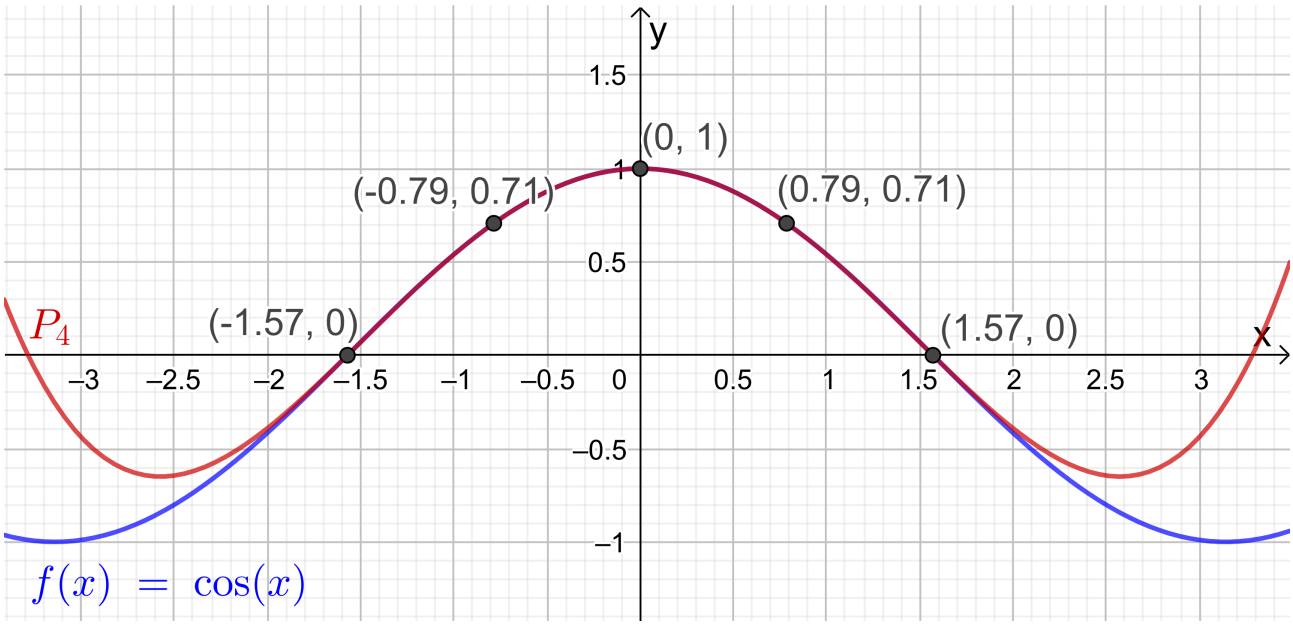
UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Modellieren und Anwendungen:

Numerische Analysis

Kerstin Hesse

Universität Paderborn, Sommersemester 2024



Funktion $\cos(x)$ und ihr Interpolationspolynom P_4 bzgl. der fünf Datenpunkte.

Dieses Skript wurde in seiner ersten Version für das Sommersemester 2020 erstellt. Für das Sommersemester 2021 wurde das Kapitel 5 über numerische Intregration ergänzt. Als Vorlage diente dabei (für alle Kapitel) insbesondere das englischsprachige Buch „Elementary Numerical Analysis“ von Kendall Atkinson und Weimin Han (John Wiley & Sons, Inc., 2004, 3. Auflage), dessen ersten fünf Kapiteln dieses Skript (mit verschiedenen Auslassungen) weitgehend folgt.

Paderborn, März 2024

Kerstin Hesse

Einleitung

Bei dem Thema **Numerik** geht es darum, wie man mathematische Probleme (angenähert) mit einem Computer löst. Hier sind zwei Beispiele, an denen deutlich wird, worum es geht und warum dieses interessant ist:

Beispiel 1: Lösen linearer Gleichungssysteme. Angenommen, man hat ein lineares Gleichungssystem mit 100.000 Unbekannten. Sicher möchte man dieses nicht per Hand lösen, sondern ein geeignetes numerisches Verfahren (einen Algorithmus) auf einem Computer dazu nutzen. Dabei erhält man als Ergebnis oft nur eine Annäherung an die exakte Lösung, denn es treten einerseits Rundungsfehler auf und andererseits wird man zur Lösung oft kein direktes sondern ein iteratives Verfahren verwenden (dieses berechnet eine Folge von Näherungen der Lösung), welches man abbricht, wenn die Lösung hinreichend gut angenähert worden ist.

Beispiel 2: Funktion finden, die einen Datensatz beschreibt. Angenommen, die Temperatur y wurde an einem bestimmten Ort stündlich ein Jahr lang gemessen. Dann erhalten wir $365 \cdot 24 = 8.760$ Messdaten $(t_i; y_i)$, $1 = 1, 2, \ldots, 8.760$, wobei y_i die zum Zeitpunkt t_i gemessene Temperatur ist. Man möchte nun gerne eine Funktion $y(t)$ bestimmten, deren Graph (genau oder auch nur angenähert) durch alle Datenpunkte $(t_i; y_i)$, $i = 1, 2, \ldots, 8.760$, geht. Es stellt sich die Frage, wie gut diese Funktion die Temperatur $y(t)$ zu anderen Zeitpunkten t beschreibt.

Das erste Bespiel gehört in den Bereich der **Numerischen Linearen Algebra**, und das zweite Beispiel gehört in den Bereich der **Numerischen Analysis**. In dieser Vorlesung werden Themen aus der Numerischen Analysis behandelt. Neben dem Thema **Interpolation und Approximation** (siehe Beispiel 2 oben) werden wir uns mit **numerischer Integration** und sehr intensiv mit dem Thema der **Nullstellenbestimmung** bei Funktionen beschäftigen.

Dieses detailliert ausgearbeitete Skript können (und sollten) Sie wie ein Lehrbuch verwenden. Es folgt der Vorlesung ganz genau und enthält häufig zusätzliche Erklärungen und weitere Beispiele.

Ich freue mich auf Ihre Teilnahme an der „Numerischen Analysis“!

