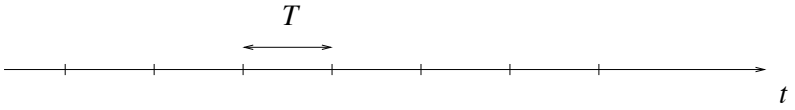
Wenn $f \in \mathscr{F}$, also insbesondere stückweise stetig differenzierbar ist, dann gilt für jedes $t \in \mathbb{R}$:

$$\boxed{f(t) = \sum_{k=-\infty}^{\infty} c_k(f) e^{2\pi i k t/T} \ .} \tag{5.14}$$

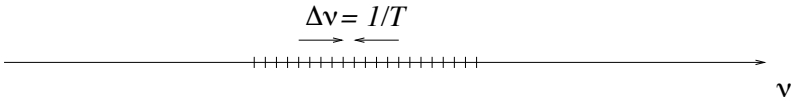
das heißt die Fourierreihe „konvergiert punktweise“ gegen f.

Man beobachtet ein reziprokes Verhältnis der „Zeitskala“ zur „Frequenzskala“: je größer die Periode T wird, desto enger rücken die Frequenzen $\{k/T\}_{k=-\infty}^{\infty}$ zusammen.

Zeit – Skala:



Frequenz – Skala:



Wir spekulieren nun darüber, was passiert, wenn $T \to \infty$, das heißt, wenn wir a-periodische Funktionen betrachten. Dazu tun wir zwei Dinge

(a): wir schieben den Faktor $1/T$ von (5.13) zu (5.14) und

(b): wir führen die neue diskrete Variable $\nu := k/T$ statt $k \in \mathbb{Z}$ ein und schreiben $\Delta\nu = 1/T$.

Dann geht (5.13) über in

$$c(\nu) = \int_{\text{„Periode“}} f(t) e^{-2\pi i \nu t}\, dt$$

und (5.14) geht über in

$$f(t) = \sum_{\text{„alle } \nu\text{“}} c(\nu) e^{2\pi i \nu t}\, \Delta\nu$$

Demnach lässt sich vermuten, dass für $T \to \infty$

$$c(\nu) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \nu t}\, dt$$

(dies ist eine Definition) und

$$f(t) = \int_{-\infty}^{\infty} c(\nu) e^{+2\pi i \nu t}\, d\nu$$

(das ist zu beweisen!).

Tatsächlich ist die Vermutung unter „milden Annahmen“ über f richtig, was wir noch als Satz formulieren werden.

Den Übergang $f(t) \rightsquigarrow c(\nu)$ nennt man eine **Transformation**, weil er auch rückwärts zu machen ist: $c(\nu) \rightsquigarrow f(t)$. Es ensteht also kein Informationsverlust.

Ab sofort führen wir statt $c(\nu)$ eine neue Bezeichnung ein, die die Beziehung zu $f(t)$ deutlicher macht:

Definition 5.13 (Fourier-Transformation)
Für eine Funktion $f : \mathbb{R} \to \mathbb{C}$ definieren wir ihre **Fourier-Transformierte** (oder **Spektralfunktion**) $F : \mathbb{R} \to \mathbb{C}$ durch

$$F(\nu) := \int_{-\infty}^{\infty} f(t)e^{-2\pi i\nu t}\,dt\,, \qquad \textbf{(„Hin“)}$$

sofern dieses Integral für alle $\nu \in \mathbb{R}$ existiert.

Abkürzend schreibt man das oft auch in der Form

$$f(t) \circ\!\!-\!\!\bullet\, F(\nu)$$

und spricht von „Fourier-Korrespondenz“.

Die **inverse Fourier-Transformierte** von $F : \mathbb{R} \to \mathbb{C}$ ist die Funktion

$$t \mapsto \int_{-\infty}^{\infty} F(\nu)e^{2\pi i\nu t}\,d\nu, \qquad \textbf{(„Rück“)}$$

sofern das Integral für alle $t \in \mathbb{R}$ existiert.

Bemerkung. Die auftretenden Integrale sind uneigentlich, wobei *beide* Integralgrenzen kritisch sind. Im allgemeinen ist für die Existenz eines Integrals $\int_{-\infty}^{\infty} \cdots$ zu fordern, dass

$$\lim_{\alpha\to-\infty} \int_{\alpha}^{c} \cdots + \lim_{\beta\to\infty} \int_{c}^{\beta} \cdots$$

für ein beliebig gewähltes endliches $c \in \mathbb{R}$ existiert. Die beiden Grenzwerte sind also *unabhängig voneinander* zu bestimmen. *Die Definition der Fouriertransformierten $F(\nu)$* **(„Hin“)** *ist in diesem Sinn zu verstehen.* Daneben gibt es aber noch einen schwächeren Grenzwertbegriff, bei dem lediglich gefordert wird, dass der sogenannte **Hauptwert** eines Integrals $\int_{-\infty}^{\infty} \cdots$ existiert, nämlich

$$\lim_{N\to\infty} \int_{-N}^{N} \cdots$$

Untere und obere Grenze gehen hier *symmetrisch* gegen ∞. Die erwartete Formel der Rücktransformation

$$f(t) = \int_{-\infty}^{\infty} F(\nu)e^{2\pi i\nu t}\,d\nu$$

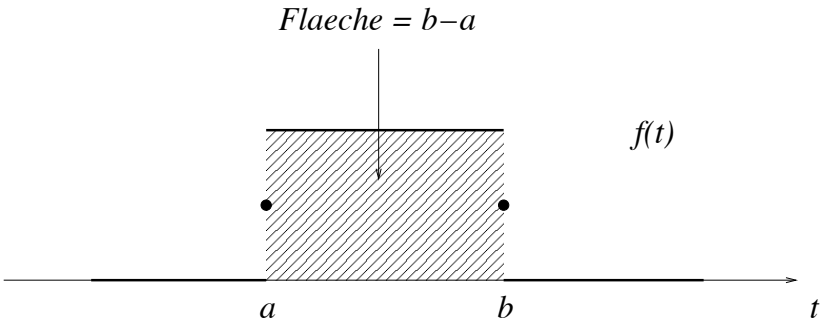
gilt im allgemeinen nur, wenn man das hier auftretende Integral im Sinn des Hauptwerts versteht, und deswegen soll die inverse Fourier-Transformation **(„Rück“)** in diesem Sinn verstanden werden.

Wir bringen jetzt Beispiele, in denen man sieht, dass

$$\text{„Hin“} + \text{„Rück“} = \text{Identität}$$

Beispiel(e) 5.14
Wir betrachten die „Rechtecksfunktion“

$$f(t) = \begin{cases} 0, & \text{für } t < a \text{ oder } t > b \\ 1/2, & \text{für } t = a \text{ oder } t = b \\ 1, & \text{für } a < t < b \end{cases}$$



Die Fourier-Transformierte ist

$$F(\nu) = \int_a^b e^{-2\pi i\nu t}\, dt = \frac{e^{-2\pi i\nu a} - e^{-2\pi i\nu b}}{2\pi i\nu}\ ,$$

wobei die rechte Seite für $\nu = 0$ im Sinn des Grenzübergangs $\nu \to 0$ zu verstehen ist, der den Wert $b - a$ liefert (das sieht man zum Beispiel aus der Potenzreihe für die Exponentialfunktion).

Für das Folgende benötigen wir das Integral

$$\sigma(a) := \int_{-\infty}^{\infty} \frac{\sin(ax)}{x}\, dx = \begin{cases} +\pi, & \text{falls } a > 0 \\ 0, & \text{falls } a = 0 \\ -\pi, & \text{falls } a < 0 \end{cases}$$

Damit bekommen wir für die inverse Transformation von F

$$\int_{-\infty}^{\infty} F(\nu) e^{+2\pi i\nu t}\, d\nu = \int_{-\infty}^{\infty} \frac{e^{2\pi i\nu(t-a)} - e^{2\pi i\nu(t-b)}}{2\pi i\nu}\, d\nu$$

$$= \int_{-\infty}^{\infty} \frac{i\sin[2\pi\nu(t-a)] - i\sin[2\pi\nu(t-b)]}{2\pi i\nu}\, d\nu$$

$$= \frac{1}{2\pi}\left(\sigma(t-a) - \sigma(t-b)\right) = \begin{cases} \frac{1}{2\pi}(-\pi - (-\pi)), & t < a < b \\ \frac{1}{2\pi}(+\pi - (-\pi)), & a < t < b \\ \frac{1}{2\pi}(+\pi - \pi), & a < b < t \end{cases}$$

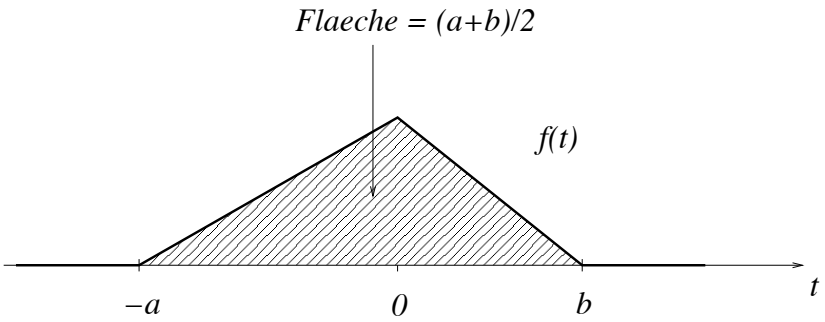
$$= \begin{cases} 1, & \text{falls } a < t < b \\ 1/2, & \text{falls } t = a \text{ oder } t = b \\ 0, & \text{falls } t < a \text{ oder } t > b \end{cases}$$

(Beim Übergang zur zweiten Zeile entfallen die cos-Terme, weil sie wegen der Division durch ν zu punktsymmetrischen Integranden mit Integralwert 0 führen).

Es zeigt sich also $f(t) = \displaystyle\int_{-\infty}^{\infty} F(\nu) e^{2\pi i\nu t}\, d\nu$.

Beispiel(e) 5.15
Wir betrachten die „Dachfunktion“

$$f(t) = \begin{cases} 0, & \text{für } t < -a \text{ oder } t > b \\ 1 + t/a, & \text{für } -a \le t \le 0 \\ 1 - t/b, & \text{für } 0 \le t \le b \end{cases}$$



Die Fourier-Transformierte ist

$$F(\nu) = \int_{-a}^{0} (1 + \frac{t}{a}) e^{-2\pi i \nu t}\, dt + \int_{0}^{b} (1 - \frac{t}{b}) e^{-2\pi i \nu t}\, dt$$

$$= \frac{1 - e^{2\pi i \nu a}}{4\pi^2 \nu^2 a} + \frac{1 - e^{-2\pi i \nu b}}{4\pi^2 \nu^2 b}\ ,$$

wobei die rechte Seite für $\nu = 0$ im Sinn des Grenzübergangs $\nu \to 0$ zu verstehen ist, der den Wert $(a+b)/2$ liefert (das sieht man wiederum aus der Potenzreihe für die Exponentialfunktion).

Für das Folgende benötigen wir das Integral

$$\tau(a) := \int_{-\infty}^{\infty} \frac{1 - \cos(ax)}{x^2}\, dx = \pi \cdot |a|$$

(zum Beweis Substitution $u = ax$ und partielle Integration).

Damit bekommen wir für die inverse Transformation von F:

$$\int_{-\infty}^{\infty} F(\nu) e^{2\pi i \nu t}\, d\nu$$

$$= \int_{-\infty}^{\infty} \left(\frac{\cos(2\pi\nu t) - \cos(2\pi\nu(t+a))}{4\pi^2\nu^2 a} + \frac{\cos(2\pi\nu t) - \cos(2\pi\nu(t-b))}{4\pi^2\nu^2 b} \right) d\nu$$

Die $\sin$-Terme entfallen, da sie zu punktsymmetrischen Integranden mit Integralwert 0 führen. Jetzt kann man im Zähler der Integranden neutral den Wert $0 = -1+1$ addieren und die Formel für das Integral τ anwenden. Damit erhält man

$$\int_{-\infty}^{\infty} F(\nu) e^{2\pi i \nu t}\, d\nu = \frac{1}{2a} \left(-|t| + |t+a| \right) + \frac{1}{2b} \left(-|t| + |t-b| \right)\ .$$

Der erste Summand ist getrennt nach den Fällen $t \le -a$, $-a \le t \le 0$ und $0 \le t$ zu betrachten und der zweite Summand getrennt nach den Fällen $t \le 0$, $0 \le t \le b$ und $b \le t$. Daraus erkennt man wiederum $f(t) = \int_{-\infty}^{\infty} F(\nu) e^{2\pi i \nu t}\, d\nu$.

Aus den beiden voranstehenden Beispielen ergibt sich, dass auch für die folgende Funktionen punktweise die Gleichung „Hin“ + „Rück“ = Identität bezüglich der Fouriertransformation gilt:

- Treppenfunktionen,
- Polygonzüge und
- Funktionen, die abschnittsweise Geradenstücke sind.

Es ist jeweils wie bei Fourierreihen vorausszusetzen, dass an jeder Sprungstelle t einer solchen Funktion f

$$f(t) = \frac{1}{2}\left(f(t-0) + f(t+0)\right)$$

gilt.

Satz 5.16 (Fourier-Integraltheorem)
Für eine große Klasse von Funktionen $f : \mathbb{R} \to \mathbb{C}$ gilt mit der Fourier-Transformierten

$$F(\nu) = \int_{-\infty}^{\infty} f(t)e^{-2\pi i\nu t}\,dt, \quad \nu \in \mathbb{R}, \qquad \textbf{(„Hin“)}$$

dass

$$f(t) = \int_{-\infty}^{\infty} F(\nu)e^{2\pi i\nu t}\,d\nu, \quad t \in \mathbb{R}. \qquad \textbf{(„Rück“)}$$

Bemerkungen.

(a): Das zweite Integral ist im Sinn eines Hauptwerts zu verstehen, wie im Anschluß an Definition 5.13 erläutert.

(b): Je weniger Annahmen man bezüglich f macht, desto schwieriger wird der Beweis.

(c): Erforderlich ist die **absolute Integrierbarkeit** von f, also

$$\int_{-\infty}^{\infty} |f(t)|\,dt < \infty$$

(dieses uneigentliche Integral von $|f(t)|$ existiert im regulären Sinn immer schon dann, wenn der Cauchysche Hauptwert existiert). Dies ist zum Beispiel gewährleistet, wenn f stückweise stetig ist und für „große“ Werte t

$$|f(t)| \le \frac{C}{t^{1+\alpha}}$$

mit festen positiven Konstanten C und α. **Diese Voraussetzung reicht aus für die Existenz der Hin-Transformation, aber noch nicht für die Gültigkeit der Rück-Transformation!**

(d): An jeder Sprungstelle t_j von $f(t)$ muss

$$f(t_j) = \frac{1}{2}\left(f(t_j-0) + f(t_j+0)\right)$$

gelten.

(e): Hinreichend für die Gültigkeit des Satzes sind neben den beiden zuletzt genannten die beiden zusätzlichen Bedingungen

- f habe in jedem endlichen Intervall nur endlich viele Sprungstellen t_j
- auf jedem Intervall (t_j, t_{j+1}) zwischen zwei Spungstellen seien f und $\dot{f}$ stetig und stetig fortsetzbar auf $[t_j, t_{j+1}]$.

(f): In der Literatur findet man die Terme $e^{-2\pi i\nu t}$ und $e^{+2\pi i\nu t}$ oft auch vertauscht, das heißt es ist nicht eindeutig geklärt, was die „Hin"- und was die „Rück"-Transformation ist.

(g): Speziell Mathematiker und Physiker verwenden oft auch die Integrations-Variable $\omega = 2\pi\nu$ statt ν in der Formel „Rück". Damit bekommt man

$$F(\omega) := \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt$$

in der „Hin"-Transformation und

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}\,d\omega$$

in der „Rück"-Transformation. Manchmal wird dann auch noch der Faktor $1/(2\pi)$ in die „Hin"-Transformation verschoben – oder aber man teilt auf: einen Faktor $1/\sqrt{2\pi}$ in „Hin" *und* „Rück".

Im ganzen restlichen Kapitel werden wir – sofern nichts anderes gesagt ist – annehmen, dass wir es mit Funktionen zu tun haben, die folgende Bedingungen erfüllen, welche hinreichend für Satz 5.16 sind:

Voraussetzungen zu Satz 5.16. Funktionen $f : \mathbb{R} \to \mathbb{C}$ sollen die folgenden Eigenschaften haben

(a): stückweise Stetigkeit und stetige Differenzierbarkeit auf jedem endlichen Intervall

(b): $\displaystyle\int_{-\infty}^{\infty} |f(t)|\,dt \;<\; \infty$

(c): an jeder Sprungstelle sei

$$f(t_j) = \frac{1}{2}\left(f(t_j - 0) + f(t_j + 0)\right)$$

Wir geben nun einige Regeln an, die bei der Berechnung und bei der Anwendung der Fourier-Transformation dienlich sind. Dabei folgen wir stets der Konvention, dass die Fourier-Transfomierte einer Funktion $f = f(t)$ mit Großbuchstaben $F = F(\nu)$ bezeichnet wird, also

$$f(t) \circ\!\!-\!\!\bullet\, F(\nu), \quad g(t) \circ\!\!-\!\!\bullet\, G(\nu) \quad \text{usw.}$$

1. Linearität.

Für alle $\alpha, \beta \in \mathbb{C}$ ist

$$\boxed{\alpha \cdot f(t) + \beta \cdot g(t) \circ\!\!-\!\!\bullet\, \alpha \cdot F(\nu) + \beta \cdot G(\nu)}$$

Dies folgt aus der Linearität des Integrals.

2. Spiegelung.

Definiert man die Funktion $g : \mathbb{R} \to \mathbb{C}$ durch $g(t) = f(-t)$, dann ist mit $f(t) \circ\!\!-\!\!\bullet\, F(\nu)$

$$\boxed{g(t) = f(-t) \circ\!\!-\!\!\bullet\, G(\nu) = F(-\nu)}$$

Zur Begründung berechnet man mit der Substitution $t' = -t$, $dt' = -dt$

$$\begin{aligned} G(\nu) &= \int_{-\infty}^{\infty} g(t) e^{-2\pi i \nu t}\, dt = \int_{-\infty}^{\infty} f(-t) e^{-2\pi i \nu t}\, dt = \\ &= \int_{-\infty}^{\infty} f(t') e^{2\pi i \nu t'}\, dt' = F(-\nu)\ . \end{aligned}$$

3. Symmetrie.

Die Eigenschaften

$$\boxed{\begin{aligned} f(-t) = f(t) \quad &\Longleftrightarrow \quad F(-\nu) = F(\nu) \\ f(-t) = -f(t) \quad &\Longleftrightarrow \quad F(-\nu) = -F(\nu) \end{aligned}}$$

sieht man genauso ein wie die unter „Spiegelung“ aufgeführten.

4. Konjugation.

Es ist

$$\boxed{g(t) := \overline{f(t)} \circ\!\!-\!\!\bullet\, G(\nu) = \overline{F(-\nu)}\ .}$$

Begründung:

$$G(\nu) = \int_{-\infty}^{\infty} \bar{f}(t) e^{-2\pi i \nu t}\, dt = \overline{\int_{-\infty}^{\infty} f(t) e^{2\pi i \nu t}\, dt}\ .$$

5. Reelle gerade / ungerade Funktionen.

Aus der Kombination von 3. und 4. ergibt sich

$$\boxed{\begin{aligned} f(-t) = f(t) \in \mathbb{R} \quad &\Longrightarrow \quad F(-\nu) = F(\nu) \in \mathbb{R} \\ f(-t) = -f(t) \in \mathbb{R} \quad &\Longrightarrow \quad F(-\nu) = -F(\nu) \in i\mathbb{R} \end{aligned}}$$

6. Zerlegung.

Immer ist

$$f(t) = \underbrace{\frac{1}{2}\left(f(t) + f(-t)\right)}_{\text{„gerade“}} + \underbrace{\frac{1}{2}\left(f(t) - f(-t)\right)}_{\text{„ungerade“}}$$

Das bedeutet für reellwertiges f, dass man immer eine Zerlegung der Fouriertransformierten in Real- und Imaginärteil erzielen kann.

7. Streckung und Stauchung.

Für $\gamma > 0$ sei $g(t) := f(\gamma t)$. Es ist

$$\boxed{g(t) := f(\gamma t) \circ\!\!-\!\!\bullet\, G(\nu) = \frac{1}{\gamma} F\left(\frac{\nu}{\gamma}\right) .}$$

Der Beweis ergibt sich durch Substitution $t' = t\gamma$ im Fourierintegral.

Diese Eigenschaft zeigt die Reziprozität der t- und der ν-Skala: Funktionen f mit „kurzer Lebenszeit“ (Impulse) haben ein breites Spektrum.

8. Translation.

Für ein $T \in \mathbb{R}$ ergibt sich

$$\boxed{g(t) := f(t - T) \circ\!\!-\!\!\bullet\, G(\nu) = e^{-2\pi i \nu T} F(\nu) .}$$

Zum Beweis muss man nur $t' = t - T$ im Fourierintegral substituieren.

9. Phasenverschiebung.

Für ein $\lambda \in \mathbb{R}$ ergibt sich

$$\boxed{g(t) := e^{2\pi i \lambda t} f(t) \circ\!\!-\!\!\bullet\, G(\nu) = F(\nu - \lambda) .}$$

Zum Beweis muss man lediglich das Fourierintegral ausschreiben.

Die Regeln 8 und 9 sind offenbar zueinander invers.

10. Ableitung nach t.

Es wird vorausgesetzt, dass f überall differenzierbar sei und auch seine Ableitung $\dot{f}$ die Voraussetzungen für das Fourier-Integraltheorem erfülle (etwa absolute Integrierbarkeit und stückweise stetige Differenzierbarkeit). Dann ist

$$\boxed{g(t) := \dot{f}(t) \circ\!\!-\!\!\bullet\, G(\nu) = 2\pi i \nu F(\nu) .}$$

Der Beweis besteht in der partiellen Integration des Fourierintegrals für $\dot{f}$.

Aufgrund dieses Zusammenhangs lassen sich lineare Differentialgleichungen mit konstanten Koeffizienten in algebraische Gleichungen „im Frequenzraum“ überführen. Darauf kommen wir im Abschnitt über die Laplace-Transformation zurück.

11. Ableitung nach ν.

Es wird vorausgesetzt, dass f „genügend stark abklingt“, das heißt: auch die Funktion $tf(t)$ soll die Voraussetzungen für das Fourier-Integraltheorem erfüllen (insbesondere soll sie absolut integrierbar sein). Dann ist

$$\boxed{g(t) := tf(t) \circ\!\!-\!\!\bullet\, G(\nu) = \frac{i}{2\pi} \frac{d}{d\nu} F(\nu) .}$$

Der Beweis ergibt sich aus

$$\begin{aligned} G(\nu) &= \int_{-\infty}^{\infty} t f(t) e^{-2\pi i \nu t}\, dt = \frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{d}{d\nu} \left(f(t) e^{-2\pi i \nu t} \right)\, dt \\ &= \frac{i}{2\pi} \frac{d}{d\nu} \int_{-\infty}^{\infty} f(t) e^{-2\pi i \nu t}\, dt . \end{aligned}$$

Hier wurde Differentiation und *uneigentliche* Integration vertauscht, was man ohne Weiteres nicht tun darf. Unter den Voraussetzungen

$$\int_{-\infty}^{\infty} |f(t)|\,dt < \infty \qquad \text{und} \qquad \int_{-\infty}^{\infty} |tf(t)|\,dt < \infty$$

ist die Vertauschung aber erlaubt.

Wenn also $tf(t)$ genügend schnell abklingt, ist die Fourier-Transformierte eine differenzierbare Funktion.

12. Skalarprodukt

Für Funktionen $f, g : \mathbb{R} \to \mathbb{C}$, die auf jedem endlichen Intervall stückweise stetig sind und $\int_{-\infty}^{\infty} |f(t)|\,dt < \infty$ bzw. $\int_{-\infty}^{\infty} |g(t)|\,dt < \infty$ erfüllen, definieren wir ein **Skalarprodukt**

$$\langle f \mid g \rangle \quad := \quad \int_{-\infty}^{\infty} \overline{f(t)} g(t)\,dt$$

und haben $\langle f \,|\, g \rangle < \infty$: mit f und g ist nämlich auch $\bar{f} \cdot g$ absolut integrierbar (ohne Beweis). Die beiden angeführten Voraussetzungen reichen auch aus für die Existenz der Fourier-Transformierten $F \bullet\!\!-\!\!\circ f$ und $G \bullet\!\!-\!\!\circ g$ (sie reichen *nicht* aus für die Gültigkeit des Fourier-Integraltheorems) und es gilt

$$\boxed{\langle f \mid g \rangle \quad = \quad \langle F \mid G \rangle\ .}$$

Der Beweis ist einfach, wenn man, wie im Folgenden getan, die Reihenfolge der Integration vertauschen darf: dies ist unter den gemachten Voraussetzungen der absoluten Integrierbarkeit von f und g tatsächlich erlaubt (ohne Beweis).

$$\begin{aligned}
\langle F \mid G \rangle &= \int_{-\infty}^{\infty} \overline{F(\nu)} G(\nu)\,d\nu = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \overline{f(t)} e^{2\pi i \nu t}\,dt \right) G(\nu)\,d\nu \\
&= \int_{-\infty}^{\infty} \overline{f(t)} \left(\int_{-\infty}^{\infty} e^{2\pi i \nu t} G(\nu)\,d\nu \right)\,dt = \int_{-\infty}^{\infty} \overline{f(t)} g(t)\,dt \\
&= \langle f \mid g \rangle
\end{aligned}$$

Aus der 12. Regel ergibt sich als unmittelbare Folgerung der

Satz von Plancherel: Unter den Voraussetzungen des Fourier-Integraltheorems ist

$$\int_{-\infty}^{\infty} |f(t)|^2\,dt \quad = \quad \int_{-\infty}^{\infty} |F(\nu)|^2\,d\nu$$

Der Satz von Plancherel entspricht der Parsevalschen Gleichung

$$\frac{1}{T} \int_0^T |f(t)|^2\,dt = \sum_{k=-\infty}^{\infty} |c_k(f)|^2$$

für periodische Funktionen.

In den Anwendungen ist $f(t)$ der zeitliche Verlauf eines Signals und $F(\nu)$ die „Signal-Verteilung“ im Frequenz-Band. Dann werden $|f(t)|^2$ als zeitliche Verteilung und $|F(\nu)|^2$ als Frequenz-Verteilung der Signalenergie aufgefasst. Die Verteilungen können unterschiedlich sein, aber die Gesamtenergie bleibt gleich.

5.3 Laplace-Transformation

Für viele Funktionen $f : (0, \infty) \to \mathbb{C}$ ist das uneigentliche Integral

$$\int_0^\infty e^{-st} f(t)\, dt = \lim_{\substack{\varepsilon \to +0 \\ \lambda \to +\infty}} \int_\varepsilon^\lambda e^{-st} f(t)\, dt$$

definiert. Dieses hängt dann vom Parameter s ab und definiert eine neue Funktion:

$$\boxed{F(s) := \int_0^\infty e^{-st} f(t)\, dt} \tag{5.15}$$

Die Funktion F heißt die **Laplace-Transformierte** von f.

Andere Schreibweisen für die Laplace-Transformierte sind

$$F(s) = \mathscr{L}\{f(t)\} \tag{5.16}$$

oder

$$f(t) \circ\!\!-\!\!\bullet\, F(s)\ . \tag{5.17}$$

Die Schreibweise (5.16) ist mathematisch bedenklich: transfomiert wird eine Funktion f in eine andere Funktion F und nicht ein Funktionswert $f(t)$; eigentlich sollte man besser schreiben $F = \mathscr{L}(f)$ und für die Funktionswerte $F(s) = (\mathscr{L}(f))(s)$, wir bleiben aber mit (5.16) bei der üblichen Konvention. Auch die zweite Schreibweise (5.17) ist nicht unkritisch, da sie genauso bereits für die Fourier-Transformierte verwendet wurde. Man sollte sie nur verwenden, wenn man sich sicher ist, dass keine Verwechslung mit der Fourier-Transformierten passieren kann.

Man nennt (5.17) die **L-Korrespondenz** mit $F(s)$ als **Bildfunktion** von $f(t)$ und $f(t)$ als **Urbildfunktion** von $F(s)$.

Wir führen einige Beispiele von Laplace-Transformierten an, die man einfach nachrechnen kann. Bis auf die vorletzte sind alle der folgenden Funktionen f stetig bei $t = 0$, so dass das Integral an seiner unteren Grenze nicht uneigentlich ist.

Weitere Beispiele findet man in Formelsammlungen.

$f(t)$	$F(s)$	Definitionsbereich
1	$\frac{1}{s}$	$s > 0$
t^n	$\frac{n!}{s^{n+1}}$	$s > 0$
$e^{\lambda t}$	$\frac{1}{s-\lambda}$	$s > \lambda$
$e^{i\omega t}$	$\frac{1}{s-i\omega}$	$s > 0$
$\cos(\omega t)$	$\frac{s}{s^2+\omega^2}$	$s > 0$
$\sin(\omega t)$	$\frac{\omega}{s^2+\omega^2}$	$s > 0$
$f(t) := \begin{cases} 1, & \text{für } 0 \le t < a \\ 0, & \text{sonst} \end{cases}$	$\frac{1-e^{-as}}{s}$	$s > 0$
$\frac{1}{\sqrt{t}}$	$\sqrt{\frac{\pi}{s}}$	$s > 0$
e^{t^2}	existiert nicht!	

Wir können in der Definition (5.15) ohne weiteres auch komplexe Zahlen $s \in \mathbb{C}$ zulassen. Mit $s = \sigma + i\omega$, $\sigma, \omega \in \mathbb{R}$ wird dann aus (5.15)

$$F(s) = \int_0^\infty e^{-\sigma t - i\omega t} f(t)\, dt = \int_0^\infty e^{-\sigma t} \cos(\omega t) f(t)\, dt - i \int_0^\infty e^{-\sigma t} \sin(\omega t) f(t)\, dt\,.$$

Zunächst sollen zwei Fragen beantwortet werden:

(a): Unter welchen Bedingungen an f ist die Definition (5.15) sinnvoll?

(b): Für welche komplexen Werte s ist die Bildfunktion F definiert?

Diese Fragen beantwortet der folgende

Satz 5.17 (Existenz der Laplace-Transfomierten)
Falls $f : (0, \infty) \to \mathbb{C}$ die folgenden Eigenschaften hat

(a): $|f(t)| \leq C_1/t^{1-\varepsilon}$ für feste Konstanten $C_1, \varepsilon > 0$ und für $t \to 0$,

(b): stückweise stetig mit Sprüngen höchstens an Stellen $0 < t_1 < t_2 < \ldots$, wobei die Folge (t_j) keinen Häufungspunkt haben darf (nur endlich viele Sprünge in jedem endlichen Intervall) und

(c): $|f(t)| \leq C_2 e^{\lambda t}$ für feste Konstanten $C_2, \lambda \in \mathbb{R}$ und für $t \to \infty$,

dann existiert das Integral (5.15) für alle komplexen $s \in \mathbb{C}$ mit Re $s > \lambda$.

Beweis. Die zweite Eigenschaft sichert die Existenz des Integrals $\int_\alpha^\beta \cdots$ für alle Werte $0 < \alpha < \beta < \infty$. Bleibt also zu überprüfen, ob $\lim\limits_{\alpha \overset{>}{\to} 0} \int_\alpha^c \cdots$ und $\lim\limits_{\beta \to \infty} \int_d^\beta \cdots$ existieren. Im ersten Fall ist wegen der ersten Voraussetzung für jedes beliebige s und für dazu genügend klein gewähltes $c > 0$

$$\left| \int_\alpha^c e^{-st} f(t)\, dt \right| \quad \leq \quad 2 \int_\alpha^c |f(t)|\, dt \quad \leq \quad \frac{2C_1}{\varepsilon} \left[t^\varepsilon \right]_\alpha^c \quad \overset{\alpha \to 0}{\to} \quad \frac{2C_1}{\varepsilon} c^\varepsilon \quad < \quad \infty \,,$$

also existiert das Integral an der unteren Grenze. An der oberen Grenze haben wir für $s = \sigma + i\omega \in \mathbb{C}$ mit Re $s = \sigma > \lambda$ für alle genügend großen t wegen der dritten Voraussetzung

$$\left| e^{-st} f(t) \right| \quad \leq C_2 e^{-\Delta t}, \quad \Delta := \sigma - \lambda > 0,$$

also für genügend groß gewähltes d

$$\left| \int_d^\beta e^{-st} f(t)\, dt \right| \quad \leq \quad \frac{C_2}{-\Delta} \left[e^{-\Delta t} \right]_d^\beta \quad \overset{\beta \to \infty}{\to} \quad \frac{C_2}{\Delta} e^{-\Delta d} \quad < \quad \infty \,,$$

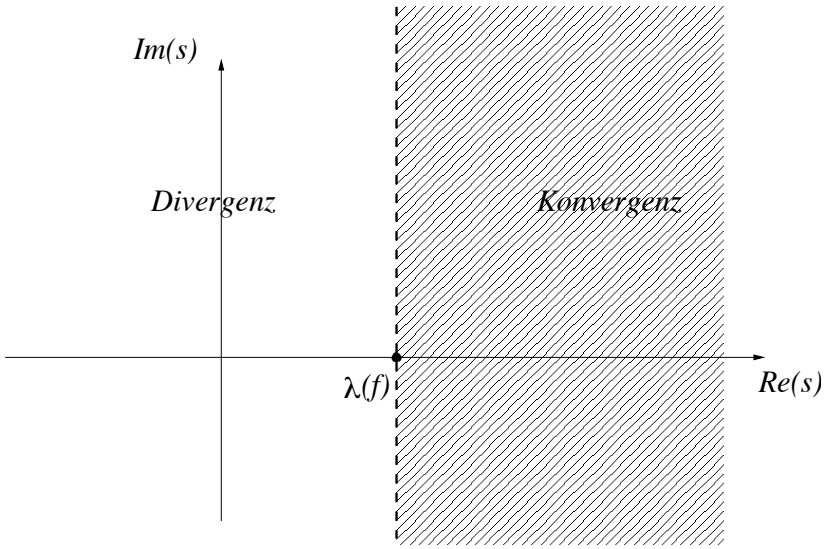
also existiert auch dieses Integral. q.e.d.

Nach Satz 5.17 existiert die Laplace-Transfomierte, wenn die drei genannten Voraussetzungen erfüllt sind. Diese Voraussetzungen sind *hinreichend*, aber nicht *notwendig*, das heißt es kann durchaus sein, dass die Laplace-Transformierte einer Funktion existiert, die die obigen Voraussetzungen verletzt. Dennoch beschränken wir uns im Folgenden auf Funktionen, die alle Voraussetzungen von Satz 5.17 erfüllen.

Mit der dritten Voraussetzung definieren wir

$$\boxed{\lambda(f) := \inf\{\lambda \in \mathbb{R};\ \lambda \text{ erfülle 3. in Satz 5.17}\}\,.} \tag{5.18}$$

Gemäß dem für Satz 5.17 erbrachten Beweis konvergiert das Integral (5.15) für alle $s \in \mathbb{C}$ mit Re $s > \lambda(f)$. Genauso divergiert es für alle $s \in \mathbb{C}$ mit Re $s < \lambda(f)$. Damit haben wir das folgende Konvergenzgebiet:



Auf der Grenze Re $s = \lambda(f)$ des Konvergenzgebiets kann (5.15) für manche Werte s existieren und für andere nicht. Im Sonderfall $\lambda(f) = -\infty$ existiert die Laplace-Transformation für alle $s \in \mathbb{C}$.

Die Frage der Umkehrbarkeit der Laplace-Transformation ist von großer Bedeutung, da wir später Differentialgleichungen für eine Funktion f in Gleichungen für die Transformierte $F(s) \bullet\!\!-\!\!\circ f(t)$ umwandeln wollen und die gefundenen Lösungen danach wieder in den Ausgangsraum zurücktransformieren müssen. Eine Umkehrformel für die Laplace-Transformation lässt sich mithilfe der Umkehrformel für die Fourier-Transformation angeben. Die entsprechende Aussage lautet:

Satz 5.18 (Umkehrung der Laplace-Transformierten)
Die Funktion $f : [0, \infty) \mapsto \mathbb{C}$ erfülle die folgenden beiden Bedingungen:

(i) f sei stetig differenzierbar und

(ii) es sei Bedingung (3) von Satz 5.17 erfüllt.

Dann gilt für die L-Korrespondenz $f(t) \circ\!\!-\!\!\bullet F(s)$ und für $s = \sigma + i\omega \in \mathbb{C}$ mit $\sigma > \lambda(f)$ die Umkehrformel

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{(\sigma+i\omega)t} F(\sigma + i\omega)\, d\omega \; .$$

Man muss also die Bildfunktion auf einer Geraden $\sigma + i\mathbb{R}$ mit $\sigma > \lambda(f)$ kennen, um einen Punkt $f(t)$ der Urbildfunktion ausrechnen zu können.

Beweis. Wir setzen zuerst f durch die Festlegung

$$\tilde{f}(t) := \begin{cases} f(t), & t \geq 0 \\ 0, & t < 0 \end{cases}$$

auf die ganze reelle Achse fort. Statt $\tilde{f}$ schreiben wir im Folgenden aber wieder f und haben

$$G(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} \left(e^{-\sigma t} f(t)\right)\, dt = F(\sigma + i\omega) \; ,$$

wobei $G(\omega)$ die Fourier-Transformierte der Funktion $\exp(-\sigma t)f(t)$ ist.[1] Nach den Voraussetzungen des Satzes ist $\exp(-\sigma t)f(t)$ absolut integrierbar und stetig differenzierbar, so dass nach der Umkehrformel für die Fourier-Transformierte (Satz 5.16, Bemerkung 7.) folgt

$$\begin{aligned} e^{-\sigma t}f(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega t}G(\omega)\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega t}F(\sigma+i\omega)\,d\omega\ . \end{aligned}$$

q.e.d.

Hierzu ein paar Bemerkungen.

(a): Wenn man nur voraussetzt, f sei *stückweise* stetig differenzierbar mit maximal endlich vielen Sprungstellen/Knicken in jedem endlichen Intervall, dann gilt der Satz immer noch, außer eventuell an den Sprungstellen/Knicken.

(b): Aus der Umkehrformel folgt insbesondere auch, dass eine Funktion durch ihre Laplace-Transformierte – außer an Sprungstellen – eindeutig festgelegt ist. In Formeln: unter den Voraussetzungen des Satzes sei $f_1(t) \circ\!\!-\!\!\bullet\ F_1(s)$ und $f_2(t) \circ\!\!-\!\!\bullet\ F_2(s)$. Dann gilt

$$F_1(s) = F_2(s) \quad \forall s \quad \text{mit Re } s > \max\{\lambda(f_1), \lambda(f_2)\} \quad \Rightarrow \quad f_1(t) = f_2(t)\ ,$$

wobei eventuelle Sprungstellen ausgenommen werden müssen.

(c): Es ist gut zu wissen, dass es eine Umkehrformel gibt, in der Praxis benutzt wird sie aber kaum. Stattdessen schlägt man Korrespondenzen $f(t) \circ\!\!-\!\!\bullet\ F(s)$ in Tabellen nach und benutzt die folgenden Rechenregeln, auf die der große Nutzen der Laplace-Transformation vor allem zurückzuführen ist.

Wir geben 8 Rechenregel für die Laplace-Transformation an, die uns in den Anwendungen nützlich sein werden. **Generell wird angenommen, dass alle Funktionen f, die transformiert werden sollen, die Voraussetzungen des Satzes 5.17 erfüllen. Wo nötig, soll darüber hinaus $\lim_{t\to 0} f(t) =: f(0+)$ existieren.**

(A) Linearität.

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

$$\boxed{\mathscr{L}\{\alpha f(t) + \beta g(t)\} = \alpha\mathscr{L}\{f(t)\} + \beta\mathscr{L}\{g(t)\}\ .}$$

Beispiel: $\mathscr{L}\{\cosh\omega t\} = \frac{1}{2}\mathscr{L}\{e^{\omega t}\} + \frac{1}{2}\mathscr{L}\{e^{-\omega t}\} = \frac{s}{s^2-\omega^2}$ für Re $s > \omega$.

(B) Streckung.

Es ist für $c > 0$

$$\boxed{\mathscr{L}\{f(ct)\} = \frac{1}{c}F\left(\frac{1}{c}s\right)\ , \quad \text{Re } s > c\lambda(f)\ .}$$

Zum Beweis macht man die Substitution $t' = ct$ im Integral (5.15).

[1] Es handelt sich um die Fourier-Transformierte gemäß der Bemerkung 7. im Anschluß an Satz 5.16

Beispiel: man kann in einer L-Tabelle die Korrespondenz $\cos t \;\circ\!\!-\!\!\bullet\; \dfrac{s}{s^2+1}$ finden und bekommt daraus die Korrespondenz $\cos \omega t \;\circ\!\!-\!\!\bullet\; \dfrac{1}{\omega}\left(\dfrac{s/\omega}{(s/\omega)^2+1}\right) = \dfrac{s}{s^2+\omega^2}$ ohne weitere Rechnung.

(C) Transformation von Ableitung und Integral.

Für differenzierbares f ist

$$\boxed{\dot f(t) \;\circ\!\!-\!\!\bullet\; sF(s) - f(0+)\,, \quad \text{Re } s > \lambda(f)\,,}$$

wobei $f(0+)$ den rechtsseitigen Funktionswert von f an der Stelle 0 bezeichnet.

Beweis. Es ist (5.15) partiell zu integrieren:

$$\int_0^\infty e^{-st}\dot f(t)\,dt = \lim_{c\to\infty,\varepsilon\to 0+}\left[e^{-st}f(t)\right]_\varepsilon^c - \int_0^\infty (-s)e^{-st}f(t)\,dt\,.$$

q.e.d.

Das lässt sich dann auch gleich verallgemeinern für die n-te Ableitung:

$$\boxed{f^{(n)}(t) \;\circ\!\!-\!\!\bullet\; s^nF(s) - s^{n-1}f(0+) - s^{n-2}\dot f(0+) - \ldots - f^{(n-1)}(0+)\,,}$$

wenn f genügend oft differenzierbar ist und die Grenzwerte der Ableitungen bei 0 existieren.

Für die Integration ist

$$\boxed{\int_0^t f(\tau)\,d\tau \;\circ\!\!-\!\!\bullet\; \frac{1}{s}F(s)\,.}$$

Beweis. Man beachte, dass

$$g(t) := \int_0^t f(\tau)\,d\tau \quad\Rightarrow\quad \dot g(t) = f(t),\ g(0+) = 0$$

und wende die Formel für die Korrespondenz der Ableitung an. q.e.d.

Genau diese Formel für Korrespondenz von Ableitungen ist es, die die L-Transformation so wichtig zur Behandlung von DGL macht. DGL im Zeitbereich werden in einfache algebraische Gleichungen im Bildbereich verwandelt, dort gelöst und dann die Lösung rücktransformiert. Bevor wir diese Anwendung bringen, führen wir erst noch unsere Liste von Rechenregeln zu Ende.

(D) Ableitung und Integration der Bildfunktion.

Es ist, jeweils für Re $s > \lambda(f)$ und für ein $\varepsilon > 0$ in der letzten Formel,

$$\boxed{\begin{aligned}
\mathscr{L}\{tf(t)\} &= -\frac{d}{ds}F(s)\\
\mathscr{L}\{t^nf(t)\} &= (-1)^n\frac{d^n}{ds^n}F(s)\\
\mathscr{L}\{\frac{1}{t}f(t)\} &= \int_s^\infty F(s')\,ds' \quad \text{falls } |f(t)| \le C_1t^\varepsilon \text{ für } t \overset{>}{\to} 0
\end{aligned}}$$

Beweis. Wir bringen den Beweis nur für reelles s. Mit Differentiation unter dem Integralzeichen ergibt sich aus (5.15):

$$\frac{d}{ds}F(s) = \int_0^\infty \frac{\partial}{\partial s} e^{-st} f(t)\, dt = -\int_0^\infty e^{-st} t f(t)\, dt\ .$$

Hier haben wir unter dem Integral differenziert, was trotz des uneigentlichen Integrals erlaubt ist, weil es für $s > \lambda(f) = \lambda$ ein $\varepsilon > 0$ gibt mit $s > \lambda + \varepsilon$ und damit

$$\left|\frac{\partial}{\partial s} e^{-st} f(t)\right| \le \left|C_2 e^{(\lambda - s)t} t\right| \le C_2 t e^{-\varepsilon t}$$

und das uneigentliche Integral dieser von s unabhängigen Funktion konvergiert. Wenn man die Differentiation n-mal wiederholt, ergibt sich die zweite Formel.

Zum Beweis der dritten Formel wird (5.15) integriert:

$$\begin{aligned}\int_s^\infty F(s')\, ds' &= \int_s^\infty \int_0^\infty e^{-s't} f(t)\, dt\, ds' = \int_0^\infty \left(\int_s^\infty e^{-s't}\, ds'\right) f(t)\, dt \\ &= \int_0^\infty \left[-\frac{1}{t} e^{-s't}\right]_s^\infty = \int_0^\infty e^{-st} \frac{1}{t} f(t)\, dt\ .\end{aligned}$$

Wir bemerken noch, dass bei der dritten Formel eine hinreichende Bedingung zur Existenz des Integrals $\mathscr{L}\{f(t)/t\}$ an der unteren Grenze 0 die folgende Bedingung ist:

$$\left|\frac{1}{t} f(t)\right| \le \frac{C_1}{t^{1-\varepsilon}} \quad \Leftrightarrow \quad |f(t)| \le C_1 t^\varepsilon\ .$$

q.e.d.

Als Beispiel zur Regel (D) betrachte man die Korrespondenzen für Re $s >$ Re a:

$$\begin{aligned} e^{at} &\quad\circ\!\!-\!\!\bullet\quad \frac{1}{s-a} \\ te^{at} &\quad\circ\!\!-\!\!\bullet\quad \frac{1}{(s-a)^2} \\ \frac{t^2}{2} e^{at} &\quad\circ\!\!-\!\!\bullet\quad \frac{1}{(s-a)^3} \\ \vdots &\qquad\quad \vdots \\ \frac{t^n}{n!} e^{at} &\quad\circ\!\!-\!\!\bullet\quad \frac{1}{(s-a)^{n+1}} \end{aligned}$$

(E) Dämpfung und Verschiebung.

In den folgenden Formeln sei stets $f(t) = 0$ für $t < 0$ vereinbart, so dass $f(t-a)$ nur für $t \ge a$ ungleich Null sein kann:

$$\boxed{\begin{aligned}\mathscr{L}\{e^{-at} f(t)\} &= F(s+a), \quad \text{Re } s > \lambda(f) - a \\ \mathscr{L}\{f(t-a)\} &= e^{-as} F(s), \quad \text{Re } s > \lambda(f), a \ge 0\end{aligned}}$$

Beweis. Beide Aussagen ergeben sich wieder direkt aus (5.15). Zunächst die erste

$$\int_0^\infty e^{-st} e^{-at} f(t)\, dt = \int_0^\infty e^{-(s+a)t} f(t)\, dt$$

und dann, mit der Substitution $t' = t - a$, auch die zweite:

$$\int_0^\infty e^{-st} f(t-a)\,dt = \int_a^\infty e^{-st} f(t-a)\,dt = \int_0^\infty e^{-s(t'+a)} f(t')\,dt' = e^{-sa} \int_0^\infty e^{-st} f(t)\,dt\ .$$

(F) Periodisches f.

Wenn

$$f(t-T) = f(t) \qquad \forall t \geq T > 0$$

heißt f periodisch mit Periode T. Genauer gesagt wird die Periodizität nur im Bereich der positiven Achse gefordert. In diesem Fall gilt für die L-Transformierte

$$F(s) = \sum_{n=0}^\infty \int_{nT}^{(n+1)T} e^{-st} f(t)\,dt = \sum_{n=0}^\infty e^{-snT} \int_0^T e^{-st} f(t)\,dt\ .$$

Folglich ist

$$\boxed{F(s) = \frac{1}{1-e^{-sT}} \int_0^T e^{-st} f(t)\,dt, \quad \text{Re } s > 0\ .}$$

(G) Faltung.

Zu je zwei Funktionen $f, g : (0,\infty) \to \mathbb{R}$ definiert man das Faltungsprodukt (kurz: die **Faltung**) als Funktion $f * g : (0,\infty) \to \mathbb{R}$ über

$$\boxed{(f*g)(t) := \int_0^t f(\tau) g(t-\tau)\,d\tau = \int_0^t f(t-\tau) g(\tau)\,d\tau\ .}$$

Einige Bemerkungen zur Faltung:

(a): Erfüllen f und g die Bedingungen von Satz (5.17), so auch $f * g$.

(b): Es gelten die Rechenregeln $f*g = g*f$, $f*(g*h) = (f*g)*h$, $f*(g+h) = f*g + f*h$ und $0 * f = f * 0 = 0$.

(c): Es ist $1 * f = f * 1 \neq f$. Das Einselement des Faltungsprodukts ist also *nicht* die Funktion $g \equiv 1$.

(d): Es gibt andere Definitionen von den Faltungen, etwa

$$\int_{-\infty}^\infty f(\tau) g(t-\tau)\,d\tau, \quad \text{falls} \quad \int_{-\infty}^\infty |f(t)|\,dt < \infty,\ \int_{-\infty}^\infty |g(t)|\,dt < \infty,$$

als Faltung von f und g definiert. In diesem Fall gilt eine zur folgenden Formel (5.19 analoge Formel, wenn man statt der Laplace- die Fourier-Transformation benutzt.

Für die Faltung gilt die Produktformel:

$$\boxed{(f*g)(t) \circ\!\!-\!\!\bullet\ F(s)G(s), \quad \text{Re } s > \max\{\lambda(f), \lambda(g)\}\ .} \tag{5.19}$$

Beweis:

$$\begin{aligned}
\mathscr{L}\{(f*g)(t)\} &= \int_0^\infty e^{-st}\left(\int_0^t f(\tau)g(t-\tau)\,d\tau\right)\,dt \\
&= \int_0^\infty\int_0^t e^{-s\tau}e^{-s(t-\tau)}f(\tau)g(t-\tau)\,d\tau\,dt \\
&= \int_0^\infty\int_\tau^\infty e^{-s\tau}e^{-s(t-\tau)}f(\tau)g(t-\tau)\,dt\,d\tau \\
&= \int_0^\infty e^{-s\tau}f(\tau)\left(\int_\tau^\infty e^{-s(t-\tau)}g(t-\tau)\,dt\right)\,d\tau \\
&= \int_0^\infty e^{-s\tau}f(\tau)\left(\int_0^\infty e^{-st'}g(t')\,dt'\right)\,d\tau \\
&= \int_0^\infty e^{-s\tau}f(\tau)\,d\tau\cdot\int_0^\infty e^{-st'}g(t')\,dt' = F(s)G(s)
\end{aligned}$$

q.e.d.

(H) Diracsche Delta-Funktion.

Die L-Transformierte von

$$f_\varepsilon(t) := \begin{cases} 0, & t<0 \\ \frac{1}{\varepsilon}, & 0\le t<\varepsilon \\ 0, & \varepsilon\le t<\infty \end{cases}$$

ergibt sich zu

$$F_\varepsilon(s) = \int_0^\varepsilon e^{-st}\frac{1}{\varepsilon}\,dt = \frac{1-e^{-\varepsilon s}}{\varepsilon s}, \quad \text{Re } s>0\ .$$

Es ist also

$$\lim_{\varepsilon\to 0} F_\varepsilon(s) = 1, \quad \text{Re } s>0\ ,$$

obwohl

$$\lim_{\varepsilon\to 0} f_\varepsilon(t)$$

nicht existiert. Formell wird das Urbild der 1 unter der L-Transformation dennoch eingeführt als sogenannte „Deltafunktion“ δ, indem man die folgende Korrespondenz angibt:

$$\delta(t) \circ\!\!-\!\!\bullet\ 1\ .$$

Diese „Funktion“ ist nicht wirklich eine, sondern eine sogenannte „Distribution“, die sinnvoll nur in Integralen mit stetigen Funktionen geschrieben werden kann:

$$\int_{-\infty}^\infty \delta(t)f(t)\,dt = f(0)\ .$$

Aus dieser letzten Formel gewinnt man dann auch formal direkt die Einsfunktion als Laplace-Transformierte der Delta-Funktion. Zusammen mit Regel (G) ergibt sich auch, dass die Delta-Funktion das neutrale Element des Faltungsprodukts ist.

Anwendung der Laplace-Transformation: Lösen von Differentialgleichungen

Wir benutzen die Laplace-Transformation als ein Hilfsmittel zur Lösung von Anfangswertproblemen für lineare Differentialgleichungen mit konstanten Koeffzienten. Dies ist im Prinzip auch schon mit der Fourier-Transformation möglich, wie wir bereits früher gesehen haben, allerdings

gibt es dabei eine große Einschränkung: damit die Fourier-Transformierte einer Funktion f überhaupt existiert, muss $\int_{-\infty}^{\infty} |f(t)|\, dt < \infty$ sein, muss die Funktion f also sehr schnell abklingen. Genau diese Einschränkung existiert bei der Laplace-Transformation nicht mehr.

Wir betrachten das folgende AWP für DGL mit konstanten Koeffizienten: gesucht ist eine Funktion $y(t),\ t \geq 0$, die die Gleichung

$$a_n y^{(n)}(t) + \ldots + a_1 \dot{y}(t) + a_0 y(t) = b(t)$$

erfüllt und die folgenden Anfangswerte hat:

$$y(0) = y_0, \dot{y}(0) = y_1, \ldots, y^{(n-1)}(0) = y_{n-1}\ .$$

Es wird vorausgesetzt, dass alle Koeffizienten a_i konstant sind und $a_n \neq 0$.

Setzt man wie üblich

$$B(s) \bullet\!\!-\!\!\circ\, b(t), \qquad Y(s) \bullet\!\!-\!\!\circ\, y(t)$$

als die L-Transformierten von b und (dem unbekannten) y an, dann ist nach Regel (C)

$$\begin{array}{rcl} y(t) & \circ\!\!-\!\!\bullet & Y(s), \\ \dot{y}(t) & \circ\!\!-\!\!\bullet & sY(s) - y_0, \\ \vdots & & \vdots \\ y^{(n-1)}(t) & \circ\!\!-\!\!\bullet & s^{n-1}Y(s) - y_{n-2} - \ldots - s^{n-2}y_0, \\ y^{(n)}(t) & \circ\!\!-\!\!\bullet & s^n Y(s) - y_{n-1} - \ldots - s^{n-2}y_1 - s^{n-1}y_0, \end{array}$$

also ist

$$a_n y^{(n)}(t) + \ldots + a_1 \dot{y}(t) + a_0 y(t) \circ\!\!-\!\!\bullet\, P(s)Y(s) - P_1(s)y_0 - \ldots - P_n(s)y_{n-1},$$

wobei

$$\begin{array}{rcl} P(s) & = & a_n s^n + \ldots + a_1 s + a_0, \\ P_1(s) & = & a_n s^{n-1} + \ldots + a_1, \\ \vdots & & \vdots \\ P_n(s) & = & a_n\ . \end{array}$$

Man erhält demnach die Formel

$$Y(s) = \frac{1}{P(s)} \left(B(s) + P_1(s)y_0 + \ldots + P_n(s)y_{n-1}\right) \tag{5.20}$$

und muss jetzt nur noch rücktransformieren: $Y(s) \bullet\!\!-\!\!\circ\, y(t)$. Wenn die rechte Seite in (5.20) eine rationale Funktion ist, kann man (nach einer Partialbruchzerlegung) die Rücktransformation in Tabellen nachschlagen und muss sie nicht berechnen.

Beispiel(e) 5.19
Gesucht ist eine Lösung des AWP

$$\ddot{y} + 4y = \sin(\omega t), \quad y(0) = c_1, \quad \dot{y}(0) = c_2, \quad \omega > 0 \; .$$

Im ersten Schritt wird das Problem transformiert:

$$s^2 Y(s) + 4Y(s) = \frac{\omega}{s^2 + \omega^2} + sc_1 + c_2 \; ,$$

wobei $y(t) \circ\!\!-\!\!\bullet\, Y(s)$ und $\sin(\omega t) \circ\!\!-\!\!\bullet\, \omega/(s^2 + \omega^2)$ benutzt wurde.

Diese Gleichung wird dann nach $Y(s)$ aufgelöst:

$$Y(s) = \frac{\omega}{(s^2 + \omega^2)(s^2 + 4)} + c_1 \frac{s}{s^2 + 4} + c_2 \frac{1}{s^2 + 4} \; .$$

Die Rücktransformation findet man eventuell direkt in einer L-Tabelle – ansonsten mit PBZ:

$$\frac{\omega}{(s^2 + \omega^2)(s^2 + 4)} = \begin{cases} \dfrac{\omega}{4 - \omega^2} \left(\dfrac{1}{s^2 + \omega^2} - \dfrac{1}{s^2 + 4} \right), & \omega^2 \neq 4 \\ \dfrac{2}{(s^2 + 4)^2}, & \omega^2 = 4 \end{cases}$$

und der Formel

$$\frac{1}{(s^2 + a^2)^2} \bullet\!\!-\!\!\circ\, \frac{\sin(at) - at\cos(at)}{2a^3}$$

sowie der L-Tabelle vom Anfang des Kapitels. Daraus erhält man

$$y(t) = c_1 \cos(2t) + \frac{c_2}{2} \sin(2t) + \begin{cases} \dfrac{1}{4 - \omega^2} \left(\sin(\omega t) - \dfrac{\omega}{2} \sin(2t) \right), & \omega^2 \neq 4 \\ \dfrac{1}{8}(\sin(2t) - 2t\cos(2t)), & \omega^2 = 4 \end{cases}$$

Auch Systeme von DGL mit konstanten Koeffizienten können wir mit derselben Methode behandeln.

Beispiel(e) 5.20
Wir betrachten

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix} \mathbf{x} + e^{5t} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \quad \mathbf{x}(0) = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}.$$

Hier wenden wir die L-Transformation auf jede Komponente von $\mathbf{x}$ und der rechten Seite $\mathbf{b}$ an – also

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b}(t) \circ\!\!-\!\!\bullet\ s\mathbf{X}(s) - \mathbf{x}(0) = A\mathbf{X}(s) + \mathbf{B}(s)$$

und erhalten so das LGS

$$\begin{pmatrix} s & -1 & 1 \\ 2 & s-3 & 1 \\ 1 & -1 & s-1 \end{pmatrix} \mathbf{X}(s) = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} + \frac{1}{s-5} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}.$$

Dieses LGS muss gelöst werden, etwa mithilfe der Gauß-Elimination. Man erhält so für die drei Komponenten X_i von $\mathbf{X}$ die folgenden L-Korrespondenzen

$$\begin{aligned} X_3(s) = \frac{2(s-2)}{(s-1)(s-2)} = \frac{2}{s-1} &\quad \bullet\!\!-\!\!\circ \quad x_3(t) = 2e^t \\ X_2(s) = \frac{1}{s-5} - \frac{2}{(s-1)^2} &\quad \bullet\!\!-\!\!\circ \quad x_2(t) = e^{5t} - 2te^t \\ X_1(s) = -\frac{2}{(s-1)^2} &\quad \bullet\!\!-\!\!\circ \quad x_1(t) = -2te^t \end{aligned}$$

Teil III

Mathematik 3

Kapitel 6

Analysis mehrerer reeller Veränderlicher

6.1 Differentiation

Im Folgenden untersuchen wir Funktionen $f : \mathbb{D} \to \mathbb{R}^m$, $\mathbb{D} \subseteq \mathbb{R}^n$, die jedem Spaltenvektor $x \in \mathbb{R}^n$ einen mit $f(x)$ bezeichneten Spaltenvektor im $\mathbb{R}^m$ zuordnen:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto f(x) = f(x_1, \ldots, x_n) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}. \tag{6.1}$$

Ist $n = 1$, also $f : \mathbb{D} \to \mathbb{R}^m$, $\mathbb{D} \subseteq \mathbb{R}$, so spricht man von Kurven im $\mathbb{R}^m$,
ist $m = 1$, also $f : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} \subseteq \mathbb{R}^n$, so spricht man von Skalarenfeldern, ansonsten von Vektorfeldern.
Zunächst betrachten wir spezielle Kurven im $\mathbb{R}^m$, nämlich auf einem Intervall $I \subseteq \mathbb{R}$ definierte Funktionen $\xi : I \to \mathbb{R}^m, t \mapsto \xi(t)$. Der für die Analysis grundlegende Konvergenzbegriff wird komponentenweise erklärt:

$$\lim_{t \to t_0} \xi(t) = c \in \mathbb{R}^m :\Longleftrightarrow \lim_{t \to t_0} \xi_i(t) = c_i \in \mathbb{R} \quad \text{für alle } i = 1, \ldots, m. \tag{6.2}$$

Daher heißt ξ in $t_0 \in I$ stetig bzw. differenzierbar, wenn alle Komponentenfunktionen in $t_0 \in I$ stetig bzw. differenzierbar sind. Somit ist auch die Ableitung komponentenweise zu berechnen.

$$\dot{\xi}(t) := \frac{d}{dt}\xi(t) = \lim_{h \to 0} \frac{1}{h}(\xi(t+h) - \xi(t)) = \begin{pmatrix} \dot{\xi}_1(t) \\ \vdots \\ \dot{\xi}_m(t) \end{pmatrix}. \tag{6.3}$$

Der Vektor $\dot{\xi}(t)$ wird als Tangentialvektor von ξ an der Stelle $t \in I$ bezeichnet. Für zwei Kurven $\xi, \eta : I \to \mathbb{R}^m$ gilt offensichtlich für alle $\alpha, \beta \in \mathbb{R}$:

(a): $\frac{d}{dt}(\alpha\xi(t) + \beta\eta(t)) = \alpha\dot{\xi}(t) + \beta\dot{\eta}(t)$

(b): $\frac{d}{dt}(\xi(t)^\top \eta(t)) = \dot{\xi}(t)^\top \eta(t) + \xi(t)^\top \dot{\eta}(t)$

(c): für $m = 3$: $\frac{d}{dt}(\xi(t) \times \eta(t)) = \dot{\xi}(t) \times \eta(t) + \xi(t) \times \dot{\eta}(t)$

(d): $\frac{d}{dt}(\alpha(t) \cdot \xi(t)) = \dot{\alpha}(t) \cdot \xi(t) + \alpha(t) \cdot \dot{\xi}(t)$, wobei $\alpha : I \to \mathbb{R}$ eine differenzierbare Funktion darstellt.

