GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Analysis und Numerische Mathematik für die Informatik

Prof. Dr. Bastian von Harrach-Sammet

Goethe-Universität Frankfurt am Main
Institut für Mathematik

Sommersemester 2025

http://numerical.solutions

Inhaltlich finale Version des SoSe25 (Tippfehler werden aber weiterhin korrigiert).
Letzte Aktualisierung: 15. August 2025

Inhaltsverzeichnis

Kapitel 1

Grenzwerte von Folgen und Reihen

1.1 Motivation: Reelle Zahlen und die Wurzel aus 2

Die Menge der *natürlichen Zahlen* wird gebildet, indem beginnend mit 1 zu jeder enthaltenen Zahl auch die nächstgrößere (d.h. um 1 erhöhte) Zahl hinzugenommen wird

$$\mathbb{N} := \{1,2,3,4,5,\ldots\}.$$

$\mathbb{N}$ enthält unendlich viele Elemente. Beginnt man diese Konstruktion mit 0, so erhält man die *natürlichen Zahlen mit Null*

$$\mathbb{N}_0 := \{0,1,2,3,4,5,\ldots\}.$$

Durch Hinzunahme der negativen Zahlen erhält man die Menge der *ganzen Zahlen*

$$\mathbb{Z} := \{\ldots,-3,-2,-1,0,1,2,3,\ldots\}$$

und durch Hinzunahme von Brüchen erhält man die Menge der *rationalen Zahlen*

$$\mathbb{Q} := \{\frac{p}{q} : \ p,q \in \mathbb{Z},\ q \neq 0\}.$$

Dabei ergeben zwei Brüche $\frac{p_1}{q_1}$, $\frac{p_2}{q_2} \in \mathbb{Q}$ genau dann die gleiche rationale Zahl, wenn sie sich durch Kürzen bzw. Erweitern ineinander überführen lassen. Dies gilt genau dann, wenn $p_1 q_2 = p_2 q_1$.

Man kann zeigen, dass sich jede rationale Zahl entweder als Dezimalzahl mit endlich vielen Kommastellen oder als Dezimalzahl mit unendlich vielen, sich aber periodisch wiederholenden Kommastellen geschrieben werden kann. Durch Hinzunahme von Dezimalzahlen mit unendlich vielen, sich dabei aber nicht periodisch

wiederholenden Dezimalzahlen erhält man die Menge der *reellen Zahlen* $\mathbb{R}$. In der Dezimaldarstellung wird eine periodisch wiederholte 9 mit der dadurch beliebig genau approximierten Zahl identifiziert, also z.B. $0,\overline{9} = 1$ und $20,24\overline{9} = 20,25$.

Um den Übergang von den rationalen zu den reellen Zahlen zu motivieren, betrachten wir ein rechtwinkliges Dreieck bei dem die am rechten Winkel anliegenden Seiten (Katheten) die Länge $a = 1$ und $b = 1$ besitzen. Für die Seitenlänge der dritten Seite (Hypotenuse) gilt dann nach dem Satz des Pythagoras

$$c^2 = a^2 + b^2 = 1^2 + 1^2 = 2.$$

Mit Hilfe der Primfaktorzerlegung kann man jedoch beweisen, dass keine rationale Zahl $c \in \mathbb{Q}$ mit der Eigenschaft $c^2 = 2$ existiert, $\sqrt{2} \notin \mathbb{Q}$.

Man kann $\sqrt{2}$ jedoch mit der sogenannten *Intervallhalbierung* beliebig genau annähern. Wegen $1^2 = 1 < 2$ und $2^2 = 4 > 2$ muss $\sqrt{2}$ zwischen 1 und 2 liegen. Für den Mittelwert gilt[1] $1.5^2 = 2.25$, also gilt $\sqrt{2} \in]1, 1.5[$. Wir fahren so fort und erhalten:

$$\begin{aligned}
1^2 = 1 < 2, \qquad & 2^2 = 4 > 2 && \Longrightarrow \quad \sqrt{2} \in]1,2[\\
\frac{1+2}{2} = 1.5, \qquad & 1.5^2 = 2.25 > 2 && \Longrightarrow \quad \sqrt{2} \in]1,1.5[\\
\frac{1+1.5}{2} = 1.25, \qquad & 1.25^2 = 1.5625 < 2 && \Longrightarrow \quad \sqrt{2} \in]1.25,1.5[\\
u.s.w.
\end{aligned}$$

Die untere und die obere Grenze des enstehenden Intervalls nähern sich offenbar immer mehr aneinander an und schließen dabei immer $\sqrt{2}$ ein. So lassen sich daher beliebig viele Kommastellen von $\sqrt{2}$ berechnen, z.B. erhält man nach 10 Schritten

$$\sqrt{2} \in]1.4140625, 1.4150390625[,$$

womit $\sqrt{2}$ bis auf 0.1%-Genauigkeit bestimmt ist.

Das mathematische Teilgebiet der *Analysis* befasst sich theoretisch mit solchen Grenzwertprozessen und Veränderungen auf beliebig kleinen Stücken. Die *Numerische Mathematik* befasst sich mit der Konstruktion von Algorithmen, die mathematische Probleme praktisch (und insbesondere mit Hilfe des Computers) lösen. Oft ist es dabei wie bei obiger Berechnung von $\sqrt{2}$ so, dass ein numerischer Algorithmus die exakte Lösung nur als Grenzwert unendlich vieler Berechnungsschritte erreichen kann und zur praktischen Verwendung bei ausreichender Genauigkeit gestoppt wird.

[1] Im Deutschen schreibt man die Dezimalstellen üblicherweise durch ein Komma abgetrennt. Wir werden in dieser Vorlesung zur besseren Lesbarkeit (etwa bei Kommazahlen als Intervallgrenzen) aber auch häufig die im Englischen übliche Variante der Abtrennung duch einen Punkt verwenden.

1.2 Folgen und ihre Grenzwerte

Definition 1.1 (Folgen)
Sei M eine Menge. Eine Folge in M ist eine Abbildung der natürlichen Zahlen $\mathbb{N}$ nach M, also

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

Üblich ist die Schreibweise $(a_n)_{n\in\mathbb{N}} \subseteq M$ oder $a_1, a_2, a_3, \ldots \in M$. Eine Folge in $M = \mathbb{R}$ nennen wir auch reelle Folge.

Manchmal ist es nützlich, die Nummerierung der Folgenglieder mit 0 zu starten, dann schreiben wir entsprechend $(a_n)_{n\in\mathbb{N}_0} \subseteq M$ und alle im Folgenden gezeigten Aussagen gelten entsprechend auch für diesen Fall. Wenn die Indexmenge aus dem Zusammenhang bekannt ist schreibt man oft auch kurz $(a_n)_n$ oder (a_n).

Beispiele 1.2 (Einfache reelle Folgen)
(a) Mit $a_n := n$ ergibt sich die Folge der natürlichen Zahlen $1, 2, 3, \ldots$

(b) Mit $a_n := 2n-1$ ergibt sich die Folge der ungeraden Zahlen $1, 3, 5, 7, \ldots$

(c) Mit $a_n := n^2$ so ergibt sich die Folge der Quadratzahlen $1, 4, 9, 16, \ldots$

(d) Mit $a_n := (-1)^n$ ergibt sich die alternierende Folge $1, -1, 1, -1, \ldots$

(e) Mit $a_n := \frac{1}{n}$ ergibt sich die harmonische Folge $1, \frac{1}{2}, \frac{1}{3}, \ldots$

(f) Mit $a_n := a$ (für ein $a \in \mathbb{R}$) ergibt sich die konstante Folge $a, a, a, \ldots$

(g) Mit $a_n := a^n$ (für ein $a \in \mathbb{R}$) ergibt sich die Folge der Potenzen $a, a^2, a^3, \ldots$

Die natürlichen Zahlen $\mathbb{N}$ sind *rekursiv* dadurch definiert, dass $1 \in \mathbb{N}$ liegt und zu jeder Zahl $n \in \mathbb{N}$ auch der Nachfolger $n+1$ in $\mathbb{N}$ liegt. Entsprechend können wir eine Folge auch *rekursiv* definieren, indem wir das erste Folgenglied a_1 angeben zusammen mit einer Vorschrift, wie jeweils aus a_n das nächste Folgenglied a_{n+1} berechnet werden kann (*einstufige Rekursion*). Bei *mehrstufigen Rekursion* werden zur Berechnung von a_{n+1} mehrere Vorgänger (z.B. a_n und a_{n-1}) benötigt und entsprechend auch mehrere erste Folgenglieder (z.B. a_1 und a_2).

Beispiele 1.3 (Rekursiv definierte Folgen)
(a) Mit $a_1 := 1$ und $a_{n+1} := a_n + 2$ (für alle $n \in \mathbb{N}$) ergibt sich wieder die Folge der ungeraden Zahlen $1, 3, 5, 7, \ldots$.

(b) Mit $a_1 := 1$ und $a_n := n a_{n-1}$ ergibt sich die Folge der Fakultäten $a_n = n!$

(c) Mit $a_1 := 1$, $a_2 := 1$ und $a_{n+1} := a_n + a_{n-1}$ ergibt sich die Fibonacci-Folge

$$1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21, \ldots$$

(d) Seien $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ die Folge der unteren und die Folge der oberen Intervallgrenzen aus der in Abschnitt 1.1 eingeführten Berechnung von $\sqrt{2}$ (nach jeweils n Intervallhalbierungen). Diese lassen sich als rekursive Folge in $\mathbb{R}^2$ schreiben mit

$$(a_0, b_0) := (1,2), \quad (a_{n+1}, b_{n+1}) := \begin{cases} (a_n, \frac{a_n+b_n}{2}) & \text{falls } \left(\frac{a_n+b_n}{2}\right)^2 > 2, \\ (\frac{a_n+b_n}{2}, b_n) & \text{falls } \left(\frac{a_n+b_n}{2}\right)^2 < 2. \end{cases}$$

Wir geben nun eine mathematisch rigorose Definition dafür, dass sich eine Folge immer mehr einem Grenzwert annähert.

Definition 1.4
Sei $(a_n)_{n\in\mathbb{N}}$ eine reelle Folge. $a \in \mathbb{R}$ heißt Grenzwert (auch: Limes) der Folge, falls

$$\forall \varepsilon > 0 : \ \exists N \in \mathbb{N} : \ |a_n - a| < \varepsilon \quad \forall n \geq N.$$

Wir bezeichnen das auch mit $(a_n)_{n\in\mathbb{N}}$ konvergiert gegen a und schreiben

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

Eine gegen $a = 0$ konvergente Folge heißt auch Nullfolge. Konvergiert eine Folge nicht, so nennen wir sie divergent.

Bemerkung 1.5
Wir sagen, dass eine Eigenschaft für fast alle Folgenglieder gilt, wenn sie für alle bis auf endlich viele gilt. Das Interval $]a-\varepsilon, a+\varepsilon[$ nennen wir auch ε-Umgebung von a. Mit diesen Begriffen ist die Konvergenz einer Folge gegen einen Grenzwert gleichbedeutend zu folgender Aussage:

In jeder ε-Umgebung des Grenzwertes liegen fast alle Folgenglieder,

vgl. das in der Vorlesung gemalte Bild.

Wir geben einige Beispiele für die Anwendung dieser Definition auf einfache Folgen.

Beispiel 1.6

(a) Sei $a \in \mathbb{R}$. Die konstante Folge $a_n = a$ konvergiert gegen a.

Beweis: *Sei $\varepsilon > 0$. Wähle $N := 1$, dann gilt für alle $n \geq N$*

$$|a_n - a| = 0 < \varepsilon.$$

Es ist also $\lim_{n\to\infty} a_n = a$. □

(b) Die Folge $a_n := \frac{1}{n}$ ist eine Nullfolge.

Beweis: *Sei $\varepsilon > 0$. Wähle $N \in \mathbb{N}$, so dass $N \geq \frac{1}{\varepsilon}$, z.B. $N := \text{ceil}(\frac{1}{\varepsilon})$, dann gilt für alle $n \geq N$*

$$|a_n - 0| = \left|\frac{1}{n}\right| = \frac{1}{n} \leq \frac{1}{N} \leq \frac{1}{\frac{1}{\varepsilon}} = \varepsilon.$$

Es ist also $\lim_{n\to\infty} a_n = 0$. □

(c) Die Folge $a_n := n$ besitzt keinen Grenzwert $a \in \mathbb{R}$.

Beweis: *Angenommen es wäre $a = \lim_{n\to\infty} n \in \mathbb{R}$. Dann gäbe es für $\varepsilon := 1$ ein $N \in \mathbb{N}$, so dass für alle $n \in \mathbb{N}$ gilt $n - a \leq |n - a| = |a_n - a| < 1$. Dann wäre aber $n \leq a + 1$ für alle $n \in \mathbb{N}$ und das widerspricht der Unbeschränktheit der natürlichen Zahlen.* □

Motiviert durch die ins Unendliche strebende Folge $a_n := n$ in Beispiel 1.6(c) definieren wir wie folgt auch ∞ und $-\infty$ als mögliche Grenzwerte.

Definition 1.7

Eine reelle Folge $(a_n)_{n\in\mathbb{N}}$ heißt

(a) bestimmt divergent gegen ∞, falls für alle $C > 0$ ein $N \in \mathbb{N}$ existiert, so dass

$$a_n \geq C \quad \text{für alle } n \geq N.$$

Wir schreiben in diesem Fall $\lim_{n\to\infty} a_n = \infty$.

(b) bestimmt divergent gegen $-\infty$, falls für alle $C > 0$ ein $N \in \mathbb{N}$ existiert, so dass

$$a_n \leq -C \quad \text{für alle } n \geq N.$$

Wir schreiben in diesem Fall $\lim_{n\to\infty} a_n = -\infty$.

Beispiel 1.8

Die durch $a_n := n$ definierte Folge der natürlichen Zahlen divergiert bestimmt gegen ∞, es gilt $\lim_{n\to\infty} n = \infty$.

Die durch $a_n := (-1)^n$ definierte Folge divergiert, aber sie divergiert nicht bestimmt. $\lim_{n\to\infty}(-1)^n$ existiert nicht.

1.3 Rechenregeln für Grenzwerte

Für die Untersuchung des Grenzwertverhaltens komplizierterer Folgen ist es oft sinnvoll, diese in einfachere Folgen (mit bekanntem Grenzwertverhalten) zu zerlegen. Wir geben dazu erste Rechenregeln an. Weitere werden sich aus der Eigenschaft der Stetigkeit in Abschnitt 2.2 ergeben. Für diese Rechenregeln sind die folgenden Eigenschaften der Betragsfunktion und der Begriff der Beschränktheit nützlich.

Lemma 1.9
(a) Für alle $a \in \mathbb{R}$ gilt $a \leq |a|$ und $-a \leq |a|$.

(b) Für alle $a, b \in \mathbb{R}$ gilt

$$|a+b| \leq |a| + |b|.$$

Diese Ungleichung heißt auch Dreiecksungleichung.

Beweis: (a) Wir betrachten die beiden möglichen Fälle $a \geq 0$ und $a < 0$:

(i) Für $a \geq 0$ ist $|a| = a$ und $-a \leq 0 \leq |a|$. (a) ist also erfüllt.

(ii) Für $a \leq 0$ ist $|a| = -a \geq 0 \geq a$. Auch in diesem Fall ist (a) erfüllt.

(b) Wir betrachten die beiden möglichen Fälle $a+b \geq 0$ und $a+b < 0$ und verwenden (a):

(i) Für $a+b \geq 0$ ist $|a+b| = a+b \leq |a| + |b|$ und damit (b) erfüllt.

(ii) Für $a+b < 0$ ist $|a+b| = -(a+b) = -a+(-b) \leq |a| + |b|$ und damit (b) erfüllt. □

Definition 1.10
Eine reelle Folge $(a_n)_{n\in\mathbb{N}}$ heißt

(a) nach oben beschränkt, falls ein $C > 0$ existiert mit

$$a_n \leq C \quad \text{für alle } n \in \mathbb{N}.$$

(b) nach unten beschränkt, falls ein $C > 0$ existiert mit

$$a_n \geq -C \quad \text{für alle } n \in \mathbb{N}.$$

(c) beschränkt, falls ein $C > 0$ existiert mit

$$|a_n| \leq C \quad \text{für alle } n \in \mathbb{N}.$$

Dies gilt offenbar genau dann, wenn die Menge der Folgenglieder $\{a_1, a_2, \ldots\}$ eine nach oben beschränkte, nach unten beschränkte, bzw. beschränkte Menge ist.

Satz 1.11
Jede konvergente reelle Folge $(a_n)_{n\in\mathbb{N}}$ ist auch beschränkt. Eine Schranke der Folge ist auch Schranke des Grenzwerts, d.h.

$$c \le a_n \le C \quad \forall n \in \mathbb{N} \quad \Longrightarrow \quad c \le \lim_{n\to\infty} a_n \le C.$$

Beweis: Konvergiert $(a_n)_{n\in\mathbb{N}}$ gegen $a \in \mathbb{R}$, so liegen nach Bemerkung 1.5 für jede Wahl von $\varepsilon > 0$ alle bis auf endlich viele Folgenglieder in der ε-Umgebung von a. Insbesondere gilt das für $\varepsilon := 1$ und es sind also alle bis auf endlich viele Folgenglieder kleiner als $a+1$. Unter den verbleibenden endlich vielen Folgengliedern ist eines das größte, nennen wir dieses $a_{\max}$. Damit sind alle Folgenglieder kleiner oder gleich $\max\{a_{\max}, a+1\}$. Genauso zeigt man die Beschränktheit nach unten.

Gilt $\lim_{n\to\infty} a_n > C$, dann verwenden wir Bemerkung 1.5 mit der Wahl

$$\varepsilon := \frac{1}{2}(\lim_{n\to\infty} a_n - C)$$

und erhalten, dass fast alle, und damit insbesondere ein Folgenglied a_n in der ε-Umgebung von $\lim_{n\to\infty} a_n$ liegt. Es gilt also

$$\lim_{n\to\infty} a_n > C \quad \Longrightarrow \quad \exists n \in \mathbb{N} : \; a_n \ge \lim_{n\to\infty} a_n - \varepsilon = \frac{1}{2}(\lim_{n\to\infty} a_n + C) > C.$$

Durch Kontraposition folgt damit

$$a_n \le C \quad \forall n \in \mathbb{N} \quad \Longrightarrow \quad \lim_{n\to\infty} a_n \le C.$$

Die entsprechende Aussage über untere Schranken zeigt man genauso. □

Bemerkung 1.12
Die Umkehrung der ersten Aussage von Satz 1.11 gilt nicht. Beispielsweise ist $a_n := (-1)^n$ eine beschränkte aber nicht konvergente Folge.

Die zweite Aussage in Satz 1.11 gilt nicht mit $<$ anstelle von $\le$. Beispielsweise ist

$$\frac{1}{n} > 0 \quad \textit{aber} \quad \lim_{n\to\infty} \frac{1}{n} = 0.$$

Satz 1.13 (Rechenregeln für Grenzwerte)
(a) Sind $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ konvergente reelle Folgen, dann ist auch die durch $c_n := a_n + b_n$ definierte Folge konvergent und ihr Grenzwert ist $a+b$. Kurz:

$$\lim_{n\to\infty} a_n = a \quad \textit{und} \quad \lim_{n\to\infty} b_n = b \quad \Longrightarrow \quad \lim_{n\to\infty} a_n + b_n = a + b.$$

Außerdem gilt

$$\begin{aligned}
&\lim_{n\to\infty} a_n = \infty && \text{und} && (b_n)_{n\in\mathbb{N}} \text{ beschränkt} && \Longrightarrow && \lim_{n\to\infty} a_n + b_n = \infty,\\
&\lim_{n\to\infty} a_n = -\infty && \text{und} && (b_n)_{n\in\mathbb{N}} \text{ beschränkt} && \Longrightarrow && \lim_{n\to\infty} a_n + b_n = -\infty,\\
&\lim_{n\to\infty} a_n = \infty && \text{und} && \lim_{n\to\infty} b_n = \infty && \Longrightarrow && \lim_{n\to\infty} a_n + b_n = \infty,\\
&\lim_{n\to\infty} a_n = -\infty && \text{und} && \lim_{n\to\infty} b_n = -\infty && \Longrightarrow && \lim_{n\to\infty} a_n + b_n = -\infty.
\end{aligned}$$

(b) Es gilt

$$\begin{aligned}
&\lim_{n\to\infty} a_n = a && \text{und} && \lim_{n\to\infty} b_n = b && \Longrightarrow && \lim_{n\to\infty} a_n b_n = ab,\\
&\lim_{n\to\infty} a_n = 0 && \text{und} && (b_n)_{n\in\mathbb{N}} \text{ beschränkt} && \Longrightarrow && \lim_{n\to\infty} a_n b_n = 0,\\
&\lim_{n\to\infty} a_n = \infty && \text{und} && \lim_{n\to\infty} b_n = b > 0 && \Longrightarrow && \lim_{n\to\infty} a_n b_n = \infty,\\
&\lim_{n\to\infty} a_n = \infty && \text{und} && \lim_{n\to\infty} b_n = b < 0 && \Longrightarrow && \lim_{n\to\infty} a_n b_n = -\infty,\\
&\lim_{n\to\infty} a_n = -\infty && \text{und} && \lim_{n\to\infty} b_n = b > 0 && \Longrightarrow && \lim_{n\to\infty} a_n b_n = -\infty,\\
&\lim_{n\to\infty} a_n = -\infty && \text{und} && \lim_{n\to\infty} b_n = b < 0 && \Longrightarrow && \lim_{n\to\infty} a_n b_n = \infty.
\end{aligned}$$

Die letzten vier Aussagen gelten auch bereits, wenn man $\lim_{n\to\infty} b_n = b > 0$ *ersetzt durch*

$$\exists c > 0 : \quad b_n \geq c \quad \text{für fast alle } n \in \mathbb{N}$$

bzw. $\lim_{n\to\infty} b_n = b < 0$ *ersetzt durch*

$$\exists c > 0 : \quad b_n \leq -c \quad \text{für fast alle } n \in \mathbb{N},$$

also insbesondere auch für $\lim_{n\to\infty} b_n = \infty$ *bzw.* $\lim_{n\to\infty} b_n = -\infty$.

(c) Für eine Folge $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}\setminus\{0\}$ *gilt*

$$\lim_{n\to\infty} a_n = a \neq 0 \quad \Longrightarrow \quad \lim_{n\to\infty} \frac{1}{a_n} = \frac{1}{a},$$

und

$$\lim_{n\to\infty} |a_n| = \infty \quad \Longrightarrow \quad \lim_{n\to\infty} \frac{1}{a_n} = 0.$$

Beweis: Wir beweisen nur exemplarisch die erste Aussage in (a). Dazu sei $\varepsilon > 0$ und wir müssen zeigen, dass ein $N \in \mathbb{N}$ existiert mit

$$|a_n + b_n - (a+b)| < \varepsilon \quad \text{für alle } n \geq N.$$

Wegen $\lim_{n\to\infty} a_n = a$ existiert zu jedem $\varepsilon_a > 0$ ein $N_a \in \mathbb{N}$ so dass

$$|a_n - a| < \varepsilon_a \quad \text{für alle } n \geq N_a.$$

Ebenso existiert wegen $\lim_{n\to\infty} b_n = b$ zu jedem $\varepsilon_b > 0$ ein $N_b \in \mathbb{N}$ so dass

$$|b_n - b| < \varepsilon_b \quad \text{für alle } n \geq N_b.$$

Wir verwenden dies mit der Wahl $\varepsilon_a := \varepsilon/2$ und $\varepsilon_b := \varepsilon/2$. Dann existieren also $N_a, N_b \in \mathbb{N}$ mit

$$|a_n + b_n - (a+b)| \leq |a_n - a| + |b_n - b| < \varepsilon_a + \varepsilon_b = \varepsilon.$$

für alle $n \geq \max\{N_a, N_b\} =: N$.

Die anderen Aussagen in (a), sowie (b) und (c) lassen sich mit ähnlichen Argumenten zeigen. □

Bemerkung 1.14

Satz 1.13 enthält als Spezialfall auch die konstante Folge. Damit folgt beispielsweise aus $\lim_{n\to\infty} a_n = a$ *und* $\lim_{n\to\infty} b_n = b$ *auch*

$$\lim_{n\to\infty} a_n - b_n = \lim_{n\to\infty} a_n + (-1) * b_n = \lim_{n\to\infty} a_n + \lim_{n\to\infty} (-1) * \lim_{n\to\infty} b_n = a - b$$

und aus $\lim_{n\to\infty} a_n = \infty$, $\lim_{n\to\infty} b_n > 0$ *(mit* $b_n \neq 0\ \forall n \in \mathbb{N}$*) auch*

$$\lim_{n\to\infty} \frac{1}{b_n} > 0 \quad \textit{und daher} \quad \lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} a_n * \frac{1}{b_n} = \infty.$$

Beispiel 1.15

Wir betrachten den Quotienten aus zwei bestimmt divergenten Folgen

$$a_n = \frac{3n^6 + 4n^4 + n}{-n^5 + 17n^4 + 3n^3 + 2n^2 + 5}.$$

Durch Kürzen von n^5 *ergibt sich*

$$a_n = \frac{3n + 4\frac{1}{n} + \frac{1}{n^4}}{-1 + 17\frac{1}{n} + 3\frac{1}{n^2} + 2\frac{1}{n^3} + 5\frac{1}{n^5}}$$

Für den gekürzten Zähler und Nenner erhalten wir

$$3n + 4\frac{1}{n} + \frac{1}{n^4} \to \infty \quad \textit{und} \quad -1 + 17\frac{1}{n} + 3\frac{1}{n^2} + 2\frac{1}{n^3} + 5\frac{1}{n^5} \to -1,$$

insgesamt also $a_n \to -\infty$.

Genauso folgt im allgemeinen Fall des Quotienten zweier Polynome in n mit Höchstgrad j im Zähler und k im Nenner

$$a_n = \frac{\alpha_j n^j + \alpha_{j-1} n^{j-1} + \ldots + \alpha_1 n^1 + \alpha_0}{\beta_k n^k + \beta_{k-1} n^{k-1} + \ldots + \beta_1 n^1 + \beta_0}, \quad \textit{mit } \alpha_j \neq 0,\ \beta_k \neq 0$$

(a) Für $j > k$:

$$\lim_{n\to\infty} a_n = \infty, \quad \textit{falls } \frac{\alpha_j}{\beta_k} > 0, \quad \textit{und} \quad \lim_{n\to\infty} a_n = -\infty, \quad \textit{falls } \frac{\alpha_j}{\beta_k} < 0.$$

(b) Für $j = k$:

$$\lim_{n\to\infty} a_n = \frac{\alpha_j}{\beta_k}.$$

(c) Für $j < k$:

$$\lim_{n\to\infty} a_n = 0.$$

1.4 Vollständigkeit der reellen Zahlen

Wir geben noch zwei Kriterien an, mit denen die Konvergenz einer Folge sichergestellt werden kann, auch ohne ihren Grenzwert konkret zu kennen.

Definition 1.16
Eine reelle Folge $(a_n)_{n\in\mathbb{N}}$ heißt Cauchy-Folge, falls

$$\forall \varepsilon > 0: \quad \exists N \in \mathbb{N}: \quad |a_n - a_m| < \varepsilon \quad \forall n, m \geq N.$$

Für eine Cauchy-Folge existiert also für jedes noch so kleine $\varepsilon > 0$ jeweils eine Umgebung der Breite ε, in der dann fast alle Folgenglieder liegen. In der Definition kann man äquivalenterweise auch $m \geq n \geq N$ schreiben, da wegen $|a_n - a_m| = |a_m - a_n|$ sich der andere Fall $n \geq m \geq N$ einfach durch Vertauschen von m und n ergibt.

Beispiele 1.17
(a) Die konstante Folge $a_n := a$ ist offensichtlich eine Cauchy-Folge.

(b) Für die durch $a_n := \frac{1}{n}$ definierte Folge gilt für alle $m \geq n \geq N$

$$|a_n - a_m| = \left|\frac{1}{n} - \frac{1}{m}\right| = \frac{1}{n} - \frac{1}{m} \leq \frac{1}{n} \leq \frac{1}{N}.$$

Da sich für jedes $\varepsilon > 0$ ein hinreichend großes $N \in \mathbb{N}$ finden lässt mit $\frac{1}{N} < \varepsilon$, ist $(a_n)_{n\in\mathbb{N}}$ also eine Cauchy-Folge.

(c) Die Folge $s_n := 1 + \frac{1}{2} + \ldots + \frac{1}{n}$ erfüllt

$$|s_{n+1} - s_n| = \frac{1}{n+1}.$$

Der Abstand zweier aufeinanderfolgenden Folgenglieder wird also beliebig klein. Trotzdem ist $(s_n)_{n\in\mathbb{N}}$ keine Cauchy-Folge (die geforderte Eigenschaft ist zwar für $m = n+1$, aber nicht für alle $m \geq n$ erfüllt). Es gilt nämlich für beliebig große $n \in \mathbb{N}$ mit $m := 2n$

$$|s_m - s_n| = \frac{1}{n+1} + \frac{1}{n+2} + \ldots + \frac{1}{2n} \geq n\frac{1}{2n} = \frac{1}{2}.$$

Es ist leicht zu zeigen (und gilt genauso z.B. auch für Folgen in $\mathbb{Q}$), dass konvergente Folgen immer auch Cauchy-Folgen sind. Dass auch umgekehrt Cauchy-Folgen immer einen Grenzwert besitzen ist jedoch die Besonderheit der der Menge der reellen Zahlen (die sogenannte *Vollständigkeit*), die diese von $\mathbb{Q}$ unterscheidet.

Satz 1.18 (Vollständigkeit der reellen Zahlen)
Eine reelle Folge $(a_n)_{n\in\mathbb{N}}$ besitzt genau dann einen Grenzwert $a \in \mathbb{R}$, wenn sie eine Cauchy-Folge ist.

(a) „$\Longrightarrow$": Sei $\lim_{n\to\infty} a_n = a$. Zum Beweis der Cauchy-Folgeneigenschaft müssen wir zeigen, dass zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ existiert, so dass

$$|a_m - a_n| < \varepsilon \quad \text{für alle } m \geq n \geq N.$$

Da $\lim_{n\to\infty} a_n = a$ existiert zu jedem $\varepsilon' > 0$ ein $N' \in \mathbb{N}$ mit

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

Wir verwenden dies mit der Wahl $\varepsilon' := \frac{\varepsilon}{2}$ und erhalten mit der Dreiecksungleichung

$$|a_m - a_n| = |a_m - a + (a - a_n)| \leq |a_m - a| + |a - a_n| < 2\varepsilon' = \varepsilon$$

für alle $m \geq n \geq N' =: N$.