Kerstin Hesse März 2024

Literaturverzeichnis

Bei der Erstellung dieses Skripts wurde die unten aufgelistete Literatur verwendet:

[1] Kendall Atkinson, Weimin Han: Elementary Numerical Analysis (3. Auflage). John Wiley & Sons, Inc., 2004.

[2] Kerstin Hesse: Mathematik für Chemiker. Vorlesungsskript, Universität Paderborn, 2019.

[3] Kerstin Hesse: Höhere Mathematik A für Elektrotechniker. Vorlesungsskript, Universität Paderborn, 2018.

[4] Hans Rudolf Schwarz: Numerische Mathematik (3. Auflage). B. G. Teubner, Stuttgart, 1993.

Inhaltsverzeichnis

*Dieses Teilkapitel wird im Sommersemester 2024 nicht behandelt und ist damit auch nicht klausurrelevant.

*Dieses Teilkapitel ist nicht klausurrelevant.

KAPITEL 1

Taylor-Polynome

In Teilkapitel 1.1 wiederholen wir einige wichtige Fakten über die **Ableitung** und lernen den **Mittelwertsatz der Differentialrechnung** kennen. In Teilkapitel 1.2 wird das **Taylor-Polynom vom Grad** n mit dem Entwicklungspunkt x_0 als einfachste Näherung einer hinreichend oft differenzierbaren Funktion in der Nähe eines Punktes x_0 hergeleitet. In Teilkapitel 1.3 lernen wir den **Satz von Taylor** kennen. Dieser liefert Informationen darüber, wie gut eine hinreichend oft differenzierbare Funktion durch seine Taylor-Polynome mit dem Entwicklungspunkt x_0 angenähert wird. In Teilkapitel 1.4 interessieren wir uns für die **effiziente Berechnung** von Polynomen, und in Teilkapitel 1.5 werden wir als eine Anwendung des Taylor-Polynoms mit diesem **Integrale einfach angenähert berechnen**.

Das Taylor-Polynom einer hinreichend oft differenzierbaren Funktion f mit dem Entwicklungspunkt x_0 als **einfachste Näherung** (oder **einfachste Approximation**) der Funktion f in der Nähe von x_0 illustriert viele **Problemstellungen, Fragen und Ideen**, mit denen wir uns in diesem Kurs beschäftigen wollen, und bietet daher einen guten Einstieg in die Numerische Analysis.

1.1 Ableitung und Mittelwertsatz

In diesem Teilkapitel wiederholen wir einige wichtige Fakten über die Ableitung einer differenzierbaren Funktion und lernen den Mittelwertsatz der Differentialrechnung kennen.

Sei I ein offenes Intervall. Ist eine Funktion $f : I \to \mathbb{R}$ auf dem Intervall I

differenzierbar, so gibt die **Ableitung $f'(x_0)$ im Punkt** $x_0 \in I$ die **Steigung der Tangente an den Graphen im Punkt** $\big(x_0; f(x_0)\big)$ an (siehe Abbildung 1.1).

Angenähert wird die Ableitung $f'(x_0)$ durch die Steigung der Sekante durch die Punkte $\big(x_0; f(x_0)\big)$ und $\big(x_0 + h; f(x_0 + h)\big)$ mit kleinem h (vgl. Abbildung 1.1):

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0)}{(x_0 + h) - x_0} = \frac{f(x_0 + h) - f(x_0)}{h}\,.$$

Bildet man den Grenzwert für $h \to 0$, so erhält man die Steigung der Tangente

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}\,.$$

Die Ableitungen der folgenden Funktionen sollten Sie unbedingt kennen:

Definitionsmenge	Funktion	Ableitung
$\mathbb{R}$	$f(x) = c$	$f'(x) = 0$
$\mathbb{R}$	$f(x) = x^n$ mit $n \in \mathbb{N}$	$f'(x) = n\,x^{n-1}$
$\mathbb{R} \setminus \{0\}$	$f(x) = x^n$ mit $n \in \mathbb{Z} \setminus \mathbb{N}_0$	$f'(x) = n\,x^{n-1}$
$]0; \infty[$	$f(x) = x^r$ mit $r \in \mathbb{R} \setminus \mathbb{Z}$	$f'(x) = r\,x^{r-1}$
$\mathbb{R}$	$f(x) = e^x$	$f'(x) = e^x$
$]0; \infty[$	$f(x) = \ln(x)$	$f'(x) = \dfrac{1}{x}$
$\mathbb{R}$	$f(x) = \sin(x)$	$f'(x) = \cos(x)$
$\mathbb{R}$	$f(x) = \cos(x)$	$f'(x) = -\sin(x)$

Für die Ableitung gelten die folgenden Rechenregeln:

Satz 1.1. (Rechenregeln für die Ableitung)
Sei I ein offenes Intervall. Seien $f : I \to \mathbb{R}$ und $g : I \to \mathbb{R}$ differenzierbar. Dann gelten die folgenden Rechenregeln:

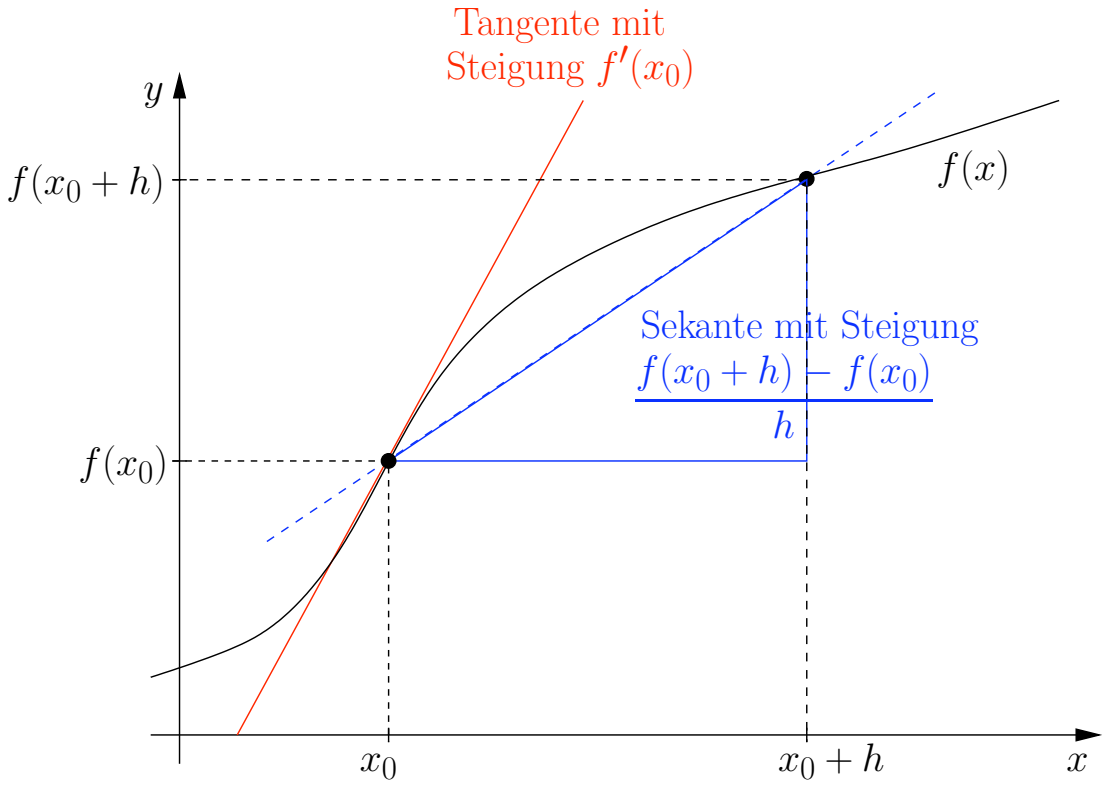


Abb. 1.1: Veranschaulichung der Ableitung: Die Steigung der Tangente $f'(x_0)$ lässt sich über die Steigung der Sekanten annähern.

(1) Für $\alpha \in \mathbb{R}$ ist $\alpha \cdot f = \alpha f$ differenzierbar, und es gilt

$$(\alpha f)'(x) = \alpha f'(x).$$

(2) $f + g$ ist differenzierbar, und es gilt

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

(3) $f \cdot g$ ist differenzierbar, und es gilt

$$(f \cdot g)'(x) = (f\,g)'(x) = f'(x)\,g(x) + f(x)\,g'(x). \qquad \textit{(**Produktregel**)}$$

(4) Ist $g(x) \neq 0$ für alle $x \in I$, so ist f/g differenzierbar, und es gilt

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)\,g(x) - f(x)\,g'(x)}{\big[g(x)\big]^2}. \qquad \textit{(**Quotientenregel**)}$$

Satz 1.2. (Kettenregel)
Seien I, J offene Intervalle. Seien $f : I \to \mathbb{R}$ und $g : J \to \mathbb{R}$ differenzierbare Funktionen mit $f(I) = \{f(x) \ : \ x \in I\} \subseteq J$. Dann ist die Verkettung $g \circ f : I \to \mathbb{R}$, $(g \circ f)(x) = g\big(f(x)\big)$, differenzierbar, und es gilt

$$(g \circ f)'(x) = \underbrace{g'\big(f(x)\big)}_{\substack{\text{äußere} \\ \text{Ableitung}}} \ \underbrace{f'(x)}_{\substack{\text{innere} \\ \text{Ableitung}}} .$$

Hier sind einige Beispiele zur Anwendung dieser Rechenregeln.

Beispiel 1.3. (Rechenregeln für die Ableitung)

(a) Nach Satz 1.1 (1) hat $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 13 \sin(x)$, die Ableitung

$$f'(x) = 13 \sin'(x) = 13 \cos(x).$$

(b) Nach Satz 1.1 (2) hat $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \cos(x) + \sin(x)$, die Ableitung

$$f'(x) = \cos'(x) + \sin'(x) = -\sin(x) + \cos(x).$$

(c) Nach Satz 1.1 (3) hat $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \cos(x) \sin(x)$, die Ableitung

$$\begin{aligned} f'(x) &= \cos'(x) \sin(x) + \cos(x) \sin'(x) \\ &= -\sin(x) \sin(x) + \cos(x) \cos(x) = -\sin^2(x) + \cos^2(x). \end{aligned}$$

Dabei sind $\sin^2(x) = \big(\sin(x)\big)^2$ und $\cos^2(x) = \big(\cos(x)\big)^2$.

(d) Nach Satz 1.1 (4) hat $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{x^3 + 2\,x + 1}{x^2 + 1}$, die Ableitung

$$\begin{aligned} f'(x) &= \frac{(x^3 + 2\,x + 1)'\,(x^2 + 1) - (x^3 + 2\,x + 1)\,(x^2 + 1)'}{(x^2 + 1)^2} \\ &= \frac{(3\,x^2 + 2)\,(x^2 + 1) - (x^3 + 2\,x + 1)\,2\,x}{(x^2 + 1)^2} . \end{aligned}$$

(e) Nach Satz 1.2 hat $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sin(3\,x^2 + x)$, die Ableitung

$$f'(x) = \sin'(3\,x^2 + x) \cdot (3\,x^2 + x)' = \cos(3\,x^2 + x) \cdot (6\,x + 1).$$

Auf dem ersten Übungszettel werden wir das Berechnen von Ableitungen mit weiteren Beispielen wiederholen. ♠

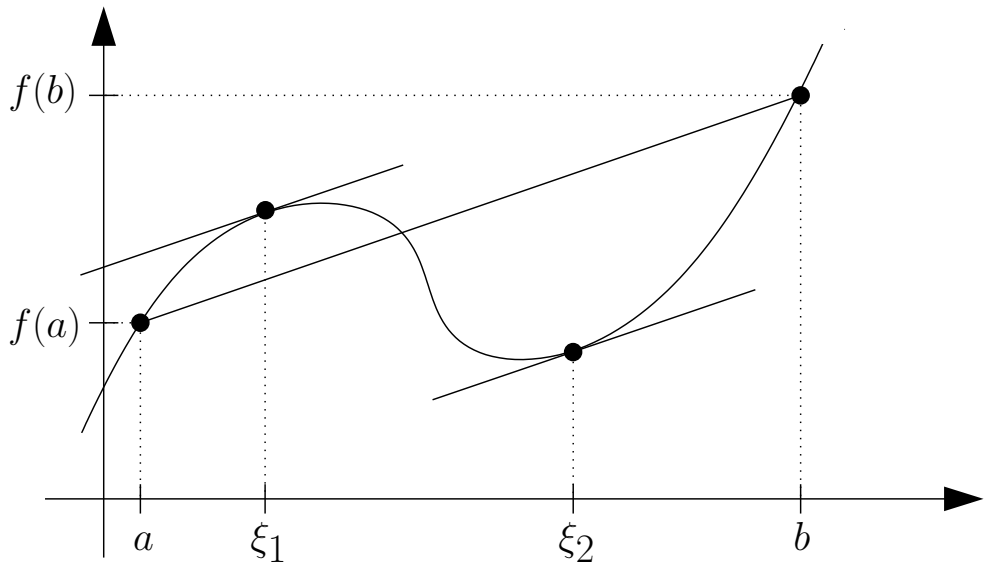


Abb. 1.2: Veranschaulichung des Mittelwertsatzes der Differentialrechnung.