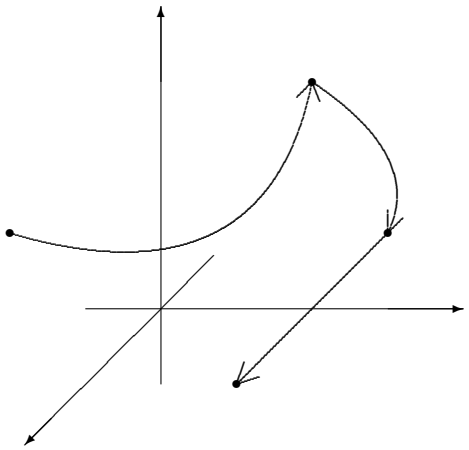
(6.4)

Aus b) folgt:

$$|\xi(t)| \equiv c \Longleftrightarrow \frac{d}{dt}|\xi(t)| = \frac{d}{dt}|\xi(t)|^2 \equiv 0 \Longleftrightarrow \xi(t)^\top \dot{\xi}(t) \equiv 0. \tag{6.5}$$

Definition 6.1 (Kurvenstück, regulär, Spur)
Sei $\mathbb{G} \subseteq \mathbb{R}^m$ und $[a,b] \subset \mathbb{R}$ ein abgeschlossenes Intervall, so wird jede stetig differenzierbare Funktion $\xi : [a,b] \to \mathbb{G}$ als Kurvenstück in $\mathbb{G}$ mit Anfangspunkt $\xi(a)$ und Endpunkt $\xi(b)$ bezeichnet. $\dot{\xi}(a)$ und $\dot{\xi}(b)$ sind als einseitige Ableitungen zu verstehen. Ein Kurvenstück ξ heißt regulär, falls $\dot{\xi}(t) \neq 0$ für alle $t \in [a,b]$. Die Menge $\{\xi(t); a \leq t \leq b\}$ heißt Spur des Kurvenstückes ξ.

Unter einer Kurve versteht man im allgemeinen eine Kette von Kurvenstücken



(6.6)

Ausgehend von einem regulären Kurvenstück $\xi : [a,b] \to \mathbb{R}^m$ heißt

$$s(t) := \int_a^t |\dot{\xi}(\tau)| d\tau \tag{6.7}$$

die Bogenlänge des Kurvenstückes über $[a,t]$, $t \in [a,b]$, und der Vektor

$$T(t) := \frac{\dot{\xi}(t)}{|\dot{\xi}(t)|} \tag{6.8}$$

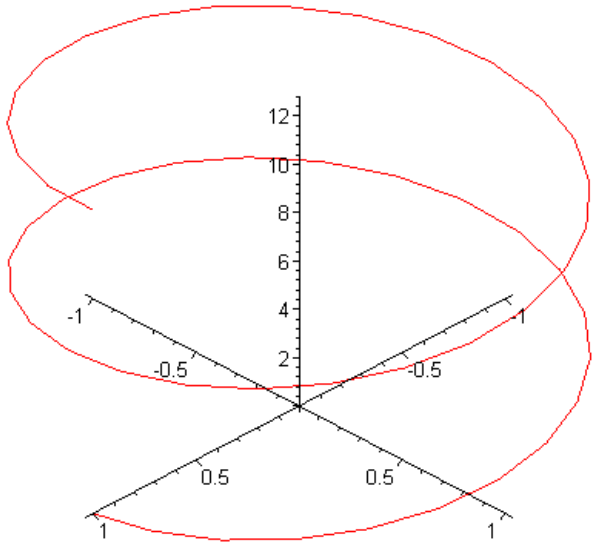
heißt Tangenteneinheitsvektor in $t \in [a,b]$. Mit

$$\kappa(t) := \frac{|\dot{T}(t)|}{|\dot{\xi}(t)|} \tag{6.9}$$

wird die Krümmung an der Stelle t bezeichnet, wobei ξ nun zweimal stetig differenzierbar sein muss. Die Krümmung ist ein Maß für die Änderung des Tangenteneinheitsvektors.

Beispiel(e) 6.2

$$\xi : [0, 2n\pi] \to \mathbb{R}^3,\ t \mapsto \begin{pmatrix} r \cdot \cos(t) \\ r \cdot \sin(t) \\ ht \end{pmatrix},\ n \in \mathbb{N}$$



(6.10)

$$|\dot{\xi}(t)| = \sqrt{r^2(\sin^2(t) + \cos^2(t)) + h^2} = \sqrt{r^2 + h^2}. \tag{6.11}$$

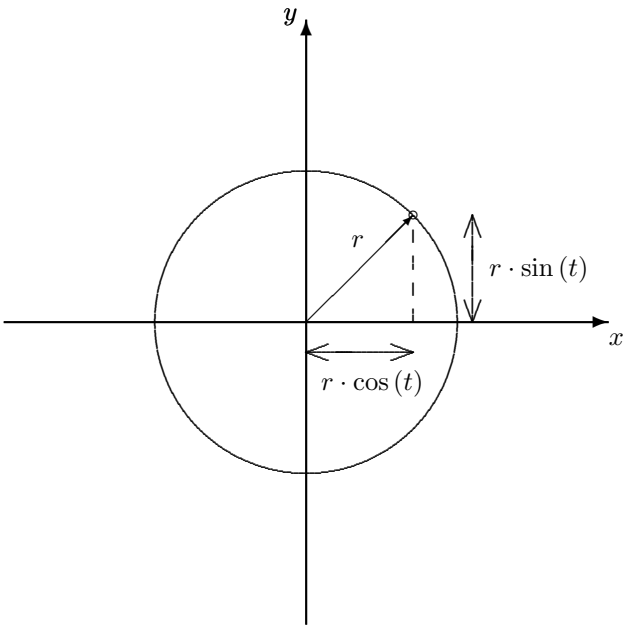
Somit beschreibt

$$s : [0, 2n\pi] \to \mathbb{R}, \quad t \mapsto t\sqrt{r^2 + h^2} \quad \text{die funktionale Darstellung der Bogenlänge} \tag{6.12}$$

und

$$\kappa : [0, 2n\pi] \to \mathbb{R}, \quad t \mapsto \frac{r}{r^2 + h^2} \quad \text{die funktionale Darstellung der Krümmung.} \tag{6.13}$$

Beispiel(e) 6.3
$\xi : [0, 2\pi] \to \mathbb{R}^2$, $t \mapsto \begin{pmatrix} r \cdot \cos(t) \\ r \cdot \sin(t) \end{pmatrix}$



(6.14)

Spur von ξ ist ein Kreis um $(0, 0)$ mit Radius r. Ferner gilt:

$$\dot{\xi} : [0, 2\pi] \to \mathbb{R}^2, \quad t \mapsto \begin{pmatrix} -r \cdot \sin(t) \\ r \cdot \cos(t) \end{pmatrix} \quad |\dot{\xi}(t)| = r \quad \text{für alle } t \in [0, 2\pi], \tag{6.15}$$

Bogenlänge und Krümmung:

$$s : [0, 2\pi] \to \mathbb{R}, \quad t \mapsto t \cdot r, \quad \kappa : [0, 2\pi] \to \mathbb{R}, \quad t \mapsto \frac{r}{r^2} = \frac{1}{r}. \tag{6.16}$$

Nun betrachten wir Skalarenfelder oder reellwertige Funktionen mehrerer reeller Veränderlicher $f : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} \subseteq \mathbb{R}^n$. Um Abbildungen dieser Art diskutieren zu können, hat man zunächst wie im Fall $\mathbb{D} \subseteq \mathbb{R}$ den Grenzwert und Stetigkeitsbegriff einzuführen. Dazu benötigen wir einige Begriffe aus der Topologie.

Definition 6.4 (innerer Punkt, offene Menge, Rand, abgschlossene Menge)
Sei $\mathbb{D} \subseteq \mathbb{R}^n$. Ein Punkt $a \in \mathbb{D}$ heißt innerer Punkt von $\mathbb{D}$, falls es ein $r > 0$ gibt mit

$$\{x \in \mathbb{R}^n; |x - a| < r\} \subseteq \mathbb{D}. \tag{6.17}$$

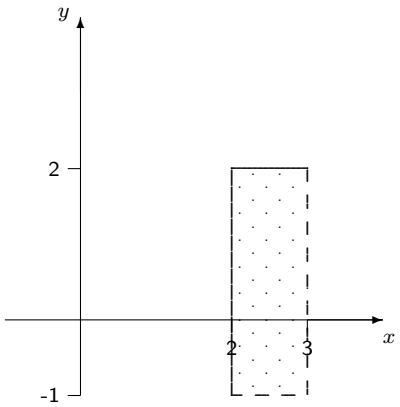
$\mathbb{D}$ heißt offen, falls jeder Punkt von $\mathbb{D}$ ein innerer Punkt ist. Ein Punkt $b \in \mathbb{R}^n$ heißt Randpunkt von $\mathbb{D}$, falls jede r-Kugel

$$K_r(b) := \{x \in \mathbb{R}^n;\ |x - b| < r\}, \quad r > 0, \tag{6.18}$$

mindestens einen Punkt aus $\mathbb{D}$ und mindestens einen nicht zu $\mathbb{D}$ gehörenden Punkt enthält.
Die Menge aller Randpunkte von $\mathbb{D}$ heißt Rand von $\mathbb{D}$ und wird mit $\partial\mathbb{D}$ bezeichnet. Eine Menge heißt abgeschlossen, wenn sie alle ihre Randpunkte enthält.

Beispiel(e) 6.5
Sei $r > 0$:

- $\{x \in \mathbb{R}^n;\ |x| \leq r\}$ ist abgeschlossen,
- $\{x \in \mathbb{R}^n;\ |x| < r\}$ ist offen,
- $\{(x, y) \in \mathbb{R}^2;\ 2 \leq x < 3,\ -1 < y \leq 2\}$ ist weder offen noch abgeschlossen.



(6.19)

Nun sind wir in der Lage, Grenzwert und Stetigkeit für Skalarenfelder zu definieren.

Definition 6.6 (Grenzwert, Stetigkeit)
Sei $f : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} \subseteq \mathbb{R}^n$ und $a \in \mathbb{D} \cup \partial\mathbb{D}$.

(a): f hat in a den Grenzwert $c \in \mathbb{R}$, in Zeichen

$$\lim_{x \to a} f(x) = c \quad (\text{oder } f(x) \to c \quad \text{für } x \to a), \tag{6.20}$$

falls es zu jedem $\varepsilon > 0$ eine r-Kugel

$$K_r(a) := \{x \in \mathbb{R}^n;\ |x - a| < r\}, \quad r > 0, \tag{6.21}$$

gibt, sodass

$$|f(x) - c| < \varepsilon \quad \text{für alle } x \in \mathbb{D} \cap K_r(a). \tag{6.22}$$

(b): f heißt stetig in $a \in \mathbb{D}$, falls $\lim_{x \to a} f(x) = f(a)$ gilt.

(c): f heißt auf $\mathbb{D}$ stetig, falls f in allen $a \in \mathbb{D}$ stetig ist.

Die aus der Analysis einer reellen Veränderlichen bekannten Rechenregeln für Grenzwerte und stetige Funktionen gelten auch hier.

Beispiel(e) 6.7

- Die Projektionen $p_i : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto x_i$ sind stetig.
  Daher sind auch die linearen Funktionen $l : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto a^\top x$ und jedes Polynom $p : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \sum_{1 \le k_i \le m} a_{k_1 k_2 \dots k_n} x_1^{k_1} x_2^{k_2} \cdot \ldots \cdot x_n^{k_n}$ stetig. Jede rationale Funktion $f : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \frac{p(x)}{q(x)}$ mit $q(x) \neq 0$, $x \in \mathbb{R}$, ist stetig.
- Die Funktion
$$f : \mathbb{R}^2 \to \mathbb{R}, \quad (x, y) \mapsto \begin{cases} \frac{2xy^2}{x^2+y^4} & \text{falls } x > 0 \\ 0 & \text{falls } x \le 0 \end{cases} \tag{6.23}$$
ist nicht stetig in $(0,0)^\top$, denn es gilt für $x = y^2$
$$\lim_{y^2 \to 0} f(y^2, y) = 1 \neq f(0,0) = 0. \tag{6.24}$$

Definition 6.8 (beschränkt, kompakt)
Eine Menge $B \subseteq \mathbb{R}^n$ heißt beschränkt, falls es ein $K > 0$ gibt mit $|x| < K$ für alle $x \in B$. Ist eine Menge $B \subseteq \mathbb{R}^n$ abgeschlossen und beschränkt, so wird sie als kompakt bezeichnet.

Die r-Kugeln $\{x \in \mathbb{R}^n;\ |x - a| \le r\}$ mit Rand sind für alle $a \in \mathbb{R}^n$ und alle $r > 0$ kompakt.
Völlig analog zum eindimensionalen Fall gilt der wichtige Satz, dass jede auf einer kompakten Menge $\mathbb{D} \subset \mathbb{R}^n$ betrachtete stetige Funktion $f : \mathbb{D} \to \mathbb{R}$ auf $\mathbb{D}$ ein Minimum und Maximum annimmt, dass es also $a, b \in \mathbb{D}$ gibt mit

$$f(a) \le f(x) \le f(b) \quad \text{für alle } x \in \mathbb{D}. \tag{6.25}$$

Seien nun $\mathbb{D} \subseteq \mathbb{R}^n$ offen, $f : \mathbb{D} \to \mathbb{R}$ eine Funktion und $a \in \mathbb{D}$.
Existiert die Ableitung der „partiellen“ Funktion einer reellen Variablen

$$x \mapsto f(a_1, a_2, \dots, a_{i-1}, x, a_{i+1}, \dots, a_n) \tag{6.26}$$

an der Stelle $x = a_i$, so nennt man diese die partielle Ableitung von f nach x_i im Punkte a. Sie wird mit

$$\left.\frac{\partial f(x)}{\partial x_i}\right|_{x=a} \quad \text{oder} \quad \frac{\partial f}{\partial x_i}(a) \tag{6.27}$$

bezeichnet. Wie im eindimensionalen Fall betrachtet man die Ableitungsfunktionen

$$f_{x_i} : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \frac{\partial f}{\partial x_i}(x) := \lim_{t \to 0} \frac{1}{t}(f(x_1, \dots, x_{i-1}, x_i + t, x_{i+1}, \dots, x_n) - f(x)). \tag{6.28}$$

Die Funktion f heißt partiell differenzierbar (bzw. stetig partiell differenzierbar), wenn alle partiellen Ableitungen f_{x_i}, $i = 1, \dots, n$, existieren (und stetig sind). Fasst man die partiellen Ableitungen einer Funktion an der Stelle x zu einem Spaltenvektor zusammen, so erhält man den Gradienten von f an der Stelle x:

$$\nabla f := \text{grad } f : \mathbb{D} \to \mathbb{R}^n, \quad x \mapsto \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \tag{6.29}$$

Beispiel(e) 6.9

- $f : \mathbb{R}^3 \to \mathbb{R}$, $(x, y, z) \mapsto e^{x+2y} + 2x\sin(z) + z^2xy$

$$\text{grad } f : \mathbb{R}^3 \to \mathbb{R}^3, \quad (x, y, z) \mapsto \begin{pmatrix} e^{x+2y} + 2\sin(z) + z^2y \\ 2e^{x+2y} + z^2x \\ 2x\cos(z) + 2zxy \end{pmatrix}$$

- $f : (0, \infty) \times \mathbb{R} \to \mathbb{R}$, $(x, y) \mapsto x^2y^3 + y\ln(x)$

$$\text{grad } f : (0, \infty) \times \mathbb{R} \to \mathbb{R}^2, \quad (x, y) \mapsto \begin{pmatrix} 2xy^3 + \frac{y}{x} \\ 3x^2y^2 + \ln(x) \end{pmatrix}$$

$$f_{xx} := (f_x)_x : (0, \infty) \times \mathbb{R} \to \mathbb{R}, \quad (x, y) \mapsto 2y^3 - \frac{y}{x^2}$$

$$f_{xy} := (f_x)_y : (0, \infty) \times \mathbb{R} \to \mathbb{R}, \quad (x, y) \mapsto 6xy^2 + \frac{1}{x} = f_{yx}(x, y)$$

$$f_{yy} := (f_y)_y : (0, \infty) \times \mathbb{R} \to \mathbb{R}, \quad (x, y) \mapsto 6x^2y.$$

Eine Funktion $f : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} \subseteq \mathbb{R}^n$ offen, heißt k-mal partiell differenzierbar, wenn alle k-ten Ableitungen $f_{x_{i_1},\ldots,x_{i_k}}$, $i_1, \ldots, i_k \in \{1, \ldots, n\}$, existieren.
Sind alle diese Ableitungen stetig, so heißt f k-mal stetig partiell differenzierbar ($k \in \mathbb{N}$). Wir bezeichnen alle auf einer offenen Menge $\mathbb{D} \subseteq \mathbb{R}^n$ k-mal stetig partiell differenzierbaren Funktionen $f : \mathbb{D} \to \mathbb{R}$ mit $\mathcal{C}^k(\mathbb{D}, \mathbb{R})$ und nennen $f \in \mathcal{C}^k(\mathbb{D}, \mathbb{R})$ eine $\mathcal{C}^k$-Funktion.
Mit $\mathcal{C}^0(\mathbb{D}, \mathbb{R})$ bezeichnen wir alle stetigen Funktionen $f : \mathbb{D} \to \mathbb{R}$.
Im folgenden Satz betrachten wir eine hinreichende Bedingung für die Vertauschbarkeit partieller Ableitungen.

Satz 6.10 (Vertauschbarkeit partieller Ableitungen)
Sei $f : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} \subseteq \mathbb{R}^n$ offen und $f \in \mathcal{C}^2(\mathbb{D}, \mathbb{R})$, so gilt:

$$f_{x_ix_j} = f_{x_jx_i} \quad \text{für } 1 \leq i, j \leq n. \tag{6.30}$$

Wie im eindimensionalen Fall versucht man auch bei Skalarenfeldern, Funktionen durch einfachere „Ersatzfunktionen“ zu approximieren. Gilt für zwei Funktionen $f, g : \mathbb{D} \to \mathbb{R}$ mit $\mathbb{D} \subseteq \mathbb{R}^n$:

$$\lim_{x \to x_0} \frac{f(x) - g(x)}{|x - x_0|^k} = 0 \tag{6.31}$$

für ein $x_0 \in \mathbb{D}$ und ein $k \in \mathbb{N}$, so schreibt man dafür

$$f(x) = g(x) + o(|x - x_0|^k). \tag{6.32}$$

Diese Gleichung besagt, dass bei der Approximation von f durch g der entstehende Fehler $R(x, x_0) := f(x) - g(x)$ in der Nähe des Punktes x_0 klein ist im Vergleich zu $|x - x_0|^k$ in dem Sinne, dass

$$\lim_{x \to x_0} \frac{R(x, x_0)}{|x - x_0|^k} = 0. \tag{6.33}$$

Die Approximation einer differenzierbaren Funktion einer Variablen $f : I \to \mathbb{R}$ ($I \subseteq \mathbb{R}$ offenes Intervall) in der Umgebung von $x_0 \in I$ durch die lineare Funktion

$$g : I \to \mathbb{R}, \quad x \mapsto f(x_0) + f'(x_0)(x - x_0) \tag{6.34}$$

wird mit dem o-Symbol geschrieben als:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(|x - x_0|). \tag{6.35}$$

Für Funktionen von mehreren Variablen gewährleistet die partielle Differenzierbarkeit allein noch nicht die analoge Vorgehensweise. Dazu benötigt man die folgende Eigenschaft:

Definition 6.11 (Total differenzierbar)
Sei $\mathbb{D} \subseteq \mathbb{R}^n$ offen. Eine Funktion $f : \mathbb{D} \to \mathbb{R}$ heißt in $x_0 \in \mathbb{D}$ total differenzierbar, wenn es einen Spaltenvektor $a \in \mathbb{R}^n$ mit

$$f(x) = f(x_0) + a^\top (x - x_0) + o(|x - x_0|) \tag{6.36}$$

gibt.

Ist nun f in x_0 total differenzierbar mit

$$f(x) = f(x_0) + a^\top (x - x_0) + o(|x - x_0|), \tag{6.37}$$

so ist f in x_0 stetig, partiell differenzierbar und es gilt $a = \text{grad } f(x_0)$. Somit ist „total differenzierbar in x_0“ eine stärkere Bedingung als „partiell differenzierbar in x_0“.
Allerdings ist jede $\mathcal{C}^1(\mathbb{D}, \mathbb{R})$-Funktion $f : \mathbb{D} \to \mathbb{R}$ mit $\mathbb{D} \subseteq \mathbb{R}^n$ offen, in allen Punkten $x_0 \in \mathbb{D}$ total differenzierbar.

Beispiel(e) 6.12
$f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^4 + 2x^3y^2 + y$, $x_0 = (1, 1)$

$$\begin{aligned} f(x,y) &= f(1,1) + (\text{grad } f(1,1))^\top \begin{pmatrix} x-1 \\ y-1 \end{pmatrix} + o(\sqrt{(x-1)^2 + (y-1)^2}) \\ &= 4 + 10(x-1) + 5(y-1) + o(\sqrt{(x-1)^2 + (y-1)^2}), \end{aligned} \tag{6.38}$$

da grad $f(1,1) = \begin{pmatrix} 10 \\ 5 \end{pmatrix}$.

Anwendung: Das Newton-Verfahren
Gegeben sind n $\mathcal{C}^1$-Funktionen $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, n$.
Gesucht ist ein $x \in \mathbb{R}^n$ mit

$$\begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} = 0 \in \mathbb{R}^n. \tag{6.39}$$

Sei $x_0 \in \mathbb{R}^n$ ein Startpunkt, so betrachtet man anstelle von

$$\begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} = 0 \tag{6.40}$$

das linearisierte Problem

$$\begin{pmatrix} f_1(x_0) + (\text{grad } f_1(x_0))^\top (x - x_0) \\ \vdots \\ f_n(x_0) + (\text{grad } f_n(x_0))^\top (x - x_0) \end{pmatrix} = 0 \tag{6.41}$$

Dies ist ein lineares Gleichungssystem $Ax = b$ mit

$$A = \begin{pmatrix} (\text{grad } f_1(x_0))^\top \\ \vdots \\ (\text{grad } f_n(x_0))^\top \end{pmatrix} \in \mathbb{R}^{n,n} \quad \text{und} \quad b = Ax_0 - \begin{pmatrix} f_1(x_0) \\ \vdots \\ f_n(x_0) \end{pmatrix} \tag{6.42}$$

Unter der Annahme, dass dieses lineare Gleichungssystem genau eine Lösung x_1 besitzt, kann nun x_1 als Startpunkt verwendet und die Vorgehensweise wiederholt werden. Insgesamt soll eine Folge $\{x_n\}$, $n \in \mathbb{N}_0$, erzeugt werden, sodass

$$\lim_{n \to \infty} |x_n - \bar{x}| = 0 \tag{6.43}$$

und

$$\begin{pmatrix} f_1(\bar{x}) \\ \vdots \\ f_n(\bar{x}) \end{pmatrix} = 0. \tag{6.44}$$

Sei $n = 1$: $\quad f_1 : \mathbb{R} \to \mathbb{R}$, $x \mapsto \cos(x)$, $x_0 = 1.0$

$$x_1 = 1.0 + \frac{\cos(1.0)}{\sin(1.0)} \approx 1.6420926 \text{ (Lösung von } Ax = b) \tag{6.45}$$

$$x_2 = 1.6420926 + \frac{\cos(1.6420926)}{\sin(1.6420926)} \approx 1.5706753 \tag{6.46}$$

$$x_2 = 1.5706753 + \frac{\cos(1.5706753)}{\sin(1.5706753)} \approx 1.5707963 \tag{6.47}$$

zum Vergleich:

$$\frac{\pi}{2} \approx 1.5707963267\ldots \tag{6.48}$$

$$\cos\left(\frac{\pi}{2}\right) = 0. \tag{6.49}$$

Die partiellen Ableitungen f_{x_i} einer Funktion $f : \mathbb{D} \to \mathbb{R}$ geben lokale Änderungen der Funktionswerte in Richtung der Koordinatenachsen e_i an.

Allgemein nennt man zu jedem $v \in \mathbb{R}^n$, $v \neq 0$, den Grenzwert

$$\partial_v f(x) := \lim_{t \to 0} \frac{1}{t}(f(x + t \cdot v) - f(x)) \tag{6.50}$$

(falls existent) die Ableitung von f an der Stelle x längs v. Ist v ein Einheitsvektor (also $|v| = 1$), dann heißt $\partial_v f(x)$ die Richtungsableitung von f an der Stelle x in Richtung v. Die entsprechende Funktion

$$\partial_v f : \mathbb{D} \to \mathbb{R} \quad x \mapsto \lim_{t \to 0} \frac{1}{t}(f(x + t \cdot v) - f(x)) \tag{6.51}$$

heißt (Richtungs)ableitung von f in Richtung v.
Ist $\mathbb{D} \subseteq \mathbb{R}^n$ offen und $f : \mathbb{D} \to \mathbb{R}$ auf $\mathbb{D}$ total differenzierbar, so gilt für jeden Vektor $v \in \mathbb{R}^n$, $v \neq 0$ und für $x \in \mathbb{D}$:

$$\lim_{t \to 0} \frac{1}{t}(f(x + t \cdot v) - f(x)) = (\text{grad } f(x))^\top v. \tag{6.52}$$

Höhere Ableitungen längs v werden mit $\partial_v^k f$ bezeichnet und sind durch die höheren Ableitungen der Funktion

$$g : I \subseteq \mathbb{R} \to \mathbb{R}, \quad t \mapsto f(x + tv) \tag{6.53}$$

an der Stelle $t = 0$ definiert.

Beispiel(e) 6.13

- $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^2 + y^2$. Sei $v = (\cos(\alpha), \sin(\alpha))^\top$, $\alpha \in [0, 2\pi]$ fest gewählt, so ist $|v| = 1$. Für $(x, y) = (1, 1)$ gilt:

$$\begin{aligned}
\partial_v f(1,1) &= \lim_{t \to 0} \frac{1}{t}\left(f(1 + t\cos(\alpha), 1 + t \cdot \sin(\alpha)) - f(1,1) = \right. \\
&= \lim_{t \to 0} \frac{(1 + t \cdot \cos(\alpha))^2 + (1 + t \cdot \sin(\alpha))^2 - 2}{t} = \\
&= \lim_{t \to 0} \frac{t^2\cos^2(\alpha) + 2t\cos(\alpha) + t^2\sin^2(\alpha) + 2t\sin(\alpha)}{t} = \\
&= 2\cos(\alpha) + 2\sin(\alpha) = \\
&= (\text{grad } f(1,1))^\top v.
\end{aligned}$$

Verbindet man Skalarenfelder mit Kurvenstücken, so kann man unter gewissen Voraussetzungen ein Kettenregel analog zum eindimensionalen Fall herleiten.
Sei also $f : \mathbb{D} \to \mathbb{R}$, $x \mapsto f(x)$, differenzierbar, wobei $\mathbb{D} \subseteq \mathbb{R}^n$ offen ist, und sei $\xi : [a, b] \to \mathbb{D}$, $t \mapsto \xi(t)$ ein Kurvenstück, so gilt für die stetig differenzierbare Funktion

$$g : [a, b] \to \mathbb{R}, \quad t \mapsto f(\xi(t)) : \tag{6.54}$$

$$\dot{g} : [a, b] \to \mathbb{R}, \quad t \mapsto (\text{grad } f(\xi(t)))^\top \dot{\xi}(t). \tag{6.55}$$

Die Kettenregel kommt insbesondere dann zur Anwendung, wenn man im $\mathbb{R}^2$ oder $\mathbb{R}^3$ zu Polarkoordinaten übergeht:

- im $\mathbb{R}^2$:
  Seien

$$\begin{aligned}
f : \mathbb{R}^2 \to \mathbb{R}, \quad (x, y) &\mapsto f(x, y) && (6.56)\\
\xi : \mathbb{R}_0^+ \times [0, 2\pi] \to \mathbb{R}, \quad (r, \varphi) &\mapsto r \cdot \cos(\varphi) && (6.57)\\
\eta : \mathbb{R}_0^+ \times [0, 2\pi] \to \mathbb{R}, \quad (r, \varphi) &\mapsto r \cdot \sin(\varphi) && (6.58)
\end{aligned}$$

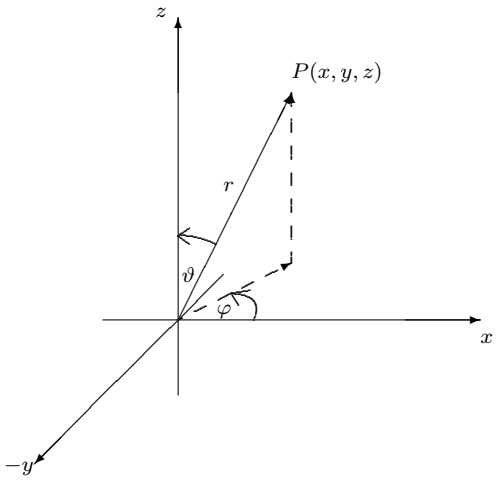
  differenzierbare Funktionen. Mit $F : \mathbb{R}_0^+ \times [0, 2\pi] \to \mathbb{R}$, $(r, \varphi) \mapsto f(\xi(r, \varphi), \eta(r, \varphi))$ folgt:

$$\begin{aligned}
F_r : \mathbb{R}_0^+ \times [0, 2\pi] \to \mathbb{R}, \quad (r, \varphi) &\mapsto f_x(\xi(r, \varphi), \eta(r, \varphi)) \cdot \cos(\varphi) + f_y(\xi(r, \varphi), \eta(r, \varphi)) \cdot \sin(\varphi)\\
F_\varphi : \mathbb{R}_0^+ \times [0, 2\pi] \to \mathbb{R}, \quad (r, \varphi) &\mapsto f_x(\xi(r, \varphi), \eta(r, \varphi))(-r \cdot \sin(\varphi)) + f_y(\xi(r, \varphi), \eta(r, \varphi)) \cdot r \cdot \cos(\varphi).
\end{aligned}$$

- im $\mathbb{R}^3$:
  Seien

$$\begin{aligned} f : \mathbb{R}^3 \to \mathbb{R}, \quad (x,y,z) &\mapsto f(x,y,z) && (6.59)\\ \xi : \mathbb{R}_0^+ \times [0,2\pi] \times [0,\pi] \to \mathbb{R}, \quad (r,\varphi,\vartheta) &\mapsto r \cdot \cos(\varphi) \cdot \sin(\vartheta) && (6.60)\\ \eta : \mathbb{R}_0^+ \times [0,2\pi] \times [0,\pi] \to \mathbb{R}, \quad (r,\varphi,\vartheta) &\mapsto r \cdot \sin(\varphi) \cdot \sin(\vartheta) && (6.61)\\ \zeta : \mathbb{R}_0^+ \times [0,2\pi] \times [0,\pi] \to \mathbb{R}, \quad (r,\varphi,\vartheta) &\mapsto r \cdot \cos(\vartheta) && (6.62) \end{aligned}$$

  differenzierbare Funktionen.



(6.63)

  Mit

$$F : \mathbb{R}_0^+ \times [0,2\pi] \times [0,\pi] \to \mathbb{R}, \quad (r,\varphi,\vartheta) \mapsto f(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \qquad (6.64)$$

  folgt:

$$\begin{aligned} F_r : \mathbb{R}_0^+ \times [0,2\pi] \times [0,\pi] \to \mathbb{R}, \quad (r,\varphi,\vartheta) &\mapsto f_x(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot \cos(\varphi) \cdot \sin(\vartheta) + \\ &+ f_y(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot \sin(\varphi) \cdot \sin(\vartheta) + \\ &+ f_z(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot \cos(\vartheta) \\ F_\varphi : \mathbb{R}_0^+ \times [0,2\pi] \times [0,\pi] \to \mathbb{R}, \quad (r,\varphi,\vartheta) &\mapsto f_x(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot (-r \cdot \sin(\varphi) \cdot \sin(\vartheta)) + \\ &+ f_y(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot r \cdot \cos(\varphi) \cdot \sin(\vartheta) \\ F_\vartheta : \mathbb{R}_0^+ \times [0,2\pi] \times [0,\pi] \to \mathbb{R}, \quad (r,\varphi,\vartheta) &\mapsto f_x(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot r \cdot \cos(\varphi) \cdot \cos(\vartheta) + \\ &+ f_y(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot r \cdot \sin(\varphi) \cdot \cos(\vartheta) + \\ &+ f_z(\xi(r,\varphi,\vartheta), \eta(r,\varphi,\vartheta), \zeta(r,\varphi,\vartheta)) \cdot (-r \cdot \sin(\vartheta)). \end{aligned}$$

Für eine differenzierbare Funktion $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ ist der Anstieg im Punkt $\widetilde{x} \in \mathbb{D}$ in Richtung eines Vektors $v \in \mathbb{R}^n$, $|v| = 1$ gegeben durch

$$\partial_v f(\widetilde{x}) = (\text{grad } f(\widetilde{x}))^\top v = |\text{grad } f(\widetilde{x})| \cdot \cos(\alpha), \qquad (6.65)$$

wobei α den Winkel zwischen grad $f(\widetilde{x})$ und v bezeichnet. Dieser Anstieg ist maximal, wenn $\cos(\alpha) = 1$, also $\alpha = 0$ ist.
Somit gibt grad $f(\widetilde{x})$ die Richtung des steilsten Anstieges von f in $\widetilde{x}$ an. Da f genau dann ansteigt, wenn $-f$ abnimmt, bezeichnet $-$grad $f(\widetilde{x})$ die Richtung des stärksten Abnehmens der Funktion f in $\widetilde{x}$. Insbesondere erhält man analog zum eindimensionalen Fall Kandidaten für Extremalstellen von f durch Lösung des (im allgemeinen nichtlinearen Gleichungssystems)

$$\text{grad } f(x) = 0 \qquad (6.66)$$

Die Interpretation der Gradienteninformation wird in der nichtlinearen Optimierung dazu verwendet, das folgende Verfahren zur numerischen Berechnung von Kandidaten für Minimalstellen herzuleiten:
Ausgehend von einem Startpunkt x_0 bestimmt man einen Punkt

$$x_1 = x_0 - \sigma \cdot \text{grad } f(x_0) \tag{6.67}$$

mit einer dem Problem angepassten Schrittweite $\sigma \geq 0$. Ist $f(x_1) > f(x_0)$, so war σ zu groß gewählt und man versucht

$$x_1 = x_0 - \frac{\sigma}{2} \cdot \text{grad } f(x_0) \tag{6.68}$$

Ist $f(x_1) < f(x_0)$, so nimmt man x_1 als neuen Startwert und wiederholt das Verfahren.
Aus der Tatsache, dass $-$grad $f(x_0)$ die Richtung des steilsten Abstieges angibt, folgt aus

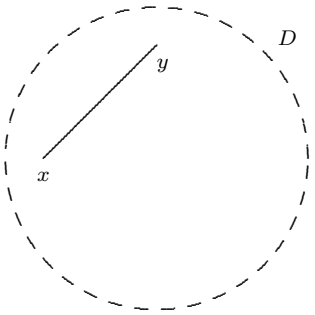
$$-\text{grad } f(x_0) \neq 0, \tag{6.69}$$

dass es ein $\overline{\sigma} > 0$ gibt mit $f(x_1) < f(x_0)$ für alle $\sigma \in (0, \overline{\sigma})$.
Im Folgenden verallgemeinern wir die Taylor-Formel für Funktionen einer reellen Veränderlichen auf $\mathcal{C}^{k+1}(\mathbb{D}, \mathbb{R})$-Funktionen $f : \mathbb{D} \to \mathbb{R}$, wobei wir für $\mathbb{D} \subseteq \mathbb{R}^n$ voraussetzen:

(a): $\mathbb{D}$ ist offen

(b): Aus $x, y \in \mathbb{D}$ folgt:

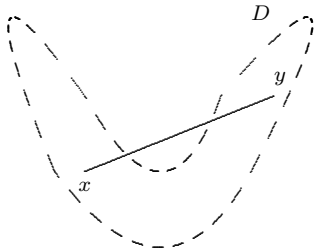
$$x + t(y - x) \in \mathbb{D} \quad \text{für alle } t \in [0, 1] \tag{6.70}$$



(6.71)

$x + t(y - x)$ ist die Strecke zwischen x und y.

Eine Menge mit diesen beiden Eigenschaften bezeichnet man als konvexes Gebiet.
Gegenbeispiel:



(6.72)

Bei einem konvexen Gebiet $\mathbb{D}$ ist mit zwei Punkten $x, v \in \mathbb{D}$ auch die Verbindungsstrecke $x + t(v - x)$, $t \in [0, 1]$, in $\mathbb{D}$. Sei nun $h : [0, 1] \to \mathbb{R}$, $t \mapsto f(x + t \cdot v)$ mit $x, v \in \mathbb{D}$. Die Taylor-Formel für h mit Entwicklungspunkt $t = 0$ lautet:

$$h : [0,1] \to \mathbb{R}, \quad t \mapsto h(0) + \dot{h}(0)t + \frac{1}{2}\ddot{h}(0)t^2 + \ldots + \frac{1}{k!}h^{(k)}(0)t^k + \frac{1}{(k+1)!}h^{(k+1)}(\xi)t^{k+1} \tag{6.73}$$

Für $t = 1$ erhalten wir:

$$h(1) = h(0) + \dot{h}(0) + \ldots + \frac{1}{k!}h^k(0) + \frac{1}{(k+1)!}h^{(k+1)}(\xi), \quad \text{wobei } 0 \leq \xi \leq 1. \tag{6.74}$$

Da nun $h(0) = f(x)$, $\dot{h}(0) = \partial_v f(x)$, $\ddot{h}(0) = \partial_v^2 f(x)$, ..., $h^{(k)}(0) = \partial_v^k f(x)$, $h^{(k+1)}(\xi) = \partial_v^{k+1}(x + \xi v)$ und $h(1) = f(x + v)$, gilt für $x \in \mathbb{D}$:

$$f(x+v) = f(x) + \partial_v f(x) + \frac{1}{2}\partial_v^2 f(x) + \ldots + \frac{1}{k!}\partial_v^k f(x) + \frac{1}{(k+1)!}\partial_v^{k+1} f(x + \xi v) \tag{6.75}$$

mit $0 \leq \xi \leq 1$.
Die Taylor-Formel dient zur Approximation von f in der Umgebung eines festen Punktes $x_0 \in \mathbb{D}$ durch das Taylor-Polynom (ersetze x durch x_0, dann v durch $x - x_0$):

$$\begin{aligned} f : \mathbb{D} \to \mathbb{R}, \quad x \quad \mapsto \quad & \underbrace{f(x_0) + \partial_{(x-x_0)} f(x_0) + \frac{1}{2}\partial_{(x-x_0)}^2 f(x_0) + \ldots + \frac{1}{k!}\partial_{(x-x_0)}^k f(x_0)}_{\text{Taylor-Polynom } p(x)} + \\ + \quad & \underbrace{\frac{1}{(k+1)!}\partial_{(x-x_0)}^{k+1} f(x_0 + \xi(x - x_0))}_{\text{Restglied}}. \end{aligned} \tag{6.76}$$

Es gilt:

$$f(x) = p(x) + o(|x - x_0|^k) \tag{6.77}$$

Für eine $\mathcal{C}^m(\mathbb{D}, \mathbb{R})$-Funktion $(m \geq 2)$ heißt die symmetrische Matrix

$$H_f(\overline{x}) := \begin{pmatrix} f_{x_1x_1}(\overline{x}) & \ldots & f_{x_1x_n}(\overline{x}) \\ \vdots & & \vdots \\ f_{x_nx_1}(\overline{x}) & \ldots & f_{x_nx_n}(\overline{x}) \end{pmatrix} \tag{6.78}$$

die Hesse-Matrix von f im Punkt $\overline{x}$.
Die eben hergeleitete Taylor-Formel lautet somit für die Spezialfälle

$k = 0$:

$$f : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \underbrace{f(x_0)}_{p(x)} + \underbrace{(\text{grad } f(x_0 + \xi(x - x_0)))^\top (x - x_0)}_{o(|x-x_0|^0)} \tag{6.79}$$

$k = 1$:

$$f : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \underbrace{f(x_0) + (\text{grad } f(x_0))^\top (x - x_0)}_{p(x)} + \underbrace{\frac{1}{2}(x - x_0)^\top H_f(x + \xi(x - x_0))(x - x_0)}_{o(|x-x_0|)} \tag{6.80}$$

$k = 2$:

$$f : \mathbb{D} \to \mathbb{R}, \quad x \mapsto \underbrace{f(x_0) + (\text{grad } f(x_0))^\top (x - x_0) + \frac{1}{2}(x - x_0)^\top H_f(x_0)(x - x_0)}_{p(x)} + o(|x - x_0|^2) \tag{6.81}$$

Beispiel(e) 6.14
Der Affensattel $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^3 - 3xy^2$ hat im Punkt $(0, 0)$:

$$\text{grad } f(0,0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad H_f(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{6.82}$$

Wie im eindimensionalen Fall interessiert man sich auch bei Skalarenfeldern für lokale und globale Minimierer. Sei $\mathbb{D} \subseteq \mathbb{R}^n$ und $f : \mathbb{D} \to \mathbb{R}$, so heißt $\overline{x} \in \mathbb{D}$ lokale Minimalstelle (oder lokaler Minimierer), falls es eine r-Kugel $K_r := \{x \in \mathbb{R}^n; |x - \overline{x}| < r\}$ um $\overline{x}$ mit $r > 0$ gibt und

$$f(x) \geq f(\overline{x}) \quad \text{für alle } x \in K_r \cap \mathbb{D}. \tag{6.83}$$

Gilt die obige Ungleichung für alle $x \in \mathbb{D}$, so heißt $\overline{x} \in \mathbb{D}$ globale Minimalstelle (oder globaler Minimierer) (analoge Definitionen für Maximierer). Ist $f \in \mathcal{C}^1(\mathbb{D}, \mathbb{R})$, $\mathbb{D}$ offen und $\overline{x} \in \mathbb{D}$ eine Minimalstelle, so wissen wir bereits, dass dann grad $f(\overline{x}) = 0$ gilt.
Ist ein innerer Punkt a von $\mathbb{D}$ keine Minimalstelle, gilt aber dennoch grad $f(a) = 0$, so heißt dieser Punkt Sattelpunkt.
Sei nun $f : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} \subseteq \mathbb{R}^n$ offen, $f \in \mathcal{C}^2(\mathbb{D}, \mathbb{R})$ und grad $f(\overline{x}) = 0$, so gilt:

(a): Alle Eigenwerte von $H_f(\overline{x})$ sind größer als $0 \Longrightarrow \overline{x}$ ist Minimalstelle

(b): Alle Eigenwerte von $H_f(\overline{x})$ sind kleiner als $0 \Longrightarrow \overline{x}$ ist Maximalstelle (6.84)

(c): Es gibt positive und negative Eigenwerte von $H_f(\overline{x}) \Longrightarrow \overline{x}$ ist Sattelpunkt.

Symmetrische Matrizen mit allen Eigenwerten größer Null heißen positiv definit.
Symmetrische Matrizen mit allen Eigenwerten kleiner Null heißen negativ definit.
Sind alle Eigenwerte einer symmetrischen Matrix größer oder gleich (kleiner oder gleich) Null, so spricht man von positiv (negativ) semidefiniten Matrizen. Eine symmetrische Matrix heißt indefinit, falls mindestens ein Eigenwert größer Null und mindestens ein Eigenwert kleiner Null ist.

Beispiel(e) 6.15

- $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^2 + y^2$

$$\text{grad } f(x,y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix} \qquad H_f(x,y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}. \tag{6.85}$$

  Die Stelle $(0, 0)$ ist globaler Minimierer.

- $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^2 + y^4$

$$\text{grad } f(x,y) = \begin{pmatrix} 2x \\ 4y^3 \end{pmatrix} \qquad H_f(x,y) = \begin{pmatrix} 2 & 0 \\ 0 & 12y^2 \end{pmatrix}. \tag{6.86}$$

  Die Stelle $(0, 0)$ ist globaler Minimierer mit positiver semidefiniter Hessematrix $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$.

- $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^2 - y^2$

$$\text{grad } f(x,y) = \begin{pmatrix} 2x \\ -2y \end{pmatrix} \qquad H_f(x,y) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}. \tag{6.87}$$

  Die Stelle $(0, 0)$ ist ein Sattelpunkt, da $H_f(0, 0)$ indefinit ist.

- $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^2 - y^4$

$$\text{grad } f(x,y) = \begin{pmatrix} 2x \\ -4y^3 \end{pmatrix} \qquad H_f(x,y) = \begin{pmatrix} 2 & 0 \\ 0 & -12y^2 \end{pmatrix}. \tag{6.88}$$

  Die Stelle $(0, 0)$ ist Sattelpunkt, obwohl $H_f(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ positiv semidefinit ist.

- $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto (x - y)^2$

$$\text{grad } f(x,y) = \begin{pmatrix} 2(x-y) \\ -2(x-y) \end{pmatrix} \qquad H_f(x,y) = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}. \tag{6.89}$$

  Die Stelle $(0, 0)$ ist ein nichtisolierter globaler Minimierer, da alle Punkte (x, y) mit $x = y$ globale Minimalstellen sind.
  Für diese Minimalstellen ist H_f positiv semidefinit.

- $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto \frac{1}{2}(x - y^2)(x - 2y^2)$,

$$\text{grad } f(x,y) = \begin{pmatrix} \frac{1}{2}(x-2y^2) + \frac{1}{2}(x-y^2) \\ -y(x-2y^2) - 2y(x-y^2) \end{pmatrix}, \tag{6.90}$$

$$H_f(x,y) = \begin{pmatrix} 1 & -3y \\ -3y & -3x + 12y^2 \end{pmatrix}, \tag{6.91}$$

$$\text{grad } f(0,0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad H_f(0,0) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ pos. semidefinit.} \tag{6.92}$$

  Die Funktion f steigt von $(0, 0)$ aus entlang **jeder** Richtung zunächst an, dennoch ist $(0, 0)$ ein Sattelpunkt, denn entlang der Kurve $k : \mathbb{R} \to \mathbb{R}^2$, $t \mapsto (\frac{3}{2}t^2, t)$ ist f streng monoton fallend, wobei $k(0) = (0, 0)$.

Für zwei Variable kann somit der folgende Extremstellen-Test durchgeführt werden:
Sei $f : \mathbb{D} \subseteq \mathbb{R}^2 \to \mathbb{R}$, $\mathbb{D}$ offen, $f \in \mathcal{C}^2(\mathbb{D}, \mathbb{R})$ und $\overline{x} \in \mathbb{D}$ mit grad $f(\overline{x}) = 0$, so gilt:

(a): Ist $f_{xx}(\overline{x}) > 0$ und $\det(H_f(\overline{x})) > 0$, so ist $H_f(\overline{x})$ positiv definit und $\overline{x}$ eine (lokale) Minimalstelle.

(b): Ist $f_{xx}(\overline{x}) < 0$ und $\det(H_f(\overline{x})) > 0$, so ist $H_f(\overline{x})$ negativ definit und $\overline{x}$ eine (lokale) Maximalstelle. (6.93)

(c): Ist $\det(H_f(\overline{x})) < 0$, so ist $H_f(\overline{x})$ indefinit und $\overline{x}$ ein Sattelpunkt.

Nach Kurven und Skalarenfeldern betrachten wir nun Vektorfelder $f : \mathbb{D} \to \mathbb{R}^m$, $\mathbb{D} \subseteq \mathbb{R}^n$. Dabei übertragen wir alle wichtigen Begriffe für Skalarenfelder auf die Komponentenfunktionen:

$$f_1, \ldots, f_m : \mathbb{D} \to \mathbb{R}, \quad x \mapsto f_i(x), \quad \text{wobei } f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix} \text{ gilt.} \tag{6.94}$$

Definition 6.16 (Grenzwert und Anwendungen)
Für $f : \mathbb{D} \to \mathbb{R}^m$, $\mathbb{D} \subseteq \mathbb{R}^n$ und $x, x_0 \in \mathbb{D}$ gilt:

(a):

$$\lim_{x \to x_0} f(x) := \begin{pmatrix} \lim\limits_{x \to x_0} f_1(x) \\ \vdots \\ \lim\limits_{x \to x_0} f_m(x) \end{pmatrix}; \quad \frac{\partial f}{\partial x_i}(x) := \begin{pmatrix} \frac{\partial f_1}{\partial x_i}(x) \\ \vdots \\ \frac{\partial f_m}{\partial x_i}(x) \end{pmatrix} \tag{6.95}$$

$f(x) = o(|x - x_0|) \iff f_k(x) = o(|x - x_0|), \quad k = 1, \ldots, m.$

(b): f ist genau dann stetig, partiell differenzierbar bzw. eine $\mathcal{C}^p(\mathbb{D}, \mathbb{R}^m)$-Funktion, falls alle Komponentenfunktionen f_k von f, $k = 1, \ldots, m$, stetig, partiell differenzierbar bzw. $\mathcal{C}^p(\mathbb{D}, \mathbb{R})$-Funktionen sind.

(c): f heißt in $x_0 \in \mathbb{D}$ total differenzierbar oder auch linear approximierbar, falls es eine $m \times n$ Matrix A und eine Kugel $K_r(x_0) := \{x \in \mathbb{R}^n; |x - x_0| < r\}$, $r > 0$, in $\mathbb{D}$ gibt, sodass für alle $x \in K_r(x_0)$ gilt:

$$f(x) = f(x_0) + A(x - x_0) + o(|x - x_0|) \tag{6.96}$$

Da aufgrund der obigen Definition für eine total differenzierbare Funktion $f : \mathbb{D} \to \mathbb{R}^m$, $\mathbb{D} \subseteq \mathbb{R}^n$, folgt, dass alle Komponentenfunktionen $f_i : \mathbb{D} \to \mathbb{R}$, $i = 1, \ldots, m$, total differenzierbar sind mit

$$f_i(x) = f_i(x_0) + (\text{grad } f_i(x_0))^\top (x - x_0) + o(|x - x_0|), \tag{6.97}$$

definiert man die Jacobi-Matrix (oder Funktionalmatrix) von f in x_0 als

$$J_f(x_0) = \begin{pmatrix} (\text{grad } f_1(x_0))^\top \\ \vdots \\ (\text{grad } f_m(x_0))^\top \end{pmatrix} \tag{6.98}$$

und erhält:

$$f(x) = f(x_0) + J_f(x_0)(x - x_0) + o(|x - x_0|). \tag{6.99}$$

Beispiel(e) 6.17

- Jede lineare Abbildung $f : \mathbb{R}^n \to \mathbb{R}^m$ mit $x \mapsto Ax$, $A \in \mathbb{R}^{m\times n}$, ist total differenzierbar, wobei die Jacobi-Matrix gerade die Abbildungsmatrix ist:

$$J_f(x) = A \text{ für alle } x \in \mathbb{R}^n. \tag{6.100}$$

- Polarkoordinaten: Der Streifen $\mathbb{D} : \{(r,\varphi) \in \mathbb{R}^2;\ 0 \le r \text{ und } 0 \le \varphi < 2\pi\}$ wird durch $f : \mathbb{D} \to \mathbb{R}^2$, $(r,\varphi) \mapsto \begin{pmatrix} r\cdot\cos(\varphi) \\ r\cdot\sin(\varphi)\end{pmatrix}$ so auf den $\mathbb{R}^2$ abgebildet, dass zu jedem $(x,y) \neq 0$ genau ein $(r,\varphi) \in \mathbb{D}$ gehört.
  Die Jacobi-Matrix dieser Abbildung lautet:

$$J_f(r,\varphi) = \begin{pmatrix} \cos(\varphi) & -r\cdot\sin(\varphi) \\ \sin(\varphi) & r\cdot\cos(\varphi)\end{pmatrix}. \tag{6.101}$$

Sei nun $f : \mathbb{R}^n \to \mathbb{R}$, $v, w : \mathbb{R}^n \to \mathbb{R}^m$, $\alpha, \beta \in \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$, $v, w \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^m)$, so gilt:

$$J_{\alpha v+\beta w}(x) = \alpha J_v(x) + \beta J_w(x) \tag{6.102}$$

$$J_{f\cdot v}(x) = f(x)\cdot J_v(x) + v(x)\cdot(\text{grad } f(x))^\top \tag{6.103}$$

$$J_{v^\top w}(x) = v(x)^\top J_w(x) + w(x)^\top J_v(x). \tag{6.104}$$

Für $m = 3$:

$$J_{v\times w}(x) = v(x)\times J_w(x) - w(x)\times J_v(x), \tag{6.105}$$

wobei für einen Vektor $v \in \mathbb{R}^3$ und eine Matrix $M \in \mathbb{R}^{3\times 3}$, $M = (m_1, m_2, m_3)$, $m_i \in \mathbb{R}^3$, gilt:

$$v \times M := \begin{pmatrix} v\times m_1, & v\times m_2, & v\times m_3\end{pmatrix}. \tag{6.106}$$

Anwendung: Basiswechsel: Sei $(b_1,\ldots,b_n)$ eine Basis des $\mathbb{R}^n$ und $(w_1,\ldots,w_m)$ eine Basis des $\mathbb{R}^m$. Sei ferner $f : \mathbb{R}^n \to \mathbb{R}^m$, so wird für f zunächst immer die Basis $(e_1,\ldots,e_n) \in \mathbb{R}^n$ bzw. $(e_1,\ldots,e_m)$ für den $\mathbb{R}^m$ zugrundelegt. Es stellt sich die Frage, wie sich die Darstellung von f bei einem Basiswechsel von $(e_1,\ldots,e_n)$ zu $(b_1,\ldots,b_n)$ und von $(e_1,\ldots,e_m)$ zu $(w_1,\ldots,w_m)$ ändert.
Es gilt mit $W = (w_1,\ldots,w_m) \in \mathbb{R}^{m\times m}$, $B = (b_1,\ldots,b_n) \in \mathbb{R}^{n\times n}$ und $\xi := B^{-1}x$:

$$f(x) = f(B\xi) = Wg(\xi), \tag{6.107}$$

wobei $g(\xi)$ den Vektor $f(x)$ in der Basis W darstellt.
Gesucht ist die Funktion g. Es gilt:

$$g : \mathbb{R}^n \to \mathbb{R}^m, \quad \xi \mapsto W^{-1}f(B\xi) \tag{6.108}$$

und

$$J_g(\xi) = W^{-1}J_f(B\xi)B. \tag{6.109}$$

Wichtig sind nun Eigenschaften von f, die unabhängig von der Wahl der Basen des $\mathbb{R}^n$, $\mathbb{R}^m$ sind (die also für f und g gelten).
Dazu gehören Stetigkeit, Differenzierbarkeit und spezielle geometrische Eigenschaften.
Analog zu den bisher betrachteten Funktionstypen gibt es auch für Vektorfelder $f : \mathbb{D} \to \mathbb{R}^m$, $\mathbb{D} \subseteq \mathbb{R}^n$, eine Kettenregel:

Satz 6.18 (Kettenregel für Vektorfelder)
Seien $f : \mathbb{D} \to \mathbb{G}$, $\mathbb{D} \subseteq \mathbb{R}^n$, $\mathbb{G} \subseteq \mathbb{R}^m$ und $g : \mathbb{G} \to \mathbb{R}^q$ zwei Vektorfelder, wobei f in $x_0 \in \mathbb{D}$ und g in $f(x_0)$ total differenzierbar sind, dann ist auch die Funktion $g \circ f : \mathbb{D} \to \mathbb{R}^q$ in x_0 total differenzierbar und es gilt:

$$J_{g\circ f}(x_0) = J_g(f(x_0))\cdot J_f(x_0) \tag{6.110}$$

Speziellen Skalaren- und Vektorfeldern werden wichtige Abbildungen zugeordnet, die in der Physik eine entscheidende Rolle spielen.

(a): Jedem $\mathcal{C}^1(\mathbb{D},\mathbb{R})$-Skalarenfeld $f:\mathbb{D}\to\mathbb{R}$, $\mathbb{D}\subseteq\mathbb{R}^n$, wird das Vektorfeld

$$\text{grad } f:\mathbb{D}\to\mathbb{R}^n,\quad x\mapsto\begin{pmatrix}\frac{\partial f}{\partial x_1}(x)\\ \vdots\\ \frac{\partial f}{\partial x_n}(x)\end{pmatrix} \tag{6.111}$$

zugeordnet.

(b): Jedem $\mathcal{C}^2(\mathbb{D},\mathbb{R})$-Skalarenfeld $f:\mathbb{D}\to\mathbb{R}$, $\mathbb{D}\subseteq\mathbb{R}^n$, wird das Skalarenfeld

$$\Delta f:\mathbb{D}\to\mathbb{R},\quad x\mapsto\sum_{i=1}^{n}\frac{\partial^2 f}{\partial x_i^2}(x)=\sum_{i=1}^{n}f_{x_ix_i}(x) \tag{6.112}$$

zugeordnet. Δ wird als Laplace-Operator bezeichnet.

(c): Speziell für $\mathbb{D}\subseteq\mathbb{R}^3$ wird einem $\mathcal{C}^1(\mathbb{D},\mathbb{R}^3)$-Vektorfeld $f:\mathbb{D}\to\mathbb{R}^3$ das Skalarenfeld „Divergenz“

$$\text{div } f:\mathbb{D}\to\mathbb{R},\quad x\mapsto\frac{\partial f_1}{\partial x_1}(x)+\frac{\partial f_2}{\partial x_2}(x)+\frac{\partial f_3}{\partial x_3}(x) \tag{6.113}$$

und das Vektorfeld „Rotation“

$$\text{rot } f:\mathbb{D}\to\mathbb{R}^3,\quad x\mapsto\begin{pmatrix}\frac{\partial f_3}{\partial x_2}(x)-\frac{\partial f_2}{\partial x_3}(x)\\ \frac{\partial f_1}{\partial x_3}(x)-\frac{\partial f_3}{\partial x_1}(x)\\ \frac{\partial f_2}{\partial x_1}(x)-\frac{\partial f_1}{\partial x_2}(x)\end{pmatrix} \tag{6.114}$$

zugeordnet.

Die Bedeutung dieser Abbildungen liegt in Invarianzeigenschaften bei Basiswechseln. Sei $(e_1,\dots,e_n)$ die kanonische Basis des $\mathbb{R}^n$ und $B=(b_1,\dots,b_n)$ eine weitere orthonormale Basis des $\mathbb{R}^n$, (also $b_i\perp b_j$, $i\neq j$, und $|b_i|=1$ für alle $i=1,\dots,n$), so gilt für ein $\mathcal{C}^1(\mathbb{D},\mathbb{R})$-Skalarenfeld $f:\mathbb{D}\to\mathbb{R}$:

$$|\text{grad } f(x)|=|B^{-1}\text{grad } f(x)|=|\text{grad } g(y)|, \tag{6.115}$$

wobei $g:\mathbb{R}^n\to\mathbb{R}$ die Funktion f in der Basis B repräsentiert ($y=B^{-1}x$)
Für ein $\mathcal{C}^1(\mathbb{D},\mathbb{R}^3)$-Vektorfeld $v:\mathbb{D}\to\mathbb{R}^3$, $\mathbb{D}\subseteq\mathbb{R}^3$, gilt analog:

$$|\text{rot } v(x)|=|\text{rot } w(y)|, \tag{6.116}$$

wobei w das Vektorfeld v bezüglich der Basis $(b_1,\dots,b_n)$ repräsentiert. (also: $w(y)=B^{-1}v(x), y=B^{-1}x$)
Für die Divergenz gilt sogar:

$$\text{div } v(x)=\text{div } w(y). \tag{6.117}$$

Rechenregeln:

$$\begin{aligned}
&\text{(a): } \text{rot}\circ(\text{grad } f)\equiv 0,\\
&\text{(b): } \text{div}\circ(\text{rot } v)\equiv 0,\\
&\text{(c): } \text{div}\circ(\text{grad } f)=\Delta f,\\
&\text{(d): } \text{div}\circ(f\cdot v)=(\text{grad } f)\cdot v+f\cdot\text{div } v,\\
&\text{(e): } \text{rot}\circ(\text{rot } v)=\text{grad}\circ(\text{div } v)-\begin{pmatrix}\Delta v_1\\ \Delta v_2\\ \Delta v_3\end{pmatrix}.
\end{aligned} \tag{6.118}$$

für entsprechende Abbildungen $f : \mathbb{D} \subseteq \mathbb{R}^3 \to \mathbb{R}$, $v : \mathbb{G} \subseteq \mathbb{R}^3 \to \mathbb{R}^3$.
Eine der wichtigsten Anwendungen der Differentiation ist die Betrachtung von implizit definierten Funktionen. Häufig können zu untersuchende Abbildungen $f : \mathbb{D} \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, nicht explizit in der Form $x \mapsto \ldots$ angegeben werden, aber mit $F : \mathbb{G} \subset \mathbb{R}^{m+1}$ durch eine implizite Definition der Form

$$F(x, f(x)) = 0 \quad \text{für alle } x \in \mathbb{D}. \tag{6.119}$$

Es stellt sich die Frage, unter welchen Voraussetzungen eine Funktion implizit definiert ist und wie man f gegebenenfalls approximieren kann. Zur Betrachtung dieser Frage dient der folgende Satz, der ein überaus wichtiges Resultat der Mathematik darstellt.