(b) „$\Longleftarrow$": Sei $(a_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge. Wir müssen zeigen, dass dann ein Grenzwert $a \in \mathbb{R}$ existiert. Dies begründen wir im Folgenden mit unserer (nicht ganz mathematisch rigorosen, aber anschaulichen) Definition der reellen Zahlen als unendlich lange Kommazahlen. Da $(a_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist, gibt es für jedes $\varepsilon = 10^{-k-1}$ ein $N_k \in \mathbb{N}$, so dass sich ab dem N_k-ten Folgenglied alle weiteren jeweils paarweise nur um 10^{-k-1} unterscheiden, also (mit Ausnahme des unten behandelten Falles) alle die gleichen ersten k Kommastellen besitzen. Wir verwenden dies für alle $k \in \mathbb{N}$ und bilden so eine unendlich lange Kommazahl $a \in \mathbb{R}$. Für dieses a gilt dann aber, dass alle

Folgenglieder ab dem N_k-ten Folgenglied in den ersten k Kommastellen mit a übereinstimmen und $|a_n - a| < 10^{-k}$ für alle $n \geq N_k$ gilt. In jeder noch so kleinen Umgebung von a liegen also fast alle Folgenglieder und damit gilt $\lim_{n\to\infty} a_n = a$.

Wir haben dabei den Fall ignoriert, dass z.B. auch 1 und $0,999$ den Abstand 10^{-3} besitzen, aber nicht in den ersten beiden Kommastellen übereinstimmen. Für immer kleinere Abstände 10^{-k} tritt dieser Fall aber entweder irgendwann nicht mehr auf, oder die Kommastelle 9 tritt periodisch auf, was wir dann aber als die gleiche Zahl ansehen (z.B. $1 = 0.\overline{9}$).

Definition und Satz 1.19
Eine reelle Folge $(a_n)_{n\in\mathbb{N}}$ heißt monoton wachsend, falls $a_{n+1} \geq a_n$ für alle $n \in \mathbb{N}$ gilt. Gilt sogar $a_{n+1} > a_n$ für alle $n \in \mathbb{N}$ dann heißt sie streng monoton wachsend. Analog definieren wir (streng) monoton fallende Folgen.

Ist eine monoton wachsende Folgen nach oben beschränkt, so konvergiert sie. Gleiches gilt für monoton fallende, nach unten beschränkte Folgen.

Beweis: Wir beweisen die Behauptung durch Widerspruch, indem wir zeigen dass eine monoton wachsende, aber nicht konvergente Folge nicht nach oben beschränkt sein kann. Sei also $(a_n)_{n\in\mathbb{N}}$ monoton wachsend, aber nicht konvergent. Dann kann $(a_n)_{n\in\mathbb{N}}$ nach Satz 1.18 keine Cauchy-Folge sein. Durch die Negation der definierenden Eigenschaft einer Cauchy-Folge erhalten wir, dass dann also gelten muss

$$\exists \varepsilon > 0: \quad \forall N \in \mathbb{N}: \quad \exists m \geq n \geq N: \quad |a_n - a_m| \geq \varepsilon.$$

Wir beginnen mit der Wahl $N = 1$ und erhalten $m_1 \geq n_1 \geq 1$ so dass

$$a_{m_1} - a_1 \geq a_{m_1} - a_{n_1} = |a_{m_1} - a_{n_1}| \geq \varepsilon$$

wobei wir die Monotonie der Folge verwendet haben. Wir wiederholen dies mit der Wahl $N = m_1$ und erhalten $m_2 \geq n_2 \geq m_1$, so dass (wieder unter Verwendung der Monotonie)

$$a_{m_2} - a_{m_1} \geq a_{m_2} - a_{n_2} = |a_{m_2} - a_{n_2}| \geq \varepsilon,$$

also $a_{m_2} \geq a_1 + 2\varepsilon$. Wir fahren so fort und erhalten im k-ten Schritt ein Folgenglied a_{m_k} mit $a_{m_k} \geq a_1 + k\varepsilon$. Da $a_1 + k\varepsilon \to \infty$ für $k \to \infty$, können die Folgenglieder nicht nach oben beschränkt sein. Für monoton fallende Folgen folgt die entsprechende Aussage durch Negation. □

Beispiel 1.20
Die durch Intervallhalbierung zur Berechnung von $\sqrt{2}$ konstruierten Folgen der Intervallgrenzen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ erfüllen nach Konstruktion

$$a_1 \leq a_2 \leq a_3 \leq \ldots \leq b_3 \leq b_2 \leq b_1.$$

Die unteren Grenzen $(a_n)_{n\in\mathbb{N}}$ bilden also eine monoton wachsende, beschränkte Folge und die oberene Grenzen $(b_n)_{n\in\mathbb{N}}$ eine monoton fallende, beschränkte Folge. Nach Definition und Satz 1.19 konvergieren sie also. Außerdem gilt nach Konstruktion immer $b_n - a_n = \frac{1}{2^n}$ und damit erhalten wir aus den Rechenregeln für Grenzwerte (Satz 1.13)

$$c := \lim_{n\to\infty} b_n = \lim_{n\to\infty} a_n + \lim_{n\to\infty} \frac{1}{2^n} = \lim_{n\to\infty} a_n.$$

Aus Satz 1.13 folgt außerdem, dass $a_n^2 = a_n * a_n$ und $b_n^2 = b_n * b_n$ beide gegen c^2 konvergieren. Da nach Konstruktion gilt

$$a_n^2 < 2 < b_n^2$$

erhalten wir mit Satz 1.11

$$c^2 = \lim_{n\to\infty} a_n^2 \leq 2 \leq \lim_{n\to\infty} b_n^2 = c^2$$

und damit $c^2 = 2$. Die Folgen der unteren und der oberen Grenzen konvergieren also beide gegen den selben Grenzwert c und dieser ist $c = \sqrt{2}$.

Definition und Satz 1.21 (Maximum und Supremum)
Sei $M \subseteq \mathbb{R}$ eine Menge reeller Zahlen. Jedes $\hat{x} \in \mathbb{R}$ mit der Eigenschaft

$$x \leq \hat{x} \quad \text{für alle } x \in M$$

nennen wir obere Schranke für M. Offenbar existieren obere Schranken genau dann, wenn M nach oben beschränkt ist.

Liegt eine obere Schranke in M, so nennen wir sie das Maximum von M. Jede Menge kann höchstens ein Maximum besitzen, wir schreiben es als $\max M$, $\max\{x : x \in M\}$ oder auch $\max_{x\in M} x$.

Die kleinste obere Schranke von M, d.h. ein $x_{sup} \in \mathbb{R}$ mit der Eigenschaft

$$x \leq x_{sup} \quad \text{für alle } x \in M \quad \text{und} \quad \forall \varepsilon > 0 : \quad \exists x \in M : \quad x > x_{sup} - \varepsilon,$$

heißt Supremum von M. Jede nicht-leere, nach oben beschränkte Menge besitzt genau ein Supremum, wir schreiben es als $\sup M$ und verwenden außerdem die

Notation $\sup M = \infty$ *für eine nicht nach oben beschränkte Menge und* $\sup \emptyset = -\infty$ *für die leere Menge* $\emptyset$*. Falls ein Maximum existiert, dann gilt* $\max M = \sup M$*.*

Analog definieren wir das Minimum einer Menge, als eine in der Menge enthaltene untere Schranke, und das Infimum einer Menge als die größte untere Schranke. Alle obigen Aussagen gelten dafür analog.

Beweis: Die Tatsachen, dass es höchstens ein Maximum geben kann, und dass ein Maximum auch automatisch das Supremum ist, sind trivial.

Um zu zeigen, dass jede nicht-leere, nach oben beschränkte Menge M ein Supremum besitzt, wählen wir zunächst ein $a_1 \in \mathbb{R}$, welches keine obere Schranke ist (das existiert, da $M \neq \emptyset$) und ein $b_1 \in \mathbb{R}$ das eine obere Schranke ist (das existiert, da M nach oben beschränkt ist). Gemäß der Idee der Intervallhalbierung überprüfen wir dann, ob der Mittelwert $c_1 := \frac{a_1+b_1}{2}$ eine obere Schranke ist und ersetzen damit entsprechend entweder a_1 oder b_1, d.h. wir definieren $a_2, b_2 \in \mathbb{R}$ durch

$$a_2 := \begin{cases} a_1 & \text{falls } c_1 \text{ obere Schranke,} \\ c_1 & \text{sonst,} \end{cases} \qquad b_2 := \begin{cases} c_1 & \text{falls } c_1 \text{ obere Schranke,} \\ b_1 & \text{sonst.} \end{cases}$$

Wie in Beispiel 1.20 folgt dann, dass $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ gegen den selben Grenzwert $c \in \mathbb{R}$ konvergieren. Als Grenzwert der oberen Schranken b_n ist auch c eine obere Schranke (Satz 1.11) und aus $\lim_{n\to\infty} a_n = c$ folgert man leicht, dass es keine kleine obere Schranke geben kann. □

Wir haben in Satz 1.11 und Bemerkung 1.12 gesehen, dass konvergente Folgen beschränkt sind, aber beschränkte Folgen nicht notwendigerweise konvergent. Es gelten jedoch die folgenden beiden nützlichen Aussagen.

Satz 1.22 (Satz von Bolzano-Weierstraß)
Jede beschränkte reelle Folge $(a_n)_{n\in\mathbb{N}}$ *besitzt eine konvergente Teilfolge, d.h. es existieren unendlich viele Indizes*

$$n_1 < n_2 < n_3 < n_4 < \ldots,$$

so dass die Folge $(a_{n_k})_{k\in\mathbb{N}}$ *konvergiert.*

Beweis: Da die Folge beschränkt ist existiert ein Intervall, in dem alle unendlich vielen Folgenglieder liegen. Wir halbieren das Intervall, dann müssen in einer der beiden Hälften unendlich viele Folgenglieder liegen. Aus diesem Teilintervall wählen wir a_{n_1}. Dann halbieren wir das Teilintervall und es müssen wiederum in einer der Hälften unendlich viele Folgenglieder liegen. Aus dieser Hälfte des Teilintervalls wählen wir a_{n_2} mit $n_2 > n_1$. Wir halbieren dann wiederum, fahren so fort, und erhalten eine Folge $a_{n_1}, a_{n_2}, a_{n_3}, \ldots$, die aus Elementen aus immer

kleineren Teilintervallen besteht. Ist $C > 0$ die Breite des ersten Intervalls, so gilt nach Konstruktion für alle $j \geq k$, dass a_{n_j} und a_{n_k} in einem Teilintervall der Breite $2^{-k}C$ liegen, also

$$|a_{n_j} - a_{n_k}| \leq 2^{-k}C \quad \text{für alle} \quad j \geq k.$$

Die Teilfolge $(a_{n_k})_{k\in\mathbb{N}}$ ist also eine Cauchy-Folge und daher konvergent. □

Satz 1.23
Ist $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ eine beschränkte Folge und besitzt jede konvergente Teilfolge von $(a_n)_{n\in\mathbb{N}}$ denselben Grenzwert $a \in \mathbb{R}$, dann ist die ganze Folge konvergent und es gilt

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

Beweis: Angenommen $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ konvergiere nicht gegen a. Dann gibt es eine ε-Umgebung von a, außerhalb derer unendliche viele Folgenglieder liegen. Diese unendlich vielen Folgenglieder bilden dann aber eine beschränkte Folge, die nach Satz 1.22 eine konvergente Teilfolge besitzt, deren Grenzwert (wegen Satz 1.11) außerhalb der ε-Umgebung von a liegt. Dies widerspricht der Voraussetzung, dass alle konvergente Teilfolgen gegen a konvergieren. □

1.5 Reihen

Definition 1.24
Sei $(a_n)_{n\in\mathbb{N}}$ eine reelle Folge. Durch Summation der jeweils ersten n Folgenglieder entstehen die Partialsummen

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

Der Grenzwert der Folge der Partialsummen entspricht der unendlichen Summe (auch: Reihe) aller Folgenglieder a_n, $n \in \mathbb{N}$ (analog im Falle $n \in \mathbb{N}_0$).

Wir schreiben dies als $\sum_{k=1}^{\infty} a_k$ und bezeichnen damit sowohl die Folge der Partialsummen $(s_n)_{n\in\mathbb{N}}$ als auch ihren Grenzwert (im Falle der Konvergenz oder bestimmten Divergenz)

$$\sum_{k=1}^{\infty} a_k := \lim_{n\to\infty} \sum_{k=1}^{n} a_k = \lim_{n\to\infty} s_n.$$

Beispiele 1.25

(a) Jede reelle Zahl kann als Reihe geschrieben werden, z.B.

$$\frac{1}{3} = 0.333\ldots = 3*10^{-1}+3*10^{-2}*3*10^{-3}+\ldots = \sum_{k=1}^{\infty} 3*10^{-k},$$

$$\sqrt{2} = 1.414\ldots = 1*10^{0}+4*10^{-1}+1*10^{-2}*4*10^{-3}+\ldots = \sum_{k=0}^{\infty} z_k 10^{-k},$$

wobei $z_k \in \{0,1,\ldots,9\}$ die k-te Kommastelle von $\sqrt{2}$ bezeichne. Insbesondere ist jede reelle Zahl ein Grenzwert von rationalen Zahlen. ($\mathbb{Q}$ liegt dicht in $\mathbb{R}$).

(b) Jede Reihe ist eine Folge (der Partialsummen). Umgekehrt kann auch jede Folge $(a_n)_{n\in\mathbb{N}_0}$ als Reihe geschrieben werden (Teleskopsummenargument)

$$a_n = (a_n - a_{n-1}) + (a_{n-1} - a_{n-2}) + \ldots + (a_1 - a_0) + a_0 = a_0 + \sum_{k=1}^{n} (a_k - a_{k-1}).$$

Wir geben einige wichtige Reihen und die dazugehörigen Summenformeln an.

Satz 1.26

(a) Gaußsche Summenformel: Für jedes $n \in \mathbb{N}$ gilt

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

Offenbar gilt $\sum_{k=1}^{\infty} k = \infty$.

(b) Geometrische Summenformel: Für jedes $q \in \mathbb{R}$, $q \neq 1$ und $n \in \mathbb{N}$ gilt

$$\sum_{k=0}^{n} q^k = 1+q+q^2+q^3+\ldots = \frac{1-q^{n+1}}{1-q}.$$

Für die zugehörige geometrische Reihe gilt:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \quad \text{falls } |q| < 1.$$

Für $q \geq 1$ ist $\sum_{k=0}^{\infty} q^k = \infty$ (also bestimmt divergent) und für $q \leq -1$ divergiert $\sum_{k=0}^{\infty} q^k$ unbestimmt.

(c) Harmonische Reihe: Es gilt

$$\sum_{k=1}^{\infty} \frac{1}{k} = 1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots = \infty.$$

(d) *Alternierende harmonische Reihe: Es gilt*

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots = \ln(2).$$

(e) *Basler Problem: Es gilt*

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \ldots = \frac{\pi^2}{6}.$$

(f) *Allgemeine harmonische Reihe: Die Reihe*

$$\sum_{k=1}^{\infty} \frac{1}{k^{\alpha}}, \quad \alpha \in \mathbb{R},$$

konvergiert für jedes $\alpha > 1$ und divergiert für jedes $\alpha \leq 1$.

Beweis: (a) Zur anschaulichen Herleitung schreiben wir die Summe gemäß folgendem Schema zweimal auf (einmal davon in umgekehrter Reihenfolge) und addieren untereinander stehende Zahlen:

$$\begin{array}{cccccc} 1 & +2 & +3 & +4 & +\ldots & +n \\ +n & +(n-1) & +(n-2) & +(n-3) & +\ldots & +1 \\ \hline = n+1 & +(n+1) & +(n+1) & +(n+1) & +\ldots & +(n+1). \end{array}$$

So erhalten wir

$$2\sum_{k=1}^{n} k = n * (n+1)$$

und damit die Gaußsche Summenformel.

(b) Zur anschaulichen Herleitung vergleichen wir die Summe gemäß folgendem Schema mit der mit q multiplizierten Summe:

$$\begin{aligned} \sum_{k=1}^{n} q^k = &\quad 1 \quad +q \quad +q^2 \quad +q^3 \quad +\ldots \quad +q^n \\ q\sum_{k=1}^{n} q^k = &\quad\quad +q \quad +q^2 \quad +q^3 \quad +\ldots \quad +q^n \quad +q^{n+1} \end{aligned}$$

Durch Subtraktion erhalten wir

$$(1-q)\sum_{k=1}^{n} q^k = \sum_{k=1}^{n} q^k - q\sum_{k=1}^{n} q^k = 1 - q^{n+1}$$

und damit die geometrische Summenformel. Das Grenzwertverhalten der Reihe folgt dann daraus, dass $q^n \to 0$ für $|q| < 1$, $q^n \to \infty$ für $q > 1$ und dass für $q \leq -1$ die Folge q^n unbestimmt divergiert. Dies zeigen wir auf dem 4. Übungsblatt. Für den verbleibenden Fall $q = 1$ ist $\sum_{k=0}^{\infty} q^k = 1 + 1 + \ldots = \infty$.

(c) Wir verwenden das gleiche Argument wie in Beispiel 1.17(c). Gemäß dem Schema

$$\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{\geq 2*\frac{1}{4}=\frac{1}{2}} + \underbrace{\frac{1}{5} + \ldots + \frac{1}{8}}_{\geq 4*\frac{1}{8}=\frac{1}{2}} + \underbrace{\frac{1}{9} + \ldots + \frac{1}{16}}_{\geq 8*\frac{1}{16}=\frac{1}{2}} + \ldots$$

finden wir in der harmonischen Reihe unendlich oft jeweils 2^k Summanden größer gleich $\frac{1}{2^{k+1}}$, die jeweils zusammen eine Zahl $\geq \frac{1}{2}$ ergeben. Die harmonische Reihe ist daher größer als unendlich oft $\frac{1}{2}$ aufzusummieren und damit gilt $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

(d) Wir zeigen nur die Konvergenz. Die Summanden in $s_n = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}$ werden betragsmäßig immer kleiner und alternieren dabei im Vorzeichen. Für jedes ungerade k vergrößert sich die Summe (aber weniger als sich im vorherigen Schritt verkleinert hat) und mit jedem geraden k verkleinert sich die Summe (aber weniger als sich im vorherigen Schritt vergrößert hat). Mit dieser Idee (*Leibniz-Kriterium*) kann man zeigen, dass

$$s_2 \leq s_4 \leq s_6 \leq \ldots s_5 \leq s_3 \leq s_1.$$

Da der Abstand von s_n und s_{n+1} gegen Null geht und alle nachfolgenden s_m, $m \geq n$ sich zwischen s_n und s_{n+1} befinden, ist die Folge der Partialsummen eine Cauchy-Folge und damit konvergent.

(e) und (f) beweisen wir in dieser Vorlesung nicht. □

Definition und Satz 1.27 (Absolute Konvergenz)
Eine Reihe $\sum_{n=0}^{\infty} a_n$ heißt absolut konvergent, falls $\sum_{n=0}^{\infty} |a_n|$ konvergiert. Jede absolut konvergente Reihe konvergiert.

Beweis: Seien $s_n := \sum_{k=0}^{n} a_k$ und $S_n := \sum_{k=0}^{n} |a_n|$ die jeweiligen Folgen von Partialsummen. Für $m > n$ ist

$$|s_m - s_n| = |a_{n+1} + a_{n+2} + \ldots + a_m| \leq |a_{n+1}| + |a_{n+2}| + \ldots + |a_m| = |S_m - S_n|.$$

Deswegen gilt

$$S_n \text{ Cauchy-Folge} \quad \Longrightarrow \quad s_n \text{ Cauchy-Folge}$$

und daher folgt aus der Konvergenz der S_n auch die der s_n. □

Bemerkung 1.28
Für endliche Summen ist nach dem Kommutativgesetz die Reihenfolge der Summanden beliebig vertauschbar. Man kann zeigen, dass dies auch für absolut konvergente Reihen gilt. Diese können beliebig umgeordnet werden, d.h. statt

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + \dots$$

addiert man

$$\sum_{n=1}^{\infty} a_{p(n)} = a_{p(1)} + a_{p(2)} + a_{p(3)} + \dots$$

mit einer beliebigen bijektiven Abbildung (Permutation) $p : \mathbb{N} \to \mathbb{N}$ ohne dass sich dadurch das Konvergenzverhalten oder der Grenzwert ändert.

Für konvergente, aber nicht absolut konvergente Reihen gilt dies nicht. Man kann sogar zeigen, dass für solche (auch bedingt konvergent genannte) Reihen, durch Umordnung jeder beliebige Grenzwert sowie auch Divergenz gegen ∞ oder $-\infty$ erreicht werden kann (Riemannscher Umordnungssatz).

1.6 Die Exponentialreihe

Zur Motivation der Exponentialfunktion und -reihe betrachten wir eine Geldanlage von 100.000 EUR mit jährlichem *Nominalzins* von $x = 2,4\% = 0,024$. Bei jährlicher Verzinsung erhält der Anleger am Ende eines Jahres

$$100.000 * (1 + 0,024) = 102.400.$$

Werden die Zinsen unterjährig gezahlt, z.B. halbjährlich $x/2 = 1,2\%$, quartalsweise $x/4 = 0,6\%$ oder monatlich $x/12 = 0,2\%$, so profitiert der Anleger zusätzlich vom Zinseszinseffekt (auf Cent gerundet):

$$\begin{aligned} \text{Halbjährlich: } & 100.000 * (1 + 0.012)^2 = 102.414,40. \\ \text{Quartalsweise: } & 100.000 * (1 + 0.006)^4 = 102.421,69. \\ \text{Monatlich: } & 100.000 * (1 + 0.002)^{12} = 102.426,58. \end{aligned}$$

und erhält damit einen *Effektivzins* von $2,4144\%$, $2,42169\%$, bzw. $2,42658\%$.

Bei n Zinszahlungen im Jahr erhält der Anleger entsprechend

$$100.000 * (1 + x/n)^n.$$

Wir betrachten nun den Grenzwert immer häufigerer Zinszahlungen (*kontinuierliche Verzinsung*)

Definition und Satz 1.29
Für jedes $x \in \mathbb{R}$ konvergiert die reelle Folge $(1+\frac{x}{n})^n$. Wir definieren damit die Exponentialfunktion

$$\exp: \mathbb{R} \to \mathbb{R}, \quad \exp(x) := \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n.$$

und die Eulersche Zahl $e := \exp(1) = 2,718\ldots$

Beweis: Aus der obigen Motivation ist anschaulich klar, dass die Folge $(1+\frac{x}{n})^n$ monoton wachsend ist (für $x \geq 0$ erhöhen häufigere Zinszahlungen den Zinseszinseffekt und für $x < 0$ reduzieren häufigere Abzüge den effektiven Negativzins). Auf dem 4. Übungsblatt beweisen wir dies auch mathematisch rigoros.

Man kann zeigen, dass sie auch beschränkt ist. Wir zeigen das im Folgenden für den Spezialfall $x := 1$ und behandeln den allgemeinen Fall $x \in \mathbb{R}$ auf dem 4. Übungsblatt. Nach dem Binomischen Lehrsatz (Übungsaufgabe 1.2(b)) gilt

$$\begin{aligned}\left(1+\frac{1}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \frac{1}{n^k} = 1 + \sum_{k=1}^{n} \binom{n}{k} \frac{1}{n^k} \leq 1 + \sum_{k=1}^{n} \frac{1}{k!} \leq 1 + \sum_{k=1}^{n} \frac{1}{2^{k-1}} \\ &= 1 + \sum_{k=0}^{n} \frac{1}{2^k} \leq 1 + \sum_{k=0}^{\infty} \frac{1}{2^k} = 1 + \frac{1}{1-\frac{1}{2}} = 3.\end{aligned}$$

wobei wir im letzten Schritt die geometrische Reihe aus Satz 1.26 verwendet haben und für die ersten beiden Abschätzungen ausgenutzt haben, dass für alle $n \geq k \geq 1$ gilt

$$\binom{n}{k} = \frac{1}{k!} \underbrace{n(n-1)\ldots(n-k+1)}_{k \text{ Faktoren} \leq n} \leq \frac{n^k}{k!} \quad \text{und} \quad k! = \underbrace{k(k-1)\ldots*2}_{k-1 \text{ Faktoren} \geq 2} * 1 \geq 2^{k-1}.$$

Aus der Monotonie und Beschränktheit folgt dann die Konvergenz mit Definition und Satz 1.19. □

Satz 1.30
Die Exponentialfunktion kann als absolut konvergente Reihe geschrieben werden (Exponentialreihe). Es gilt

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

und sie erfüllt die Funktionalgleichung

$$\exp(x+y) = \exp(x)\exp(y).$$

Außerdem gilt $\exp(0) = 1$, $\exp(-x) = \frac{1}{\exp(x)}$ und $\exp(x) > 0$ für alle $x \in \mathbb{R}$.

Beweis: (a) Wir zeigen zuerst die absolute Konvergenz der Exponentialreihe für jedes $x \in \mathbb{R}$. Die Folge der Partialsummen

$$\sum_{k=0}^{n} \left| \frac{x^k}{k!} \right| = \sum_{k=0}^{n} \frac{|x|^k}{k!}$$

ist offenbar monoton wachsend. Sei $K \in \mathbb{N}$ so groß, dass $K+1 > 2|x|$. Dann gilt für alle $k > K$

$$\frac{|x|^k}{k!} = \underbrace{\frac{|x|}{1}}_{\leq |x|} \cdot \underbrace{\frac{|x|}{2}}_{\leq |x|} \cdots \underbrace{\frac{|x|}{K}}_{\leq |x|} \cdot \underbrace{\frac{|x|}{K+1}}_{\leq \frac{1}{2}} \cdot \underbrace{\frac{|x|}{K+2}}_{\leq \frac{1}{2}} \cdots \underbrace{\frac{|x|}{k}}_{\leq \frac{1}{2}} \leq |x|^K \left(\frac{1}{2}\right)^{k-K} = (2|x|)^K \cdot \frac{1}{2^k}$$

und mit $\sum_{k=K+1}^{n} \frac{1}{2^k} \leq \sum_{k=0}^{\infty} \frac{1}{2^k} = 2$ folgt damit

$$\begin{aligned} \sum_{k=0}^{n} \frac{|x|^k}{k!} &\leq \sum_{k=0}^{K} \frac{|x|^k}{k!} + \sum_{k=K+1}^{n} \frac{|x|^k}{k!} \leq \sum_{k=0}^{K} \frac{|x|^k}{k!} + (2|x|)^K \sum_{k=K+1}^{n} \frac{1}{2^k} \\ &\leq \sum_{k=0}^{K} \frac{|x|^k}{k!} + 2(2|x|)^K. \end{aligned}$$

Das zeigt die Beschränktheit der Folge der Partialsummen und damit die absolute Konvergenz der Exponentialreihe.