In dieser Vorlesung brauchen wir wiederholt den Mittelwertsatz der Differentialrechnung, der nun eingeführt und erklärt wird.

Satz 1.4. (Mittelwertsatz der Differentialrechnung (MWS))
Seien $a, b \in \mathbb{R}$ mit $a < b$. Die Funktion f sei stetig in $[a; b]$ und differenzierbar in $]a; b[$. Dann existiert ein $\xi \in]a; b[$ mit

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}\,. \tag{1.1}$$

Veranschaulichung des Mittelwertsatzes der Differentialrechnung: Der Quotient $\frac{f(b)-f(a)}{b-a}$ auf der rechten Seite von (1.1) ist die Steigung der Sekante durch $\big(a; f(a)\big)$ und $\big(b; f(b)\big)$ (also der Geraden durch die Punkte $\big(a; f(a)\big)$ und $\big(b; f(b)\big)$). Die Ableitung $f'(\xi)$ ist die Steigung der Tangente an f im Punkt ξ. Also besagt (1.1), dass man mindestens einen Punkt $\xi \in]a; b[$, findet in dem die Tangente parallel zu dieser Sekante ist. In dem Beispiel in Abbildung 1.2 gibt es sogar zwei Punkte $\xi_1, \xi_2 \in]a; b[$ mit $f'(\xi_1) = f'(\xi_2) = \frac{f(b)-f(a)}{b-a}$.

Zu beachten ist, dass der Mittelwertsatz der Differentialrechnung (siehe Satz 1.4) nur eine **Existenzaussage** ist, denn er garantiert uns nur die Existenz eines Punktes $\xi \in]a; b[$ mit der Eigenschaft (1.1). Er gibt uns aber auch für eine konkrete Funktion und ein konkretes Intervall $]a; b[$ keinerlei Information darüber, wo der Punkt $\xi \in]a; b[$ mit der Eigenschaft (1.1) in $]a; b[$ liegt.

Der Mittelwertsatz der Differentialrechnung ist daher besonders für theoretische Überlegungen nützlich. Insbesondere kann man mit dem Mittelwertsatz der Diffe-

rentialrechnung auch nützliche Abschätzungen für konkrete Funktionen beweisen (siehe Beispiel 1.5 unten). Weiter benötigt man den Mittelwertsatz der Differentialrechnung, um das Wachstumsverhaltens einer differenzierbaren Funktion mit Hilfe der Ableitung zu charakterisieren (siehe Satz 1.6 weiter unten).

Beispiel 1.5. (Mittelwertsatz der Differentialrechnung)
Es gilt $|\sin(x)| \leq |x|$ für alle $x \in \mathbb{R}$, denn:

- *Fall 1:* Sei $x = 0$. Hier ist $|\sin(0)| = 0 \leq |0|$.
- *Fall 2:* Sei $x \neq 0$. Nach dem Mittelwertsatz der Differentialrechnung existiert ein ξ zwischen x und 0 mit

$$\cos(\xi) = \sin'(\xi) = \frac{\sin(x) - \sin(0)}{x - 0} = \frac{\sin(x)}{x}\,.$$

Also gilt durch Multiplizieren auf beiden Seiten mit x

$$\cos(\xi) \cdot x = \sin(x) \qquad \Longleftrightarrow \qquad \sin(x) = \cos(\xi) \cdot x.$$

Wir wenden den Absolutbetrag auf beiden Seiten an und nutzen danach $|\cos(x)| \leq 1$ für alle $x \in \mathbb{R}$ aus:

$$|\sin(x)| = |\cos(\xi) \cdot x| = \underbrace{|\cos(\xi)|}_{\leq 1} \cdot |x| \leq |x|.$$

Also folgt

$$|\sin(x)| \leq |x| \qquad \text{für alle } x \neq 0.$$

Aus Fall 1 und Fall 2 zusammen erhalten wir

$$|\sin(x)| \leq |x| \qquad \text{für alle } x \in \mathbb{R}.$$

Damit ist die Abschätzung bewiesen. ♠

Aus dem Mittelwertsatz der Differentialrechnung (siehe Satz 1.4) folgen Aussagen über das Monotonieverhalten einer differenzierbaren Funktion.

Satz 1.6. (Monotonieverhalten und erste Ableitung)
Seien I ein offenes Intervall und $f : I \to \mathbb{R}$ differenzierbar. Dann gelten:

(1) $f'(x) = 0$ *für alle* $x \in I$. $\Longleftrightarrow$ f *ist* ***konstant*** *auf* I.

(2) $f'(x) \geq 0$ *für alle* $x \in I$. $\Longleftrightarrow$ f *ist* ***monoton wachsend*** *auf* I.

(3) $f'(x) \leq 0$ *für alle* $x \in I$. $\iff$ f *ist* ***monoton fallend*** *auf* I.

(4) $f'(x) > 0$ *für alle* $x \in I$. $\underset{\not\Longleftarrow}{\Longrightarrow}$ f *ist* ***streng monoton wachsend*** *auf* I.

(5) $f'(x) < 0$ *für alle* $x \in I$. $\underset{\not\Longleftarrow}{\Longrightarrow}$ f *ist* ***streng monoton fallend*** *auf* I.

Statt „monoton wachsend“ ist auch die Bezeichnung „monoton steigend“ üblich.