Satz 6.19 (Satz über implizite Funktionen)
Seien $\mathbb{D}_1 \subseteq \mathbb{R}^n$ und $\mathbb{D}_2 \subseteq \mathbb{R}^m$ offen und

$$F : \mathbb{D}_1 \times \mathbb{D}_2 \to \mathbb{R}^m, \quad (x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m}) \mapsto F(x_1, \ldots, x_{n+m}) \tag{6.120}$$

eine $\mathcal{C}^1(\mathbb{D}_1 \times \mathbb{D}_2, \mathbb{R}^m)$-Funktion Sei ferner $(\xi, \eta) \in \mathbb{D}_1 \times \mathbb{D}_2$ ein Punkt mit

$$F(\xi, \eta) = 0 \quad \text{und} \quad D_2F(\xi, \eta) := \begin{pmatrix} \frac{\partial F_1}{\partial x_{n+1}}(\xi, \eta) & \ldots & \frac{\partial F_1}{\partial x_{n+m}}(\xi, \eta) \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_{n+1}}(\xi, \eta) & \ldots & \frac{\partial F_m}{\partial x_{n+m}}(\xi, \eta) \end{pmatrix} \text{ derart,} \tag{6.121}$$

dass $D_2F(\xi, \eta)$ invertierbar ist, so gibt es eine offene Menge $\mathbb{D} \subseteq \mathbb{D}_1$ mit $\xi \in \mathbb{D}$ und eine $\mathcal{C}^1(\mathbb{D}, \mathbb{R}^m)$-Funktion $f : \mathbb{D} \to \mathbb{R}^m$, $(x_1, \ldots, x_n) \mapsto f(x)$ mit:

$$F(x_1, \ldots, x_n, f(x_1, \ldots, x_n)) = 0 \quad \text{für alle } (x_1, \ldots, x_n) \in \mathbb{D}. \tag{6.122}$$

Ferner gilt neben $f(\xi) = \eta$:

$$J_f(\xi) = -D_2F(\xi, \eta)^{-1} \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(\xi, \eta) & \ldots & \frac{\partial F_1}{\partial x_n}(\xi, \eta) \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_1}(\xi, \eta) & \ldots & \frac{\partial F_m}{\partial x_n}(\xi, \eta) \end{pmatrix}. \tag{6.123}$$

Obwohl also die in Satz 6.19 betrachtete Funktion $f : \mathbb{D} \to \mathbb{R}^m$ nur implizit gegeben ist, ist es dennoch möglich, durch

$$f(x) = \eta - D_2F(\xi, \eta)^{-1} \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(\xi, \eta) & \ldots & \frac{\partial F_1}{\partial x_n}(\xi, \eta) \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_1}(\xi, \eta) & \ldots & \frac{\partial F_m}{\partial x_n}(\xi, \eta) \end{pmatrix} (x - \xi) + o(|x - \xi|) \tag{6.124}$$

eine Approximation dieser Funktion anzugeben.

Beispiel(e) 6.20
$F : \mathbb{R}^3 \to \mathbb{R}^2$, $(x_1, x_2, x_3) \mapsto \begin{pmatrix} x_1^3 + x_2^3 + x_3^3 - 7 \\ x_1x_2 + x_2x_3 + x_3x_1 + 2 \end{pmatrix}$
Es gilt: $F(2, -1, 0) = 0$. Für $n = 1$ und $m = 2$ erhalten wir:
$\xi = 2$, $\eta = (-1, 0)$, $\mathbb{D}_1 = \mathbb{R}$, $\mathbb{D}_2 = \mathbb{R}^2$, $D_2F(\xi, \eta) = \begin{pmatrix} 3 & 0 \\ 2 & 1 \end{pmatrix}$.
Es ergibt sich somit die Existenz einer implizit definierten Funktion $f : \mathbb{D} \to \mathbb{R}^2$, $\mathbb{D} \subseteq \mathbb{R}$ offen mit $2 \in \mathbb{D}$, $f(2) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ und

$$F(x, f_1(x), f_2(x)) = 0 \quad \text{für alle } x \in \mathbb{D}. \tag{6.125}$$

Für f lässt sich die folgende Approximation angeben:

$$f(x) = \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} 3 & 0 \\ 2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 12 \\ -1 \end{pmatrix} (x - 2) + o(|x - 2|). \tag{6.126}$$

Der Satz über implizite Funktionen spielt zum Beispiel bei implizit gegebenen Lösungen gewöhnlicher Differentialgleichungen eine wichtige Rolle.

6.2 Integration

Im Folgenden betrachten wir Funktionen, deren Funktionswerte durch die Integration gegebener Funktionen bestimmt werden. Dies kann auf verschiedene Weisen geschehen:

$$F : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}; \qquad x \mapsto \int_0^\infty t^{x-1} e^{-t} dt, \tag{6.127}$$

$$F : \mathbb{R} \to \mathbb{R}; \qquad x \mapsto \frac{1}{\pi} \int_0^\pi \cos(x \cdot \sin(t) - t) dt, \tag{6.128}$$

$$F : \mathbb{R} \to \mathbb{R}; \qquad x \mapsto \int_{-\infty}^x e^{-\frac{t^2}{2}} dt. \tag{6.129}$$

Integrale dieser Form werden als Parameterintegrale bezeichnet, da die Stelle x im Integral als Parameter vorkommt. Betrachten wir nun den Fall, dass der Parameter x nicht in den Integrationsgrenzen auftritt:

Satz 6.21 (Grenzwertvertauschung bei bestimmtem Parameterintegral)
Seien $\mathbb{D} := [a,b] \times [c,d] = \{(x,y) \in \mathbb{R}^2;\ a \le x \le b \text{ und } c \le y \le d\}$ und $f : \mathbb{D} \to \mathbb{R}$ stetig. Dann gilt für die Funktion

$$F : [a,b] \to \mathbb{R}, \quad x \mapsto \int_c^d f(x,y)dy : \tag{6.130}$$

(a): F ist auf $[a,b]$ stetig,

(b):

$$\int_a^b F(x)dx = \int_a^b \left(\int_c^d f(x,y)dy \right) dx = \int_c^d \left(\int_a^b f(x,y)dx \right) dy \tag{6.131}$$

(Satz von Fubini)

(c): Ist zusätzlich f auf $[a,b]$ stetig partiell nach x differenzierbar, dann ist F differenzierbar mit

$$F' : [a,b] \to \mathbb{R}; \quad x \mapsto \frac{d}{dx} \int_c^d f(x,y)dy = \int_c^d \frac{\partial f}{\partial x}(x,y)dy. \tag{6.132}$$

Beispiel(e) 6.22

- $F : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \int_1^\pi \frac{\sin(tx)}{t} dt,$

$$F' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \int_1^\pi \cos(tx)dt, \quad F'' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -\int_1^\pi t\sin(tx)dt \tag{6.133}$$

- Die Bessel-Funktionen

$$J_n : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{\pi} \int_0^\pi \cos(x \cdot \sin(t) - nt)dt, \quad n \in \mathbb{Z}. \tag{6.134}$$

Die Bessel-Funktionen wurden von F.W. Bessel (1784-1846) zur Berechnung von Planetenbahnen verwendet. Es gilt:

$$J_n' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{n}{\pi} \int_0^\pi \cos(t)\cos(x\cdot\sin(t) - nt)dt - \frac{x}{\pi} \int_0^\pi \cos^2(t)\cos(x\cdot\sin(t) - nt)dt. \tag{6.135}$$

Die Bedeutung der Bessel-Funktionen liegt in der Tatsache, dass sie Lösungen spezieller Differentialgleichungen sind, die insbesondere in der Astronomie eine wichtige Rolle spielen.
Hängen im Parameterintegral auch die Integrationsgrenzen von x ab, gibt es also stetig differenzierbare Funktionen $g, h : \mathbb{R} \to \mathbb{R}$ und eine stetig partiell nach x differenzierbare Funktion

$f : \mathbb{R}^2 \to \mathbb{R}$, so gilt mit $F : \mathbb{R} \to \mathbb{R}$, $x \mapsto \int_{g(x)}^{h(x)} f(x,y)dy$:

$$F' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \int_{g(x)}^{h(x)} f_x(x,y)dy + f(x,h(x)) \cdot h'(x) - f(x,g(x)) \cdot g'(x) \text{ (Leibniz-Regel).} \tag{6.136}$$

Beispiel(e) 6.23

- $F : \mathbb{R} \to \mathbb{R}$, $x \mapsto \int_0^{x^2} \cos(xy^2)dy$

$$F' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -\int_0^{x^2} \sin(xy^2)y^2dy + \cos(x^5) \cdot 2x. \tag{6.137}$$

- $F : \mathbb{R} \to \mathbb{R}$, $x \mapsto \frac{1}{k}\int_0^x q(y)\sin(k(x-y))dy$ (mit $q : \mathbb{R} \to \mathbb{R}$ stetig)

$$F' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \int_0^x q(y)\cos(k(x-y))dy \tag{6.138}$$

$$F'' : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -k\int_0^x q(y)\sin(k(x-y))dy + q(x) \tag{6.139}$$

Somit gilt:

$$F''(x) + k^2 \cdot F(x) = q(x), \quad F(0) = F'(0) = 0 \quad \text{(Schwingungsgleichung).} \tag{6.140}$$

Die Aussagen von Satz 6.21 sind nicht ohne Zusatzeinschränkungen auf uneigentliche Integrale anwendbar. Für die Praxis ist der folgende Satz ausreichend.

Satz 6.24 (Grenzwertvertauschung bei uneigentlichen Integralen)
Seien $a, b, c \in \mathbb{R}$, $d \in \mathbb{R} \cup \{\infty\}$, $\mathbb{D} := \{(x, y) \in \mathbb{R}^2; a \leq x \leq b \text{ und } c \leq x < d\}$ und $f : \mathbb{D} \to \mathbb{R}$ stetig und stetig partiell nach x differenzierbar.
Seien ferner $g, h : [c, d) \to \mathbb{R}$ Funktionen mit

(a): $|f(x, y)| \leq g(y)$ und $|f_x(x, y)| \leq h(y)$ für alle $(x, y) \in \mathbb{D}$

(b): Die uneigentlichen Integrale

$$\int_c^d g(y)dy, \quad \int_c^d h(y)dy \tag{6.141}$$

sind konvergent,

dann existiert die Funktion

$$F : [a, b] \to \mathbb{R}, \quad x \mapsto \int_c^d f(x, y)dy \tag{6.142}$$

und ist differenzierbar mit

$$F' : [a, b] \to \mathbb{R}, \quad x \mapsto \int_c^d f_x(x, y)dy. \tag{6.143}$$

Beispiel(e) 6.25
$\Gamma : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}$, $x \mapsto \int_0^\infty e^{-t}t^{x-1}dt$

$$\Gamma' : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}, \quad x \mapsto \int_0^\infty e^{-t}t^{x-1}\ln(t)dt \tag{6.144}$$

$$\Gamma'' : \mathbb{R}_0^+ \setminus \{0\} \to \mathbb{R}, \quad x \mapsto \int_0^\infty e^{-t}t^{x-1}(\ln(t))^2dt \tag{6.145}$$

Für das uneigentliche Integral $\int_0^\infty e^{-xy}\cos(y)dy$ mit $x \in \mathbb{R}_0^+ \setminus \{0\}$ gilt einerseits mit zweifacher partieller Integration:

$$\lim_{x\to 0} \int_0^\infty e^{-xy}\cos(y)dy = 0 \tag{6.146}$$

Andererseits würde bei Vertauschung von Grenzwert und Integration gelten:

$$\lim_{x\to 0} \int_0^\infty e^{-xy}\cos(y)dy = \int_0^\infty (\lim_{x\to 0} e^{-xy})\cos(y)dy = \int_0^\infty \cos(y)dy \text{ (nicht definiert).} \tag{6.147}$$

Im Folgenden betrachten wir Kurvenintegrale, die in der Physik etwa bei Schwerpunktsberechnungen eine wichtige Rolle spielen.

Definition 6.26 (Kurvenintegrale)
Seien $\mathbb{D} \subseteq \mathbb{R}^n$ offen, $w : [a, b] \to \mathbb{D}$ ein Kurvenstück und $f : \mathbb{D} \to \mathbb{R}$ eine Funktion derart, dass $f \circ w$ stetig ist, dann heißt

$$\int_w f ds := \int_a^b f(w(t)) \cdot |\dot{w}(t)| dt \tag{6.148}$$

das Kurvenintegral von f längs w.

Durchlaufen zwei Kurvenstücke $w : [a, b] \to \mathbb{R}^n$ und $u : [c, d] \to \mathbb{R}^n$ dieselbe Spur

$$\{w(t); a \leq t \leq b\} = \{u(t); c \leq t \leq d\} \tag{6.149}$$

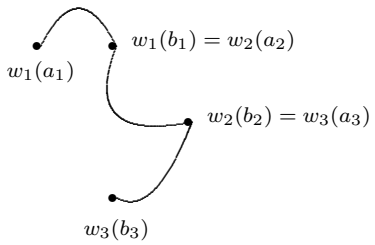
genau gleich oft, dann gilt:

$$\int_a^b f(w(t))|\dot{w}(t)|dt = \int_c^d f(u(t)) \cdot |\dot{u}(t)|dt. \tag{6.150}$$

Die Schreibweise $\int_w f ds$ verdeutlicht diese Eigenschaft.

Es ist zweckmäßig, den Begriff des Kurvenintegrals in naheliegender Weise auf Integrationswege zu erweitern, die aus Kurvenstücken zusammengesetzt sind. Sei also eine Kurve w in $\mathbb{D} \subseteq \mathbb{R}^n$ derart gegeben, dass es (stetig differenzierbare) Kurvenstücke $w_i : [a_i, b_i] \to \mathbb{D}$, $i = 1, \ldots, k$, mit $b_i = a_{i+1}$ und $w_i(b_i) = w_{i+1}(a_{i+1})$, $i = 1, \ldots, k-1$ und $w : [a_1, b_k]$, $t \mapsto w_i(t)$, falls $t \in [a_i, b_i]$ gilt, so ist das Kurvenintegral $\int_w f ds$ für eine stetige Funktion $f : \text{Spur}(w) \to \mathbb{R}$ gegeben durch

$$\int_w f ds = \sum_{i=1}^{k} \int_{w_i} f ds. \tag{6.151}$$



(6.152)

Beispiel(e) 6.27

- Die Bogenlänge L von w ergibt sich durch die Wahl $f \equiv 1$. Somit erhält man die bereits bekannte Formel

$$L(w) = \int_a^b |\dot{w}(t)|dt \tag{6.153}$$

- Der geometrische Schwerpunkt $(\overline{x}, \overline{y}, \overline{z})$ eines Kurvenstückes $w : [a, b] \to \mathbb{R}^3$ berechnet sich zu

$$\overline{x} = \frac{1}{L}\int_w x ds, \quad \overline{y} = \frac{1}{L}\int_w y ds, \quad \overline{z} = \frac{1}{L}\int_w z ds, \tag{6.154}$$

  wobei L die Bogenlänge des Kurvenstückes w bezeichnet.
  Für eine Schraubenlinie $w : [0, 2\pi] \to \mathbb{R}^3$, $t \mapsto (r \cdot \cos(t), r \cdot \sin(t), h \cdot t)$ der Höhe $h \cdot 2\pi$ erhält man den geometrischen Schwerpunkt:

$$\overline{x} = \frac{1}{L}\int_0^{2\pi} r \cdot \cos(t) \cdot \sqrt{r^2 + h^2}dt = 0 \tag{6.155}$$

$$\overline{y} = \frac{1}{L}\int_0^{2\pi} r \cdot \sin(t) \cdot \sqrt{r^2 + h^2}dt = 0 \tag{6.156}$$

$$\overline{z} = \frac{1}{L}\int_0^{2\pi} h \cdot t \cdot \sqrt{r^2 + h^2}dt = \frac{1}{2\pi\sqrt{r^2 + h^2}} \cdot \left(h\sqrt{r^2 + h^2}\right) 2\pi^2 = \pi h \tag{6.157}$$

Als Verallgemeinerung des Kurvenintegrals kann dieses auch für Vektorfelder definiert werden:

Definition 6.28 (Kurvenintegral für Vektorfelder)
Seien $\mathbb{D} \subseteq \mathbb{R}^n$ offen, $w : [a, b] \to \mathbb{D}$ eine reguläre Kurve und $v : \mathbb{D} \to \mathbb{R}^n$ ein stetiges Vektorfeld, so heißt

$$\int_w v dx := \int_a^b v(w(t))^\top \dot{w}(t)dt \tag{6.158}$$

das Kurvenintegral von v längs w.

Für Kurvenintegrale gilt die Linearität:

$$\int_w (u + v)dx = \int_w u dx + \int_w v dx \tag{6.159}$$

$$\int_w (\lambda \cdot v)dx = \lambda \int_w v dx \tag{6.160}$$

für stetige Vektorfelder u, v, eine reguläre Kurve w und ein $\lambda \in \mathbb{R}$.
Ist $w(a) = w(b)$ (geschlossene Kurve), so schreibt man

$$\oint_w f ds \text{ für } \int_w f ds \quad \text{bzw.} \quad \oint_w v dx \text{ für } \int_w v dx \tag{6.161}$$

Die von einer Kraft $K : \mathbb{R}^3 \to \mathbb{R}^3$ längs eines Weges $w : [a, b] \to \mathbb{R}^3$ geleistete Arbeit A ist

gegeben durch

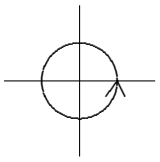
$$A = \int_w K dx = \int_a^b K(w(t))^\top \dot{w}(t) dt \tag{6.162}$$

Im Gegensatz zu Kurvenintegralen über Skalarenfelder hängt aber das Kurvenintegral über Vektorfelder davon ab, wie die Spur der Kurve durchlaufen wird:

Sei $v : \mathbb{R}^2 \setminus \{(0,0)\} \to \mathbb{R}^2$, $(x,y) \mapsto \begin{pmatrix} \frac{-y}{x^2+y^2} \\ \frac{x}{x^2+y^2} \end{pmatrix}$

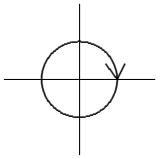
und

$$w_1 : [0, 2\pi] \to \mathbb{R}^2, \quad t \mapsto \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix} \tag{6.163}$$



(6.164)

$$w_2 : [0, 2\pi] \to \mathbb{R}^2, \quad t \mapsto \begin{pmatrix} \cos(t) \\ -\sin(t) \end{pmatrix} \tag{6.165}$$



(6.166)

so gilt:

$$\begin{aligned}
\int_{w_1} v dx &= \int_0^{2\pi} \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}^\top \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix} dt = \\
&= \int_0^{2\pi} (\sin^2(t) + \cos^2(t)) dt = -\int_0^{2\pi} (-\sin^2(t) - \cos^2(t)) dt \\
&= \int_0^{2\pi} \begin{pmatrix} \sin(t) \\ \cos(t) \end{pmatrix}^\top \begin{pmatrix} -\sin(t) \\ -\cos(t) \end{pmatrix} dt = -\int_{w_2} v dx
\end{aligned}$$

Ein wichtiger Spezialfall von Kurvenintegralen über Vektorfelder liegt vor, falls die betrachteten Vektorfelder als Gradienten von Skalarenfeldern gegeben sind.

Definition 6.29 (Gebiet, Gradientenfeld)
Eine Teilmenge $\mathbb{G} \subseteq \mathbb{R}^n$ heißt Gebiet, falls $\mathbb{G}$ offen ist und falls $\mathbb{G}$ zusammenhängend ist, falls also mit zwei Punkten x_0, y_0 aus $\mathbb{G}$ eine reguläre Kurve $w : [a,b] \to \mathbb{G}$ mit $w(a) = x_0$ und $w(b) = y_0$ existiert. Ein auf einem Gebiet $\mathbb{G}$ definiertes stetiges Vektorfeld $v : \mathbb{G} \to \mathbb{R}^n$ heißt stetiges Gradientenfeld, falls es eine Funktion $f : \mathbb{G} \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{G}, \mathbb{R})$ gibt mit:

$$v = \text{grad } f. \tag{6.167}$$

f heißt Stammfunktion von v.

Kurvenintegrale über stetige Gradientenfelder haben nun interessante Eigenschaften

Satz 6.30 (1. Hauptsatz für Kurvenintegrale)
Sei $v : \mathbb{G} \to \mathbb{R}^n$ ein stetiges Gradientenfeld auf dem Gebiet $\mathbb{G} \subseteq \mathbb{R}^n$ mit einer Stammfunktion f, so gilt für jede stückweise reguläre Kurve w in $\mathbb{G}$ mit Anfangspunkt $w(a)$ und Endpunkt $w(b)$:

$$\int_w v dx = f(w(b)) - f(w(a)). \tag{6.168}$$

Charakteristisch für das Kurvenintegral über Gradientenfelder ist also die Tatsache, dass dieses Integral nur vom Anfangs- und Endpunkt der Kurve abhängt.
Diese Aussage lässt sich sogar umkehren, denn es gilt für ein stetiges Vektorfeld $v : \mathbb{G} \to \mathbb{R}^n$ auf einem Gebiet $\mathbb{G} \subseteq \mathbb{R}^n$: v ist genau dann ein Gradientenfeld, falls für alle regulären Kurven w in $\mathbb{G}$ das Integral $\int_w v dx$ nur vom Anfangs- und Endpunkt von w abhängt. (Man sagt, das Integral ist wegunabhängig). Dies ist genau dann der Fall, falls für alle geschlossenen regulären Kurven w in $\mathbb{G}$ gilt:

$$\oint_w v dx = 0. \tag{6.169}$$

Wie bei der eindimensionalen Integration unterscheiden sich zwei Stammfunktionen eines Gradientenfeldes nur um eine additive Konstante. Um nun entscheiden zu können, ob ein gegebenes v wirklich ein Gradientenfeld ist, betrachtet man eine Kurve w in $\mathbb{G}$ zwischen zwei Punkten $x_0, x \in \mathbb{G}$ und berechnet

$$f : \mathbb{G} \to \mathbb{R}, \quad x \mapsto \int_w v dx. \tag{6.170}$$

Ist v ein Gradientenfeld, so ist f eine Stammfunktion.
Ist $\mathbb{G}$ konvex, so ist etwa $w : [0, 1] \to \mathbb{R}^n$, $t \mapsto x_0 + t(x - x_0)$ zu wählen. Man erhält:

$$f : \mathbb{G} \to \mathbb{R}, \quad x \mapsto \int_0^1 v(x_0 + t(x - x_0))^\top (x - x_0) dt. \tag{6.171}$$

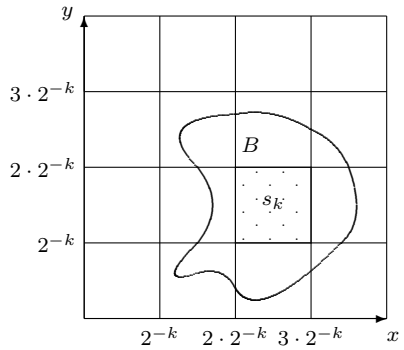
Neben Kurvenintegralen lassen sich auch Integrale über spezielle Teilmengen des $\mathbb{R}^n$ als Integrationsbereiche definieren. Wir beginnen mit Teilmengen der Ebene.

Definition 6.31 (regulärer Bereich im $\mathbb{R}^2$)
Eine Teilmenge $\mathbb{B} \subseteq \mathbb{R}^2$ heißt regulärer Bereich im $\mathbb{R}^2$, falls

- der Rand $\partial\mathbb{B}$ aus endlich vielen regulären Kurvenstücken besteht,
- das Innere $\mathbb{B} \setminus \partial\mathbb{B}$ ein nichtleeres, beschränktes Gebiet ist,
- $\mathbb{B}$ abgeschlossen ist (also $\partial\mathbb{B} \subseteq \mathbb{B}$).

Einem regulären Bereich im $\mathbb{R}^2$ kann man immer einen Flächeninhalt zuordnen.
Dazu zerlegen wir für festes $k \in \mathbb{N}$ den $\mathbb{R}^2$ durch die Geraden $x = n \cdot 2^{-k}$ und $y = n \cdot 2^{-k}$, $n \in \mathbb{Z}$,

in Quadrate der Fläche 2^{-2k}.

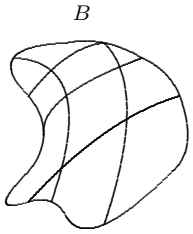


(6.172)

Sei nun $s_k(\mathbb{B})$ die gemeinsame Fläche aller Quadrate, die einschließlich ihres Randes ganz in $\mathbb{B}$ liegen und $S_k(\mathbb{B})$ die gemeinsame Fläche alle Quadrate, die mindestens einen Punkt von $\mathbb{B}$ enthalten, so gilt:

$$s_k(\mathbb{B}) \leq S_k(\mathbb{B}), \quad s_k(\mathbb{B}) \leq s_{k+1}(\mathbb{B}), \quad S_{k+1}(\mathbb{B}) \leq S_k(\mathbb{B}) \tag{6.173}$$

und $\lim\limits_{k\to\infty} s_k(\mathbb{B}) = \lim\limits_{k\to\infty} S_k(\mathbb{B}) =: F(\mathbb{B})$. $F(\mathbb{B})$ heißt der Flächeninhalt von $\mathbb{B}$.
Im Folgenden lassen wir als Integrationsbereiche Teilmengen des $\mathbb{R}^2$ zu, die aus regulären Bereichen $\mathbb{B} \subset \mathbb{R}^2$ bestehen, aus denen gegebenenfalls Bereiche $\mathbb{L}$ mit Flächeninhalt Null (also Punkte oder Kurvenstücke) herausgeschnitten sind. Wir schreiben dafür $\mathbb{B} \setminus \mathbb{L}$. Sei nun $f : \mathbb{B} \to \mathbb{R}$ eine beschränkte Funktion (also $m \leq f(x,y) \leq M$ für alle $(x,y) \in \mathbb{B}$), die auf $\mathbb{B} \setminus \mathbb{L}$ stetig ist, so betrachtet man eine Zerlegung $\mathbb{B}_1, \ldots, \mathbb{B}_n$ des Bereiches $\mathbb{B}$ durch reguläre Kurven



(6.174)

und betrachtet die Summe

$$Z_n := \sum_{i=1}^{n} f(x_i^*, y_i^*) F(\mathbb{B}_i), \tag{6.175}$$

wobei $(x_i^*, y_i^*) \in \mathbb{B}_i$ und $F(\mathbb{B}_i)$ die Fläche von $\mathbb{B}_i$ darstellt.
Wählt man nun für jedes $n \in \mathbb{N}$ die Zerlegung $(\mathbb{B}_1, \ldots, \mathbb{B}_n)$ von $\mathbb{B}$ derart, dass die Flächen $F(\mathbb{B}_1), \ldots, F(\mathbb{B}_n)$ für wachsende n gegen Null konvergieren, so konvergiert die Summe Z_n und wird mit $\int\limits_{\mathbb{B}} f(x,y)dF$ oder $\int\limits_{\mathbb{B}} f dF$ bezeichnet. Es gilt also

$$\int_{\mathbb{B}} f(x,y)dF = \lim_{n\to\infty} Z_n \tag{6.176}$$

Wegen

$$F(\mathbb{B} \setminus \mathbb{L}) = F(\mathbb{B}) \tag{6.177}$$

gilt:

$$\int_{\mathbb{B}} f(x,y)dF = \int_{\mathbb{B}\setminus\mathbb{L}} f(x,y)dF. \tag{6.178}$$

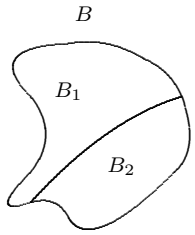
Geometrische Interpretation:

- Für $f \equiv 1$ gilt:
$$\int_{\mathbb{B}} f(x,y)dF = F(\mathbb{B}). \tag{6.179}$$
- Für $f(x,y) \geq 0$ für alle $x, y \in \mathbb{B}$ gilt:
$$V(K) := \int_{\mathbb{B}} f(x,y)dF \tag{6.180}$$
ist das Volumen des Körpers $K := \{(x,y,z) \in \mathbb{R}^3;\ (x,y) \in \mathbb{B} \text{ und } 0 \leq z \leq f(x,y)\}$.

Rechenregeln:

- $\int_{\mathbb{B}} (af + bg)dF = a\int_{\mathbb{B}} fdF + b\int_{\mathbb{B}} gdF, \quad a,b \in \mathbb{R}, \quad f,g : \mathbb{B} \to \mathbb{R}$
- $\int_{\mathbb{B}} fdF \leq \int_{\mathbb{B}} gdF$, falls $f(x,y) \leq g(x,y)$ für alle $(x,y) \in \mathbb{B}$ (Monotonie)
- $\int_{\mathbb{B}} fdF = \int_{\mathbb{B}_1} fdF + \int_{\mathbb{B}_2} fdF$,
falls $\mathbb{B}$ durch eine stückweise reguläre Kurve in zwei Teilbereiche $\mathbb{B}_1$ und $\mathbb{B}_2$ zerlegt wird (Additivität)

(6.181)



Zerlegt man die (x,y)-Ebene durch ein achsenparalleles Gitter, so entstehen Rechtecke der Seitenlängen $\Delta x, \Delta y$ und Flächeninhalt ΔF. Verwendet man für die Definition von $\int_{\mathbb{B}} fdF$ Zerlegungen, die auf derartigen Gittern basieren, so schreibt man dafür $\int_{\mathbb{B}} f(x,y)dxdy$. Analog zur Riemann-Integration hängt der Wert des Integrals nicht von den geometrischen Eigenschaften der verwendeten Zerlegung ab. Es stellt sich die Frage, für welche Integrationsbereiche das Integral $\int_{\mathbb{B}} f(x,y)dxdy$ durch zwei nacheinander auszuführende eindimensionale Riemann-Integrale berechnet werden kann.

Definition 6.32 (Normalbereich im $\mathbb{R}^2$)
Eine Menge $\mathbb{B}_1 \subseteq \mathbb{R}^2$ heißt Normalbereich vom Typ I, wenn es zwei reelle Zahlen a, b und zwei $\mathcal{C}^1([a,b],\mathbb{R})$-Funktionen $g, h : [a,b] \to \mathbb{R}$ gibt mit $g(x) \leq h(x)$ für alle $x \in [a,b]$ und

$$\mathbb{B}_1 : \{(x,y) \in \mathbb{R}^2; a \leq x \leq b, g(x) \leq y \leq h(x)\}. \tag{6.182}$$

Eine Menge $\mathbb{B}_2 \subseteq \mathbb{R}^2$ heißt Normalbereich vom Typ II, wenn es zwei reelle Zahlen c, d und zwei $\mathcal{C}^1([c,d],\mathbb{R})$-Funktionen $l, r : [c,d] \to \mathbb{R}$ gibt mit

$$l(y) \leq r(y) \quad \text{für alle } y \in [c,d] \tag{6.183}$$

und

$$\mathbb{B}_2 := \{(x,y) \in \mathbb{R}^2; l(y) \leq x \leq r(y), c \leq y \leq d\}. \tag{6.184}$$

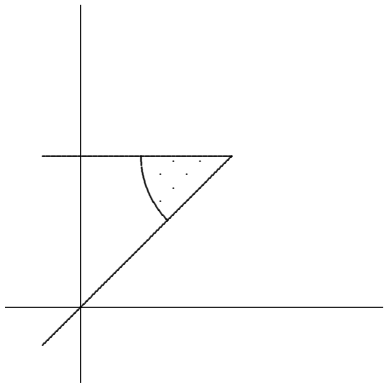
Die Integration über Normalbereiche lässt sich nun durch zweifache Riemann-Integration durchführen:

Satz 6.33 (Integration über Normalbereiche)
Seien $f : \mathbb{B}_1 \to \mathbb{R}$ eine stetige, auf einem Normalbereich vom Typ I definierte Funktion und $k : \mathbb{B}_2 \to \mathbb{R}$ eine stetige, auf einem Normalbereich vom Typ II definierte Funktion, so gilt:

$$\int\limits_{\mathbb{B}_1} f(x,y)dxdy = \int\limits_a^b \left(\int\limits_{g(x)}^{h(x)} f(x,y)dy \right) dx \tag{6.185}$$

$$\int\limits_{\mathbb{B}_2} k(x,y)dxdy = \int\limits_c^d \left(\int\limits_{l(y)}^{r(y)} k(x,y)dx \right) dy \tag{6.186}$$

Beispiel(e) 6.34
Sei $\mathbb{B}$ berandet von $y = x$, $xy = 1$ und $y = 2$



(6.187)

$\mathbb{B}$ ist ein Normalbereich vom Typ II, denn es gilt:

$$\mathbb{B} = \left\{ (x,y) \in \mathbb{R}^2, 1 \leq y \leq 2, \frac{1}{y} \leq x \leq y \right\} \tag{6.188}$$

Die Fläche von $\mathbb{B}$ ergibt sich zu

$$F(\mathbb{B}) = \int\limits_{\mathbb{B}} dxdy = \int\limits_1^2 \left(\int\limits_{\frac{1}{y}}^{y} 1dx \right) dy = \int\limits_1^2 \left(y - \frac{1}{y} \right) dy = \frac{3}{2} - \ln(2) \tag{6.189}$$

Beispiel(e) 6.35
Der geometrische Schwerpunkt des Normalbereichs

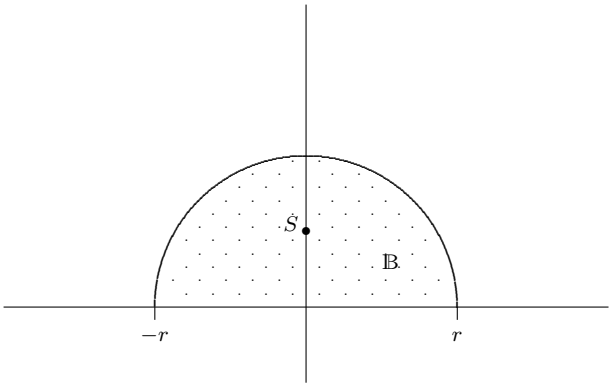
$$\mathbb{B} = \{(x,y) \in \mathbb{R}^2; a \leq x \leq b, 0 \leq y \leq g(x)\} \tag{6.190}$$

hat die Koordinaten

$$\overline{x} = \frac{1}{F}\int\limits_{\mathbb{B}} x dxdy, \quad \overline{y} = \frac{1}{F}\int\limits_{\mathbb{B}} y dxdy, \text{ mit } F = F(\mathbb{B}) = \int\limits_{\mathbb{B}} dxdy \tag{6.191}$$

Speziell für $a = -r$, $b = r$, $g : [-r, r] \to \mathbb{R}$, $x \mapsto \sqrt{r^2 - x^2}$, $r > 0$, erhält man:

$$F(\mathbb{B}) = \int\limits_{-r}^{r} \sqrt{r^2 - x^2} dx = \frac{r^2\pi}{2} \tag{6.192}$$



(6.193)

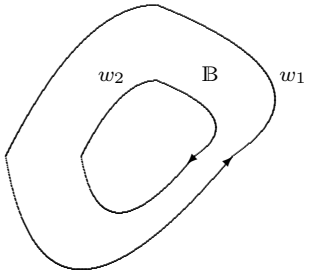
$$\overline{x} = \frac{2}{r^2\pi}\int\limits_{-r}^{r}\left(\int\limits_{0}^{g(x)} x dy\right) dx = \frac{2}{r^2\pi}\int\limits_{-r}^{r} x \cdot g(x) dx = \frac{2}{r^2\pi}\int\limits_{-r}^{r} x\sqrt{r^2 - x^2} dx = 0 \tag{6.194}$$

$$\overline{y} = \frac{2}{r^2\pi}\int\limits_{-r}^{r}\left(\int\limits_{0}^{g(x)} y dy\right) dx = \frac{2}{r^2\pi}\int\limits_{-r}^{r} \frac{g^2(x)}{2} dx = \frac{2}{r^2\pi}\int\limits_{-r}^{r} \frac{r^2 - x^2}{2} dx = \frac{4r}{3\pi} \tag{6.195}$$

Unter gewissen Voraussetzungen an den Bereich $\mathbb{B} \subseteq \mathbb{R}^2$ lassen sich Integrale mit Integrationsgebiet $\mathbb{B}$ auch durch Kurvenintegrale berechnen. Dies ist die Aussage des folgenden wichtigen Satzes, der nach GEORGE GREEN (1793 - 1841) benannt ist.

Satz 6.36 (Integralsatz von Green)
Seien $\mathbb{B} \subseteq \mathbb{R}^2$ ein regulärer Bereich, dessen Rand $\partial\mathbb{B}$ aus endlich vielen geschlossenen, stückweise regulären Kurven $w_1, \ldots, w_n$ besteht.
Die Parametrisierung sei so, dass $\mathbb{B}$ stets links zur Laufrichtung liegt.
Dann gilt für ein $\mathcal{C}^1(\mathbb{B}, \mathbb{R})$-Vektorfeld $v : \mathbb{D} \to \mathbb{R}^2$ mit $\mathbb{B} \subseteq \mathbb{D}$, $\mathbb{D}$ offen:

$$\int\limits_{\partial\mathbb{B}} v dx := \int\limits_{w_1} v dx + \ldots + \int\limits_{w_n} v dx = \int\limits_{\mathbb{B}} \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right) dxdy \tag{6.196}$$



(6.197)

Sonderfälle:

- $v : \mathbb{D} \to \mathbb{R}^2$, $(x, y) \mapsto \begin{pmatrix} 0 \\ x \end{pmatrix}$

$$F(\mathbb{B}) = \int\limits_{\mathbb{B}} dxdy = \int\limits_{\partial\mathbb{B}} v dx. \tag{6.198}$$

- $v : \mathbb{D} \to \mathbb{R}^2$, $(x, y) \mapsto \begin{pmatrix} -y \\ 0 \end{pmatrix}$

$$F(\mathbb{B}) = \int\limits_{\mathbb{B}} dxdy = \int\limits_{\partial\mathbb{B}} v dx. \tag{6.199}$$

Beispiel(e) 6.37

$$w_1 : [0, 2\pi] \to \mathbb{R}^2, \quad t \mapsto \begin{pmatrix} 2\cos(t) \\ 2\sin(t) \end{pmatrix} \tag{6.200}$$

$$w_2 : [0, 2\pi] \to \mathbb{R}^2, \quad t \mapsto \begin{pmatrix} \cos(t) \\ -\sin(t) \end{pmatrix} \tag{6.201}$$

$$F(\mathbb{B}) = \int\limits_0^{2\pi} 4\cos^2(t) dt - \int\limits_0^{2\pi} \cos^2(t) dt = 3\pi \tag{6.202}$$

Für die Anwendung genügt es nicht, nur Integrale über Teilmengen des $\mathbb{R}^2$ einzuführen. Daneben benötigt man auch einen Integralbegriff über Flächen im Raum $\mathbb{R}^3$. Dazu betrachten wir die folgenden Flächenstücke:

Definition 6.38 (reguläres Flächenstück)
Sei $\mathbb{D}$ ein regulärer Bereich in einem Gebiet $\mathbb{G} \subseteq \mathbb{R}^2$. Sei ferner

$$\xi : \mathbb{G} \to \mathbb{R}^3, \quad (u,v) \mapsto (x(u,v), y(u,v), z(u,v))^\top \tag{6.203}$$

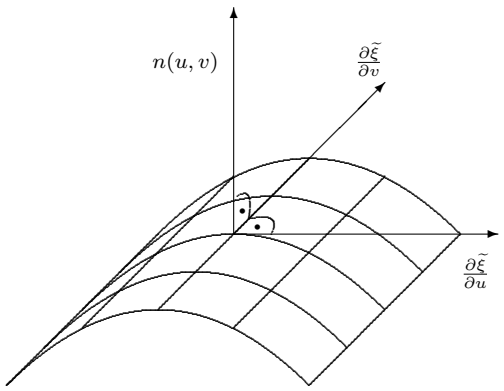
eine $\mathcal{C}^1(\mathbb{G}, \mathbb{R}^3)$-Funktion auf $\mathbb{G}$, so heißt die Einschränkung $\widetilde{\xi} := \xi_{|\mathbb{D}} : \mathbb{D} \to \mathbb{R}^3$ von ξ auf $\mathbb{D}$ Parameterdarstellung eines regulären Flächenstückes, falls gilt:

(a): Für $(u,v) \neq (u',v')$ aus $\mathbb{D}$ gilt $\widetilde{\xi}(u,v) \neq \widetilde{\xi}(u',v')$

(b): $\frac{\partial \widetilde{\xi}}{\partial u}(u,v) \times \frac{\partial \widetilde{\xi}}{\partial v}(u,v) \neq 0$ für alle $(u,v) \in \mathbb{D}$

Die Punktmenge $S = \{\widetilde{\xi}(u,v) \in \mathbb{R}^3; (u,v) \in \mathbb{D}\}$ heißt reguläres Flächenstück.

Der Vektor $n(u,v) := \frac{\frac{\partial \widetilde{\xi}}{\partial u}(u,v) \times \frac{\partial \widetilde{\xi}}{\partial v}(u,v)}{\left|\frac{\partial \widetilde{\xi}}{\partial u}(u,v) \times \frac{\partial \widetilde{\xi}}{\partial v}(u,v)\right|}$ heißt Flächennormale in (u,v)



(6.204)

Die Oberfläche der in den technischen Anwendungen auftretenden dreidimensionalen Körper kann in der Regel nicht durch ein einziges reguläres Flächenstück dargestellt werden (zum Beispiel ein Würfel). Daher ist eine stückweise reguläre Fläche die Vereinigung endlich vieler regulärer Flächenstücke S_i, von denen je zwei längs einem oder mehreren gemeinsamen Randstücken aneinanderstoßen, aber sonst keine anderen Punkte gemeinsam haben. Der Rand ∂S von $S = S_1 \cup \ldots \cup S_n$ wird aus allen Randstücken gebildet, die nur zu einem der regulären Flächenstücke S_i gehören, wobei

$$\partial S_i := \{\widetilde{\xi}(u,v); (u,v) \in \partial \mathbb{D}\}. \tag{6.205}$$

S heißt geschlossen, falls $\partial S = 0$. Für einen Würfel gilt $\partial S = 0$.

Beispiel(e) 6.39

- Ebenen: Für $a, b \in \mathbb{R}^3$, $a \times b \neq 0$ ist

$$\widetilde{\xi}: \mathbb{D} \to \mathbb{R}^3, \quad (u,v) \mapsto x_0 + u \cdot a + v \cdot b \tag{6.206}$$

  ein Ebenenstück.

- Graphen: Sei $h : \mathbb{D} \to \mathbb{R}$, $(x,y) \mapsto h(x,y)$ und $h \in \mathcal{C}^1(\mathbb{D}, \mathbb{R})$, so ist

$$\widetilde{\xi}: \mathbb{D} \to \mathbb{R}^3, \quad (x,y) \mapsto \begin{pmatrix} x \\ y \\ h(x,y) \end{pmatrix} \tag{6.207}$$

  ein reguläres Flächenstück.

- Drehflächen: Aus einem regulären Kurvenstück $t \to \begin{pmatrix} x(t) \\ 0 \\ z(t) \end{pmatrix}$ $t_0 \leq t \leq t_1$ ohne Doppelpunkte mit $x(t) > 0$ entsteht durch Drehung um die z-Achse mit dem Winkel φ_0 $(0 < \varphi_0 \leq 2\pi)$ ein reguläres Flächenstück

$$\widetilde{\xi}: [t_0, t_1] \times [0, \varphi_0] \to \mathbb{R}^3, \quad (t, \varphi) \mapsto \begin{pmatrix} x(t)\cos(\varphi) \\ x(t)\sin(\varphi) \\ z(t) \end{pmatrix}. \tag{6.208}$$

Betrachtet man ein reguläres Flächenstück $\widetilde{\xi} : \mathbb{D} \to \mathbb{R}^3$,

$$(u,v) \mapsto (x(u,v), y(u,v), z(u,v))^\top \tag{6.209}$$

so erhält man (Taylor-Entwicklung):

$$\widetilde{\xi}(u,v) = \widetilde{\xi}(u_0,v_0) + \underbrace{\begin{pmatrix} x_u(u_0,v_0) & x_v(u_0,v_0) \\ y_u(u_0,v_0) & y_v(u_0,v_0) \\ z_u(u_0,v_0) & z_v(u_0,v_0) \end{pmatrix} \begin{pmatrix} u-u_0 \\ v-v_0 \end{pmatrix}}_{\text{Parallelogramm mit Fläche } |\widetilde{\xi}_u(u_0,v_0) \times \widetilde{\xi}_v(v_0,v_0)|(u-u_0)\cdot(v-v_0)} + R \tag{6.210}$$

Wird nun das durch $\widetilde{\xi}$ gegebene Flächenstück S in n Teilstücke zerlegt, die durch n Punkte $(u_1,v_1), \ldots, (u_n,v_n)$ repräsentiert werden, und wird die Fläche jedes dieser n Teilstücke durch die Fläche ΔO_i des entsprechenden Parallelogramms approximiert, so erhält man mit

$$\lim_{n\to\infty} \sum_{i=1}^{n} |\widetilde{\xi}_u(u_i,v_i) \times \widetilde{\xi}_v(u_i,v_i)|(u-u_i)\cdot(v-v_i) = \int_{\mathbb{D}} |\widetilde{\xi}_u \times \widetilde{\xi}_v| du dv \tag{6.211}$$

den Flächeninhalt von S, falls für wachsende n die Flächen der Parallelogramme gegen Null konvergieren.
Für die Oberfläche eines Graphen, gegeben durch $z = h(x,y)$ ergibt sich

$$O = \int_{\mathbb{D}} \sqrt{1 + h_x^2 + h_y^2} dx dy \tag{6.212}$$

Für die Oberfläche einer Drehfläche erhält man

$$O = \int_{\mathbb{D}} \sqrt{x^2(\dot{x}^2 + \dot{z}^2)} dt d\varphi = 2\pi \int_{t_0}^{t_1} x\sqrt{\dot{x}^2 + \dot{z}^2} dt. \tag{6.213}$$

Mit Hilfe des Flächeninhalts von regulären Flächenstücken lassen sich nun Oberflächenintegrale skalarer Funktionen definieren

Definition 6.40 (Oberflächenintegral skalarer Funktionen)
Seien $\mathbb{D} \subseteq \mathbb{R}^2$ ein regulärer Bereich und

$$\widetilde{\xi} : \mathbb{D} \to \mathbb{R}^3, \quad (u,v) \mapsto (x(u,v), y(u,v), z(u,v))^\top \tag{6.214}$$

die Parametrisierung eines regulären Flächenstücks S. Sei ferner $f : \widetilde{\xi}(\mathbb{D}) \to \mathbb{R}$ ein stetiges Skalarenfeld, so heißt

$$\int\limits_S f dO = \int\limits_{\mathbb{D}} f(x(u,v), y(u,v), z(u,v)) \cdot |\widetilde{\xi}_u(u,v) \times \widetilde{\xi}_v(u,v)| du dv \tag{6.215}$$

das Oberflächenintegral von f über S.
Für eine aus regulären Flächenstücken $S_1, \ldots, S_n$ bestehende, stückweise reguläre Fläche S definiert man:

$$\int\limits_S f dO := \sum_{i=1}^{n} \int\limits_{S_i} f dO. \tag{6.216}$$

Für Oberflächenintegrale gelten offenbar die üblichen Rechenregeln Linearität, Monotonie, Additivität und der Mittelwertsatz:
Es existiert ein $\widetilde{\xi}(u^*, v^*) \in S$ mit

$$f(\widetilde{\xi}(u^*, v^*)) = \frac{1}{O(S)} \int\limits_S f dO, \tag{6.217}$$

wobei $O(S)$ die Oberfläche von S bezeichnet.

Beispiel(e) 6.41
Sei $\mathbb{D} = \{(x,y) \in \mathbb{R}^2;\ 0 \le x \le 1 \text{ und } 0 \le y \le 1-x\}$.
Für die Fläche $\widetilde{\xi} : \mathbb{D} \to \mathbb{R}^3,\ (u,v) \mapsto \begin{pmatrix} u \\ v \\ \frac{2}{3}\left(u^{\frac{3}{2}} + v^{\frac{3}{2}}\right) \end{pmatrix}$ ist der geometrische Schwerpunkt $(\overline{x}, \overline{y}, \overline{z})$ gesucht. Es gilt:

$$\overline{x} = \frac{1}{O(S)} \int\limits_S x(u,v) dO, \quad \overline{y} = \frac{1}{O(S)} \int\limits_S y(u,v) dO, \quad \overline{z} = \frac{1}{O(S)} \int\limits_S z(u,v) dO$$

Wegen der Symmetrie bezüglich x und y gilt:

$$\begin{aligned}
\overline{x} &= \overline{y} = \frac{1}{O(S)} \int\limits_S x(u,v) dO = \frac{\int\limits_S x(u,v) dO}{\int\limits_S 1 dO} = \\
&= \frac{\int\limits_{\mathbb{D}} u \cdot \sqrt{1+u+v} du dv}{\int\limits_{\mathbb{D}} \sqrt{1+u+v} du dv} = \frac{\int\limits_0^1 \left(\int\limits_0^{1-u} u \cdot \sqrt{1+u+v} dv \right) du}{\int\limits_0^1 \left(\int\limits_0^{1-u} \sqrt{1+u+v} dv \right) du} \\
&= \frac{26 - 15\sqrt{2}}{14} \\
\overline{z} &= \frac{\int\limits_{\mathbb{D}} z(u,v) \cdot \sqrt{1+u+v} du dv}{\int\limits_{\mathbb{D}} \sqrt{1+u+v} du dv} = \frac{\int\limits_0^1 \left(\int\limits_0^{1-u} \frac{2}{3}\left(u^{\frac{3}{2}} + v^{\frac{3}{2}}\right) \cdot \sqrt{1+u+v} dv \right) du}{\int\limits_0^1 \left(\int\limits_0^{1-u} \sqrt{1+u+v} dv \right) du} \\
&= \frac{61\sqrt{2} - 15\ln(1+\sqrt{2})}{96(\sqrt{2}+1)},
\end{aligned}$$

denn

$$\widetilde{\xi}_u = \begin{pmatrix} 1 \\ 0 \\ u^{\frac{1}{2}} \end{pmatrix}, \quad \widetilde{\xi}_v = \begin{pmatrix} 0 \\ 1 \\ v^{\frac{1}{2}} \end{pmatrix}, \quad \widetilde{\xi}_u \times \widetilde{\xi}_v = \begin{pmatrix} -u^{\frac{1}{2}} \\ -v^{\frac{1}{2}} \\ 1 \end{pmatrix}$$

$$|\widetilde{\xi}_u \times \widetilde{\xi}_v| = \sqrt{1+u+v}$$

Oberflächenintegrale lassen sich auch über Vektorfelder definieren

Definition 6.42 (Oberflächenintegrale über Vektorfelder, Fluss)
Seien S durch eine Parameterdarstellung $\widetilde{\xi} : \mathbb{D} \to \mathbb{R}^3,\ (u,v) \mapsto \widetilde{\xi}(u,v)$, gegebenes reguläres Flächenstück mit $n = \frac{\widetilde{\xi}_u \times \widetilde{\xi}_v}{|\widetilde{\xi}_u \times \widetilde{\xi}_v|}$, und $v : \widetilde{\xi}(\mathbb{D}) \to \mathbb{R}^3$ ein auf S stetiges Vektorfeld, so heißt das Oberflächenintegral

$$\int\limits_S v dO := \int\limits_S (v^\top n) dO \tag{6.218}$$

der Fluss von v durch S.

Beispiel(e) 6.43
Der Fluss eines elektrischen Feldes $E : \mathbb{R}^3 \setminus \{0\} \to \mathbb{R}^3$, $x \mapsto \frac{c \cdot q}{|x|^3} x$ einer Punktladung q im Ursprung durch die Sphäre

$$S \subseteq \mathbb{R}^3, \quad S = \{(x, y, z) \in \mathbb{R}^3; x^2 + y^2 + z^2 = R\} \tag{6.219}$$

beträgt wegen $E^\top n = \frac{c \cdot q}{R^2}$ auf S:

$$\int_S (E^\top n) dO = \int_S \frac{c \cdot q}{R^2} dO = \frac{c \cdot q}{R^2} \int_S dO = 4\pi c q. \tag{6.220}$$

Für bestimmte, stückweise reguläre Flächen lässt sich der Satz von Green folgendermaßen verallgemeinern (Satz von Stokes):

$$\oint_{\partial S} v dx = \int_S ((\text{rot } v)^\top n) dO, \tag{6.221}$$

wobei $v : \mathbb{U} \to \mathbb{R}^3$ ein $\mathcal{C}^1(\mathbb{U}, \mathbb{R}^3)$-Vektorfeld darstellt und $\mathbb{U}$ eine offene Teilmenge des $\mathbb{R}^3$ mit $S \subseteq \mathbb{U}$ ist. Die stückweise reguläre Fläche S muss dabei die Eigenschaft besitzen, dass sich die Oberseiten der Flächenstücke S_k so wählen lassen, dass der Umlaufsinn um den Normalenvektor nicht gewechselt werden muss, wenn man über eine Kante von einem Flächenstück S_i zum benachbarten Flächenstück S_k wechselt (zweiseitige Flächenstücke). Die Oberseite eines Flächenstückes ist dadurch gegeben, dass durch eine einheitliche Wahl $n := \pm \frac{\widetilde{\xi}_u \times \widetilde{\xi}_v}{|\widetilde{\xi}_u \times \widetilde{\xi}_v|}$ der Normalenvektoren n aus der Oberfläche herauszeigt. Analog zum Satz von Green muss im Satz von Stokes die geschlossene Randkurve ∂S so durchlaufen werden, dass S „links“ liegt und der Umlaufsinn zusammen mit n eine Rechtsschraubung ergibt.

Beispiel(e) 6.44
Ist S ein ebener Bereich in der (x, y)-Ebene und

$$v : \mathbb{U} \to \mathbb{R}^3, \quad (v_1, v_2, v_3) \mapsto \begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix}, \tag{6.222}$$

so gilt:

$$\oint_{\partial S} v ds = \int_S \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right) dO \quad \text{(Satz von Green)} \tag{6.223}$$

Analog zur Definition des Integrals über ebene Bereiche kann man auch Integrale über Integrationsgebiete $\mathbb{H} \subseteq \mathbb{R}^3$ betrachten. Dazu muss zunächst einer beschränkten Teilmenge $\mathbb{H}$ des $\mathbb{R}^3$ eine Volumen zugeordnet werden. Für $x = n \cdot 2^{-k}$, $y = n \cdot 2^{-k}$, $z = n \cdot 2^{-k}$, $(n \in \mathbb{Z})$ erhalten wir Würfel im $\mathbb{R}^3$ mit Volumen 2^{-3k}. Für $\mathbb{H}$ bezeichne $s_k(\mathbb{H})$ bzw. $S_k(\mathbb{H})$ das Gesamtvolumen aller ganz in $\mathbb{H}$ enthaltenen Würfel bzw. aller Würfel, die mit $\mathbb{H}$ wenigstens einen Punkt gemeinsam haben. Man nennt $\mathbb{H}$ Riemann-messbar und $V(\mathbb{H})$ das Volumen von $\mathbb{H}$, falls

$$V(\mathbb{H}) := \lim_{k \to \infty} s_k(\mathbb{H}) = \lim_{k \to \infty} S_k(\mathbb{H}). \tag{6.224}$$

Für die Anwendungen genügt die Betrachtung der folgenden Teilmengen $\mathbb{B}$ des $\mathbb{R}^3$ (reguläre Bereiche):

(a): Der Rand $\partial\mathbb{B}$ (die „Oberfläche“ von $\mathbb{B}$) besteht aus endlich vielen stückweise regulären Flächen

(b): $\mathbb{B} \setminus \partial\mathbb{B}$ (das Innere von $\mathbb{B}$) ist ein nicht leeres, beschränktes Gebiet im $\mathbb{R}^3$ (offen und zusammenhängend)

(c): $\mathbb{B}$ ist abgeschlossen, d.h. $\partial\mathbb{B} \subset \mathbb{B}$.

Jeder reguläre Bereich $\mathbb{B}$ ist Riemann-messbar mit Volumen $V(\mathbb{B})$.
Völlig analog zur Definition eines Integrals über eine Teilmenge des $\mathbb{R}^2$ betrachten wir nun für jedes $n \in \mathbb{N}$ eine Zerlegung $\mathbb{B}_1, \ldots, \mathbb{B}_n$ eines regulären Bereichs $\mathbb{B} \subset \mathbb{R}^3$ in reguläre Teilbereiche $\mathbb{B}_1, \ldots, \mathbb{B}_n \subset \mathbb{R}^3$ mit den Volumina $V(\mathbb{B}_1), \ldots, V(\mathbb{B}_n)$. Für ein auf $\mathbb{B}$ stetiges Skalarenfeld $f : \mathbb{B} \to \mathbb{R}$ definiert man

$$\int\limits_{\mathbb{B}} f dV := \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*, y_i^*, z_i^*) V(\mathbb{B}_i) \tag{6.225}$$

mit $(x_i^*, y_i^*, z_i^*) \in \mathbb{B}_i$, wobei die Zerlegungen so zu wählen sind, dass die Volumina der regulären Teilbereiche gegen Null konvergieren.
Für

$$\mathbb{B} = \{(x, y, z) \in \mathbb{R}^3;\ a \le x \le a',\ b \le y \le b',\ c \le z \le c'\} \tag{6.226}$$

gilt:

$$\begin{aligned}
\int\limits_{\mathbb{B}} f dV &= \int\limits_a^{a'} \left(\int\limits_b^{b'} \left(\int\limits_c^{c'} f(x,y,z) dz \right) dy \right) dx = \\
&= \int\limits_b^{b'} \left(\int\limits_a^{a'} \left(\int\limits_c^{c'} f(x,y,z) dz \right) dx \right) dy = \\
&= \ldots \\
&= \int\limits_c^{c'} \left(\int\limits_b^{b'} \left(\int\limits_a^{a'} f(x,y,z) dx \right) dy \right) dz
\end{aligned} \tag{6.227}$$

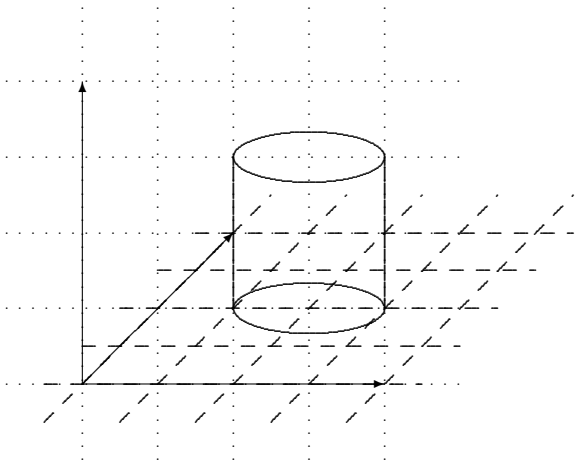
Für reguläre Bereiche der Form

$$\mathbb{B} = \left\{(x, y, z) \in \mathbb{R}^3,\ a \le x \le b,\ u(x) \le y \le v(x),\ g(x, y) \le z \le h(x, y)\right\} \tag{6.228}$$

gilt:

$$\int\limits_{\mathbb{B}} f dV = \int\limits_a^b \left(\int\limits_{u(x)}^{v(x)} \left(\int\limits_{u(x,y)}^{v(x,y)} f(x,y,z) dz \right) dy \right) dx. \tag{6.229}$$

Wählt man die Zerlegungen $\mathbb{B}_1, \ldots, \mathbb{B}_n$ von $\mathbb{B}$ durch die Koordinatenebenen,



(6.230)

so schreibt man $\int\limits_{\mathbb{B}} f dxdydz$ für $\int\limits_{\mathbb{B}} f dV$.
Offensichtlich ist das eben beschriebene Volumenintegral linear, monoton, additiv und es gilt der Mittelwertsatz: Es existiert ein $(x^*, y^*, z^*) \in \mathbb{B}$ mit:

$$f(x^*, y^*, z^*) = \frac{1}{V(\mathbb{B})} \int\limits_{\mathbb{B}} f dV \tag{6.231}$$

Anwendungen:

- $V(\mathbb{B}) = \int\limits_{\mathbb{B}} dV$
- geometrischer Schwerpunkt $(\overline{x}, \overline{y}, \overline{z})$:

$$\overline{x} = \frac{1}{V(\mathbb{B})} \int\limits_{\mathbb{B}} x dV, \quad \overline{y} = \frac{1}{V(\mathbb{B})} \int\limits_{\mathbb{B}} y dV, \quad \overline{z} = \frac{1}{V(\mathbb{B})} \int\limits_{\mathbb{B}} z dV. \tag{6.232}$$

Seien nun $\mathbb{B}, \mathbb{U} \subset \mathbb{R}^3$ reguläre Bereiche, $f : \mathbb{B} \to \mathbb{R}$ ein stetiges Skalarenfeld und

$$\zeta : \mathbb{U} \to \mathbb{B}, \quad (u, v, w) \mapsto (x(u, v, w), y(u, v, w), z(u, v, w))^\top \tag{6.233}$$

mit:

(a): ζ ist bijektiv und $\mathcal{C}^1(\mathbb{U}, \mathbb{B})$

(b): Für jeden Vektor $(u, v, w) \in \mathbb{U}$ gilt:

$$\frac{\partial \zeta}{\partial u}(u, v, w), \quad \frac{\partial \zeta}{\partial v}(u, v, w), \quad \frac{\partial \zeta}{\partial w}(u, v, w) \quad \text{linear unabhängig.} \tag{6.234}$$

so gilt **der Transformationssatz**:

$$\begin{aligned} \int\limits_{\mathbb{B}} f(x, y, z) dxdydz &= \int\limits_{\mathbb{U}} f(x(u, v, w), y(u, v, w), z(u, v, w)) \cdot \\ &\cdot \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{pmatrix} \right| dudvdw. \end{aligned} \tag{6.235}$$

Eine analoge Formel gilt auch im zweidimensionalen Fall mit regulären Bereichen $\mathbb{B}, \mathbb{U} \in \mathbb{R}^2$, $f : \mathbb{B} \to \mathbb{R}$ und

$$\zeta : \mathbb{U} \to \mathbb{B}, \quad (u, v) \mapsto (x(u, v), y(u, v))^\top \tag{6.236}$$

und analogen Bedingungen an x:

$$\int\limits_{\mathbb{B}} f(x, y) dxdy = \int\limits_{\mathbb{U}} f(x(u, v), y(u, v)) \cdot \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \right| dudv. \tag{6.237}$$

Beispiel(e) 6.45

- Affine Koordinaten:

$$\mathbb{R}^3 \to \mathbb{R}^3, \quad (u,v,w) \mapsto \begin{pmatrix} x(u,v,w) \\ y(u,v,w) \\ z(u,v,w) \end{pmatrix} = A \begin{pmatrix} u \\ v \\ w \end{pmatrix} + x_0 \tag{6.238}$$

mit $A \in \mathbb{R}^{3\times 3}$, A regulär, $x_0 \in \mathbb{R}^3$.