(b) Als nächstes zeigen wir, dass

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Aus dem Binomischen Lehrsatz (Übungsaufgabe 1.2(b)) erhalten wir

$$\left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \frac{x^k}{n^k}$$

Dabei ist

$$\begin{aligned} \binom{n}{k} \frac{x^k}{n^k} &= \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} \frac{x^k}{n^k} \\ &= \frac{(1-\frac{1}{n})(1-\frac{2}{n})\dots(1-\frac{n-k+1}{n})x^k}{k!} \leq \frac{x^k}{k!}. \end{aligned}$$

Mit $P_{kn} := 1 - (1-\frac{1}{n})(1-\frac{2}{n})\dots(1-\frac{n-k+1}{n})$ folgt

$$0 \leq \sum_{k=0}^{n} \frac{x^k}{k!} - \sum_{k=0}^{n} \binom{n}{k} \frac{x^k}{n^k} = \sum_{k=0}^{n} \frac{x^k}{k!} P_{kn}.$$

Um zu zeigen, dass diese Differenz für $n \to \infty$ gegen Null konvergiert, sei $\varepsilon > 0$. Dann existiert wegen der schon gezeigten absoluten Konvergenz der Exponentialreihe ein $K \in \mathbb{N}$, so dass

$$\sum_{k=K}^{n} \frac{x^k}{k!} P_{kn} \leq \sum_{k=K}^{n} \frac{x^k}{k!} \leq \frac{\varepsilon}{2} \quad \text{für alle } n > K$$

Außerdem gilt

$$\sum_{k=1}^{K-1} \frac{x^k}{k!} P_{kn} \to 0.$$

Daher existiert ein $N > K$ so dass $\sum_{k=1}^{K-1} \frac{x^k}{k!} P_{kn} < \frac{\varepsilon}{2}$ für alle $n \geq N$. Insgesamt ist daher für alle $n \geq N$

$$\sum_{k=0}^{n} \frac{x^k}{k!} P_{kn} = \sum_{k=0}^{K-1} \frac{x^k}{k!} P_{kn} + \sum_{k=K}^{n} \frac{x^k}{k!} P_{kn} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

und damit haben wir gezeigt, dass

$$\exp(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

(c) Zum Beweis der Funktionalgleichung betrachten wir

$$\begin{aligned} \exp(x)\exp(y) &= \left(\sum_{j=0}^{\infty} \frac{x^j}{j!}\right)\left(\sum_{k=0}^{\infty} \frac{y^k}{k!}\right) \\ &= \left(\frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots\right)\left(\frac{y^0}{0!} + \frac{y^1}{1!} + \frac{y^2}{2!} + \frac{y^3}{3!} + \dots\right). \end{aligned}$$

Beim ausmultiplizieren diese beiden unendlichen Summen ergeben sich unendlich viele Summanden entsprechend des folgenden Schemas:

$$\begin{array}{cccccc} \frac{x^0}{0!}\frac{y^0}{0!} & +\frac{x^0}{0!}\frac{y^1}{1!} & +\frac{x^0}{0!}\frac{y^2}{2!} & +\frac{x^0}{0!}\frac{y^3}{3!} & +\dots \\ +\frac{x^1}{1!}\frac{y^0}{0!} & +\frac{x^1}{1!}\frac{y^1}{1!} & +\frac{x^1}{1!}\frac{y^2}{2!} & +\frac{x^1}{1!}\frac{y^3}{3!} & +\dots \\ +\frac{x^2}{2!}\frac{y^0}{0!} & +\frac{x^2}{2!}\frac{y^1}{1!} & +\frac{x^2}{2!}\frac{y^2}{2!} & +\frac{x^2}{2!}\frac{y^3}{3!} & +\dots \\ +\frac{x^3}{3!}\frac{y^0}{0!} & +\frac{x^3}{3!}\frac{y^1}{1!} & +\frac{x^3}{3!}\frac{y^2}{2!} & +\frac{x^3}{3!}\frac{y^3}{3!} & +\dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

Man kann zeigen, dass das Produkt zweier absolut konvergenter Reihen ebenfalls absolut konvergiert und wir diese unendlich vielen Summanden in einer von uns gewählten Reihenfolge aufsummieren können. Wir summieren entlang der (im obigen Schema grau hinterlegten) endlichen Diagonalen (*Cauchy-Produkt*) und erhalten

$$\exp(x)\exp(y) = \frac{x^0}{0!}\frac{y^0}{0!} + \left(\frac{x^0}{0!}\frac{y^1}{1!} + \frac{x^1}{1!}\frac{y^0}{0!}\right) + \left(\frac{x^0}{0!}\frac{y^2}{2!} + \frac{x^1}{1!}\frac{y^1}{1!} + \frac{x^2}{2!}\frac{y^0}{0!}\right) + \ldots$$

Die Summe über die $n+1$-te Diagonale lautet dabei

$$\sum_{k=0}^{n} \frac{x^k}{k!}\frac{y^{n-k}}{(n-k)!},$$

und mit $\binom{n}{k} = \frac{n!}{k!}(n-k)!$ und dem Binomischen Lehrsatz ergibt sich

$$\sum_{k=0}^{n} \frac{x^k}{k!}\frac{y^{n-k}}{(n-k)!} = \frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k} = \frac{1}{n!}(x+y)^n.$$

Insgesamt folgt also mit Summation über alle Diagonalen die Funktionalgleichung

$$\exp(x)\exp(y) = \sum_{n=0}^{\infty}\sum_{k=0}^{n}\frac{x^k}{k!}\frac{y^{n-k}}{(n-k)!} = \sum_{k=0}^{n}\frac{(x+y)^n}{n!} = \exp(x+y).$$

Offenbar gilt $\exp(0) = 1$ und $\exp(x) > 0$ für alle $x \geq 0$. Aus der Funktionalgleichung erhalten wir außerden, dass

$$\exp(x)\exp(-x) = \exp(x-x) = \exp(0) = 1, \quad \text{also} \quad \exp(-x) = \frac{1}{\exp(x)},$$

und damit auch $\exp(x) > 0$ für alle $x \in \mathbb{R}$. □

Folgerung 1.31
Für die Eulersche Zahl gilt nach Definition $e^1 = exp(1)$ und wegen der Funktionalgleichung auch für jede natürliche Zahl $n \in \mathbb{N}$

$$e^n = e * \ldots * e = \exp(1) * \ldots * \exp(1) = \exp(1+\ldots+1) = \exp(n).$$

Für jede positive rationale Zahl $\frac{p}{q} \in \mathbb{Q}$ mit $p \in \mathbb{N}$, $q \in \mathbb{N}$ gilt

$$\underbrace{\exp\left(\frac{p}{q}\right) * \ldots * \exp\left(\frac{p}{q}\right)}_{q \text{ mal}} = \exp\underbrace{\left(\frac{p}{q}+\ldots+\frac{p}{q}\right)}_{q \text{ mal}} = \exp(p),$$

also

$$\exp\left(\frac{p}{q}\right) = \sqrt[q]{\exp(p)} = \sqrt[q]{e^p} = e^{\frac{p}{q}}.$$

Dies gilt auch für nicht-positive rationale Zahlen, da

$$\exp(0) = 1 \quad \textit{und} \quad \exp(-r) = \frac{1}{\exp(r)} = \frac{1}{e^r} = e^{-r} \quad \forall r \in \mathbb{Q},\ r > 0.$$

Insgesamt gilt also $\exp(x) = e^x$ *für alle* $x \in \mathbb{Q}$.

Die Exponentialfunktion ist auch für nicht-rationale Argumente definiert. In Definition und Satz 2.18 werden wir damit nicht-rationale Potenzen der Eulerschen Zahl definieren durch

$$e^x := \exp(x) \quad \forall x \in \mathbb{R} \setminus \mathbb{Q}.$$

Auch die Sinus- und Kosinusfunktion lassen sich über Reihen definieren. Man kann zeigen, dass die folgenden Reihen tatsächlich die aus der Geometrie bekannten Winkelfunktionen liefern. Die große Ähnlichkeit zur Exponentialreihe suggeriert einen Zusammenhang zur Exponentialfunktion, dieser wird aber erst in Abschnitt 6.3 mit der Einführung der komplexen Zahlen deutlich.

Definition und Satz 1.32 (Sinus- und Kosinusreihe)
Wir definieren die Funktionen $\sin : \mathbb{R} \to \mathbb{R}$ *und* $\cos : \mathbb{R} \to \mathbb{R}$ *durch*

$$\sin(x) := \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \quad \textit{und} \quad \cos(x) := \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}.$$

Die Reihen konvergieren für jedes $x \in \mathbb{R}$ *absolut.*

Beweis: In beiden Fällen ist die Folge der Partialsummen der Absolutbeträge offenbar monoton wachsend und nach oben beschränkt durch $\exp(x)$ und daher konvergent. □

1.7 Asymptotisches Verhalten von Folgen

In praktischen Anwendungen ist es wichtig zu charakterisieren, wie schnell der Fehler eines Algorithmus gegen Null konvergiert oder wie schnell der Aufwand eines Algorithmus gegen Unendlich strebt. Wir führen deshalb die folgende Notation für das asympotische Verhalten von Folgen ein.

Definition 1.33 (Landau-Notation für Folgen)
Sei $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$. Wir schreiben

$$a_n \in o(b_n), \quad \text{falls} \quad \frac{a_n}{b_n} \to 0 \quad \text{für } n \to \infty,$$

und

$$a_n \in O(b_n), \quad \text{falls} \quad \exists C > 0 : \ |a_n| \leq C|b_n| \quad \text{für fast alle } n \in \mathbb{N}.$$

Insbesondere bedeutet $a_n \in O(1)$, dass die Folge $(a_n)_{n\in\mathbb{N}}$ beschränkt ist, und $a_n \in o(1)$ bedeutet, dass $(a_n)_{n\in\mathbb{N}}$ eine Nullfolge ist.

Anschaulich bedeutet $a_n \in o(b_n)$, dass a_n schneller als b_n gegen Null konvergiert oder dass a_n langsamer gegen Unendlich strebt als b_n. $a_n \in O(b_n)$ bedeutet anschaulich, dass a_n so schnell wie b_n gegen Null konvergiert oder nicht schneller als b_n gegen Unendlich strebt.

Beispiele 1.34
(a) $2n^2+30n+1 \in O(n^2)$ *und* $2n^2+30n+1 \in o(n^3)$.

(b) $\frac{n+1}{2n^2+30n} \in O(\frac{1}{n})$ *und* $\frac{n+1}{2n^2+30n} \in o(\frac{1}{\sqrt{n}})$

(c) Für die durch Intervallhalbierung zur Berechnung von $\sqrt{2}$ entstehenden Folgen a_n, b_n gilt

$$a_n - \sqrt{2} \in O(2^{-n}) \quad \text{und} \quad b_n - \sqrt{2} \in O(2^{-n}).$$

Bemerkung 1.35
Es ist bei dieser Notation auch üblich, Gleichheitszeichen zu verwenden und feste Teile der Folge auf die rechte Seite zu schreiben, etwa in Beispiel 1.34(c)

$$a_n = \sqrt{2} + O(2^{-n}) \quad \text{und} \quad b_n = \sqrt{2} + O(2^{-n}).$$

Die Schreibweise ist sehr intuitiv, sowohl a_n als auch b_n stimmen mit $\sqrt{2}$ bis auf einen Fehler der Größenordnung 2^{-n} überein. Die üblichen Rechenregeln für das Gleichheitszeichen gelten dabei jedoch nicht, es folgt natürlich nicht $a_n = b_n$, sondern die Folgen (a_n) und (b_n) liegen nur beide in der gleichen Menge (nämlich der Menge der Folgen, deren Abstand zu $\sqrt{2}$ asymptotisch mindest so schnell wie 2^{-n} fällt). Daher ist die Schreibweise als Element mathematisch präziser.

Genauso ist beispielsweise $n^2 = O(n^2)$ aber auch $n^2 = O(n^3)$ ohne dass daraus $O(n^2) = O(n^3)$ folgt. Mathematisch präziser ist auch hier die Mengenschreibweise

$$n^2 \in O(n^2) \subseteq O(n^3).$$

1.8 Zahldarstellung im Computer

Mit n Bit $d_0,\dots,d_{n-1}\in\{0,1\}$ lassen sich die folgenden Zahlen kodieren (*unsigned integer*)

$$\sum_{k=0}^{n-1} d_k 2^k,$$

also die Zahlen $0,\ 1,\dots,\sum_{k=0}^{n-1}2^k=\frac{1-2^n}{1-2}=2^n-1$. Mit einem Byte bestehend aus 8 Bit erhält man also die Zahlen

$$\begin{aligned} 0&=\boxed{0\,0\,0\,0\,0\,0\,0\,0}\\ 1&=\boxed{0\,0\,0\,0\,0\,0\,0\,1}\\ &\vdots\\ 255&=\boxed{1\,1\,1\,1\,1\,1\,1\,1} \end{aligned}$$

und mit zwei, drei und vier Byte entsprechend die Zahlen von 0 bis $2^8-1=255$, bis $2^{16}-1=65.535$, bis $2^{24}-1=16.777.216$ und bis $2^32-1=4.294.967.296$.

Negative Zahlen (*signed integer*) lassen sich dadurch kodieren, dass man ein Bit für das Vorzeichen verwendet, so erhält man mit n Bit

$$(-1)^{d_{n-1}}\sum_{k=0}^{n-2} d_k 2^k,$$

also die Zahlen von $-2^{n-1}+1$ bis $2^{n-1}-1$. Damit hat man allerdings die Null zweimal kodiert, z.B. mit einem Byte

$$\begin{aligned} 0&=\boxed{0\,|\,0\,0\,0\,0\,0\,0\,0} & -0&=\boxed{1\,|\,0\,0\,0\,0\,0\,0\,0}\\ &\vdots & &\vdots\\ 127&=\boxed{0\,|\,1\,1\,1\,1\,1\,1\,1} & -127&=\boxed{1\,|\,1\,1\,1\,1\,1\,1\,1}. \end{aligned}$$

Alternativ könnte man einen festen *Bias* abziehen, und mit n Bit die Zahlen

$$-2^{n-1}+\sum_{k=0}^{n-1} d_k 2^k,$$

kodieren, z.B. mit einem Byte $-128,\dots,127$.

Gebräuchlich ist auch die Darstellung im sogenannten *Zweierkomplement*

$$\begin{cases} \sum_{k=0}^{n-2} d_k 2^k & \text{für } d_{n-1}=0,\\ -\sum_{k=0}^{n-2}(1-d_k)2^k-1 & \text{für } d_{n-1}=1. \end{cases}$$

Damit erhält man mit $d_{n-1} = 0$ die Zahlen $0, \ldots, 2^{n-1} - 1$ und mit $d_{n-1} = 1$ die Zahlen $-1, \ldots, -2^{n-1}$, z.B. mit einem Byte

$$\begin{aligned} 0 &= \boxed{0\,|\,0\,0\,0\,0\,0\,0\,0} & -128 &= \boxed{1\,|\,0\,0\,0\,0\,0\,0\,0} \\ &\vdots & &\vdots \\ 127 &= \boxed{0\,|\,1\,1\,1\,1\,1\,1\,1} & -1 &= \boxed{1\,|\,1\,1\,1\,1\,1\,1\,1}. \end{aligned}$$

So wird die Null nicht doppelt kodiert und die Addition zweier Zahlen ist im Zweierkomplement besonders einfach durch bitweise Addition durchzuführen, z.B. ist mit einem Byte

$$\begin{aligned} 13 &= \boxed{0\,|\,0\,0\,0\,1\,1\,0\,1} \\ -6 &= \boxed{1\,|\,1\,1\,1\,1\,0\,1\,0} \\ 13 + (-6) &= \boxed{0\,|\,0\,0\,0\,0\,1\,1\,1} = 7. \end{aligned}$$

Prinzipiell lässt sich mit beliebig großen ganzen Zahlen rechnen, indem bei Bedarf dynamisch Speicherplatz allokiert wird. Entsprechend kann man prinzipiell fehlerfrei mit rationalen Zahlen rechnen, indem bei jeder Rechnung gegebenenfalls der Speicher für Zähler und Nenner erhöht wird (*Arithmetik mit unendlicher Präzision / arbitrary-precision arithmetic*).

In den meisten Anwendungen begnügt man sich aber mit endlicher Präzision und stellt sehr (betrags-)große ganze Zahlen, rationale und reelle Zahlen als *Gleitkommazahlen* dar. Vereinfacht ausgedrückt würde man im Zehnersystem eine Zahl als

$$a * 10^b$$

schreiben und a und b jeweils als signed integer kodieren. Der verwendete Speicherplatz für den *Exponenten* b steuert dabei, wie groß und wie klein die Zahlen werden können, und der Speicherplatz für die *Mantisse* a steuert die Genauigkeit der Darstellung. Beispielsweise könnte man so im Zehnersystem mit $a \in \{-999, \ldots, 999\}$ und $b \in \{-9, \ldots, 9\}$, Zahlen von $1 * 10^{-9}$ bis $9{,}99 * 10^{11}$ jeweils mit drei Stellen Genauigkeit darstellen, hätte aber noch Doppelkodierungen z.B. $10 * 10^0 = 1 * 10^1$.

Der seit 1985 bestehende (und letzmalig 2019 geringfügig überarbeiteten) IEEE Standard for Floating-Point Arithmetic (IEEE 754) kodiert *double precision* Zahlen mit 64 Bit indem 1 Bit $S \in \{0,1\}$ für das Vorzeichen, 11 Bit $e_0, \ldots, e_{10} \in \{0,1\}$ für den Exponenten und 52 Bit $m_0, \ldots, m_{51} \in \{0,1\}$ für die Mantisse verwendet werden. Aus den Exponenten- und Mantissenbits ergeben sich

$$E := \sum_{j=0}^{10} e_j 2^j \quad \text{und} \quad M := \sum_{j=0}^{51} e_j 2^j$$

und für $0 < E < 2^{11} - 1 = 2047$ kodiert man damit die sogenannten *normalisierten Zahlen*[2]

$$(-1)^S(1 + M * 2^{-52}) * 2^{E-1023}$$

d.h. M wird als Kommastellen hinter 1 gehangen und negative Werte im Exponent werden durch einen Bias erreicht. Die betragsmäßig kleinste normalisierte Zahl erhalten wir mit $M = 0$ und $E = 1$

$$1 * 2^{-1022} \approx 2 * 10^{-308}.$$

Die größte normaliserte Zahl erhalten wir mit Exponent $E = 2046$ und Mantisse $M = \sum_{j=0}^{51} 2^j = 2^{52} - 1$

$$(2 - 2^{-52}) * 2^{1023} \approx 1,8 * 10^{308}.$$

Die Genauigkeit dieser Zahldarstellung kann dadurch charakterisiert werden, dass der Abstand zwischen 1 und der nächstgrößeren darstellbaren Zahl $1 + \texttt{eps}$ angegeben wird. Bei den normalisierten double precision Zahlen ist

$$\texttt{eps} = 2^{-52} \approx 2,22 * 10^{-16}$$

Rechenergebnisse zwischen 1 und $1 + \texttt{eps}$ müssen auf eine dieser beiden Zahlen gerundet werden, wobei ein Fehler entsteht der bis zu $\texttt{eps}/2 \approx 1,11 * 10^{-16}$ (*Maschinengenauigkeit*) betragen kann. Durch die Gleitkommadarstellung gilt für Zweierpotenzen entsprechendes, z.B. liegt zwischen 1024 und $1024 + 1024\texttt{eps}$ keine darstellbare Zahl. In diesem Sinne können wir mit dem Format double precision Zahlen mit etwa 16 Stellen Genauigkeit darstellen.

Beispiel 1.36
Die Folge $(1 + \frac{1}{n})^n$ konvergiert für $n \to \infty$ gegen die Eulersche Zahl $e \approx 2,72$. Ab ca. $n = 10^{16}$ wird jedoch die Zahl $1 + \frac{1}{n}$ im double precision Format auf 1 gerundet und es ergibt sich die unbrauchbare Näherung $1^n = 1$. Tatsächlich werden die so berechneten Näherungen schon ab ca. $n = 10^8$ schlecht, da sich Fehler durch die Potenzierung mit immer größerem n immer mehr verstärken.

[2]Dazu kommen noch die folgenden Sonderfälle: Mit $E = 0$ und $M = 0$ wird die Null kodiert (mit Vorzeichen, also als $+0$ und -0 und mit $E = 0$ und $M > 0$ werden äquidistant Zahlen zwischen 0 und der betragskleinsten normalisierte Zahl aufgefüllt (*denormalisierte Zahlen*). Außerdem kodiert man mit $E = 2^{11} - 1$ und $M = 0$ unendlich, und mit $E = 2^{11} - 1$ und $M > 0$ sogenannte Nicht-Zahlen (*NaN*).

Kapitel 2

Stetige Funktionen

2.1 Motivation: Intervallhalbierung

Wir betrachten das Problem, eine Gleichung nach x aufzulösen, d.h. die Lösung $x \in \mathbb{R}$ von $f(x) = y$ zu bestimmen, wobei $f : \mathbb{R} \to \mathbb{R}$, $y \in \mathbb{R}$. Falls wir zwei Werte $a_0 < b_0 \in \mathbb{R}$ finden mit $f(a_0) < y < f(b_0)$, so können wir vermuten, dass zwischen diesen Werten eine Lösung von $f(x) = y$ liegt (analog falls $f(a_0) > y > f(b_0)$). Wie bei der Bestimmung von $\sqrt{2}$ (also der Lösung der Gleichung $f(x) = 2$ mit $f(x) = x^2$) können wir versuchen, uns dieser Lösung durch Intervallhalbierung anzunähern. Wir definieren also[1] für alle $n \in \mathbb{N}_0$

$$(a_{n+1}, b_{n+1}) := \begin{cases} (a_n, \frac{a_n+b_n}{2}) & \text{falls } f\left(\frac{a_n+b_n}{2}\right) \geq y, \\ (\frac{a_n+b_n}{2}, b_n) & \text{falls } f\left(\frac{a_n+b_n}{2}\right) < y. \end{cases}$$

Wie im letzten Abschnitt gezeigt erhalten wir so zwei Folgen

$$a_0 \leq a_1 \leq \ldots \leq a_n \leq \ldots \leq b_n \leq \ldots \leq b_1 \leq b_0,$$

wobei $(a_n)_{n\in\mathbb{N}}$ montonon wachsend und $(b_n)_{n\in\mathbb{N}}$ monoton fallend ist und beide gegen den selben Grenzwert konvergieren

$$\hat{x} := \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

[1] Wenn für den Intervallmittelwert gilt $f\left(\frac{a_n+b_n}{2}\right) = y$, dann hat man glücklicherweise die exakte Lösung gefunden und man kann das Verfahren abbrechen. Der Kürze halber werden wir diesen Fall aber nicht immer separat abfangen.

Erfüllt dieser Grenzwert dann wirklich $f(\hat{x}) = y$? Nach Konstruktion wissen wir, dass

$$f(a_n) < y \leq f(b_n) \quad \text{für alle } n \in \mathbb{N}_0.$$

Falls sich die Konvergenz von $a_n, b_n \to \hat{x}$ auch auf die Funktionswerte überträgt, also $f(a_n) \to f(\hat{x})$ und $f(b_n) \to f(\hat{x})$ dann folgt (mit Satz 1.11)

$$f(\hat{x}) \leq y \leq f(\hat{x})$$

und damit tatsächlich $f(\hat{x}) = y$.

Das kann aber nicht für alle Funktionen zutreffen, z.B. ist für die Sprungfunktion (auch Heavyside-Funktion)

$$H(x) := \begin{cases} 0 & \text{für } x < 0 \\ 1 & \text{für } x \geq 0 \end{cases}$$

offenbar $H(-1) < \frac{1}{2} \leq H(1)$, aber $H(x) = \frac{1}{2}$ besitzt gar keine Lösung.

Der in diesem Abschnitt eingeführte Begriff der *Stetigkeit* bedeutet anschaulich, dass solche Sprünge nicht auftreten und lässt sich mathematisch über die Übertragung der Grenzwerteigenschaft auf die Funktionswerte einer Folge definieren. Die Konvergenz des Intervallhalbierungsverfahrens lässt sich damit garantieren.

2.2 Stetigkeit

Im Folgenden sei stets $f: D \to \mathbb{R}$ eine auf einer reellen Teilmenge $D \subseteq \mathbb{R}$ definierte Funktion. D nennen wir auch die *Definitionsmenge* der Funktion.

Definition 2.1
Wir definieren den Abschluss einer Menge $D \subseteq \mathbb{R}$ durch die Hinzunahme aller Punkte, die als Grenzwerte von Folgen in D erreicht werden können, d.h.

$$\overline{D} := \{x \in \mathbb{R}, \quad \exists (x_n)_{n\in\mathbb{N}} \subseteq D : \ x_n \to x\}.$$

Beispiele 2.2
(a) Für jedes $x \in D$ konvergiert $x_n := x$ gegen x. Es gilt also $D \subseteq \overline{D}$.

(b) Für das offene Intervall $D :=]a,b[$ ist $\overline{D} = [a,b]$.

(c) Für das abgeschlossene Intervall $D := [a,b]$ ist $\overline{D} = D = [a,b]$.

(d) Für $D = \mathbb{N}$ ist $\overline{D} = D$.

Definition 2.3 (Grenzwert einer Funktion)
Sei $a \in \overline{D}$ und $\hat{y} \in \mathbb{R}$. Falls für jede Folge $(x_n)_{n\in\mathbb{N}} \subseteq D$ auch die Folge $(f(x_n))_{n\in\mathbb{N}}$ konvergiert und immer derselbe Grenzwert $\hat{y} = \lim_{n\to\infty} f(x_n)$ entsteht, dann nennen wir $\hat{y}$ den Grenzwert von f für $x \to a$ und schreiben

$$\lim_{x\to a} f(x) = \hat{y}.$$

Entsprechend schreiben wir auch $\lim_{x\to a} f(x) = \infty$ oder $\lim_{x\to a} f(x) = -\infty$, falls für jede Folge $x_n \to a$ die Folge $f(x_n)$ bestimmt divergent nach ∞ bzw. $-\infty$ ist. Ist D nach oben oder unten unbeschränkt, so definieren wir $\lim_{x\to\infty}$ und $\lim_{x\to-\infty}$ außerdem analog über die Betrachtung bestimmt divergenter Folge $x_n \to \infty$, bzw. $x_n \to \infty$.

Ebenfalls analog definiert man einseitige Grenzwerte, indem nur Folgen $x_n \to a$ mit $x_n \leq a$ bzw. nur Folgen mit $x_n \geq a$ betrachtet werden und bezeichnet diese mit

$$\lim_{x\to a^-} f(x) = \lim_{x\nearrow a} f(x) = \lim_{x\uparrow a} f(x) \quad \text{bzw.} \quad \lim_{x\to a^+} f(x) = \lim_{x\searrow a} f(x) = \lim_{x\downarrow a} f(x).$$

Beispiel 2.4
(a) Für die konstante Funktion $f: \mathbb{R} \to \mathbb{R}$, $f(x) := 1$ gilt für jedes $a \in \mathbb{R}$, dass

$$\lim_{x\to a} f(x) = 1, \quad \lim_{x\to\infty} = 1, \quad \text{und} \quad \lim_{x\to-\infty} = 1.$$

(b) Für die Funktion $f: \mathbb{R} \to \mathbb{R}$, $f(x) := x$ gilt für jedes $a \in \mathbb{R}$, dass

$$\lim_{x\to a} f(x) = a, \quad \lim_{x\to\infty} = \infty, \quad \text{und} \quad \lim_{x\to-\infty} = -\infty.$$

(c) Für die Sprungfunktion $H: \mathbb{R} \to \mathbb{R}$ gilt für

$$\lim_{x\to a} H(x) := \begin{cases} 0 & \text{für } a < 0, \\ 1 & \text{für } a > 0. \end{cases}$$

Außerdem ist $\lim_{x\to-\infty} H(x) = 0$ und $\lim_{x\to\infty} H(x) = 1$.
Der Grenzwert $\lim_{x\to 0} H(x)$ existiert jedoch nicht, da $-\frac{1}{n}, \frac{1}{n} \to 0$, aber

$$\lim_{n\to\infty} H(-1/n) = 0 \neq 1 = \lim_{n\to\infty} H(1/n).$$

Es existieren aber die einseitigen Grenzwerte

$$\lim_{x\to 0^-} H(x) = 0 \quad \text{und} \quad \lim_{x\to 0^+} H(x) = 1.$$

(d) Für die Funktion $f:\ \mathbb{R}\to\mathbb{R}$, $f(x):=1/x$ *gilt für jedes* $a\in\mathbb{R}$, *dass*

$$\lim_{x\to a} f(x) = 1/a, \quad \lim_{x\to\infty} = 0, \quad \textit{und} \quad \lim_{x\to-\infty} = 0,$$

Der Grenzwert $\lim_{x\to 0} f(x)$ *existiert nicht, aber es existieren die einseitigen Grenzwerte*

$$\lim_{x\to 0^-} 1/x = -\infty \quad \textit{und} \quad \lim_{x\to 0^+} \frac{1}{x} = \infty.$$

(e) Für das Polynom $p(x) = a_m x^m + a_{m-1}x^{m-1} + \ldots a_1 x + a_0 \in \Pi_m$ *mit Koeffizienten* $a_0,\ldots,a_m\in\mathbb{R}$, $a_m > 0$, $m\in\mathbb{N}$ *gilt*

$$\lim_{x\to\infty} p(x) = \infty, \quad \lim_{x\to 0} p(x) = a_0, \quad \lim_{x\to-\infty} p(x) = \begin{cases} \infty & \textit{für } m \textit{ gerade,} \\ -\infty & \textit{für } m \textit{ ungerade.} \end{cases}$$

Definition 2.5 (Stetigkeit)
Sei $a\in D$. *Die Funktion* $f:\ D\to\mathbb{R}$ *heißt stetig in* a, *falls*

$$\lim_{x\to a} f(x) = f(a).$$

f *heißt stetig (auch: stetig in* D*), falls dies für alle* $a\in D$ *gilt.*

Beispiele 2.6
(a) Für jedes $m\in\mathbb{N}_0$ *ist* $f(x):=x^m$ *stetig auf* $\mathbb{R}$.

(b) Die Sprungfunktion $H(x)$ *ist stetig für alle* $x\in\mathbb{R}\setminus 0$, *aber nicht stetig in* $x=0$.

(c) $f(x)=1/x$ *ist stetig auf* $D:=\mathbb{R}\setminus 0$.

(d) Die ReLU-Funktion (rectified linear unit), auch: Rampenfunktion, ist definiert durch

$$f:\ \mathbb{R}\to\mathbb{R}, \quad f(x):=\max\{0,x\} = \begin{cases} 0 & \textit{für } x<0, \\ x & \textit{für } x\geq 0. \end{cases}$$

Dies ist eine stetige Funktion auf $\mathbb{R}$.

Definition und Satz 2.7
Seien $f,g:\ D\to\mathbb{R}$ *zwei Funktionen. Dann definieren wir*

(a) für $r\in\mathbb{R}$ *das skalare Vielfache* $rf:\ D\to\mathbb{R}$ *durch* $(rf)(x):=rf(x)$.

(b) die Summe $f+g:\ D\to\mathbb{R}$ *durch* $(f+g)(x):=f(x)+g(x)$.

(c) das Produkt $fg:\ D\to\mathbb{R}$ *durch* $(fg)(x):=f(x)g(x)$.

(d) den Quotient $f/g: D' \to \mathbb{R}$ *durch* $(f/g)(x) := f(x)/g(x)$ *auf dem Definitionsbereich* $x \in D' := \{x \in D : g(x) \neq 0\}$

Sind f *und* g *stetig in* $a \in D$*, dann sind für alle* $r \in \mathbb{R}$ *auch* rf*,* $f+g$ *und* fg *stetig in* a*. Gilt zusätzlich* $g(a) \neq 0$*, so ist auch* f/g *stetig in* a*.*

Beweis: Dies folgt leicht aus den Rechenregeln für Grenzwerte von Folgen. □

Definition und Satz 2.8
Seien $f: D_f \to \mathbb{R}$ *und* $g: D_g \to \mathbb{R}$ *zwei Funktionen und es gelte*

$$g(D_g) := \{g(x) : x \in D_g\} \subseteq D_f.$$

Dann definieren wir die Hintereinanderausführung (auch: Konkatenation)

$$f \circ g: D_g \to \mathbb{R}, \quad (f \circ g)(x) := f(g(x)).$$

Ist g *stetig in* $a \in D_g$ *und ist* f *stetig in* $g(a) \in D_f$*, so ist* $f \circ g$ *stetig in* a*.*

Beweis: Um zu zeigen, dass $f \circ g$ stetig ist in $a \in D_g$ müssen wir zeigen, dass für jede Folge $x_n \to a$ gilt

$$f(g(x_n)) = (f \circ g)(x_n) \to (f \circ g)(a) = f(g(a)).$$

Da g stetig ist in a konvergiert die Folge $y_n := g(x_n) \to g(a)$. Da f stetig ist in $y := g(a)$ konvergiert damit aber auch $f(y_n) \to f(y)$, also $f(g(x_n)) \to f(g(a))$. □

Aus den obigen Sätzen folgt dass jedes Polynom (als endliche Summe von skalaren Vielfachen von Potenzen) und jeder Quotient aus Polynomen (definiert außerhalb der Nullstellen des Nenners) stetige Funktionen sind. Die Stetigkeit von durch (unendliche) Potenzreihen definierte Funktionen lässt sich mit dem *Abelschen Grenzwertsatz* untersuchen. Wir führen dies in dieser Vorlesung aber nicht aus und geben nur für wichtige Funktionen die Stetigkeit ohne Beweis an.