Die durchgestrichenen Folgepfeile $\not\Longleftarrow$ in Satz 1.6 (4) und (5) bedeuten, dass die Folgerung mit $\Longleftarrow$ nicht gilt.

Beispiel 1.7. (Anwendung von Satz 1.6)

(a) $f : \mathbb{R} \to \mathbb{R}$, $f(x) := x^2$ $\quad\Longrightarrow\quad$ $f'(x) = 2\,x \begin{cases} < 0 & \text{für } x < 0, \\ = 0 & \text{für } x = 0, \\ > 0 & \text{für } x > 0 \end{cases}$

In $]-\infty; 0[$ ist f streng monoton fallend.

In $]0; \infty[$ ist f streng monoton wachsend.

(b) $f : \mathbb{R} \to \mathbb{R}$, $f(x) := x^3$ $\quad\Longrightarrow\quad$ $f'(x) = 3\,x^2$

Da $f'(x) = 3\,x^2 \geq 0$ für alle $x \in \mathbb{R}$ ist, ist f in ganz $\mathbb{R}$ monoton wachsend.

Aber: Obwohl in Beispiel (b) $f'(0) = 0$ gilt, ist f in ganz $\mathbb{R}$ sogar streng monoton wachsend, denn für alle $x_1, x_2 \in \mathbb{R}$ folgt aus $x_1 < x_2$, dass $f(x_1) = x_1^3 < x_2^3 = f(x_2)$ gilt. Man sieht an Beispiel (b), warum in Satz 1.6 (4) nicht die Rückrichtung „$\Longleftarrow$“ gelten kann. ♠

Natürlich kann man Funktionen auch zweimal differenzieren, wenn die Ableitung der Ableitung existiert.

Definition 1.8. (höhere Ableitungen)
Sei I *ein offenes Intervall, und seien* $f : I \to \mathbb{R}$, $x_0 \in I$ *und* $k \in \mathbb{N}$.

(1) Ist f *differenzierbar in* I *und ist* f' *in* x_0 *differenzierbar, so heißt* f ***zweimal differenzierbar in*** x_0.

$$f^{(2)}(x_0) = f''(x_0) = (f')'(x_0)$$

heißt dann die ***zweite Ableitung von*** f ***in*** x_0. *Ist* f *in jedem* $x_0 \in I$ *zweimal differenzierbar, so heißt* f ***zweimal differenzierbar in*** I.

(2) Allgemein heißt f ***in*** *x_0* ***k-mal differenzierbar****, wenn f in I $(k-1)$-mal differenzierbar ist und $f^{(k-1)}$ in x_0 differenzierbar ist.*

$$f^{(k)}(x_0) = (f^{(k-1)})'(x_0)$$

heißt dann die ***k-te Ableitung von f in*** *x_0.*

(3) Man schreibt: $f^{(0)} = f$, $f^{(1)} = f'$.
Die Ableitung f' von f ist also eigentlich die erste Ableitung.

Beispiel 1.9. (k-mal stetig differenzierbare Funktionen)

(a) Alle Polynome (vgl. Definition 1.10) sind beliebig oft stetig differenzierbar. Für das Polynom

$$p : \mathbb{R} \to \mathbb{R}, \qquad p(x) = 3 + 7\,x + 13\,x^2,$$

gilt beispielsweise

$$p'(x) = 7 + 26\,x, \qquad p''(x) = 26, \qquad p^{(3)} = 0, \qquad p^{(k)} = 0 \text{ für alle } k \geq 4.$$

Allgemein gilt für Polynome: Die k-ten Ableitungen eines Polynoms vom Grad n sind für $k > n$ alle gleich der Nullfunktion.

(b) Die Exponentialfunktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, ist beliebig oft differenzierbar, und für alle Ableitungen gilt $f^{(k)}(x) = e^x$ (für $k \in \mathbb{N}$). Mit $f^{(0)}(x) = e^x$ gilt also $f^{(k)}(x) = e^x$ für $k \in \mathbb{N}_0$.

Wir werden im weiteren Verlauf des Kapitels weitere Beispiele kennenlernen. ♠

1.2 Taylor-Polynome

Problem: Die meisten mathematischen Funktionen können ohne die Hilfe eines Computers oder Taschenrechner nicht einfach in einem Punkt x ausgewertet werden. Wie wollen Sie z.B. e^x, $\cos(x)$ oder $\sqrt{x}$ für $x = 3$ ohne die Hilfe eines Taschenrechners berechnen?

Numerische Lösung des Problems: Um solche Funktionen einfach berechnen zu können, nähert man sie durch eine „einfachere“ Funktion an, die sich leicht und effizient berechnen lässt. Die am häufigsten genutzten „einfacheren“ Funktionen sind **Polynome** oder **stückweise Polynome**. (Stückweise Polynome sind

Funktionen, die stückweise definiert sind und auf jedem in der Fallunterscheidung aufgeführten Teilintervall ihrer Definitionsmenge jeweils ein Polynom sind). Das einfachste Polynom, mit dem man eine hinreichend oft differenzierbare Funktion f in der Nähe eines Punkte x_0 gut annähern kann, ist ein **Taylor-Polynom dieser Funktion** (passenden Grades) **mit Entwicklungspunkt** x_0. Taylor-Polynome sind das Thema dieses Teilkapitels.

Wir erinnern uns zunächst an den Begriff eines Polynoms.

Definition 1.10. (Polynom)

(1) Seien $n \in \mathbb{N}_0$ und reelle „Koeffizienten“ $a_0, a_1, a_2, \ldots, a_n \in \mathbb{R}$ mit $a_n \neq 0$ gegeben. Dann heißt die Funktion

$$p : \mathbb{R} \to \mathbb{R}, \qquad p(x) = a_0 + a_1\, x + a_2\, x^2 + \ldots + a_n\, x^n = \sum_{k=0}^{n} a_k\, x^k,$$

*ein **Polynom vom Grad** n. Wir schreiben dann* $\mathrm{Grad}(p) = n$.

*(2) Die **Menge aller Polynome vom Grad** $\leq n$ wird mit $\mathbb{P}_n$ bezeichnet.*

Betrachten wir einige Beispiele für Polynome.