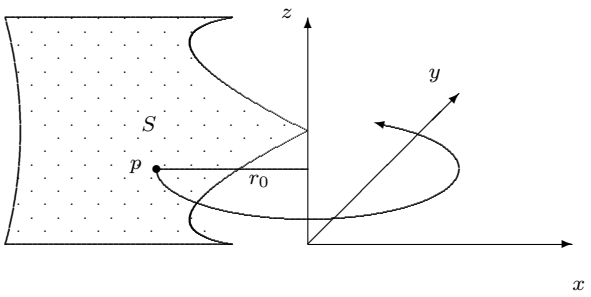
$$\int\limits_{\mathbb{B}} f dV = \int\limits_{\mathbb{U}} f \left(A \begin{pmatrix} u \\ v \\ w \end{pmatrix} + x_0 \right) \cdot |\det A| du dv dw. \tag{6.239}$$

- Kugelkoordinaten:

$$[0,\infty) \times [0,2\pi] \times [0,\pi] \to \mathbb{R}^3, \quad (r,\varphi,\vartheta) \mapsto \begin{pmatrix} x(r,\varphi,\vartheta) \\ y(r,\varphi,\vartheta) \\ z(r,\varphi,\vartheta) \end{pmatrix} = \begin{pmatrix} r \cdot \sin(\vartheta)\cos(\varphi) \\ r \cdot \sin(\vartheta)\sin(\varphi) \\ r \cdot \cos(\vartheta) \end{pmatrix}. \tag{6.240}$$

$$\int\limits_{\mathbb{B}} f dV = \int\limits_{\mathbb{U}} f(x(r,\varphi,\vartheta), y(r,\varphi,\vartheta), z(r,\varphi,\vartheta)) \cdot r^2 \sin(\vartheta) dr d\varphi d\vartheta. \tag{6.241}$$

- Regel von Guldin:
  Sei S eine Teilmenge der (x,z)-Ebene mit Fläche $F(S)$ und Schwerpunkt P. Die Rotation dieser Fläche um die z-Achse mit Abstand r_0 zu P ergibt ein dreidimensionales Gebilde mit Volumen $V = 2\pi r_0 F$.



(6.242)

Zum Abschluss der Integrationstheorie betrachten wir eine wichtige Beziehung zwischen einem Volumenintegral und einem Oberflächenintegral.

Satz 6.46 (Divergenzsatz von Gauss)
Sei $\mathbb{B}$ ein regulärer räumlicher Bereich mit Oberfläche $\partial\mathbb{B}$, die aus endlich vielen, stückweise regulären zweiseitigen Flächen besteht, die sich höchstens in Randpunkten berühren. Sei ferner n die aus allen regulären Oberflächenstücken von $\mathbb{B}$ nach außen weisende Einheitsnormale, dann gilt für alle auf $\mathbb{U} \subseteq \mathbb{R}^3$ definierten $\mathcal{C}^1(\mathbb{B},\mathbb{U})$-Vektorfelder v mit $\mathbb{B} \subset \mathbb{U}$, $\mathbb{U}$ offen:

$$\int\limits_{\mathbb{B}} (\text{div } v) dV = \int\limits_{\partial\mathbb{B}} (v^\top n) dO. \tag{6.243}$$

Beispiel(e) 6.47
$\mathbb{B} = \{(x,y,z) \in \mathbb{R}^3;\ x^2+y^2+z^2 \leq r^2\}$, $v : \mathbb{R}^3 \to \mathbb{R}^3$, $(x,y,z) \mapsto \begin{pmatrix} 2z \\ x+y \\ 0 \end{pmatrix}$

$$\int\limits_{\partial\mathbb{B}} (v^\top n) dO = \int\limits_{\mathbb{B}} (\mathsf{div}\ v) dV = \int\limits_{\mathbb{B}} dV = \frac{4}{3} r^3 \pi. \tag{6.244}$$

Für $v : \mathbb{U} \to \mathbb{R}^3$ mit $(\mathsf{div}\ v) = 1$ gilt:

$$V(\mathbb{B}) = \int\limits_{\mathbb{B}} dV = \int\limits_{\partial\mathbb{B}} (v^\top n) dO. \tag{6.245}$$

Kapitel 7

Funktionentheorie

Funktionentheorie ist die „Theorie von Funktionen einer komplexen Variablen“, also

$$f : D \to \mathbb{C} \quad \text{mit} \quad D \subseteq \mathbb{C}.$$

Im ersten Kapitel haben wir die Menge $\mathbb{C}$ mit der Menge $\mathbb{R}^2$ identifiziert:

$$\mathbb{C} = \mathbb{R}^2, \qquad \mathbb{C} \ni z = x + iy = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$$

mit der imaginären Einheit i, dem Realteil $\mathrm{Re}(z) = x$ und dem Imaginärteil $\mathrm{Im}(z) = y$ von z. Die Identifikation $\mathbb{C} = \mathbb{R}^2$ ist zulässig, solange man jedes $z \in \mathbb{C}$ lediglich geometrisch als einen Punkt in der Zahlenebene auffasst.

Mit der Identifikation $\mathbb{C} = \mathbb{R}^2$ kann man jede Teilmenge $D \subseteq \mathbb{C}$ auch als eine Teilmenge $D \subseteq \mathbb{R}^2$ auffassen. Dann entspricht eine Funktion

$$f : D \to \mathbb{C}, \quad z = x + iy \mapsto f(z) = \mathrm{Re}(f(x+iy)) + i\mathrm{Im}(f(x+iy))$$

gerade einer Funktion

$$F : D \to \mathbb{R}^2, \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix}$$

mit $u(x,y) = \mathrm{Re}(f(x+iy))$ und $v(x,y) = \mathrm{Im}(f(x+iy))$. Es fragt sich, ob es überhaupt nötig ist, Funktionen einer komplexen Variablen einzuführen oder ob man nicht gemäß der Identifikation von f mit F ganz bei reellen Funktionen bleiben kann.

Nun kann man mit komplexen Zahlen auch rechnen, das heißt man kann der Menge $\mathbb{C}$ eine algebraische Struktur geben. Das tut man auch mit $\mathbb{R}^2$: dies ist ein Vektorraum mit Vektoraddition und Skalarmultiplikation

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \end{pmatrix}, \quad \lambda \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix}, \quad \lambda \in \mathbb{R}\,.$$

$\mathbb{R}^2$ ist darüber hinaus ein normierter Vektorraum:

$$\left| \begin{pmatrix} x \\ y \end{pmatrix} \right| = \sqrt{x^2 + y^2}\,.$$

Ganz analog wird auch in $\mathbb{C}$ addiert und mit Skalaren $\lambda \in \mathbb{R}$ multipliziert

$$(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2), \quad \lambda(x + iy) = (\lambda x) + i(\lambda y)\,.$$

Auch einen Betrag komplexer Zahlen definiert man in Analogie zur oben angegebenen (Euklidischen) Norm auf $\mathbb{R}^2$:

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

Auch als Vektorräume lassen sich $\mathbb{C}$ und $\mathbb{R}^2$ somit identifizieren.

Einen wichtigen Unterschied gibt es jedoch: in Vektorräumen ist im allgemeinen nicht die Multiplikation und die Division zweier Vektoren definiert. So auch nicht in $\mathbb{R}^2$. In $\mathbb{C}$ hingegen haben wir sehr wohl eine Multiplikation, wie im ersten Kapitel definiert:

$$(x+iy)(u+iv) = (xu-yv)+i(xv+yu)$$

und ebenso eine Division

$$\frac{x+iy}{u+iv} = \frac{xu+yv}{u^2+v^2} + i\frac{yu-xv}{u^2+v^2} \quad \text{für} \quad u+iv \neq 0\,.$$

Da eine Division existiert, kann man auch Differenzenquotienten komplexer Funktionen und damit eine „komplexe Differenzierbarkeit“ erklären. Es wird sich zeigen, dass die komplexe Differenzierbarkeit eine sehr viel weiter führende Eigenschaft als die reelle Differenzierbarkeit ist.

7.1 Folgen, Stetigkeit und Holomorphie

Eine Folge $\{z_n\}$, $n \in \mathbb{N}_0$, komplexer Zahlen heißt konvergent gegen $a \in \mathbb{C}$, in Zeichen: $\lim\limits_{n\to\infty} z_n = a$, falls es zu jedem $\epsilon > 0$ ein $n_0 \in \mathbb{N}$ gibt mit:

$$|z_n - a| < \epsilon \quad \text{für alle } n \geq n_0. \tag{7.1}$$

Konvergiert eine komplexe Folge $\{z_n\}$ gegen a, so konvergieren die reellen Folgen $\{\mathrm{Re}(z_n)\}$ bzw. $\{\mathrm{Im}(z_n)\}$, $n \in \mathbb{N}_0$, gegen $\mathrm{Re}(a)$ bzw. gegen $\mathrm{Im}(a)$. Eine Funktion $f : \mathbb{C} \to \mathbb{C}$ heißt in $z_0 \in \mathbb{C}$ stetig, falls

$$\lim_{z\to z_0} f(z) = f(z_0), \tag{7.2}$$

wobei wie im Reellen $\lim\limits_{z\to z_0} f(z) = f(z_0)$ genau dann, wenn für jede Folge $\{z_n\}$, $n \in \mathbb{N}_0$, die gegen z_0 konvergiert, gilt:

$$\lim_{n\to\infty} f(z_n) = f(z_0). \tag{7.3}$$

Mit dem Grenzwertbegriff lässt sich analog zum Reellen auch der Differenzierbarkeitsbegriff einführen.

Definition 7.1 ((komplex) differenzierbar, Ableitung, holomorph)
Sei $\mathbb{G} \subseteq \mathbb{C}$ ein Gebiet (also offen und zusammenhängend). Eine Funktion

$$f : \mathbb{G} \to \mathbb{C} \tag{7.4}$$

heißt in z_0 (komplex) differenzierbar, falls

$$\lim_{z\to z_0} \frac{f(z)-f(z_0)}{z-z_0} \tag{7.5}$$

existiert. Dieser Grenzwert wird als Ableitung von f an der Stelle $z_0 \in \mathbb{G}$ bezeichnet und mit $f'(z_0)$ notiert. Ist f an jeder Stelle $z_0 \in \mathbb{G}$ (komplex) differenzierbar, so heißt f auf $\mathbb{G}$ **holomorph**.

So wie oben betrachten wir $\mathbb{G}$ sowohl als Teilmenge von $\mathbb{C}$ als auch als Teilmenge von $\mathbb{R}^2$ und zerlegen die Funktion $f : \mathbb{G} \to \mathbb{C}$ in Real- und Imaginärteil:

$$f(x+iy) = u(x,y) + iv(x,y) \tag{7.6}$$

mit Funktionen

$$u, v : \mathbb{G} \to \mathbb{R}, \quad (x, y) \mapsto u(x, y) \quad \text{bzw. } v(x, y) \; .$$

Die Frage ist nun, was die komplexe Differenzierbarkeit von f mit der Differenzierbarkeit von u und v zu tun hat. Dazu sei an folgende Begriffe aus dem vierten Kapitel erinnert:

- u heißt in $(x_0, y_0) \in \mathbb{G}$ **(stetig) partiell differenzierbar nach** x, wenn die Funktion $x \mapsto u(x, y_0)$ bei festgehaltenem y_0 (stetig) differenzierbar in x_0 ist. Man schreibt $\partial u/\partial x(x_0, y_0)$ oder $u_x(x_0, y_0)$ für die entsprechende Ableitung, die man **partielle Ableitung von** u **nach** x nennt. Vergleiche hierzu die Formeln (6.26)-(6.28).

- u heißt in $(x_0, y_0) \in \mathbb{G}$ **(stetig) partiell differenzierbar** wenn sowohl $x \mapsto u(x, y_0)$ als auch $y \mapsto u(x_0, y)$ in x_0 bzw. y_0 (stetig) differenzierbar sind. Man fasst dann beide partiellen Ableitungen als Vektor zusammen und nennt

  $$\nabla u(x_0, y_0) = \text{grad } u(x_0, y_0) = \begin{pmatrix} \dfrac{\partial u}{\partial x}(x_0, y_0) \\ \dfrac{\partial u}{\partial y}(x_0, y_0) \end{pmatrix} = \begin{pmatrix} u_x(x_0, y_0) \\ u_y(x_0, y_0) \end{pmatrix}$$

  den Gradienten von u in (x_0, y_0). Vergleiche hierzu (6.29).

- Man nennt u **(stetig) partiell differenzierbar auf** $\mathbb{G}$, wenn u in jedem Punkt $(x, y) \in \mathbb{G}$ (stetig) partiell differenzierbar ist. Die Menge aller stetig partiell differenzierbaren Funktionen von $\mathbb{G}$ nach $\mathbb{R}$ bezeichnet man mit $\mathcal{C}^1(\mathbb{G}, \mathbb{R})$.

- Man nennt u **differenzierbar** (oder auch: **total differenzierbar**) in $(x_0, y_0) \in \mathbb{G}$, wenn u in (x_0, y_0) linear approximierbar ist im Sinn von

  $$u(x, y) = u(x_0, y_0) + a^\top \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} + o\left(\left| \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \right| \right)$$

  für einen Vektor $a \in \mathbb{R}^2$. Vergleiche hierzu die Definition 6.11 und zur Bedeutung des Symbols o insbesondere auch (6.31)-(6.33).

- Ist u in $(x_0, y_0) \in \mathbb{G}$ differenzierbar, dann ist u dort auch partiell differenzierbar und es muss

  $$a = \text{grad } u(x_0, y_0)$$

  gelten. Umgekehrt folgt aus der partiellen aber nicht die totale Differenzierbarkeit.

- Ist u in $(x_0, y_0) \in \mathbb{G}$ *stetig* partiell differenzierbar, dann ist u dort auch total differenzierbar. Die Umkehrung hiervon gilt nicht.

Alles gerade über u Gesagte gilt selbstverständlich ebenso für v.

Wir betrachten nun den Grenzwert (7.5) unter Benutzung von (7.6). Setzt man voraus, dass f in z_0 komplex differenzierbar ist, dann muss der Grenzwert *für jede* Folge $z \to z_0$ existieren. So können wir einerseits $z_0 = x_0 + iy_0$ und $z = x_0 + h + iy_0$ für reelles h betrachten:

$$\begin{aligned} \frac{f(z) - f(z_0)}{z - z_0} &= \frac{u(x_0 + h, y_0) - u(x_0, y_0) + i(v(x_0 + h, y_0) - v(x_0, y_0))}{h} \\ &\longrightarrow u_x(x_0, y_0) + iv_x(x_0, y_0) \quad \text{für } h \longrightarrow 0 \; . \end{aligned}$$

Genauso gut können wir $z = x_0 + i(y_0 + h)$ wiederum für reelles h betrachten. Damit ergibt sich

$$\begin{aligned} \frac{f(z) - f(z_0)}{z - z_0} &= \frac{u(x_0, y_0 + h) - u(x_0, y_0) + i(v(x_0, y_0 + h) - v(x_0, y_0))}{ih} \\ &\longrightarrow v_y(x_0, y_0) - iu_y(x_0, y_0) \quad \text{für } h \longrightarrow 0 \; . \end{aligned}$$

Da der Grenzwert in beiden Fällen gleich sein muss (es war die komplexe Differenzierbarkeit vorausgesetzt), lässt sich aus dem Vergleich von Real- und Imaginärteil folgern, dass im Punkt (x_0, y_0) die sogenannten **Cauchy-Riemannschen-Differentialgleichungen**

$$u_x = v_y \qquad \text{und} \qquad v_x = -u_y$$

gelten müssen. Hiervon gilt sogar die Umkehrung:

Satz 7.2 (Cauchy-Riemannsche Differentialgleichungen)
Es sei $\mathbb{G} \subseteq \mathbb{C}$ ein Gebiet. Die Funktion $f : \mathbb{G} \to \mathbb{C}$, $f(x+iy) = u(x,y) + iv(x,y)$ ist holomorph

$$\iff \quad u \text{ und } v \text{ total differenzierbar und}$$

$$u_x \;=\; v_y, \qquad v_x \;=\; -u_y$$

In diesem Fall ist

$$f'(z) = u_x(x,y) + iv_x(x,y) = v_y(x,y) - iu_y(x,y)\ .$$

Beweis. Es sei $z_0 \in \mathbb{G}$ fest gewählt und für beliebiges $z \in \mathbb{G}$ sei $z - z_0 = \Delta x + i\Delta y$.

„$\Longrightarrow$“: Nach Voraussetzung existiert $f'(z) = a + ib$ und es ist

$$\frac{f(z) - f(z_0)}{z - z_0} - (a + ib) \to 0 \quad \text{für } z - z_0 \to 0\ .$$

Also ist nach Multiplikation mit $z - z_0$

$$f(z) = f(z_0) + (a + ib)(\Delta x + i\Delta y) + o(\sqrt{(\Delta x)^2 + (\Delta y)^2})\ .$$

Mit dem **Landau-Symbol** o wird zum Ausdruck gebracht, dass $o(\sqrt{(\Delta x)^2 + (\Delta y)^2})$ ein Ausdruck ist, der schneller als $\sqrt{(\Delta x)^2 + (\Delta y)^2}$ und damit insbesondere schneller als Δx und schneller als Δy gegen Null geht.

Wir wissen nach der Vorbemerkung zum Satz bereits, dass $a = u_x = v_y$ und ebenso, dass $b = v_x = -u_y$, so dass wir die letzte Gleichung wie folgt getrennt nach Real- und Imaginärteil schreiben können

$$\begin{aligned} u(x,y) &= u(x_0,y_0) + a\Delta x - b\Delta y + o(\ldots) \\ &= u(x_0,y_0) + u_x(x_0,y_0)\Delta x + u_y(x_0,y_0)\Delta y + o(\ldots) \\ v(x,y) &= v(x_0,y_0) + b\Delta x + a\Delta y + o(\ldots) \\ &= v(x_0,y_0) + v_x(x_0,y_0)\Delta x + v_y(x_0,y_0)\Delta y + o(\ldots) \end{aligned}$$

Das ist gerade die totale Differenzierbarkeit von u und v im Punkt (x_0, y_0) im Sinn der linearen Approximierbarkeit gemäß Definition 6.11.

„$\Longleftarrow$“: Man kann alle oben vollzogenen Umformungsschritte auch in der anderen Richtung von unten nach oben lesen. q.e.d

Beispiel(e) 7.3
Für die Funktion

$$f : \mathbb{C} \to \mathbb{C}, \quad z \mapsto \bar{z} = x - iy$$

ist $u(x, y) = x$ und $v(x, y) = -y$. Offenbar sind u und v beide auf ganz $\mathbb{R}^2$ stetig partiell und damit total differenzierbar. Jedoch ist

$$u_x \equiv 1 \neq -1 \equiv v_y \ ,$$

das heißt f ist nirgends komplex differenzierbar.

Der Satz 7.2 liefert neben der Definition 7.1 eine zweite Möglichkeit nachzuprüfen, ob eine Funktion komplex differenzierbar ist. Daneben stehen folgende Rechenregeln für holomorphe Funktionen $f, g : \mathbb{G} \to \mathbb{C}$ zur Verfügung:

(a):

$$f(z) \equiv c \quad \Longrightarrow \quad f'(z) \equiv 0, \tag{7.7}$$

(b):

$$f(z) = z \quad \Longrightarrow \quad f'(z) \equiv 1, \tag{7.8}$$

(c):

$$(f \pm g)' = f' \pm g', \tag{7.9}$$

(d):

$$(f \cdot g)' = f' \cdot g + f \cdot g', \tag{7.10}$$

(e):

$$\left(\frac{f}{g}\right)' = \frac{f'g - g'f}{g^2}, \quad g(z) \neq 0, \tag{7.11}$$

(f): Sei $h : f(\mathbb{G}) \to \mathbb{C}$ holomorph, so gilt:

$$(h \circ f)'(z) = h'(f(z))f'(z). \tag{7.12}$$

Die Cauchy-Riemannschen DGL zeigen, dass der Realteil und der Imaginärteil einer komplex differenzierbaren Funktion ganz stark zusammenhängen. Kennt man den einen, dann steht auch der andere bis auf eine additive Konstante fest, wie wir gleich am folgenden Beispiel sehen werden. Selbst für sich allein genommen unterliegen der Realteil (und ebenso der Imaginärteil) einer holomorphen Funktion einer sehr starken Einschränkung. Wenn etwa u zweimal stetig differenzierbar ist, dann gilt aufgrund der Cauchy-Riemannschen DGL und aufgrund des Satzes 6.10 über die Vertauschbarkeit partieller Ableitungen, dass

$$u_{xx} = (v_y)_x = v_{yx} = v_{xy} = -u_{yy} \quad \Longleftrightarrow \quad u_{xx} + u_{yy} = 0$$

und ebenso gilt für den Imaginärteil v einer holomorphen Funktion, dass

$$v_{xx} + v_{yy} = 0 \ .$$

Man nennt eine Funktion $\varphi : \mathbb{G} \to \mathbb{R}$ mit $\varphi_{xx} + \varphi_{yy} = 0$ **harmonisch**. Es lässt sich zeigen, dass jede harmonische Funktion Realteil oder Imaginärteil einer holomorphen Funktion ist.

Beispiel(e) 7.4
Gegeben sei die Funktion

$$u : \mathbb{R}^2 \to \mathbb{R}, \quad (x, y) \mapsto x^2 - y^2 \ .$$

Es ist zu untersuchen, ob u Realteil einer holomorphen Funktion ist. Gegebenenfalls sind der zugehörige Imaginärteil v sowie eine geeignete holomorphe Funktion $z \mapsto f(z)$ anzugeben.

Man stellt zunächst fest, dass $u_{xx} \equiv 2$ und $u_{yy} \equiv -2$, so dass in der Tat $u_{xx} + u_{yy} = 0$. Also ist u Realteil einer holomorphen Funktion.

Für den unbekannten Imaginärteil $v = v(x, y)$ von f müssen die Cauchy-Riemannschen DGL gelten, also

$$u_x = 2x \stackrel{!}{=} v_y \quad \Rightarrow \quad v(x, y) = 2xy + c(x) \ ,$$

wobei $c = c(x)$ eine nur von x abhängige Funktion ist. Ableiten nach x liefert

$$v_x = 2y + c'(x) \stackrel{!}{=} -u_y = 2y \ .$$

Also ist

$$c'(x) = 0 \quad \Rightarrow \quad c(x) = C = \text{ const.}$$

und wir haben für die Wahl $C = 0$

$$f(x + iy) = x^2 - y^2 + 2ixy \ . \tag{7.13}$$

Will man diese Funktion in der komplexen Variable $z = x + iy$ schreiben, kann man folgenden Trick benutzen. Wenn eine „Formel in z“ für $f(z)$ existiert, muss diese insbesondere für $z = x+i0$ gelten. Setzt man $y = 0$ in (7.13), dann ergibt sich

$$f(x + i0) = x^2 = (x + i0)^2$$

und daraus folgert man, dass $f(z) = z^2$.

7.2 Integration

A. Das unbestimmte Integral.

Genau wie im Reellen wird das unbestimmte Integral – die **Stammfunktion** von f – definiert gemäß

$$F(z) = \int f(z)\, dz \quad :\Longleftrightarrow \quad F'(z) = f(z) \ .$$

$F(z)$ ist wie im Reellen eindeutig bestimmt bis auf die Addition einer komplexen Konstanten C.

Wir haben auch die gleichen Rechenregeln wie im Reellen:

- $\int f'(z)dz = f(z)+C$, denn nach Definition des unbestimmten Integrals ist $\left(\int f'(z)\, dz \right)' = f'(z)$.

- Man darf partiell integrieren:

$$\int g(z)h'(z)\,dz = g(z)h(z) - \int g'(z)h(z)\,dz + C\,,$$

  denn beim Differenzieren gilt die Produktregel.

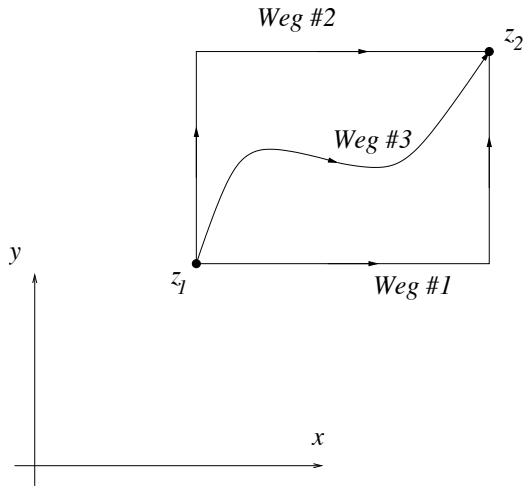
- Man darf im Integral substituieren:

$$\int f(z)\,dz = \int f(g(\xi))g'(\xi)\,d\xi\,,$$

  denn beim Differenzieren gilt die Kettenregel.

B. Integral längs Kurven

Es handelt sich um das Analogon zu den bestimmten Integralen im Reellen: von einem Punkt $z_1 \in \mathbb{C}$ wird zu einem Punkt $z_2 \in \mathbb{C}$ integriert. Allerdings gibt es jetzt die Möglichkeit, auf verschiedenen „**Integrationswegen**“ von z_1 nach z_2 zu kommen:



Integrationswege in $\mathbb{C}$ werden mathematisch beschrieben durch stetig differenzierbare Abbildungen

$$\gamma : [a, b] \to \mathbb{C}\,,$$

sogenannte **Parametrisierungen** für das glatte Kurvenstück

$$K := \{\gamma(t) = x(t) + iy(t);\ a \le t \le b\}\,.$$

Mit der geometrischen Gleichsetzung $\mathbb{C} = \mathbb{R}^2$ entspricht das gerade den in der reellen Analysis behandelten ebenen Kurven(stücken):

$$\gamma(t) = x(t) + iy(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}\,.$$

Die stetige Differenzierbarkeit von γ bedeutet, dass die Funktionen $x = x(t)$ und $y = y(t)$ stetig differenzierbar sein müssen. Für die Ableitung ist dann

$$\dot\gamma(t) = \begin{pmatrix} \dot x(t) \\ \dot y(t) \end{pmatrix} = \dot x(t) + i\dot y(t)\,.$$

Durch die Parametrisierung wird nicht nur eine Punktmenge in $\mathbb{C}$ beschrieben, vielmehr wird der Kurve K noch eine **Orientierung** im Sinn einer Durchlaufrichtung gegeben: die Kurve „beginnt" bei $\gamma(a)$ und „endet" bei $\gamma(b)$. In der obigen Skizze ist diese Durchlaufrichtung durch Pfeile gekennzeichnet.

Beispiel(e) 7.5
Durch die Abbildung

$$\gamma : [0, 2\pi] \to \mathbb{C}, \quad \gamma(t) = a + re^{it},$$

wird ein Kreis mit Radius r um $a \in \mathbb{C}$ parametrisiert. Der Kreis wird im mathematisch positiven Sinn durchlaufen. Durch $\hat{\gamma} : [0, \pi] \to \mathbb{C}$, definiert durch $\hat{\gamma}(t) = a + re^{2it}$, wird diesselbe Kurve beschrieben, die allerdings anders parametrisiert wird. Die Parametrisierung $\tilde{\gamma} : [0, 2\pi] \to \mathbb{C}$, gegeben durch $\tilde{\gamma}(t) = a + re^{-it}$, beschreibt dieselbe Punktmenge, aber nicht dieselbe Kurve: der Durchlaufsinn hat sich geändert.

Definition 7.6 (Integral über f längs K)
Es sei $\mathbb{G} \subseteq \mathbb{C}$ ein Gebiet, $f : \mathbb{G} \to \mathbb{C}$ stetig und K ein glattes Kurvenstück, das durch $\gamma : [a, b] \to \mathbb{C}$ parametrisiert wird. Dann definiert man

$$\int_K f(z)\, dz := \int_a^b f(\gamma(t))\dot{\gamma}(t)\, dt \tag{7.14}$$

Mit $z = x + iy$, $f(z) = u(x, y) + iv(x, y)$ und $\gamma(t) = x(t) + iy(t)$, also $\dot{\gamma}(t) = \dot{x}(t) + i\dot{y}(t)$, lässt sich die rechte Seite als Summe zweier bestimmter reeller Integrale schreiben:

$$\begin{aligned}\int_a^b &[u(x(t), y(t)) + iv(x(t), y(t))] \cdot [\dot{x}(t) + i\dot{y}(t)]\ dt \\ &= \int_a^b [u\dot{x} - v\dot{y}]\ dt + i \int_a^b [v\dot{x} + u\dot{y}]\ dt\end{aligned}$$

Hinweise.

- Das Kurvenstück K heißt in diesem Zusammenhang der **Integrationsweg**
- Im Komplexen hat das Kurvenintegral immer zwei Anteile, den Real- und den Imaginärteil
- Das Resultat ist **unabhängig von der gewählten Parametrisierung des glatten Kurvenstücks, sofern der Durchlaufsinn nicht geändert wird**:

  Bei einer Umparametrisierung $\varphi : [a', b'] \to [a, b]$, $t = \varphi(\tau)$, mit monoton steigendem φ ändert sich das Integral nicht:

$$\begin{aligned}\int_a^b f(\gamma(t))\dot{\gamma}(t)\, dt &= \int_{a'}^{b'} f(\gamma(\varphi(\tau)))\dot{\gamma}(\varphi(\tau))\varphi'(\tau)\, d\tau \\ &= \int_{a'}^{b'} f(\psi(\tau))\psi'(\tau)\, d\tau\ .\end{aligned}$$

  Hier haben wir in der ersten Zeile die Transformationsformel für Integrale benutzt und in der zweiten Zeile die Kettenregel für die neu definierte Abbildung

$$\psi : [a', b'] \to \mathbb{C}, \quad \psi(\tau) = \gamma(\varphi(\tau)) = \gamma(t)\ ,$$

die eine andere Parametrisierung von K darstellt. Die Ableitung nach τ wird mit einem Strich gekennzeichnet, die nach t mit einem Punkt.

Weil φ monoton steigend ist, bleibt der Durchlaufsinn der Kurve K erhalten.

Ein triviales Beispiel ist der Kreis um 0 mit Radius $r > 0$. Er kann zum Beispiel durch $\gamma : [0, 2\pi] \to \mathbb{C}$, $\gamma(t) = re^{it}$, parametrisiert werden, aber ebenso durch $\psi : [0, \pi] \to \mathbb{C}$, $\psi(\tau) = re^{2i\tau}$. Die Umparametrisierung ist in diesem Fall $t = 2\tau$.

- Man kann glatte Kurvenstücke aneinanderhängen und kommt so zu einer stückweise glatten Kurve als Integrationsweg:
$$K = K_1 \cup K_2 \cup \ldots \cup K_m$$
und
$$\int_K f(z)\, dz := \int_{K_1} f(z)\, dz + \ldots + \int_{K_m} f(z)\, dz\ .$$

- Umgekehrt kann man ein Kurvenintegral $\displaystyle\int_K f(z)\, dz$ aufspalten in mehrere Teilintegrale längs glatter Kurvenstücke

- Im Sonderfall, dass eine Kurve K **geschlossen** ist, das heißt dass Anfangs- und Endpunkt der Parametrisierung gleich sind – $\gamma(a) = \gamma(b)$ – schreibt man auch
$$\oint_K f(z)\, dz \quad \text{statt} \quad \int_K f(z)\, dz$$

- Eine geschlossene, doppelpunktfreie (das meint: die Kurve schneidet bzw. berührt sich nicht selbst) Kurve K zerteilt $\mathbb{C} \setminus K$ in zwei Gebiete: das beschränkte **Innere** und das unbeschränkte **Äußere** von K. (Ein Sachverhalt, der in der Mathematik als „**Jordanscher Kurvensatz**“ bekannt ist). Das Innere wird von K **positiv umlaufen**, wenn es (in Durchlaufrichtung gesehen) links von K liegt. Wird das Innere genau einmal positiv umlaufen, dann schreibt man auch
$$\oint_K f(z)\, dz \quad \text{statt} \quad \int_K f(z)\, dz\ .$$
Ist der Integrationsweg der einmal positive umlaufene Kreis $K = \{|z - a| = r\}$, so schreibt man auch
$$\oint_{|z-a|=r} f(z)\, dz$$

Um komplexe Integrale auszurechnen, kann man die folgenden Schritte ausführen:

(a): K bzw. $K_1, \ldots, K_m$ parametrisieren, das heißt: wähle $\gamma(t) = x(t) + iy(t)$ und $a \le t \le b$ passend

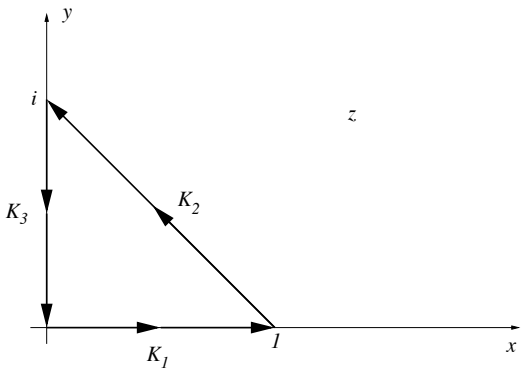
(b): $\gamma(t)$ in $f(z)$ einsetzen: $u(x(t), y(t)) + iv(x(t), y(t))$. $\dot\gamma(t) = \dot x(t) + i\dot y(t)$ berechnen

(c): Integranden $f(\gamma(t))\dot\gamma(t)$ sortieren in Real- und Imaginärteil

(d): Beide bestimmte Integrale berechnen und Beiträge akkumulieren.

Beispiel(e) 7.7

Es ist $\oint_K \bar{z}\,dz$ zu berechnen für folgenden Integrationsweg K:



Bezüglich K_1:
$\gamma(t) = t$, also $f(\gamma(t)) = t$ und $\dot{\gamma}(t) = 1$. Integrationsintervall: $0 \le t \le 1$. Folglich

$$\int_{K_1} \bar{z}\,dz = \int_0^1 t\,dt = \left[\frac{t^2}{2}\right]_0^1 = \frac{1}{2}\,.$$

Bezüglich K_2:
$\gamma(t) = 1-t+it$, also $f(\gamma(t)) = 1-(1+i)t$ und $\dot{\gamma}(t) = -1+i$. Integrationsintervall: $0 \le t \le 1$. Folglich

$$\begin{aligned}\int_{K_2} \bar{z}\,dz &= \int_0^1 [1-(1+i)t](-1+i)\,dt \\ &= \left[(-1+i)t - (1+i)(-1+i)\frac{t^2}{2}\right]_0^1 = -1+i+\frac{1-i^2}{2} = i\,.\end{aligned}$$

Bezüglich K_3:
$\gamma(t) = i-it$, also $f(\gamma(t)) = -i+it$ und $\dot{\gamma}(t) = -i$. Integrationsintervall: $0 \le t \le 1$. Folglich

$$\int_{K_3} \bar{z}\,dz = \int_0^1 (-i+it)(-i)\,dt = \left[i^2 t - i^2\frac{t^2}{2}\right]_0^1 = -\frac{1}{2}\,.$$

Alle Teilintegrale zusammen ergeben

$$\oint_K \bar{z}\,dz = i\,.$$

Beispiel(e) 7.8 (Das Fundamentalintegral)
Zu berechnen ist

$$F_m := \oint\limits_{|z-a|=r} (z-a)^m \, dz, \quad m \in \mathbb{Z} \, .$$

Mit der Parametrisierung $\gamma(t) = a + re^{it}$, $0 \le t \le 2\pi$ ergibt sich $\dot{\gamma}(t) = ire^{it}$ und daraus

$$\begin{aligned} F_m &= \int_0^{2\pi} \left(re^{it}\right)^m ire^{it} \, dt = ir^{m+1} \int_0^{2\pi} e^{i(m+1)t} \, dt \\ &= \begin{cases} 2\pi i, & \text{falls } m = -1 \\ 0, & \text{sonst} \end{cases} \end{aligned}$$

Die wichtigsten Rechenregeln für Kurvenintegrale sind die folgenden

- **Linearität**:
$$\int_K (\alpha f(z) + \beta g(z)) \, dz = \alpha \int_K f(z) \, dz + \beta \int_K g(z) \, dz$$
- **Weg-Additivität**:
$$\int_{K_1 \cup K_2} f(z) \, dz = \int_{K_1} f(z) \, dz + \int_{K_2} f(z) \, dz$$
- **Orientierung**: Zu gegebenem K sei K_* die in entgegengesetzter Richtung durchlaufene Kurve. Wird K durch $\gamma : [a,b] \to \mathbb{C}$ parametrisiert, so K_* durch $\gamma_* : [a,b] \to \mathbb{C}$, $\gamma_*(t) = \gamma(a+b-t)$. Es ist
$$\int_{K_*} f(z) \, dz = -\int_K f(z) \, dz$$
- **Integralabschätzung**:
$$\left| \int_K f(z) \, dz \right| \le \max_{z \in K} |f(z)| \cdot \text{ Länge}(K)$$
Die Länge der durch $\gamma(t) = x(t) + iy(t)$ mit $a \le t \le b$ parametrisierten Kurve K ergibt sich genauso wie die der entsprechenden reellen ebenen i Kurve zu
$$\text{Länge}(K) = \int_a^b \sqrt{\dot{x}^2(t) + \dot{y}^2(t)} \, dt \, .$$

C. Integralsatz und Integralformel von Cauchy

Der Integrationsweg kommt in der Definition des Kurvenintegrals vor und deswegen hängt das Kurvenintegral im allgemeinen von ihm ab.

Es kann jedoch vorkommen, dass die Integration längs zweier verschiedener Wege von z_1 nach z_2 das gleiche Resultat ergibt.

Frage: unter welchen Bedingungen ist

$$\int_{K_1} f(z)\,dz = \int_{K_2} f(z)\,dz\ ,$$

wenn K_1 und K_2 zwei verschiedene Integrationswege mit dem gleichen Anfangs- und Endpunkt sind?

Gleichbedeutend ist die Frage, unter welchen Bedingungen

$$\oint_K f(z)\,dz = 0$$

ist für geschlossene Kurven K.

Eine hinreichende Bedingung hierfür gibt der folgende Satz.

Satz 7.9 (Cauchyscher Integralsatz; Hauptsatz der Funktionentheorie)
Es sei $\mathbb{G} \subseteq \mathbb{C}$ ein Gebiet und $f : \mathbb{G} \to \mathbb{C}$ holomorph. K sei eine geschlossene, stückweise stetig differenzierbare, doppelpunktfreie Kurve, deren Inneres ganz in $\mathbb{G}$ enthalten sei.

Dann gilt

$$\oint_K f(z)\,dz \ = \ 0\ .$$

Satz 7.10 (Cauchy-Integralformel)
Es sei $\mathbb{G} \subseteq \mathbb{C}$ ein Gebiet und $f : \mathbb{G} \to \mathbb{C}$ holomorph. Es sei K eine stückweise stetig differenzierbare, einfach geschlossene Kurve in $\mathbb{G}$, deren Inneres ganz in $\mathbb{G}$ liegt. Dann gilt für alle $a \in \mathbb{G}$, die im Inneren von K liegen:

$$f(a) = \frac{1}{2\pi i} \oint_K \frac{f(z)}{z-a}\,dz\ . \tag{7.15}$$

Beispiel(e) 7.11
Mit Partialbruchzerlegung erhält man

$$\begin{aligned} \oint\limits_{|z|=3} \frac{e^z}{z^2+2z}\,dz &= \frac{1}{2} \oint\limits_{|z|=3} \frac{e^z}{z}\,dz - \frac{1}{2} \oint\limits_{|z|=3} \frac{e^z}{z+2}\,dz \\ &= \pi i e^0 - \pi i e^{-2} = \pi i(1-e^{-2}) \end{aligned}$$

Die Cauchy-Integralformel behält ihre Gültigkeit, wenn man die Voraussetzungen abschwächt: die Kurve K darf auch auf dem Rand des Gebiets $\mathbb{G}$ liegen, sofern ihr Inneres nach wie vor ganz in $\mathbb{G}$ liegt und das auf $\mathbb{G}$ holomorphe f auf dem Rand $\partial\mathbb{G}$ von $\mathbb{G}$ wenigstens noch stetig ist.

7.3 Potenzreihen, Laurentreihen und Residuensatz

Potenzreihen lassen sich nicht nur in $\mathbb{R}$, sondern genauso gut in $\mathbb{C}$ definieren.

Definition 7.12 ((Komplexe) Potenzreihe)
Sei $a \in \mathbb{C}$ und $\{a_n\}$, $n \in \mathbb{N}_0$, eine Folge komplexer Zahlen, so heißt die Folge $\{p_n\}$, $n \in \mathbb{N}_0$, komplexer Funktionen

$$p_n : \mathbb{C} \to \mathbb{C}, \quad z \mapsto \sum_{k=0}^{n} a_k (z-a)^k \tag{7.16}$$

(komplexe) Potenzreihe und wird mit $\sum_{k=0}^{\infty} a_k (z-a)^k$ notiert.

Die Schreibweise $\sum_{k=0}^{\infty} a_k (z-a)^k$ steht also für eine Folge. *Gleichzeitig* benutzt man sie, um für festes z den Wert der Reihe mit Gliedern $a_k(z-a)^k$ auszudrücken.

Die erste Frage ist, für welche Werte von z die Reihe $\sum_{k=0}^{\infty} a_k (z-a)^k$ konvergiert. Erstaunlicherweise können dabei nur drei Fälle auftreten.

1. Fall: Konvergenz ausschließlich für $z = a$.

Beispiel(e) 7.13
Betrachte für $a = 0$ und irgendein $z \neq 0$ die Reihe $\sum_{k=0}^{\infty} k! z^k$. Zur Untersuchung der Konvergenz wird das Quotientenkriterium angewendet, das auch im Komplexen gilt:

$$\left| \frac{(k+1)! z^{k+1}}{k! z^k} \right| = (k+1)|z| \longrightarrow \infty \quad \text{für } k \longrightarrow \infty.$$

Die Reihe konvergiert also nicht für $z \neq 0$.

2. Fall: Konvergenz für alle $z \in \mathbb{C}$.

Beispiel(e) 7.14
Betrachte für $a = 0$ und irgendein $z \in \mathbb{C}$ die Reihe $\sum_{k=0}^{\infty} z^k/k!$. Zur Untersuchung der Konvergenz wird wieder das Quotientenkriterium angewendet:

$$\left| \frac{z^{k+1} k!}{(k+1)! z^k} \right| = \frac{|z|}{k+1} \longrightarrow 0 \quad \text{für } k \longrightarrow \infty.$$

Die Reihe konvergiert also für beliebiges z.

3. Fall: Es gibt ein $R > 0$ so, dass die Reihe für alle z mit $|z-a| < R$ konvergiert und für alle z mit $|z-a| > R$ divergiert. Man nennt dann R den **Konvergenzradius** der Reihe.

Beispiel(e) 7.15
Betrachte für $a = 0$ und $z \in \mathbb{C}$ die Reihe $\sum_{k=0}^{\infty} z^k$. Zur Untersuchung der Konvergenz wird wieder das Quotientenkriterium angewendet:

$$\left|\frac{z^{k+1}}{z^k}\right| = |z| \longrightarrow |z| \quad \text{für } k \longrightarrow \infty.$$

Die Reihe konvergiert also für $|z| < 1$ und sie divergiert für $|z| > 1$, der Konvergenzradius ist $R = 1$.

Man kann formal auch in den beiden ersten Fällen einen Konvergenzradius festsetzen:

- Im 1. Fall sagt man, die Reihe habe Konvergenzradius $R = 0$.
- Im 2. Fall sagt man, die Reihe habe Konvergenzradius $R = \infty$.

In jedem Fall ist das Konvergenzgebiet also eine (eventuell auf einen Punkt zusammengeschrumpfte oder auf ganz $\mathbb{C}$ ausgedehnte) Kreisscheibe.

Potenzreihen dürfen im Inneren ihres Konvergenzgebiets gliedweise differenziert werden. Wenn die Potenzreihe

$$\sum_{k=0}^{\infty} a_k (z-a)^k \tag{7.17}$$

Konvergenzradius $R > 0$ (eventuell $R = \infty$) hat, wenn sie also auf

$$\mathbb{G} = \{z \in \mathbb{C};\ |z-a| < R\}$$

konvergiert, dann definiert

$$f : \mathbb{G} \to \mathbb{C}, \quad z \mapsto \sum_{k=0}^{\infty} a_k (z-a)^k, \tag{7.18}$$

eine holomorphe Funktion mit

$$f'(z) = \sum_{k=1}^{\infty} k a_k (z-a)^{k-1} \quad \text{für alle } z \in \mathbb{G}\ . \tag{7.19}$$

(7.19) selbst ist wieder eine Potenzreihe und hat denselben Konvergenzradius R wie (7.17), so dass für $z \in \mathbb{G}$ erneut gliedweise differenziert werden darf. Man bekommt so für die ℓ-te Ableitung von f

$$f^{(\ell)}(z) = \sum_{k=\ell}^{\infty} k(k-1)\cdots(k-\ell+1) a_k (z-a)^{k-\ell} \quad \text{für alle } z \in \mathbb{G}\ . \tag{7.20}$$

Nun gilt sogar, dass nicht nur jede Potenzreihe eine holomorphe Funktion ist, sondern es ist auch umgekehrt jede holomorphe Funktion in eine Potenzreihe entwickelbar.

Satz 7.16 (Holomorphie und Potenzreihenentwicklung)

(a): Sei $\sum_{k=0}^{\infty} a_k(z-a)^k$ eine Potenzreihe mit Konvergenzradius $R > 0$, so ist die Funktion

$$f : \{z \in \mathbb{C}; |z-a| < R\} \to \mathbb{C}, \quad z \mapsto \sum_{k=0}^{\infty} a_k(z-a)^k \tag{7.21}$$

holomorph und es gilt:

$$f'(z) = \sum_{k=1}^{\infty} a_k k(z-a)^{k-1}. \tag{7.22}$$

(b): Sei $f : \mathbb{G} \to \mathbb{C}$ im Gebiet $\mathbb{G}$ holomorph, so ist f um jede Stelle $a \in \mathbb{G}$ in eine Potenzreihe

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(z-a)^k \tag{7.23}$$

entwickelbar. Die Reihe konvergiert im Inneren des größten Kreises

$$\{z \in \mathbb{C}; |z-a| < r\} \tag{7.24}$$

um a, in dem f holomorph ist.

(c): Unter den Voraussetzungen des Satzes 7.10 gilt die (7.15) verallgemeinernde Formel

$$f^{(k)}(a) = \frac{k!}{2\pi i} \oint_K \frac{f(z)}{(z-a)^{k+1}}\, dz \tag{7.25}$$

für die Koeffizienten in (7.23).

An diesem Satz ist bemerkenswert, dass bei komplexen Funktionen die Holomorphie, d.h. die einmalige Differenzierbarkeit, für eine Potenzreihenentwicklung hinreicht. Bei reellen Funktionen reicht nicht einmal die beliebig häufige Differenzierbarkeit.

Im Reellen gibt es einige Funktionen, die sich auf ganz $\mathbb{R}$ als Potenzreihen darstellen lassen:

$$\sin : \mathbb{R} \to \mathbb{R}, \qquad x \mapsto \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1}, \tag{7.26}$$

$$\cos : \mathbb{R} \to \mathbb{R}, \qquad x \mapsto \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}, \tag{7.27}$$

$$\exp : \mathbb{R} \to \mathbb{R}, \qquad x \mapsto \sum_{k=0}^{\infty} \frac{1}{k!} x^k. \tag{7.28}$$

Daher erhält man durch Ersetzung von $x \in \mathbb{R}$ durch $z \in \mathbb{C}$ die folgenden holomorphen Funktionen:

$$\sin : \mathbb{C} \to \mathbb{C}, \qquad z \mapsto \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k+1}, \tag{7.29}$$

$$\cos : \mathbb{C} \to \mathbb{C}, \qquad z \mapsto \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} z^{2k}, \tag{7.30}$$

$$\exp : \mathbb{C} \to \mathbb{C}, \qquad z \mapsto \sum_{k=0}^{\infty} \frac{1}{k!} z^k =: e^z. \tag{7.31}$$

Aus

$$e^{iz} = \sum_{k=0}^{\infty} \frac{1}{k!} (iz)^k = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} z^{2k} + i \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k+1} \tag{7.32}$$

folgt:

$$e^{iz} = \cos(z) + i \sin(z) \quad \text{bzw.} \tag{7.33}$$

$$e^{ix} = \cos(x) + i \sin(x), \quad x \in \mathbb{R} \quad \text{(wurde in Kapitel 3 so definiert).} \tag{7.34}$$

Da nun für die Funktion

$$g : \mathbb{C} \to \mathbb{C}, \quad z \mapsto \exp(z_1 + z_2 - z) \exp(z) \tag{7.35}$$

mit $z_1, z_2 \in \mathbb{C}$ gilt:

$$g'(z) = -\exp(z_1 + z_2 - z) \exp(z) + \exp(z_1 + z_2 - z) \exp(z) \equiv 0, \tag{7.36}$$

folgt:

$$g(z) \equiv \tilde{z} \in \mathbb{C}. \tag{7.37}$$

Aus $g(0) = g(z_2)$ folgt die wichtige Beziehung:

$$\exp(z_1 + z_2) = \exp(z_1) \exp(z_2). \tag{7.38}$$

Mit der komplexen Exponentialfunktion lässt sich auch ihre Umkehrung definieren, der komplexe Logarithmus. Dazu sei zuerst an die Darstellung einer komplexen Zahl mittels Polarkoordinaten erinnert:

$$z = re^{i\varphi} = r(\cos\varphi + i \sin\varphi)$$

mit $r = |z|$ und $\varphi = \arg(z)$, dem Argument von z. Im Fall $z = 0$ ist das Argument völlig beliebig und selbst im Fall $z \neq 0$ ist es nur bis auf Vielfache von 2π festgelegt, da $\cos$ und $\sin$ beides Funktionen mit Periode 2π sind. Man kann wenigstens für $z \neq 0$ Eindeutigkeit erzwingen, wenn man festlegt, dass das Argument immer aus dem halboffenen Intervall $(-\pi, \pi]$ sein soll. Dies nennt man den **Hauptwert des Arguments von** $z \neq 0$ und bezeichnet diesen mit

$$\mathrm{Arg}(z) \in (-\pi, \pi], \quad z \in \mathbb{C}\backslash\{0\} \ .$$

Nun können wir den **Hauptwert des komplexen Logarithmus** definieren als Funktion

$$\mathrm{Ln} : \mathbb{C}\backslash\{0\} \to \mathbb{C}, \quad z \mapsto \mathrm{Ln}(z) = \ln(|z|) + i\mathrm{Arg}(z) \ . \tag{7.39}$$

Hier bezeichnet $\ln(|z|)$ den reellen, natürlichen Logarithmus der reellen Zahl $|z| > 0$. Die Definition ist so gemacht, dass

$$\exp(\mathrm{Ln}(z)) = \exp(\ln|z|) \cdot \exp(i\mathrm{Arg}(z)) = |z|e^{i\varphi} = z$$

für $\mathrm{Arg}(z) = \varphi$. $\exp(z)$ kann nie den Wert 0 annehmen und entsprechend kann die Umkehrfunktion Ln nicht für $z = 0$ definiert sein.

Beispiel(e) 7.17
Wir berechnen einige Werte des komplexen Logarithmus.

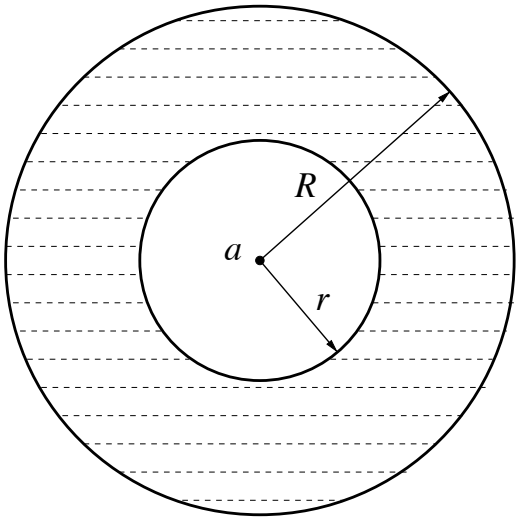
$$\begin{aligned} \mathsf{Ln}(1) &= \ln(1) + i \cdot 0 = 0 \,, \\ \mathsf{Ln}(-1) &= \ln(1) + i \cdot \pi = i\pi \,, \\ \mathsf{Ln}(i) &= \ln(1) + i \cdot \frac{\pi}{2} = i\frac{\pi}{2} \,. \end{aligned}$$

Man beachte, dass der komplexe Logarithmus auf der negativen reellen Achse unstetig ist.

Wir betrachten jetzt Funktionen, die auf Kreisringen

$$K_{r,R}(a) := \{z \in \mathbb{C};\ r < |z - a| < R\}, \quad 0 \le r < R \le \infty,$$

definiert und holomorph sind.



Ein dabei sehr häufiger Fall ist $r = 0$, das heißt

$$K_{0,R}(a) := \{z \in \mathbb{C};\ 0 < |z - a| < R\} \,.$$

$K_{0,R}(a)$ nennt man „punktierte Kreisscheibe“. Punktierte Kreisscheiben treten auf, wenn f in a nicht definiert oder nicht komplex differenzierbar ist. Eine solche Ausnahmestellen a nennt man **Singularität** von f. Wenn $f : K_{0,R}(a) \to \mathbb{C}$ holomorph ist für ein $R > 0$, dann nennt man die Singularität **isoliert**.

Beispiel(e) 7.18

(a): Die Funktion

$$f : \mathbb{C} \setminus \{0\} \to \mathbb{C}, \quad z \mapsto \frac{1}{z}$$

besitzt eine Singularität in $z = 0$.

(b): Die Funktion

$$f : \mathbb{C} \setminus \{0\} \to \mathbb{C}, \quad z \mapsto \sin\left(\frac{1}{z}\right)$$

besitzt ebenfalls eine Singularität in $z = 0$.

(c): Die Funktion

$$f : \mathbb{C} \to \mathbb{C}, \quad z \mapsto \begin{cases} z^2 \sin\left(\frac{1}{z}\right) & , z \neq 0 \\ 0 & , z = 0 \end{cases}$$

ist zwar auf ganz $\mathbb{C}$ definiert, aber in $z = 0$ nicht (komplex) differenzierbar und hat deswegen ebenfalls eine Singularität in $z = 0$.

(d): Der Hauptzweig des komplexen Logarithmus ist auf der negativen reellen Achse $\{x + iy \in \mathbb{C};\ x \leq 0,\ y = 0\}$ unstetig, insbesondere also nicht differenzierbar. Jeder Punkt auf der negativen Halbachse ist somit eine Singularität und keine dieser Singularitäten ist isoliert.

In Verallgemeinerung von Taylor-Reihen gilt der folgende

Satz 7.19 (Laurentreihe)

Es sei $f : K_{r,R}(a) \to \mathbb{C}$ holomorph. Dann gibt es Koeffizienten $c_n \in \mathbb{C}$, $n \in \mathbb{Z}$, mit

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z-a)^n = \sum_{n=0}^{\infty} c_n (z-a)^n + \sum_{n=1}^{\infty} \frac{c_{-n}}{(z-a)^n} \,. \tag{7.40}$$

Die Reihe darf in jedem Teilring $A := \{z;\ r < r_1 < |z-a| < R_1 < R\}$ gliedweise differenziert werden. Ebenso dürfen Integrale längs Kurven, die in A verlaufen, gliedweise berechnet werden.

Die Koeffizienten c_n sind eindeutig bestimmt durch

$$c_n = \frac{1}{2\pi i} \oint\limits_{|\zeta - a| = \varrho} \frac{f(\zeta)}{(\zeta - a)^{(n+1)}} \, d\zeta, \quad n \in \mathbb{Z},\ r < \varrho < R \,. \tag{7.41}$$

Beispiel(e) 7.20

(a): Die Funktion

$$f : \mathbb{C} \setminus \{0\} \to \mathbb{C}, \quad z \mapsto \frac{1}{z}$$

besitzt um $z = 0$ die folgende Entwicklung in eine Laurentreihe:

$$f(z) = \frac{1}{z}.$$

(b): Die Funktion

$$f : \mathbb{C} \setminus \{0\} \to \mathbb{C}, \quad z \mapsto \sin\left(\frac{1}{z}\right)$$

besitzt um $z = 0$ die folgende Entwicklung in eine Laurentreihe:

$$f(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} \frac{1}{z^{2k+1}} ,$$

die sich ganz einfach durch Einsetzen von $1/z$ in die Sinus-Reihe ergibt.

(c): Die Funktion

$$f : \mathbb{C} \to \mathbb{C}, \quad z \mapsto \begin{cases} z^2 \sin\left(\frac{1}{z}\right) & , z \neq 0 \\ 0 & , z = 0 \end{cases}$$

besitzt um $z = 0$ die folgende Entwicklung in eine Laurentreihe:

$$f(z) = z - \frac{1}{3!z} + \frac{1}{5!z^3} - \frac{1}{7!z^5} \pm \dots .$$

(d): Der Hauptzweig des komplexen Logarithmus besitzt keine Laurentreihe um $z = 0$, da er auf keinem Kreisring um 0 holomorph ist.

Betrachten wir nun ein f mit einer isolierten Singularität in $z = a$. Es sei also

$$A = \{z \in \mathbb{C};\ 0 < |z - a| < R\}, \quad 0 < R \leq \infty$$

und $f : A \to \mathbb{C}$ holomorph, so dass eine Laurentreihe wie in (7.40) existiert:

$$f(z) = \dots + \frac{c_{-2}}{(z-a)^2} + \frac{c_{-1}}{z-a} + c_0 + c_1(z-a) + c_2(z-a)^2 + \dots \tag{7.42}$$

Für $0 < r < R$ berechnen wir dann gemäß Satz 7.19 das Kurvenintegral

$$\oint\limits_{|z-a|=r} f(z)\, dz$$

gliedweise und erhalten aus dem Fundamentalintegral (vergleiche Beispiel 7.8):

$$\oint\limits_{|z-a|=r} f(z)\, dz = \dots + 0 + 2\pi i c_{-1} + 0 + 0 + 0 + \dots$$

Der Koeffizient c_{-1} der Laurent-Entwicklung um eine isolierte Singularität bekommt einen besonderen Namen:

$$c_{-1} = \frac{1}{2\pi i} \oint\limits_{|z-a|=r} f(z)\, dz =: \ \mathsf{Res}(f, a) \tag{7.43}$$

heißt **Residuum** von f in a.