Satz 2.9
Die Funktionen $\exp: \mathbb{R} \to \mathbb{R}$*,* $\sin: \mathbb{R} \to \mathbb{R}$ *und* $\cos: \mathbb{R} \to \mathbb{R}$ *sind stetig.*

Ebenfalls ohne Beweis geben wir noch eine äquivalente Charakterisierung der Stetigkeit an und definieren zwei stärkere Stetigkeitsbegriffe.

Definition und Satz 2.10
(a) f *ist genau dann stetig in* $a \in D$*, falls*

$$\forall \varepsilon > 0 \quad \exists \delta > 0: \quad |f(x) - f(a)| < \varepsilon \quad \forall x \in D \text{ mit } |x - a| < \delta.$$

Diese Eigenschaft heißt auch ε-δ*-Definition der Stetigkeit, vgl. das in der Vorlesung gemalte Bild.*

(b) f heißt gleichmäßig stetig in D, falls

$$\forall \varepsilon > 0 \quad \exists \delta > 0 : \quad |f(x) - f(y)| < \varepsilon \quad \forall x, y \in D \text{ mit } |x - y| < \delta.$$

(c) f heißt Lipschitz-stetig in D, falls

$$\exists L > 0 : \quad |f(x) - f(y)| \leq L|x - y| \quad \forall x, y \in D.$$

Es gilt

$$f \text{ Lipschitz stetig in } D \implies f \text{ glm. stetig in } D \implies f \text{ stetig in } D.$$

2.3 Maxima, Minima und der Zwischenwertsatz

Unsere Intuition, dass stetige Funktionen nicht springen, können wir nun mathematisch präzise formulieren.

Satz 2.11 (Zwischenwertsatz)
Sei $f : [a,b] \to \mathbb{R}$ eine stetige Funktion. Sei $y \in \mathbb{R}$ ein Wert zwischen $f(a)$ und $f(b)$, also

$$f(a) \leq y \leq f(b) \quad \text{oder} \quad f(a) \geq y \geq f(b).$$

Dann existiert (mindestens) ein $\hat{x} \in [a,b]$ mit $f(x) = y$.

Beweis: Für $f(a) = y$ oder $f(b) = y$ ist die Aussage trivial. Wir betrachten also den Fall $f(a) < y < f(b)$. Dazu verwenden wir wieder wie in Abschnitt 2.1 beschrieben die Idee der Intervallhalbierung und erhalten zwei Folgen

$$a = a_0 \leq a_1 \leq \ldots \leq a_n \leq \ldots \leq b_n \leq \ldots \leq b_1 \leq b_0 = b,$$

die (die eine monoton wachsend und die andere monoton fallend) beide gegen den selben Grenzwert konvergieren

$$\hat{x} := \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

Wegen Satz 1.11 gilt

$$a \leq \hat{x} \leq b.$$

Außerdem erfüllen die Folgen nach Konstruktion

$$f(a_n) < y \leq f(b_n) \quad \text{für alle } n \in \mathbb{N}_0.$$

Wegen der Stetigkeit von f und wiederum Satz 1.11 wissen wir nun aber auch

$$y \geq \lim_{n\to\infty} f(a_n) = f(\lim_{n\to\infty} a_n) = f(\hat{x}) = f(\lim_{n\to\infty} b_n) = \lim_{n\to\infty} f(b_n) \geq y.$$

Es muss also gelten $f(\hat{x}) = y$. □

Definition 2.12
Eine Funktion $f: D \to \mathbb{R}$ heißt nach oben beschränkt, falls ihre Wertemenge nach oben beschränkt ist, also wenn ein $C > 0$ existiert mit

$$f(x) \leq C \quad \forall x \in D.$$

Analog definieren wir die Begriffe nach unten beschränkt und beschränkt.

Wir sagen, dass die Funktion f auf D ihr Maximum annimmt, falls ein $\hat{x} \in D$ existiert mit $f(\hat{x}) = \max\{f(x): x \in D\}$, also

$$f(x) \leq f(\hat{x}) \quad \forall x \in D.$$

Analog definieren wir, dass die Funktion ihr Minimum annimmt.

Satz 2.13
Eine stetige Funtkion $f: [a,b] \to \mathbb{R}$ auf einem abgeschlossenen Intervall $[a,b]$ nimmt sowohl ihr Maximum als auch ihr Minimum an, d.h.

$$\exists x_{max} \in [a,b]: \quad f(x_{max}) = \max_{x\in[a,b]} f(x) = \sup_{x\in[a,b]} f(x),$$
$$\exists x_{min} \in [a,b]: \quad f(x_{min}) = \min_{x\in[a,b]} f(x) = \inf_{x\in[a,b]} f(x).$$

Beweis: Ist f nach oben beschränkt, so existiert nach Definition und Satz 1.21 das Supremum

$$y_{\max} = \sup\{f(x): x \in [a,b]\} < \infty.$$

Gemäß der Definition des Supremums existiert für jedes $\varepsilon > 0$ ein $x \in [a,b]$ mit $f(x) > y_{\max} - \varepsilon$. Wir verwenden dies mit der Wahl $\varepsilon := \frac{1}{n}$ und erhalten so für alle $n \in \mathbb{N}$ ein $x_n \in [a,b]$ mit $f(x_n) > y_{\max} - \frac{1}{n}$ und damit eine *maximierende Folge* $(x_n)_{n\in\mathbb{N}} \subset [a,b]$ für die gilt $f(x_n) \to y_{\max}$. Ist f nach oben unbeschränkt, so existiert entsprechend eine maximierende Folge mit $f(x_n) \to \infty$.

Mit dem Satz von Bolzano-Weierstraß (Satz 1.22) folgt, dass sich aus $(x_n)_{n\in\mathbb{N}}$ eine konvergente Teilfolge $(x_{n_k})_{k\in\mathbb{N}}$ extrahieren lässt. Sei $x_{\max} := \lim_{k\to\infty} x_{n_k}$. Für die Teilfolge gilt noch immer $f(x_{n_k}) \to y_{\max}$ (bzw. $f(x_{n_k}) \to \infty$) und mit der Stetigkeit von f folgt dann, dass der Fall $f(x_{n_k}) \to \infty$ nicht möglich ist und dass gilt

$$f(x_{\max}) = f(\lim_{k\to\infty} x_{n_k}) = \lim_{k\to\infty} f(x_{n_k}) = y_{\max}.$$

Die Funktion f nimmt also ihr Maximum an. Die Annahme des Minimums folgt durch Anwendung des bereits gezeigten auf die Funktion $-f$. □

Wir geben ohne Beweis noch zwei nützliche Eigenschaften stetiger Funktionen auf abgeschlossenen Intervallen an. Diese sind automatisch auch gleichmäßig stetig und können nicht beliebig stark oszillieren können, d.h. auf kleinen Teilintervallen liegen Maxima und Minima nah beeinander.

Satz 2.14
Eine stetige Funktion $f:\ [a,b]\to\mathbb{R}$ auf einem abgeschlossenen Intervall $[a,b]$ ist auch gleichmäßig stetig. Für jedes $\varepsilon>0$ existiert eine Intervallbreite $\delta>0$, so dass für jedes Teilintervall $[c,d]\subseteq[a,b]$ mit $d-c\le\delta$ gilt

$$\max_{x\in[c,d]} f(x) - \min_{x\in[c,d]} f(x) < \varepsilon.$$

2.4 Die Umkehrfunktion

Wir erinnern an den Begriff der Umkehrfunktion. Eine Funktion $f:\ X\to Y$ zwischen zwei Mengen X und Y heißt

- *injektiv*, falls zwei verschiedenen Elemente $x,x'\in X$ nicht das gleiche y zugeordnet wird, d.h.

$$\forall x,x'\in X:\ f(x)=f(x') \quad\Longrightarrow\quad x=x'.$$

- *surjektiv*, falls jedes Element von Y erreicht wird, d.h. in Formelsprache:

$$\forall y\in Y:\ \exists x\in X:\ f(x)=y.$$

- *bijektiv*, falls f injektiv und surjektiv ist, d.h. in Formelsprache:

$$\forall y\in Y:\ \exists! x\in X:\ f(x)=y.$$

Für bijektive Funktionen können wir die *Umkehrfunktion* f^{-1} definieren durch

$$f^{-1}:\ Y\to X,\quad f^{-1}(y):=x,\quad \text{wobei } x\in X \text{ erfüllt } f(x)=y.$$

Es gilt

$$f^{-1}(f(x))=x\quad\forall x\in X \qquad\text{und}\qquad f(f^{-1}(y))=y\quad\forall y\in Y.$$

Definition 2.15
Eine Funktion $f:\ D\to\mathbb{R}$ heißt

(a) monoton wachsend, falls

$$\forall x,y \in D: \quad x < y \quad \Longrightarrow \quad f(x) \leq f(y)$$

(b) streng monoton wachsend, falls

$$\forall x,y \in D: \quad x < y \quad \Longrightarrow \quad f(x) < f(y)$$

Sie heißt monoton fallend (bzw. streng monoton fallend), falls $-f$ monoton wachsend (bzw. streng monoton wachsend) ist.

Satz 2.16
Sei $f: [a,b] \to \mathbb{R}$ eine stetige, streng monoton wachsende Funktion. Dann ist f eine bijektive Abbildung von $[a,b]$ nach $[f(a),f(b)]$. Die Umkehrfunktion

$$f^{-1}: [f(a),f(b)] \to [a,b]$$

ist ebenfalls stetig und streng monoton wachsend.

Für eine stetige, streng monoton fallende Funktion existiert die Umkehrfunktion

$$f^{-1}: [f(b),f(a)] \to [a,b]$$

und sie ist stetig und streng monoton fallend.

Beweis: Sei $f: [a,b] \to \mathbb{R}$ stetig und streng monoton wachsend. Der Zwischenwertsatz 2.11 zeigt, dass f jeden Wert zwischen $f(a)$ und $f(b)$ erreicht, d.h. $f: [a,b] \to [f(a),f(b)]$ ist surjektiv. Wegen der strengen Monotonie ist f zusätzlich auch injektiv, also existiert die Umkehrfunktion

$$f^{-1}: [f(a),f(b)] \to [a,b].$$

Um die Monotonie von f^{-1} zu zeigen, seien $y_1,y_2 \in [f(a),f(b)]$ mit $y_1 < y_2$. Wir zeigen $f^{-1}(y_1) < f^{-1}(y_2)$ mit Widerspruchsbeweis und nehmen an, dass $f^{-1}(y_1) = f^{-1}(y_2)$ oder $f^{-1}(y_1) > f^{-1}(y_2)$. Im ersten Fall wäre dann aber wegen der Bijektivität auch $y_1 = y_2$ und im zweiten Fall wegen der Monotonie von f

$$y_2 = f(f^{-1}(y_2)) < f(f^{-1}(y_1)) = y_1.$$

Beides widerspricht $y_1 < y_2$, so dass $f^{-1}(y_1) < f^{-1}(y_2)$ gezeigt ist.

Zum Beweis der Stetigkeit von f^{-1} sei $(y_n)_{n\in\mathbb{N}}$ eine Folge in $[f(a),f(b)]$ mit $y_n \to y \in [f(a),f(b)]$. Falls der Grenzwert $x := \lim_{n\to\infty} f^{-1}(y_n)$ existiert, dann folgt aus der Stetigkeit von f

$$f(x) = f(\lim_{n\to\infty} f^{-1}(y_n)) = \lim_{n\to\infty} f(f^{-1}(y_n)) = \lim_{n\to\infty} y_n = y$$

und damit $x = f^{-1}(f(x)) = f^{-1}(y)$. Genauso folgt, dass auch jede konvergente Teilfolge von $(f^{-1}(y_n))_{n\in\mathbb{N}}$ gegen denselben Grenzwert $f^{-1}(y)$ konvergieren muss. Da die Folge $(f^{-1}(y_n))_{n\in\mathbb{N}} \subset [a,b]$ beschränkt ist, folgt mit Satz 1.23, dass

$$\lim_{n\to\infty} f^{-1}(y_n) = f^{-1}(y),$$

womit die Stetigkeit gezeigt ist. □

Beispiele 2.17

(a) Die konstante Funktion $f:\ \mathbb{R}\to\mathbb{R},\ f(x) := 1$ ist monoton wachsend und monoton fallend, aber weder streng monoton wachsend noch streng monoton fallend.

(b) Die Funktion $f:\mathbb{R}\to\mathbb{R},\ f(x) := x$ ist streng monoton wachsend und besitzt die Umkehrfunktion $f^{-1}:\ \mathbb{R}\to\mathbb{R},\ f^{-1}(x) := x$.

(c) Die Funktion $f:\ \mathbb{R}\to\mathbb{R},\ f(x) := x^2$ ist auf jedem Intervall $[a,b]$ mit $b > a \geq 0$ streng monoton wachsend und stetig. Für jedes solche Intervall existiert daher die Umkehrfunktion $f^{-1}:\ [a^2,b^2]\to[a,b],\ f^{-1}(y) := \sqrt{y}$ Mit $a := 0$ und da b beliebig groß gewählt werden kann, kann die Umkehrfunktion definiert werden auf

$$f^{-1}:\ [0,\infty[\to[0,\infty[,\quad f^{-1}(y) = \sqrt{y}$$

und f^{-1} ist stetig und monoton wachsend.

Definition und Satz 2.18 (Logarithmus und allgemeine Potenzen)

(a) Die Exponentialfunktion $\exp:\ \mathbb{R}\to\mathbb{R}$ ist stetig und streng monoton wachsend und besitzt daher für jedes Intervall $[a,b]\subset\mathbb{R}$ eine Umkehrfunktion von $[\exp(a),\exp(b)]$ nach $[a,b]$. Es gilt

$$\lim_{x\to-\infty}\exp(x) = 0 \quad \textit{und} \quad \lim_{x\to\infty}\exp(x) = \infty,$$

so dass die Umkehrfunktion auf $]0,\infty[$ definiert werden kann. Wir nennen sie den (natürlichen) Logarithmus und schreiben

$$\ln:\ ]0,\infty[\to\mathbb{R},\quad x\mapsto\ln(x).$$

Die Funktion $\ln$ ist stetig und streng monoton wachsend.

(b) Für $a > 0$ definieren wir die Exponentialfunktion zur Basis a durch

$$\exp_a:\ \mathbb{R}\to\mathbb{R},\quad \exp_a(x) := \exp(x\ln(a)).$$

$\exp_a$ *ist für* $0 < a < 1$ *streng monoton fallend, für* $a = 1$ *konstant* 1 *und für* $a > 1$ *streng monoton wachsend. Sie ist stetig und erfüllt*

$$\exp_a(p/q) = \sqrt[q]{a^p} = a^{p/q} \quad \text{für alle } p \in \mathbb{Z},\ q \in \mathbb{N}.$$

Wir definieren damit die nicht-rationalen Potenzen

$$a^x := \exp_a(x) \quad \text{für alle } x \in \mathbb{R} \setminus \mathbb{Q}$$

und schreiben im Folgenden immer a^x *statt* $\exp_a$. *Insbesondere gilt also auch* $\exp(x) = e^x$.

(c) Für $a > 0$, $a \neq 1$ *definieren wir die Logarithmusfunktion zur Basis a durch*

$$\log_a :]0,\infty[\to \mathbb{R}, \quad \log_a(x) := \frac{\ln(x)}{\ln(a)}.$$

Sie ist die Umkehrfunktion zu $\exp_a$ *und damit stetig und streng monoton fallend (für* $0 < a < 1$*) bzw. steigend (für* $a > 1$*). Es gilt* $\ln(x) = \log_e(x)$.

(d) Es gelten die Potenzgesetze (für alle $a > 0$ *und* $x, y \in \mathbb{R}$*)*

$$a^x a^y = a^{x+y} \quad \text{und} \quad (a^x)^y = a^{xy}$$

und die Logarithmengesetze (für alle $a, x, y > 0$, $a \neq 1$*)*

$$\log_a(x * y) = \log_a(x) + \log_a(y) \quad \text{und} \quad \log_a(x^y) = y \log_a(x).$$

Beweis: Die Stetigkeit der Exponentialfunktion haben wir in Satz 2.9 festgestellt. Die Monotonie folgt für $x > 0$ aus der Reihendarstellung in Satz 1.30 und für $x < 0$ aus $\exp(-x) = \frac{1}{\exp(x)}$. Ebenfalls aus der Reihendarstellung erhalten wir

$$\mathrm{e} = \exp(1) = \sum_{k=0}^{\infty} \frac{1}{k!} \geq \sum_{k=0}^{1} \frac{1}{k!} = 2$$

und (zusammen mit der Funktionalgleichung)

$$\exp(n) = \exp(1 + \ldots + 1) = \exp(1) \cdot \ldots \cdot \exp(1) \geq 2^n \to \infty \quad \text{für } n \to \infty,$$

also auch

$$\lim_{x\to\infty} \exp(x) \to \infty \quad \text{und} \quad \lim_{x\to-\infty} \exp(x) = \lim_{x\to\infty} \exp(-x) = \lim_{x\to\infty} \frac{1}{\exp(x)} = 0.$$

Die Eigenschaften der Umkehrfunktionen folgen aus Satz 2.16 und der Zusammenhang zu rationalen Potenzen lässt sich wie in Folgerung 1.31 zeigen. Das erste Potenzgesetz $a^x a^y = a^{x+y}$ folgt aus der Funktionalgleichung für die Exponentialfunktion. Das zweite Potenzgesetz folgt für rationale $y \in \mathbb{Q}$ aus dem ersten und durch Grenzwertbildung und Stetigkeit gilt es dann auch für alle $y \in \mathbb{R}$. Die Logarithmengesetze folgen aus den Potenzgesetzen. □

2.5 Asymptotisches Verhalten von Funktionen

Die in Abschnitt 1.7 für Folgen eingeführte Landau-Notation verwenden wir auch für Funktionen.

Definition 2.19 (Landau-Notation für Funktionen)
Seien $f, g: D \to \mathbb{R}$ und $a \in \overline{D}$. Wir schreiben

$$f \in o(g), \quad \text{falls} \quad \frac{f(x)}{g(x)} \to 0 \quad \text{für } x \to a,$$

und wir schreiben $f \in O(g)$, falls f durch g in einer Umgebung von a beschränkt ist, d.h.

$$\exists C > 0, \varepsilon > 0: \ |f(x)| \leq C|g(x)| \quad \text{für alle } x \in]a-\varepsilon, a+\varepsilon[\cap D.$$

Für eine nach oben unbeschränkte Definitionsmenge D verwenden wir die Notation auch für $x \to \infty$, also

$$f \in o(g), \quad \text{falls} \quad \frac{f(x)}{g(x)} \to 0 \quad \text{für } x \to \infty,$$

und $f \in O(g)$, falls f durch g für hinreichend große x beschränkt ist, d.h.

$$\exists C > 0, \ K > 0: \ |f(x)| \leq C|g(x)| \quad \text{für alle } x \in D, \ x \geq K.$$

Analog definieren wir die Notation für $x \to -\infty$.

Mit dieser Notation gilt offenbar, dass $f: D \to \mathbb{R}$ genau dann in $a \in D$ stetig ist, wenn

$$f(x) \in f(a) + o(1).$$

Wir können das so interpretieren, dass $f(a)$ in der Nähe von $x = a$ eine gute Approximation an $f(x)$ ist, in dem Sinne, dass der Abstand von $f(a)$ zu $f(x)$ beliebig klein ist, wenn nur x nah genug an a liegt. Die konstante Funktion $x \mapsto f(a)$ ist in diesem Sinne die bestapproximierende konstante Funktion, vgl. die in der Vorlesung gemalte Skizze. Bessere Approximationen erhalten wir mit den im nächsten Abschnitt eingeführten Ableitungen von f.

Kapitel 3

Differenzierbarkeit

3.1 Der Ableitungsbegriff und wichtige Regeln

Definition 3.1
Eine Funktion $f:\ [a,b] \to \mathbb{R}$ heißt in $x \in]a,b[$ differenzierbar, falls der Grenzwert der (für betragsmäßig hinreichend kleine $h \neq 0$ definierten) Funktion

$$h \mapsto \frac{f(x+h)-f(x)}{h}$$

für $h \to 0$ existiert. In dem Fall nennen wir

$$f'(x) := \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$$

die Ableitung von f in $x \in]a,b[$. Für Randpunkte $x = a$ oder $x = b$ definieren wir entsprechend die Differenzierbarkeit durch die einseitigen Grenzwerte

$$f'(a) := \lim_{h\to 0^+} \frac{f(a+h)-f(a)}{h} \quad \text{und} \quad f'(b) := \lim_{h\to 0^-} \frac{f(b+h)-f(b)}{h}.$$

f heißt differenzierbar in $[a,b]$, wenn es in allen $x \in [a,b]$ differenzierbar ist. f heißt stetig differenzierbar in $[a,b]$, wenn es differenzierbar ist und die Funktion

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

stetig in $[a,b]$ ist. Wir verwenden die Begriffe auch analog für offene, halbabgeschlossene oder unbeschränkte Intervalle.

Bemerkung 3.2
In der Definition der Ableitung können wir auch äquivalent schreiben

$$f'(x) = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}$$

oder mit Landau-Notation

$$f(x+h) \in f(x) + f'(x)h + o(h) \quad \text{bzw.} \quad f(y) \in f(x) + f'(x)(y-x) + o(y-x).$$

Die Funktion $y \mapsto f(x) + f'(x)(y-x)$ ist die einzige lineare Funktion, die dies erfüllt und (in diesem Sinne) die bestapproximierende lineare Funktion, vgl. die in der Vorlesung gemalte Skizze.

Beispiele 3.3
(a) Für die Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x$ gilt für jedes $x \in \mathbb{R}$

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{x + h - x}{h} = 1.$$

f ist also stetig differenzierbar und $f'(x) = 1$.

(b) Für die Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ gilt für jedes $x \in \mathbb{R}$

$$\frac{f(x+h) - f(x)}{h} = \frac{(x+h)^2 - x^2}{h} = \frac{x^2 + 2hx + h^2 - x^2}{h} = 2x + h$$

und damit $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = 2x$.
f ist also stetig differenzierbar und $f'(x) = 2x$.

(c) Die Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$ ist für jedes $x \neq 0$ differenzierbar mit $f'(x) = 1$ für alle $x > 0$ und $f'(x) = -1$ für alle $x < 0$.
Für $x = 0$ konvergiert

$$\frac{f(x+h_n) - f(x)}{h_n} = \frac{|h_n|}{h_n}$$

für die Folge $h_n := \frac{1}{n}$ gegen 1 und für die Folge $h_n := -\frac{1}{n}$ gegen -1. Gemäß Definition 2.3 existiert also der Grenzwert der Funktion

$$\frac{f(x+h_n) - f(x)}{h_n} = \frac{|h_n|}{h_n}$$

für $h \to 0$ nicht. f ist daher in $x = 0$ ist f nicht differenzierbar.

Wir fassen die wichtigsten Ableitungsregeln (Summenregel, Produktregel, Quotientenregel, Kettenregel und Ableitung der Umkehrfunktion) in folgendem Satz zusammen:

Satz 3.4 (Ableitungsregeln)

(a) Ist $f:\ [a,b]\to\mathbb{R}$ in $x\in[a,b]$ differenzierbar, dann ist f auch stetig in x.

(b) Es seien $f,g:\ [a,b]\to\mathbb{R}$ in $x\in[a,b]$ differenzierbar und $c\in\mathbb{R}$. Dann gilt

(i) $f+g$ ist in x differenzierbar und $(f+g)'(x)=f'(x)+g'(x)$.

(ii) cf ist in x differenzierbar und $(cf)'(x)=cf'(x)$.

(iii) fg ist in x differentierbar und $(fg)'(x)=f'(x)g(x)+f(x)g'(x)$.

(iv) Ist zusätzlich $g(x)\neq 0$, dann ist f/g in x differenzierbar und

$$\left(\frac{f}{g}\right)(x)=\frac{f'(x)g(x)-f(x)g'(x)}{g^2(x)}.$$

(c) Es seien $f:\ [a,b]\to\mathbb{R}$ und $g:\ [c,d]\to\mathbb{R}$ mit $g([c,d])\subseteq[a,b]$. Ist g in $x\in[c,d]$ differenzierbar und f in $g(x)\in[a,b]$ differenzierbar, so ist $f\circ g$ in x differenzierbar und

$$(f\circ g)'(x)=f'(g(x))g'(x).$$

(d) $f:\ [a,b]\to\mathbb{R}$ besitze eine Umkehrfunktion $f^{-1}:\ f([a,b])\to\mathbb{R}$. Ist f in x differenzierbar und ist $f'(x)\neq 0$, dann ist f^{-1} in $y:=f(x)$ differenzierbar und es gilt

$$(f^{-1})'(y)=\frac{1}{f'(f^{-1}(y))}=\frac{1}{f'(x)}.$$

Beweis: (a) Ist f differenzierbar in x, dann gilt

$$\lim_{y\to x}f(y)=\lim_{y\to x}\left(f(x)+\frac{f(y)-f(x)}{y-x}(y-x)\right)=f(x)+f'(x)0=f(x),$$

also ist f auch stetig in x.

(b) Beweis der Summen-, Produkt- und Quotientenregel:

(i) folgt aus

$$\begin{aligned}&\lim_{h\to 0}\frac{(f+g)(x+h)-(f+g)(x)}{h}\\&=\lim_{h\to 0}\frac{f(x+h)+g(x+h)-f(x)-g(x)}{h}\\&=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}+\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}=f'(x)+g'(x).\end{aligned}$$

(ii) folgt genauso leicht.

(iii) folgt aus

$$\begin{aligned}
&\lim_{h\to 0}\frac{(fg)(x+h)-(fg)(x)}{h}=\lim_{h\to 0}\frac{f(x+h)g(x+h)-f(x)g(x)}{h}\\
&=\lim_{h\to 0}\frac{f(x+h)g(x+h)-f(x+h)g(x)+f(x+h)g(x)-f(x)g(x)}{h}\\
&=\lim_{h\to 0}\left(f(x+h)\frac{g(x+h)-g(x)}{h}+\frac{f(x+h)-f(x)}{h}g(x)\right)\\
&=\lim_{h\to 0}f(x+h)\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}+\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}g(x)\\
&=f(x)g'(x)+f'(x)g(x),
\end{aligned}$$

wobei wir die in (a) bewiesene Stetigkeit verwendet haben.

(iv) Wir betrachten zunächst den Spezialfall $f=1$ (also die Differenzierbarkeit von $1/g$):

$$\begin{aligned}
&\lim_{h\to 0}\frac{(1/g)(x+h)-(1/g)(x)}{h}=\lim_{h\to 0}\frac{1}{h}\left(\frac{1}{g(x+h)}-\frac{1}{g(x)}\right)\\
&=\lim_{h\to 0}\frac{g(x)-g(x+h)}{hg(x+h)g(x)}=-\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}\lim_{h\to 0}\frac{1}{g(x+h)g(x)}\\
&=-\frac{g'(x)}{g^2(x)},
\end{aligned}$$

wobei wir wieder die in (a) bewiesene Stetigkeit verwendet haben.

Mit der in (b)(iii) bewiesenen Produktregel folgt jetzt die Quotientenregel

$$\begin{aligned}
\left(\frac{f}{g}\right)'(x)&=\left(f\frac{1}{g}\right)'(x)=f'(x)\frac{1}{g(x)}+f(x)\left(\frac{1}{g}\right)'(x)\\
&=\frac{f'(x)}{g(x)}-\frac{f(x)g'(x)}{g^2(x)}=\frac{f'(x)g(x)-f(x)g'(x)}{g^2(x)}.
\end{aligned}$$

(c) Zum Beweis der Kettenregel definieren wir die Funktion

$$f^*:\ [a,b]\to\mathbb{R},\quad f^*(y)=\begin{cases}\frac{f(y)-f(g(x))}{y-g(x)} & \text{für } y\neq g(x),\\ f'(g(x)) & \text{für } y=g(x).\end{cases}$$

Dann gilt wegen der Differenzierbarkeit von f, dass

$$\lim_{y\to g(x)}f^*(y)=f'(g(x))\quad\text{und damit}\quad\lim_{h\to 0}f^*(g(x+h))=f'(g(x)).$$

Für alle $y \neq g(x)$ ist außerdem $f(y) - f(g(x)) = f^*(y)(y - g(x))$. Damit erhalten wir

$$\lim_{h\to 0} \frac{f(g(x+h)) - f(g(x))}{h} = \lim_{h\to 0} \frac{f^*(g(x+h))(g(x+h) - g(x))}{h}$$
$$= \lim_{h\to 0} f^*(g(x+h)) \lim_{h\to 0} \frac{g(x+h) - g(x)}{h} = f'(g(x))g'(x).$$

(d) Sei $(y_n)_{n\in\mathbb{N}} \subset f([a,b])$ eine Folge mit $y_n \to y$. Dann gilt

$$x_n := f^{-1}(y_n) \to f^{-1}(y) = x, \quad f(x_n) = y_n, \quad \text{und} \quad f(x) = y.$$

Damit erhalten wir

$$\frac{f^{-1}(y_n) - f^{-1}(y)}{y_n - y} = \frac{x_n - x}{f(x_n) - f(x)} = \frac{1}{\frac{f(x_n)-f(x)}{x_n - x}} \to \frac{1}{f'(x)},$$

und damit ist $(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}$. □

Definition und Satz 3.5 (Höhere Ableitungen)
Ist $f\colon [a,b] \to \mathbb{R}$ differenzierbar und ist auch $f'\colon [a,b] \to \mathbb{R}$ differenzierbar, dann heißt $f'' := (f')'$ die zweite Ableitung von f. Entsprechend definieren wir die n-te Ableitung rekursiv als $f^{(n)} = (f^{(n-1)})'$ mit $f^{(0)} := f$. Wir schreiben dafür auch

$$f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x) \quad \text{und} \quad f^{(n)}(x) = \frac{\mathrm{d}^n}{\mathrm{d}x^n} f(x).$$

Sind $f, g\colon [a,b] \to \mathbb{R}$ zwei n-mal differenzierbare Funktionen und $c \in \mathbb{R}$, so sind auch $f+g$, cf sowie fg jeweils n-mal differenzierbar. Es gilt

$$(f+g)^{(n)}(x) = f^{(n)}(x) + g^{(n)}(x)$$
$$(cf)^{(n)}(x) = cf^{(n)}(x)$$

und die Leibnizregel

$$(fg)^{(n)}(x) = \sum_{k=0}^{n} \binom{n}{k} f^{(k)}(x) g^{(n-k)}(x).$$

Die Menge aller n-mal stetig differenzierbaren Funktionen bezeichnen wir mit $C^n([a,b])$. Mit $C^0([a,b])$ (auch: $C([a,b])$) bezeichnen wir die Menge aller stetigen Funktionen und mit $C^\infty([a,b])$ die Menge aller beliebig oft differenzierbaren Funktionen. Aufgrund der gezeigten Regeln sind dies jeweils Vektorräume.