Beispiel 1.11. (Polynome)

(a) $p : \mathbb{R} \to \mathbb{R}$, $p(x) = 13 - 55\, x^3 - 3\, x + 9\, x^6 - x = 13 - 3\, x - 55\, x^3 + 9\, x^6$, ist ein Polynom vom Grad 6.

(b) $p : \mathbb{R} \to \mathbb{R}$, $p(x) = 7 + 0\, x^4 = 7$, ist ein Polynom vom Grad 0.

(c) $p : \mathbb{R} \to \mathbb{R}$, $p(x) = 2 + x^2 + 5\, x - x^2 = 2 + 5\, x$, ist ein Polynom vom Grad 1.

(d) $p : \mathbb{R} \to \mathbb{R}$, $p(x) = 0$, ist das Nullpolynom. Dieses hat per Definition den Grad $-\infty$.

Man definiert den Grad des Nullpolynoms als $-\infty$, damit Aussagen über den Grad eines Produkts von Polynomen auch für das Nullpolynom stimmen. ♠

Bemerkung 1.12. (gleiche Polynome)

Zwei Polynome $p : \mathbb{R} \to \mathbb{R}$, $p(x) = a_0 + a_1\, x + a_2\, x^2 + \ldots + a_n\, x^n$ mit $a_n \neq 0$, und $q : \mathbb{R} \to \mathbb{R}$, $q(x) = b_0 + b_1\, x + b_2\, x^2 + \ldots + b_m\, x^m$ mit $b_m \neq 0$, sind genau dann **gleich**, wenn sie den **gleichen Grad** haben und **ihre Koeffizienten**

übereinstimmen, also wenn gilt

$$n = m \qquad \text{und} \qquad a_0 = b_0, \quad a_1 = b_1, \quad a_2 = b_2, \quad \ldots, \quad a_n = b_n.$$

Notation 1.13. (Polynome)
Es gelten die folgenden Bezeichnungen für Polynome bis zum Grad 3:

Grad 0 :	$p(x) = a_0$ mit $a_0 \neq 0$	konstantes Polynom
Grad 1 :	$p(x) = a_0 + a_1\, x$ mit $a_1 \neq 0$	lineares Polynom
Grad 2 :	$p(x) = a_0 + a_1\, x + a_2\, x^2$ mit $a_2 \neq 0$	quadratisches Polynom
Grad 3 :	$p(x) = a_0 + a_1\, x + a_2\, x^2 + a_3\, x^3$ mit $a_3 \neq 0$	kubisches Polynom

Wie brauchen noch eine wichtige Aussage über Nullstellen von Polynomen.

Satz 1.14. (Nullstellen von Polynomen)

*(1) Ein Polynom p vom Grad $n \geq 1$ hat **höchstens** n **Nullstellen**.*

*(2) Gilt für ein Polynom $q \in \mathbb{P}_n$ (also q hat $\mathrm{Grad}(q) \leq n$) mit $n \in \mathbb{N}_0$, dass es **mehr als** n **Nullstellen** (also mindestens $n+1$ Nullstellen) hat, so ist q das **Nullpolynom**.*

Beispiel 1.15. (Nullstellen von Polynomen)

(a) $p(x) = x^2 - 1$ hat genau zwei verschiedene reelle Nullstellen, nämlich $x_1 = -1$ und $x_2 = 1$, da (nach der dritten binomischen Formel) gilt

$$p(x) = x^2 - 1 = (x+1)\,(x-1).$$

(b) $p(x) = x^2 - 4\,x + 4 = (x-2)^2$ hat genau eine reelle Nullstelle in $x_1 = 2$. Diese hat allerdings die „Vielfachheit“ 2.

(c) $p(x) = x^2 + 1$ hat keine reellen Nullstellen, da $p(x) = x^2 + 1 \geq 1$ für alle $x \in \mathbb{R}$ gilt.

(d) $p(x) = x^3 + x$ hat genau eine reelle Nullstelle $x_1 = 0$, da gilt

$$p(x) = x^3 + x = x\,(x^2 + 1)$$

und da $x^2 + 1 \geq 1$ für alle $x \in \mathbb{R}$ ist.

(e) Ein konstantes Polynom $p(x) = c$ hat entweder keine Nullstellen, wenn $c \neq 0$ ist, oder ist das Nullpolynom, wenn $c = 0$ ist. Alle reellen Zahlen sind Nullstellen des Nullpolynoms $p(x) = 0$.

Das Faktorisieren eines Polynoms vom Grad 2 wird in dieser Vorlesung wiederholt benötigt. Sie sollten dieses daher üben, falls erforderlich. ♠

Wir führen nun das Taylor-Polynom (einer n-mal differenzierbaren Funktion f) vom Grad n mit dem Entwicklungspunkt x_0 ein.

Definition 1.16. (Taylor-Polynom)
Seien I ein offenes Intervall, $x_0 \in I$ und $f : I \to \mathbb{R}$ eine (mindestens) n-mal differenzierbare Funktion. Das ***Taylor-Polynom von f vom Grad n mit dem Entwicklungspunkt x_0*** *(oder* ***um x_0****) ist das eindeutig bestimmte Polynom $p_n \in \mathbb{P}_n$ (also vom Grad $\leq n$), welches die folgenden Bedingungen alle erfüllt:*

$$p_n(x_0) = f(x_0), \quad p_n'(x_0) = f'(x_0), \quad p_n''(x_0) = f''(x_0),$$
$$p_n^{(3)}(x_0) = f^{(3)}(x_0), \quad \ldots, \quad p_n^{(n)}(x_0) = f^{(n)}(x_0). \tag{1.2}$$

Mit der Konvention, dass die 0-te Ableitung $f^{(0)}$ einer Funktion f die Funktion selber ist, also $f^{(0)} = f$, können wir (1.2) auch kürzer schreiben als

$$\boxed{p_n^{(k)}(x_0) = f^{(k)}(x_0) \qquad \text{für alle } k = 0, 1, 2, \ldots, n.} \tag{1.3}$$

In der Regel liefert das Taylor-Polynom p_n von f vom Grad n um x_0 bereits für kleine Werte von n in der Nähe von x_0 eine gute Näherung von f.

Wir wollen nun aus den Bedingungen (1.2) bzw. (1.3) die Formel für das Taylor-Polynom von f um x_0 vom Grad 0 bzw. vom Grad 1 herleiten.