Satz 7.21 (Residuensatz)
Sei $\mathbb{G} \subseteq \mathbb{C}$ ein Gebiet, seien $a_1, \ldots, a_N$ isoliert und

$$f : \mathbb{G}\backslash\{a_1, \ldots, a_N\} \to \mathbb{C} \tag{7.44}$$

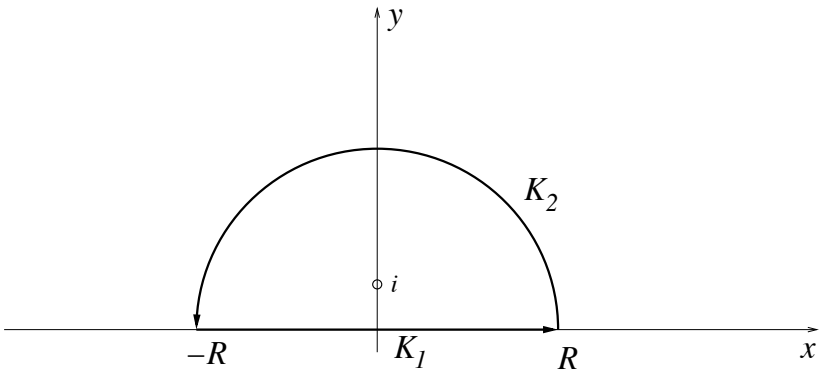
holomorph (in $a_1, \ldots, a_N$ dürfen also Singularitäten von f liegen). Sei K eine einfach geschlossene, stückweise stetig differenzierbare Kurve, die in $\mathbb{G}\backslash\{a_1, \ldots, a_N\}$ verläuft (also durch keinen der Punkte $a_1, \ldots, a_N$ geht) und die die Punkte $a_1, \ldots, a_n$ in ihrem Inneren positiv umläuft. Dann gilt

$$\oint_K f(z)dz = 2\pi i \sum_{k=1}^{n} \mathrm{Res}(f, a_k) \; . \tag{7.45}$$

Der Residuensatz wird häufig zur Berechnung reeller Integrale verwendet.
Sei beispielsweise

$$f : \mathbb{C}\backslash\{\pm i\} \to \mathbb{C}, \quad z \mapsto \frac{1}{1+z^2}, \tag{7.46}$$

mit isolierten Singularitäten $a_1 = i$ und $a_2 = -i$. Für ein $R > 1$ betrachten wir die Kurve K, die sich wie im Folgenden skizziert aus den stetig differenzierbaren Kurvenstücken K_1 und K_2 zusammensetzt: $K = K_1 \cup K_2$.



Nach dem Residuensatz ist

$$\oint_K f(z)dz = 2\pi i \cdot \mathrm{Res}(f, i) \; .$$

Zur Berechnung des Residuums kann man eine Partialbruchzerlegung machen:

$$f(z) = \frac{1}{1+z^2} = \frac{i}{2}\frac{1}{z+i} - \frac{i}{2}\frac{1}{z-i},$$

wobei der 1. Summand in einer Umgebung von $z = i$ holomorph ist, so dass mit dem Integralsatz von Cauchy

$$\oint_K \frac{i}{2}\frac{1}{z+i}dz = 0 \; .$$

Der 2. Summand ist seine eigene Laurententwicklung um $z = i$, so dass

$$\mathrm{Res}\left(-\frac{i}{2}\frac{1}{z-i}, i\right) = -\frac{i}{2} \; .$$

Somit ergibt sich

$$\oint_K f(z)dz = 2\pi i \left(-\frac{i}{2}\right) = \pi \; . \tag{7.47}$$

Parametrisiert man K_2 durch $\gamma(t) = Re^{it}$ für $0 \leq t \leq \pi$, so ergibt sich

$$\begin{aligned} \left| \int_{K_2} \frac{dz}{1+z^2} \right| &= \left| \int_0^\pi \frac{Rie^{it}}{1+R^2e^{2it}}\, dt \right| \\ &\leq \int_0^\pi \frac{R}{R^2-1}\, dt = \frac{\pi R}{R^2-1} \longrightarrow 0 \quad \text{für } R \to \infty. \end{aligned}$$

Folglich ist mit (7.47)

$$\pi = \oint_K f(z)dz \longrightarrow \int_{-\infty}^{\infty} \frac{dx}{1+x^2} \quad \text{für } R \longrightarrow \infty \; .$$

Teil IV

Mathematik 4

Kapitel 8

Numerische Mathematik

8.1 Grundbegriffe der Numerik

Es ist ein weit verbreiteter Irrtum zu glauben, dass Computer stets korrekte Ergebnisse liefern. In Wahrheit ist es eine Kunst für sich, Computer so zu programmieren, dass man ihren Ergebnissen trauen darf.

Am 25. Februar 1991 berechnete im ersten Golfkrieg der Waffensystemrechner der Patriot-Batterie die Flugbahn einer irakischen Scud-Rakete falsch. Da scheinbar keine Gefahr drohte, wurde keine Patriot-Rakete zur Zerstörung der Scud-Rakete gestartet. Durch einen Volltreffer wurden 28 US-Soldaten getötet, über 100 US-Soldaten wurden verletzt.

Um zu verstehen, wie es zu diesem Unfall kommen konnte, muss man verstehen, wie Computer rechnen und warum sie dabei Fehler machen *müssen*.

Die Menge $\mathbb{R}$ der reellen Zahlen ist unbegrenzt. Zu jeder reellen Zahl gibt es noch größere und noch kleinere reelle Zahlen, und zu jedem Paar von zwei verschiedenen reellen Zahlen gibt es weitere, dazwischenliegende reelle Zahlen. Im Gegensatz dazu ist die Menge $\mathbb{M}$ der in einer Maschine exakt darstellbaren reellen Zahlen immer endlich, also beschränkt und diskret.

Die meisten Rechner verwenden sogenannten Gleitpunktzahlen (*floating point numbers*) als Maschinenzahlen. Ungeachtet der tatsächlichen Implementierung kann man die *Menge* $\mathbb{M}$ dieser Zahlen durch vier Parameter beschreiben.

Definition 8.1 (Menge $\mathbb{M}$ der Maschinenzahlen)

$$\mathbb{M} := \mathbb{M}(B,t,\alpha,\beta) := \{x \in \mathbb{R};\quad x = S \cdot B^E \quad \text{mit} \quad S = 0 \quad \text{oder} \quad B^{t-1} \le |S| < B^t, \tag{8.1}$$

$$\alpha \le E \le \beta, \quad S, E \in \mathbb{Z}\} \tag{8.2}$$

mit den Parametern $\alpha, \beta \in \mathbb{Z}$, $B \in \mathbb{N} \setminus \{1\}$ (B heißt Basis) und der Stellenzahl $t \in \mathbb{N}$. Die ganzzahligen Variablen S und E heißen Signifikant und Exponent.

Die Parameter B, t, α, β sind durch die Implementierung festgelegt und werden daher nicht gespeichert. Es gibt eine kleinste positive Maschinenzahl σ und eine größte Maschinenzahl λ:

$$\begin{aligned} \sigma &:= B^{t-1} \cdot B^{\alpha} \\ \lambda &:= (B^t - 1) \cdot B^{\beta} \end{aligned}$$

Bemerkung. Es gibt auch Rechner, die mit **Fixpunktzahlen** arbeiten, also mit Zahlen, die alle ein- und denselben Exponenten E aufweisen. Der Zahlenvorrat ist dann die Menge

$$\mathbb{F}_{B,E,t} = \{S \cdot B^E;\ -B^t < S < B^t\},$$

mit festen Werten B, E und t.

Beispiel: Die durch $B = 10$, $E = 0$ und $t = 2$ gegebenen Fixpunktzahlen sind gerade die ganzen Zahlen zwischen -99 und 99.

Im Folgenden sind mit Maschinenzahlen aber immer Zahlen gemäß Definition 8.1 gemeint, wo nicht explizit etwas anderes gesagt ist.

Beispiel(e) 8.2
Der Standard ANSI/IEEE-Std-754-1985 (international: IEC-60559) für Gleitpunktarithmetik (*floating point arithmetic*) legt Parameter für verschiedene Sätze von Maschinenzahlen fest, die dem Rechnen in verschiedenen Genauigkeitsstufen entsprechen.

Genauigkeitsstufe	B	t	α	β
single precision	2	24	-149	104
double precision	2	53	-1074	971
extended precision	2	64	-16445	16320

Der 2008 verabschiedete Nachfolge-Standard IEEE-754-2008 sieht weitere Genauigkeitsstufen vor, unter anderem auch eine „halbe“ Genauigkeit von 16 Bit und eine vierfache Genauigkeit von 128 Bit, die meistens allerdings nur in Software emuliert wird.

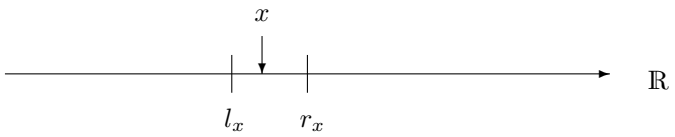
Von einer Maschinenzahl $g \in \mathbb{M}$ werden nur S und E gespeichert.

Beispiel(e) 8.3
$B = 10$, $t = 3$, $\alpha = -1$, $\beta = 1$,

$$\mathbb{M} = \{0, \underbrace{\pm 10.0, \pm 10.1, \ldots, \pm 99.9}_{E=-1}, \underbrace{\pm 100, \pm 101, \ldots, \pm 999}_{E=0}, \tag{8.3}$$

$$\underbrace{\pm 1000, \pm 1010, \ldots, \pm 9990}_{E=1}\}. \tag{8.4}$$

Jede reelle Zahl $x \in [-\lambda, \lambda]$ hat genau eine Maschinenzahl l_x als Nachbar zu ihrer Linken und genau eine Maschinenzahl r_x als Nachbar zu ihrer Rechten, d.h. es gilt für $x \notin \mathbb{M}$: $l_x < x < r_x$ und $(l_x, r_x) \cap \mathbb{M} = \emptyset$



Für $x \in \mathbb{M}$ gilt: $l_x = r_x = x$.

Jedes $x \in \mathbb{R}$ wird im Rechner ersatzweise durch seinen linken oder rechten Nachbarn dargestellt. Formal stellt dies eine Abbildung $rd : \mathbb{R} \to \mathbb{M}$ dar, die man als **Rundung** bezeichnet. Genauer muss man dazu noch zwei symbolische Zahlen „$-\infty$“ kleiner als $-\lambda$ und „∞“ größer als λ einführen (diese sind auch im IEEE-Format vorgesehen und heißen dort +Inf und -Inf) und kann dann auf vier gebräuchliche Arten runden:

- Abrunden: $rd_- : \mathbb{R} \to \mathbb{M} \cup \{\pm\infty\}$

$$x \mapsto \begin{cases} -\infty & \text{falls} \quad x < -\lambda \\ l_x & \text{falls} \quad x \geq -\lambda \end{cases}. \tag{8.5}$$

- Aufrunden: $rd_+ : \mathbb{R} \to \mathbb{M} \cup \{\pm\infty\}$

$$x \mapsto \begin{cases} \infty & \text{falls} \quad x > \lambda \\ r_x & \text{falls} \quad x \leq \lambda \end{cases}. \tag{8.6}$$

- Abhacken: $rd_0 : \mathbb{R} \to \mathbb{M} \cup \{\pm\infty\}$

$$x \mapsto \begin{cases} l_x & \text{falls} \quad x \geq 0 \\ r_x & \text{falls} \quad x \leq 0 \end{cases}, \tag{8.7}$$

  wobei $l_x = \lambda$ für alle $x > \lambda$ und $r_x = -\lambda$ für alle $x < -\lambda$.

- Korrektes Runden: $rd_* : \mathbb{R} \to \mathbb{M} \cup \{\pm\infty\}$

$$x \mapsto \begin{cases} \infty & \text{falls} \quad x > \lambda \\ -\infty & \text{falls} \quad x < -\lambda \\ l_x & \text{falls} \quad x < \frac{l_x + r_x}{2} \\ r_x & \text{falls} \quad x > \frac{l_x + r_x}{2} \end{cases}. \tag{8.8}$$

Reelle Zahlen $\frac{l_x+r_x}{2}$ werden so gerundet, dass $rd_*(\frac{l_x+r_x}{2})$ einen geraden Signifikanten S hat.

Die ersten drei Modi bezeichnet man als **gerichtetes Runden**. Runden führt zu einem (relativen) Fehler in der Darstellung einer Zahl, dem sogenannten **Rundungsfehler** $(rd(x)-x)/x$ (für $0 = x \in \mathbb{M}$ gibt es keinen Rundungsfehler: $rd(x) = x$). Für Zahlen $x > \lambda$, $x < \lambda$ und $-\sigma < x < \sigma$ kann der Rundungsfehler den maximalen Wert 1 erreichen, das heißt $rd(x)$ kann zu 100% falsch sein. Wenn aber $\sigma \leq |x| \leq \lambda$, dann bleibt der maximale Rundungsfehler beim gerichteteten Runden beschränkt durch den maximalen relativen Abstand zweier Maschinenzahlen, beim korrekten Runden bleibt er beschränkt durch den halben maximalen relativen Abstand zweier Maschinenzahlen. Zwei benachbarte Maschinenzahlen $g = S \cdot B^E$ und $\bar{g} = (S+1) \cdot B^E$, $S < B^t - 1$ haben den relativen Abstand:

$$\frac{\bar{g}-g}{g} = \frac{B^E}{S \cdot B^E} = \frac{1}{S}, \quad g \neq 0. \tag{8.9}$$

Der maximale relative Abstand zweier benachbarter Maschinenzahlen aus $\mathbb{M}(B,t,\alpha,\beta)$ *ist gegeben durch* $\varrho = B^{1-t}$.

Für den Rundungsfehler haben wir demnach, sofern $\sigma \leq |x| \leq \lambda$ und $x \neq 0$:

$$\frac{rd(x)-x}{x} = \varepsilon,\ |\varepsilon| \leq \varepsilon_{mach} := \begin{cases} \frac{1}{2}B^{1-t} & \text{für} \quad rd_* \\ B^{1-t} & \text{für} \quad rd_-, rd_+, rd_0 \end{cases} \tag{8.10}$$

Die Zahl ε_{mach} heißt **Maschinengenauigkeit**. Statt (8.10) können wir auch

$$rd(x) = x(1+\varepsilon) \quad \text{mit} \quad |\varepsilon| \leq \varepsilon_{mach} \tag{8.11}$$

schreiben und diese Darstellung ist für alle $\sigma \leq |x| \leq \lambda$ richtig.

Beispiel(e) 8.4
Der Rechner des Patriot-Systems benutzte binäre Fixpunktzahlen mit 24 Stellen. Die Zahl $\frac{1}{10}$ wurde auf `0.00011001100110011001100` (im Binärsystem) gerundet, die von $\frac{1}{10}$ um etwa $9.5 \cdot 10^{-8}$ abweicht. Nach 100 Betriebsstunden ohne Korrektur der Zeit ergab sich ein Fehler von

$$100 \cdot 60 \cdot 60 \cdot 10 \cdot 9.5 \cdot 10^{-8} \approx 0.34 \text{ sec.}$$

In diesen ca. 0.34 sec. legte die Scud-Rakete etwa 570 m zurück - für die Betroffenen entscheidende 570 m. Besonders tragisch: In der Betriebsanleitung der Patriot-Rakete wurde ein wesentlich kürzeres Zeitintervall bis zur Zeitkorrektur vorgeschrieben; ferner sollte die Zeit in $\frac{1}{8}$ Sekunden (binär: 0.001) gemessen werden, was keinerlei Rundungsfehler zur Folge gehabt hätte. Ein Manager hat dies später aber auf die fatalen $\frac{1}{10}$ Sekunden korrigiert.

Rundungsfehler treten nicht nur dann auf, wenn eine gegebene Zahl $x \in \mathbb{R}$ durch eine Maschinenzahl approximiert werden muss. Vielmehr ist auch die Verknüpfung zweier Maschinenzahlen nicht unbedingt wieder eine Maschinenzahl.

So können sich im Lauf einer Rechnung Werte ergeben, die außerhalb des Bereiches von $\mathbb{M}$ liegen, etwa bei der Addition der größten Maschinenzahl λ mit sich selbst: $(\lambda + \lambda) \notin \mathbb{M}$. In diesem Fall spricht man von Bereichsüberschreitung oder Exponentenüberlauf. Ein anderer Fall von Bereichsüberschreitung wäre der Exponentenunterlauf. Zum Beispiel ergibt sich für $x = 5{\cdot}2^{-3} \in \mathbb{M}(2,3,-3,0)$ und $y = 4{\cdot}2^{-3} \in \mathbb{M}(2,3,-3,0)$, dass $x-y = 1{\cdot}2^{-3} \notin \mathbb{M}(2,3,-3,0)$. In diesem Fall ist $0 \neq |x-y| < \sigma$. An gefährdeten Stellen in Numerik-Programmen (z.B. bei der numerischen Berechnung von Determinanten) muss Vorsorge für Bereichsüberschreitungen

getroffen werden.

Auch abgesehen von Bereichsüberschreitungen können die vier arithmetischen Grundoperationen im allgemeinen nicht exakt ausgeführt werden. Zum Beispiel ist für $x = 5 \cdot 2^{-3} \in \mathbb{M}(2,3,-3,0)$ und $y = 4 \cdot 2^{-3} \in \mathbb{M}(2,3,-3,0)$ das Ergebnis $x + y = 9 \cdot 2^{-3} = 9/8 \notin \mathbb{M}(2,3,-3,0)$, aber $\sigma \leq |x+y| \leq \lambda$. Der Computer muss in solchen Fällen ein Ersatzresultat berechnen. Für jede der Grundoperationen $* \in \{+,-,\times,/\}$ und für alle Zahlen $a, b \in \mathbb{M}$ haben wir

$$\begin{array}{rl} \text{ein exaktes Resultat} & a * b \in \mathbb{R}, \text{ i.a. } \notin \mathbb{M} \text{ und} \\ \text{ein berechnetes Ersatzresultat} & a \overset{\bullet}{*} b \in \mathbb{M}. \end{array}$$

Im besten Fall gilt:

$$a \overset{\bullet}{*} b = \mathsf{rd}(a * b), \qquad \forall a, b \in \mathbb{M} \tag{8.12}$$

Das bedeutet, dass das Ergebnis jeder arithmetischen Operation gleich dem exakten Ergebnis ist, gerundet auf eine Maschinenzahl. Dies ist das theoretisch bestmögliche Ergebnis. Wenn (8.12) erfüllt ist, spricht man deshalb von einer **idealen Arithmetik**.

Beispiel(e) 8.5
$B = 10$, $t = 4$, rd_*, $\alpha = -3$, $\beta = 0$:

- $a = 20.29; \quad b = 49.31$
  $a \times b = 1000.4999 \quad a \overset{\bullet}{\times} b = 1000 = rd_*(a \times b).$
- $a = 99.99; \quad b = 1.020$
  $a + b = 101.01 \quad a \overset{\bullet}{+} b = 101.0 = rd_*(a + b).$

Wichtig ist, dass man die ideale Arithmetik realisieren kann, *ohne* vorher das exakte Ergebnis $a * b$ ausrechnen zu müssen. Im schon erwähnten IEEE-Standard 754 wird deswegen die Einhaltung von (8.12) für alle vier arithmetischen Grundoperationen (und ebenso für die Berechnung der Quadratwurzel) gefordert und zwar für die Rechnung in allen Genauigkeitsstufen. Die Rundung darf in irgendeinem der vier genannten Modi stattfinden. Wir können somit bei einer idealen Arithmetik stets unterstellen, dass

$$a \overset{\bullet}{+} b = (a+b)(1+\alpha), \tag{8.13}$$

$$a \overset{\bullet}{-} b = (a-b)(1+\sigma), \tag{8.14}$$

$$a \overset{\bullet}{\times} b = (a \times b)(1+\mu), \tag{8.15}$$

$$a \overset{\bullet}{/} b = (a/b)(1+\delta), \quad b \neq 0, \tag{8.16}$$

$$\overset{\bullet}{\sqrt{}} a = \sqrt{a}(1+\omega), \quad a \geq 0. \tag{8.17}$$

Sofern keine Bereichsüberschreitungen auftreten ist

$$|\alpha|, |\sigma|, |\mu|, |\delta|, |\omega| \leq \varepsilon_{mach} = \begin{cases} \frac{1}{2}B^{1-t} & \text{für} \quad rd_* \\ B^{1-t} & \text{für} \quad rd_-, rd_+, rd_0 \end{cases} \tag{8.18}$$

für alle entsprechenden $a, b \in \mathbb{M}$.

Auch wenn ein Prozessor die ideale Arithmetik implementiert hat, bleibt es natürlich dabei:

Jede einfache arithmetische Operation, die ein Rechner ausführt, kann fehlerhaft sein. Alte Fehler werden weitergegeben und bei jeder Operation kann ein neuer Fehler hinzukommen. Jeder verantwortungsvolle Entwickler numerischer Software muss sich Gedanken über die Auswirkung von Rundungsfehlern machen, zumindest wo es sich um sicherheitsrelevante Software handelt.

Die Abschätzung des Einflusses von Rundungsfehlern auf das Ergebnis einer Berechnung nennt man **Rundungsfehleranalyse**. Zur Formalisierung definieren wir zunächst: unter einem **Algorithmus** (Rechenverfahren) versteht man in der Numerik eine endliche Folge von Grundoperationen, deren Reihenfolge beim Ablauf eindeutig festliegt. Für Maschinenzahlen ausgeführt verfälschen Rundungsfehler die Zwischen- und Endresultate. Eine a priori Analyse dieser Fehler geht von den mathematisch exakten Identitäten (8.13)-(8.17) über die Rundungsfehler der Grundoperationen aus und ermittelt mit Hilfe von (8.18) Schranken für den Gesamtfehler. Da dabei immer die Verkettung der ungünstigsten Umstände unterstellt werden muss, ist der tatsächliche Fehler in einer konkreten Anwendung meist viel kleiner.

In der a priori Rundungsfehleranalyse hat sich die sogenannte Rückwärts-Analyse durchgesetzt, welche das berechnete Resultat als das exakte Resultat zu geeignet veränderten Eingabedaten interpretiert. Ausgangspunkt sind wieder die Gleichungen (8.13)-(8.17), die sich auch leicht verändert schreiben lassen:

$$a \overset{\bullet}{+} b = a(1+\alpha) + b(1+\alpha) \tag{8.19}$$

$$a \overset{\bullet}{-} b = a(1+\sigma) - b(1+\sigma) \tag{8.20}$$

$$a \overset{\bullet}{\times} b = a\sqrt{1+\mu} \times b\sqrt{1+\mu}, \qquad -1 < \mu \tag{8.21}$$

$$a \overset{\bullet}{/} b = a\sqrt{1+\delta} / \frac{b}{\sqrt{1+\delta}}, \qquad -1 < \delta, \tag{8.22}$$

$$\overset{\bullet}{\sqrt{a}} = \sqrt{a(1+\omega)}, \quad a \geq 0. \tag{8.23}$$

Die Interpretation hiervon ist: das vom Computer berechnete Ergebnis einer arithmetischen Operation ist das Ergebnis einer exakten Operation mit leicht veränderten Eingabedaten. Die in (8.19)-(8.23) benutzte Technik lässt sich auf die Kombination von Verknüpfungen ausdehnen, zum Beispiel

$$((a \overset{\bullet}{+} b) \overset{\bullet}{+} c) = ((a+b)(1+\alpha_1)+c)(1+\alpha_2) = a\underbrace{(1+\alpha_1)(1+\alpha_2)}_{=:\,(1+\varepsilon)} + b\underbrace{(1+\alpha_1)(1+\alpha_2)}_{=(1+\varepsilon)} + c(1+\alpha_2).$$

Wir haben $\varepsilon = \alpha_1 + \alpha_2 + \alpha_1\alpha_2$. Da $|\alpha_i| \leq \varepsilon_{mach}$ ist der Term $\alpha_1\alpha_2$ vernachlässigbar klein und man schreibt „in erster Näherung“ $|\varepsilon| \dot{\leq} 2\varepsilon_{mach}$. Auch hier also wieder das Ergebnis: das vom Computer (mit idealer Arithmetik bei Ausschluss von Bereichsüberschreitungen) berechnete Ergebnis entspricht einer exakten Berechnung mit leicht veränderten Daten.

Wir führen eine Rückwärts-Analyse an einem etwas komplizierteren Beispiel durch.

Beispiel(e) 8.6
Hornerschema: Auswertung eines Polynoms $y = P(x) = c_n x^n + c_{n-1}x^{n-1} + \ldots + c_1 x + c_0$.

$y = c_n$
`for (i=n-1; i>=0; i--) {`
$\quad y = y \cdot x + c_i;$
}

Unter dem Einfluss von Rundungsfehlern ergibt sich

$\tilde{y} = c_n \cdot (1 + \alpha_n) \qquad [\alpha_n := 0]$
`for (i=n-1; i>=0; i--) {`
$\quad \tilde{y} = (\tilde{y} \cdot x \cdot (1 + \mu_i) + c_i) \cdot (1 + \alpha_i);$
}

mit $|\mu_i|, |\alpha_i| \leq \varepsilon_{mach}$. Berechnet wird also im Endergebnis

$$\tilde{y} = c_n x^n (1 + \varepsilon_n) + \ldots + c_1 x (1 + \varepsilon_1) + c_0 (1 + \varepsilon_0) \tag{8.24}$$

mit

$$1 + \varepsilon_k := (1 + \mu_{k-1}) \cdots (1 + \mu_0) \cdot (1 + \alpha_k) \cdots (1 + \alpha_0) \; .$$

Wir können $\tilde{y}$ interpretieren als exakten Wert eines Polynoms mit geänderten Koeffizienten:

$$\tilde{y} = \tilde{c}_n x^n + \ldots + \tilde{c}_1 x + \tilde{c}_0, \quad \tilde{c}_k = c_k (1 + \varepsilon_k) \; .$$

In erster Näherung (unter Berücksichtigung von $\alpha_n = 0$) ergibt sich

$$|\varepsilon_k| \leq 2n\varepsilon_{mach} \; .$$

Man könnte $|\varepsilon_k|$ auch exakt abschätzen, denn es lässt sich zeigen: wenn

$$1 + \varepsilon := \prod_{i=1}^{n} (1 + \varepsilon_i)^{\pm 1} \quad \text{mit} \quad |\varepsilon_i| \leq \varepsilon_{mach} \; ,$$

dann gilt

$$|\varepsilon| \leq \frac{n\varepsilon_{mach}}{1 - n\varepsilon_{mach}} \; . \tag{8.25}$$

Aus (8.25) folgert man

$$|\tilde{c}_k - c_k| \leq \frac{(2k+1)\varepsilon_{mach}}{1 - (2k+1)\varepsilon_{mach}} |c_k| \leq \frac{2n\varepsilon_{mach}}{1 - 2n\varepsilon_{mach}} |c_k| \tag{8.26}$$

Offenbar stellt (8.26) *keine* Abschätzung des Fehlers im berechneten Ergebnis dar. Die Interpretation ist vielmehr, dass die bei Anwendung des Horner-Schemas begangenen Rundungsfehler das exakte Ergebnis nur so weit verfälschen, wie es auch relative Unsicherheiten der Größe $2n\varepsilon_{mach}/(1 - 2n\varepsilon_{mach}) \approx 2n\varepsilon_{mach}$ in den Eingangsdaten täten.

Nun muss man aber Fehler der Größenordnung ε_{mach} in den Eingangsdaten in der Praxis immer unterstellen, weil die Polynomkoeffizienten selbst Ergebnis einer vorangegangenen Rechnung sind oder auch nur, weil sie auf Maschinenzahlen gerundet wurden. Nach der Rückwärts-Analyse wissen wir also zwar *nicht*, wie groß der Fehler im berechneten Polynomwert tatsächlich ist, aber wir wissen immerhin eines: Rundungsfehler beim Hornerschema wirken sich auf das Endergebnis nicht

schlimmer aus, als es die zu unterstellenden Eingangsfehler in den Daten tun. Diese Eigenschaft des Algorithmus „Horner-Schema“ bezeichnet man als seine **numerische Stabilität** oder auch **Gutartigkeit**. Nicht jeder Algorithmus ist numerisch stabil.

Beispiel(e) 8.7
Für $p > 0$ und $q > 0$ sei die betragsgrößte Nullstelle des Polynoms $x^2 + 2px - q$ zu bestimmen, also $\lambda = \sqrt{p^2 + q} - p$. Dazu werden zwei Algorithmen betrachtet.

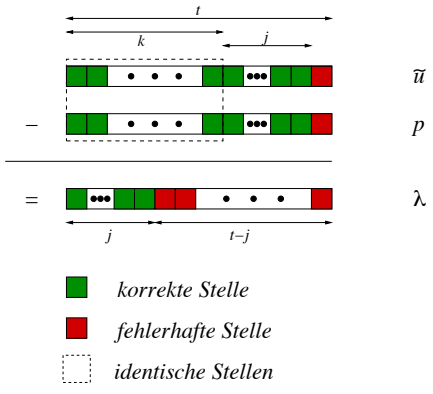
Algorithmus 1 bestehe in der direkten Berechnung von $\lambda = \sqrt{p^2 + q} - p$, also

$$\begin{aligned} s &= p^2 \\ t &= s + q \\ u &= \sqrt{t} \\ \lambda &= u - p. \end{aligned}$$

Algorithmus 2 bestehe in der mathematisch äquivalenten Berechnung $\lambda = q/(\sqrt{p^2 + q} + p)$, also

$$\begin{aligned} s &= p^2 \\ t &= s + q \\ u &= \sqrt{t} \\ v &= u + p \\ \lambda &= q/v. \end{aligned}$$

Eine Analyse zeigt, dass Algorithmus 2 stabil ist, nicht aber Algorithmus 1. Wenn nämlich $p \gg q$, dann ist $\sqrt{p^2 + q} \approx p$. Auch wenn das Zwischenergebnis $\tilde{u}$ mit großer Genauigkeit berechnet sein mag, muss man von einem Fehler wenigstens in der letzten Stelle ausgehen: $\tilde{u} - u = u(1+\varepsilon)$ mit $\varepsilon = \mathcal{O}(\varepsilon_{mach})$. Alle führenden (und korrekten) Stellen von $\tilde{u}$, die mit p übereinstimmen, werden jedoch durch die Subtraktion $\tilde{u} - p$ vernichtet (*cancellation*)!



Es bleibt freilich weiterhin die Frage bestehen, wie stark sich Fehler bzw. Unsicherheiten in den Eingabedaten auf das Ergebnis einer Berechnung auswirken. Dazu die folgende

Definition 8.8 (Kondition eines Problems, Konditionszahlen)
Seien $\mathbf{x} \in \mathbb{D} \subseteq \mathbb{R}^n$ ein Vektor von Eingabedaten und $\mathbf{y} \in \mathbb{R}^m$ ein Vektor von Ergebnissen gegeben durch eine Funktion $\mathbf{p} : \mathbb{D} \to \mathbb{R}^m, \quad \mathbf{x} \mapsto \mathbf{y} = \mathbf{p}(\mathbf{x})$, $m, n \in \mathbb{N}$, so heißt das Problem repräsentiert durch die Abbildung $\mathbf{p}$ gut konditioniert, falls kleine Änderungen $\delta\mathbf{x}$ in den Eingabedaten zu kleinen Änderungen $\delta\mathbf{y}$ in den Resultaten führen ($\mathbf{y} + \delta\mathbf{y} = \mathbf{p}(\mathbf{x} + \delta\mathbf{x})$). Eine Spezifikation des Begriffs „klein“ hängt vom Problem (Skalierung) ab. Ist jede Komponente p_i von $\mathbf{p}$ nach jeder Variablen x_j differenzierbar (mit der Ableitung $(p_i)_{x_j}$), so heißen für alle i, j die Zahlen

$$\left|(p_i)_{x_j}(\mathbf{x})\right| \text{ bzw. } \|\mathbf{Jp}(\mathbf{x})\| := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left((p_i)_{x_j}(\mathbf{x})\right)^2} \quad \text{Konditionszahlen.} \tag{8.27}$$

Die Konditionszahlen sind ein Maß für die Kondition des Problems. Es ist sehr wichtig festzuhalten, dass die Kondition ein Attribut eines Problems und nicht eines numerischen Verfahrens zur Lösung dieses Problems ist. Ist ein Problem schlecht konditioniert, so heißt dies, dass man bei **jedem** Algorithmus zur numerischen Lösung des Problems damit rechnen muss, dass Rundungsfehler das Ergebnis völlig verfälschen. Rundungsfehler treten ja im allgemeinen immer auf und werden als kleine Änderungen der Eingabedaten interpretiert. Diese veränderten Eingabedaten führen aber wegen der schlechten Kondition des Problems zu stark veränderten Resultaten.

Beispiel(e) 8.9
Sei $\mathbf{A}$ die symmetrische 12×12 Matrix mit

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & & & \cdots & 0 \\ 1 & 2 & 1 & 0 & & \cdots & 0 \\ 0 & 1 & 3 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 10 & 1 & 0 \\ 0 & \ldots & & 0 & 1 & 11 & 1 \\ 0 & \ldots & & & 0 & 1 & 12 \end{pmatrix} \tag{8.28}$$

Berechnet man nun die Eigenwerte dieser Matrix als Nullstellen des charakteristischen Polynoms **exakt**, wobei die Koeffizienten des charakteristischen Polynoms erstens exakt berechnet werden (Fall 1) und zweitens im IEEE single precision Format mit rd_*-Rundung (Fall 2), so erhält man:

EW	Fall 1	Fall 2	EW	Fall 1	Fall 2
λ_1	0,253805817...	0,253805802...	λ_7	7,000007953...	7,349137838...
λ_2	1,789321352...	1,789342507...	λ_8	8,000225676...	0,827589263...
λ_3	2,961058880...	2,960859724...	λ_9	9,003952002...	9,696317177...
λ_4	3,996047997...	3,988928696...	λ_{10}	10,038941119...	0,821801891...
λ_5	4,999774323...	5,132691281...	λ_{11}	11,210678647...	11,453800408...
λ_6	5,999992046...	5,615392728...	λ_{12}	12,746194182...	12,714268818...

Eigenwerte über das Ersatzproblem „Nullstellen des charakteristischen Polynoms“ zu berechnen ist also schlecht konditioniert, da die Berechnung der Nullstellen von Polynomen schlecht konditioniert ist.

Definition 8.8 ist im konkreten Fall oft nicht leicht handhabbar. Betrachtet man etwa das nume-

rische Problem, ein lineares Gleichungssystem

$$Ax = b, \quad A \in \mathbb{R}^{n,n} \text{ invertierbar und } b \in \mathbb{R}^n$$

zu lösen, so ist im Sinn der Definition 8.8 die Abbildung $(A, b) \mapsto x = A^{-1}b$ zu betrachten und nach den Eingangsgrößen $a_{i,j}$ und b_i (den Komponenten von A und b) abzuleiten. Wir wählen einen anderen Zugang. Kondition bedeutet die Empfindlichkeit der Lösung eines linearen Gleichungssystems gegenüber Änderungen der Matrix-Koeffizienten und der rechten Seite. Also muss die Lösung x von $Ax = b$ verglichen werden mit der Lösung $x+\delta x$ von $(A+\delta A)(x+\delta x) = b+\delta b$. Wir setzen voraus, dass A invertierbar und die Störung δA so klein ist, dass auch $A+\delta A$ invertierbar bleibt. Mit der in Definition 8.8 eingeführten „Matrixnorm“

$$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}^2} \quad \text{sowie mit der Euklidischen Norm} \quad \|b\| := \sqrt{\sum_{i=1}^{n} b_i^2}$$

für $A \in \mathbb{R}^{m,n}$ und $b \in \mathbb{R}^n$ lässt sich zeigen: es sei $A \in \mathbb{R}^{n,n}$ invertierbar und $\delta A \in \mathbb{R}^{n,n}$. Dann gilt

$$\|\delta A\| < 1/\|A^{-1}\| \quad \Longrightarrow \quad (A + \delta A) \quad \text{invertierbar.} \tag{8.29}$$

Außerdem ist

$$\|Ab\| \le \|A\| \cdot \|b\| \, .$$

Seien nun also $Ax = b$ und $(A + \delta A)(x + \delta x) = b + \delta b$ und sei $\|\delta A\| < 1/\|A^{-1}\|$. Daraus folgt

$$\begin{aligned} A\delta x &= b + \delta b - (Ax + \delta A \cdot x + \delta A \cdot \delta x) = \delta b - \delta A \cdot x - \delta A \cdot \delta x, \\ \delta x &= A^{-1}(\delta b - \delta A \cdot x - \delta A \cdot \delta x), \\ \|\delta x\| &\le \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|} \cdot (\|\delta b\| + \|\delta A\| \cdot \|x\|) \, . \end{aligned}$$

Mit dieser Schranke für die absolute Änderung δx der Lösung bei Änderung der Daten lässt sich auch eine Schranke für die relative Änderung angeben. Mit der **Konditionszahl**

$$\kappa = \kappa(A) = \|A\| \cdot \|A^{-1}\| \tag{8.30}$$

erhält man

$$\boxed{\begin{aligned} \frac{\|\delta A\|}{\|A\|} &\le \varepsilon \quad \text{und} \quad \frac{\|\delta b\|}{\|b\|} \le \varepsilon \quad \Longrightarrow \\ \frac{\|\delta x\|}{\|x\|} &\le \frac{\varepsilon\kappa}{1 - \varepsilon\kappa} \cdot \left(\frac{\|b\|}{\|A\| \cdot \|x\|} + 1 \right) \, . \end{aligned}} \tag{8.31}$$

(man beachte hierbei $\|b\| = \|Ax\| \le \|A\| \cdot \|x\|$).

Die Interpretation von (8.31) ist die folgende. Steht ε für die relative Unsicherheit in den Eingangsdaten, dann verstärkt sich diese um den Faktor $\kappa = \kappa(A)$. Wenn $\kappa\varepsilon$ in die Nähe der Größe 1 gerät, kann es zu völlig falschen Ergebnissen kommen, unabhängig vom verwendeten numerischen Verfahren.

8.2 Lineare Gleichungssysteme

Dies ist das wichtigste Problem der Numerik und immer noch Gegenstand aktiver Forschung. *Die* Standardmethode zur Lösung des linearen Gleichungssystems

$$Ax = b, \quad A \in \mathbb{R}^{n,n}, \quad b \in \mathbb{R}^n$$

ist der Gauß-Algorithmus mit Zeilenvertauschungen, wie beschrieben in Abschnitt 2.1. Die erreichbare Genauigkeit wird durch die Konditionszahl von A abgeschätzt, siehe (8.31). Für die Stabilität kommt es entscheidend auf die Zeilenvertauschungen an, die man beim Gauß-Algorithmus verwendet. Es genügt nicht, bei der Vorwärtselimination zu erreichen, dass nach der Zeilenvertauschung $\diamond = \alpha_{11} \neq 0$, vergleiche (2.26). Meistens tauscht man eine Zeile nach oben, so dass anschließend $\alpha_{11} \geq \alpha_{i1}$ für $i \geq 1$. Dies nennt man „Spaltenpivotsuche“. In allen weiteren Eliminationsschritten hält man es genauso: immer das maximale Element der führenden Spalte der Restmatrix nach oben tauschen.

Da man die exakte Lösung nicht kennt, kann man in der Regel den Fehler $\|x - \tilde{x}\|/\|x\|$ des berechneten Ergebnisses $\tilde{x}$ nicht abschätzen. Man kann aber eine „Einsetzprobe“ machen und das sogenannte $\mathbf{Residuum}$ berechnen:

$$r = b - A\tilde{x} \ .$$

Wenn r ein Vektor mit kleinen Komponenten ist, dann kann man wegen

$$A\tilde{x} = b + r =: \tilde{b}$$

sagen: das berechnete $\tilde{x}$ ist die exakte Lösung eines linearen Gleichungssystems $A\tilde{x} = \tilde{b}$, bei dem lediglich die rechte Seite (ein Eingabedatum) leicht abgeändert wurde. Es ist also ein Ergebnis, wie es im Rahmen der zu unterstellenden Unsicherheit in den Eingabedaten jederzeit möglich ist, auch wenn man hypothetisch gar keine Rundungsfehler gemacht hätte. Mehr kann man nicht erwarten.

Es gibt leider Ausnahmen, aber wenn man die Spaltenpivotsuche durchführt, dann liefert der Gauß-Algorithmus fast immer eine Lösung $\tilde{x}$ mit kleinem Residuum, wie gewünscht.

8.3 Nichtlineare Gleichungssysteme

Gesucht wird ein $\bar{x} \in \mathbb{R}^n$ das die n Gleichungen

$$\begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} = 0 \in \mathbb{R}^n.$$

erfüllt.

Es handelt sich um ein sehr schwieriges Problem. Auch hier gibt es eine Standardmethode zur Lösung, das Newton-Verfahren, wie schon besprochen in den Formeln (6.39) ff. Das Newton-Verfahren basiert auf der Idee einer iterativen Linearisierung und hat zwei große Nachteile

(a): Es muss nicht immer gegen eine Lösung konvergieren, selbst wenn das nichtlineare Gleichungssystem eine Lösung hat.

(b): Es ist sehr aufwändig. In jedem Schritt des Verfahrens sind n^2 Ableitungen zu berechnen und ein lineares Gleichungssystem zu lösen.

Diese beiden Schwierigkeiten treten ganz analog bei einem ähnlichen, aber noch wichtigeren Problem der Numerik auf, nämlich dem der Optimierung. Auch die Ideen, den beiden oben genannten Problemen entgegen zu wirken sind ähnlich. Wir beschränken uns für diese Vorlesung auf eine kurz Erörterung der Optimierung in Abschnitt 8.8.

8.4 Polynominterpolation

Betrachtet man eine stetige Funktion $f : [a, b] \to \mathbb{R}$, $x \mapsto f(x)$, und Stützstellen $x_1, \ldots, x_n \in [a, b]$ mit $x_1 < \ldots < x_n$, so besteht das Interpolationsproblem in der Bestimmung eines Polynoms (n-1)-ter Ordnung $p : \mathbb{R} \to \mathbb{R}$, $x \mapsto \sum_{i=0}^{n-1} a_i x^i$ derart, dass:

$$p(x_i) = f(x_i), \text{ für alle } \quad i = 1, \ldots, n. \tag{8.32}$$

Der Sinn dieser Aufgabe besteht darin, eine Approximation von f an einer Stelle $\bar{x} \in [a, b]$, $\bar{x} \neq x_i$, $i = 1, \ldots, n$, durch $p(\bar{x})$ zu erhalten, falls von f nur die Funktionswerte $f(x_1), \ldots, f(x_n)$ bekannt sind. Als Bedingungen an die Koeffizienten $a_0, \ldots, a_{n-1}$ des Polynoms p erhält man:

$$\begin{aligned} a_0 + a_1 x_1 + \ldots + a_{n-1} x_1^{n-1} &= f(x_1) && (8.33)\\ a_0 + a_1 x_2 + \ldots + a_{n-1} x_2^{n-1} &= f(x_2) && (8.34)\\ &\vdots && (8.35)\\ a_0 + a_1 x_n + \ldots + a_{n-1} x_n^{n-1} &= f(x_n) && (8.36) \end{aligned}$$

beziehungsweise:

$$\underbrace{\begin{pmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \tag{8.37}$$

Dieses lineare Gleichungssystem hat immer genau eine Lösung, da die Matrix $\mathbf{A}$ den Rang n besitzt. Es gibt also stets genau ein Polynom p, das die gestellte Aufgabe löst.

Wesentlich besser geeignet als der obige ist der Newton-Ansatz (3.27) für das gesuchte Polynom, der auf einen gestaffeltes Gleichungssystem führt, welches sehr einfach aufgelöst werden kann, vergleiche (3.29).

Im folgenden untersuchen wir die Frage, wie dieses Polynom am besten darzustellen ist, um mit dem Rechner möglichst effizient einen Interpolationswert $p(\bar{x})$ zu erhalten. Eine Darstellung des Interpolanten p ist durch die Lagrange-Polynome $L_k : \mathbb{R} \to \mathbb{R}$, $k = 1, \ldots, n$:

$$L_k(x) = \begin{cases} 1 & \text{für} \quad n = 1 \\ \frac{x-x_1}{x_k-x_1} \cdot \ldots \cdot \frac{x-x_{k-1}}{x_k-x_{k-1}} \cdot \frac{x-x_{k+1}}{x_k-x_{k+1}} \cdot \ldots \cdot \frac{x-x_n}{x_k-x_n} & \text{für} \quad n > 1 \end{cases} \tag{8.38}$$

mit

$$p(x) = \sum_{k=1}^{n} f(x_k) L_k(x) \tag{8.39}$$

gegeben, denn es gilt für $n > 1$:

$$L_k(x_i) = \begin{cases} 0 & \text{für} \quad i \neq k \\ 1 & \text{für} \quad i = k \end{cases}, \quad i = 1, \ldots, n. \tag{8.40}$$

Somit ist es möglich, p an einer Stelle x auszuwerten, ohne die Koeffizienten $a_0, \ldots, a_{n-1}$ explizit zu berechnen.
Die Interpolationsaufgabe hat als Eingabedaten die Größen

$$x_1, \ldots, x_n, y_1 = f(x_1), \ldots, y_n = f(x_n) \tag{8.41}$$

und die zu interpolierende Stelle $\bar{x}$, und als Resultat den Wert $p(\bar{x})$. Für die Konditionszahlen bezüglich $y_1, \ldots, y_n$ gilt:

$$\frac{\partial \left(\sum_{k=1}^{n} y_k L_k(x) \right)}{\partial y_i}(\bar{x}) = L_i(\bar{x}),\ i = 1, \ldots, n. \tag{8.42}$$

Sind zum Beispiel $n = 101$ und die Stützstellen $x_1, \ldots, x_{101}$ äquidistant, so ergibt sich $L_{51}(x) \approx 10^{26}$ für $\bar{x}$ zwischen x_1 und x_2 bzw. zwischen x_{100} und x_{101}. Polynominterpolation ist daher für große n unbrauchbar.

8.5 Polynom-Splines

Im letzten Abschnitt wurde offensichtlich, dass Polynominterpolation für eine große Zahl von Stützstellen völlig unbrauchbar ist. Auf der anderen Seite stehen häufig sehr große Datenmengen zur Verfügung. Der Wunsch, einerseits große Datenmengen verarbeiten zu können, andererseits die vorteilhaften Eigenschaften von Polynomen ausnutzen zu können, führt auf die Idee, für jedes Teilintervall $[x_i, x_{i+1}]$ zweier benachbarter Stützstellen ein eigenes Polynom zu verwenden.

Definition 8.10 (Polynom-Spline)
Seien $x_1 < \ldots < x_n \in \mathbb{R}$ und $m \in \mathbb{N}$. Eine Funktion $s : [x_1, x_n] \to \mathbb{R}$ heißt Polynom-Spline der Ordnung m (bzw. vom Grad (m-1)), falls

$$s(x) = p_i(x) \quad \text{für} \quad x_i \leq x < x_{i+1}, \tag{8.43}$$

und falls s (m-2)-mal stetig differenzierbar ist (entfällt bei $m = 1$, Stetigkeit bei $m = 2$), wobei p_i für $i = 1, \ldots, n-1$ ein Polynom vom Grad $(m-1)$ ist.
$m = 2$: Polygone,
$m = 3$: Quadratische Splines,
$m = 4$: Kubische Splines.

Für die Interpolation einer Funktion $f : [a, b] \to \mathbb{R}$ mit den Stützstellen $a \leq x_1 < x_2 < \ldots < x_n \leq b$ und den Funktionswerten $f(x_1), \ldots, f(x_n)$ fordert man natürlich:

$$s(x_i) = p_i(x_i) = f(x_i), \quad i = 1, \ldots, n-1 \tag{8.44}$$

und

$$s(x_n) = p_{n-1}(x_n) = f(x_n). \tag{8.45}$$

Der für die Praxis häufigste Fall ist $m = 4$, also kubische Splines. Zu n Stützstellen $x_1 < \ldots < x_n$ hat man $(n-1)$ Polynome dritten Grades $p_j : \mathbb{R} \to \mathbb{R}$, $x \mapsto a_j x^3 + b_j x^2 + c_j x + d_j$, zu finden mit:

$$\begin{aligned}
p_i(x_i) &= f(x_i), \quad i = 1, \ldots, n-1 && (8.46)\\
p_i(x_{i+1}) &= f(x_{i+1}), \quad i = 1, \ldots, n-1 && (8.47)\\
p_i'(x_{i+1}) &= p_{i+1}'(x_{i+1}), \quad i = 1, \ldots, n-2 && (8.48)\\
p_i''(x_{i+1}) &= p_{i+1}''(x_{i+1}), \quad i = 1, \ldots, n-2. && (8.49)
\end{aligned}$$

Die ersten beiden Bedingungen dienen dazu, dass die zu $p_1, \ldots, p_{n-1}$ gehörige Splinefunktion $s : [x_1, x_n] \to \mathbb{R}$,

$$x \mapsto \begin{cases} p_i(x) & \text{für} \quad x \in [x_i, x_{i+1}) \\ p_{n-1}(x_n) & \text{für} \quad x = x_n \end{cases} \tag{8.50}$$

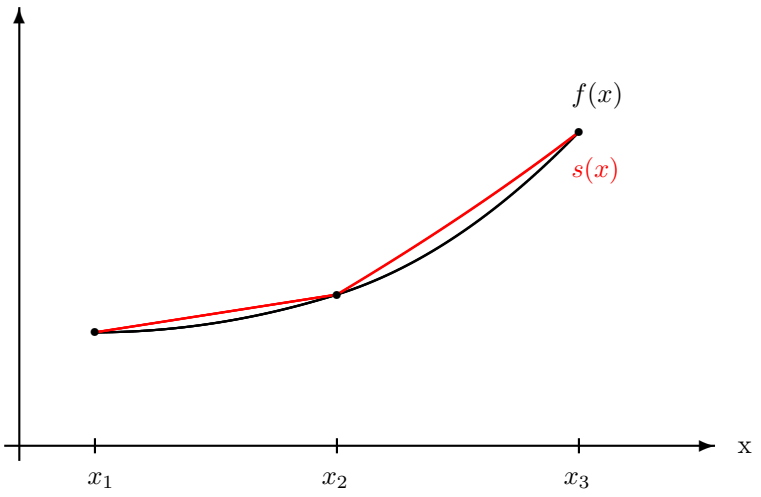


Abbildung 8.1: Interpolation durch Polygone

die Eigenschaft $s(x_i) = f(x_i)$, $i = 1, \ldots, n$, besitzt. Die dritte und die vierte Bedingung garantieren, dass die Funktion s zweimal stetig differenzierbar ist. Insgesamt hat man also $4n - 6$ Gleichungen für $4n - 4$ Unbekannte. Es können also für s noch zwei Bedingungen hinzugefügt werden. Häufig wählt man:

Typ 1: $s''(x_1) = s''(x_n) = 0$ (einseitige Grenzwerte)

Typ 2: $s'(x_1) = s'(x_n), \quad s''(x_1) = s''(x_n).$ (einseitige Grenzwerte)

Insgesamt erhält man somit:

$$\left.\begin{aligned} a_1x_1^3 + b_1x_1^2 + c_1x_1 + d_1 &= f(x_1) \\ &\vdots \\ a_{n-1}x_{n-1}^3 + b_{n-1}x_{n-1}^2 + c_{n-1}x_{n-1} + d_{n-1} &= f(x_{n-1}) \end{aligned}\right\} \tag{8.51}$$

$$\left.\begin{aligned} a_1x_2^3 + b_1x_2^2 + c_1x_2 + d_1 &= f(x_2) \\ &\vdots \\ a_{n-1}x_n^3 + b_{n-1}x_n^2 + c_{n-1}x_n + d_{n-1} &= f(x_n) \end{aligned}\right\} \tag{8.52}$$

$$\left.\begin{aligned} 3a_1x_2^2 + 2b_1x_2 + c_1 &= 3a_2x_2^2 + 2b_2x_2 + c_2 \\ &\vdots \\ 3a_{n-2}x_{n-1}^2 + 2b_{n-2}x_{n-1} + c_{n-2} &= 3a_{n-1}x_{n-1}^2 + 2b_{n-1}x_{n-1} + c_{n-1} \end{aligned}\right\} \tag{8.53}$$

$$\left.\begin{aligned} 6a_1x_2 + 2b_1 &= 6a_2x_2 + 2b_2 \\ &\vdots \\ 6a_{n-2}x_{n-1} + 2b_{n-2} &= 6a_{n-1}x_{n-1} + 2b_{n-1} \end{aligned}\right\} \tag{8.54}$$

$$\left.\begin{aligned} 6a_1x_1 + 2b_1 &= 0 \\ 6a_{n-1}x_n + 2b_{n-1} &= 0 \end{aligned}\right\} \text{(Typ 1)} \tag{8.55}$$

$$\left.\begin{aligned} 3a_1x_1^2 + 2b_1x_1 + c_1 &= 3a_{n-1}x_n^2 + 2b_{n-1}x_n + c_{n-1} \\ 6a_1x_1 + 2b_1 &= 6a_{n-1}x_n + 2b_{n-1} \end{aligned}\right\} \text{(Typ 2)} \tag{8.56}$$

Die linearen Gleichungssysteme (für Typ 1 bzw. Typ 2) in

$$a_1, \ldots, a_{n-1}, b_1, \ldots, b_{n-1}, c_1, \ldots, c_{n-1}, d_1, \ldots, d_{n-1} \tag{8.57}$$

sind eindeutig lösbar und somit existiert genau eine Splinefunktion mit $m = 4$. Es läßt sich mit speziellen Hilfsmitteln aus der Numerischen Mathematik zeigen, dass die Interpolation mit kubischen Splines gerade für äquidistante Stützstellen unabhängig von n sehr gut konditioniert ist.
In der Praxis betrachtet man im allgemeinen nicht das obige lineare Gleichungssystem, sondern man wählt eine spezielle Darstellung der kubischen Polynome. Die Anzahl der zu berechnenden Parameter kann dadurch zum Beispiel bei äquidistanten Stützstellen sehr einfach auf n reduziert werden, wobei das entsprechende lineare Gleichungssystem eine sehr einfache Form besitzt.

8.6 Numerische Quadratur

Unter Quadratur versteht man die Berechnung eines bestimmten Integrals

$$I(f) = \int_a^b f(x)dx. \tag{8.58}$$

Häufig ist eine (analytische) Berechnung von $I(f)$ nicht möglich, so dass man auf eine numerische Approximation ausweichen muss. Da viele numerische Schemata zur Berechnung von $I(f)$ (bzw. deren Analyse) auch Ableitungen von f verwenden, ist im allgemeinen eine Aufbereitung der zu berechnenden Integrale nötig.

Beispiel(e) 8.11

- Zerlegung:

$$\int_{-2}^{1} |x|dx = \int_{-2}^{0} (-x)dx + \int_{0}^{1} xdx \tag{8.59}$$

  (die Funktion $f : [-1, 1] \to \mathbb{R}, x \mapsto |x|$ ist nur stetig, aber die Funktionen $h : [-1, 0] \to \mathbb{R}$, $x \mapsto -x$ und $k : [0, 1] \to \mathbb{R}, x \mapsto x$ sind beliebig oft stetig differenzierbar.)

- Substitution:

$$\int_0^{\pi} \underbrace{\frac{\sin(x)}{\sqrt{x}}}_{f(x)} dx = \int_0^{\sqrt{\pi}} \underbrace{2\sin(t^2)}_{g(t)} dt \tag{8.60}$$

  (Substitution: $x = t^2$, die Funktion g ist beliebig oft stetig differenzierbar; der einseitige Differenzenquotient für f an der Stelle $x_0 = 0$ existiert nicht.)

In einer Näherungsformel

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} g_i f(\rho_i) = \tilde{I}(f) \tag{8.61}$$

werden die reellen g_i, $i = 1, \ldots, n$, als Gewichte und die Stellen

$$a \leq \rho_i \leq b \tag{8.62}$$

als Auswertungsstellen bezeichnet. Der Fehler (oder das Restglied) ergibt sich zu

$$R(f) = \sum_{i=1}^{n} g_i f(\rho_i) - \int_a^b f(x)dx. \tag{8.63}$$

Die Auswertungsstellen $\rho_1, \ldots, \rho_n$ werden im allgemeinen in Abhängigkeit von einer Diskretisierung

$$a = x_0 < \ldots < x_m = b \tag{8.64}$$

mit Stützstellen $x_0, \ldots, x_m$ gewählt.

- **Rechtecksregel:**

$$\tilde{I}(f) = \sum_{i=1}^{m} (x_i - x_{i-1}) f\left(\frac{x_i + x_{i-1}}{2}\right). \tag{8.65}$$

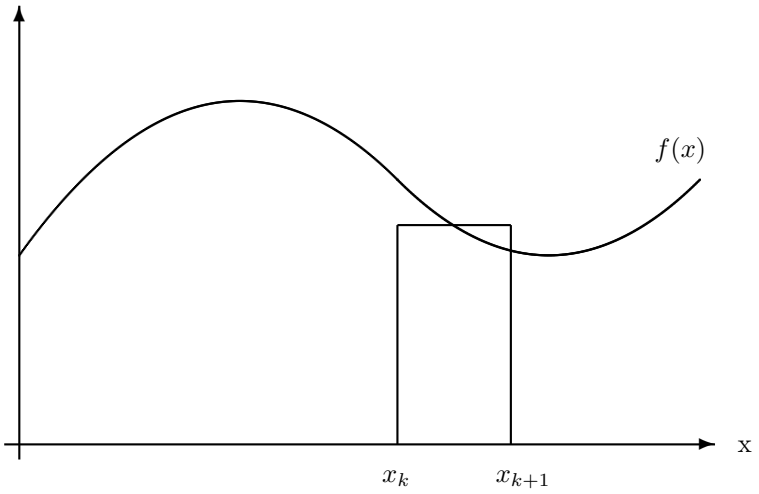


Abbildung 8.2: Rechtecksregel

- **Trapezregel:**

$$\tilde{I}(f) = \sum_{i=1}^{m} (x_i - x_{i-1}) \frac{f(x_i) + f(x_{i-1})}{2} = \tag{8.66}$$

$$= \sum_{i=1}^{m} \frac{x_i - x_{i-1}}{2} f(x_i) + \sum_{i=m+1}^{2m} \frac{x_{i-m} - x_{i-m-1}}{2} f(x_{i-m-1}). \tag{8.67}$$

- **Fassregel (J. Kepler: Nova stereometrica doliorum vinariorum, Linz 1615):**

$$\tilde{I}(f) = \sum_{i=1}^{m} (x_i - x_{i-1}) \frac{f(x_i) + 4f(\frac{x_i + x_{i-1}}{2}) + f(x_{i-1})}{6} \tag{8.68}$$

Kepler gewann die Fassregel aus einer Mischung aus Rechtecks- und Trapezregel.

Um die Güte einer numerischen Quadraturformel bewerten zu können, betrachtet man eine äquidistante Zerlegung

$$a = x_0 < \ldots < x_m = b \quad \text{mit} \quad h := x_i - x_{i-1} \tag{8.69}$$

und berechnet den Integrationsfehler in Abhängigkeit von h.