Beweis: Die Differenzierbarkeitsregeln für die n-te Ableitung folgen aus den jeweiligen Regeln für die erste Ableitung durch vollständige Induktion. □

Satz 3.6 (Ableitung einiger wichtiger Funktionen)

(a) *Das Monom* $f:\ \mathbb{R}\to\mathbb{R}$, $f(x):=x^n$ *ist (für alle* $n\in\mathbb{N}$*) eine* C^∞*-Funktion. Es gilt*

$$f'(x)=nx^{n-1},\ f''(x)=n(n-1)x^{n-2},\ \ldots,\ f^{(n)}(x)=n!$$

und

$$f^{(n+1)}(x)=f^{(n+2)}(x)=\ldots=0.$$

(b) $f:\ \mathbb{R}\setminus\{0\}\to\mathbb{R}$, $f(x)=\frac{1}{x^n}:=x^{-n}$ *ist (für alle* $n\in\mathbb{N}$*) eine* C^∞*-Funktion. Es gilt*

$$f'(x)=-nx^{-n-1}=\frac{-n}{x^{n+1}},\quad f''(x)=-n(-n-1)x^{-n-2}=\frac{n(n+1)}{x^{n+2}},\ldots$$

(c) *Die Funktion* $f:\ \mathbb{R}\to\mathbb{R}$, $f(x):=e^x$ *ist eine* C^∞*-Funktion. Es gilt*

$$f'(x)=f''(x)=\ldots=f^{(n)}(x)=e^x\quad \text{für alle } n\in\mathbb{N}.$$

(d) *Die Funktion* $f:\ ]0,\infty[\to\mathbb{R}$, $f(x):=\ln(x)$ *ist eine* C^∞*-Funktion. Es gilt*

$$f'(x)=\frac{1}{x},\quad f''(x)=-\frac{1}{x^2},\quad \ldots$$

(e) *Für* $a>0$ *ist die Funktion* $f:\ \mathbb{R}\to\mathbb{R}$, $f(x):=a^x$ *eine* C^∞*-Funktion. Es gilt*

$$f'(x)=\ln(a)a^x,\quad f''(x)=\ln(a)^2a^x,\quad \ldots$$

(f) *Für* $a>0$ *ist die Funktion* $f:\ ]0,\infty[\to\mathbb{R}$, $f(x):=\log_a(x)$ *eine* C^∞*-Funktion. Es gilt*

$$f'(x)=\frac{1}{\ln(a)x},\quad f''(x)=-\frac{1}{\ln(a)^2x^2},\quad \ldots$$

(g) *Die Funktion* $f:\ ]0,\infty[\to\mathbb{R}$, $f(x):=x^a$ *ist auch für* $a\in\mathbb{R}\setminus\mathbb{Z}$ *eine* C^∞*-Funktion und es gilt*

$$f'(x)=ax^{a-1},\quad f''(x)=a(a-1)x^{a-2},\ldots$$

(h) *Die Funktionen* $\sin,\cos:\ \mathbb{R}\to\mathbb{R}$ *sind* C^∞*-Funktionen. Es gilt*

$$\frac{\mathrm{d}}{\mathrm{d}x}\sin(x)=\cos(x)\quad \text{und}\quad \frac{\mathrm{d}}{\mathrm{d}x}\cos(x)=-\sin(x).$$

Beweis: (a) und (b) folgen mit vollständiger Induktion aus der Produkt- und Quotientenregel. Zum Beweis von (c) und (h) kann man zeigen, dass die Exponentialreihe und die Sinus- und Kosinusreihen jeweils summandenweise differenziert werden können. (d) folgt aus der Ableitungsregel für die Umkehrfunktion. (e), (f) und (g) folgen mit aus der Kettenregel. □

3.2 Mittelwertsatz und Taylorentwicklung

Das Maximum und Minimum einer Funktion haben wir in Definition 2.12 eingeführt. Wir nennen es im Folgenden auch das *globale Maximum* bzw. *globale Minimum* und führen die lokalen Varianten ein, vgl. die in der Vorlesung gemalten Skizzen

Definition 3.7 (Lokales Maximum und Minimum)
f besitzt in $\hat{x} \in]a,b[$ ein lokales Maximum, falls ein $\varepsilon > 0$ existiert mit

$$f(x) \leq f(\hat{x}) \quad \text{für alle } x \in (\hat{x}-\varepsilon, \hat{x}+\varepsilon).$$

Entsprechend besitzt f in $\hat{x} \in]a,b[$ ein lokales Minimum, falls ein $\varepsilon > 0$ existiert mit

$$f(x) \geq f(\hat{x}) \quad \text{für alle } x \in (\hat{x}-\varepsilon, \hat{x}+\varepsilon).$$

Satz 3.8
Besitzt f in $\hat{x} \in]a,b[$ ein lokales Maximum oder Minimum, und ist f differenzierbar in $\hat{x}$, dann gilt

$$f'(\hat{x}) = 0.$$

Beweis: Im Falle eines lokalen Maximums erhalten wir aus der Definition der Ableitung mit der Folge $h_n = \frac{1}{n}$

$$f'(\hat{x}) = \lim_{n\to\infty} \frac{f(\hat{x}+\frac{1}{n}) - f(\hat{x})}{\frac{1}{n}} \leq 0$$

und mit der Folge $h_n = -\frac{1}{n}$

$$f'(\hat{x}) = \lim_{n\to\infty} \frac{f(\hat{x}-\frac{1}{n}) - f(\hat{x})}{-\frac{1}{n}} \geq 0,$$

also insgesamt $f'(\hat{x}) = 0$. Genauso folgt die Aussage für lokale Minima. □

Wir verwenden Satz 3.8 nun zum Beweis der anschaulichen Tatsache, dass zu jeder Sekante zwischen zwei Punkten auf der Funktion f ein Zwischenwert mit paralleler Tangentensteigung liegen muss, vgl. die in der Vorlesung gemalte Skizze.

Satz 3.9 (Mittelwertsatz der Differentialrechnung)
Sei $f: [a,b] \to \mathbb{R}$ eine differenzierbare Funktion. Dann existiert ein $\bar{x} \in]a,b[$ mit

$$\frac{f(b)-f(a)}{b-a} = f'(\bar{x}).$$

Insbesondere gilt

$$\begin{aligned} f'(x) \geq 0 \quad \forall x \in]a,b[\quad &\Longrightarrow \quad f(x) \geq f(a) \quad \forall x \in [a,b], \\ f'(x) \leq 0 \quad \forall x \in]a,b[\quad &\Longrightarrow \quad f(x) \leq f(a) \quad \forall x \in [a,b], \\ f'(x) = 0 \quad \forall x \in]a,b[\quad &\Longrightarrow \quad f(x) = f(a) \quad \forall x \in [a,b]. \end{aligned}$$

Beweis: Wir beweisen die Aussage zunächst für $f(a) = 0 = f(b)$ (*Satz von Rolle*). Da f stetig auf $[a,b]$ ist, nimmt f nach Satz 2.13 sein Maximum und Minimum auf $[a,b]$ and. Ist $f = 0$, dann ist die Aussage für jedes $\bar{x} \in (a,b)$ erfüllt. Ansonsten müssen entweder Maximum oder Minimum in $]a,b[$ liegen, dann bezeichnen wir dieses mit $\bar{x} \in]a,b[$ und erhalten aus Satz 3.8

$$f'(\bar{x}) = 0 = \frac{f(b) - f(a)}{b-a}.$$

Im allgemeinen Fall (ohne $f(a) = 0 = f(b)$ vorauszusetzen) definieren wir die Funktion

$$g : \ [a,b] \to \mathbb{R}, \quad g(x) := f(x) - f(a) - \frac{f(b) - f(a)}{b-a}(x-a).$$

Diese ist offenbar differenzierbar in $[a,b]$ und erfüllt $g(a) = 0 = g(b)$. Mit dem schon gezeigten Satz von Rolle existiert also ein $\bar{x} \in]a,b[$ mit

$$\frac{g(b) - g(a)}{b-a} = 0 = g'(\bar{x}) = f'(\bar{x}) - \frac{f(b) - f(a)}{b-a}$$

und damit gilt $f'(\bar{x}) = \frac{f(b)-f(a)}{b-a}$. □

Wir haben bereits gesehen, dass wir die Stetigkeits- und Diffenzbarkeitseigenschaften im Sinne bestapproximierender konstanter, bzw. linearer Funktionen interpretieren können:

$$f(y) = f(x) + o(1) \quad \text{bzw.} \quad f(y) = f(x) + f'(x)(y-x) + o(|y-x|).$$

Mit dem Mittelwertsatz können wir diese Abschätzungen noch verbessern. Für stetig differenzierbare Funktionen erhalten wir

$$f(y) = f(x) + f'(\bar{x})(y-x) = f(x) + O(|y-x|),$$

mit einer Zwischenstelle $\bar{x}$ zwischen x und y, wobei wir für die O-Schreibweise ausgenutzt haben, dass die stetige Ableitung ihr Betragsmaximum $[x,y]$ bzw. $[y,x]$ annimmt.

Für zweimal stetig differenzierbare Funktionen können wir mit einer Erweiterung des Mittelwertsatzes zeigen, dass ein $\bar{x}$ zwischen x und y existiert mit

$$f(y) = f(x) + f'(x)(y-x) + \frac{f''(\bar{x})}{2}(y-x)^2 = f(x) + f'(x)(y-x) + O(|y-x|^2).$$

Das Argument kann für immere höhere Ableitungen verwendet werden und so f immer besser (vgl. aber Bemerkung 3.12) durch ein Polynom approximiert werden. Wir geben die finale Formel ohne detaillierten Beweis an:

Definition und Satz 3.10 (Satz von Taylor)
Sei $f \in C^{n+1}([a,b])$ und $x, y \in [a,b]$. Dann existiert $\bar{x}$ zwischen x und y so dass

$$f(y) = \sum_{k=0}^{n} \frac{f^{(k)}}{k!}(x)(y-x)^k + \frac{f^{(n+1)}(\bar{x})}{(n+1)!}(y-x)^{n+1}.$$

Das Polynom

$$T_n(y) := \sum_{k=0}^{n} \frac{f^{(k)}}{k!}(x)(y-x)^k$$

heißt Taylorpolynom n-ter Ordnung im Entwicklungspunkt x. Für $f \in C^\infty([a,b])$ heißt die unendliche Summe $\sum_{k=0}^{\infty} \frac{f^{(k)}}{k!}(x)(y-x)^k$ Taylorreihe im Entwicklungspunkt x.

Der finale Summand $\frac{f^{(n+1)}(\bar{x})}{(n+1)!}(y-x)^{n+1}$ heißt auch Lagrange-Restglied.

Die Auswertung der $n+1$-ten Ableitung an der Zwischenstelle $\hat{x}$ im Restglied wird in in Anwendungen der Taylorformel oft durch das Maximum abgeschätzt. Hierzu schreiben wir für eine stetige Funktion $f: [a,b] \to \mathbb{R}$ das Betragsmaximum als

$$\|f\|_{[a,b]} := \max_{x \in [a,b]} |f(x)|$$

und fassen die wichtigen Taylorentwicklungen nullter, erster und zweiter Ordnung sowie ihre Fehler noch einmal zusammen.

Folgerung 3.11
(a) Für $f \in C^1([a,b])$ gilt für alle $x, y \in [a,b]$ die Taylorentwicklung 0. Ordnung

$$f(y) = f(x) + O(|y-x|)$$

und für den Fehler gilt

$$|f(y) - f(x)| \leq |y-x| \left\|f'\right\|_{[a,b]}.$$

(b) Für $f \in C^2([a,b])$ gilt für alle $x,y \in [a,b]$ die Taylorentwicklung 1. Ordnung

$$f(y) = f(x) + f'(x)(y-x) + O(|y-x|^2)$$

und für den Fehler gilt

$$|f(y) - f(x) - f'(x)(y-x)| \leq \frac{1}{2}|y-x|^2 \left\|f''\right\|_{[a,b]}.$$

(c) Für $f \in C^3([a,b])$ gilt für alle $x,y \in [a,b]$ die Taylorentwicklung 2. Ordnung

$$f(y) = f(x) + f'(x)(y-x) + \frac{1}{2}f''(x)(y-x)^2 + O(|y-x|^3)$$

und für den Fehler gilt

$$|f(y) - f(x) - f'(x)(y-x) - \frac{1}{2}f''(x)(y-x)^2| \leq \frac{1}{6}|y-x|^3 \left\|f'''\right\|_{[a,b]}.$$

Bemerkung 3.12
Aus dem Satz von Taylor folgt für $f \in C^\infty([a,b])$, dass

$$f(y) = T_n(y) + O(|y-x|^{n+1}).$$

Durch Verwendung des Funktionswertes und der ersten n-Ableitungen von f im Punkt x können wir daher ein Polynom aufstellen, dass sich für größere n immer besser an die Funktion anschmiegt, vgl. die in der Vorlesung gemalte Skizzze. Dabei stimmen Funktionswert und die Ableitungen des Polynomes T_n im Entwicklungspunkt x mit denen der Funktion f überein. Das Polynom approximiert die Funktion in dem Sinne immer besser, dass bei größerem n der Fehler zwischen f und Polynom bei Annäherung an den Punkt x immer schneller gegen Null fällt.

Hieraus folgt aber im Allgemeinen auch für C^∞-Funktionen nicht, dass im Grenzwert $n \to \infty$ die Polynome außerhalb von x gegen die Funktion f konvergieren. Funktionen, die mit ihrer (unendlichen) Taylorreihe übereinstimmen heißen analytisch, aber nicht jede C^∞-Funktion ist analytisch.

3.3 Lipschitz-stetigkeit und Fehlerfortpflanzung

Sei $f \in C^1([a,b])$. Aus der Taylorentwicklung 0. Ordnung in Folgerung 3.11(a) erhalten wir, dass f Lipschitz-stetig ist,

$$|f(y) - f(x)| \leq L|y-x| \quad \text{für alle } x,y \in [a,b],$$

mit Lipschitz-Konstante $L := \|f'\|_{[a,b]}$.

Die Lipschitz-Konstante können wir auch als Maß für die Fehlerverstärkung durch f interpretieren. Möchten wir $f(x)$ berechnen, so kennen wir aufgrund von Rundungsfehlern oder vorhergehenden Rechenfehlern typischerweise x nicht exakt, sondern $x+\delta$ mit einem betragskleinen δ. Selbst bei exakter Ausführung von f erhalten wir daher nicht $f(x)$ sondern $f(x+\delta)$ und es gilt

$$|f(x+\delta)-f(x)| \leq L|\delta|$$

Der Fehler δ im Argument $x \in [a,b]$ wird also schlimmstenfalls um den Faktor $L = \|f'\|_{[a,b]}$ verstärkt. Dieses Maß der Fehlerverstärkung heißt auch *absolute Kondition* auf $[a,b]$. Statt der schlimmstmöglichen Abschätzung auf dem ganzen Intervall $[a,b]$ betrachtet man oft auch nur eine kleine Umgebung eines festen x und definiert

$$\kappa_{\text{abs}} := |f'(x)|.$$

In der üblicherweise verwendeten Gleitkommadarstellung (siehe Abschnitt 1.8) ist der Rundungsfehler kein absoluter Wert sondern abhängig von der Größe der dargestellten Zahl. Betrachten wir daher stattdessen die *relative Kondition*, also die Fehlerverstärkung relativ zur Größe von x bzw. $f(x)$ so erhalten wir entsprechend

$$\frac{|f(x+\delta)-f(x)|}{|f(x)|} \leq \frac{L|\delta|}{|f(x)|} = \frac{L|x|}{|f(x)|}\frac{|\delta|}{|x|}.$$

Auf $[a,b]$ ist die relative Fehlerverstärkung also höchstens $L\sup_{x\in[a,b]}\frac{|x|}{|f(x)|}$ und bei Betrachtung einer kleinen Umgebung eines festen x erhalten wir

$$\kappa_{\text{rel}} := \frac{|f'(x)|\,|x|}{|f(x)|}.$$

Beispiel 3.13
Wir betrachten noch einmal die näherungsweise Berechnung der Eulerschen Zahl e durch die Folge $(1+\frac{1}{n})^n$. In Beispiel 1.36 haben wir gesehen, dass durch Rundungsfehler ab ca. $n = 10^{16}$ die Zahl $(1+\frac{1}{n})$ im double precision format auf 1 gerundet wird und sich dann unbrauchbare Näherungen $1^n = 1$ ergeben.

Schon vorher tritt aber der Effekt der Fehlerverstärkung auf. Die Potenzbildung $f:\ x \mapsto x^n$ besitzt in $x = 1$ die absolute Kondition $\kappa_{abs} = |f'(1)| = n$. Ein Rundungsfehler $|\delta| \approx 10^{-16}$ führt also bei der Potenzbildung zu einem Fehler der Größenordnung $n\delta$. Tatsächlich beobachtet man für $n = 10^8$ einen Fehler von etwa $n|\delta| \approx 10^{-8}$, der dann mit größerem n weiter ansteigt.

3.4 Das eindimensionale Newton-Verfahren

Wir betrachten das Problem der Nullstellenbestimmung. Zu einer gegebenen reellen Funktion $f: \mathbb{R} \to \mathbb{R}$ suchen wir ein $x \in \mathbb{R}$ mit

$$f(x) = 0.$$

Offenbar lässt sich jede Gleichung $f(x) = y$ in diese Form bringen, z.B. ist die Gleichung $x^2 = 2$ äquivalent zu $x^2 - 2 = 0$. Mit dem Intervallhalbierungsverfahren haben wir bereits (ein auf der Stetigkeit von f basierendes) Verfahren zur Lösung solcher Probleme kennengelernt. Wir stellen jetzt noch ein weiteres (auf der Differenzierbarkeit basierendes) Verfahren vor, das schneller konvergiert und sich später auch auf mehrere Unbekannte erweitern lässt.

Wir starten dazu in einem Punkt $x_0 \in \mathbb{R}$ und ersetzen f durch sein lineares Taylor-Polynom (also durch die Tangente in x_0) entsprechend der Taylorentwicklung 1. Ordnung in Folgerung 3.11(b)

$$f(x_0) + f'(x_0)(x - x_0) \approx f(x) \stackrel{!}{=} 0.$$

Die Nullstelle dieser linearen Approximation ist (falls $f'(x_0) \neq 0$)

$$x_1 := x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Wir wiederholen diesen Schritt mit der linearen Approximation in x_1 und erhalten mit

$$x_2 := x_1 - \frac{f(x_1)}{f'(x_1)}, \quad \ldots \quad x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)},$$

eine Folge $(x_n)_{n\in\mathbb{N}}$ von der wir anschaulich erwarten, dass sie sich immer mehr einer Nullstelle von f annähert, vgl. die in der Vorlesung gemalte Skizze.

Beispiel 3.14
Sei $a > 0$. Die Berechnung von $\sqrt{a}$ ist äquivalent zur Lösung der Nullstellenaufgabe $f(x) := x^2 - a = 0$. Die Iterationsvorschrift des Newton-Verfahren hierzu lautet

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$$

und ist als Heron-Verfahren bekannt. Für die Berechnung von $\sqrt{2}$ beginnend mit

$x_0 := 1$ *erhalten wir*

$$\begin{aligned}
x_0 &= \mathbf{1}\\
x_1 &= \mathbf{1},5\\
x_2 &= \mathbf{1},\mathbf{41}666666666666666666666666667\\
x_3 &= \mathbf{1},\mathbf{4142}1568627450980392156862745 1\\
x_4 &= \mathbf{1},\mathbf{41421356237}46899106262955788901\\
x_5 &= \mathbf{1},\mathbf{4142135623730950488016896}235025
\end{aligned}$$

und beobachten, dass die Anzahl der korrekten Stellen (fett gedruckt) sich mit jedem Iterationsschritt etwa verdoppelt. Zum Vergleich: Beim Intervallhalbierungsverfahren halbiert sich der Fehler in jedem Schritt, so dass jeweils etwa 3-4 *Schritte für eine weitere korrekte Stelle benötigt werden.*

Satz 3.15
Die Funktion $f : \mathbb{R} \to \mathbb{R}$ *sei zweimal stetig differenzierbar und besitze eine Nullstelle* $\hat{x} \in \mathbb{R}$ *mit* $f'(\hat{x}) \neq 0$*. Dann gibt es ein* $\delta > 0$*, so dass für jeden Startwert* $x_0 \in]\hat{x}-\delta, \hat{x}+\delta[$ *das Newton-Verfahren*

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)}$$

durchführbar ist (d.h. es gilt immer $f'(x_n) \neq 0$*) und die damit konstruierte Folge* $(x_n)_{n\in\mathbb{N}}$ *erfüllt*

$$x_n \to \hat{x} \quad \text{und} \quad x_{n+1} - \hat{x} = O(|x_n - \hat{x}|^2).$$

Beweis: Wir skizzieren nur die wesentliche Beweisidee. Mit der Iterationsvorschrift des Newton-Verfahrens und $f(\hat{x}) = 0$ erhalten wir

$$\begin{aligned}
x_{n+1} - \hat{x} = x_n - \frac{f(x_n)}{f'(x_n)} - \hat{x} &= \frac{-f(x_n) - f'(x_n)(\hat{x} - x_n)}{f'(x_n)}\\
&= \frac{f(\hat{x}) - f(x_n) - f'(x_n)(\hat{x} - x_n)}{f'(x_n)}.
\end{aligned}$$

Im Zähler steht die Differenz zwischen der Funktion und ihrem linearen Taylorpolynom. Da $f \in C^2(\mathbb{R})$ verhält sich diese (nach Bemerkung 3.12) wie $O(|x_n - \hat{x}|^2)$.

Da f' stetig ist, folgt aus $f'(\hat{x}) \neq 0$ auch dass $f'(x)$ in einer Umgebung von $\hat{x}$ nicht Null wird und $\frac{1}{f'(x)}$ beschränkt bleibt. Damit ergibt sich

$$x_{n+1} - \hat{x} = O(|x_n - \hat{x}|^2)$$

und wenn die Iterierten nah genug an $\hat{x}$ liegen, dann folgt daraus auch $x_n \to \hat{x}$. $\square$

Die Eigenschaft $x_{n+1} - \hat{x} = O(|x_n - \hat{x}|^2)$ zeigt, dass sich der Fehler in jedem Schritt etwa quadriert, d.h. beispielsweise aus einem Fehler von 10^{-1} im nächsten Schritt ein Fehler von 10^{-2} wird, dann ein Fehler von 10^{-4}, dann 10^{-8}, etc. Dies entspricht dem beobachteten Verhalten der Verdoppelung der richtigen Stellen pro Iterationsschritt und heißt *quadratische Konvergenz*. Werden dagegen, wie beim Intervallhalbierungsverfahren pro richtiger Kommastelle eine feste Anzahl von Iterationsschritten gebraucht, so spricht man von *linearer Konvergenz*.

Ein Nachteil des Newton-Verfahrens ist, dass die Konvergenz nur für Startwerte in einer Umgebung der Nullstelle garantiert ist (*lokale Konvergenz*). Beim Heron-Verfahren kann man Konvergenz für jeden Startwert $x_0 > 0$ konvergiert, aber das folgende Beispiel zeigt, dass dies nicht immer der Fall ist.

Beispiel 3.16
Wir betrachten die Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) := \frac{x}{\sqrt{x^2+1}}$. *Es ist*

$$f'(x) = \frac{1}{\sqrt{x^2+1}} - \frac{x^2}{(x^2+1)^{3/2}} = \frac{1}{(x^2+1)^{3/2}}.$$

Das Newton-Verfahrens zur Lösung von $f(x) = 0$ *lautet daher*

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - x_n(x_n^2+1) = -x_n^3.$$

Für jeden Startwert $x_0 \in (-1,1)$ *konvergieren die Iterierten gegen die Nullstelle* $\hat{x} = 0$ *(sogar mit sogenannter kubischer Konvergenzgeschwindigkeit, d.h. Verdreifachung der richtigen Stellen pro Iterationsschritt). Für* $x_0 = 1$ *und* $x_0 = -1$ *springen die Iterierten jedoch zwischen* -1 *und* 1 *hin und her und für* $|x_0| > 1$ *divergieren die Iterierten mit alternierenden Vorzeichen betragsmäßig gegen unendlich.*

3.5 Eindimensionale Optimierungsprobleme

Wir skizzieren noch die Grundidee, wie sich die Taylorentwicklungen 1. und 2. Ordnung zur Lösung unrestringierter, kontinuierlicher, eindimensionaler Optimierungsaufgabe einsetzen lassen. Zu einer hinreichend oft stetig differenzierbaren Funktion $f : \mathbb{R} \to \mathbb{R}$ suchen wir ein (möglicherweise nur lokales) Minimum. Dies schreibt man auch als

$$f(x) \to \min! \quad \text{u.d.N.} \quad x \in \mathbb{R}$$

und nennt f die zu minimierende *Zielfunktion* (engl.: *cost functional*). Maximierungsprobleme lassen sich durch Negation der Zielfunktion in Minimierungsprobleme überführen, so dass wir sie nicht separat behandeln müssen.

Beginnend mit einer Startnäherung $x_0 \in \mathbb{R}$ erhalten wir mit der Taylorentwicklungen 1. Ordnung

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0).$$

Für $f'(x_0) > 0$ sollte man entsprechend dieser Näherung also x kleiner als x_0 zu wählen und für $f'(x_0) < 0$ sollte man x kleiner als x_0 wählen. Da die Taylorentwicklung nur in einer Nähe des Entwicklungspunkts eine gute Approximation der Funktion f ist, sollte man sich aber auch nicht zu weit von x_0 entfernen. Dieser Idee folgend wählt man im k-ten Schritt

$$x_{k+1} = x_k - s_k f'(x_k)$$

mit einer *Schrittweite* $s_k > 0$ (im Kontext machinellen Lernens heißt s_k auch *Lernrate*) die aufgrund der Approximationseigenschaft der Taylor-Näherung nicht zu groß gewählt werden darf, aber auch nicht zu klein (sonst benötigt das Verfahren zuviele Schritte oder kommt gar nicht mehr vorran). Man stoppt das Verfahren, wenn $f'(x_k) = 0$ ist (bzw. wenn $f'(x_k)$ betragsmäßig unter einer vorgegebenen Toleranzschwelle liegt). Dieses Vorgehen heißt *Gradientenverfahren*[1].

Ein weiteres Optimierungsverfahren erhalten wir durch Verwendung der Taylorentwicklung 2. Ordnung in jedem Iterationsschritt

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k) + \frac{f''(x_k)}{2}(x - x_k)^2.$$

Ist die Parabel $f(x_k) + f'(x_k)(x - x_k) + \frac{f''(x_k)}{2}(x - x_k)^2$ nach oben geöffnet, so ist ihr Minimum ihr Scheitelpunkt. Wir verwenden daher diesen Scheitelpunkt als neue Iterierte und erhalten durch leichtes Nachrechnen die Iterationsvorschrift

$$x_{k+1} := x_k - \frac{f'(x_k)}{f''(x_k)}.$$

Man sieht sofort, dass dies der Anwendung des Newton-Verfahrens auf das Nullstellenproblem

$$f'(x) = 0$$

entspricht, man dieses Vorgehen deshalb auch das *Newton-Verfahren für Optimierungsprobleme*.

[1] Der Gradient bezeichnet eine Verallgemeinerung der Ableitung im Mehrdimensionalen, wir werden den Begriff in Abschnitt 5.1 einführen.

Sowohl Gradienten- als auch Newtonverfahren stoppen in Punkten mit verschwindender Ableitung, können also im Allgemeinen nur zur Bestimmung lokaler Minima eingesetzt werden. (Und auch das ist nicht garantiert. Falls ein Iterationsschritt zufälligerweise genau ein lokales Maximum trifft, so stoppt das Verfahren auch da.) Das Newton-Verfahren findet lokale Minima typischerweise schneller als das Gradientenverfahren (mit quadratischer statt linearer Konvergenzgeschwindigkeit), unterliegt aber zusätzlich dem Problem lokaler Konvergenz, dass sich der Startwert schon hinreichend nah am Minimum befinden muss.

Kapitel 4

Integration und Approximation

4.1 Das Riemann-Integral

Definition 4.1 (Integral von Treppenfunktionen)
Eine Funktion $f:\ [a,b] \to \mathbb{R}$ heißt Treppenfunktion, wenn eine Partition des Intervalls $[a,b]$ existiert bezüglich der f stückweise konstant ist, d.h. es existieren

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b,$$

und $c_1, \ldots, c_n \in \mathbb{R}$ mit

$$f(x) = c_k \quad \text{für alle } x \in]x_{k-1}, x_k[, \quad k = 1, \ldots, n.$$

Die Menge der Treppenfunktionen bezeichnen wir mit $T([a,b])$.

Das Integral einer Treppenfunktion $f \in T([a,b])$ definieren wir als

$$\int_a^b f(x)\,\mathrm{d}x := \sum_{k=1}^{n} c_k (x_k - x_{k-1}).$$

Das Integral einer Treppenfunktion stimmt mit dem (gerichteten) Flächeninhalt zwischen Funktionsgraph und x-Achse überein. Die Werte der Treppenfunktion an den endlich vielen Übergangsstellen $x_0, \ldots, x_n$ spielen dabei keine Rolle, vgl. das in der Vorlesung gemalte Bild.

Für allgemeinere Funktionen lässt sich der Integralbegriff durch Einschachtelung mit Treppenfunktionen definieren.