Herleitung der Formel des Taylor-Polynoms von f vom Grad 0 um x_0: Wir machen den Ansatz $p_0(x) = a_0$, da das Taylor-Polynom p_0 vom Grad 0 ein konstantes Polynom ist. Die Bedingung $p_0(x_0) = f(x_0)$ aus (1.2) bzw. (1.3) liefert:

$$p_0(x_0) = a_0 \stackrel{!}{=} f(x_0) \qquad \Longrightarrow \qquad a_0 = f(x_0)$$

Also finden wir

$$\boxed{\text{Taylor-Polynom von } f \text{ vom Grad 0 um } x_0: \qquad p_0(x) = f(x_0)} \tag{1.4}$$

Herleitung der Formel des Taylor-Polynoms von f vom Grad 1 um x_0:
Da das Taylor-Polynom p_1 eine lineare Funktion ist, machen wir den Ansatz

$$p_1(x) = a_0 + a_1\, x, \tag{1.5}$$

und es müssen nach (1.2) bzw. (1.3) die Bedingungen

$$p_1(x_0) = f(x_0), \qquad p_1'(x_0) = f'(x_0)$$

gelten. Einsetzen des Ansatzes für p_1 in diese Bedingungen liefert:

$$p_1(x_0) = a_0 + a_1\, x_0 \stackrel{!}{=} f(x_0) \qquad \text{(I)}$$
$$p_1'(x_0) = a_1 \stackrel{!}{=} f'(x_0) \qquad \text{(II)}$$

Aus (II) folgt direkt $a_1 = f'(x_0)$. Einsetzen von $a_1 = f'(x_0)$ in (I) liefert:

$$a_0 + f'(x_0)\, x_0 = f(x_0) \qquad \Longleftrightarrow \qquad a_0 = f(x_0) - f'(x_0)\, x_0$$

Einsetzen der Formeln für die Koeffizienten a_0 und a_1 in den Ansatz (1.5) liefert:

$$p_1(x) = f(x_0) - f'(x_0)\, x_0 + f'(x_0)\, x = f(x_0) + f'(x_0)\,(x - x_0).$$

Also finden wir

$$\boxed{\begin{array}{rl} \text{Taylor-Polynom von } f \\ \text{vom Grad 1 um } x_0\text{:} & p_1(x) = f(x_0) + f'(x_0)\,(x - x_0) \end{array}} \tag{1.6}$$

Dieses ist die Gleichung der Tangente an f im Punkt x_0.

Beispiel 1.17. (Taylor-Polynome vom Grad 0 und 1)

(a) Seien $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und $x_0 = 0$. Dann gilt $f'(x) = e^x$. Also sind die Taylor-Polynome von $f(x) = e^x$ vom Grad 0 bzw. Grad 1 um $x_0 = 0$ gegeben durch

$$p_0(x) = f(0) = e^0 = 1,$$
$$p_1(x) = f(0) + f'(0)\,(x - 0) = e^0 + e^0\,(x - 0) = 1 + x.$$

(b) Seien $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und $x_0 = 1$. Dann gilt $f'(x) = e^x$. Also sind die Taylor-Polynome von $f(x) = e^x$ vom Grad 0 bzw. Grad 1 um $x_0 = 1$ gegeben durch

$$p_0(x) = f(1) = e^1 = e,$$
$$p_1(x) = f(1) + f'(1)\,(x - 1) = e^1 + e^1\,(x - 1) = e + e\,(x - 1) = e\,x.$$

(c) Seien $f :]0;\infty[\to \mathbb{R}$, $f(x) = \sqrt{x} = x^{1/2}$, und $x_0 = 2$. Dann gilt $f'(x) = \frac{1}{2}\, x^{-1/2} = \frac{1}{2\sqrt{x}}$. Also sind die Taylor-Polynome von $f(x) = \sqrt{x}$ vom Grad 0 bzw. Grad 1 um $x_0 = 2$ gegeben durch

$$p_0(x) = f(2) = \sqrt{2},$$
$$p_1(x) = f(2) + f'(2)\,(x-2) = \sqrt{2} + \frac{1}{2\sqrt{2}}\,(x-2).$$

Überlegen Sie sich selber weitere Beispiele zum Üben. ♠

Idee zur Herleitung der Formel des Taylor-Polynoms von f vom Grad 2 um x_0: Das Taylor-Polynom p_2 von f vom Grad 2 um x_0 ist ein quadratisches Polynom, d.h. wir machen dann Ansatz

$$p_2(x) = a_0 + a_1\, x + a_2\, x^2. \tag{1.7}$$

Das Einsetzen dieses Ansatzes in die Bedingungen (vgl. (1.2) bzw. (1.3))

$$p_2(x_0) = f(x_0), \qquad p_2'(x_0) = f'(x_0) \qquad \text{und} \qquad p_2''(x_0) = f''(x_0).$$

ermöglicht es, die Koeffizienten a_0, a_1, a_2 im Ansatz (1.7) eindeutig zu bestimmen. Dieses führt zu der folgenden Formel für p_2 (Details als Übungsaufgabe):

Taylor-Polynom von f vom Grad 2 um x_0:	$p_2(x) = f(x_0) + f'(x_0)\,(x-x_0) + \frac{1}{2}\, f''(x_0)\,(x-x_0)^2$

(1.8)

Beispiel 1.18. (Taylor-Polynome vom Grad 2)

(a) Seien $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und $x_0 = 0$. Dann gelten $f'(x) = e^x$ und $f''(x) = e^x$. Also ist das Taylor-Polynom von $f(x) = e^x$ vom Grad 2 um $x_0 = 0$ gegeben durch

$$\begin{aligned} p_2(x) &= f(0) + f'(0)\,(x-0) + \frac{1}{2}\, f''(0)\,(x-0)^2 \\ &= e^0 + e^0\,(x-0) + \frac{1}{2}\, e^0\,(x-0)^2 = 1 + x + \frac{1}{2}\, x^2. \end{aligned}$$