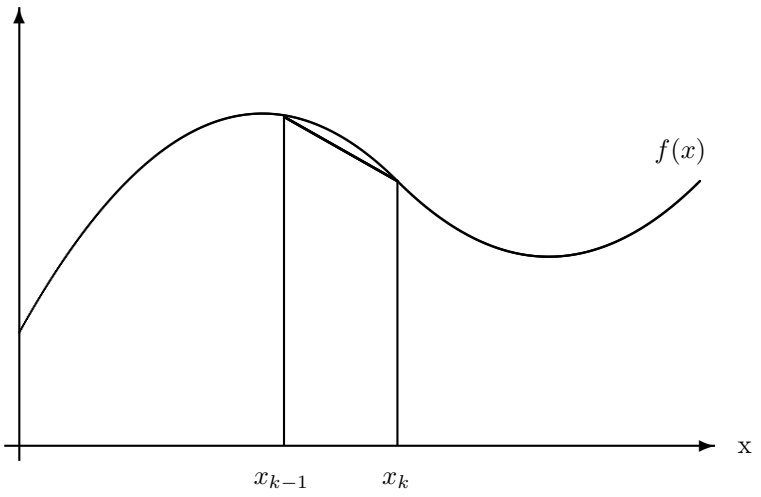


Abbildung 8.3: Trapez-Regel

- Rechteckregel: Sei f zweimal stetig differenzierbar:

$$\begin{aligned}
\tilde{I}(f) - I(f) &= \sum_{i=1}^{m} h \cdot f\left(\frac{x_i + x_{i-1}}{2}\right) - \int_a^b f(x)dx && (8.70)\\
&= \sum_{i=1}^{m} h \cdot f\left(\frac{x_i + x_{i-1}}{2}\right) - \sum_{i=1}^{m} \int_0^h f(x_{i-1} + t)dt && (8.71)\\
&= h\sum_{i=1}^{m} f\left(x_{i-1} + \frac{h}{2}\right) - \sum_{i=1}^{m} \int_0^h f(x_{i-1} + t)dt && (8.72)
\end{aligned}$$

Da (dank Taylorentwicklung) gilt:

$$\begin{aligned}
f(x) &= f\left(x_{i-1} + \frac{h}{2}\right) + f'\left(x_{i-1} + \frac{h}{2}\right)\left(x - \left(x_{i-1} + \frac{h}{2}\right)\right) + && (8.73)\\
&+ \frac{1}{2}f''(\xi_i)\left(x - \left(x_{i-1} + \frac{h}{2}\right)\right)^2, && (8.74)
\end{aligned}$$

erhält man:

$$\begin{aligned}
\tilde{I}(f) - I(f) &= h\sum_{i=1}^{m} f\left(x_{i-1} + \frac{h}{2}\right) -\\
&\quad -\sum_{i=1}^{m} \int_0^h \left(f\left(x_{i-1} + \frac{h}{2}\right) + f'\left(x_{i-1} + \frac{h}{2}\right)\left(t - \frac{h}{2}\right) + \frac{1}{2}f''(\xi_i)\left(t - \frac{h}{2}\right)^2\right) dt\\
&= -\sum_{i=1}^{m} \int_0^h \frac{1}{2}f''(\xi_i)\left(t - \frac{h}{2}\right)^2 dt\\
&= -(b-a) \cdot \frac{h^2}{24} \cdot f''(\hat{\xi}), \quad \hat{\xi} \in [a,b].
\end{aligned}$$

- Trapezregel: Sei f zweimal stetig differenzierbar:

$$\tilde{I}(f) - I(f) = (b-a) \cdot \frac{h^2}{12} \cdot f''(\hat{\xi}), \quad \hat{\xi} \in [a,b]. \qquad (8.75)$$

(Rechnung schwierig)

- Fassregel: Die Fassregel von Kepler kann man auf „moderne“ Art herleiten, wenn man in jedem Integral in

$$I(f) = \sum_{i=1}^{m} \int_0^h f(x_{i-1} + t)dt \tag{8.76}$$

die Funktionen $g_{i-1} : [0, h] \to \mathbb{R}, t \mapsto f(x_{i-1} + t)$ durch den quadratischen Interpolanten an den Stützstellen $0, \frac{h}{2}, h$ ersetzt. Dieser Zugang erlaubt für viermal stetig differenzierbare Funktionen f die Fehlerabschätzung:

$$\tilde{I}(f) - I(f) = (b - a) \cdot \frac{h^4}{2880} \cdot f''''(\hat{\xi}), \quad \hat{\xi} \in [a, b]. \tag{8.77}$$

Im folgenden kommen wir auf die Trapezregel zurück. Es gilt die Trapezsummenformel:

$$\begin{aligned} \tilde{I}(f) &= h \sum_{i=1}^{m} \frac{f(x_i) + f(x_{i-1})}{2} && (8.78)\\ &= h \left(\frac{1}{2} f(a) + f(x_1) + \ldots + f(x_{m-1}) + \frac{1}{2} f(b) \right) && (8.79)\\ &= h \left(\frac{1}{2} f(a) + f(a + h) + \ldots + f(a + (m - 1)h) + \frac{1}{2} f(b) \right). && (8.80) \end{aligned}$$

Ist nun f $(2k)$-mal stetig differenzierbar für $k \in \mathbb{N}$, so läßt sich die Trapezsummenformel $\tilde{I}(f)$ folgendermaßen in Abhängigkeit von h darstellen (der Beweis erfordert viel Mathematik)

$$\tilde{I}(f) = \underbrace{T(h) = \tau_0 + \tau_1 h^2 + \ldots + \tau_{k-1} h^{2k-2} + \tau_k(h) h^{2k}}_{\text{Euler-Maclaurinsche Summenformel}}, \tag{8.81}$$

wobei $\tau_0, \ldots, \tau_{k-1}$ nicht von h abhängen,

$$\tau_0 = \int_a^b f(x)dx \tag{8.82}$$

gilt und $\tau_k(h)$ eine in h beschränkte Funktion ist. Somit erhalten wir für den Fehler:

$$R(f)(h) = T(h) - \int_a^b f(x)dx = \tau_1 h^2 + \ldots + \tau_{k-1} h^{2k-2} + \tau_k(h) h^{2k} \tag{8.83}$$

Verwendet man nun zwei verschiedene äquidistante Diskretisierungen von $[a, b]$, einmal mit h_1 als Abstand und einmal mit h_2, so erhält man für die jeweiligen Trapezsummen:

$$\begin{aligned} T(h_1) &= \tau_0 + \tau_1 h_1^2 + \ldots + \tau_{k-1} h_1^{2k-2} + \tau_k(h_1) h_1^{2k} && (8.84)\\ T(h_2) &= \tau_0 + \tau_1 h_2^2 + \ldots + \tau_{k-1} h_2^{2k-2} + \tau_k(h_2) h_2^{2k} && (8.85) \end{aligned}$$

Eliminiert man aus diesen beiden linearen Gleichungen in $\tau_1, \ldots, \tau_{k-1}$ die Variable τ_1, so erhält man:

$$\begin{aligned} \hat{I}(f) = \frac{h_1^2 T(h_2) - h_2^2 T(h_1)}{h_1^2 - h_2^2} &= \tau_0 + 0 - \tau_2 h_1^2 h_2^2 + \ldots && (8.86)\\ &= \int_a^b f(x)dx + O(h_1^2 h_2^2). && (8.87) \end{aligned}$$

Analog dazu kann man für $l \leq (k-1)$ Schrittweiten $h_1, \ldots, h_l$ durch

$$\begin{aligned} T(h_1) &= \tau_0 + \tau_1 h_1^2 + \ldots + \tau_{k-1} h_1^{2k-2} + \tau_k(h_1) h_1^{2k} && (8.88)\\ T(h_2) &= \tau_0 + \tau_1 h_2^2 + \ldots + \tau_{k-1} h_2^{2k-2} + \tau_k(h_2) h_2^{2k} && (8.89)\\ &\vdots && (8.90)\\ T(h_l) &= \tau_0 + \tau_1 h_l^2 + \ldots + \tau_{k-1} h_l^{2k-2} + \tau_k(h_l) h_l^{2k} && (8.91) \end{aligned}$$

die Größen $\tau_1, \ldots, \tau_{l-1}$ eliminieren und erhält somit ein numerisches Schema der Ordnung $O(h_1^2 \cdot \ldots \cdot h_l^2)$.

Für das Integral

$$\int_1^2 \frac{1}{x} dx = \ln(2) \approx 0.6931471806 \tag{8.92}$$

erhält man:

h_i	$T(h_i)$	τ_1 eliminiert	τ_1, τ_2 eliminiert	τ_1, τ_2, τ_3 eliminiert
1	0.7500000000			
1/2	0.7083333333	0.6944444444		
1/4	0.6970238095	0.6932539683	0.6931746032	
1/8	0.6941218504	0.6931545307	0.6931479015	0.6931474776

8.7 Numerische Lösung von Anfangswertproblemen gew. Differentialgleichungen

Seien für $\mathbf{f} : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$, $(x, \mathbf{z}) \mapsto \mathbf{f}(x, \mathbf{z})$, der Existenz- und Eindeutigkeitssatz (Mathematik II, Satz 4.13) erfüllt. Sei ferner $\mathbf{y} : [a,b] \to \mathbb{R}^n$ die Lösung des Anfangswertproblems

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(a) = \mathbf{y}_0, \tag{8.93}$$

wobei an den Intervallgrenzen jeweils der einseitige Differentialquotient zu wählen ist. Bekanntlich besitzt (8.93) eine äquivalente Integraldarstellung

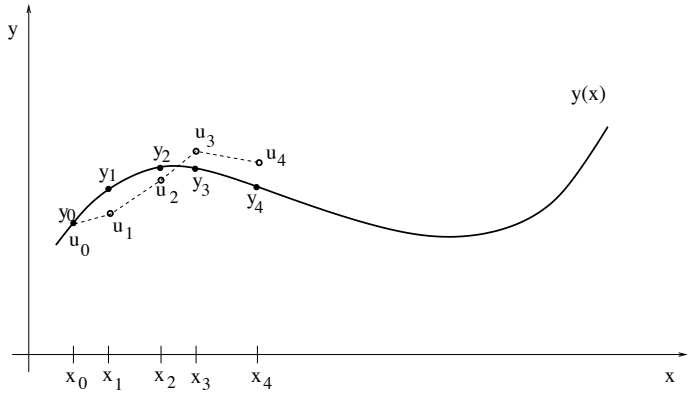
$$\mathbf{y}(x) = \mathbf{y}_0 + \int_{x_0}^{x} \mathbf{f}(t, \mathbf{y}(t)) dt \,, \tag{8.94}$$

die von Vorteil zur Herleitung numerischer Verfahren ist.

Wie bereits aus der Theorie der Anfangswertprobleme gewöhnlicher Differentialgleichungen bekannt ist, sind Probleme dieser Art nur in den seltensten Fällen analytisch lösbar. Daher ist man im allgemeinen auf numerische Verfahren angewiesen. Diese Verfahren verwenden eine **Diskretisierung** mittels eines **Gitters** $a = x_0 < x_1 < \ldots < x_N = b$ des Intervalls $[a,b]$ mit der **Schrittweite** $h_k := x_k - x_{k-1}$, $k = 1, \ldots, N$. Dann wird entweder der Differentialquotient in (8.93) oder das Integral in (8.94) durch eine Näherung ersetzt, die sich auf die Gitterpunkte stützt. Im Folgenden werden die so berechneten Näherungswerte für die exakte Lösung $\mathbf{y}$ an den Stellen x_k mit $\mathbf{u}_k$ bezeichnet, also

$$\begin{aligned} \mathbf{u}_k &\approx \mathbf{y}(x_k) = \mathbf{y}_k \\ \mathbf{u}_k' \equiv \mathbf{f}_k &:= \mathbf{f}(x_k, \mathbf{u}_k) \approx \mathbf{f}(x_k, \mathbf{y}(x_k)) = \mathbf{y}'(x_k) = \mathbf{y}_k' \,. \end{aligned}$$

Die Bezeichnungen $\mathbf{f}_k$ und $\mathbf{u}'_k$ sind gleichwertig — manchmal scheint die eine, manchmal die andere suggestiver. Nachstehend wird eine mögliche Lage der Näherungswerte im Vergleich zu den exakten Werten $\mathbf{y}_k$ skizziert.



Wir geben jetzt zwei konkrete Verfahren an.

- **Euler-Cauchy-Verfahren:**

$$\mathbf{u}_k = \mathbf{u}_{k-1} + h_k \mathbf{f}(x_{k-1}, \mathbf{u}_{k-1}), \quad \mathbf{u}_0 = \mathbf{y}_0 = \mathbf{y}(a). \tag{8.95}$$

Eine Erklärung für dieses Verfahren: ist schon der Näherungswert $\mathbf{u}_{k-1}$ für $\mathbf{y}_{k-1} = \mathbf{y}(x_{k-1})$ bestimmt, betrachtet man die Lösung $\mathbf{z}$ von

$$\mathbf{z}'(x) = \mathbf{f}(x, \mathbf{z}(x)),\ \mathbf{z}(x_{k-1}) = \mathbf{u}_{k-1} \quad \Longleftrightarrow \quad \mathbf{z}(x) = \mathbf{u}_{k-1} + \int_{x_{k-1}}^{x} \mathbf{f}(t, \mathbf{z}(t))dt \tag{8.96}$$

und berechnet eine Näherung $\mathbf{u}_k \approx \mathbf{z}(x_k) \approx \mathbf{y}(x_k)$, indem man das Integral $\int_{x_{k-1}}^{x} \mathbf{f}(t, \mathbf{z}(t))dt$ für $x = x_k$ durch $h_k \mathbf{f}(x_{k-1}, \mathbf{u}_{k-1})$ ersetzt wie in der folgenden Skizze illustriert.

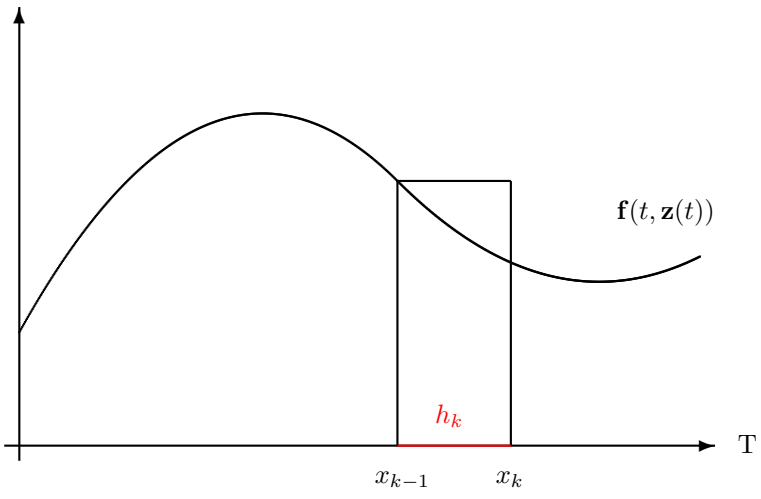


Abbildung 8.4: Euler-Cauchy-Verfahren

Das Euler-Cauchy-Verfahren heißt **explizit**, weil die neu zu berechnende Näherung $\mathbf{u}_k$ in einer expliziten Verfahrensvorschrift (Formel) angegeben ist. Man spricht genauer von einem expliziten **Einschrittverfahren**, weil die neue Näherung nur mithilfe *einer* vorherigen Näherung berechnet wird (nämlich nur mithilfe von $\mathbf{u}_{k-1}$, aber nicht mithilfe von $\mathbf{u}_{k-2}$).

- **Implizites Euler-Verfahren:**

$$\mathbf{u}_k = \mathbf{u}_{k-1} + h_k \mathbf{f}(x_k, \mathbf{u}_k), \quad \mathbf{u}_0 = \mathbf{y}_0 = \mathbf{y}(a). \tag{8.97}$$

Dieses Verfahren heißt **implizit**, weil keine Formel angegeben ist, um die neue Näherung $\mathbf{u}_k$ zu berechnen. Vielmehr wird $\mathbf{u}_k$ indirekt durch ein im allgemeinen nichtlineares Gleichungssystem bestimmt, das erst noch nach dem unbekannten Vektor $\mathbf{u}_k$ aufgelöst werden muss.

Das Verfahren (8.97) lässt sich analog zum Euler-Verfahren mit einer Funktion $\mathbf{z}$ wie in (8.96) herleiten, nur wird jetzt das Integral anders approximiert, siehe die folgende Skizze.

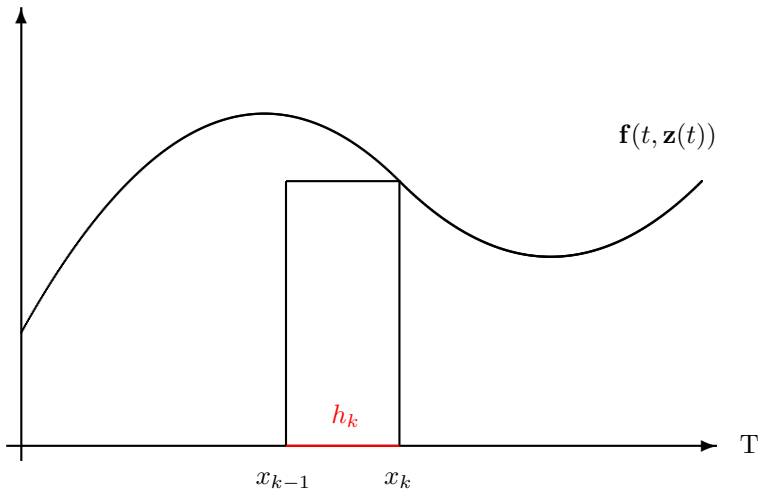


Abbildung 8.5: Implizites Euler-Verfahren

Das implizite Euler-Verfahren soll anhand zweier Beispiele illustriert werden.

Beispiel(e) 8.12

$$y'(x) = y(x) + 1 = f(x, y(x)), \quad y(0) = 0. \tag{8.98}$$

Das implizite Euler-Verfahren führt auf

$$u_k = u_{k-1} + h_k f(x_{k-1}, u_{k-1}) = u_{k-1} + h_k u_k,$$

was sich in diesem Fall recht einfach nach u_k auflösen lässt:

$$u_k = \frac{1}{1 - h_k} \left(u_{k-1} + 1\right).$$

Im nächstem Beispiel wird das implizite Euler-Verfahren auf ein nichtlineares DGL-System angewendet.

Beispiel(e) 8.13

$$\mathbf{y}'(x) = \begin{pmatrix} y_1'(x) \\ y_2'(x) \end{pmatrix} = \begin{pmatrix} y_2^2(x) \cdot \sin(x) \\ -y_1(x) \cdot \cos(x) \end{pmatrix} = \mathbf{f}(x, \mathbf{y}(x))$$

Das implizite Euler-Verfahren führt auf

$$\mathbf{u}_k = \begin{pmatrix} u_k, 1 \\ u_k, 2 \end{pmatrix} = \mathbf{u}_{k-1} + h_k \mathbf{f}(x_k, \mathbf{u}_k) = \begin{pmatrix} u_{k-1,1} \\ u_{k-1,2} \end{pmatrix} + h_k \begin{pmatrix} u_{k,2}^2 \cdot \sin(x_k) \\ -u_{k,1} \cdot \cos(x_k) \end{pmatrix}$$

Schreiben wir zur Abkürzung $h = h_k, c = \cos(x_k), s = \sin(x_k), p = u_{k-1,1}, q = u_{k-1,2}$ sowie $u = u_{k,1}$ und $v = u_{k,2}$, dann ist folgende Gleichung zu lösen:

$$F(u, v) := \begin{pmatrix} u - hs \cdot v^2 - p \\ v + hc \cdot u - q \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Zur Lösung nichtlinearer Gleichungen wie im letzten Beispiel kann man das Newton-Verfahren verwenden. Dieses versucht, eine gegen die Lösung (u, v) der Gleichung $F(u, v) = 0$ konvergente Folge $(u_j, v_j)_j$ zu finden, indem in jedem Iterationspunkt (u_{j-1}, v_{j-1}) die nichtlineare Gleichung $F = 0$ durch eine „Linearisierung“ ersetzt und (u_j, v_j) als *deren* Lösung definiert wird. Also: (u_j, v_j) ist gegeben als Lösung von

$$F(u_{j-1}, v_{j-1}) + J_F(u_{j-1}, v_{j-1}) \cdot \begin{pmatrix} u - u_{j-1} \\ v - v_{j-1} \end{pmatrix} = 0 \,. \tag{8.99}$$

Hier ist $J_F(p, q)$ die Jacobi-Matrix der Funktion $F = F(u, v)$ an der Stelle (p, q), also

$$J_F(p, q) = \begin{pmatrix} \dfrac{\partial F_1}{\partial u}(p, q) & \dfrac{\partial F_1}{\partial v}(p, q) \\ \dfrac{\partial F_2}{\partial u}(p, q) & \dfrac{\partial F_2}{\partial v}(p, q) \end{pmatrix}$$

Im obigen Beispiel 8.13 hätten wir demnach

$$J_F(u_{j-1}, v_{j-1}) = \begin{pmatrix} 1 & -2hs \cdot v_{j-1} \\ hc & 1 \end{pmatrix}$$

Man sieht, wie aufwändig das implizite Euler-Verfahren ist: *pro Schritt* ist ein nichtlineares Gleichungssystem zu lösen, das seinerseits ein iteratives Verfahren erfordert, welches für jeden *seiner* Schritte die Lösung eines linearen Gleichungssystems entsprechend (8.99) erfordert.

Es folgen einige weitere Verfahren.

- **Trapez-Regel:**

$$\mathbf{u}_k = \mathbf{u}_{k-1} + \frac{h_k}{2}(\mathbf{f}(x_{k-1}, \mathbf{u}_{k-1}) + \mathbf{f}(x_k, \mathbf{u}_k)), \quad \mathbf{u}_0 = \mathbf{y}_0 = \mathbf{y}(a). \tag{8.100}$$

  Die Trapezregel ergibt sich durch eine gegenüber dem (impliziten) Euler-Verfahren andere Art, Integrale näherungsweise zu berechnen, siehe die nachfolgende Skizze. Die Trapezregel ist ein implizites Einschrittverfahren: die nichtlineare Gleichung (8.100) muss nach dem unbekannten Vektor $\mathbf{u}_k$ aufgelöst werden.

- **Heun-Verfahren:**

$$\Delta_1 = h_k \mathbf{f}(x_{k-1}, \mathbf{u}_{k-1}) \tag{8.101}$$

$$\Delta_2 = h_k \mathbf{f}(x_k, \mathbf{u}_{k-1} + \Delta_1) \tag{8.102}$$

$$\mathbf{u}_k = \mathbf{u}_{k-1} + \frac{1}{2}(\Delta_1 + \Delta_2) \tag{8.103}$$

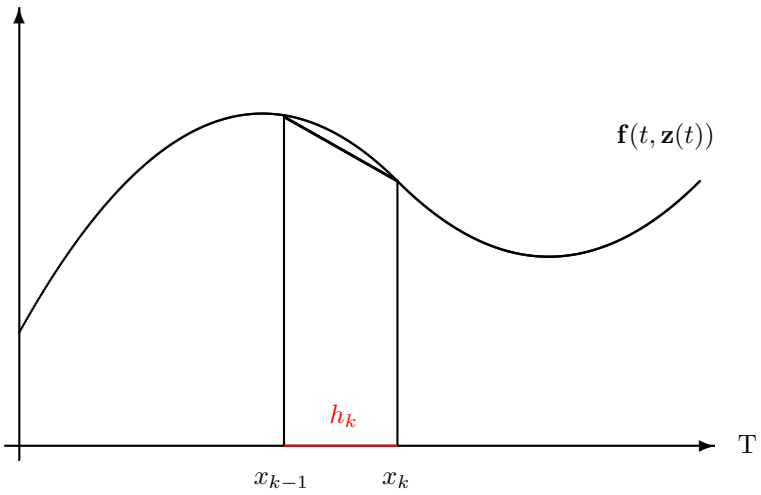


Abbildung 8.6: Trapez-Verfahren

- **Zwischenpunktsregel:**

$$\mathbf{u}_{k+1} = \mathbf{u}_{k-1} + (h_k + h_{k+1})\mathbf{f}(x_k, \mathbf{u}_k), \quad \text{ab } k = 1, \quad \tilde{\mathbf{y}}(x_0) = \mathbf{y}_0 = \mathbf{y}(a). \tag{8.104}$$

  Die Zwischenpunktsregel ist ein explizites Zweischrittverfahren (Mehrschrittverfahren). Die Näherung $\mathbf{u}_1$ kann mit ihr nicht berechnet werden – dafür muss ein Schritt eines Einschrittverfahrens ausgeführt werden.

- **Runge-Kutta-Verfahren:**

$$\begin{aligned}
\Delta_1 &:= h_k \mathbf{f}(x_{k-1}, \mathbf{u}_{k-1}) && (8.105)\\
\Delta_2 &:= h_k \mathbf{f}\left(x_{k-1} + \frac{h_k}{2}, \mathbf{u}_{k-1} + \frac{\Delta_1}{2}\right) && (8.106)\\
\Delta_3 &:= h_k \mathbf{f}\left(x_{k-1} + \frac{h_k}{2}, \mathbf{u}_{k-1} + \frac{\Delta_2}{2}\right) && (8.107)\\
\Delta_4 &:= h_k \mathbf{f}(x_{k-1} + h_k, \mathbf{u}_{k-1} + \Delta_3) && (8.108)\\
\mathbf{u}_k &= \mathbf{u}_{k-1} + \frac{1}{6}(\Delta_1 + 2\Delta_2 + 2\Delta_3 + \Delta_4), \quad \mathbf{u}_0 = \mathbf{y}_0 = \mathbf{y}(a). && (8.109)
\end{aligned}$$

Neben den angegebenen gibt es noch eine Vielzahl weiterer Verfahren, die sich in verschiedener Hinsicht unterscheiden. Eine besonders wichtige Frage bei der Beurteilung eines Verfahrens ist seine „Genauigkeit“. Etwas formaler: Unter dem **globalen Diskretisierungsfehler (GDF)** an einer Stelle x_k versteht man die Differenz $\mathbf{e}(x_k) := \mathbf{y}(x_k) - \mathbf{u}_k$. Der Betrag $\|\mathbf{e}(x_k)\|$ ist ein unmittelbares Maß für die Genauigkeit der Näherungslösung. [1]

Für konstante Schrittweiten $h_k \equiv h$ schreibt man auch $\mathbf{e}(h, x)$ für den globalen Diskretisierungsfehler, wobei $x = x_0 + nh$, $n \in \mathbb{N}$ (n Schritte). Gemeint ist also, dass man für ein beliebiges $x \in [a, b]$ eine Schrittweite h genau so wählt, dass man in n gleich großen Schritten von x_0 nach x kommt. Von einem sinnvollen Verfahren wird man erwarten, dass $\mathbf{e}(h, x)$ für jedes x gegen 0 geht, wenn $h \to 0$ oder gleichwertig $n \to \infty$. Verfahren können sich jedoch darin unterscheiden, *wie schnell* $\mathbf{e}(h, x)$ gegen 0 geht. Dazu betrachtet man alle möglichen Argumente $x \in [a, b]$ und

[1] Ein rein technischer Begriff, der zur Analyse von Verfahren eingeführt wurde, ist der des lokalen Diskretisierungsfehlers (LDF). Da wir hier keine Analyse von Verfahren durchführen, lassen wir die Definition des lokalen Diskretisierungsfehlers weg.

zu jedem x alle möglichen konstanten Schrittweiten h, so dass eine Näherung $\mathbf{u}(x) = \mathbf{u}_n$ an der Stelle x von x_0 aus (also alle $h = \frac{x-x_0}{n}$, $n \in \mathbb{N}$) berechnet wird. Gilt für alle $x \in [a, b]$ mit den entsprechenden $h > 0$:

$$|\mathbf{e}(h, x)| \leq s(x)h^p \tag{8.110}$$

mit beschränktem $s : [a, b] \to \mathbb{R}_0^+$ und $p \in \mathbb{N}$, so heißt p die Konvergenzordnung des Verfahrens.

Euler-Cauchy-Verfahren: $p = 1$,
Heun-Verfahren: $p = 2$,
Runge-Kutta-Verfahren: $p = 4$.

Das folgende Beispiel zeigt die Bedeutung einer hohen Konvergenzordnung.

Beispiel(e) 8.14

$$y'(x) = \sqrt{x^2 + y^2}, \quad y(0) = 0. \tag{8.111}$$

Man erhält als Betrag des globalen Diskretisierungsfehlers bei der Approximation des Wertes $y(2) = 2.588607526700...$.

N	Euler-Cauchy-Verfahren $h = 2/N$	Heun-Verfahren $h = 4/N$	Runge-Kutta-Verfahren $h = 8/N$
4	0.9812	0.24331354	0.1755687199278
8	0.5676	0.07925875	0.0229122136473
16	0.3098	0.02192620	0.0021689612996
32	0.1628	0.00564107	0.0001790363350
64	0.0834	0.00141657	0.0000136986430
128	0.0422	0.00035365	0.0000010030984
256	0.0213	0.00008825	0.0000000715001
512	0.0107	0.00002203	0.0000000050061
1024	0.0053	0.00000550	0.0000000003460
2048	0.0027	0.00000138	0.0000000000237
4096	0.0013	0.00000034	0.0000000000016

Hier bezeichnet N die Anzahl von Auswertungen der rechten Seite $\mathbf{f}$ der DGL, die erforderlich sind. Der Vergleich der Verfahren ist also fair: es wird die Güte der berechneten Approximation verglichen, die in allen vier Fällen mit gleichem Rechenaufwand erzielt wurde.

In der Praxis wird die Schrittweite h nicht konstant, sondern *adaptiv*, das heißt an das Problem angepasst gewählt: wenige, große Schritte dort, wo die Lösung „glatt“ ist und viele, kleine Schritte dort, wo die Lösung oszilliert oder „Knicke“ aufweist. Ziel der Schrittweitensteuerung ist es, über den gesamten zeitlichen Verlauf der Lösung hinweg eine gleich gute Qualität der Näherung zu erzielen. Hat man zum Beispiel für das Anfangswertproblem

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(a) = \mathbf{y}_0. \tag{8.112}$$

bereits eine Näherung $\mathbf{u}_{k-1} \approx \mathbf{y}(\hat{x})$, $\hat{x} \in [a, b)$, berechnet und hat man die Absicht, mit einem (der vorgestellten) Verfahren den Wert $\mathbf{y}(\hat{x} + h)$, $h \leq b - \hat{x}$, zu approximieren, so wählt man sich eine reelle Zahl $\delta > 0$ und kann dann die Schrittweite h folgendermaßen bestimmen:

- Bestimme einen Startwert $\hat{h}$, $\hat{h} \leq b - \hat{x}$, für die Schrittweite.
- Berechne eine Näherung $\tilde{\mathbf{u}}_k \approx \mathbf{y}(\hat{x} + \hat{h})$ für einen Schritt der Weite $\hat{h}$.

- Berechne eine Näherung $\bar{\mathbf{u}}_{k-\frac{1}{2}} \approx \mathbf{y}(\hat{x} + \frac{\hat{h}}{2})$ für einen Schritt der Weite $\frac{\hat{h}}{2}$ und basierend darauf $\bar{\mathbf{u}}_k \approx \mathbf{y}(\hat{x} + \hat{h})$ mit einem weiteren $\frac{\hat{h}}{2}$-Schritt mit Startpunkt $\bar{\mathbf{u}}_{k-\frac{1}{2}}$.
- Ist $\|\bar{\mathbf{u}}_k - \tilde{\mathbf{u}}_k\| \leq \delta$, so ist die Schrittweite $h = \hat{h}$ akzeptiert und $\bar{\mathbf{u}}_k$ als Approximation für $\mathbf{y}(\hat{x}+h)$ (in Abhängigkeit von δ) berechnet. Ansonsten: Wähle $\hat{h} := \frac{\hat{h}}{2}$ und wiederhole das Verfahren.

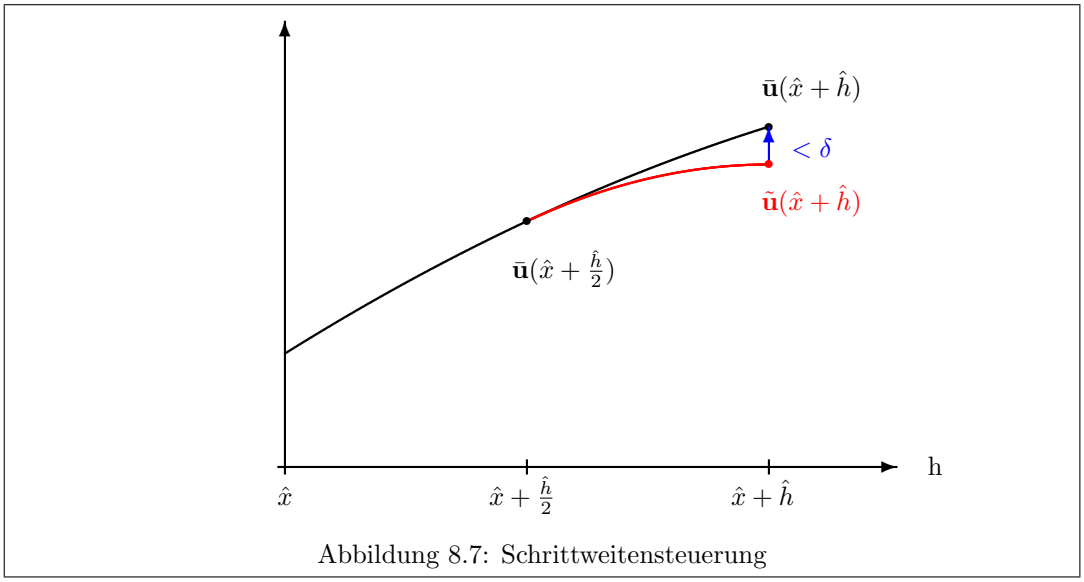


Abbildung 8.7: Schrittweitensteuerung

Eine „naive“ Erklärung für diese Vorgehensweise ist die folgende. Da man zumindest für $p > 1$ und genügend kleines $\hat{h}$ erwartet, dass $\bar{\mathbf{u}}_k$ eine bessere Näherung für $\mathbf{y}(\hat{x}+\hat{h})$ als $\tilde{\mathbf{u}}_k$ ist, bedeutet die Bedingung $\|\bar{\mathbf{u}}_k - \tilde{\mathbf{u}}_k\| \leq \delta$, dass bereits der $\hat{h}$-Schritt ausreichend genau ist. Hier kommt es natürlich sehr auf die Wahl von δ an, die wir aber nicht besprechen. Eine mathematisch saubere Begründung der Schrittweitenwahl stützt sich auf den oben erwähnten lokalen Diskretisierungsfehler. Erwähnt werden muss noch

- Das obige Verfahren sieht noch keine Vergrößerung der Schrittweite vor. Dies könnte man erreichen, indem man als Startwert für $\hat{h}$ immer erst die erfolgreiche Schrittweite des letzten Schritts verdoppelt.
- Die in der Praxis verwendeten Schrittweitensteuerungen sind wesentlich augeklügelter als das oben dargestellte Verfahren.

Die Konvergenzordnung eines Verfahrens ist nicht das einzige Gütekriterium. In Abschnitt 4.4 hatten wir die Stabilität von Gleichgewichtslagen, also von stationären Zuständen dynamischer Systeme betrachtet, vergleiche Defintion 4.24. Dabei ergab sich, dass ein stationärer Zustand eines linearen DGL-Systems $\mathbf{y}(x)' = M\mathbf{y}(x)$ asymptotisch stabil („anziehend“) ist, wenn alle Eigenwerte von M negativen Realteil haben, vergleiche Satz 4.26. Für nichtlineare Systeme $\mathbf{y}(x)' = \mathbf{f}(\mathbf{y}(x))$ mit Gleichgewichtslage $\mathbf{y}(x) \equiv \mathbf{a}$ gilt das gleiche Resultat, wenn man $M = J_{\mathbf{f}}(\mathbf{a})$ setzt (Jacobi-Matrix von $\mathbf{f}$ in $\mathbf{a}$).

Im folgenden Beispiel wird ein lineares DGL-System betrachtet, dessen Gleichgewichtslage $\mathbf{y}(x) \equiv 0$ asymptotisch stabil ist. Es gilt, dass jede Lösung $\mathbf{y}(x)$ des DGL-Systems für $x \to \infty$ gegen 0 konvergiert. Eine numerisch berechnete Lösung sollte natürlich das gleiche Verhalten aufweisen. Das heißt wir interessieren uns jetzt nicht für die Konvergenzfrage von oben, ob nämlich für ein

endliches x die berechnete Näherung gegen die exakte Lösung konvergiert, wenn die Schrittweite gegen 0 strebt, sondern vielmehr dafür, ob die berechnete Lösung bei endlicher Schrittweite das gleiche Verhalten wie die exakte Lösung hat, wenn die Anzahl der endlich weiten Schritte immer größer wird.

Beispiel(e) 8.15

$$\mathbf{y}'(x) = \underbrace{\begin{pmatrix} -500.5 & 499.5 \\ 499.5 & -500.5 \end{pmatrix}}_{\mathbf{M}} \mathbf{y}(x), \quad \mathbf{y}(0) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}. \tag{8.113}$$

Für das Euler-Cauchy-Verfahren mit konstanter Schrittweite erhält man mit $\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$:

$$\begin{aligned} \mathbf{u}_i &= \mathbf{u}_{i-1} + h\mathbf{M}\mathbf{u}_{i-1} = & (8.114)\\ &= (\mathbf{I} + h\mathbf{M})\mathbf{u}_{i-1} = & (8.115)\\ &= (\mathbf{I} + h\mathbf{M})^i \begin{pmatrix} 2 \\ 0 \end{pmatrix}. & (8.116) \end{aligned}$$

Da für die exakte Lösung $\mathbf{y} : [0, \infty) \to \mathbb{R}^2$ gilt:

$$\mathbf{y} : [0, \infty) \to \mathbb{R}^2, x \mapsto \begin{pmatrix} e^{-x} + e^{-1000x} \\ e^{-x} - e^{-1000x} \end{pmatrix}, \tag{8.117}$$

folgt $\lim_{x \to \infty} \mathbf{y}(x) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Andererseits konvergiert $\mathbf{u}_i$ für $i \to \infty$ nur dann gegen $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, wenn gilt:

$$|1 + h\lambda_1| < 1 \quad \text{und} \quad |1 + h\lambda_2| < 1 \tag{8.118}$$

mit $\lambda_1 = -1$ und $\lambda_2 = -1000$ (Eigenwerte von $\mathbf{M}$), also wenn $0 < h < 0.002$.
Es zeigt sich:
Obwohl in der exakten Lösung

$$\mathbf{y} : [0, \infty) \to \mathbb{R}^2, x \mapsto \begin{pmatrix} e^{-x} + e^{-1000x} \\ e^{-x} - e^{-1000x} \end{pmatrix}, \tag{8.119}$$

der Term e^{-1000x} praktisch keine Rolle spielt, denn für $x > 0.02$ ist dieser Term kleiner als 10^{-8}, bremst dieser Anteil der Lösung in der Numerik die Schrittweite, da gerade der Eigenwert $\lambda_2 = -1000$ die Schrittweite h auf $h < 0.002$ reduziert.

Ein lineares DGL-System $y' = My$ mit konstanten Koeffizienten heißt steif, wenn M mindestens einen Eigenwert mit stark negativem Realteil aufweist. In diesem Fall hat die exakte Lösung eine sehr schnell gegen Null abklingende additive Komponente, die eine drastische Beschränkung der maximal erlaubten Schrittweite erzwingt. Eine Überschreitung dieser Schrittweite führt zu einem Aufschaukeln der numerisch berechneten Näherungslösung, die zu dem fälschlichen Schhluss verleiten könnte, es liege eine Instabilität des untersuchten dynamischen Systems vor (während es sich in Wirklichkeit um eine rein „numerische Instabilität“ handelt). Dies haben wir oben für das Euler-Verfahren gesehen, gleiches gilt jedoch für alle expliziten Verfahren.

Geeignete Methoden zur numerischen Lösung steifer Differentialgleichungen leiten sich aus impli-

ziten Verfahren ab. Für das obige Beispiel erhält man für das implizite Eulerverfahren:

$$\begin{aligned} \mathbf{u}_i &= \mathbf{u}_{i-1} + h\mathbf{M}\mathbf{u}_i = && (8.120)\\ &= (\mathbf{I} - h\mathbf{M})^{-1}\mathbf{u}_{i-1} = && (8.121)\\ &= (\mathbf{I} - h\mathbf{M})^{-i} \begin{pmatrix} 2 \\ 0 \end{pmatrix}. && (8.122) \end{aligned}$$

Jetzt konvergiert $\mathbf{u}_i$ für $i \to \infty$ gegen $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, falls

$$|1 - h\lambda_1| > 1 \quad \text{und} \quad |1 - h\lambda_2| > 1, \tag{8.123}$$

also für alle $h > 0$.

Um nun den für die impliziten Verfahren erforderlichen hohen Rechenaufwand zu reduzieren ohne deren Stabilitätseigenschaften zu verlieren, wurden sogenannte semi-implizite Verfahren entwickelt. Dabei handelt es sich um implizite Verfahren, bei denen zur näherungsweisen Auflösung des nichtlinearen Gleichungssytems, das die nächste Iterierte $\mathbf{u}_k$ beschreibt, ein einziger Schritt eines Newton-Verfahrens aussgeführt wird. Wir beschreiben das Prinzip am Beispiel des semi-impliziten Euler-Verfahren. Sei

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(a) = \mathbf{y}_0 \tag{8.124}$$

und sei $\mathbf{f} : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$, $(x, \mathbf{z}) \mapsto \mathbf{f}(x, \mathbf{z})$ stetig und für alle $x \in [a, b]$ stetig nach $\mathbf{z}$ differenzierbar, so erhält man mit

$$\mathbf{J}_\mathbf{z}\mathbf{f}(x, \bar{\mathbf{z}}) = \begin{pmatrix} (f_1)_{z_1}(x, \bar{\mathbf{z}}) & \cdots & (f_1)_{z_n}(x, \bar{\mathbf{z}}) \\ \vdots & & \vdots \\ (f_n)_{z_1}(x, \bar{\mathbf{z}}) & \cdots & (f_n)_{z_n}(x, \bar{\mathbf{z}}) \end{pmatrix} \tag{8.125}$$

das semi-implizite Euler-Verfahren:

$$\begin{aligned} \mathbf{u}_k &= \mathbf{u}_{k-1} + h_k(\mathbf{I}_n - h_k\mathbf{J}_\mathbf{z}\mathbf{f}(x_k, \mathbf{u}_{k-1}))^{-1}\mathbf{f}(x_k, \mathbf{u}_{k-1}), && (8.126)\\ \mathbf{u}_0 &= \mathbf{y}_0 = \mathbf{y}(a), && (8.127) \end{aligned}$$

wobei $\mathbf{I}_n$ die n-dim. Einheitsmatrix bezeichnet.
Dieses Verfahren erhält man aus dem impliziten Euler-Verfahren

$$\mathbf{u}_k = \mathbf{u}_{k-1} + h_k\mathbf{f}(x_k, \mathbf{u}_k) \tag{8.128}$$

dadurch, dass das nichtlineare Gleichungssystem

$$\mathbf{F}(\mathbf{z}) = \mathbf{z} - \mathbf{u}_{k-1} - h_k\mathbf{f}(x_k, \mathbf{z}) = \mathbf{0} \tag{8.129}$$

betrachtet wird. Linearisierung von $\mathbf{F}$ um $\mathbf{z}_0 = \mathbf{u}_{k-1}$ liefert:

$$\mathbf{F}(\mathbf{z}_0) + \mathbf{J}_\mathbf{z}\mathbf{F}(\mathbf{z}_0)(\mathbf{z} - \mathbf{z}_0) = \mathbf{0} \tag{8.130}$$

beziehungsweise:

$$-h_k\mathbf{f}(x_k, \mathbf{u}_{k-1}) + (\mathbf{I}_n - h_k\mathbf{J}_\mathbf{z}\mathbf{f}(x_k, \mathbf{u}_{k-1}))(\mathbf{z} - \mathbf{u}_{k-1}) = \mathbf{0}. \tag{8.131}$$

Daraus folgt für die Lösung $\tilde{\mathbf{z}}$:

$$\tilde{\mathbf{z}} := \mathbf{u}_k = \mathbf{u}_{k-1} + h_k(\mathbf{I}_n - h_k\mathbf{J}_\mathbf{z}\mathbf{f}(x_k, \mathbf{u}_{k-1}))^{-1}\mathbf{f}(x_k, \mathbf{u}_{k-1}). \tag{8.132}$$

Das semi-implizite Euler-Verfahren vereinigt die Vorteile eines expliziten Verfahrens ($\mathbf{u}_k$ kommt nur auf der linken Seite des Verfahrensschemas in Form $\mathbf{u}_k = \ldots$ vor) mit dem Vorteil, dass damit auch steife Differentialgleichungen numerisch behandelt werden können. Ein Nachteil ist die Verwendung von $\mathbf{J}_\mathbf{z}\mathbf{f}$. Die Schrittweite h_k kann immer so gewählt werden, dass die Inverse zu $(\mathbf{I}_n - h_k\mathbf{J}_\mathbf{z}\mathbf{f}(x_k, \mathbf{u}_{k-1}))$ existiert.

8.8 Lokale Minimierung ohne Nebenbedingungen

In diesem Abschnitt untersuchen wir iterative Methoden zur lokalen Minimierung einer gegebenen Zielfunktion $f : \mathbb{R}^k \to \mathbb{R}$, $f \in C^2$, ohne Nebenbedingungen. Es werden Algorithmen betrachtet, die Folgen $\{\mathbf{x}_j\}$, $\{\mathbf{s}_j\}$ und $\{\sigma_j\}$ mit $j \in \mathbb{N}_0$ derart erzeugen, dass für jedes $j \in \mathbb{N}_0$ gilt:

$$\begin{aligned} \mathbf{x}_{j+1} &= \mathbf{x}_j - \sigma_j \mathbf{s}_j & (8.133)\\ f(\mathbf{x}_{j+1}) &< f(\mathbf{x}_j), & (8.134) \end{aligned}$$

wobei $\mathbf{x}_j, \mathbf{s}_j \in \mathbb{R}^k, \sigma_j \in \mathbb{R}^+$. Die Verwendbarkeit dieser Algorithmen entscheidet sich im wesentlichen durch die Wahl der Suchrichtung $-\mathbf{s}_j$. Aufgrund der obigen Festlegung ist für jedes $j \in \mathbb{N}_0$

$$\nabla f(\mathbf{x}_j)^\top \mathbf{s}_j > 0 \tag{8.135}$$

(-$\mathbf{s}_j$ sind Abstiegsrichtungen) zu fordern. Ziel ist es, Algorithmen zu konstruieren, die entweder superlinear gegen einen lokalen Minimierer $\bar{\mathbf{x}}$ von f konvergieren, für die also gilt:

$$\lim_{j\to\infty} \frac{\|\mathbf{x}_{j+1} - \bar{\mathbf{x}}\|_2}{\|\mathbf{x}_j - \bar{\mathbf{x}}\|_2} = 0, \tag{8.136}$$

oder Algorithmen zu konstruieren, die quadratisch gegen einen lokalen Minimierer $\bar{\mathbf{x}}$ konvergieren, für die also gilt:

$$\lim_{j\to\infty} \frac{\|\mathbf{x}_{j+1} - \bar{\mathbf{x}}\|_2}{\|\mathbf{x}_j - \bar{\mathbf{x}}\|_2^2} \leq \alpha,\ \alpha \in \mathbb{R}_0^+. \tag{8.137}$$

Beim Newton-Verfahren wird die Suchrichtung $\mathbf{s}_j$ als Lösung des linearen Gleichungssystems

$$Hf(\mathbf{x}_j)\mathbf{s}_j = \nabla f(\mathbf{x}_j) \tag{8.138}$$

für alle $j \in \mathbb{N}_0$ bestimmt. Die Schrittweite σ_j wird im allgemeinen durch den folgenden Algorithmus von Armijo-Goldstein berechnet: Sei $0 < \delta < 0.5$, dann wird das kleinste $\nu \in \mathbb{N}_0$ bestimmt, sodass mit $\sigma_j := (0.5)^\nu$ gilt:

$$f(\mathbf{x}_j) - \delta\sigma_j \nabla f(\mathbf{x}_j)^\top \mathbf{s}_j \geq f(\mathbf{x}_j - \sigma_j \mathbf{s}_j). \tag{8.139}$$

Befindet man sich nun in einer geeigneten Umgebung eines lokalen Minimierers $\bar{\mathbf{x}}$ von f mit positiv definiter Hessematrix, so ist das Newton-Verfahren durchführbar und die Folge $\{\mathbf{x}_j\}$ konvergiert gegen $\bar{\mathbf{x}}$, wobei nach endlich vielen Schritten stets $\nu = 0$ (Schrittweite gleich 1) gilt. Kann man zudem für die zweite Ableitung von f in dieser Umgebung eine Lipschitzbedingung nachweisen (gegebenenfalls bei $f \in C^3$), so konvergiert die vom Newton-Verfahren erzeugte Folge $\{\mathbf{x}_j\}$ quadratisch gegen $\bar{\mathbf{x}}$. Offensichtlich können mit dieser Methode im allgemeinen nur lokale Minimierer berechnet werden. Die quadratische Konvergenz wird durch einen relativ hohen Aufwand (Berechnung des Gradienten und der Hessematrix in jedem Iterationsschritt) erkauft. Daher versucht man, die Information zweiter Ordnung durch bisher berechnete Gradienten zu approximieren. Eines der wichtigsten Verfahren in diesem Zusammenhang bildet das BFGS-Verfahren (Broyden, Fletcher, Goldfarb, Shanno-Verfahren). Die Suchrichtung wird durch

$$\mathbf{s}_j = \mathbf{H}_j \nabla f(\mathbf{x}_j) \tag{8.140}$$

für alle $j \in \mathbb{N}_0$ berechnet (mit $\mathbf{H}_0 = \mathbf{I}_k$ oder $\mathbf{H}_0$ gleich einer wählbaren, positiv definiten Matrix). Die Matrix $\mathbf{H}_{j+1}$ berechnet sich durch:

$$\mathbf{H}_{j+1} = \mathbf{H}_j + \frac{\mathbf{d}_j^\top \mathbf{p}_j + \mathbf{d}_j^\top \mathbf{H}_j \mathbf{d}_j}{(\mathbf{d}_j^\top \mathbf{p}_j)^2} \mathbf{p}_j \mathbf{p}_j^\top - \frac{\mathbf{p}_j \mathbf{d}_j^\top \mathbf{H}_j + \mathbf{H}_j \mathbf{d}_j \mathbf{p}_j^\top}{\mathbf{d}_j^\top \mathbf{p}_j}, \tag{8.141}$$

wobei $\mathbf{d}_j := \frac{\nabla f(\mathbf{x}_j) - \nabla f(\mathbf{x}_{j+1})}{\|\sigma_j \mathbf{s}_j\|_2}$, $\mathbf{p}_j := \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|_2}$. Die Schrittweite $\sigma_j, j \in \mathbb{N}_0$, wird durch folgenden Algorithmus berechnet: Wähle γ_1 und γ_2 mit $0 < \gamma_1 < \gamma_2 \leq 0.5$ und bestimme σ_j so, dass folgendes gilt:

- $\nabla f(\mathbf{x}_{j+1})^\top \mathbf{s}_j \leq \gamma_2 \nabla f(\mathbf{x}_j)^\top \mathbf{s}_j$ (somit ist $\mathbf{H}_{j+1}$ positiv definit)
- $f(\mathbf{x}_{j+1}) \leq f(\mathbf{x}_j) - \gamma_1 \sigma_j \nabla f(\mathbf{x}_j)^\top \mathbf{s}_j$
- $\sigma_j = 1$, falls möglich.

Ziel der BFGS-Methode ist es, in den positiv definiten Matrizen $\mathbf{H}_j$ die für die Konvergenzgeschwindigkeit relevanten Informationen der zweiten Ableitung von f durch Verwendung der Gradienten

$$\nabla f(\mathbf{x}_0), \nabla f(\mathbf{x}_1), \nabla f(\mathbf{x}_2), \ldots \tag{8.142}$$

zu approximieren. Befindet man sich in einer geeigneten Umgebung eines lokalen Minimierers $\bar{\mathbf{x}}$ von f mit positiv definiter Hessematrix, so konvergiert die vom BFGS-Verfahren erzeugte Folge $\{\mathbf{x}_j\}$ superlinear gegen $\bar{\mathbf{x}}$, falls die zweite Ableitung von f in dieser Umgebung einer Lipschitzbedingung genügt. Die benötigten Auswertungen der Gradienten und Hessematrizen an den entsprechenden Iterationspunkten zur Realisierung des Newton- beziehungsweise BFGS-Verfahrens können durch verschiedene Verfahren der rechnergestützten Differentiation (numerisch, symbolisch oder automatisch) berechnet werden.

BFGS-Verfahren

Schritt 0: (Initialisierung)

Wähle $\mathbf{x}_0$, eine positiv definite Matrix $\mathbf{H}_0$ und γ_1, γ_2 mit:

$$0 < \gamma_1 < \gamma_2 \leq 0.5. \tag{8.143}$$

Berechne $\nabla f(\mathbf{x}_0)$.
Falls $\nabla f(\mathbf{x}_0) = \mathbf{0}$, dann STOP; sonst: j := 0, gehe zu Schritt 1.

Schritt 1: (Berechnung der Suchrichtung)

Berechne

$$\mathbf{s}_j := \mathbf{H}_j \nabla f(\mathbf{x}_j), \tag{8.144}$$

gehe zu Schritt 2.

Schritt 2: (Berechnung der Schrittweite)

Berechne σ_j derart, dass gilt:

$$\nabla f(\mathbf{x}_{j+1})^\top \mathbf{s}_j \leq \gamma_2 \nabla f(\mathbf{x}_j)^\top \mathbf{s}_j, \tag{8.145}$$

$$f(\mathbf{x}_{j+1}) \leq f(\mathbf{x}_j) - \gamma_1 \sigma_j \nabla f(\mathbf{x}_j)^\top \mathbf{s}_j, \tag{8.146}$$

$$\sigma_j = 1, \text{falls möglich.} \tag{8.147}$$

Berechne

$$\mathbf{x}_{j+1} = \mathbf{x}_j - \sigma_j \mathbf{s}_j. \tag{8.148}$$

Falls $\nabla f(\mathbf{x}_{j+1}) = \mathbf{0}$, dann STOP; sonst: gehe zu Schritt 3.

Schritt 3: (Berechnung von $\mathbf{H}_{j+1}$)

Setze

$$\mathbf{d}_j := \frac{\nabla f(\mathbf{x}_j) - \nabla f(\mathbf{x}_{j+1})}{\|\sigma_j \mathbf{s}_j\|_2}, \quad \mathbf{p}_j := \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|_2}. \tag{8.149}$$

Berechne

$$\mathbf{H}_{j+1} = \mathbf{H}_j + \frac{\mathbf{d}_j^\top \mathbf{p}_j + \mathbf{d}_j^\top \mathbf{H}_j \mathbf{d}_j}{(\mathbf{d}_j^\top \mathbf{p}_j)^2} \mathbf{p}_j \mathbf{p}_j^\top - \frac{\mathbf{p}_j \mathbf{d}_j^\top \mathbf{H}_j + \mathbf{H}_j \mathbf{d}_j \mathbf{p}_j^\top}{\mathbf{d}_j^\top \mathbf{p}_j}. \tag{8.150}$$

Ersetze j durch j+1,
gehe zu Schritt 1.

Kapitel 9

Partielle Differentialgleichungen

Die Behandlung partieller DGL (PDGL) ist wesentlich komplizierter als die gewöhnlicher DGL und es gibt hierfür keine einheitliche Theorie. Ja selbst zur Lösung ein- und derselben PDGL muss man manchmal unterschiedliche Methoden anwenden, je nachdem, auf welchem Gebiet und unter welchen Nebenbedingungen (d.h. zu welchen Rand- und/oder Anfangswerten) eine Lösung gesucht ist. Demgemäß muss der folgende Überblick recht grob ausfallen.

9.1 Beispiele und Grundbegriffe

Eine **partielle Differentialgleichung** (PDGL) ist eine DGL für Funktionen in mehreren Veränderlichen. In der allgemeinsten Form handelt es sich um eine Gleichung

$$F\left(x_1, x_2, \ldots, x_n, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}, \ldots, \frac{\partial^p u}{\partial x_1^k \cdots \partial x_n^s}\right) = 0. \tag{9.1}$$

Hier sind $x_1, \ldots, x_n$ unabhängige Variable, manchmal zusammengefasst zum Vektor $\mathbf{x} \in \mathbb{R}^n$, und gesucht ist eine Funktion $u : G \to \mathbb{R}, \mathbf{x} \mapsto u(x_1, \ldots, x_n) = u(\mathbf{x})$, die die Gleichung erfüllt, wobei $G \subseteq \mathbb{R}^n$ ein Gebiet ist. Für die partiellen Ableitungen schreibt man auch

$$\frac{\partial u}{\partial x_i} = u_{x_i}, \quad \frac{\partial^2 u}{\partial x_i \partial x_j} = u_{x_i x_j} \quad \text{usw.}$$

Häufig ist u eine Funktion, die von bis zu drei Raumdimensionen und eventuell noch von der Zeit abhängt. Dann schreibt man für die Variablen in der Regel t (Zeit) und x, y und z (Raum) anstatt $x_1, \ldots, x_4$. Zum Beispiel könnte $u = u(x, y)$ ein elektrostatisches Potential in 2 Raumdimensionen beschreiben, das einer PDGL der Form

$$u_{xx}(x, y) + u_{yy}(x, y) = f(x, y)$$

genügt. Oder es könnte $u = u(t, x, y, z)$ die Ausbreitung einer Schallwelle in einem homogenen dreidimensionalen Medium beschreiben: dann hätte man eine PDGL der Form

$$u_{tt}(t, x, y, z) = c^2[u_{xx}(t, x, y, z) + u_{yy}(t, x, y, z) + u_{zz}(t, x, y, z)] + f(t, x, y, z). \tag{9.2}$$

Die PDGL (9.1) hat die **Ordnung** p, wenn $p = k + \ldots + s$ die höchste tatsächlich auftretende Ableitung von u ist. Eine Lösung u der PDGL nennt man manchmal auch **Integralfläche** oder **Lösungsfläche**.

Die PDGL (9.1) ist **implizit**. Wenn die Gleichung nach einer der höchsten auftretenden Ableitungen aufgelöst werden kann, dann heißt die PDGL **explizit**. Explizit ist etwa die PDGL 2.

Ordnung aus (9.2).

Treten in dem Ausdruck $F(\ldots)$ aus (9.1) die Funktion u und ihre Ableitungen nur linear auf, dann heißt die PDGL **linear**. Beispielsweise sind (9.2) oder auch

$$x^2 \cdot u_{xx} + \sin x \cdot u_{xy} + \ln y \cdot u_y + u = \sin x \tag{9.3}$$

lineare PDGL.

Die obige lineare PDGL ist **inhomogen**. Wenn nur Terme autreten, die u oder Ableitungen von u enthalten, dann heißt eine lineare PDGL **homogen**. Die PDGL (9.3) wäre homogen, wenn man den $\sin$-Term auf der rechten Seite durch Null ersetzen würde.

In „milder Form“ nichtlinear sind **quasilineare** PDGL. Dies sind PDGL, bei denen wenigstens alle partiellen Ableitungen der höchsten Ordnung nur linear auftreten (partielle Ableitungen niedrigerer Ordnung können nichtlinear auftreten). Die allgemeinste quasilineare PDGL 2.Ordnung in 2 Variablen hat die Form

$$a(x,y,u,u_x,u_y)u_{xx} + b(x,y,u,u_x,u_y)u_{xy} + c(x,y,u,u_x,u_y)u_{yy} + d(x,y,u,u_x,u_y) = 0\,. \tag{9.4}$$

Hängen die Funktionen a, b und c nur von x und y ab, dann heißt die PDGL (9.4) **semilinear**.

Neben skalaren PDGL, also PDGL für *skalarwertige* Funktionen $u : \mathbb{R}^n \to \mathbb{R}$, werden auch Systeme von PDGL, also PDGL für *vektorwertige* Funktionen $\mathbf{u} : \mathbb{R}^n \to \mathbb{R}^m$ betrachtet. In der Regel ist dabei $m = n$, zum Beispiel $m = n = 2$ bei den Cauchy-Riemannschen DGL

$$\begin{aligned} u_x &= v_y \\ u_y &= -v_x \end{aligned}$$

für die vektorwertige Funktion

$$\mathbf{u}(x,y) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix}.$$

Fast alle in der Praxis vorkommenden PDGL sind von der Ordnung 1 oder 2. Wir listen nun ein paar Beispiele auf.

Beispiel(e) 9.1
Die Transportgleichung (Burgers-Gleichung) in 2 Variablen x und t

$$u_t + uu_x = 0.$$

ist eine quasilineare PDGL 1. Ordnung.

Beispiel(e) 9.2
Die **Diffusionsgleichung** bzw. **Wärmeleitungsgleichung**

$$\Delta u = \frac{1}{\kappa} u_t,$$

ist eine lineare PDGL 2. Ordnung mit konstanten Koeffzienten für eine Funktion u in der Zeitvariable t und den Ortsvariablen x (1D), (x, y) (2D) oder (x, y, z) (3D). Nach bei PDGL üblicher Konvention ist der **Laplace-Operator** Δ so zu verstehen, dass er sich nur auf die Ortsvariablen der Funktion u erstreckt, nicht auf die Zeitvariable. Beispielsweise ist im $\mathbb{R}^3$ („3-dimensionale Diffusionsgleichung“) $u = u(t, x, y, z)$ und

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \ .$$

Beispiel(e) 9.3
Die **Wellengleichung**

$$\Delta u = \frac{1}{c^2} u_{tt}$$

ist ebenfalls eine lineare PDGL 2. Ordnung mit konstanten Koeffizienten für eine Funktion u in der Zeitvariablen t und den Ortsvariablen x, (x, y) oder (x, y, z). Auch hier bezieht sich der Laplace-Operator nur auf die Ortsvariablen, nicht auf die Zeitvariable.

Beispiel(e) 9.4
Die **Potentialgleichung**

$$\Delta u = 0$$

ist eine weitere lineare PDGL 2. Ordnung mit konstanten Koeffizienten, diesmal jedoch für eine Funktion u allein in den Ortsvariablen x, y (und eventuell z).

Beispiel(e) 9.5
Die **Schrödingergleichung**

$$\Delta u - \frac{2m}{\hbar^2} V u = -\frac{2im}{\hbar} u_t$$

ist eine lineare PDGL 2. Ordnung für eine Funktion u in der Zeitvariablen t und den Ortsvariablen x, y und z. Die Koeffizienten sind hier nicht alle konstant, vielmehr ist $V = V(r)$ eine Funktion, die vom Abstand $r = \|\mathbf{x}\|_2$ von $\mathbf{x} = (x, y, z)$ vom Nullpunkt, also von den Ortsvariablen abhängt.

Beispiel(e) 9.6
Die **Telegrafengleichungen**

$$\begin{aligned} u_x + L \cdot v_t + R \cdot v &= 0 \\ v_x + C \cdot u_t + G \cdot u &= 0\,. \end{aligned}$$

sind ein System von PDGL 1. Ordnung. Hier ist $u(t,x)$ die Spannung und $v(t,x)$ die Stromstärke im Punkt x eines Kabels zur Zeit t (L: Induktivität, R: Ohmscher Widerstand, C: Kapazität, G: Ohmscher Leitwert).

Aus den Telegrafengleichungen lassen sich zwei entkoppelte Gleichungen 2. Ordnung für u und für v gewinnen. Dazu leitet man die erste Gleichung nach x ab und setzt die zweite Gleichung ein:

$$u_{xx} + L \cdot v_{xt} - R \cdot (C \cdot u_t + G \cdot u) = 0.$$

Wenn man jetzt noch die nach t abgeleitete zweite Telegrafengleichung einsetzt, dann erhält man

$$u_{xx} - LC \cdot u_{tt} - (LG + RC) \cdot u_t - RG \cdot u = 0.$$

Eine Lösung dieser (Wellen-)Gleichung kann mit einem „Separationsansatz“ gefunden werden, siehe Abschnitt 9.2. Für v geht man analog vor. Die Lösungen der ursprünglichen Telegrafengleichungen lassen sich unter den Lösungen der gewonnenen Wellengleichungen finden.