Definition 4.2 (Riemann-Integral)
Sei $f:\ [a,b] \to \mathbb{R}$ eine beschränkte Funktion. Wir definieren

(a) das Oberintegral von f durch

$$\overline{\int_a^b} f(x)\,\mathrm{d}x := \inf\left\{\int_a^b g(x)\,\mathrm{d}x :\ g \in T([a,b]),\ g(x) \geq f(x)\ \forall x \in [a,b]\right\}$$

(b) das Unterintegral von f durch

$$\underline{\int_a^b} f(x)\,\mathrm{d}x := \sup\left\{\int_a^b g(x)\,\mathrm{d}x :\ g \in T([a,b]),\ g(x) \leq f(x)\ \forall x \in [a,b]\right\}$$

Stimmen Ober- und Unterintegral überein, dann heißt f Riemann-integrierbar und wir definieren ihr Riemann-Integral durch

$$\int_a^b f(x)\,\mathrm{d}x := \overline{\int_a^b} f(x)\,\mathrm{d}x = \underline{\int_a^b} f(x)\,\mathrm{d}x$$

Satz 4.3
Es seien $f, g :\ [a,b] \to \mathbb{R}$

(a) Sind f und g Riemann-integrierbar und ist $c \in \mathbb{R}$, so sind auch cf und $f+g$ Riemann-integrierbar und es gilt

$$\int_a^b cf(x)\,\mathrm{d}x = c\int_a^b f(x)\,\mathrm{d}x,$$
$$\int_a^b (f(x)+g(x))\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x.$$

(b) Ist f Riemann-integrierbar und gilt $f(x) = g(x)$ für fast alle $x \in [a,b]$ (d.h. alle bis auf endlich viele), dann ist auch g Riemann-integrierbar und es gilt

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b g(x)\,\mathrm{d}x.$$

(c) Sind f und g Riemann-integrierbar und gilt $f(x) \leq g(x)$ für fast alle $x \in [a,b]$, dann gilt

$$\int_a^b f(x)\,\mathrm{d}x \leq \int_a^b g(x)\,\mathrm{d}x.$$

Außerdem gilt für $f :\ [a,b] \to \mathbb{R}$, $g :\ [b,c] \to \mathbb{R}$, $c > b > a$.

(d) Sind f und g Riemann-integrierbar, so ist es auch die stückweise zusammengesetzte Funktion

$$h(x) := \begin{cases} f(x) & \text{für } a \le x \le b \\ g(x) & \text{für } b < x \le c \end{cases}$$

und es gilt

$$\int_a^c h(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_b^c g(x)\,\mathrm{d}x.$$

Umgekehrt ist auch jede Einschränkung einer Riemann-integrierbaren Funktion $f : [a,b] \to \mathbb{R}$ auf ein Teilintervall $[\alpha,\beta] \subseteq [a,b]$ Riemann-integrierbar.

Beweis: Dies folgt jeweils leicht aus Definition 4.2. Wir skizzieren nur die jeweilige Beweisidee. Für (a) zeigt man, dass Summe und skalare Vielfache von Treppenfunktionen wieder Treppenfunktionen sind. Für (b) nimmt man die endlich vielen Punkte als Stützstellen in die Partition der Treppenfunktionen auf. Für (c) nutzt man, dass jede Treppenfunktion, die g von oben beschränkt, dies auch für f tut (und umgekehrt mit unteren Schranken). (d) folgt, da sich Treppenfunktionen abschnittsweise zu einer auf einem größeren Intervall definierten Treppenfunktion zusammensetzen lassen und umgekehrt auch die Einschränkung einer Treppenfunktion auf ein Teilintervall eine Treppenfunktion bleibt. □

Satz 4.4
Jede stetige, jede monoton wachsende und auch jede monton fallende Funktion $f : [a,b] \to \mathbb{R}$ ist Riemann-integrierbar. Jede (auf endlich vielen Stücken) stückweise aus solchen Funktionen zusammengesetzte Funktion ist Riemann-integrierbar.

Beweis: Sei $f : [a,b] \to \mathbb{R}$ eine stetige Funktion. Wir zerlegen $[a,b]$ in n äquidistante Teilintervalle

$$a = x_0 < x_1 < \ldots < x_{n-1} = x_n = b, \quad \text{mit} \quad x_k := a + k\frac{b-a}{n}, \quad k = 0,\ldots,n.$$

Auf dieser Partition definieren wir Treppenfunktionen $g_n, h_n \in T([a,b])$ durch

$$g_n(x) = \min_{x\in[x_{k-1},x_k]} f(x) \quad \text{für } x \in]x_{k-1},x_k], \quad k = 1,\ldots,n,$$
$$h_n(x) = \max_{x\in[x_{k-1},x_k]} f(x) \quad \text{für } x \in]x_{k-1},x_k], \quad k = 1,\ldots,n,$$

sowie $g_n(a) := \min_{x\in[a,x_1]} f(x)$ und $h_n(a) := \max_{x\in[a,x_1]} f(x)$.
Dann gilt $g_n(x) \le f(x) \le h_n(x)$ für alle $x \in [a,b]$ und daher

$$\int_a^b g_n(x)\,\mathrm{d}x \le \underline{\int_a^b} f(x)\,\mathrm{d}x \le \overline{\int_a^b} f(x)\,\mathrm{d}x \le \int_a^b h_n(x)\,\mathrm{d}x.$$

Für die Integrale über g_n und h_n erhalten wir aber aus Satz 2.14

$$\begin{aligned}\int_a^b h_n(x)\,\mathrm{d}x - \int_a^b g_n(x)\,\mathrm{d}x &= \frac{b-a}{n}\sum_{k=1}^{n}\left(\max_{x\in[x_{k-1},x_k]} f(x) - \min_{x\in[x_{k-1},x_k]} f(x)\right)\\ &\le (b-a)\max_{k=1,\dots,n}\left(\max_{x\in[x_{k-1},x_k]} f(x) - \min_{x\in[x_{k-1},x_k]} f(x)\right)\\ &\to 0.\end{aligned}$$

Daher gilt $\underline{\int_a^b} f(x)\,\mathrm{d}x = \overline{\int_a^b} f(x)\,\mathrm{d}x$. f ist also Riemann-integrierbar.

Zum Beweis der Riemann-integrierbarkeit einer monoton wachsende Funktion gehen wir genauso vor und erhalten im letzten Schritt mit einem Teleskopsummenargument

$$\begin{aligned}&\int_a^b h_n(x)\,\mathrm{d}x - \int_a^b g_n(x)\,\mathrm{d}x = \frac{b-a}{n}\sum_{k=1}^{n}\left(\max_{x\in[x_{k-1},x_k]} f(x) - \min_{x\in[x_{k-1},x_k]} f(x)\right)\\ &= \frac{b-a}{n}\sum_{k=1}^{n}\left(f(x_k)-f(x_{k-1})\right)\\ &= \frac{b-a}{n}\left(f(x_1)-f(x_0)+f(x_2)-f(x_1)+f(x_3)-f(x_2)+\ldots+f(x_n)-f(x_{n-1})\right)\\ &= \frac{b-a}{n}\left(f(b)-f(a)\right)\to 0.\end{aligned}$$

Damit folgt wiederum die Riemann-integrierbarkeit. Für monoton fallende Funktionen und für stückweise zusammengesetzte Funktionen folgt die Aussage durch Anwendung von Satz 4.3(a) und (d). □

Beispiel 4.5
Ein Beispiel für eine nicht Riemann-integrierbare Funktion ist die sogenannte Dirichlet-Funktion

$$D(x) := \begin{cases} 1 & \text{falls } x\in\mathbb{Q},\\ 0 & \text{falls } x\in\mathbb{R}\setminus\mathbb{Q}.\end{cases}$$

Für sie gilt

$$\overline{\int_a^b} D(x)\,\mathrm{d}x = 1 > 0 = \underline{\int_a^b} D(x)\,\mathrm{d}x.$$

Bemerkung 4.6
Man definiert auch

$$\int_a^a f(x)\,\mathrm{d}x := 0 \quad \textit{und} \quad \int_b^a f(x)\,\mathrm{d}x := -\int_a^b f(x)\,\mathrm{d}x.$$

Ist $f :]a,b] \to \mathbb{R}$ (also f in a nicht definiert), so definiert man außerdem das uneigentliche Integral durch

$$\int_a^b f(x)\,\mathrm{d}x := \lim_{\alpha \to a^+} \int_\alpha^b f(x)\,\mathrm{d}x,$$

falls f auf jedem Teilintervall $[\alpha,b]$, $\alpha > a$, integrierbar ist und dieser Grenzwert existiert. Ist $f :]a,b[\to \mathbb{R}$ (also an beiden Randpunkten nicht definiert), so wählt man $a < c < b$ und definiert

$$\int_a^b f(x)\,\mathrm{d}x := \lim_{\alpha \to a^+} \int_\alpha^c f(x)\,\mathrm{d}x + \lim_{\beta \to b^-} \int_c^\beta f(x)\,\mathrm{d}x,$$

falls f auf jedem $[\alpha,c]$ und jedem $[c,\beta]$, mit $\alpha > a$ und $\beta < b$, integrierbar ist und die beiden Grenzwerte existieren. Man kann zeigen, dass dies für jede Wahl von $c \in]a,b[$ denselben Wert ergibt. Analog definiert man Integrale mit Grenzen ∞ und/oder $-\infty$.

4.2 Hauptsatz der Differential- & Integralrechnung

Für konstante Funktionen $f : [a,b] \to \mathbb{R}$, $f(x) = c$ folgt (da diese insbesondere Treppenfunktionen sind) aus der Definition der Riemann-Integrals, dass

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)c.$$

Für kompliziertere Funktionen lässt sich das Integral oft (aber nicht immer) über den folgenden fundamentalen Zusammenhang zur Differentialrechnung bestimmen.

Satz 4.7
Sei $f : [a,b] \to \mathbb{R}$ eine stetige Funktion.

(a) Die durch

$$F : [a,b] \to \mathbb{R}, \quad F(x) := \int_a^x f(t)\,\mathrm{d}t.$$

definierte Funktion F ist stetig differenzierbar und es gilt $F'(x) = f(x)$ für alle $x \in [a,b]$.

(b) Ist $F : [a,b] \to \mathbb{R}$ eine stetig differenzierbare Funktion mit $F'(x) = f(x)$ für alle $x \in [a,b]$ (solche F nennen wir auch Stammfunktion von f), dann gilt

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

Man schreibt auch $F(x)\big|_a^b := F(b) - F(a)$ oder $[F(x)]_a^b := F(b) - F(a)$.

Beweis: Zum Beweis von (a) sei $x \in [a,b[$. Aus Satz 4.3(d) folgt für alle $h > 0$ (mit h hinreichend klein, so dass $x+h \in [a,b]$)

$$\begin{aligned}\frac{F(x+h)-F(x)}{h} - f(x) &= \frac{1}{h}\left(\int_a^{x+h} f(t)\,\mathrm{d}t - \int_a^x f(t)\,\mathrm{d}t\right) - f(x)\\ &= \frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t - f(x).\end{aligned}$$

Durch Abschätzung von f gegen sein Maximum und Minimum auf $[x,x+h]$ erhalten wir

$$\min_{t\in[x,x+h]} f(t) \le \frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t \le \max_{t\in[x,x+h]} f(t).$$

Mit

$$\min_{t\in[x,x+h]} f(t) \le f(x) \le \max_{t\in[x,x+h]} f(t)$$

erhalten wir insgesamt

$$\min_{t\in[x,x+h]} f(t) - \max_{t\in[x,x+h]} f(t) \le \frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t - f(x) \le \max_{t\in[x,x+h]} f(t) - \min_{t\in[x,x+h]} f(t)$$

und mit Satz 2.14 folgt

$$\lim_{h\to 0^+} \frac{F(x+h)-F(x)}{h} - f(x) = 0.$$

Für $h < 0$ gilt analoges, so dass $F'(x) = f(x)$ für alle $x \in [a,b]$ gezeigt ist.

Zum Beweis von (b) definieren wir

$$G(x) = \int_a^x f(t)\,\mathrm{d}t.$$

Nach (a) gilt dann $G'(x) = f(x) = F'(x)$ für alle $x \in [a,b]$ und aus Satz 3.9 folgt damit, dass $G - F$ konstant ist, also insbesondere

$$G(b) - F(b) = G(a) - F(a) = -F(a).$$

Damit folgt $\int_a^b f(t)\,\mathrm{d}t = G(b) = F(b) - F(a)$. □

Aus der Produkt- und Kettenregel der Differentialrechnung erhalten wir mit diesem Zusammenhang folgende Integrationsregeln.

Satz 4.8

(a) Partielle Integration: Sind $f, g\colon [a,b] \to \mathbb{R}$ stetig differenzierbare Funktionen, so gilt

$$\int_a^b f(x)g'(x)\,\mathrm{d}x = f(x)g(x)\Big|_a^b - \int_a^b f'(x)g(x)\,\mathrm{d}x.$$

(b) Integration durch Substitution: Ist $g\colon [a,b] \to \mathbb{R}$ eine stetig differenzierbare Funktion mit $g([a,b]) \subseteq [c,d]$ und ist $f\colon [c,d] \to \mathbb{R}$ eine stetige Funktion, so gilt

$$\int_a^b f(g(x))g'(x)\,\mathrm{d}x = \int_{g(a)}^{g(b)} f(t)\,\mathrm{d}t$$

Beweis: Die Aussage (a) folgt, da nach der Produktregel in Satz 3.4 die Funktion $fg\colon [a,b] \to \mathbb{R}$ stetig differenzierbar ist und aus

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x)$$

folgt, dass

$$\int_a^b (f'(x)g(x) + f'(x)g(x))\,\mathrm{d}x = f(x)g(x)\Big|_a^b.$$

Zum Beweis von (b) definieren wir eine Stammfunktion von f durch

$$F\colon [c,d] \to \mathbb{R}, \quad F(x) := \int_c^x f(t)\,\mathrm{d}t.$$

Dann ist nach der Kettenregel in Satz 3.4 die Funktion

$$F \circ g\colon [a,b] \to \mathbb{R}, \quad (F \circ g)(x) = F(g(x))$$

stetig differenzierbar und es gilt

$$(F \circ g)'(x) = F'(g(x))g'(x) = f(g(x))g'(x).$$

Damit folgt

$$\begin{aligned}\int_a^b F'(g(x))g'(x)\,\mathrm{d}x = F(g(x))\Big|_a^b &= F(g(b)) - F(g(a)) \\ &= \int_c^{g(b)} f(t)\,\mathrm{d}t - \int_c^{g(a)} f(t)\,\mathrm{d}t = \int_{g(a)}^{g(b)} f(t)\,\mathrm{d}t.\end{aligned}$$

4.3 Numerische Quadratur: Erste Verfahren

Im Folgenden sei stets $a, b \in \mathbb{R}$, $a < b$ und $f : [a,b] \to \mathbb{R}$ integrierbar. Die Berechnung eines Integrals

$$I[f] = \int_a^b f(x)\,\mathrm{d}x$$

gelingt oft nicht durch Suche nach einer Stammfunktion. Beispielsweise spielen in der Stochastik Integrale der Form

$$\int_a^b e^{-x^2}\,\mathrm{d}x$$

eine wesentliche Rolle. Für die *Gauß-Funktion* $f(x) = e^{-x^2}$ ist jedoch keine (elementare) Stammfunktion bekannt. Darüber hinaus ist in vielen Anwendungen die zu integrierende Funktion f gar nicht explizit bekannt, sondern es existiert lediglich ein Algorithmus mit dem $f(x)$ für gegebenes x ausgerechnet werden kann.

Man kann das Integral jedoch näherungsweise berechnen, indem f durch einfach integrierbare (z.B. stückweise konstante oder stückweise lineare) Funktionen ersetzt wird (numerische Quadratur). So erhalten wir als erste einfache Quadraturverfahren

- **Mittelpunktsformel:**

$$I[f] = \int_a^b f(x)\,\mathrm{d}x \approx (b-a) f\left(\frac{a+b}{2}\right) =: M[f]$$

- **Trapezformel:**

$$I[f] = \int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b) =: T[f]$$

Offenbar liefern beide Formeln sowohl für konstante als auch für lineare Funktionen den richtigen Integralwert. Für allgemeine Funktionen können wir erwarten, dass der Fehler jeweils von der Krümmung der Funktion, also von ihrer zweiten Ableitung abhängt. Tatsächlich erhalten wir die folgende Fehlerabschätzung.

Satz 4.9
Für $f \in C^2([a,b])$ ist

$$|I[f] - M[f]| \le \frac{1}{24} \left\|f''\right\|_{[a,b]} (b-a)^3,$$
$$|I[f] - T[f]| \le \frac{1}{12} \left\|f''\right\|_{[a,b]} (b-a)^3.$$

Beweis: Es gilt

$$\int_a^b f'(\tfrac{a+b}{2})(x-\tfrac{a+b}{2})\,\mathrm{d}x = \frac{f'(\frac{a+b}{2})}{2}(x-\tfrac{a+b}{2})^2\Big|_a^b$$
$$= \frac{f'(\frac{a+b}{2})}{2}(\tfrac{b-a}{2})^2 - \frac{f'(\frac{a+b}{2})}{2}(\tfrac{a-b}{2})^2 = 0$$

und damit

$$I[f]-M[f] = \int_a^b f(x)\,\mathrm{d}x - (b-a)f(\tfrac{a+b}{2}) = \int_a^b \big(f(x)-f(\tfrac{a+b}{2})\big)\,\mathrm{d}x$$
$$= \int_a^b \big(f(x)-f(\tfrac{a+b}{2})-f'(\tfrac{a+b}{2})(x-\tfrac{a+b}{2})\big)\,\mathrm{d}x.$$

Der Integrand entspricht dem Fehler der linearen Taylorentwicklung. Mit Folgerung 3.11(b) gilt daher für alle $x \in [a,b]$

$$|f(x)-f(\tfrac{a+b}{2})-f'(\tfrac{a+b}{2})(x-\tfrac{a+b}{2})| \le \frac{1}{2}(x-\tfrac{a+b}{2})^2 \left\|f''\right\|_{[a,b]}.$$

Daher erhalten wir

$$|I[f]-M[f]| \le \frac{1}{2}\left\|f''\right\|_{[a,b]} \int_a^b (x-\tfrac{a+b}{2})^2\,\mathrm{d}x$$
$$= \frac{1}{6}\left\|f''\right\|_{[a,b]} (x-\tfrac{a+b}{2})^3\Big|_a^b = \frac{1}{24}\left\|f''\right\|_{[a,b]} (b-a)^3.$$

Die Fehlerabschätzung für die Trapezformel werden wir in Beispiel 4.20 aus einem allgemeineren Resultat erhalten. □

Beispiel 4.10
Für die Funktion $f(x) = x^2$ auf $[a,b] := [-1,1]$ ist

$$I[f] = \int_{-1}^1 x^2\,\mathrm{d}x = 2/3, \quad (b-a)^3 = 8, \quad \left\|f''\right\|_{[a,b]} = 2$$

und offenbar gilt $M[f] = 0$ und $T[f] = 2$. Damit ist

$$|I[f]-M[f]| = 2/3 = \frac{1}{24}\left\|f''\right\|_{[a,b]}(b-a)^3,$$
$$|I[f]-T[f]| = 4/3 = \frac{1}{12}\left\|f''\right\|_{[a,b]}(b-a)^3.$$

Die Fehlerschranken sind für dieses Beispiel mit Gleichheit erfüllt.

Um das Integral genauer zu approximieren, zerlegen wir $[a,b]$ in n gleich große Teilintervalle

$$x_i := a+(i-1)h, \quad i=1,\dots,n+1, \quad h := \frac{b-a}{n},$$

schreiben also

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x,$$

und wenden auf jedes Teilintervall die Mittelpunkts- bzw. Trapezformel an. So erhalten wir die folgenden zusammengesetzten Formeln.

(a) **Zusammengesetzte Mittelpunktsformel:**

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} (x_{i+1}-x_i) f(\tfrac{x_i+x_{i+1}}{2}) = h \sum_{i=1}^{n} f(\tfrac{x_i+x_{i+1}}{2}) =: M_n[f].$$

(b) **Zusammengesetzte Trapezformel:**

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &\approx \sum_{i=1}^{n} \frac{x_{i+1}-x_i}{2}\left(f(x_i)+f(x_{i+1})\right)\\ &= \frac{h}{2}f(a) + h\sum_{i=2}^{n} f(x_i) + \frac{h}{2}f(b) =: T_n[f].\end{aligned}$$

Satz 4.11 (Fehler der zusammengesetzten Mittelpunkt- und Trapezformel) *Sei $f \in C^2([a,b])$. Dann gilt*

$$\begin{aligned}|I[f]-M_n[f]| &\le \frac{b-a}{24}\left\|f''\right\|_{[a,b]} h^2,\\ |I[f]-T_n[f]| &\le \frac{b-a}{12}\left\|f''\right\|_{[a,b]} h^2.\end{aligned}$$

Beweis: Wir verwenden die Fehlerabschätzungen in Satz 4.9 für jedes Teilintervall $[x_i,x_{i+1}]$, $i=1,\dots,n$, mit Breite $x_{i+1}-x_i = h$ und erhalten

$$\begin{aligned}|I[f]-M_n[f]| &\le \sum_{i=1}^{n} \frac{1}{24}\left\|f''\right\|_{[x_i,x_{i+1}]} h^3 \le \frac{1}{24}\left\|f''\right\|_{[a,b]} h^3 n\\ &= \frac{b-a}{12}\left\|f''\right\|_{[a,b]} h^2\end{aligned}$$

und

$$\begin{aligned}|I[f]-T_n[f]| &\le \sum_{i=1}^{n} \frac{1}{12}\left\|f''\right\|_{[x_i,x_{i+1}]} h^3 \le \frac{1}{12}\left\|f''\right\|_{[a,b]} h^3 n\\ &= \frac{b-a}{12}\left\|f''\right\|_{[a,b]} h^2,\end{aligned}$$

womit die Abschätzungen gezeigt sind. □

Beide Formeln erreichen mit jeweils $O(n)$ Funktionauswertungen, Additionen und Multiplikationen eine Genauigkeit von $O(h^2) = O(n^{-2})$. Um noch schneller konvergierende Formeln zu erhalten, ist es naheliegend, die Funktion stückweise durch Polynome höheren Grades zu approximieren, etwa durch Parabeln oder kubische Polynome. Dafür könnte man das Taylorpolynom verwenden, aber dann bräuchte man zusätzlich zu Funktionsauswertungen auch Werte der Ableitung. Wir verfolgen stattdessen die Idee, die Polynome so zu wählen, dass sie an möglichst vielen Stellen mit der Funktion übereinstimmen. Solche Polynome sind auch außerhalb von Quadraturverfahren wichtig, deshalb widmen wir ihnen einen eigenen Abschnitt.

4.4 Polynominterpolation

Wir betrachten die folgende *Interpolationsaufgabe*. Gegeben seien m paarweise verschiedene Knoten x_i und Werte y_i, $i = 1,\dots,m$. Gibt es ein Polynom p, das diese Werte interpoliert, d.h.

$$p(x_i) = y_i \quad \forall i \in \{1,\dots,m\}?$$

Wir bezeichnen die Menge der Polynome vom Höchstgrad $k \in \mathbb{N}$ mit

$$\Pi_k := \{p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_k x^k : \quad \alpha_0,\dots,\alpha_k \in \mathbb{R}\}.$$

Da jedes Polynom in Π_k durch seine $k+1$ Koeffizienten $\alpha_0,\dots,\alpha_k \in \mathbb{R}$ eindeutig bestimmt ist und jeder Wert $p(x_i)$ linear von dieser Koeffizienten abhängt, können wir das Interpolationsproblem als m lineare Gleichungen für $k+1$ Unbekannte schreiben. Tatsächlich existiert für $k+1 = m$ auch immer eine eindeutige Lösung. Wir geben die im Folgenden sogar explizit an.

Definition 4.12
Zu einem Gitter aus m paarweise verschiedenen, aufsteigend angeordneten Knoten

$$\{x_1, x_2, \dots, x_m\} \subset \mathbb{R}, \qquad x_1 < x_2 < \ldots < x_m,$$

definieren wir das Knotenpolynom

$$\omega(x) = \prod_{i=1}^{m} (x - x_i) \in \Pi_m$$

und die Lagrange-Grundpolynome $(i = 1,\dots,m)$

$$l_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^{m} \frac{x - x_j}{x_i - x_j} \in \Pi_{m-1}.$$

Offenbar gilt für alle $i, j = 1, \ldots, m$

$$l_i(x_j) = \delta_{ij} := \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j. \end{cases}$$

Satz 4.13 (Interpolationspolynom)
Zu einem Gitter

$$\{x_1, x_2, \ldots, x_m\} \subset \mathbb{R}, \qquad x_1 < x_2 < \ldots < x_m,$$

und Werten $y_i \in \mathbb{R}$, $i = 1, \ldots, m$ *existiert genau ein Interpolationspolynom vom Höchstgrad* $m-1$, *d.h. genau ein*

$$p \in \Pi_{m-1} \quad \text{mit} \quad p(x_i) = y_i \quad \text{für alle} \quad i \in \{1, \ldots, m\},$$

nämlich

$$p(x) = \sum_{i=1}^{m} y_i l_i(x).$$

Beweis: Offenbar gilt $p \in \Pi_{m-1}$ und wegen

$$p(x_j) = \sum_{i=1}^{m} y_i l_i(x_j) = \sum_{i=1}^{m} y_i \delta_{ij} = y_j$$

löst p auch die Interpolationsaufgabe. Die Eindeutigkeit folgt ebenfalls, da für ein lineares Gleichungssystem mit gleicher Anzahl Unbekannter und Gleichungen genau dann immer eine Lösung existiert, wenn die jeweilige Lösung eindeutig ist. □

Bemerkung 4.14

(a) Jedes Polynom $p \in \Pi_{m-1}$ *interpoliert seine eigenen Funktionsauswertungen* $y_i = p(x_i)$. *Aus Satz 4.13 folgt also*

$$p(x) = \sum_{i=1}^{m} p(x_i) l_i(x) \quad \forall p \in \Pi_{m-1}. \tag{4.1}$$

(b) Der Polynomraum Π_{m-1} *bildet offensichtlich einen Vektorraum der Dimension m. Eine Basis bilden z.B. die Monome* $(x^0, x^1, x^2, \ldots, x^{m-1})$.

Satz 4.13 zeigt, dass auch die Lagrange-Grundpolynome $(l_1(x), \ldots, l_m(x))$ *eine Basis des* Π_{m-1} *bilden. (Die Erzeugendeneigenschaft folgt aus (4.1) und die lineare Unabhängigkeit aus* $l_i(x_j) = \delta_i$.*)*

Satz 4.15 (Interpolationsfehler)
Sei $f \in C^m([a,b])$ und $p \in \Pi_{m-1}$ das Interpolationspolynom zu m paarweise verschiedenen Knoten

$$x_1, \ldots, x_m \in [a,b], \qquad x_1 < x_2 < \ldots < x_m,$$

und Werten $y_i = f(x_i)$, $i = 1, \ldots, m$.
Dann gibt es zu jedem $x \in [a,b]$ ein $\xi \in [\min\{x, x_1\}, \max\{x, x_m\}] \subseteq [a,b]$ mit

$$f(x) - p(x) = \frac{f^{(m)}(\xi)}{m!}\omega(x).$$

Beweis: Für $x = x_i$ ist die Behauptung offenbar richtig. Sei also $x \in [a,b]$ mit $x \notin \{x_1, \ldots, x_m\}$.

Betrachte (zu diesem festen x) die Funktion

$$h(t) := f(t) - p(t) - \frac{\omega(t)}{\omega(x)}(f(x) - p(x)), \quad t \in \mathbb{R}.$$

Dann ist $h \in C^m([a,b])$ und h besitzt (mindestens) $m+1$ paarweise verschiedene Nullstellen, nämlich $t = x_i$, $i = 1, \ldots, m$, und $t = x$. Das kleinste Intervall, das alle diese Nullstellen enthält, ist gegeben durch

$$I := [\min\{x, x_1\}, \max\{x, x_m\}].$$

Zwischen je zwei benachbarten Nullstellen, also an (mindestens) m verschiedenen Werten im Intervall I, liegt nach dem Mittelwertsatz (Satz 3.9) eine Nullstelle der Ableitung $h'(t)$. Zwischen je zwei benachbarten Nullstellen der Ableitung, also an (mindestens) $m-1$ verschiedenen Werten in I, liegt wiederum nach dem Mittelwertsatz eine Nullstelle der zweiten Ableitung h''. Wir fahren so fort und erhalten $m-2$ verschiedene Nullstellen von h''', $m-3$ verschiedene Nullstellen von $h^{(4)}$, usw., und schließlich eine Nullstelle $\xi \in I$ von $h^{(m)}$.

Wegen $p \in \Pi_{m-1}$ ist $p^{(m)} = 0$. Außerdem ist

$$\omega(t) = \prod_{i=1}^{m}(t - x_i) = t^m + q(t), \qquad \text{mit } q \in \Pi_{m-1}$$

und damit $\omega^{(m)}(t) = m!$. Es folgt, dass

$$0 = h^{(m)}(\xi) = f^{(m)}(\xi) - \frac{m!}{\omega(x)}(f(x) - p(x)),$$

womit die Behauptung gezeigt ist. □

Bemerkung 4.16
Aus Satz 4.15 folgt nicht, dass der Interpolationsfehler für $m \to \infty$ gegen Null konvergiert. Tatsächlich lassen sich selbst auf äquidistanten Gittern Beispiele konstruieren, bei denen das Interpolationspolynom nicht in jedem Punkt gegen die interpolierte Funktion konvergiert. In der Praxis beobachtet man bei hohem Interpolationsgrad meist unerwünschte starke Oszillationen zwischen den Interpolationspunkten.

Aus Satz 4.15 folgt aber

$$|f(x) - p(x)| \leq \frac{\left\| f^{(m)} \right\|_{[a,b]}}{m!} |b-a|^m.$$

Für festes m konvergiert der Interpolationsfehler also für kleiner werdende Intervallbreite $(b-a) \to 0$ gegen Null, und m steuert, wie schnell dies geschieht.

In vielen Teilgebieten der Numerik werden deshalb Funktionen nicht global durch ein Polynom möglichst hohen Grades angenähert, sondern auf möglichst kleinen Stücken durch Polynome von festem (oder an die Größe der Stücke angepasstem) Grad approximiert.

Bei der Interpolation mehrerer vorgegebener Punkte ergibt sich ein besonders natürlich wirkender Kurvenverlauf durch sogenannte kubische Splines. Diese bestehen stückweise aus Polynomen dritten Grades, die so gewählt sind, dass beim Übergang von einem Stück zum nächsten im Funktionswert, der ersten sowie der zweiten Ableitung jeweils kein Sprung auftritt.

4.5 Newton-Cotes-Quadraturformeln

Wir kehren zurück zur Frage, wie sich ein Integral numerisch berechnen lässt und untersuchen dafür jetzt eine allgemeine *Quadraturformel* der Form

$$I[f] = \int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=1}^{m} w_i f(x_i) =: Q[f]$$

mit noch zu bestimmenden *Gewichten* $w_i \in \mathbb{R}$ und (paarweise verschiedenen) *Knoten* $x_i \in [a,b]$, $i = 1,\ldots,m$.