Beispiel(e) 9.7
Die **Maxwellschen Gleichungen** für das elektrische Feld $\mathbf{E}(t,\mathbf{x})$ und das magnetische Feld $\mathbf{B}(t,\mathbf{x})$ lauten im Vakuum (mit magnetischer Feldkonstante μ_0 und elektrischer Feldkonstante ϵ_0)

$$\begin{aligned} \text{rot } \mathbf{B} &= \epsilon_0\mu_0 \frac{\partial \mathbf{E}}{\partial t}, \qquad & \text{div } \mathbf{B} = 0, \\ \text{rot } \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, & \text{div } \mathbf{E} = 0\,. \end{aligned}$$

Aus den Maxwellschen Gleichungen lassen sich 6 entkoppelte Gleichungen 2. Ordnung (Wellengleichungen) gewinnen. Man wende dazu auf die beiden linken Gleichungen den Differentialoperator rot an und berücksichtige dabei

$$\text{rot}(\text{rot } \mathbf{u}) = \begin{pmatrix} \frac{\partial}{\partial x_1}(\text{div } \mathbf{u}) - \Delta u_1 \\ \frac{\partial}{\partial x_2}(\text{div } \mathbf{u}) - \Delta u_2 \\ \frac{\partial}{\partial x_3}(\text{div } \mathbf{u}) - \Delta u_3 \end{pmatrix}.$$

Da die Divergenz beider Vektorfelder nach den Gleichungen der rechten Seite verschwindet, erhält man etwa für $\mathbf{B}$

$$-\begin{pmatrix} \Delta B_1 \\ \Delta B_2 \\ \Delta B_3 \end{pmatrix} = \epsilon_0\mu_0 \frac{\partial}{\partial t}(\text{rot } \mathbf{E}) = -\epsilon_0\mu_0 \frac{\partial}{\partial t}\left(\frac{\partial \mathbf{B}}{\partial t}\right).$$

Analog geht man für $\mathbf{E}$ vor.

Die allgemeine Lösung einer gewöhnlichen DGL vom Grad n hängt von n Integrationskonstanten ab. Bei PDGL zeigt sich ein wesentlich komplizierteres Verhalten der allgemeinen Lösung. Wir betrachten dazu ein Beispiel, die eindimensionale Wellengleichung

$$u_{tt} - c^2 u_{xx} = f(x,t), \quad c > 0\,.$$

Im Gegensatz zur oben zitierten homogenen Form der Wellengleichung haben wir hier noch eine Störterm f auf der rechten Seite berücksichtigt. Diese PDGL kann nach d'Alembert durch eine Variablentransformation auf eine gewöhnliche DGL zurückgeführt werden:

$$\xi = x - ct, \quad \tau = x + ct, \qquad U(\xi,\tau) := u(x,t)$$

führt mit der Kettenregel

$$\begin{aligned} u_t = U_\xi \xi_t + U_\tau \tau_t = -cU_\xi + cU_\tau \quad &\Rightarrow \quad u_{tt} = c^2 U_{\xi\xi} - 2c^2 U_{\xi\tau} + c^2 U_{\tau\tau} \\ u_x = U_\xi \xi_x + U_\tau \tau_x = U_\xi + U_\tau \quad &\Rightarrow \quad u_{xx} = U_{\xi\xi} + 2U_{\xi\tau} + U_{\tau\tau} \end{aligned}$$

auf

$$u_{tt} - c^2 u_{xx} = -4c^2 U_{\xi\tau} = F(\xi,\tau) := f\left(\frac{\xi+\tau}{2}, \frac{\xi-\tau}{-2c}\right)\,.$$

Man erhält eine partikuläre Lösung durch 2-malige Integration (nach ξ und nach τ) und anschließende Rücksubstitution in der Form

$$u_p(x,t) := U_p(\xi,\tau) = -\iint \frac{1}{4c^2} F(\xi,\tau)\, d\xi\, d\tau.$$

Für die allgemeine homogene Lösung ($f = 0$, $F = 0$) ergibt sich

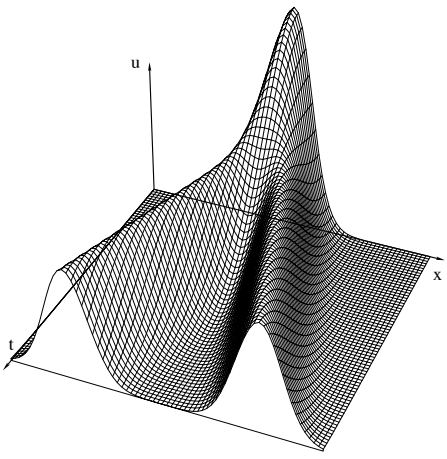
$$U_{\xi\tau} = 0 \quad \Rightarrow \quad U_\xi = c(\xi) \quad \Rightarrow \quad U = \int c(\xi)\, d\xi + h(\tau) = g(\xi) + h(\tau)$$

mit *beliebig wählbaren* zweimal stetig differenzierbaren Funktionen g und h einer Veränderlichen.

Durch Rücksubstitution erhält man daraus die allgemeine homogene Lösung

$$u_h(x,t) = g(x-ct) + h(x+ct).$$

u_h ist interpretierbar als die Überlagerung einer mit Geschwindigkeit c sich nach rechts ausbreitenden Welle $g(x-ct)$ mit einer gegenläufigen Welle $h(x+ct)$.



Nebenstehend ist dies für den Fall

$$g(z) = h(z) = e^{-z^2}$$

skizziert.

Die allgemeine Lösung der Wellengleichung hat demnach die Form

$$u(x,t) = u_p(x,t) + g(x-ct) + h(x+ct)$$

und im Gegensatz zu den 2 Integrationskonstanten einer gewöhnlichen DGL 2.Ordnung treten jetzt 2 beliebig wählbare Funktionen einer Veränderlichen auf. Dieses Verhalten ist typisch (aber nicht allgemein!) für PDGL:

Typischerweise enthält die allgemeine Lösung einer PDGL m-ter Ordnung in n Variablen m willkürlich wählbare Funktionen in $n-1$ Variablen.

Um eindeutige Lösbarkeit zu erhalten, braucht man zusätzliche Bedingungen, um die „freien Funktionen“ in der Lösung eindeutig festzulegen.

Betrachten wir dazu weiter das Beispiel der eindimensionalen homogenen Wellengleichung

$$u_{tt} = c^2 u_{xx}$$

mit der allgemeinen Lösung

$$u(x,t) = g(x-ct) + h(x+ct)\ .$$

Hier ließen sich in Verallgemeinerung der Anfangswerte bei gewöhnlichen DGL sogenannte **Anfangsbedingungen** für die gesuchte Funktion u festschreiben:

$$u(x,0) = \varphi(x), \quad u_t(x,0) = \psi(x)\ ,$$

wobei φ und ψ vorgegebene, bekannte Funktionen sein sollen.

(**Zwischenbemerkung.** Man beachte, dass es wenig sinnvoll ist, die Ableitung u_x längs der x-Achse festzuschreiben, denn diese ist ja bereits durch die Festlegung $u(x,0) = \varphi(x)$ bestimmt — man hätte bestenfalls keine neue und schlimmstenfalls eine widersprüchliche Bedingung.)

Einsetzen der Anfangsbedingungen in die allgemeine Lösung liefert für $t = 0$

$$\varphi(x) = g(x) + h(x), \qquad \psi(x) = c\,(-g'(x) + h'(x))\ .$$

Es ergeben sich also für die unbestimmten Funktionen h und g die beiden Gleichungen

$$\begin{aligned} h(x) + g(x) &= \varphi(x) \\ h(x) - g(x) &= \frac{1}{c}\int_{x_0}^{x} \psi(\zeta)\,d\zeta + C, \quad C \equiv \text{ const,} \end{aligned}$$

für ein beliebiges $x_0 \in \mathbb{R}$. Addition und Subtraktion dieser Gleichungen voneinander ergibt

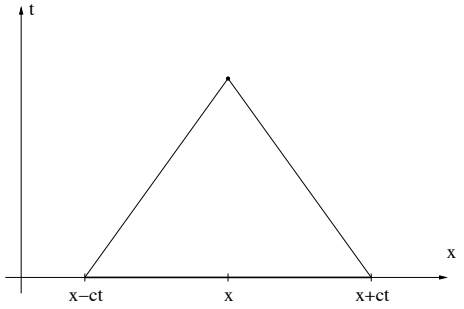
$$\begin{aligned} h(x) &= \frac{1}{2}\varphi(x) + \frac{1}{2c}\int_{x_0}^{x} \psi(\zeta)\,d\zeta + \frac{C}{2} \\ g(x) &= \frac{1}{2}\varphi(x) - \frac{1}{2c}\int_{x_0}^{x} \psi(\zeta)\,d\zeta - \frac{C}{2}\ . \end{aligned}$$

Wenn man das jetzt wieder in die allgemeine Lösung $u(x,t) = g(x-ct) + h(x+ct)$ einsetzt, dann bekommt man die **Lösungsformel von d'Alembert**

$$u(x,t) = \frac{1}{2}\left(\varphi(x-ct) + \varphi(x+ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} \psi(\zeta)\,d\zeta\ .$$

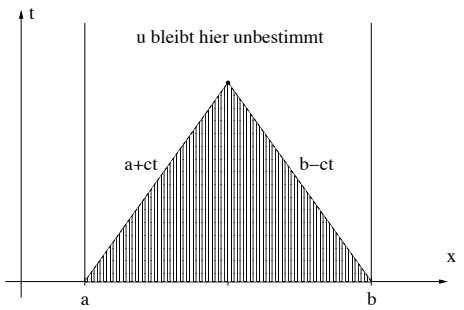
Hier hat also das Vorgeben einer Anfangsbedingung längs der x-Achse tatsächlich dazu geführt, dass die unbestimmten Funktionen der allgemeinen Lösung festgelegt werden und eine eindeutig bestimmte Lösung resultiert.

Bemerkung. Da die Lösung u im Punkt (x,t) nach der d'Alembertschen Lösungsformel abhängt von den Werten der Anfangsbedingungen im Intervall $[x-ct, x+ct]$ nennt man dieses Intervall den **Abhängigkeitsbereich** der Lösung. Dass der Abhängigkeitsbereich endlich ist bedeutet, dass sich Wellen mit endlicher Geschwindigkeit ausbreiten.



Das Vorgeben von Anfangsbedingungen geschieht in Analogie zum Vorgeben von Anfangswerten bei gewöhnlichen DGL. Entsprechend nennt man PDGL mit zusätzlichen Anfangsbedingungen auch **Anfangswertprobleme** (oder auch **Cauchy-Probleme**).

Man könnte die Wellengleichung jedoch auch für ein in den Ortsvariablen beschränktes Gebiet untersuchen, z.B. für $n = 1$ im Bereich $a \leq x \leq b$. Aufgrund des oben festgestellten Abhängigkeitsbereichs der Lösung von den Anfangswerten würde die Lösung durch Vorgabe von Anfangswerten dann aber nur, so wie nebenstehend schrafffiert skizziert, in einem endlichen Bereich eindeutig bestimmt sein, während sie im restlichen Zylinder $\mathbb{R}^+ \times [a, b]$ unbestimmt bliebe.



Abhilfe schafft in diesem Fall die Vorgabe *zusätzlicher Randbedingungen* für $x = a$ und $x = b$, das heißt man schreibt weiterhin für $t = 0$ und gegebene Funktionen φ und ψ

$$u(x, 0) = \varphi(x), \qquad u_t(x, 0) = \psi(x), \qquad a \leq x \leq b,$$

als Anfangsbedingungen und zusätzlich noch die Randbedingungen

$$u(a, t) = g(t), \qquad u(b, t) = h(t), \qquad t \geq 0,$$

vor. Man spricht in diesem Fall von einem **Rand-Anfangswertproblem**. Dieses Problem hat wiederum eine eindeutige Lösung.

Wellengleichungen oder auch Wärmeleitungsgleichungen beschreiben *dynamische*, das heißt von der Zeit abhängige Vorgänge. Dazu „passt“ die Vorgabe von Anfangsbedingungen oder Anfangs-Randbedingungen, je nachdem, ob die Lösung auf einem unbeschränkten oder beschränkten Gebiet interessiert.

Im Gegensatz dazu beschreibt die Potentialgleichung eine *statische*, das heißt zeitunabhängige Funktion. Für zeitunabhängige Funktionen sind in der Regel schwächere Nebenbedingungen, nämlich sogenannte **Randbedingungen** natürlich. Im Beispiel der Potentialgleichung $\Delta u = 0$ zur Beschreibung der statischen Temperaturverteilung in einem einfach zusammenhängendem Gebiet $D \subset \mathbb{R}^n$ wird man erwarten, dass durch Messung der Temperatur am Rand ∂D auf die Verteilung in ganz D geschlossen werden kann. Entsprechend kann man zur PDGL eine sogenannte **Randbedingung 1. Art** oder **Dirichlet-Bedingung**:

$$u(\mathbf{x}) = f(\mathbf{x}) \quad \text{für } \mathbf{x} \in \partial D$$

mit bekannter Funktion f als zusätzliche Bedingung an die Lösung u vorgegeben. Werte für die ersten Ableitungen der gesuchten Funktion u werden *nicht* vorgegeben.

Andere Arten von Randbedingungen sind die **Neumann-Bedingungen** (auch **Randbedingung 2. Art** genannt):

$$\partial_\nu u(\mathbf{x}) = 0 \quad \text{für } \mathbf{x} \in \partial D\ ,$$

wobei $\partial_\nu u = \nabla u \cdot \nu$ die Richtungsableitung von u in Richtung der nach außen zeigenden Normalen ν von ∂D ist. Die anschauliche Bedeutung hiervon: der Rand von D ist isoliert, d.h. die Werte

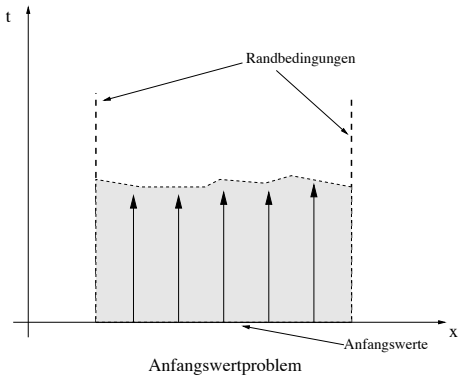
von u sollen sich über den Rand von D hinaus nicht ändern. Bei der **Randbedingung der 3. Art** wird

$$\partial_\nu u(\mathbf{x}) + \alpha u(\mathbf{x}) = f(\mathbf{x}) \quad \text{für } \mathbf{x} \in \partial D$$

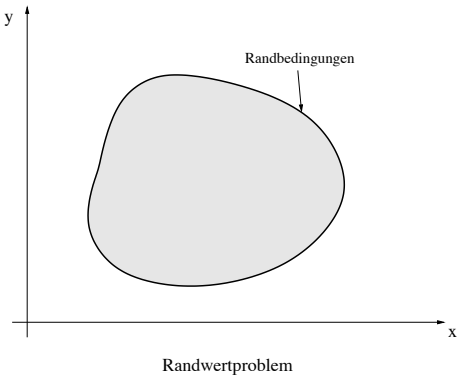
mit einem $\alpha \neq 0$ vorgegeben.

Die Unterscheidung in Rand- und Anfangswerte erscheint womöglich insofern künstlich, als Anfangswerte nichts anderes als Randwerte für die Variable t sind. Tatsächlich ist diese Unterscheidung sowohl in der Theorie als auch bei der Wahl numerischer Lösungsverfahren entscheidend. Bei dynamischen Vorgängen wird ähnlich wie bei AWP für gewöhnliche DGL versucht, das Verhalten der Lösungsfunktion u mit fortschreitender Zeit von den gegebenen Anfangswerten weg nachzuzeichnen („die PDGL in der Zeit zu integrieren"). Eines der Hauptprobleme dabei ist die Untersuchung der **Stabilität** eines numerischen Verfahrens, das heißt die Frage, ob die im Verlauf der Zeitintegration sich akkumulierenden Fehler beherrschbar klein bleiben und gegen Null gehen, wenn die Schrittweite gegen Null geht. Bei statischen Problemen mit zugehörigen Randwerten benötigt man hingegen Verfahren, die global auf dem ganzen Gebiet gleichzeitig gegen die Lösung konvergieren und insbesondere überall die richtigen Randwerte annehmen. Hierbei steht oft die Frage nach der **Effizienz** numerischer Verfahren im Vordergrund.

Wir stellen die unterschiedliche Situationen bei Anfangswertproblemen und Randwertproblemen in der nachfolgenden Skizzen gegenüber.



Anfangswertproblem



Randwertproblem

Man nennt ein Anfangs- oder Randwertproblem für eine PDGL **sachgemäß gestellt**, wenn

(a): eine Lösung existiert,

(b): die Lösung eindeutig ist und

(c): die Lösung stetig von den Nebenbedingungen abhängt.

Betrachten wir als Beispiel noch einmal die eindimensionale Wellengleichung

$$u_{tt} = c^2 u_{xx} \, .$$

Wie wir schon wissen, ist nach der d'Alembertschen Lösungsformel

$$u(x,t) = \frac{1}{2}\left(\varphi(x-ct) + \varphi(x+ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} \psi(\zeta)\, d\zeta \, .$$

die eindeutig festgelegte Lösung zu den Anfangsbedingungen

$$u(x,0) = \varphi(x), \quad u_t(x,0) = \psi(x) \, .$$

Ersetzen wir nun die Anfangsfunktionen φ und ψ durch $\varphi + \delta\varphi$ und $\psi + \delta\psi$ mit sup $(|\delta\varphi|) < \varepsilon$ und sup $(|\delta\psi|) < \varepsilon$ so bekommen wir eine neue Lösung $u + \delta u$ mit

$$|\delta u(x,t)| \leq \frac{1}{2}(\varepsilon + \varepsilon) + \frac{1}{2c}\int_{x-ct}^{x+ct} \varepsilon \, d\zeta = \varepsilon(1+t) \, .$$

Also ist das Cauchysche Anfangswertproblem für die eindimensionale Wellengleichung sachgemäß gestellt.

Die oben getroffene Unterscheidung zwischen dynamischen Problemen mit zugehörigen Anfangs-(Rand-)Werten und statischen Problemen mit zugehörigen Randwerten ist die für die numerische, praktische Behandlung von PDGL wichtigste. Daneben gibt es noch eine andere Art der Klassifikation speziell für semilineare PDGL 2. Ordnung, die wir zunächst für den Fall zweier unabhängiger Variablen x und y, also für PDGL der Bauart

$$au_{xx} + 2bu_{xy} + cu_{yy} = f(x, y, u, u_x, u_y) \tag{9.5}$$

angeben, wobei a, b und c stetige Funktionen in x und y und u eine gesuchte C^2-Funktion auf einem Gebiet $D \subseteq \mathbb{R}^2$ sein sollen. Als Spezialfall enthalten sind hierin die *linearen* PDGL 2.Ordnung

$$au_{xx} + 2bu_{xy} + cu_{yy} + du_x + eu_y + fu = g, \tag{9.6}$$

wobei $a, \ldots, g$ stetige Funktionen in x und y sind. Der Term $au_{xx} + 2bu_{xy} + cu_{yy}$ heißt der **Hauptteil** der (quasilinearen) PDGL 2.Ordnung.

Folgende Typen quasilinearer PDGL 2.Ordnung in zwei Variablen werden anhand ihres Hauptteils unterschieden:

$ac - b^2 > 0$:	die PDGL (9.5) heißt **elliptisch**,
$ac - b^2 = 0$:	die PDGL (9.5) heißt **parabolisch**,
$ac - b^2 < 0$:	die PDGL (9.5) heißt **hyperbolisch**.

Die Standardbeispiele für die verschiedenen Typen von PDGL:

- die Potentialgleichung $u_{xx} + u_{yy} = 0$ ist elliptisch ($a = 1$, $b = 0$, $c = 1$),
- die Diffusionsgleichung $u_t = \alpha^2 u_{xx}$ ist parabolisch ($a = \alpha^2$, $b = 0 = c$),
- die Wellengleichung $u_{tt} = \gamma^2 u_{xx}$ ist hyperbolisch ($a = \gamma^2$, $b = 0$, $c = -1$).

In diesen Fällen sind die Koeffizienten des Hauptteils konstant. Sind sie dies nicht, so kann es durchaus vorkommen, dass sich der Typ einer PDGL ortsabhängig (bzw. zeitabhängig) ändert. Ein Beispiel hierfür ist die **Tricomi-Gleichung**

$$u_{yy} - yu_{xx} = 0.$$

Diese ist hyperbolisch, falls $y > 0$, parabolisch, falls $y = 0$ und elliptisch, falls $y < 0$.

Eine Klassifikation wird auch für lineare PDGL in n unabhängigen Variablen $x_1, \ldots, x_n$ vorgenommen. Diese haben die Form

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu = f, \quad A = (a_{ij}) \quad \text{symmetrisch},$$

wobei a_{ij}, b_i, c und f stetige Funktionen in den x_i sind. Wegen $u_{x_i x_j} = u_{x_j x_i}$ für zweimal stetig differenzierbares u kann man die Matrix $A = (a_{ij})_{i,j=1}^n$ als symmetrisch voraussetzen. Im Fall zweier Variablen wäre also

$$A = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} .$$

Anhand der Eigenwerte der Matrix A ihres Hauptteils nennt man eine lineare PDGL

- **elliptisch**, wenn alle Eigenwerte von A das gleiche Vorzeichen haben,
- **parabolisch**, wenn genau ein Eigenwert von A verschwindet und alle anderen das gleiche Vorzeichen haben und
- **hyperbolisch**, wenn genau $n-1$ Eigenwerte von von A ein und dasselbe Vorzeichen und der verbleibende das entgegengesetzte Vorzeichen haben.

Diese Definition deckt sich mit der für den Fall $n = 2$ gegebenen. Sie ist wiederum punktweise zu verstehen, außer für konstante Koeffizienten. Man beachte, dass A als symmetrische reelle Matrix stets n reelle Eigenwerte hat.

Bemerkung. Manche PDGL werden durch diese Klassifikation nicht erfasst, zum Beispiel

$$u_t = u_{xy}$$

Solche PDGL haben in der Regel keine physikalische Bedeutung.

Der Typ einer PDGL ist entscheidend dafür, welche Art von Nebenbedingungen (Rand- oder Anfangswerte) die Wohlgestelltheit eines Problems garantieren. Dafür geben wir folgende *Faustregel* an.

- Zu *hyperbolischen PDGL* passt das *Cauchysche Anfangswertproblem*, d.h. Funktions- und erste Ableitungswerte der gesuchten Funktion $u(x_1, \ldots, x_n)$ werden auf einer $(n-1)$-dimensionalen Fläche H des $\mathbb{R}^n$ vorgeschrieben. Oft ist H die Hyperebene $x_1 = 0$, wenn $x_1 = t$ die Zeit bedeuten soll. Wenn H beschränkt ist, müssen an ihrem Rand ∂H (für alle Zeiten t, bzw. für alle x_1) zusätzliche Randwerte vorgegeben werden. Ausserdem darf H nicht beliebig gewählt sein.
- Zu *elliptischen PDGL* passt das *Randwertproblem*, bei dem die Lösung auf einem beschränkten Gebiet G gesucht ist. Auf dem Rand ∂G werden Randbedingungen der ersten, zweiten oder dritten Art vorgegeben werden, d.h. *entweder* Funktionswerte *oder* Ableitungswerte (oder eine feste Kombination aus beiden). An den Rand ∂G müssen in der Regel gewisse Glattheitsforderungen gestellt werden.
- Zu *parabolischen PDGL* passen vorgegebene Anfangswerte für die gesuchte Funktion u auf einer $(n-1)$-dimensionalen Fläche H. Ableitungswerte u_t werden nicht vorgegeben. Ist H beschränkt, kommen für ∂H Randbedingungen wie bei hyperbolischen PDGL hinzu.

Hier haben wir vorausgesetzt, die PDGL sei stets auf dem ganzen Gebiet, auf dem ihre Lösung gesucht ist, vom gleichen Typ. Das ist nicht immer der Fall und verkompliziert dann die Wahl von Nebenbedingungen.

9.2 Separationsansätze

Ein „Separationsansatz“ führt eine PDGL auf (mehrere) gewöhnliche DGL zurück. Diese Methode wurde von Fourier zur Lösung der Wärmeleitungsgleichung eingeführt und wir greifen genau dieses Beispiel auf.

Im einfachsten Fall der eindimensionalen Wärmeleitungsgleichung wird eine Funktion $u(x,t)$ gesucht, die folgende Gleichungen erfüllt

$$u_t = a^2 u_{xx}, \qquad a>0,\ t>0,\ 0\le x\le l, \qquad \text{(Homogene PDGL)} \tag{9.7}$$
$$u(x,0) = f(x), \qquad 0\le x\le l, \qquad \text{(Anfangswerte)} \tag{9.8}$$
$$u(0,t)=0, \quad u(l,t)=0, \qquad t\ge 0. \qquad \text{(Homogene Randwerte)} \tag{9.9}$$

Hier beschreibt $u(x,t)$ die Temperatur zum Zeitpunkt t im Punkt x eines Stabs der Länge l mit Enden $x=0$ und $x=l$. Die Temperatur an den Stabenden wird konstant auf 0 gehalten und die anfängliche Temperaturverteilung (zum Zeitpunkt $t=0$) im Stab wird durch die stetige Funktion f beschrieben. Natürlich wird Konsistenz von Anfangs- und Randbedingungen unterstellt:

$$f(0)=0 \quad \text{und} \quad f(l)=0.$$

Separation (oder **Trennung der Variablen**) bedeutet, dass wir eine spezielle Lösungsfunktion $u(x,t)$ unseres Problems suchen, die sich als Kombination zweier Funktionen $X=X(x)$ und $T=T(t)$ schreiben lässt. Entsprechende Separationsansätze sind beipielsweise die additive und die multiplikative Trennung der Variablen:

$$u(x,t)=X(x)+T(t) \quad \text{bzw.} \quad u(x,t)=X(x)\cdot T(t)\ .$$

In unserem Beispiel (wie in den meisten Fällen) ist es die multiplikative Trennung der Variablen, die zum Erfolg führt. Oft wird unter „Trennung der Variablen“ überhaupt nur die multiplikative Trennung verstanden.

Einsetzen einer Funktion $u(x,t)=X(x)T(t)$ liefert

$$(9.7) \quad \Leftrightarrow \quad X\dot{T} = a^2 X'' T \quad \Leftrightarrow \quad \frac{X}{X''} = a^2 \frac{T}{\dot{T}}.$$

Hier ist die linke Seite eine Funktion von x allein und die rechte Seite eine Funktion von t allein[1]. Wenn also eine Lösung $u=X\cdot T$ existieren soll, dann müssen beide Seiten konstante Funktionen sein. Ausserdem liefert Einsetzen von $u(x,t)=X(x)T(t)$

$$(9.9) \quad \Leftrightarrow \quad X(0)T(t)=0 \quad \text{und} \quad X(l)T(t)=0 \qquad \text{für alle } t\ge 0.$$

Da wir den Fall $T\equiv 0$ ausschließen wollen (sonst würden wir nur ein triviales $u\equiv 0$ bekommen können) erhalten wir, dass es eine Konstante λ geben muss, so dass

$$X''+\lambda X = 0, \qquad X(0)=0,\ X(l)=0 \tag{9.10}$$

sowie

$$\dot{T}+a^2\lambda T = 0. \tag{9.11}$$

(9.10) können wir im Prinzip wie in Kapitel 4 lösen: zunächst erhält man als Fundamentalsystem (das heißt als zwei linear unabhängige Lösungen der linearen homogenen DGL 2. Ordnung)

$$\begin{aligned} X_1 = e^{-\sqrt{-\lambda}x},\ X_2 = e^{\sqrt{-\lambda}x}, &\qquad \text{falls } \lambda<0,\\ X_1 = 1,\ X_2 = x, &\qquad \text{falls } \lambda=0,\\ X_1 = \cos(\sqrt{\lambda}x),\ X_2 = \sin(\sqrt{\lambda}x), &\qquad \text{falls } \lambda>0\ . \end{aligned}$$

[1] Bei Koeffizienten, die ebenfalls von x und t abhängen, muss also Separierbarkeit dieser Koeffizienten gefordert werden, um eine solche Situation herstellen zu können

Die allgemeine Lösung der DGL lautet somit

$$X(t) = c_1 X_1(t) + c_2 X_2(t), \quad c_1, c_2 \in \mathbb{R}$$

und die Konstanten sind so zu bestimmen, dass die Randbedingungen $X(0) = 0$ und $X(l) = 0$ erfüllt werden.

Man rechnet nach, dass für $\lambda \leq 0$ nur die (uninteressante) Lösung $X \equiv 0$ von (9.10) existiert. Nur im Fall $\lambda > 0$ und zwar genau für

$$\lambda = \lambda_n := n^2\pi^2/l^2, \quad n \in \mathbb{N},$$

gibt es nichttriviale Lösungen, nämlich die Funktionen

$$X_n(x) = \sin(\frac{n\pi}{l}x)$$

(und deren Vielfache $c \cdot X_n, c \in \mathbb{R}$). *Nur für solche Werte λ_n führt der Separationsansatz also auf nicht-triviale Lösungen* und dementsprechend lösen wir die zweite DGL (9.11) auch nur für diese Werte von λ. Die allgemeinen Lösungen dafür lauten (für $n \in \mathbb{N}$)

$$T_n(t) = c_n e^{-\alpha_n^2 t} \quad \text{mit} \quad \alpha_n = \frac{an\pi}{l}$$

und $c_n \in \mathbb{R}$, vergleiche (4.23) in Kapitel 4.

Bisher hat der Separationsansatz also auf die folgenden nichttrivialen Lösungen geführt

$$u_n(x,t) = c_n e^{-\alpha_n^2 t} \sin(\frac{n\pi}{l}x), \quad c_n \in \mathbb{R}, \quad n = 1, 2, \ldots,$$

die alle die Randbedingungen (9.9) erfüllen. Wir haben allerdings noch nicht die Anfangsbedingung (9.8) erfüllt — und keine der Funktionen u_n wird dies im allgemeinen tun. Deswegen benutzen wir jetzt noch die Linearität unseres Problems, um nach dem Superpositionsprinzip die Einzel-Lösungen u_n zu allgemeineren Lösungen linear zu kombinieren, die auch noch (9.9) erfüllen. Es reicht dabei allerdings nicht aus, nur endliche Linearkombinationen zu betrachten, vielmehr brauchen wir auch „unendliche Kombinationen“, also Funktionen der Gestalt:

$$u(x,t) = \sum_{n=1}^{\infty} c_n e^{-\alpha_n^2 t} \sin(\frac{n\pi}{l}x) \tag{9.12}$$

mit zu wählenden Koeffizienten c_n. Die Idee, die hierbei verfolgt wird, ist schnell klar: Einsetzen von $t = 0$ in (9.12) liefert eine Fourier-Sinus-Reihe

$$u(x,0) = \sum_{n=1}^{\infty} c_n \sin(\frac{n\pi}{l}x) \tag{9.13}$$

Wir entwickeln auch die Anfangsfunktion f in eine Fourier-Sinus-Reihe [2] Vorgehen: ungerade Fortsetzung von f von $[0, l]$ auf $[-l, 0]$ und dann $2l$-periodische Fortsetzung auf ganz $\mathbb{R}$ liefert Fourierkoeffizienten

$$c_n(f) = \frac{1}{2l}\int_{-l}^{l} f(t)e^{-2\pi int/(2l)}\,dt = \frac{-i}{l}\int_0^l f(t)\sin\left(\frac{\pi nt}{l}\right)\,dt\,.$$

Als ungerade reellwertige Funktion hat $f(x)$ eine reine Sinusreihen-Entwicklung

$$f(x) = \sum_{n=1}^{\infty} 2b_n \sin\left(\frac{2\pi nx}{2l}\right)$$

[2] Das geht unter geeigneten Voraussetzungen an f, zum Beispiel für eine stetige und stückweise stetig differenzierbare Funktion.

mit $2b_n = i(c_n - c_{-n})$, vergleiche (5.7) und (5.8). Man erhält also

$$f(x) = \sum_{n=1}^{\infty} f_n \sin(\frac{n\pi}{l}x), \quad x \in [0, l],$$

mit den Koeffizienten

$$f_n = 2b_n = \frac{2}{l}\int_0^l f(\tau) \sin\frac{n\pi\tau}{l}\, d\tau \ .$$

Also erfüllt $u(x,t)$ für die Wahl $c_n = f_n$ die Anfangsbedingung $u(x,0) = f(x)$, denn zwei stetige Funktionen mit identischen Fourier-Koeffizienten sind identisch (Eindeutigkeitssatz für Fourier-Reihen).

Zusatz. Zweierlei müsste noch gezeigt werden, nämlich dass erstens (9.12) für die Wahl $c_n = f_n$ tatsächlich eine konvergente Reihe in x und t darstellt und dass die so definierte Funktion zweitens die Wärmeleitungsgleichung $u_t = a^2 u_{xx}$ für $t > 0$ erfüllt. Für die zweite Eigenschaft muss man zeigen, dass die Reihe (9.12) gliedweise differenziert werden darf. Wir lassen hier beide Nachweise aus.

Bemerkung 1:
Obwohl die beschriebene Methode von Fourier Einschränkungen unterworfen ist — insbesondere funktioniert sie nur, wenn auch die Koeffizientfunktionen der PDGL sowie die Nebenbedingungen separierbar sind — stellt sie doch die wichtigste analytische Lösungsmethode für lineare PDGL 2.Ordnung dar.

Bemerkung 2:
Separationsansätze existieren analog für Funktionen in mehreren Variablen, etwa $u(x,y,t) = X(x) \cdot Y(y) \cdot T(t)$.

Bemerkung 3:
Oft ergeben sich erfolgreiche Separationsansätze erst nach geeigneten Variablentransformationen. Zum Beispiel wird man zur Lösung der Schwingungsgleichung für eine in einen Kreisring eingespannte Membran erst auf Polarkordinaten transformieren.

Zusatz: Der allgemeine, inhomogene Fall.

Die Gleichungen (9.7)-(9.9) beschreiben die Wärmeleitung in einem Stab, wenn keine Wärme von außen zugeführt wird und wenn die Randbedingungen homogen sind. Im allgemeinen Fall haben wir

$$u_t = a^2 u_{xx} + F(x,t), \qquad a > 0,\ t > 0,\ 0 \le x \le l, \qquad \text{(PDGL)} \qquad (9.14)$$

$$u(x,0) = f(x), \qquad 0 \le x \le l, \qquad \text{(Anfangswerte)} \qquad (9.15)$$

$$u(0,t) = g(t), \quad u(l,t) = h(t). \qquad \text{(Randwerte)} \qquad (9.16)$$

Hier ist $F(x,t)$ die äußere Temperaturzufuhr, die zu einer nicht mehr homogenen PDGL führt. Wir setzen F als stetig differenzierbar voraus. Die Temperatur an den Stabenden wird jetzt auf den vorgegebenen, aber beliebigen Werten $g(t)$ und $h(t)$ gehalten. Natürlich wird wieder Konsistenz der Nebenbedingungen gefordert:

$$f(0) = g(0) \quad \text{und} \quad f(l) = h(0).$$

Dieses Problem wird in drei Teilprobleme zerlegt, die nacheinander gelöst und zu einer Gesamtlösung zusammengesetzt werden.

1. Schritt:
Die zweimal stetig differenzierbare Funktion

$$w(x,t) := g(t) + \frac{x}{l}(h(t) - g(t))$$

erfüllt offenbar die Randbedingungen (9.16). Die gesuchte Gesamtlösung $u(x,t)$ kann dann geschrieben werden in der Form $u(x,t) = v(x,t) + w(x,t)$, wobei $v(x,t)$ eine Lösung sein muss von

$$\begin{aligned} u_t = a^2 u_{xx} + H(x,t) \qquad &\text{mit} \qquad H(x,t) := F(x,t) - w_t(x,t) + a^2 w_{xx}(x,t) \\ u(x,0) = \tilde{f}(x), \qquad &\text{mit} \qquad \tilde{f}(x) = f(x) - w(x,0) \\ u(0,t) = 0, \quad u(l,t) = 0. & \end{aligned}$$

2. Schritt:
Wir lösen das **homogene Teilproblem**

$$u_t = a^2 u_{xx} \qquad \text{(Homogene PDGL)} \tag{9.17}$$

$$u(x,0) = \tilde{f}(x), \qquad \text{(Anfangswerte)} \tag{9.18}$$

$$u(0,t) = 0, \quad u(l,t) = 0. \qquad \text{(Homogene Randwerte)} \tag{9.19}$$

Diesen Schritt haben wir oben schon erledigt (wobei wir f statt $\tilde{f}$ geschrieben haben).

3. Schritt:
Wir lösen das **halbhomogene Problem** (mit inhomogener PDGL, aber homogenen AW und RW)

$$u_t = a^2 u_{xx} + H(x,t) \qquad \text{(Inhomogene PDGL)} \tag{9.20}$$

$$u(x,0) = 0, \qquad \text{(Homogene Anfangswerte)} \tag{9.21}$$

$$u(0,t) = 0, \quad u(l,t) = 0. \qquad \text{(Homogene Randwerte)} \tag{9.22}$$

Wenn wir die Lösungen des homogene Teilproblems v_1 und die des halbhomogenen Teilproblems v_2 nennen, dann ergibt sich durch Einsetzen, dass $u = v_1 + v_2 + w$ eine Lösung des Originalproblems (9.14)-(9.16) ist. Jetzt führen wir nur noch den noch ausstehenden Schritt 3 aus.

Wenn wir die Steuerfunktion $H(x,t)$ der PDGL (9.20) für jedes $t \geq 0$ in eine Fourier-Sinus-Reihe auf $[0,l]$ entwickeln (ungerade und dann periodische Fortsetzung der Funktion in x wie oben für f), dann erhalten wir

$$H(x,t) = \sum_{n=1}^{\infty} H_n(t) \sin(\frac{n\pi}{l}x) \quad \text{mit} \quad H_n(t) = \frac{2}{l}\int_0^l H(\tau,t)\sin(\frac{n\pi\tau}{l})\,d\tau$$

(t spielt hier die Rolle eines Parameters — die Fourier-Reihe wird für jedes t aufgestellt).

Für die PDGL (9.20) bietet es sich dann an, den Ansatz vom Typ der rechten Seite

$$u(x,t) = \sum_{n=1}^{\infty} u_n(t) \sin(\frac{n\pi}{l}x) \tag{9.23}$$

zu probieren. Konvergenz und gliedweise Differenzierbarkeit dieser Reihe vorausgesetzt, liefert Einsetzen in (9.20) und anschließender Koeffizientenvergleich eine gewöhnliche DGL für jedes u_n:

$$\dot{u}_n = -\left(\frac{an\pi}{l}\right)^2 u_n + H_n \quad \text{mit AW} \quad u_n(0) = 0$$

wobei sich der Anfangswert aus Gleichung (9.21) und Koeffizientenvergleich ergibt:

$$u(x,0) = \sum_{n=1}^{\infty} u_n(0) \sin(\frac{n\pi}{l}x) = 0.$$

Dieses AWP für eine gewöhnliche lineare DGL 1. Ordnung kann einfach gelöst werden und seine Lösung lautet

$$u_n(t) = \int_0^t \exp\left(\left(\frac{an\pi}{l}\right)^2 (\tau - t)\right) H_n(\tau)\, d\tau.$$

Mit den so definierten Funktionen u_n wären jetzt noch die Konvergenz und gliedweise Differenzierbarkeit von (9.23) zu zeigen, was wir jedoch wieder auslassen.

9.3 Numerische Lösungsansätze

PDGL sind bedeutsam in technischen Anwendungen, aber nur in den seltensten Fällen analytisch lösbar, so dass numerischen Verfahren große Bedeutung zukommt. Die beiden bekanntesten Klassen numerischer Verfahren heißen „Finite Differenzen“ und „Methode der finiten Elemente“. Wir besprechen in aller Kürze das Verfahren der finiten Differenzen an zwei Beispielen, nämlich dem dynamischen Problem der Wärmeleitung in einem endlichen Stab und dem statischen Dirichlet-Problem für die **Helmholtz-Gleichung** auf dem Einheitsquadrat.

Differenzenverfahren überführen Differentialgleichungen näherungsweise in algebraische Gleichungen, indem Ableitungen durch Differenzenquotienten aproximiert werden. Zum Beispiel können für eine (zweimal) differenzierbare Funktion $f : \mathbb{R} \to \mathbb{R}$ auf dem Gitter

$$x_j := jh \quad \text{für} \quad j \in \mathbb{Z} \text{ und } h > 0$$

die Approximationen

$$f'(x_j) \approx \begin{cases} \dfrac{f(x_j+h) - f(x_j)}{h} & \text{oder} \\ \dfrac{f(x_j) - f(x_j-h)}{h} & \text{oder} \\ \dfrac{f(x_j+h) - f(x_j-h)}{2h} & \text{oder ...} \end{cases}$$

und

$$f''(x_j) \approx \frac{f(x_j-h) - 2f(x_j) + f(x_j+h)}{h^2}$$

(oder andere) benutzt werden.

Bei Funktionen in mehreren Variablen kann man dieselbe Technik für partielle Ableitungen benutzen. Man setze zum Beispiel

$$x_j = jh_x,\ y_k = kh_y \quad \text{für} \quad j,k \in \mathbb{Z} \text{ und } h_x, h_y > 0$$

und approximiere für zweimal differenzierbares $u : \mathbb{R}^2 \to \mathbb{R}$:

$$\begin{aligned} u_{xx}(x_j, y_k) &\approx \frac{u_x(x_j + h, y_k) - u_x(x_j, y_k)}{h} \\ &\approx \frac{\dfrac{u(x_j + h, y_k) - u(x_j, y_k)}{h} - \dfrac{u(x_j, y_k) - u(x_j - h, y_k)}{h}}{h} \\ &= \frac{u(x_j + h, y_k) - 2u(x_j, y_k) + u(x_j - h, y_k)}{h^2} \end{aligned}$$

Analoges funktioniert für u_{yy} und u_{xy}, wobei immer mehrere Möglichkeiten einer Approximation bestehen. Ebenso ist es natürlich möglich, dass die obigen Funktionen f und u nur auf Teilmengen von $\mathbb{R}$ bzw. $\mathbb{R}^2$ definiert sind.

Wärmeleitung

Zu lösen ist das Anfangsrandwertproblem

$$\begin{aligned} u_t(x,t) - a^2 u_{xx}(x,t) &= 0, \qquad x \in (0,1),\ t > 0, \\ u(x,0) &= f(x), \quad x \in [0,1], \\ u(0,t) &= g(t), \quad t \geq 0, \\ u(1,t) &= h(t), \quad t \geq 0. \end{aligned}$$

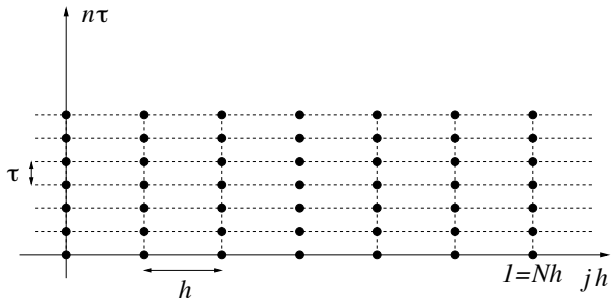
Grundlage des Differenzenverfahrens ist eine Diskretisierung des Gebiets $[0,1] \times \mathbb{R}_+$, auf dem die Lösung u des AWP gesucht ist. Dazu definieren wir mit einem $N \in \mathbb{N}$ die **Diskretisierungsfeinheiten**

$$\tau > 0 \quad \text{und} \quad h = \frac{1}{N} > 0$$

und damit das **Gitter**

$$\Omega_h^\tau := \{(x_j, t_n) = (jh, n\tau), \quad j = 0, \ldots, N,\ n \in \mathbb{N}_0\} ,$$

siehe die folgende Skizze



Eine **Gitterfunktion** $U : \Omega_h^\tau \to \mathbb{R}$, gegeben durch die Werte $U_j^n := U(x_j, t_n)$, soll die Lösung u des AWP in den Gitterpunkten approximieren:

$$U_j^n \approx u_j^n := u(x_j, t_n) .$$

Dazu können wir (neben anderen Möglichkeiten!) die Approximationen

$$u_t(x_j, t_n) \approx \frac{u_j^{n+1} - u_j^n}{\tau} \quad \text{und} \quad u_{xx}(x_j, t_n) \approx \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2} .$$

heranziehen. Entsprechend wird von der Gitterfunktion U gefordert

$$\begin{aligned}
&\frac{U_j^{n+1}-U_j^n}{\tau} = a^2\frac{U_{j+1}^n-2U_j^n+U_{j-1}^n}{h^2}, && j=0,\dots,N,\ n\in\mathbb{N}_0\\
&U_j^0 = F_j := f(x_j), && j=0,\dots,N\\
&U_0^n = G_n := g(t_n), && n=0,1,\dots\\
&U_N^n = H_n := h(t_n), && n=0,1,\dots
\end{aligned}$$

Durch Multiplikation mit τ erhalten wir hieraus mit der Abkürzung

$$\lambda := a^2\frac{\tau}{h^2} > 0$$

die Gleichung

$$U_j^{n+1} = \lambda U_{j-1}^n + (1-2\lambda)U_j^n + \lambda U_{j+1}^n\ . \tag{9.24}$$

Es handelt sich bei (9.24) um ein explizites Verfahren: die Gitterfunktion U ist zum Zeitpunkt $t_0=0$ bekannt und wird sukzessive für alle diskretisierten Zeitpunkte $t_1,t_2,\dots$ berechnet. Schreibt man $U^n=(U_j^n)_{j=0,\dots,N}$ für die Folge aller Werte von U zum Zeitpunkt $t=t_n$, dann wird also U^{n+1} aus U^n berechnet in Analogie zum Eulerverfahren bei AWP für gewöhnliche DGL.

Nimmt man an, dass $\lambda\le\frac{1}{2}$, dann ist

$$|U_j^{n+1}| \overset{(9.24)}{\le} \lambda|U_{j-1}^n| + (1-2\lambda)|U_j^n| + \lambda|U_{j+1}^n| \le \sup\{|U_j^n|,|G_n|,|H_n|;\ j=0,\dots,N\}\ .$$

Per Induktion folgt hieraus, dass

$$|U_j^n| \le \sup\{|F_j|,|G_m|,|H_m|;\ j=0,\dots,N, m=0,\dots,n\}\quad \forall j=0,\dots,N,\ n\in\mathbb{N}_0\ . \tag{9.25}$$

Die Werte $|U_j^n|$ können also nicht beliebig anwachsen (sie sind stetst durch die vorgegebenen Anfangs- und Randwerte beschränkt). Diese Eigenschaft nennt man **Stabilität**.

Für $\lambda > 1/2$ geht die Stabilität verloren. Betrachtet man etwa die Anfangswerte $F_j=(-1)^j\varepsilon$ für ein beliebig kleines $\varepsilon>0$, so kann man mit Induktion zeigen, dass

$$U_j^n = (1-4\lambda)^n(-1)^j\varepsilon \quad\Longrightarrow\quad |U_j^n| = (4\lambda-1)^n\ .$$

Offenbar ist

$$|U_j^n|\to\infty \quad\text{für}\quad n\to\infty \quad\text{falls}\quad \lambda > \frac{1}{2}\ .$$

Fehlende Stabilität hat zur Folge, dass kleine Änderungen der Anfangswerte oder auch Rundungsfehler, die bei endlicher Rechengenauigkeit unvermeidlich sind, im Lauf der Zeitintegration zu beliebiger Größe anwachsen können. Das Verfahren (9.24) funktioniert also nur, wenn die Bedingung $\lambda\le 1/2$ eingehalten wird, was eine arge Beschränkung für die Schrittweite τ darstellt. Das Verfahren (9.24) wird deswegen in der Praxis durch bessere Verfahren ersetzt.

Dirichlet-Problem für Helmholtz-Gleichung

Wir setzen $\Omega := (0,1)\times(0,1)\subset\mathbb{R}^2$. Für eine Konstante $\sigma\ge 0$ und ein stetiges $f:\Omega\to\mathbb{R}$ ist die Lösung $u:[0,1]\times[0,1]\to\mathbb{R}$ des folgenden Randwertproblems gesucht:

$$\begin{aligned}
-u_{xx}(x,y)-u_{yy}(x,y)+\sigma u(x,y) &= f(x,y), && \text{für } (x,y)\in\Omega,\\
u(x,y) &= 0, && \text{für } (x,y)\in\partial\Omega.
\end{aligned}$$

Zur Diskretisierung wird für $n \in \mathbb{N}$ und $h = 1/n$ ein Gitter eingeführt:

$$(x_i, y_j) = (ih, jh), \quad i, j = 0, \ldots, n \; .$$

Zu bestimmen sind die Werte $u_{i,j} = u(x_i, y_j)$ für die „inneren Gitterpunkte“ (mit $i = 1, \ldots, n-1$ und $j = 1, \ldots, n-1$), während aufgrund der vorgegebenen Randwerte

$$u_{0,j} = u_{n,j} = u_{i,0} = u_{i,n} = 0$$

schon fest stehen. Zweite Ableitungen werden durch finite Differenzen approximiert:

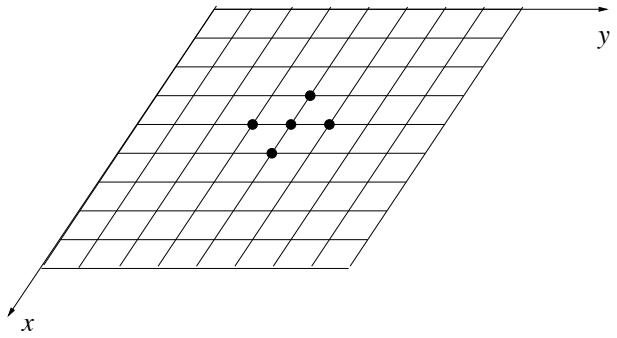
$$\begin{aligned} u_{xx}(x,y) &\approx \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} \\ u_{yy}(x,y) &\approx \frac{u(x,y+h) - 2u(x,y) + u(x,y+h)}{h^2} = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h^2}. \end{aligned}$$

Näherungen $v_{i,j} \approx u_{i,j}$ bestimmen wir dann als Lösung des Gleichungssystems

$$-\frac{v_{i+1,j} - 2v_{i,j} + v_{i-1,j}}{h^2} - \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{h^2} + \sigma v_{i,j} = f_{i,j} \; , \tag{9.26}$$

wobei $f_{i,j} = f(x_i, y_j)$ und $i, j = 1, \ldots, n-1$.

Die folgende Skizze illustriert das zweidimensionale Gitter und kennzeichnet für einen Gitterpunkt (x_i, y_j) seine vier, nördlich, südlich, westlich und östlich gelegenen Nachbarpunkte, von denen $v_{i,j}$ gemäß (9.26) abhängt.

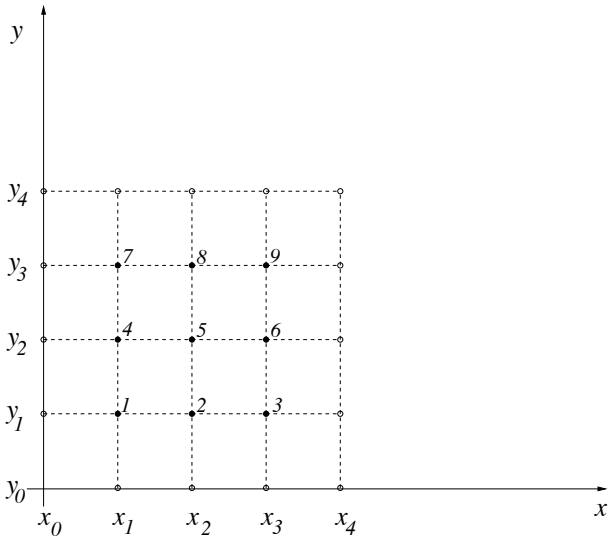


Statt des doppelten kann man einen eindimensionalen Index einführen, indem man zum Beispiel die Gitterpunkte zeilenweise nummeriert. Das führt auf Vektoren $\mathbf{v} \in \mathbb{R}^{(n-1)^2}$ und $\mathbf{f} \in \mathbb{R}^{(n-1)^2}$ für die Unbekannten und die rechte Seite und nach Multiplikation von (9.26) mit h^2 auf ein Gleichungssystem

$$A^h \mathbf{u}^h = h^2 \mathbf{f}^h$$

mit einer Matrix $A^h \in \mathbb{R}^{(n-1)^2,(n-1)^2}$.

Beispiel(e) 9.8
Im Fall $n = 4$ ergeben sich 9 innere Gitterpunkte, die zeilenweise durchnummeriert werden.



Dazu gehört im Fall $\sigma = 0$ das LGS

$$\underbrace{\begin{pmatrix} 4 & -1 & & -1 & & & & & \\ -1 & 4 & -1 & & -1 & & & & \\ & -1 & 4 & & & -1 & & & \\ -1 & & & 4 & -1 & & -1 & & \\ & -1 & & -1 & 4 & -1 & & -1 & \\ & & -1 & & -1 & 4 & & & -1 \\ & & & -1 & & & 4 & -1 & \\ & & & & -1 & & -1 & 4 & -1 \\ & & & & & -1 & & -1 & 4 \end{pmatrix}}_{A^h} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \end{pmatrix} = h^2 \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9 \end{pmatrix}$$

Für allgemeines n und $\sigma \geq 0$ haben wir mit der Hilfsmatrix

$$T_n = \begin{pmatrix} 2+\sigma h^2 & -1 & & & \\ -1 & 2+\sigma h^2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+\sigma h^2 & -1 \\ & & & -1 & 2+\sigma h^2 \end{pmatrix} \in \mathbb{R}^{n-1,n-1}$$

und mit der Einheitsmatrix $I_{n-1} \in \mathbb{R}^{n-1,n-1}$ die Block-Tridiagonalmatrix

$$A^h = \begin{pmatrix} T_n + 2I_{n-1} & -I_{n-1} & & & \\ -I_{n-1} & T_n + 2I_{n-1} & -I_{n-1} & & \\ & \ddots & \ddots & \ddots & \\ & & -I_{n-1} & T_n + 2I_{n-1} & -I_{n-1} \\ & & & -I_{n-1} & T_n + 2I_{n-1} \end{pmatrix} \in \mathbb{R}^{(n-1)^2,(n-1)^2} .$$

Die Matrix A^h hat spezielle Eigenschaften:

- sie ist positiv definit

- sie ist **dünn besiedelt**, das heißt die meisten Komponenten sind gleich Null
- sie entspricht einer Gitterstruktur

In der Numerik wurden sehr effiziente Verfahren entwickelt, um Gleichungssysteme dieser Art zu lösen, die sogenannten **Mehrgittermethoden**.

Kapitel 10

Stochastik

10.1 Einführung: Wahrscheinlichkeit und Information

Die beiden Begriffe **Nachricht** und **Information** sind Bestandteile unserer Umgangssprache, die nicht immer genau auseinandergehalten werden. Im Rahmen der Ingenieurwissenschaften und der Mathematik ist es aber unabdingbar, diese beiden Begriffe scharf zu unterscheiden. Eine Nachricht ist zunächst etwas, das stets von einem Sender ausgeht und in eine spezielle physikalische Form gebracht wird. Diese physikalische Form hängt von der Art und Weise ab, wie die entsprechende Nachricht vom Sender zu den vorgesehenen Empfängern übertragen werden soll. Bei den Indianerstämmen Nordamerikas wurden Nachrichten zum Beispiel durch spezielle Rauchzeichen übertragen. Seit etwa 1817 werden in der Schifffahrt Nachrichten unter anderem durch Flaggensignale ausgetauscht. Die Darstellung einer Nachricht in Abhängigkeit von der vorgesehenen Art der Übertragung spielt in der Kommunikationstechnik somit eine wichtige Rolle.
In der Physik gibt es Größen (zum Beispiel die **Zeit** oder die **Masse**), die als Grundbausteine der Naturbeschreibung gelten und deshalb nicht definiert werden; allerdings kann man wesentliche Eigenschaften dieser Größen benennen und man kann diese Basisgrößen auch messen.
Mit dem Begriff **Information** verhält es sich analog; eine formale Definition ist ebenfalls nicht sinnvoll, aber man kann Eigenschaften angeben und wir werden die **Informationsmenge** (oder dazu synonym den **Informationsgehalt**) im mathematischen Sinne messen. Während eine Nachricht in einem Sender entsteht und dann zu einem Empfänger übertragen wird, entsteht Information immer in einem Empfänger durch den Erhalt einer Nachricht - und zwar dann, wenn der Inhalt der empfangenen Nachricht für den Empfänger mit einem Überraschungseffekt verbunden ist. Je größer die Überraschung, die der Inhalt einer Nachricht beim Empfänger auslöst, desto mehr Informationen hat der Empfänger durch diese Nachricht erhalten (desto größer ist also die übertragene Informationsmenge). Es macht Sinn, bei der Quantifizierung des Überraschungseffekts, den eine Nachricht bei einem Empfänger auslöst, eine Vorgehensweise zu wählen, die die Anwendung einer fortgeschrittenen mathematischen Theorie erlaubt. Aus diesem Grund hat man sich entschieden, die Wahrscheinlichkeit zu betrachten, mit der ein Empfänger den Inhalt einer Nachricht erwartet. Je kleiner diese Wahrscheinlichkeit ist, desto größer ist die Überraschung beim Erhalt der Nachricht (und damit desto größer der Informationsgehalt der Nachricht für diesen Empfänger). Dadurch werden für eine zu entwickelnde Informationstheorie die Ergebnisse der Stochastik nutzbar.
Da wir von einer empfangenen Nachricht für die Bestimmung ihres Informationsgehalts nur noch die Wahrscheinlichkeit betrachten, mit der ein Empfänger diese Nachricht erwartet hat, können wir den Begriff **Nachricht** sehr weit fassen. Eine Nachricht ist jedes Ereignis, dass mit einer gewissen Wahrscheinlichkeit auftritt. Im Folgenden werden wir deshalb einen Zusammenhang zwischen einer Wahrscheinlichkeit, also einer reellen Zahl aus dem Intervall $[0, 1]$, und der dazugehörigen Informationsmenge herstellen.
Gesucht ist eine Funktion I definiert auf dem Intervall $[0, 1]$, die jeder Wahrscheinlichkeit $p \in [0, 1]$ eine Informationsmenge $I(p)$ zuordnet; von dieser Funktion I werden gewisse Eigenschaften gefordert:

(I1) Die Funktion $I : [0,1] \to [0,\infty]$ soll auf dem offenen Intervall $(0,1)$ stetig sein.
Diese Forderung bedarf wohl keiner Erklährung. Niemand wird ernsthaft Unstetigkeiten fordern oder zulassen wollen.

(I2) $I\left(\frac{1}{2}\right) = 1$.
Man kann nur messen, wenn man eine Einheit festgelegt hat (wie zum Beispiel das Urmeter als Einheit der Längenmessung in Paris). Diese Forderung legt nun als Einheit die Informationsmenge Eins für die Wahrscheinlichkeit $\frac{1}{2}$ fest.

(I3) $I(pq) = I(p) + I(q)$ für alle $p, q \in (0,1)$.
Tritt ein Ereignis A mit der Wahrscheinlichkeit p auf und tritt ein Ereignis B mit der Wahrscheinlichkeit q auf, so gelten diese Ereignisse als stochastisch unabhängig, wenn das gemeinsame Auftreten von A und B mit Wahrscheinlichkeit pq erfolgt. In diesem Fall beeinflusst das Auftreten von A nicht die Wahrscheinlichkeit für das Auftreten von B und umgekehrt. Es ist daher sinnvoll, die Informationsmenge, die das gemeinsame Auftreten von A und B beinhaltet, als Summe der einzelnen Informationsmengen (von A und von B) festzulegen.

(I4) $I(0) = \lim\limits_{\substack{p\to 0\\ p\in(0,1)}} I(p)$, $I(1) = \lim\limits_{\substack{p\to 1\\ p\in(0,1)}} I(p)$.
Diese vierte Forderung ist wiederum ein Stetigkeitsargument.

Nun soll in einem ersten Resultat gezeigt werden, dass die Funktion I auf dem Intervall $(0,1)$ durch die ersten drei Eigenschaften eindeutig festgelegt ist.

Satz 10.1 (Eindeutigkeit der Funktion I)
Es gibt genau eine Funktion

$$h : (0,1) \to (0,\infty)$$

mit:
(i) h ist stetig.
(ii) $h\left(\frac{1}{2}\right) = 1$.
(iii) $h(pq) = h(p) + h(q)$ für alle $p, q \in (0,1)$.
Diese Funktion ist die Umkehrfunktion zu

$$f : (0,\infty) \to (0,1), \quad x \mapsto 2^{-x}$$

und damit der negative Logarithmus dualis auf dem Intervall $(0,1)$ (bezeichnet mit: $-\operatorname{ld}_{(0,1)}$).
Es gilt:

$$\lim_{\substack{x\to 0\\ x>0}} x \cdot (-\operatorname*{ld}_{(0,1)}(x)) = 0.$$

Aus diesem Resultat folgt, dass unsere gesuchte Funktion I auf dem Intervall $(0,1)$ durch die Funktion $-\operatorname{ld}_{(0,1)}$ festgelegt ist. Da

$$\begin{aligned} \lim_{\substack{p\to 1\\ p\in(0,1)}} (-\operatorname*{ld}_{(0,1)}(p)) &= -\operatorname{ld}(1) = 0 \quad \text{und} \\ \lim_{\substack{p\to 0\\ p\in(0,1)}} (-\operatorname*{ld}_{(0,1)}(p)) &= \lim_{\substack{p\to 0\\ p\in(0,1)}} (-\operatorname{ld}(p)) = \infty, \end{aligned}$$

folgt:

$$I(0) = \infty \quad \text{und} \quad I(1) = 0.$$

Mit der in der Maßtheorie üblichen Festlegung

$$\infty + a = a + \infty = \infty \quad \text{für alle} \quad a \in \mathbb{R} \cup \{\infty\}$$

gilt sogar

$$I(pq) = I(p) + I(q) \quad \text{für alle} \quad p, q \in [0,1].$$

Die Informationsmenge besitzt auch eine Einheit; sie wird in **bit** gemessen. Diese Wahl ist naheliegend, wenn man sich folgende Spezialfälle betrachtet, wobei der Index „b“ bedeutet, dass das Binärsystem zugrunde gelegt ist:

$$\begin{aligned} I\left(\frac{1}{2}\right) &= I(0.1_b) = 1 \text{ bit}, \\ I\left(\left(\frac{1}{2}\right)^k\right) &= I(0.\underbrace{0\ldots01}_{k \text{ Stellen}}{}_b) = k \text{ bit}, \\ 3 \text{ bit} &= I\left(\frac{1}{8}\right) = I(0.001_b) \leq \\ &\leq I(0.\overline{0001}_b) = I\left(\frac{1}{15}\right) = 3.9069.. \text{ bit} < \\ &< I(0.0001_b) = 4 \text{ bit}. \end{aligned}$$

Die Informationsmenge einer Zahl $p \in (0,1]$ kann also mit

$$\lfloor p \rfloor := \min\left\{k \in \mathbb{N}_0;\ \left(\frac{1}{2}\right)^k \geq p\right\}$$

durch

$$\lfloor p \rfloor \leq I(p) < \lfloor p \rfloor + 1$$

abgeschätzt werden.
Da sich die Funktionen $-\operatorname{ld}_{(0,1)}$ und $-\operatorname{ld}$ auf dem Intervall $(0,1)$ nicht unterscheiden, verwenden wir im Folgenden nur noch die Funktion $-\operatorname{ld}$ bzw. ld. Hätten wir in Forderung [I2] für $\zeta > 1$ statt $I\left(\frac{1}{2}\right) = 1$ die Forderung $I\left(\frac{1}{\zeta}\right) = 1$ aufgestellt, so hätten wir als Ergebnis statt dem Logarithmus dualis den Logarithmus zur Basis ζ erhalten.
Um uns vom Begriff **Informationsmenge** gegeben durch die Funktion I eine Vorstellung machen zu können, betrachten wir folgendes Szenario:
Am 31. Mai 2010 erhalten zwei Personen, A und B, von einer dritten Person - nennen wir sie C - die Nachricht, dass heute Bundespräsident Horst Köhler zurückgetreten ist. Person A wusste das bereits, während Person B nichts wusste und den Rücktritt eines Bundespräsidenten für unmöglich hielt. Ein und dieselbe Nachricht beinhaltet somit für die beiden Personen A und B völlig unterschiedliche Mengen an Information. Für Person A war die Wahrscheinlichkeit p_A, dass Horst Köhler zurücktritt, in dem Moment, als sie die Nachricht von Person C erhält, gleich Eins, denn sie kannte den Inhalt der Nachricht bereits. Somit war die Nachricht mit keinerlei Information verbunden:

$$I(p_A) = I(1) = -\operatorname{ld}(1) = 0.$$

Für Person B war die Überraschung unendlich groß, da sie diesen Rücktritt für unmöglich hielt ($p_B = 0$):

$$I(p_B) = I(0) = \lim_{\substack{x \to 0 \\ x > 0}} -\operatorname{ld}(x) = \infty.$$

Person C hatte eine weitere Nachricht parat, nämlich dass ebenfalls an diesem Tag die israelische Armee einen Schiffskonvoi des Free Gaza Movement geentert hatte. Beide Personen A und B haben mit Wahrscheinlichkeit $q_A = q_B = 0.75$ mit dieser Handlung gerechnet, da der Staat Israel dieses Vorgehen bereits mehrfach angekündigt hatte, wussten aber noch nichts davon. Intuitiv wird man die Gesamtmenge an Information, die die Person A durch diese beiden Nachrichten erhalten hat, auf

$$I(1) + I(0.75) = 0 - \operatorname{ld}(0.75) \approx 0.415 \text{ bit}$$

festlegen. Dies liegt daran, dass sich beide Ereignisse (Rücktritt des Bundespräsidenten und Militäraktion Israels) gegenseitig nicht beeinflussen. Die Wahrscheinlichkeit für das Eintreten beider Ereignisse ist somit gleich $p_A q_A$ für Person A bzw. $p_B q_B$ für Person B und es gilt wegen (iii) für Person A:

$$I(p_A q_A) = I(p_A) + I(q_A) = 0 - \operatorname{ld}(0.75) = -\operatorname{ld}(0.75) \approx 0.415 \text{ bit.}$$

Wie sieht nun die Gesamtmenge an Information für Person B aus? Wegen $0 \cdot 0.75 = 0$ und wegen der Festlegung

$$\infty + a = a + \infty = \infty \quad \text{für alle} \quad a \in \mathbb{R} \cup \{\infty\}$$

gilt:

$$\infty = I(0) = I(p_B q_B) = I(0 \cdot 0.75) = I(0) + I(0.75) = \infty - \operatorname{ld}(0.75) = \infty.$$

10.2 Diskrete Wahrscheinlichkeitsräume

In diesem Abschnitt betrachten wir Zufallsexperimente, also Experimente, von denen man zwar einerseits genau weiß, welche Ergebnisse möglich sind, man andererseits bei der Durchführung des Experiments ein Ergebnis im Allgemeinen nicht exakt, sondern nur mit einer bestimmten Wahrscheinlichkeit vorhersagen kann. Das Besondere an den Zufallsexperimenten dieses Abschnitts ist nun, dass die nichtleere Menge der möglichen Ergebnisse, die stets mit Ω bezeichnet wird, nur endlich viele oder abzählbar unendlich viele Elemente enthalten darf. Es gibt also eine Teilmenge $N \subseteq \mathbb{N}$ derart, dass eine Bijektion $N \to \Omega$ existiert. Ist nun für jedes $\omega \in \Omega$ die Wahrscheinlichkeit $P(\{\omega\})$ dafür bekannt, dass wir als Ergebnis des Zufallsexperiments ω erhalten, so können wir jeder Teilmenge $A \subseteq \Omega$ von Ω durch

$$P(A) := \sum_{\omega \in A} P(\{\omega\})$$

eine Wahrscheinlichkeit dafür zuordnen, dass sich das Ergebnis des Zufallsexperiments in der Menge A befindet. Fordert man naheliegender Weise $P(\Omega) = 1$ und $P(\emptyset) = 0$, so erhält man eine Abbildung P mit folgenden Eigenschaften:

(P1) $P : \mathcal{P}(\Omega) \to [0, 1]$, wobei $\mathcal{P}(\Omega)$ die Potenzmenge von Ω bezeichnet.

(P2) $P(\emptyset) = 0$, $P(\Omega) = 1$.

(P3) Für jede Folge $\{A_i\}_{i \in \mathbb{N}}$ paarweise disjunkter Mengen mit $A_i \in \mathcal{P}(\Omega)$, $i \in \mathbb{N}$, gilt:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Eine Teilmenge $A \subseteq \Omega$ wird als **Ereignis** bezeichnet; daher heißt die Potenzmenge von Ω auch **Ereignismenge**. Ein einelementiges Ereignis $\{\omega\}$ heißt **Elementarereignis**; man berechnet also stets die Wahrscheinlichkeit von Ereignissen. Zusammenfassend verwenden wir als mathematisches Objekt für ein Zufallsexperiment mit höchstens abzählbar vielen verschiedenen Ergebnissen das Tripel $(\Omega, \mathcal{P}(\Omega), P)$, das einen Spezialfall eines später allgemein zu definierenden Wahrscheinlichkeitsraumes darstellt.

Definition 10.2 (diskreter Wahrscheinlichkeitsraum)
Sei Ω eine nichtleere höchstens abzählbare Menge (es gibt also eine Teilmenge $N \subseteq \mathbb{N}$ derart, dass eine Bijektion $N \to \Omega$ existiert). Sei ferner $\mathcal{P}(\Omega)$ die Potenzmenge (also die Menge aller Teilmengen) von Ω und sei P eine Abbildung mit folgenden Eigenschaften:

(P1) $P : \mathcal{P}(\Omega) \to [0,1]$,

(P2) $P(\emptyset) = 0$, $P(\Omega) = 1$,

(P3) für jede Folge $\{A_i\}_{i\in\mathbb{N}}$ paarweise disjunkter Mengen mit $A_i \in \mathcal{P}(\Omega)$, $i \in \mathbb{N}$, gilt:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),$$

dann wird $(\Omega, \mathcal{P}(\Omega), P)$ als **diskreter Wahrscheinlichkeitsraum** bezeichnet.
Ω wird als **Ergebnismenge**, $\mathcal{P}(\Omega)$ als **Ereignismenge** und P als **Wahrscheinlichkeitsmaß** auf $\mathcal{P}(\Omega)$ bezeichnet.
Für $A \in \mathcal{P}(\Omega)$ heißt die reelle Zahl $P(A)$ **Wahrscheinlichkeit** für A.

Beim Roulette erhält man für ein Spiel die Ergebnismenge

$$\Omega = \{0, 1, 2, \ldots, 35, 36\}.$$

Üblicherweise legt man für die Elementarereignisse die Wahrscheinlichkeiten

$$P(\{\omega\}) = \frac{1}{37} \quad \left(= \frac{1}{|\Omega|}\right), \quad \omega \in \Omega,$$

fest, wobei $|A|$ die Anzahl der Elemente einer Menge A (Mächtigkeit von A) bezeichnet. Somit erhalten wir durch

$$P : \mathcal{P}(\Omega) \to [0,1], \quad A \mapsto \sum_{\omega\in A} P(\{\omega\}) \quad \left(= \frac{|A|}{|\Omega|}\right)$$

ein Wahrscheinlichkeitsmaß auf $\mathcal{P}(\Omega)$. Jeder Spieler, der am Roulettetisch das Ergebnis eines Spiels zur Kenntnis nimmt, erhält dadurch die Informationsmenge

$$I\left(\frac{1}{37}\right) = -\operatorname{ld}\left(\frac{1}{37}\right) = \operatorname{ld}(37) \approx 5.21 \text{ bit}.$$

Betrachten wir nun das Ergebnis von acht Spielen, so wird man

$$\Omega_8 = \{0, 1, 2, \ldots, 35, 36\}^8$$

wählen und die Wahrscheinlichkeiten für die Elementarereignisse folgendermaßen festlegen:

$$P(\{\omega\}) = \frac{1}{37^8} \quad \left(= \frac{1}{|\Omega_8|}\right), \quad \omega \in \Omega_8.$$

Somit erhalten wir durch

$$P : \mathcal{P}(\Omega_8) \to [0,1], \quad A \mapsto \sum_{\omega\in A} P(\{\omega\}) \quad \left(= \frac{|A|}{|\Omega_8|}\right)$$

ein Wahrscheinlichkeitsmaß auf $\mathcal{P}(\Omega_8)$. Jeder Spieler, der am Roulettetisch die Ergebnisse von acht Spielen zur Kenntnis nimmt, erhält dadurch die Informationsmenge

$$I\left(\frac{1}{37^8}\right) = -\operatorname{ld}\left(\frac{1}{37^8}\right) = 8 \cdot \operatorname{ld}(37) \approx 41.68 \text{ bit}.$$

Nun ist es beim Roulette möglich, auf das Ereignis „gerade natürliche Zahl“, also auf das Ereignis $G := \{2, 4, 6, \ldots, 34, 36\}$ zu setzen. Ein Spieler setzt in jedem der acht Spiele auf das Ereignis G und will natürlich wissen, mit welcher Wahrscheinlichkeit er m-mal gewinnt ($m = 0, 1, 2, \ldots, 7, 8$). Betrachtet man die Abbildung

$$X : \Omega_8 \to \{0, 1, 2, \ldots, 8\},$$

die zählt, wie oft in einem Tupel $\omega \in \Omega_8$ eine gerade natürliche Zahl vorkommt, so ergibt sich durch

$$P_X(\{i\}) := P\left(\{\omega \in \Omega_8;\, X(\omega) = i\}\right) = \binom{8}{i} \left(\frac{18}{37}\right)^i \left(\frac{19}{37}\right)^{8-i},\ i = 0, \ldots, 8,$$

(Binomialverteilung) ein Wahrscheinlichkeitsmaß P_X auf $\mathcal{P}(\{0, 1, 2, \ldots, 7, 8\})$, wobei

$$\binom{8}{i} := \frac{8!}{(8-i)!i!} \quad \text{(Binomialkoeffizient)}.$$

Erzählt nun der Spieler seiner Frau nicht die einzelnen Ergebnisse der acht Spiele, sondern nur, wieviele von diesen acht Spielen er gewonnen hat, so ergeben sich für die Frau die folgenden möglichen Informationsmengen:

$$\begin{aligned}
I(P_X(\{0\})) &= -8 \cdot \mathrm{ld}\left(\frac{19}{37}\right) \approx 7.692 \text{ bit},\\
I(P_X(\{1\})) &= -\mathrm{ld}(8) - \mathrm{ld}\left(\frac{18}{37}\right) - 7 \cdot \mathrm{ld}\left(\frac{19}{37}\right) \approx 4.770 \text{ bit},\\
I(P_X(\{2\})) &= -\mathrm{ld}(28) - 2\,\mathrm{ld}\left(\frac{18}{37}\right) - 6 \cdot \mathrm{ld}\left(\frac{19}{37}\right) \approx 3.041 \text{ bit},\\
I(P_X(\{3\})) &= -\mathrm{ld}(56) - 3\,\mathrm{ld}\left(\frac{18}{37}\right) - 5 \cdot \mathrm{ld}\left(\frac{19}{37}\right) \approx 2.119 \text{ bit},\\
I(P_X(\{4\})) &= -\mathrm{ld}(70) - 4\,\mathrm{ld}\left(\frac{18}{37}\right) - 4 \cdot \mathrm{ld}\left(\frac{19}{37}\right) \approx 1.875 \text{ bit},\\
I(P_X(\{5\})) &= -\mathrm{ld}(56) - 5\,\mathrm{ld}\left(\frac{18}{37}\right) - 3 \cdot \mathrm{ld}\left(\frac{19}{37}\right) \approx 2.275 \text{ bit},\\
I(P_X(\{6\})) &= -\mathrm{ld}(28) - 6\,\mathrm{ld}\left(\frac{18}{37}\right) - 2 \cdot \mathrm{ld}\left(\frac{19}{37}\right) \approx 3.353 \text{ bit},\\
I(P_X(\{7\})) &= -\mathrm{ld}(8) - 7\,\mathrm{ld}\left(\frac{18}{37}\right) - \mathrm{ld}\left(\frac{19}{37}\right) \approx 5.238 \text{ bit},\\
I(P_X(\{8\})) &= -8 \cdot \mathrm{ld}\left(\frac{18}{37}\right) = 8.316 \text{ bit}.
\end{aligned}$$

Hat man einen diskreten Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), P)$, eine nichtleere Menge Ω_X und eine Abbildung

$$X : \Omega \to \Omega_X \quad \text{(eine sogenannte } \textbf{Zufallsvariable}\text{)}$$

gegeben, so erhält man durch

$$P_X : \mathcal{P}(\{X(\omega);\, \omega \in \Omega\}) \to [0, 1], \quad A \mapsto P(\{\omega \in \Omega;\, X(\omega) \in A\})$$

einen diskreten Wahrscheinlichkeitsraum $(\{X(\omega);\, \omega \in \Omega\}, \mathcal{P}(\{X(\omega);\, \omega \in \Omega\}), P_X)$. Das Wahrscheinlichkeitsmaß P_X wird als **Bildmaß** von P unter X bezeichnet.
Für eine abzählbare Teilmenge $A \subset \mathbb{C}$ und einen entsprechenden Wahrscheinlichkeitsraum $(A, \mathcal{P}(A), P)$ wird die Größe

$$\mu := \sum_{a \in A} a \cdot P(\{a\})$$

als **Erwartungswert** des Zufallsexperiments repräsentiert durch $(A, \mathcal{P}(A), P)$ bezeichnet. Der Erwartungswert muss nicht existieren. Für endliches μ werden

$$\sigma^2 := \sum_{a \in A} (a - \mu)^2 \cdot P(\{a\})$$

als **Varianz** und $\sigma = \sqrt{\sigma^2}$ als **Standardabweichung** des Zufallsexperiments repräsentiert durch $(A, \mathcal{P}(A), P)$ bezeichnet.
Ist $(\Omega, \mathcal{P}(\Omega), P)$ ein diskreter Wahrscheinlichkeitsraum und

$$X : \Omega \to \mathbb{R}$$

eine Zufallsvariable, so werden die Größen (Existenz vorausgesetzt)

$$\mu_X := \sum_{b \in \{X(\omega);\, \omega \in \Omega\}} b \cdot P_X(\{b\}), \quad \sigma_X^2 := \sum_{b \in \{X(\omega);\, \omega \in \Omega\}} (b - \mu_X)^2 \cdot P_X(\{b\}), \quad \sigma_X = \sqrt{\sigma_X^2}$$

als Erwartungswert, Varianz und Standardabweichung von X bezeichnet.
Sei $(\Omega, \mathcal{P}(\Omega), P)$ ein diskreter Wahrscheinlichkeitsraum und

$$I : \Omega \to \mathbb{R} \cup \{\infty\}, \quad \omega \mapsto -\operatorname{ld}(P(\{\omega\}) \quad (\textbf{Informationsmenge}),$$

dann wird

$$\mu_I := \sum_{b \in \{I(\omega);\, \omega \in \Omega\}} b \cdot P_I(\{b\}) \quad \text{mit der Vereinbarung } 0 \cdot \operatorname{ld}(0) = 0$$

als mittlere Informationsmenge oder **Shannon-Entropie** bezeichnet. Die Shannon-Entropie spielt eine zentrale Rolle in der Kommunikationstechnik, der Informatik und der statistischen Physik.

Poisson-Verteilung:
Ist $A = \mathbb{N}_0$ und

$$P(\{j\}) := p_j = e^{-\lambda} \frac{\lambda^j}{j!},\ j \in \mathbb{N}_0,\ \lambda > 0, \tag{10.1}$$

so spricht man von einer Poisson-Verteilung mit Parameter $\lambda (= \mu = \sigma^2)$ (D. Poisson (1781-1840)).

Gleichverteilung und Laplace-Experiment:
Ist $A = \{a_1, \ldots, a_k\}$ und

$$P(\{j\}) := p_j = \frac{1}{k},\ \text{für } j = 1, \ldots, k, \tag{10.2}$$

so wird diese Verteilung Gleichverteilung genannt. Ein Zufallsexperiment, das durch einen Wahrscheinlichkeitsraum mit Gleichverteilung repräsentiert wird, heißt nach P. S. de Laplace (1749-1827) Laplace-Experiment.

Binomial-Verteilung:
Wählt man $p \in \mathbb{R}$, $0 < p < 1$, und $A = \{0, 1, 2, \ldots, s\}$, $s \in \mathbb{N}$, so wird (mit $\binom{s}{j} := \frac{s!}{(s-j)!j!}$) die durch

$$P(\{j\}) := p_j = \binom{s}{j} p^j (1-p)^{s-j} \text{ für } j = 0, \ldots, s \tag{10.3}$$

gegebene Verteilung Binomial-Verteilung $B(s, p)$ mit Parameter s, p genannt. Die Binomialverteilung kann folgendermaßen interpretiert werden: Man betrachtet ein Zufallsexperiment, bei dem es nur zwei mögliche Ergebnisse gibt, nämlich mit Wahrscheinlichkeit p das Ergebnis 'T' (Treffer) und mit Wahrscheinlichkeit $(1 - p)$ das Ergebnis 'N' (Niete) (ein Bernoulli-Experiment). Dieses Experiment führen wir s-mal durch, ohne dass sich die Ergebnisse gegenseitig beeinflussen.

Die Wahrscheinlichkeit, dass nach diesen s Versuchen genau j Treffer auftreten, ist gegeben durch $\binom{s}{j}p^j(1-p)^{s-j}$, $0 \leq j \leq s$, $s \in \mathbb{N}$. Somit wird die s-malige Durchführung unseres Bernoulli-Experimentes durch eine Binomial-Verteilung beschrieben, falls die Ergebnisse sich nicht gegenseitig beeinflussen. Es gilt:

$$\mu = sp, \quad \sigma^2 = sp(1-p).$$

Für sehr große s und sehr kleine p ist es möglich, eine Binomial-Verteilung durch die wesentlich einfacher zu berechnende Poisson-Verteilung mit Parameter $\lambda = sp$ zu approximieren.

10.3 Allgemeine Wahrscheinlichkeitsräume

Betrachtet man einen diskreten Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), P)$, so wurden für das Wahrscheinlichkeitsmaße P die folgenden Eigenschaften gefordert:

(P1) $P : \mathcal{P}(\Omega) \to [0,1]$, wobei $\mathcal{P}(\Omega)$ die Potenzmenge von Ω bezeichnet.

(P2) $P(\emptyset) = 0$, $P(\Omega) = 1$.

(P3) Für jede Folge $\{A_i\}_{i\in\mathbb{N}}$ paarweise disjunkter Mengen mit $A_i \in \mathcal{P}(\Omega)$, $i \in \mathbb{N}$, gilt:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Eine nichtleere Menge Ω heißt **überabzählbar**, falls es keine surjektive Abbildung $\mathbb{N} \to \Omega$ gibt (in Zeichen: $|\Omega| > |\mathbb{N}|$). Es wäre nun naheliegend, für ein Wahrscheinlichkeitsmaß P die Eigenschaften (P1)-(P3) auch dann zu fordern, wenn es überabzählbar viele Ergebnisse in Ω gibt. Leider zeigt sich aber, dass es für überabzählbare Ω keine für die Praxis brauchbaren Abbildungen P dieser Art gibt; die Überabzählbarkeit von Ω schränkt die Möglichkeiten, ein P mit den Eigenschaften (P1)-(P3) finden zu können, extrem ein. Da man einerseits auf Wahrscheinlichkeitsräume mit überabzählbarer Ergebnismenge nicht verzichten kann, andererseits die durch (P1)-(P3) angegebenen Eigenschaften prinzipiell unverzichtbar sind, ist man im Rahmen der Maßheorie dazu übergegangen, die Definitionsmenge von P (im Folgenden mit $\mathcal{D}$ ($\subseteq \mathcal{P}(\Omega)$) bezeichnet) einzuschränken (also nicht mehr die Potenzmenge von Ω zu fordern), um somit die Möglichkeiten für die Wahl von P zu erweitern; ansonsten sollen die Eigenschaften (P1)-(P3) aber für $\mathcal{D}$ anstelle von $\mathcal{P}(\Omega)$ gelten. Daraus folgt natürlich sofort, dass $\mathcal{D}$ eine gewisse Minimalstruktur vorweisen muss:

(i) $\Omega, \emptyset \in \mathcal{D}$ wegen (P2).

(ii) Für jede Folge $\{A_i\}_{i\in\mathbb{N}}$ paarweise disjunkter Mengen mit $A_i \in \mathcal{D}$, $i \in \mathbb{N}$, gilt:

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{D} \quad \text{wegen (P3)}.$$

Da man einerseits darauf angewiesen ist, möglichst viele Teilmengen von Ω in $\mathcal{D}$ wiederzufinden, da man ja nur diesen Mengen eine Wahrscheinlichkeit zuordnen kann, und da man andererseits $\mathcal{D}$ nicht zu umfangreich wählen sollte, da sonst die Existenz praktisch relevanter Wahrscheinlichkeitsmaße gefährdet ist, wünscht man sich für $\mathcal{D}$ neben (i) und (ii) noch ein wichtiges Strukturmerkmal: Wählt man eine (unstrukturierte) Menge $\mathcal{M} \subset \mathcal{P}(\Omega)$ (Teilmengen von Ω, denen man unbedingt eine Wahrscheinlichkeit zuordnen will), so soll es eine **kleinste** Menge $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ geben, die (i) und (ii) erfüllt und für die $\mathcal{M} \subseteq \mathcal{D}$ gilt; mit anderen Worten: Sind $\mathcal{D}_1 \subseteq \mathcal{P}(\Omega)$ und $\mathcal{D}_2 \subseteq \mathcal{P}(\Omega)$ zwei Mengen, die (i) und (ii) erfüllen, so soll auch $\mathcal{D}_1 \cap \mathcal{D}_2$ diese beiden Eigenschaften erfüllen, denn dann gäbe es die kleinste Menge

$$\mathcal{D}(\mathcal{M}) := \bigcap_{\mathcal{D}\in\mathbf{D}} \mathcal{D},$$

die (i) und (ii) erfüllt und die die Menge $\mathcal{M}$ enthält, wobei

$$\mathbf{D} = \{\mathcal{G} \subseteq \mathcal{P}(\Omega);\ \mathcal{M} \subseteq \mathcal{G} \text{ und } \mathcal{G} \text{ erfüllt (i) und (ii)}\}.$$

Diese Forderungen führen auf die Strukturmerkmale einer σ-Algebra über Ω.

Definition 10.3 (σ-Algebra)
Sei Ω eine nichtleere Menge. Eine Menge $\mathcal{S} \subseteq \mathcal{P}(\Omega)$ heißt σ-**Algebra** über Ω, falls die folgenden Axiome erfüllt sind:

(S1) $\Omega \in \mathcal{S}$.

(S2) Aus $A \in \mathcal{S}$ folgt $A^c := \{\omega \in \Omega;\ \omega \notin A\} \in \mathcal{S}$.

(S3) Aus $A_i \in \mathcal{S}$, $i \in \mathbb{N}$, folgt $\bigcup_{i=1}^{\infty} A_i \in \mathcal{S}$.

Der große Vorteil in den Strukturmerkmalen einer σ-Algebra über Ω liegt nun nicht nur in der Verträglichkeit mit den Forderungen an die Abbildung P (also Eigenschaften (i) und (ii)), sondern in der Tatsache, dass der Schnitt zweier σ-Algebren über Ω wieder eine σ-Algebra über Ω ist. Hat man nun eine Wunschliste $\mathcal{M}$ von Teilmengen von Ω, denen man auf alle Fälle eine Wahrscheinlichkeit zuordnen will, so ist mit

$$\sigma(\mathcal{M}) := \bigcap_{\mathcal{F} \in \Sigma} \mathcal{F}$$

die kleinste σ-Algebra über Ω gegeben, die $\mathcal{M}$ enthält, wobei Σ die Menge aller σ-Algebren über Ω darstellt, die $\mathcal{M}$ enthalten.
Zusammenfassend ist ein Wahrscheinlichkeitsraum gegeben durch die Ergebnismenge Ω, eine σ-Algebra $\mathcal{S}$ über Ω und ein Wahrscheinlichkeitsmaß P, also eine Abbildung P definiert auf $\mathcal{S}$, die die Bedingungen

(P1') $P : \mathcal{S} \to [0, 1]$

(P2) $P(\emptyset) = 0$, $P(\Omega) = 1$

(P3') Für jede Folge $\{A_i\}_{i \in \mathbb{N}}$ paarweise disjunkter Mengen mit $A_i \in \mathcal{S}$, $i \in \mathbb{N}$, gilt:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

erfüllt. Für den Fall $\Omega = \mathbb{R}^n$, $n \in \mathbb{N}$, hat sich die Wahl

$$\mathcal{M} = \{A \subseteq \mathbb{R}^n;\ A \text{ offen}\}$$

bewährt. Die σ-Algebra

$$\mathcal{B}^n := \sigma(\mathcal{M})$$

wird Borelsche σ-Algebra über $\mathbb{R}^n$ genannt. Obwohl

$$\mathcal{B}^n \neq \mathcal{P}(\mathbb{R}^n),$$

sind in $\mathcal{B}^n$ alle relevanten Teilmengen des $\mathbb{R}^n$ (auch die abgeschlossenen und kompakten Teilmengen) enthalten. Ferner gibt es für alle praktisch relevanten Fragestellungen geeignete Wahrscheinlichkeitsmaße definiert auf $\mathcal{B}^n$. Ist Ω abzählbar, kann - wie bisher - stets $\mathcal{S} = \mathcal{P}(\Omega)$ gewählt werden (offensichtlich ist $\mathcal{P}(\Omega)$ immer eine σ-Algebra über Ω). Ein Tupel $(\Omega, \mathcal{S})$ bestehend aus einer nichtleeren Ergebnismenge Ω und einer σ-Algebra $\mathcal{S}$ über Ω wird als **Messraum** bezeichnet. Die Elemente der σ-Algebra $\mathcal{S}$ heißen **Ereignisse**.
Im Allgemeinen muss für jede überabzählbare Ergebnismenge Ω eine „passende“ σ-Algebra $\mathcal{S}$

gewählt werden. Im Gegensatz zu diskreten Wahrscheinlichkeitsräumen können wir bei σ-Algebren über überabzählbare Mengen Ω nicht mehr davon ausgehen, dass die Elementarereignisse Elemente der σ-Algebren sind. Ist dies doch der Fall, so kann ein Wahrscheinlichkeitsmaß

$$P : \mathcal{S} \to [0,1]$$

nicht mehr durch die Angabe der Wahrscheinlichkeiten für die Elementarereignisse festgelegt werden, da dies die Summation von überabzählbar vielen Summanden erfordern würde.
In der Wahrscheinlichkeitstheorie betrachtet man, basierend auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{S}, P)$ und einem Messraum $(\Omega', \mathcal{S}')$, **Zufallsvariable**

$$\mathbf{X} : \Omega \to \Omega',$$

also Abbildungen derart, dass gilt:

$$\mathbf{X}^{-1}(A') := \{\omega \in \Omega;\, \mathbf{X}(\omega) \in A'\} \in \mathcal{S} \quad \text{für alle} \quad A' \in \mathcal{S}'.$$

Diese Eigenschaft wird als $\mathcal{S}$-$\mathcal{S}'$-**Messbarkeit** von $\mathbf{X}$ bezeichnet. Eine Zufallsvariable dient dazu, gewisse Teilaspekte eines Zufallsexperiments gegeben durch $(\Omega, \mathcal{S}, P)$ hervorzuheben und unwichtige Teilaspekte auszublenden. Ferner ist dank der $\mathcal{S}$-$\mathcal{S}'$-Messbarkeit von $\mathbf{X}$ durch

$$P_{\mathbf{X}} : \mathcal{S}' \to [0,1], \quad A' \mapsto P(\mathbf{X}^{-1}(A'))$$

ein Wahrscheinlichkeitsmaß $P_{\mathbf{X}}$ auf $\mathcal{S}'$, das sogenannte **Bildmaß**, gegeben.

Beispiel(e) 10.4
Ein Zufallsgenerator erzeugt eine reelle Zahl im Intervall $[0,1]$. Da $\Omega = [0,1]$ Überabzählbar ist, wählen wir die σ-Algebra

$$\mathcal{S} = \{[0,1] \cap A;\ A \in \mathcal{B}\}$$

über $[0,1]$, wobei $\mathcal{B}$ die Borelsche σ-Algebra über $\mathbb{R}$ darstellt. Aus der Maßtheorie ist bekannt, dass ein Wahrscheinlichkeitsmaß P auf $\mathcal{S}$ durch Vorgabe der Wahrscheinlichkeiten

$$P((a,b]) := b - a, \quad 0 \le a < b \le 1$$

festgelegt ist. Aus den Eigenschaften (P1'), (P2) und (P3') folgt:

$$P([a,b]) = b - a, \quad 0 \le a \le b \le 1$$

und damit

$$P(\{x\}) = 0 \quad \text{für alle} \quad x \in [0,1].$$

Nun betrachten wir die Abbildung

$$\mathbf{X} : [0,1] \to \{1,2,\ldots,10\}, \quad x \mapsto \max_{k \in \{1,2,\ldots,10\}} \left\{k;\, \frac{1}{k} \ge x\right\}.$$

Da

$$\mathbf{X}^{-1}(\{k\}) = \begin{cases} \left(\frac{1}{k+1}, \frac{1}{k}\right] & \text{falls } 1 \le k \le 9 \\ \left[0, \frac{1}{10}\right] & \text{falls } k = 10 \end{cases},$$

ist $\mathbf{X}$ eine Zufallsvariable, die den Aspekt „Die Zufallszahl liegt in einem der Intervalle $\left(0, \frac{1}{10}\right], \left(\frac{1}{10}, \frac{1}{9}\right], \ldots, \left(\frac{1}{2}, 1\right]$“ hervorhebt und alles andere ausblendet.

Im Folgenden betrachten wir einige wichtige Begriffe der elementaren Wahrscheinlichkeitstheorie. Ausgangspunkt ist der Wahrscheinlichkeitsraum $(\Omega, \mathcal{S}, P)$ und zwei Mengen $A, B \in \mathcal{S}$

mit $P(B) > 0$. Auf $\mathcal{S}$ definieren wir nun ein Wahrscheinlichkeitsmaß $P^B : \mathcal{S} \to [0,1]$ durch $A \mapsto \frac{P(A \cap B)}{P(B)}$. Durch den Übergang von P zu P^B erhält die Menge B das Wahrscheinlichkeitsmaß 1. Wir interpretieren $P^B(A)$ als die Wahrscheinlichkeit von A unter der Bedingung, dass das Ereignis B eintrifft.

Formel von der totalen Wahrscheinlichkeit
Betrachtet man nun eine Partition $\{D_i \subset \Omega; i \in \mathbb{N}\}$ von Ω (also: $D_i \cap D_j = \emptyset, i \neq j$, und $\bigcup_{i=1}^{\infty} D_i = \Omega$), sodass für alle $i \in \mathbb{N}$ $D_i \in \mathcal{S}$ und $P(D_i) > 0$ gilt, so lässt sich sehr leicht die folgende Formel von der totalen Wahrscheinlichkeit nachweisen:

$$P(A) = \sum_{i=1}^{\infty} P(D_i) \cdot P^{D_i}(A) \quad \text{für alle } A \in \mathcal{S}. \tag{10.4}$$

Satz von Bayes
Gilt zusätzlich $P(A) > 0$, so folgt aus

$$P^A(D_i) = \frac{P(D_i \cap A)}{P(A)} = \frac{P^{D_i}(A) \cdot P(D_i)}{P(A)}, \ i \in \mathbb{N}, \tag{10.5}$$

der 'Satz von Bayes':

$$P^A(D_i) = \frac{P^{D_i}(A) \cdot P(D_i)}{\sum_{j=1}^{\infty} P(D_j) \cdot P^{D_j}(A)} \quad \text{für alle } i \in \mathbb{N}. \tag{10.6}$$

Analoge Formeln ergeben sich natürlich für eine endliche Partition $\{D_i \subset \Omega; i = 1, \ldots, n\}$ von Ω.

Sei $f : \mathbb{R}^n \to \mathbb{R}$, $n \in \mathbb{N}$, eine Funktion mit folgenden Eigenschaften:

- f ist bis auf endlich viele Punkte stetig,
- $f(x) \geq 0$ für alle $x \in \mathbb{R}^n$,
- $\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x)\, dx = 1$,

dann ist durch f ein Wahrscheinlichkeitsmaß P auf der Borelsche σ-Algebra $\mathcal{B}^n$ über $\mathbb{R}^n$ durch

$$P([a_1, b_1]) \times \ldots \times [a_n, b_n] = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(x)\, dx$$

gegeben. Die Funktion f wird als **Dichte** des Wahrscheinlichkeitsmaßes P bezeichnet. Für $n = 1$ bezeichnet

$$\mu := \int_{-\infty}^{\infty} x f(x)\, dx$$

den Erwartungswert (Existenz vorausgesetzt) und

$$\sigma^2 := \int_{-\infty}^{\infty} (x - \mu)^2\, dx$$

die Varianz des Zufallsexperiments $(\mathbb{R}, \mathcal{B}, P)$. Nun betrachten wir für jeden Vektor $\mu \in \mathbb{R}^n$ und für jede positiv definite Matrix $\Sigma \in \mathbb{R}^{n,n}$ die Funktion

$$\nu_{\mu,\Sigma} : \mathbb{R}^n \to \mathbb{R},\ x \mapsto \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \cdot \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1} (x-\mu)}{2}\right). \tag{10.7}$$

Offensichtlich ist $\nu_{\mu,\Sigma}(x) > 0$ für alle $\mu, x \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n,n}$, Σ positiv definit.
Aus der Analysis (Substitutionsregel, Satz von Fubini) stammt das folgende Resultat:

$$\int_{\mathbb{R}^n} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right) dx = \sqrt{(2\pi)^n \det(\Sigma)} \tag{10.8}$$

für alle $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n,n}$, Σ positiv definit. Somit können wir $\nu_{\mu,\Sigma}$ als Dichte eines Wahrscheinlichkeitsmaßes auffassen.

Definition 10.5 (Normalverteilung)
Seien $(\Omega, \mathcal{S}, P)$ ein Wahrscheinlichkeitsraum, $\mu \in \mathbb{R}^n$, $n \in \mathbb{N}$, und $\Sigma \in \mathbb{R}^{n,n}$, Σ positiv definit. Die Zufallsvariable $X_{\mu,\Sigma} : \Omega \to \mathbb{R}^n$ heißt $\mathcal{N}(\mu, \Sigma)$ normalverteilt, falls ihr Bildmaß $P_{X_{\mu,\Sigma}}$ durch die Dichte

$$\nu_{\mu,\Sigma} : \mathbb{R}^n \to \mathbb{R},\ x \mapsto \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \cdot \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right). \tag{10.9}$$

gegeben ist.

Wichtig ist der Spezialfall $n = 1$ mit $\mu \in \mathbb{R}$ und $\sigma^2 > 0$:

$$\nu_{\mu,\sigma^2} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{10.10}$$

Dabei repräsentiert μ den Erwartungswert und σ^2 die Varianz.

10.4 Mathematische Statistik

Hat man ein Zufallsexperiment durch einen Wahrscheinlichkeitsraum modelliert, so stellt die Wahrscheinlichkeitstheorie Hilfsmittel bereit, um bei bekanntem Wahrscheinlichkeitsraum Aussagen über den Ablauf des zugrundeliegenden Zufallsexperimentes machen zu können. Die mathematische Statistik behandelt die folgende Problemstellung: Das zu modellierende Zufallsexperiment wird zunächst durch einen unvollständigen Wahrscheinlichkeitsraum beschrieben. Bei dieser Beschreibung werden die Grundmenge Ω, die $\sigma-$Algebra $\mathcal{S}$ und eine Menge von Wahrscheinlichkeitsmaßen auf $\mathcal{S}$ festgelegt. Dabei wird die Menge der in Frage kommenden Wahrscheinlichkeitsmaße häufig durch einen Parameter θ aus einem Parameterraum Θ dargestellt. Die Menge aller eindimensionalen Normalverteilungen kann zum Beispiel durch den Parameterraum

$$\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ = \Theta \tag{10.11}$$

dargestellt werden. Um nun zu einer vollständigen mathematischen Beschreibung unseres Zufallsexperimentes zu kommen, müssen wir uns für ein Wahrscheinlichkeitsmaß P aus der Menge der in Frage kommenden Wahrscheinlichkeitsmaße entscheiden. Ein wesentliches Kriterium der mathematischen Statistik besteht nun darin, dass eine Entscheidung über die Wahl des Wahrscheinlichkeitsmaßes beziehungsweise über die Verkleinerung der Menge aller in Frage kommenden Wahrscheinlichkeitsmaße von Ergebnissen des Zufallsexperimentes abhängt. Somit kann die Problemstellung der mathematischen Statistik als Umkehrung der Problemstellung der Wahrscheinlichkeitstheorie aufgefasst werden. Dabei werden die Ergebnisse des Zufallsexperimentes häufig nicht unmittelbar für die zu treffende Entscheidung verwendet, sondern die von diesen Ergebnissen abhängigen Werte einer Zufallsvariablen X definiert auf Ω. Wir erhalten somit die folgende Ausgangssituation:
Gegeben ist ein Tripel $(\Omega, \mathcal{S}, \mathcal{W})$ bestehend aus einer Grundmenge Ω, einer $\sigma-$Algebra $\mathcal{S}$ und einer Menge $\mathcal{W}$ von Wahrscheinlichkeitsmaßen auf $\mathcal{S}$. Ist diese Menge $\mathcal{W}$ durch einen Parameter

$\theta \in \Theta$ beschrieben, so schreiben wir $(\Omega, \mathcal{S}, \mathcal{W}_{\theta\in\Theta})$. Ferner sind ein Messraum $(\Psi, \mathcal{G})$, eine Zufallsvariable $X : \Omega \to \Psi$ und der Funktionswert $X(\hat{\omega})$ der Zufallsvariable X für mindestens ein beobachtetes Ergebnis $\hat{\omega} \in \Omega$ des zugrundegelegten Zufallsexperimentes gegeben. Basierend auf $X(\hat{\omega})$ soll nun unter verschiedenen weiteren Vorgaben eine Entscheidung für die Wahl des Wahrscheinlichkeitsmaßes $P \in \mathcal{W}$ auf $\mathcal{S}$ ermöglicht oder zumindest vereinfacht werden. Wir werden die folgenden Problemstellungen untersuchen:

Punktschätzung
Unter der Annahme, dass unsere Menge $\mathcal{W}$ von Wahrscheinlichkeismaßen durch einen Parameter $\theta \in \Theta$ dargestellt ist, soll für eine nichtleere Menge Γ und eine vorgegebene Funktion $\gamma : \Theta \to \Gamma$ ein Funktionswert $\hat{\gamma}$ von γ ermittelt werden. Durch die Wahl von $\hat{\gamma}$ wird $\mathcal{W}_{\theta\in\Theta}$ auf $\mathcal{W}_{\theta\in\{\tau\in\Theta;\gamma(\tau)=\hat{\gamma}\}}$ reduziert. Für die Menge aller eindimensionalen Normalverteilungen mit

$$\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ = \Theta \tag{10.12}$$

könnte die Funktion γ zum Beispiel in der Projektion auf die erste Komponente bestehen:

$$\gamma : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}(= \Gamma), \quad (\mu, \sigma^2) \mapsto \mu. \tag{10.13}$$

Man interessiert sich also nur für die Festlegung des Erwartungswertes.

Bereichsschätzung
Im Gegensatz zur Punktschätzung begnügt man sich bei der Bereichsschätzung damit, die Menge $\mathcal{W}_{\theta\in\Theta}$ der möglichen Wahrscheinlichkeitsmaße durch die Wahl einer speziellen Teilmenge $B \subset \Gamma$ auf die Menge $\mathcal{W}_{\theta\in\{\tau\in\Theta;\gamma(\tau)\in B\}}$ zu reduzieren.

Testtheorie
In der Testtheorie betrachtet man eine Partition $\mathcal{W}_0, \mathcal{W}_1$ unserer Menge von Wahrscheinlichkeitsmaßen $\mathcal{W}$. Das Ergebnis eines Tests ist eine Entscheidung darüber, ob die Menge der zur Diskussion stehenden Wahrscheinlichkeitsmaße $\mathcal{W}$ auf $\mathcal{W}_0$ oder $\mathcal{W}_1$ reduziert wird.

Durch die Zufallsvariable $X : \Omega \to \Psi$ erhalten wir zu jedem $P \in \mathcal{W}$ ein Wahrscheinlichkeitsmaß P_X auf $\mathcal{G}$, das Bildmaß von X. Die zu $\mathcal{W}$ gehörige Menge aller Bildmaße P_X auf $\mathcal{G}$ bezeichnen wir mit $P_{X,\mathcal{W}}$ beziehungsweise mit $P_{X,\mathcal{W}_{\theta\in\Theta}}$, falls $\mathcal{W}$ durch einen Parameter $\theta \in \Theta$ dargestellt wird. Da wir nicht die Beobachtung $\hat{\omega} \in \Omega$ als Basis unserer Überlegungen gewählt haben, sondern $X(\hat{\omega})$, ist es sinnvoll, Aussagen über Wahrscheinlichkeitsmaße $P \in \mathcal{W}$ (auf $\mathcal{S}$) auf Aussagen über Wahrscheinlichkeitsmaße $P_X \in P_{X,\mathcal{W}}$ zu verlagern.

Definition 10.6 (Stichprobe(nraum), stat. Raum, Realisierung)
Seien $(\Omega, \mathcal{S})$ ein Messraum, $\mathcal{W}$ eine Menge von Wahrscheinlichkeitsmaßen auf $\mathcal{S}$ und $(\Omega, \mathcal{S}, \mathcal{W})$ die unvollständige Beschreibung eines Zufallsexperimentes gemäß der obigen Motivation. Seien ferner $\hat{\omega} \in \Omega$ ein Ergebnis dieses Zufallsexperimentes, $(\Psi, \mathcal{G})$ ein Messraum und $X : \Omega \to \Psi$ eine Zufallsvariable (also $\mathcal{S}$-$\mathcal{G}$-messbar), dann heißt die Menge Ψ Stichprobenraum, und der Wert $\bar{x} = X(\hat{\omega})$ Stichprobe oder Realisierung von X. Ist nun $P_{X,\mathcal{W}}$ die Menge aller Bildmaße von X in Abhängigkeit von $\mathcal{W}$, so heißt das Tupel $(\Psi, \mathcal{G}, P_{X,\mathcal{W}})$ statistischer Raum.

Im folgenden Beispiel sollen die bisherigen Begriffe eingeübt und dabei auch auf einen zentralen Begriff der mathematischen Statistik hingeführt werden.

Herstellung von Glühbirnen
Eine Firma stellt Glühbirnen her, wobei jede Glühbirne mit einer festen Wahrscheinlichkeit $\theta \in (0, 1)$ defekt ist. In einem großen Lager werden M Glühbirnen aufbewahrt. Ein Kunde bestellt

K ($K < M$) Glühbirnen, die aus dem Lager entnommen und ausgeliefert werden. Bei der Verwendung jeder dieser K Glühbirnen wird vom Kunden notiert, ob sie defekt ($\hat{=}\,1$) ist oder nicht ($\hat{=}\,0$). Die Firma hätte gerne aufgrund der Erfahrungen des Kunden eine Schätzung, mit welcher Wahrscheinlichkeit eine Glühbirne defekt ist. Das im Lager befindliche M-Tupel von Glühbirnen modellieren wir durch einen binären Vektor $\omega \in \Omega = \{0,1\}^M$, wobei wir uns die M Glühbirnen als durchnummeriert vorstellen und $\omega_i = 1$ bedeutet, dass die i-te Glühbirne defekt ist (also ist für $\omega_i = 0$ die i-te Glühbirne in Ordnung). Als σ-Algebra $\mathcal{S}$ auf $\{0,1\}^M$ können wir die Potenzmenge $\mathcal{P}(\{0,1\}^M)$ verwenden. Unter der Annahme, dass die Zustände der einzelnen Glühbirnen stochastisch unabhängig sind, erhalten wir als mögliche Wahrscheinlichkeitsmaße $P_\theta \in \mathcal{W}_{\theta\in(0,1)}$:

$$P_\theta(\omega) = \prod_{i=1}^{M} \theta^{\omega_i}(1-\theta)^{(1-\omega_i)}, \quad \theta \in (0,1). \tag{10.14}$$

Nun steht uns allerdings kein Ergebnis dieses Zufallsexperimentes unmittelbar zur Verfügung, da wir ja nur die K ausgelieferten Glühbirnen beobachten können. Seien nun $1 \le i_1 < \ldots < i_K \le M$ die Nummern derjenigen K Glühbirnen, die ausgeliefert wurden, so können wir mit $\Psi = \{0,1\}^K$ und $\mathcal{G} = \mathcal{P}(\{0,1\}^K)$ die Zufallsvariable

$$X : \Omega \to \Psi, \quad \omega \mapsto (\omega_{i_1}, \ldots, \omega_{i_K}) =: x = (x_1, \ldots, x_K) \tag{10.15}$$

angeben. Für die entsprechenden Bildmaße erhalten wir:

$$P_{X,\mathcal{W}_\theta}(x) = \prod_{i=1}^{K} \theta^{x_i}(1-\theta)^{(1-x_i)} \tag{10.16}$$

Eine Realisierung unserer Zufallsvariablen X ist somit durch einen binären Vektor $\bar{x} \in \Psi$ gegeben. Dieser Vektor ist die einzige Information, die wir zur Schätzung von θ zur Verfügung haben. Da aber

$$P_{X,\mathcal{W}_\theta}(x) = \prod_{i=1}^{K} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^{\sum_{i=1}^K x_i}(1-\theta)^{K-\sum_{i=1}^K x_i}, \tag{10.17}$$

scheint es zu genügen, sich statt $x \in \{0,1\}^K$ nur die Zahl $\sum\limits_{i=1}^{K} x_i$ zu merken. Die entscheidende Frage lautet nun: Sind alle Informationen, die in $x \in \{0,1\}^K$ über den unbekannten Parameter θ enthalten sind, auch in der Zahl $\sum\limits_{i=1}^{K} x_i$ enthalten? In diesem Fall nennt man die Abbildung

$$T : \Psi = \{0,1\}^K \to \Omega_T = \{0,1,\ldots,K\}, \quad x \mapsto \sum_{i=1}^{K} x_i \tag{10.18}$$

eine suffiziente Statistik für θ.
Rein intuitiv würde man einen Schätzwert $\hat{\theta}$ für θ folgendermaßen berechnen:

$$\hat{\theta} = \frac{\sum\limits_{i=1}^{K} \bar{x}_i}{K}. \tag{10.19}$$

Wie könnte man nun entscheiden, ob alle Informationen über θ bereits in der Abbildung T enthalten sind? Zur Beantwortung dieser Frage stellen wir zunächst fest, dass T $\mathcal{P}(\{0,1\}^K)$-$\mathcal{P}(\{0,1,\ldots,K\})$-messbar ist. Somit können wir für jede Menge

$$D_t := \{x \in \{0,1\}^K; T(x) = t\}, \quad t \in \{0,1,\ldots,K\}, \tag{10.20}$$

die bedingten Wahrscheinlichkeiten

$$P^{D_t}_{X,\mathcal{W}_\theta}(x) = \frac{P_{X,\mathcal{W}_\theta}(x \cap D_t)}{P_{X,\mathcal{W}_\theta}(D_t)} \tag{10.21}$$

für $\theta \in (0,1)$ berechnen. Es gilt:

$$P^{D_t}_{X,\mathcal{W}_\theta}(x) = \begin{cases} 0 & \text{für alle } x \text{ mit } \sum_{i=1}^{K} x_i \neq t \\ \frac{\theta^t(1-\theta)^{(K-t)}}{\binom{K}{t}\theta^t(1-\theta)^{(K-t)}} = \frac{1}{\binom{K}{t}} & \text{für alle } x \text{ mit } \sum_{i=1}^{K} x_i = t \end{cases} . \tag{10.22}$$

Entscheidend ist nun die Tatsache, dass die verschiedenen Werte $P^{D_t}_{X,\mathcal{W}_\theta}(x)$ für alle $x \in \{0,1\}^K$ und alle $t \in \{0,1,\ldots,K\}$ nicht mehr von θ abhängen. Diesen Sachverhalt interpretieren wir dahingehend, dass die Beobachtung $T(x)$ hinreichend (suffizient) dafür ist, jede Information über θ zu erhalten, die man aus der Stichprobe $\bar{x} = X(\hat{\omega})$ entnehmen kann.

Eines der wichtigsten Verfahren zur Punktschätzung ist das Maximum-Likelihood-Verfahren. Ausgangspunkt ist ein statistischer Raum $(\mathbb{R}^p, \mathcal{B}^p, P_{X,\mathcal{W}_{\theta\in\Theta}})$, eine Stichprobe $\bar{x}$ und die Forderung, dass $P_{X,\mathcal{W}_\theta}$ für jedes $\theta \in \Theta$ durch eine Dichte $f_{X,\theta} : \mathbb{R}^p \to \mathbb{R}^+_0$ gegeben ist. Man berechnet dann ein $\hat{\theta} \in \Theta$ mit:

$$f_{X,\hat{\theta}}(\bar{x}) \geq f_{X,\theta}(\bar{x}) \text{ für alle } \theta \in \Theta \tag{10.23}$$

und verwendet den Wert $\gamma(\hat{\theta})$ als Maximum-Likelihood-Schätzer für $\gamma(\theta)$. Maximum-Likelihood-Schätzer müssen nicht existieren. Die Berechnung von $\hat{\theta}$ führt auf das Gebiet der mathematischen Optimierung. Die statistische Analyse von Maximum-Likelihood-Schätzern ist ein äusserst schwieriges Problem.

Beispiel(e) 10.7
Normalverteilung
Sei

$$f_{X,\theta}(x) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x_i-\theta_1)^2}{2\theta_2}\right), \tag{10.24}$$

so ergibt sich:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{p} \bar{x}_i}{p} \quad \hat{\theta}_2 = \frac{\sum_{i=1}^{p} (\bar{x}_i - \hat{\theta}_1)^2}{p}. \tag{10.25}$$

Die Berechnung von Bereichsschätzern wollen wir anhand der Normalverteilung demonstrieren. Sei

$$f_{X,\theta}(x) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\theta)^2}{2}\right). \tag{10.26}$$

Gesucht ist ein $a \in \mathbb{R}^+_0$ derart, dass für $0 \leq c \leq 1$ gilt:

$$P_{X,\mathcal{W}_\theta}\left(\left\{x \in \Psi; \theta \in \left[\frac{1}{p}\sum_{i=1}^{p} x_i - a, \frac{1}{p}\sum_{i=1}^{p} x_i + a\right]\right\}\right) \geq c \tag{10.27}$$

für alle $\theta \in \mathbb{R}$. Dies ist äquivalent zu

$$P_{X,\mathcal{W}_\theta}\left(\left\{x \in \Psi; -a\sqrt{p} \leq \frac{1}{\sqrt{p}}\sum_{i=1}^{p} x_i - \sqrt{p}\theta \leq a\sqrt{p}\right\}\right) \geq c. \tag{10.28}$$

Da aber die Zufallsvariable

$$Z_p : \Psi \to \mathbb{R},\ x \mapsto \frac{1}{\sqrt{p}}\sum_{i=1}^{p} x_i - \sqrt{p}\theta \tag{10.29}$$

$\mathcal{N}(0,1)$ normalverteilt ist für alle $p \in \mathbb{N}$, erhalten wir mit

$$\Phi(y) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{x^2}{2}}\, dx :$$

für a:

$$\Phi(a\sqrt{p}) - \Phi(-a\sqrt{p}) \geq c \tag{10.30}$$

beziehungsweise

$$\Phi(a\sqrt{p}) \geq \frac{1+c}{2}. \tag{10.31}$$

Es ergibt sich mit der Realisierung $\bar{x}$ und mit dem aus der obigen Ungleichung bestimmten a der Bereichsschätzer:

$$\left[\frac{1}{p}\sum_{i=1}^{p} \bar{x}_i - a, \frac{1}{p}\sum_{i=1}^{p} \bar{x}_i + a\right] \quad \text{für} \quad \theta. \tag{10.32}$$

Die Testtheorie ist ein wichtiges und umfangreiches Teilgebiet der mathematischen Statistik; daher können wir nur auf Grundideen eingehen. Ausgangspunkt ist ein statistischer Raum $(\Psi, \mathcal{G}, P_{X,\mathcal{W}_{\theta\in\Theta}})$ und eine Partition Θ_0, Θ_1 von Θ (also: $\Theta_0, \Theta_1 \neq \emptyset$, $\Theta_0 \cap \Theta_1 = \emptyset$ und $\Theta_0 \cup \Theta_1 = \Theta$). Basierend auf einer Stichprobe $\bar{x} \in \Psi$ soll nun entschieden werden, ob Θ auf Θ_1 oder Θ_2 reduziert wird. In der Testtheorie wird diese Fragestellung durch die Entscheidung zwischen einer Nullhypothese

$$H_0 : \theta \in \Theta_0 \tag{10.33}$$

und einer Gegenhypothese

$$H_1 : \theta \in \Theta_1 \tag{10.34}$$

formuliert. Zu bestimmen ist also eine $\mathcal{G}$-$\mathcal{P}(\{0,1\})$-messbare Entscheidungsfunktion $\delta : \Psi \to \{0,1\}$, die jeder möglichen Stichprobe $\bar{x}$ eine Entscheidung für Θ_0 (also: $\delta(\bar{x}) = 0$) oder für Θ_1 (also: $\delta(\bar{x}) = 1$) zuordnet. Ein Test ist natürlich dann festgelegt, wenn eine Menge $\bar{K} \subseteq \Theta$ mit

$$\bar{K} = \{x \in \Psi; \delta(x) = 1\}. \tag{10.35}$$

festgelegt ist. Um nun die Menge $\bar{K}$ bestimmen zu können, benötigt man eine Vorstellung, wie man die Güte eines Tests quantifizieren kann. Dazu betrachtet man den Fehler 1. Art

$$P_{X,\mathcal{W}_\theta}(\{x \in \Psi; x \in \bar{K}\}), \quad \theta \in \Theta_0, \tag{10.36}$$

und verbunden damit das Testniveau α:

$$\alpha := \sup_{\theta\in\Theta_0} P_{X,\mathcal{W}_\theta}(\{x \in \Psi; x \in \bar{K}\}). \tag{10.37}$$

Die Zahl $\alpha \in [0,1]$ gibt die kleinste obere Schranke für die Wahrscheinlichkeit an, dass man sich für Θ_1 entscheidet, obwohl $\theta \in \Theta_0$ gilt. Es ist üblich, nur Tests mit dem selben Testniveau zu vergleichen. Ein Test ϕ_1 mit dem Testniveau α ist besser als ein Test ϕ_2 mit dem selben Testniveau, falls

$$P_{X,\mathcal{W}_\theta}(\{x \in \Psi; \phi_1(x) = 1\}) \geq P_{X,\mathcal{W}_\theta}(\{x \in \Psi; \phi_2(x) = 1\}) \text{ für alle } \theta \in \Theta_1. \tag{10.38}$$

Mit anderen Worten: Der Test ϕ_1 ist besser als der Test ϕ_2, falls der Fehler 2. Art für ϕ_1 für jedes $\theta \in \Theta_1$ kleiner ist als der Fehler 2. Art für ϕ_2:

$$P_{X,\mathcal{W}_\theta}(\{x \in \Psi; \phi_1(x) = 0\}) \leq P_{X,\mathcal{W}_\theta}(\{x \in \Psi; \phi_2(x) = 0\}) \text{ für alle } \theta \in \Theta_1. \tag{10.39}$$

Praktisch besteht nun die Vorgehensweise darin, zu einem gegebenen statistischen Raum $(\Psi, \mathcal{G}, P_{X,\mathcal{W}_{\theta\in\Theta}})$ und zu einem gewählten Testniveau α unter allen Tests vom Testniveau α denjenigen Test zu finden (also die Menge $\bar{K}$), der für alle $\theta \in \Theta_1$ den kleinsten Fehler 2. Art besitzt. Diesen Test nennt man gleichmäßig besten Niveau-α-Test. Natürlich stellen sich in diesem Zusammenhang wichtige Fragen, die allerdings im Rahmen dieser Vorlesung nicht diskutiert werden können:

- Gibt es zu einem gegebenen statistischen Raum und zu einem Testniveau α überhaupt einen gleichmäßig besten Test?
- Wenn es diesen gleichmäßig besten Test gibt, wie kann man dann die entsprechende Menge $\bar{K}$ finden?

Bevor wir für einen Spezialfall diese Fragen beantworten, soll auf einen wichtigen Punkt hingewiesen werden. Bei der Suche nach einem Testverfahren wird die kleinste obere Schranke für den Fehler 1. Art gewählt, während man mit dem entsprechenden Fehler 2. Art leben muss. Hier ist also eine Asymmetrie zwischen den Hypothesen H_0 und H_1 erkennbar. Diese Asymmetrie ist gewollt, da sie häufig den praktischen Gegebenheiten entspricht. Soll zum Beispiel aufgrund von Messungen überprüft werden, ob es gefährliche Wechselwirkungen zwischen zwei Medikamenten gibt (Hypothese H_0: Ja, Hypothese H_1: Nein), so ist der Fehler 1. Art, also die Entscheidung, dass es diese Wechselwirkungen nicht gibt, obwohl sie existieren, viel gefährlicher als der Fehler 2. Art, also die Entscheidung, dass diese Wechselwirkungen vorhanden sind, obwohl sie nicht existieren. Doch nun zum bereits erwähnten Spezialfall:

Satz 10.8 ((gleichmäßig) beste Tests bei zweipunktigem Θ)
Seien $(\mathbb{R}^n, \mathcal{B}^n, P_{X,\mathcal{W}_{\theta\in\{\theta_0,\theta_1\}}})$ ein statistischer Raum. $P_{X,\mathcal{W}_{\theta_0}}$ sowie $P_{X,\mathcal{W}_{\theta_1}}$ seien durch die Dichten $f_{X,\theta_0} : \mathbb{R}^n \to \mathbb{R}_0^+$ beziehungsweise $f_{X,\theta_1} : \mathbb{R}^n \to \mathbb{R}_0^+$ gegeben, dann ist für jedes $k \in \mathbb{R}_0^+$ der Test

$$\delta : \Psi \to \{0,1\}, x \mapsto \begin{cases} 1 \text{ für alle } x \text{ mit } f_{X,\theta_1}(x) > k f_{X,\theta_0}(x) \\ 0 \text{ für alle } x \text{ mit } f_{X,\theta_1}(x) \leq k f_{X,\theta_0}(x) \end{cases} \tag{10.40}$$

unter allen Tests vom Testniveau

$$\alpha = P_{X,\mathcal{W}_{\theta_0}}(\{x \in \Psi; f_{X,\theta_1}(x) > k f_{X,\theta_0}(x)\}) \tag{10.41}$$

der (gleichmäßig) beste.