Wie bei der Mittelpunkts- und Trapezformel erhalten wir aus so einer Quadraturformel dann ein *(zusammengesetztes) Quadraturverfahren* Q_n, indem wir $[a,b]$ in n-Teilintervalle der Länge $h = \frac{b-a}{n}$ unterteilen, in denen wir jeweils die Quadraturformel anwenden.

Die Trapez- und Quadraturformel liefern für lineare Funktionen (also Polynome bis Grad 1) jeweils den exakten Integralwert. Wir verallgemeinern diese Eigenschaft in der folgenden Definition.

Definition 4.17
Eine Quadraturformel Q für das Integral $I[\cdot]$ hat Exaktheitsgrad q, falls sie Polynome vom Höchstgrad q exakt integriert, also

$$Q[p] = I[p] \quad \forall p \in \Pi_q,$$

wobei Π_q der Raum der Polynome (mit reellen Koeffizienten) vom Höchstgrad q ist.

Eine naheliegende Idee zur Approximation des Integrals ist es, das Interpolationspolynom in den Knoten x_i und den Funktionauswertungen $f(x_i)$ aufzustellen und dann das Integral dieses Interpolationspolynoms als Näherung für $I[f]$ zu nehmen. Der folgende Satz zeigt, dass dies die einzige Möglichkeit ist, einen möglichst hohen Exaktheitsgrad zu erhalten und gibt die entstehenden Quadraturformeln an.

Satz 4.18
$Q[\cdot]$ sei eine Quadraturformel mit Knoten

$$\{x_1, x_2, \ldots, x_m\} \subset [a,b], \qquad x_1 < x_2 < \ldots < x_m, \quad m \in \mathbb{N},$$

und Gewichten $w_i \in \mathbb{R}$, $i = 1, \ldots, m$, also

$$Q[f] = \sum_{i=1}^{m} w_i f(x_i).$$

Dann hat Q genau dann Exaktheitsgrad $q \geq m-1$, wenn die Gewichte w_i erfüllen

$$w_i = \int_a^b l_i(x)\,\mathrm{d}x, \quad i = 1, \ldots, m. \tag{4.2}$$

Außerdem gilt für jede Quadraturformel mit Exaktheitsgrad $q \geq m-1$

$$Q[f] = \int_a^b p(x)\,\mathrm{d}x,$$

wobei $p \in \Pi_{m-1}$ das Interpolationspolynom mit $p(x_j) = f(x_j)$, $j = 1, \ldots, m$, ist.

Die Quadraturformeln mit äquidistante Knoten $a = x_1 < \ldots < x_m = b$ und gemäß (4.2) definierten Gewichten w_i heißen (abgeschlossene) Newton-Cotes-Formeln.

Beweis: Die Quadraturformel habe Exaktheitsgrad $q \geq m-1$. Dann integriert sie insbesondere die Lagrange-Grundpolynome $l_j \in \Pi_{m-1}$ exakt, also

$$\int_a^b l_j(x)\,\mathrm{d}x = Q[l_j] = \sum_{i=1}^{m} w_i l_j(x_i) = w_j, \quad j = 1,\ldots,m.$$

Um die Rückrichtung zu zeigen, sei $f \in \Pi_{m-1}$. Dann ist f ein Interpolationspolynom zu seinen eigenen Funktionswerten $y_i := f(x_i)$, $i = 1,\ldots,m$. Aus Satz 4.13 folgt, dass

$$f(x) = \sum_{i=1}^{m} f(x_i) l_i(x).$$

Erfüllen die Gewichte die Bedingung (4.2), so folgt wie oben

$$\begin{aligned} I[f] &= \int_a^b f(x)\,\mathrm{d}x = \int_a^b \sum_{i=1}^{m} f(x_i) l_i(x)\,\mathrm{d}x \\ &= \sum_{i=1}^{m} \underbrace{\int_a^b l_i(x) w(x)\,\mathrm{d}x}_{=w_i} f(x_i) = \sum_{i=1}^{m} w_i f(x_i) = Q[f]. \end{aligned}$$

$f \in \Pi_{m-1}$ wird also exakt integriert.

Für das Interpolationspolynom $p \in \Pi_{m-1}$ gilt in allen Knoten $p(x_i) = f(x_i)$ und daher $Q[f] = Q[p]$. Besitzt eine Quadraturformel Exaktheitsgrad $q \geq m-1$, so gilt außerdem $Q[p] = \int_a^b p(x)\,\mathrm{d}x$. □

Aus Satz 4.15 erhalten wir die folgende Fehlerabschätzung:

Satz 4.19
Sei $f \in C^m([a,b])$ und $Q[\cdot]$ sei eine Quadraturformel mit Knoten

$$\{x_1, x_2, \ldots, x_m\} \subset [a,b], \qquad x_1 < x_2 < \ldots < x_m, \quad m \in \mathbb{N},$$

und Gewichten gemäß (4.2).

Dann gilt

$$|I[f] - Q[f]| \leq \frac{\left\| f^{(m)} \right\|_{[a,b]}}{m!} \int_a^b |\omega(x)|\,\mathrm{d}x,$$

wobei $\omega(x) = \prod_{i=1}^{m}(x - x_i)$ das Knotenpolynom ist.

Beweis: Sei $f \in C^m([a,b])$ und $p \in \Pi_{m-1}$ das dazugehörige Interpolationspolynom mit $p(x_i) = f(x_i)$ für alle $i = 1,\dots,m$. Aus Satz 4.15 folgt, dass

$$|f(x) - p(x)| \leq \frac{\left\|f^{(m)}\right\|_{[a,b]}}{m!} |\omega(x)| \quad \forall x \in [a,b].$$

Wegen Satz 4.18 ist $Q[f] = Q[p] = I[p]$ und damit

$$\begin{aligned} |I[f] - Q[f]| = |I[f-p]| &\leq \int_a^b |f(x) - p(x)| \,\mathrm{d}x \\ &\leq \frac{\left\|f^{(m)}\right\|_{[a,b]}}{m!} \int_a^b |\omega(x)| \,\mathrm{d}x. \qquad \square \end{aligned}$$

Beispiel 4.20

(a) Die Trapezformel hat $m = 2$ Knoten $a = x_1 < x_2 = b$ und Exaktheitsgrad 1. Sie ist also die abgeschlossene Newton-Cotes Formel für $m = 2$ und ihre Gewichte müssen (4.2) erfüllen. Mit

$$\begin{aligned} \int_a^b |\omega(x)| \,\mathrm{d}x &= \int_a^b (x-a)(b-x) \,\mathrm{d}x \\ &= \frac{(x-a)^2}{2}(b-x)\Big|_a^b + \int \frac{(x-a)^2}{2} \,\mathrm{d}x \\ &= \frac{(x-a)^3}{6}\Big|_a^b = \frac{(b-a)^3}{6}. \end{aligned}$$

erhalten wir aus Satz 4.19 die Fehlerabschätzung

$$|I[f] - T[f]| \leq \frac{\|f''\|_{[a,b]}}{12}(b-a)^3.$$

(b) Die abgeschlossene Newton-Cotes-Formel für $m = 3$ besitzt die Knoten

$$x_1 = a, \quad x_2 = \frac{a+b}{2}, \quad x_3 = b$$

und die Gewichte

$$\begin{aligned} w_1 &= \int_a^b l_1(x) \,\mathrm{d}x = \int_a^b \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} \,\mathrm{d}x = \frac{1}{6}(b-a), \\ w_2 &= \int_a^b l_2(x) \,\mathrm{d}x = \int_a^b \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} \,\mathrm{d}x = \frac{2}{3}(b-a), \\ w_3 &= \int_a^b l_3(x) \,\mathrm{d}x = \int_a^b \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)} \,\mathrm{d}x = \frac{1}{6}(b-a). \end{aligned}$$

Dies ergibt die sogenannte Simpsonregel

$$S[f] = \tfrac{b-a}{6}\left(f(a) + 4f(\tfrac{a+b}{2}) + f(b)\right)$$

Nach Satz 4.18 besitzt sie (mindestens) Exaktheitsgrad 2 und ihr Wert entspricht dem Integral über eine Parabel, die die Funktion in den drei Knoten x_1, x_2 und x_3 interpoliert. Tatsächlich besitzt sie aus Symmetriegründen sogar Exaktheitsgrad 3.

Bemerkung 4.21

(a) Aus einer Quadraturformel Q erhalten wir ein (zusammengesetztes) Quadraturverfahren Q_n, indem wir $[a,b]$ in n-Teilintervalle der Länge $h = \frac{b-a}{n}$ unterteilen, in denen wir jeweils Q anwenden. Mit Satz 4.19 können wir den Fehler auf jedem der n Teilintervalle abschätzen durch

$$\frac{\left\|f^{(m)}\right\|_{[a,b]}}{m!} h^{m+1}$$

und wir erhalten die Fehlerabschätzung

$$|I[f] - Q_n[f]| \leq \frac{(b-a)}{m!} \left\|f^{(m)}\right\|_{[a,b]} h^m.$$

Für festes (hinreichend glattes) f fällt der Fehler (mindestens) so schnell wie h^m. Den Exponenten m bezeichnet man auch als Konsistenzordnung des Verfahrens.

Mit Satz 4.18 können wir also Quadraturformeln beliebig hohen Exaktheitsgrads und beliebig hoher Konsistenzordnung erhalten.

(b) Es ist nicht gesichert, dass die gemäß (4.2) aufgestellten Gewichte positiv sind. Bei den (abgeschlossenen) Newton-Cotes-Formeln treten ab $m = 8$ Knoten erstmalig negative Gewichte auf.

(c) Man kann zeigen, dass der höchste erreichbare Exaktheitsgrad mit m Knoten $q = 2m-1$ ist und dieser durch geschickte Wahl der Knoten (und Gewichte gemäß Satz 4.18) auch tatsächlich erreicht werden kann. Die zugehörigen Formeln heißen Gauß-Quadraturformeln, sie besitzen für beliebig hohes m positive Gewichte und können auf das komplette Intervall $[a,b]$ angewendet werden. Die Mittelpunktsformel ist die Gauß-Quadraturformel mit einem Knoten.

Kapitel 5

Analysis und Numerik im Mehrdimensionalen

5.1 Analysis im Mehrdimensionalen

In Definition 1.4 haben wir die Konvergenz einer reellen Zahlenfolge dadurch definiert, dass der Abstand zum Grenzwert jede vorgegebene Schranke $\varepsilon > 0$ für fast alle Folgenglieder unterschritt. Der Abstand zweier reeller Zahlen $x, y \in \mathbb{R}$ war dabei die Betragsdifferenz $|x-y|$. Um diese Ideen auf Funktionen mehrerer Veränderlicher zu erweitern, erinnern wir an den Begriff der Norm:

Definition 5.1
Eine Abbildung

$$\|\cdot\| : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \|x\|$$

heißt Norm, falls sie die folgenden Eigenschaften erfüllt:

(a) Positive Definitheit

$$\|x\| \geq 0 \quad \forall x \in \mathbb{R}^n \quad \textit{und} \quad \|x\| = 0 \iff x = 0.$$

(b) (Absolute) Homogenität

$$\|\alpha x\| = |\alpha| \, \|x\| \quad \forall x \in \mathbb{R}^n, \ \alpha \in \mathbb{R}.$$

(c) Dreiecksungleichung

$$\|x+y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^n.$$

Im Vektorraum $\mathbb{R}^n$ gibt es mehrere natürliche Normen. Man kann zeigen, dass für einen n-dimensionalen Vektor $x = (x_j)_{j=1}^n \in \mathbb{R}^n$ die folgenden Ausdrücke jeweils Normen definieren:

- *Betragssummennorm*: $\|x\|_1 := \sum_{j=1}^{n} |x_j|$,
- *Euklidnorm*: $\|x\| := \|x\|_2 := \left(\sum_{j=1}^{n} |x_j|^2 \right)^{1/2}$,
- *Maximumsnorm*: $\|x\|_\infty := \max_{j=1,\ldots,n} |x_j|$.
- *p-Norm*: $\|x\|_p := \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}$,

Dementsprechend gibt es mehrere natürliche Möglichkeiten, den Abstand zweier Vektoren zu messen. Es gilt jedoch der folgende Satz.

Satz 5.2 (Äquivalenz aller Normen auf dem $\mathbb{R}^n$)
Seien $\|\cdot\|_a$, $\|\cdot\|_b$ zwei Normen auf dem Vektorraum $\mathbb{R}^n$. Dann existieren $c, C > 0$, so dass

$$c\,\|x\|_b \le \|x\|_a \le C\,\|x\|_b\,.$$

Beweis: Es genügt offenbar zu zeigen, dass wir jede Norm $\|\cdot\|_a$ von oben und unten jeweils durch ein Vielfaches der Maximumsnorm $\|\cdot\|_\infty$ abschätzen können.

(a) Mit den Einheitsvektoren $e_1, \ldots, e_n \in \mathbb{R}^n$ folgt die obere Abschätzung durch die Maximumsnorm für $x = (x_1)_{i=1}^n \in \mathbb{R}^n$ aus der von jeder Norm erfüllten Dreiecksungleichung und Homogenität

$$\begin{aligned}\|x\|_a &= \|x_1 e_1 + \ldots + x_n e_n\|_a \le \|x_1 e_1\|_a + \ldots + \|x_n e_n\|_a \\ &= |x_1|\,\|e_1\|_a + \ldots + |x_n|\,\|e_n\|_a \le \|x\|_\infty (\|e_1\|_a + \ldots + \|e_n\|_a) = C\,\|x\|_\infty\,.\end{aligned}$$

wobei $C := \|e_1\|_a + \ldots + \|e_n\|_a$.

(b) Den Beweis, dass sich jede Norm von unten durch die Maximumsnorm abschätzen lässt, skizzieren wir nur. Man führt dazu mit der Maximumsnorm den Grenzwert- und den Stetigkeitsbegriff ein und zeigt dann wie in Abschnitt 2.3, dass stetige Funktionen auf der abgeschlossenen beschränkten Menge

$$K := \{x \in \mathbb{R}^n : \ \|x\|_\infty = 1\}$$

ihr Maximum und ihr Minimum annehmen. Mit (a) kann man zeigen, dass

$$x \mapsto \|x\|_a$$

eine stetige Funktion bezüglich der Maximumsnorm ist. Diese nimmt also ihr Minimum auf K an, d.h.

$$\exists x_{\min} \in K: \quad \|x\|_a \geq \|x_{\min}\|_a \quad \text{für alle } x \in \mathbb{R}^n \text{ mit } \|x\|_\infty = 1.$$

Wegen der positiven Definitheit der Normen $\|\cdot\|_\infty$ und $\|\cdot\|_a$ folgt aus $x_{\min} \in K$, dass $x_{\min} \neq 0$ und daraus dann

$$c := \|x_{\min}\|_a > 0.$$

Damit erhalten wir dann für alle $0 \neq x \in \mathbb{R}^n$

$$\|x\|_a = \left\| \frac{x}{\|x\|_\infty} \right\|_a \|x\|_\infty \geq c \|x\|_\infty,$$

womit die Abschätzung nach unten gezeigt ist. □

Aufgrund dieser Äquivalenz aller Normen sind wichtige Begriffe wie Grenzwert, Stetigkeit und Ableitungen unabhängig von der gewählten Norm. Wir fassen dies im Folgenden jeweils ohne Beweis zusammen.

Definition und Satz 5.3
Sei $(x^{(k)})_{k\in\mathbb{N}} \subset \mathbb{R}^n$ eine Folge von Vektoren

$$x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^T \in \mathbb{R}^n.$$

Ein Vektor $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ heißt Grenzwert (auch: Limes) der Folge,

$$\lim_{k\to\infty} x^{(k)} = x,$$

falls bezüglich einer Norm $\|\cdot\|_*$ auf dem $\mathbb{R}^n$ gilt

$$\left\| x^{(k)} - x \right\|_* \to 0, \quad \textit{also} \quad \forall \varepsilon > 0: \ \exists N \in \mathbb{N}: \ \left\| x^{(k)} - x \right\|_* < \varepsilon \quad \forall k \geq N.$$

Gilt dies bezüglich einer Norm, so auch bezüglich jeder anderen Norm. Außerdem ist die Konvergenz von Vektoren äquivalent zur komponentenweisen Konvergenz, d.h.

$$\lim_{k\to\infty} x^{(k)} = x \quad \Longleftrightarrow \quad \lim_{k\to\infty} x_j^{(k)} = x_j \quad \textit{für alle } j = 1, \ldots, n.$$

Sind $(x^{(k)})_{k\in\mathbb{N}},\ (y^{(k)})_{k\in\mathbb{N}} \subset \mathbb{R}^n$ und $(a_k)_{k\in\mathbb{N}} \subset \mathbb{R}$ jeweils konvergent, dann sind es auch $(x^{(k)}+y^{(k)})_{k\in\mathbb{N}}$ und $a_k(x^{(k)})_{k\in\mathbb{N}}$ und es gilt

$$\lim_{k\to\infty}(x^{(k)}+y^{(k)}) = \lim_{k\to\infty} x^{(k)} + \lim_{k\to\infty} y^{(k)} \quad \text{und} \quad \lim_{k\to\infty}(a_k x^{(k)}) = \lim_{k\to\infty} a_k \lim_{k\to\infty} x^{(k)}.$$

Mit dem Grenzwertbegriff für Folgen, können wir Grenzwerte von Funktionen und Stetigkeit definieren.

Definition 5.4
Sei $F:\ \mathbb{R}^n \to \mathbb{R}^m$, $n,m \in \mathbb{N}$ und $\hat{x} \in \mathbb{R}^m$. Falls ein $\hat{y} \in \mathbb{R}^m$ existiert, so dass

$$\text{für jede Folge } (x^{(k)})_{k\in\mathbb{N}} \subseteq \mathbb{R}^n \text{ mit } \lim_{k\to\infty} x^{(k)} = x$$

gilt $\lim_{k\to\infty} F(x^{(k)}) = \hat{y}$, dann nennen wir $\hat{y}$ den Grenzwert von F für $x \to \hat{x}$ und schreiben

$$\lim_{x\to\hat{x}} F(x) = \hat{y}.$$

Die Funktion F heißt stetig in $\hat{x}$, falls

$$\lim_{x\to\hat{x}} F(x) = F(\hat{x})$$

und sie heißt stetig, falls dies für alle $\hat{x} \in \mathbb{R}^n$ gilt.

Aus den Rechenregeln für Grenzwerte folgt, dass Sumen und skalare Vielfache stetiger Funktionen wiederum stetig sind.

Definition 5.5 (Partielle Ableitung)
Für $i \in \{1,\dots,n\}$ heißt eine Funktion

$$f:\ \mathbb{R}^n \to \mathbb{R}, \quad (x_1,\dots,x_n) \mapsto f(x_1,\dots,x_n).$$

in $x = (x_1,\dots,x_n) \in \mathbb{R}^n$ partiell nach der i-ten Komponente differenzierbar, falls der folgende Grenzwert existiert

$$\begin{aligned}\frac{\partial f}{\partial x_i}(x) &:= \lim_{h\to 0} \frac{f(x_1,\dots,x_{i-1},x_i+h,x_{i+1},\dots,x_n) - f(x_1,\dots,x_n)}{h} \\ &= \lim_{h\to 0} \frac{f(x+he_i)-f(x)}{h}.\end{aligned}$$

Dies ist offenbar äquivalent dazu, dass die Funktion

$$g:\ \mathbb{R} \to \mathbb{R}, \quad t \mapsto g(t) = f(x_1,\dots,x_{i-1},t,x_{i+1},\dots,x_n)$$

in $t = x_i$ differenzierbar ist und $\frac{\partial f}{\partial x_i}(x) = g'(x_i)$.

Beispiel 5.6
Für die Funktion $f(x_1,x_2,x_3) = x_1 + x_1 e^{x_2}$ *gilt*

$$\frac{\partial f}{\partial x_1}(x) = 1 + e^{x_2}, \quad \frac{\partial f}{\partial x_2}(x) = x_1 e^{x_2}, \quad \textit{und} \quad \frac{\partial f}{\partial x_3}(x) = 0.$$

Definition 5.7 (Jacobi-Matrix)
Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ *schreiben wir komponentenweise als*

$$F(x_1,\ldots,x_n) = \begin{pmatrix} F_1(x_1,\ldots,x_n) \\ F_2(x_1,\ldots,x_n) \\ \vdots \\ F_m(x_1,\ldots,x_n). \end{pmatrix}$$

Existieren alle partiellen Ableitungen aller Komponenten in $x \in \mathbb{R}^n$*, so nennen wir*

$$F'(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x) & \frac{\partial F_1}{\partial x_2}(x) & \ldots & \frac{\partial F_1}{\partial x_n}(x) \\ \frac{\partial F_2}{\partial x_1}(x) & \frac{\partial F_2}{\partial x_2}(x) & \ldots & \frac{\partial F_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1}(x) & \frac{\partial F_m}{\partial x_2}(x) & \ldots & \frac{\partial F_m}{\partial x_n}(x) \end{pmatrix} \in \mathbb{R}^{m\times n}$$

die Jacobi-Matrix von F.

$F'(x)$ ist mehr als nur die tabellarische Zusammenstellung aller partieller Ableitungen. Wir erinnern daran, dass im skalaren Fall die Ableitung zu einem gegebenen Punkt die beste lineare Approximation (nämlich die Tangente) an die Funktion in der Umgebung dieses Punktes darstellte. Dieser Zusammenhang gilt auch im Mehrdimensionalen. F' ordnet jedem $x \in \mathbb{R}^n$ eine lineare Approximation der Funktion von $\mathbb{R}^n$ nach $\mathbb{R}^m$, also eine $m \times n$-Matrix zu

$$F' : \mathbb{R}^n \to \mathbb{R}^{m\times n}.$$

Satz 5.8 (Totale Differenzierbarkeit)
Existieren alle partiellen Ableitungen von $F : \mathbb{R}^n \to \mathbb{R}^m$ *in allen* $x \in \mathbb{R}^n$ *und sind diese stetig (kurz:* F *ist stetig differenzierbar) , so gilt*

$$F(x) = F(x_0) + F'(x_0)(x - x_0) + o(\|x - x_0\|).$$

Die Matrixstruktur von $F'(x)$ ist außerdem für die folgende mehrdimensionale Kettenregel relevant.

Satz 5.9 (Mehrdimensionale Kettenregel)
Sind $G\colon \mathbb{R}^k \to \mathbb{R}^n$ *und* $F\colon \mathbb{R}^n \to \mathbb{R}^m$ *stetig differenzierbar, so ist auch*

$$F \circ G\colon \mathbb{R}^k \to \mathbb{R}^m, \quad (F \circ G)(x) := F(G(x))$$

stetig differenzierbar und es ist $(F \circ G)'(x) = F'(G(x))G'(x)$.

Beispiel 5.10
Die Funktion $f\colon (0,\infty) \to \mathbb{R}$, $f(x) := x^x$ *erfüllt* $f(x) = e^{\ln(x)x}$ *und besitzt daher nach der eindimensionalen Ketten- und Produktregel die Ableitung*

$$f'(x) = e^{\ln(x)x}\left(\frac{1}{x}x + \ln(x)\right) = x^x(1+\ln(x)) = x^x + x^x\ln(x).$$

Alternativ kann sie aber auch geschrieben werden als $f(x) = F(G(x))$ *mit*

$$F(x_1,x_2) := x_1^{x_2} \quad \textit{und} \quad G(x) := \begin{pmatrix} G_1(x) \\ G_2(x) \end{pmatrix} = \begin{pmatrix} x \\ x \end{pmatrix}.$$

Es ist

$$F'(x_1,x_2) = \begin{pmatrix} \frac{\partial F}{\partial x_1} & \frac{\partial F}{\partial x_2} \end{pmatrix} = \begin{pmatrix} x_2 x_1^{x_2-1} & \ln(x_1)x_1^{x_2} \end{pmatrix} \in \mathbb{R}^{1\times 2},$$
$$G'(x) = \begin{pmatrix} G_1'(x) \\ G_2'(x) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{2\times 1}.$$

Damit erhalten wir ebenfalls

$$f'(x) = F'(G(x))G'(x) = \begin{pmatrix} G_2(x)G_1(x)^{G_2(x)-1} & \ln(G_1(x))G_1(x)^{G_2(x)} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$
$$= xx^{x-1} + \ln(x)x^x = x^x + x^x\ln(x).$$

Höhere Ableitungen lassen sich ebenfalls definieren. Die partiellen Ableitungen $\frac{\partial F_j}{\partial x_k}(x)$ sind jeweils auch selbst wieder eine Funktion von $x = (x_1,\ldots,x_n) \in \mathbb{R}^n$ und können entsprechend (wenn sie differenzierbar sind) nach einem x_i partial differenziert werden. Wir bezeichnen das mit

$$\frac{\partial^2 F_j}{\partial x_i \partial x_k}.$$

Sind alle zweiten partiellen Ableitungen stetig, so kann man zeigen, dass die Reihenfolge der Differentiation (erst nach x_k dann nach x_i bzw. erst nach x_i und dann nach x_k) unerheblich ist (*Satz von Schwarz*).

Auch das Konzept der Ableitung als lineare Approximation lässt sich auf höhere Ableitungen erweitern. Da $F' : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ muss entsprechend F'' jedem $x \in \mathbb{R}^n$ eine lineare Abbildung vom $\mathbb{R}^n$ in den $\mathbb{R}^{m\times n}$ zuordnen und F''' muss jedem $x \in \mathbb{R}^n$ eine lineare Abbildung vom $\mathbb{R}^n$ in den Raum der linearen Abbildungen vom $\mathbb{R}^n$ in den $\mathbb{R}^{m\times n}$ zuordnen. Solche Abbildungen lassen sich als multilineare Abbildungen charakterisieren. Wir führen dies und die allgemeine höherdimensionale Taylor-Formel in dieser Vorlesung nicht aus, geben aber im Folgenden noch ohne Beweis für skalarwertige Funktionen die in der Optimierung wichtigen Taylor-Formel erster und zweiter Ordnung an.

Satz 5.11
Ist $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, dann gilt für alle $x, y \in \mathbb{R}$

$$f(y) = f(x) + \nabla f(x)^T (y-x) + o(\|y-x\|).$$

Ist $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, dann gilt für alle $x, y \in \mathbb{R}$

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) + o(\|y-x\|^2),$$

wobei der Gradient $\nabla f(x) \in \mathbb{R}^n$ und die Hesse-Matrix[1] $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ definiert sind durch

$$\nabla f(x) = f'(x)^T = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \quad \text{und} \quad \nabla^2 f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1^2}(x) & \cdots & \frac{\partial f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_n \partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n^2}(x) \end{pmatrix}.$$

5.2 Das Newton-Verfahren im $\mathbb{R}^n$

In Abschnitt 3.4 haben wir das eindimensionale Newton-Verfahren kennengelernt zur Lösung der eindimensionalen Nullstellenaufgabe

$$f(x) \stackrel{!}{=} 0 \quad \text{mit} \quad f : \mathbb{R} \to \mathbb{R}.$$

Dabei wurde (beginnend mit einer Startnäherung $x_0 \in \mathbb{R}$) die Funktion f jeweils durch ihre lineare Taylor-Näherung in der aktuellen Iterierten x_k ersetzt

$$f(x_k) + f'(x_k)(x - x_k) \approx f(x) \stackrel{!}{=} 0.$$

[1] Achtung: Die Notation ∇^2 wird in der Literatur nicht einheitlich verwendet. Im Kontext partieller Differentialgleichung findet sich ∇^2 auch als Notation für die Summe der zweiten Ableitungen (Laplace-Operator).

Die Nullstelle dieser linearen Näherung wurde dann als nächste Iterierte x_{k+1} verwendet, d.h.

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0, \quad \text{und damit} \quad x_{k+1} := x_k - \frac{f(x_k)}{f'(x_k)}.$$

Die gleiche Idee können wir auch auf mehrdimensionale Nullstellenaufgaben anwenden. Gegeben eine mehrdimensionale Funktion $F: \mathbb{R}^n \to \mathbb{R}^n$ betrachten wir die Nullstellenaufgabe

$$F(x) \stackrel{!}{=} 0.$$

Dies entspricht einem Gleichungssystem mit n Unbekannten und n Gleichungen

$$\begin{pmatrix} F_1(x_1, \ldots, x_n) \\ \vdots \\ F_n(x_1, \ldots, x_n) \end{pmatrix} = F(x) \stackrel{!}{=} 0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Wieder können wir beginnend mit einer Startnäherung $x^{(0)} \in \mathbb{R}^n$ in jedem Iterationsschritt F ersetzen durch die lineare Taylornäherung in der aktuellen Iterierten $x^{(k)} \in \mathbb{R}^n$

$$F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)}) \approx F(x) \stackrel{!}{=} 0$$

Die Nullstelle $x^{(k+1)} \in \mathbb{R}^n$ dieser linearen Näherung erfüllt

$$\underbrace{F'(x^{(k)})}_{\in \mathbb{R}^{n \times n}} \underbrace{(x^{(k+1)} - x^{(k)})}_{\in \mathbb{R}^n} = \underbrace{-F(x^{(k)})}_{\in \mathbb{R}^n}.$$

Dies ist ein lineares Gleichungssystem (LGS) mit n Unbekannten und n Gleichungen. Falls die Jacobi-Matrix $F'(x^{(k)})$ invertierbar ist, dann ist die Lösung[2] gegeben durch

$$x^{(k+1)} := x^{(k)} - F'(x^{(k)})^{-1} F(x^{(k)}).$$

Man kann zeigen, dass Satz 3.15 entsprechend auch für das mehrdimensionale Newton-Verfahren gilt. Ist $F: \mathbb{R}^n \to \mathbb{R}^n$ zweimal stetig differenzierbar und besitzt es eine Nullstelle mit invertierbarer Jacobi-Matrix, so konvergiert das mehrdimensionale Newton-Verfahren für alle Startwerte in einer Umgebung mit quadratischer Konvergenz gegen diese Nullstelle.

Wir führen die im Allgemeinen nur lokale Konvergenz des Newton-Verfahrens an einem Beispiel vor.

[2] Ein einzelnes LGS lässt sich schneller lösen, als die Inverse einer Matrix zu berechnen. In praktischen Implementierungen löst man daher das LGS $F'(x^{(k)})d^{(k)} = -F(x^{(k)})$ und definiert dann $x^{(k+1)} := x^{(k)} + d^{(k)}$.

Beispiel 5.12
Eine Kreis mit Mittelpunkt (m_x, m_y) und Radius r besteht aus allen Punkten $(x,y) \in R^2$ die erfüllen

$$(x-m_x)^2 + (y-m_y)^2 = r^2,$$

vgl. die in der Vorlesung gemalte Skizze.

Wir suchen nun einen solchen Kreis mit Radius 5, auf dem die zwei Punkte $(0,0)$ und $(1,-1)$ liegen. Dies führt auf zwei Gleichungen für die zwei Unbekannten m_x und m_y

$$\begin{aligned}(0-m_x)^2 + (0-m_y)^2 &= 5^2,\\ (1-m_x)^2 + (-1-m_y)^2 &= 5^2.\end{aligned}$$

Wir schreiben das als Nullstellenaufgabe für eine zweidimensionale Funktion

$$F(m_x,m_y) := \begin{pmatrix} m_x^2 + m_y^2 - 25 \\ (1-m_x)^2 + (-1-m_y)^2 - 25 \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0 \in \mathbb{R}^2.$$

Die zugehörige Jakobimatrix lautet

$$\begin{aligned} F'(m_x,m_y) &= \begin{pmatrix} \frac{\partial(m_x^2+m_y^2-25)}{\partial m_x} & \frac{\partial(m_x^2+m_y^2-25)}{\partial m_y} \\ \frac{\partial((1-m_x)^2+(-1-m_y)^2-25)}{\partial m_x} & \frac{\partial((1-m_x)^2+(-1-m_y)^2-25)}{\partial m_y} \end{pmatrix} \\ &= \begin{pmatrix} 2m_x & 2m_y \\ 2(1-m_x)(-1) & 2(-1-m_y)(-1) \end{pmatrix} = \begin{pmatrix} 2m_x & 2m_y \\ 2m_x - 2 & 2+2m_y \end{pmatrix}. \end{aligned}$$

Das Newton-Verfahren zur Lösung dieses Problems berechnet also beginnend mit einer Startnäherung $(m_x^{(0)}, m_y^{(0)}) \in \mathbb{R}^2$ die $(k+1)$-te Iterierte gemäß

$$\begin{aligned} \begin{pmatrix} m_x^{(k+1)} \\ m_y^{(k+1)} \end{pmatrix} &:= \begin{pmatrix} m_x^{(k)} \\ m_y^{(k)} \end{pmatrix} - F'(m_x^{(k)}, m_y^{(k)})^{-1} F(m_x^{(k)}, m_y^{(k)}) \\ &= \begin{pmatrix} m_x^{(k)} \\ m_y^{(k)} \end{pmatrix} - \begin{pmatrix} 2m_x^{(k)} & 2m_y^{(k)} \\ 2m_x^{(k)} - 2 & 2+2m_y^{(k)} \end{pmatrix}^{-1} \begin{pmatrix} (m_x^{(k)})^2 + (m_y^{(k)})^2 - 25 \\ (1-m_x^{(k)})^2 + (-1-m_y^{(k)})^2 - 25 \end{pmatrix}. \end{aligned}$$

Man beobachtet das folgende typische lokale Konvergenzverhalten:

(a) Beginnend mit $(m_x^{(0)}, m_y^{(0)}) = (1,1)$ konvergiert das Newton-Verfahren gegen eine richtige Lösung $(4,3)$.

(b) Beginnend mit $(m_x^{(0)}, m_y^{(0)}) = (-1,-1)$ konvergiert das Newton-Verfahren gegen eine weitere richtige Lösung $(-3,-4)$.

(c) *Beginnend mit* $(m_x^{(0)}, m_y^{(0)}) = (0,0)$ *ist das Newton-Verfahren nicht durchführbar, da die Jacobi-Matrix*

$$F'(m_x, m_y) = \begin{pmatrix} 2m_x & 2m_y \\ 2m_x - 2 & 2 + 2m_y \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ -2 & 2 \end{pmatrix}$$

nicht invertierbar ist.

5.3 Mehrdimensionale Optimierungsprobleme

Wir skizzieren noch kurz, wie die mehrdimensionalen Differentialgleichung zur Lösung von Optimierungsproblemen mit mehreren Parametern eingesetzt wird. Sei

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) = f(x_1, \ldots, x_n)$$

das zu minimierende Zielfunktional. Zur Lösung von

$$f(x) \to \min! \quad u.d.N. \quad x \in \mathbb{R}^n$$

verwenden wir die Taylor-Formeln 1. und 2. Ordnung aus Satz 5.11.

Eine naheliegende Idee ist es (beginnend mit einer Startnäherung $x^{(0)} \in \mathbb{R}^n$), im k-ten Schritt die lineare Taylorapproximation in der aktuellen Iterierten $x^{(k)}$ zu minimieren:

$$f(x) \approx f(x^{(k)}) + \nabla f(x^{(k)})^T (x - x^{(k)}) \to \min!$$

Die lineare Taylor-Approximation ist nach unten unbeschränkt, aber nur in der Nähe von $x^{(k)}$ eine gute Approximation an f. Man gibt daher zusätzlich eine Schranke an $\left\| x - x^{(k)} \right\|$ vor. Man kann zeigen, dass dann (in der linearen Taylorapproximiation) die bestmögliche Minimierung erreicht wird, wenn

$$x - x^{(k)} = -s_k \nabla f(x^{(k)}) \quad \text{mit einem } s_k > 0.$$

Man wählt daher im k-ten Schritt

$$x^{(k+1)} = x^{(k)} - s_k \nabla f(x^{(k)})$$

mit einer Schrittweite s_k, die (wie in Abschnitt 3.5) nicht zu groß gewählte werden sollte (sonst ist die Taylorapproximation keine gute Approximation mehr), aber auch nicht zu klein (sonst benötigt das Verfahren zuviele Schritte oder kommt gar nicht mehr voran). Dieses Vorgehen heißt *Gradientenverfahren* oder auch *Verfahren des steilsten Abstiegs* (da $-\nabla f$ in der linearen Näherung die Richtung des steilsten Abstiegs ist).

Die Verwendung der Taylorapproximation 2. Ordnung führt auf die Minimierung

$$f(x) \approx f(x^{(k)}) + \nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^T \nabla^2 f(x^{(k)})(x - x^{(k)}) \to \min!$$

Unter gewissen Annahmen an $\nabla^2 f(x^{(k)})$ kann man zeigen, dass das Minimum der Taylorapproximation 2. Ordnung gegeben ist durch

$$x^{(k+1)} := x^{(k)} - (\nabla^2 f(x^{(k)}))^{-1} \nabla f(x^{(k)}).$$

Die gleiche Iterationsvorschrift erhält man durch Anwendung des mehrdimensionalen Newton-Verfahrens zur Nullstellensuche der Funktion

$$F : \mathbb{R}^n \to \mathbb{R}^n, \quad F(x) := \nabla f(x).$$

Das Verfahren heißt daher *Newtonverfahren für Optimierungsprobleme*.

Kapitel 6

Komplexe Zahlen

6.1 Motivation: Auflösung linearer Rekursionen

Wir betrachten eine reelle Folge $(a_n)_{n\in\mathbb{N}}$ die durch eine einstufige lineare Rekursion definiert ist

$$a_{n+1} = ca_n \quad \text{mit} \quad a_0 \in \mathbb{R}.$$

Diese lässt sich auch explizit schreiben als $a_n = c^n a_0$.

Wir versuchen nun eine Lösung ähnlicher Form für die mehrstufig rekursiv definierte Fibonacci-Folge

$$a_{n+2} = a_{n+1} + a_n \quad \text{für } n \in \mathbb{N}_0,$$

zu finden und machen den *Ansatz* $a_n = \lambda^n$ mit noch zu bestimmendem $\lambda \in \mathbb{R}$. Falls es eine Lösung dieser Form gibt, so erfüllt λ

$$\lambda^{n+2} = \lambda^{n+1} + \lambda^n \quad \text{und damit} \quad \lambda^2 = \lambda + 1.$$

Wir nennen $p(\lambda) := \lambda^2 - \lambda - 1$ auch das *charakteristische Polynom* der (in Nullstellenform gebrachten) Rekursionsgleichung $a_{n+2} - a_{n+1} - a_n = 0$.

Die Nullstellen von $p(\lambda) = \lambda^2 - \lambda - 1$ sind

$$\lambda_1 = \frac{1}{2} + \sqrt{\frac{5}{4}} = \frac{1+\sqrt{5}}{2} \quad \text{und} \quad \lambda_2 = \frac{1}{2} - \sqrt{\frac{5}{4}} = \frac{1-\sqrt{5}}{2}.$$

Nach Konstruktion erfüllt sowohl die Folge $a_n = \lambda_1^n$ als auch die Folge $a_n = \lambda_2^n$ jeweils die Rekursionsgleichung der Fibonacci-Folge

$$a_{n+2} = a_{n+1} + a_n.$$

Die Fibonacci-Folge beginnt jedoch mit $a_0 = 1$ und $a_1 = 1$. Dies wird von λ_1^n und λ_2^n nicht erfüllt.

Da die Rekursionsgleichung linear ist, erfüllt aber auch jede Linearkombination

$$a_n = c_1\lambda_1^n + c_2\lambda_2^n \quad \text{mit } c_1, c_2 \in \mathbb{R}$$

die Rekursionsgleichung. Wir versuchen daher c_1 und c_2 so zu wählen, dass für die Anfangswerte $a_0 = a_1 = 1$ gilt.[1].

Die Fibonacci-Folge beginnt mit $a_0 = 1$ und $a_1 = 1$. Daraus erhalten wir

$$\begin{aligned} 1 = a_0 &= c_1\lambda_1^0 + c_2\lambda_2^0 = c_1 + c_2 \\ 1 = a_1 &= c_1\lambda_1^1 + c_2\lambda_2^1 = c_1\lambda_1 + (1-c_1)\lambda_2 = \lambda_2 + c_1(\lambda_1 - \lambda_2) \end{aligned}$$

und damit

$$\begin{aligned} c_1 &= \frac{1-\lambda_2}{\lambda_1 - \lambda_2} = \frac{\frac{1}{2} + \sqrt{\frac{5}{4}}}{\sqrt{5}} = \frac{1+\sqrt{5}}{2\sqrt{5}} \\ c_2 &= 1 - c_1 = \frac{\sqrt{5}-1}{2\sqrt{5}}. \end{aligned}$$

Eine explizite Formel für die Berechnung der Fibonacci-Folge

$$1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21, \ldots$$

lautet daher (*Formel von Moivre-Binet*)

$$\begin{aligned} a_n = c_1\lambda_1^n + c_n\lambda_2^n &= \frac{1+\sqrt{5}}{2\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n + \frac{\sqrt{5}-1}{2\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n \\ &= \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right). \end{aligned}$$

Dabei ist besonders bemerkenswert, dass diese Formel auch Divisionen und sogar die irrationale Zahl $\sqrt{5} \in \mathbb{R} \setminus \mathbb{Q}$ enthält, obwohl die Fibonacci-Folge nur aus Summen natürlichen Zahlen besteht. Für jedes $n \in \mathbb{N}_0$ heben sich in der obigen Formel alle Brüche und alle Wurzeln gegenseitig auf und es ensteht immer eine natürliche Zahl.

[1] Man kann zeigen, dass die Lösungen so einer zweistufigen linearen Rekursionsgleichung einen zweidimensionalen Vektorraum bilden und dass λ_1^n und λ_2^n eine Basis dieses Vektorraums bilden. Jede Lösung lässt sich daher als so eine Linearkombination darstellen

Betrachten wir jetzt noch die zweistufige Rekursion

$$a_{n+1} = 2a_n - 2a_{n-1}, \qquad \text{für } n \in \mathbb{N}.$$

Mit dem Rekursionsbeginn $a_1 = a_0 = 1$ erhalten wir die Folge ganzer Zahlen

$$1,\ 1,\ 0,\ -2,\ -4,\ -4,\ 0,\ 8, \ldots \in \mathbb{Z}.$$

Als charakteristische Polynom der Rekursionsgleichung $a_{n+1} - 2a_n + 2a_{n-1} = 0$ ergibt sich $p(\lambda) = \lambda^2 - 2\lambda + 2$ und die Nullstellen müssten erfüllen

$$\lambda_1 = 1 + \sqrt{-1} \quad \text{und} \quad \lambda_2 = 1 - \sqrt{-1}.$$

Eine Zahl $\sqrt{-1}$, die also mit sich selbst multipliziert -1 ergibt, kann es nicht geben. Das Polynom $p(\lambda)$ besitzt keine Nullstellen (in $\mathbb{R}$).

Überraschenderweise führt dieses Vorgehen trotzdem zum Erfolg, wenn man dieses Problem einfach wie folgt ignoriert. Wir tun so, als gäbe es eine Zahl

$$\mathrm{i} := \sqrt{-1}$$

(d.h. $\mathrm{i}^2 = -1$) und rechnen damit so, wie wir es von den reellen Zahlen gewohnt sind.

Dann erhalten wir als allgemeine Lösung unserer zweistufigen Rekursionsgleichung $a_{n+1} = 2a_n - 2a_{n-1}$, dass

$$a_n = c_1(1+\mathrm{i})^n + c_2(1-\mathrm{i})^n$$

und aus $a_0 = 1 = a_1$ erhalten wir

$$\begin{aligned} 1 = a_0 &= c_1(1+\mathrm{i})^0 + c_2(1-\mathrm{i})^0 = c_1 + c_2 \\ 1 = a_1 &= c_1(1+\mathrm{i})^1 + c_2(1-\mathrm{i})^1 = c_1 + c_2 + (c_1 - c_2)\mathrm{i} = 1 + (1-2c_2)\mathrm{i} \end{aligned}$$

und damit $(1-2c_2)\mathrm{i} = 0$, also $c_2 = \frac{1}{2}$ und damit $c_1 = \frac{1}{2}$. Es gilt also

$$a_n = \frac{1}{2}\left((1+\mathrm{i})^n + (1-\mathrm{i})^n\right)$$

So erhalten wir tatsächlich $a_0 = 1$, $a_1 = 1$,

$$\begin{aligned} a_2 &= \frac{1}{2}\left((1+\mathrm{i})^2 + (1-\mathrm{i})^2\right) = \frac{1}{2}\left(1 + 2\mathrm{i} + \mathrm{i}^2 + 1 - 2\mathrm{i} + \mathrm{i}^2\right) = 0 \\ a_3 &= \frac{1}{2}\left((1+\mathrm{i})^3 + (1-\mathrm{i})^3\right) = \frac{1}{2}\left(1 + 3\mathrm{i} + 3\mathrm{i}^2 + 1 + 1 - 3\mathrm{i} + 3\mathrm{i}^2 - 1\right) = -2. \end{aligned}$$

So fortfahrend kürzt sich die (mit unserem reellen Zahlenbegriff) gar nicht existierende Zahl i immer weg und die Formel $\frac{1}{2}\left((1+\mathrm{i})^n+(1-\mathrm{i})^n\right)$ liefert genau unsere rekursiv definierte Folge

$$1,\ 1,\ 0,\ -2,\ -4,\ldots$$

Die Einführung der *imaginären* Zahl i ermöglicht uns also die Lösung eines *realen* Problems.

Die naive Verwendung („wir tun so, als gäbe es i“) von $\mathrm{i}=\sqrt{-1}$ in den bekannten Rechenregeln kann jedoch auch zu Widersprüchen führen. So führt beispielsweise

$$-1=\mathrm{i}\cdot\mathrm{i}=\sqrt{-1}\sqrt{-1}=\sqrt{(-1)\cdot(-1)}=\sqrt{1}=1$$

zu der offensichtlich falsche Aussage $-1=1$. Es ist daher wichtig, i auf eine mathematisch rigorose, widerspruchsfreie Weise einzuführen, aus der hervorgeht, welche Rechenregeln für i angewendet werden dürfen.

6.2 Die komplexen Zahlen

Um widerspruchsfrei mit der Zahl $\mathrm{i}=\sqrt{-1}$ rechnen zu können, gehen wir wie folgt vor. Wir definieren eine Menge $\mathbb{C}$ zusammen mit zwei Verknüpfungen (Multiplikation und Addition), so dass $\mathbb{C}$ ein Körper wird, in dem wir sowohl $\mathbb{R}$ wiederfinden, als auch die Gleichung $\mathrm{i}^2=-1$ eine Lösung besitzt.

Im letzten Abschnitt kamen Zahlen der Form $a+b\mathrm{i}$ mit $a,b\in\mathbb{R}$ vor und unsere naive Addition und Multiplikation („rechnen so, wie wir es gewohnt sind”) führte zu

$$\begin{aligned}a_1+b_1\mathrm{i}+a_2+b_2\mathrm{i}&=(a_1+a_2)+(b_1+b_2)\mathrm{i}\\(a_1+b_1\mathrm{i})(a_2+b_2\mathrm{i})&=a_1a_2+b_1a_2\mathrm{i}+a_1b_2\mathrm{i}+b_1b_2\mathrm{i}^2\\&=(a_1a_2-b_1b_2)+(b_1a_2+a_1b_2)\mathrm{i}.\end{aligned}$$

Wir wählen daher als Menge Zweiertupel reeller Zahlen

$$\mathbb{C}:=\mathbb{R}^2=\{(a,b):\ a,b\in\mathbb{R}\}$$

und definieren darauf die komponentenweise Addition

$$+:\ \mathbb{C}\to\mathbb{C},\quad (a_1,b_1)+(a_2,b_2):=(a_1+a_2,b_1+b_2)$$

und die Multiplikation nach folgender Formel

$$*:\ \mathbb{C}\to\mathbb{C},\quad (a_1,b_1)*(a_2,b_2):=(a_1a_2-b_1b_2,b_1a_2+a_1b_2).$$

Man rechnet leicht nach, dass $\mathbb{C}$ mit diesen Verknüpfungen alle Körperaxiome erfüllt, wobei das Nullelement $(0,0)$ und das Einselement $(1,0)$ ist. Die Teilmenge

$$\{(a,0),\ a \in \mathbb{R}\} \subset \mathbb{C}$$

kann man in dem Sinne mit $\mathbb{R}$ identifizieren, dass Additionen, Multiplikation und Identitätsvergleiche in $\mathbb{R}$ das gleiche Ergebnis wie in dieser Teilmenge liefern.

Mit $\mathrm{i} := (0,1) \in \mathbb{C}$ ergibt sich außerdem

$$\mathrm{i}^2 = (0,1) * (0,1) = (-1,0),$$

was wir mit obiger Identifikation $-1 \in \mathbb{R}$ entspricht. Damit gilt auch

$$(a,b) = (a,0)(1,0) + (b,0)(0,1) = a + b\mathrm{i}$$

und

$$\mathbb{C} = \{a + b\mathrm{i} :\ a,b \in \mathbb{R}\}.$$

Die Elemente von $\mathbb{C}$ heißen *komplexe Zahlen*.

Im Körper der komplexen Zahlen besitzt das Polynom x^2+1 die Nullstellen i und i. Man kann zeigen (*Fundamentalsatz der Algebra*), dass jedes nichtkonstanten Polynom in $\mathbb{C}$ einen Nullstelle besitzt und daraus folgern, dass jedes Polynom $p \in \Pi_n$ (mit reellen oder komplexen Koeffizienten) mit Höchstkoeffizient 1 faktorisiert werden kann in

$$p(x) = (x - x_1)(x - x_2) \cdots (x - x_n)$$

und $p(x) = 0$ genau dann wenn $x \in \{x_1, \ldots, x_n\}$. Dabei können mehrere x_k übereinstimmen (mehrfache Nullstelle). In diesem Sinne besitzt im Komplexen jedes Polynom vom Grad n genau n (mit Vielfachheit gezählte) Nullstellen.

Wir führen noch einige wichtige Begriffe zum Umgang mit komplexen Zahlen ein. Für $z = a + b\mathrm{i} \in \mathbb{C}$ heißt $\mathrm{Re}(z) = a$ der *Realteil* und $\mathrm{Im}(z) = b$ der *Imaginärteil*. i heißt auch *komplexe Einheit*. Komplexe Zahlen können veranschaulicht werden in einem zweidimensionalen Koordinatensystem mit *reeller Achse* und *imaginärer Achse*, vgl. das in der Vorlesung gemalte Bild.

Zu einer komplexen Zahl $z = a + b\mathrm{i} \in \mathbb{C}$ definiert man

$$\text{den } \textit{Betrag}\ |z| = \sqrt{a^2 + b^2} \quad \text{und die } \textit{Konjugierte}\ \bar{z} = a - b\mathrm{i}.$$

Offenbar gilt

$$\mathrm{Re}(z) = \frac{1}{2}(z + \bar{z}), \quad \mathrm{Im}(z) = \frac{1}{2\mathrm{i}}(z - \bar{z}) \quad \text{und} \quad |z|^2 = z\bar{z}.$$

Außerdem rechnet man leicht nach, dass für $z_1, z_2 \in \mathbb{C}$ gilt

$$|z_1 + z_2| \leq |z_1| + |z_2|, \qquad \overline{z_1 + z_2} = \overline{z_1} + \overline{z_2},$$
$$|z_1 z_2| = |z_1||z_2|, \qquad \overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}.$$

Das *Argument* (auch: *Phase*) einer komplexen Zahl ist definiert durch den Winkel zwischen der Verbindungslinie von $a + bi$ zum Nullpunkt und der positiven reellen Achse, vgl. das in der Vorlesung gemalte Bild. Mit der Umkehrfunktion $\arctan : \mathbb{R} \to]-\pi/2, \pi/2[$ der Funktion $\tan(x) = \frac{\sin(x)}{\cos(x)}$ ergibt sich unter Betrachtung aller möglicher Fälle für $z \neq 0$

$$\arg(z) = \begin{cases} \arctan(b/a) & \text{für } a > 0, \\ \arctan(b/a) + \pi & \text{für } a < 0,\ b > 0, \\ \arctan(b/a) - \pi & \text{für } a < 0,\ b < 0, \\ \frac{\pi}{2} & \text{für } a = 0,\ b > 0, \\ -\frac{\pi}{2} & \text{für } a = 0,\ b < 0. \end{cases}$$

Diese Funktion wird auch als $\arctan 2(a,b)$ oder $\mathrm{atan2}(a,b)$ bezeichnet. Auf Werte der negativen reellen Achse bis einschließlich Null kann die Funktion offenbar nicht stetig fortgesetzt werden. Wir verwenden in dieser Vorlesung die häufig verwendete Konvention[2]

$$\arg(0) := 0 \quad \text{und} \quad \arg(a) := \pi \quad \text{für } a < 0.$$

Mit dieser Konvention (und auch den meisten anderen) gilt für $z = a + ib \in \mathbb{C}$ die sogenannte *Polardarstellung*

$$a = |z|\cos(\varphi), \quad b = |z|\sin(\varphi) \quad \text{mit } \varphi := \arg(z),$$

vgl. die in der Vorlesung gemalte Skizze.

6.3 Die komplexe Exponentialfunktion

Man kann zeigen, dass die in Satz 1.30 eingeführte Exponentialreihe

$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}$$

[2] Es sind in der Fachliteratur und in Programmiersprachen auch andere Definitionen und Implementationen üblich. Bei der Verwendung von Zahldarstellungen mit 0 und -0 werden diesen sogar manchmal verschiedene Argumente zugeordnet, z.B. ist im IEEE 754-Standard definiert, dass $\mathrm{atan2}(0,-0) = \pi$ und $\mathrm{atan2}(-0,-0) = -\pi$.

auch für alle $z \in \mathbb{C}$ konvergiert. So definiert man die *komplexe Exponentialfunktion*

$$\exp : \ \mathbb{C} \to \mathbb{C}, \quad z \mapsto \exp(z) =: \mathrm{e}^z.$$

Auch im Komplexen gilt die Funktionalgleichung der Exponentialgleichung

$$\mathrm{e}^{z_1+z_2} = \mathrm{e}^{z_1}\mathrm{e}^{z_2} \qquad \text{für alle } z_1, z_2 \in \mathbb{C}$$

und, da die Konjugation mit Summen- und Produktbildung vertauscht, gilt außerdem

$$\overline{\mathrm{e}^z} = \mathrm{e}^{\bar{z}}.$$

Für $x \in \mathbb{R}$ erhalten wir durch Zerlegung der Exponentialreihe in Summanden mit geraden und ungerade Indizes

$$\begin{aligned}\mathrm{e}^{\mathrm{i}x} &= \sum_{k=0}^{\infty} \frac{(\mathrm{i}x)^k}{k!} = \sum_{k=0}^{\infty} \frac{(\mathrm{i}x)^{2k}}{(2k)!} + \sum_{k=0}^{\infty} \frac{(\mathrm{i}x)^{2k+1}}{(2k+1)!} \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} + \mathrm{i} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}.\end{aligned}$$

Da dies die Sinus- und Kosinusreihen aus Definition und Satz 1.32 sind, erhalten wir so die *Eulersche Formel*

$$\mathrm{e}^{\mathrm{i}x} = \cos(x) + \mathrm{i}\sin(x) \qquad \text{für alle } x \in \mathbb{R}.$$

Die Werte von $\exp(\mathrm{i}x)$ liegen in der komplexen Ebene also auf einem Kreis mit Radius 1 um den Nullpunkt, es gilt

$$|\mathrm{e}^{\mathrm{i}x}| = 1 \qquad \text{für alle } x \in \mathbb{R}$$

und die Polardarstellung einer komplexen Zahl lässt sich schreiben als

$$z = a + b\mathrm{i} = |z|\cos(\varphi) + \mathrm{i}|z|\sin(\varphi) = |z|\mathrm{e}^{\mathrm{i}\varphi}$$

mit $\varphi := \arg(z)$.

Die Multiplikation zweier komplexen Zahlen lässt sich damit geometrisch interpretieren. Für $z_1, z_2 \in \mathbb{C}$ mit $\varphi_1 := \arg(z_1)$ und $\varphi_2 := \arg(z_2)$ folgt aus der Multiplikativität des komplexen Betrags und der Funktionalgleichung der Exponentialfunktion

$$z_1 z_2 = |z_1|\mathrm{e}^{\mathrm{i}\varphi_1}|z_2|\mathrm{e}^{\mathrm{i}\varphi_2} = |z_1 z_2|\mathrm{e}^{\mathrm{i}(\varphi_1+\varphi_2)}.$$

Beim Produkt zweier komplexer Zahlen multiplizieren sich also ihre Beträge und die Phasen addieren sich.

Beispiel 6.1 (Lösung einer linearen Differentialgleichung)
Wir suchen eine Funktion $y: \mathbb{R} \to \mathbb{R}$ *mit der Eigenschaft*

$$y''(x) = -2y'(x) - 2y(x) \quad \text{für alle } x \in \mathbb{R}.$$

So eine Differentialgleichung (DGL) tritt etwa bei der Beschreibung eines gedämpften Pendels auf, bei dem zu jedem Zeitpunkt $x \in \mathbb{R}$ *eine Auslenkung* $y(x)$ *zu einer entgegengerichteten Beschleunigung* $y''(x)$ *führt und auch die Geschwindigkeit* $y'(x)$ *wegen Reibungseffekten einen entgegengerichteten (also bremsenden) Einfluss auf die Beschleunigung hat. Wir fordern außerdem, dass*

$$y(0) = 1 \quad \text{und} \quad y'(0) = 0,$$

d.h. das Pendel soll zum $x = 0$ *ausgelenkt, aber ohne Anfangsgeschwindigkeit sein.*

Wir machen den Ansatz, eine Lösung der Form $y(x) = \mathrm{e}^{\lambda x}$ *zu finden. Dies erfüllt die Differentialgleichung genau dann, wenn*

$$0 = y''(x) + 2y'(x) + 2y(x) = \lambda^2 \mathrm{e}^{\lambda x} + 2\lambda \mathrm{e}^{\lambda x} + 2\mathrm{e}^{\lambda x} = (\lambda^2 + 2\lambda + 2)\mathrm{e}^{\lambda x}$$

also genau dann wenn

$$\lambda^2 + 2\lambda + 2 = 0.$$

Diese Gleichung hat keine reellen Lösungen, sie besitzt jedoch die komplexen Lösungen

$$\lambda_1 = -1 + \mathrm{i} \quad \text{und} \quad \lambda_2 = -1 - \mathrm{i}.$$

Damit erhalten wir zwei Lösungen der DGL

$$y_1(x) = \mathrm{e}^{\lambda_1 x} \quad \text{und} \quad y_2(x) = \mathrm{e}^{\lambda_2 x}.$$

Man kann zeigen, dass die Lösungen der DGL genau die Linearkombinationen dieser beiden Lösungen sind. Die allgemeine Lösung der DGL lautet daher

$$y(x) = C_1 \mathrm{e}^{\lambda_1 x} + C_2 \mathrm{e}^{\lambda_2 x}. = C_1 \mathrm{e}^{-x+\mathrm{i}x} + C_2 \mathrm{e}^{-x-\mathrm{i}x}$$

mit $C_1, C_2 \in \mathbb{C}$. *Aus der Anfangsbedingung* $y(0) = 1$ *erhalten wir*

$$1 = y(0) = C_1 \mathrm{e}^{\lambda_1 0} + C_2 \mathrm{e}^{\lambda_2 0} = C_1 + C_2 \quad \text{und damit} \quad C_2 = 1 - C_1.$$

Aus der weiteren Anfangsbedingung $y'(0) = 0$ *erhalten wir*

$$\begin{aligned} 0 = y'(0) &= C_1\lambda_1 \mathrm{e}^{\lambda_1 0} + C_2\lambda_2 \mathrm{e}^{\lambda_2 0} = \lambda_1 C_1 + \lambda_2 C_2 = C_1\lambda_1 + (1 - C_1)\lambda_2 \\ &= -C_1 + \mathrm{i}C_1 - 1 - \mathrm{i} + C_1 + \mathrm{i}C_1 = 2\mathrm{i}C_1 - 1 - \mathrm{i} \end{aligned}$$

und damit

$$C_1 = \frac{1+\mathrm{i}}{2\mathrm{i}} = \frac{1-\mathrm{i}}{2} \quad \textit{und} \quad C_2 = 1 - C_1 = \frac{1+\mathrm{i}}{2}.$$

Die Lösung der DGL mit den richtigen Anfangswerten ist also

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

Wie schon bei den linearen Rekursionsgleichungen hilft uns also auch hier das Rechnen mit komplexen Zahlen, eine reellwertige Lösung eines reellen Problems zu finden.