(b) Seien $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und $x_0 = 1$. Dann gelten $f'(x) = e^x$ und $f''(x) = e^x$. Also ist das Taylor-Polynom von $f(x) = e^x$ vom Grad 2 um $x_0 = 1$ gegeben durch

$$p_2(x) = f(1) + f'(1)\,(x-1) + \frac{1}{2}\, f''(1)\,(x-1)^2$$

$$= e^1 + e^1\,(x-1) + \frac{1}{2}\,e^1\,(x-1)^2$$
$$= e + e\,(x-1) + \frac{1}{2}\,e\,(x-1)^2.$$

(c) Seien $f :]0;\infty[\to \mathbb{R}$, $f(x) = \sqrt{x} = x^{1/2}$, und $x_0 = 2$. Dann gelten

$$f'(x) = \frac{1}{2}\,x^{\frac{1}{2}-1} = \frac{1}{2}\,x^{-1/2} = \frac{1}{2\sqrt{x}}\,,$$
$$f''(x) = \frac{1}{2}\cdot\left(-\frac{1}{2}\right)\,x^{-\frac{1}{2}-1} = -\frac{1}{4}\,x^{-3/2} = -\frac{1}{4\,x\sqrt{x}}\,.$$

Also ist das Taylor-Polynom von $f(x) = \sqrt{x}$ vom Grad 2 um $x_0 = 2$ durch

$$p_2(x) = f(2) + f'(2)\,(x-2) + \frac{1}{2}\,f''(2)\,(x-2)^2$$
$$= \sqrt{2} + \frac{1}{2\sqrt{2}}\,(x-2) - \frac{1}{16\sqrt{2}}\,(x-2)^2$$

gegeben.

Überlegen Sie sich selber weitere Beispiele zum Üben. ♠

Man kann den Prozess zum Bestimmen der Taylor-Polynome fortsetzen: Für das Taylor-Polynom p_n von f vom Grad n um x_0 machen wir den Ansatz

$$p_n(x) = a_0 + a_1\,x + a_2\,x^2 + a_3\,x^3 + \ldots + a_n\,x^n.$$

Einsetzen dieses Ansatzes in die Bedingungen (vgl. (1.2) bzw. (1.3))

$$p_n^{(k)}(x_0) = f^{(k)}(x_0) \qquad \text{für alle } k = 0, 1, 2, \ldots, n$$

ermöglicht es, die Koeffizienten $a_0, a_1, a_2, \ldots, a_n$ eindeutig zu bestimmen. Dieses führt auf die folgende Formel für das **Taylor-Polynom von f vom Grad n mit Entwicklungspunkt** x_0:

$$\boxed{p_n(x) = f(x_0) + f'(x_0)\,(x-x_0) + \frac{f''(x_0)}{2!}\,(x-x_0)^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}\,(x-x_0)^n} \tag{1.9}$$

bzw. kürzer mit der Summennotation

$$\boxed{p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}\,(x-x_0)^k.} \tag{1.10}$$

Natürlich sind die Formeln (1.4), (1.6) und (1.8) für p_0 bzw. p_1 bzw. p_2 in (1.9) als Sonderfälle für $n = 0$ bzw. $n = 1$ bzw. $n = 2$ enthalten.

Das in den Formel (1.9) und (1.10) vorkommende mathematische Objekt $k!$ heißt **k-Fakultät** und ist für $k \in \mathbb{N}_0$ wie folgt definiert:

$$0! = 1 \qquad \text{und} \qquad k! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot k \quad \text{für } k \in \mathbb{N}.$$

Es gilt die rekursive Beziehung $(k+1)! = k! \cdot (k+1)$. Wir finden also

$$1! = 1, \qquad 2! = 1 \cdot 2, \qquad 3! = \underbrace{1 \cdot 2}_{=2!} \cdot 3 = 6, \qquad 4! = \underbrace{1 \cdot 2 \cdot 3}_{=3!} \cdot 4 = 24.$$

Beispiel 1.19. (Taylor-Polynome vom Grad n)

(a) Seien $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und $x_0 = 0$. Dann gelten $f'(x) = e^x$ und $f''(x) = e^x$ und allgemeiner $f^{(k)}(x) = e^x$ für jedes $k \in \mathbb{N}_0$. Also ist das Taylor-Polynom von $f(x) = e^x$ vom Grad n um $x_0 = 0$ gegeben durch

$$\begin{aligned} p_n(x) &= f(0) + f'(0)\,(x-0) + \frac{f''(0)}{2!}\,(x-0)^2 + \ldots + \frac{f^{(n)}(0)}{n!}\,(x-0)^n \\ &= e^0 + e^0\,(x-0) + \frac{e^0}{2!}\,(x-0)^2 + \ldots + \frac{e^0}{n!}\,(x-0)^n \\ &= 1 + x + \frac{1}{2!}\,x^2 + \ldots + \frac{1}{n!}\,x^n = \sum_{k=0}^{n} \frac{1}{k!}\,x^k. \end{aligned}$$

(b) Seien $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und $x_0 = 1$. Dann gelten $f'(x) = e^x$ und $f''(x) = e^x$ und allgemeiner $f^{(k)}(x) = e^x$ für jedes $k \in \mathbb{N}_0$. Also ist das Taylor-Polynom von $f(x) = e^x$ vom Grad n um $x_0 = 1$ gegeben durch

$$\begin{aligned} p_n(x) &= f(1) + f'(1)\,(x-1) + \frac{f''(1)}{2!}\,(x-1)^2 + \ldots + \frac{f^{(n)}(1)}{n!}\,(x-1)^n \\ &= e^1 + e^1\,(x-1) + \frac{e^1}{2!}\,(x-1)^2 + \ldots + \frac{e^1}{n!}\,(x-1)^n \\ &= e + e\,x + \frac{e}{2!}\,(x-1)^2 + \ldots + \frac{e}{n!}\,(x-1)^n = \sum_{k=0}^{n} \frac{e}{k!}\,(x-1)^k. \end{aligned}$$

Wir werden das Berechnen von Taylor-Polynomen in Übungsaufgaben üben. ♠

Wir halten die Formel für das Taylor-Polynom vom Grad n mit dem Entwicklungspunkt x_0 einer (mindestens) n-mal differenzierbaren Funktion als Satz fest.

Satz 1.20. (Taylor-Polynom)
Seien I ein offenes Intervall, $x_0 \in I$ und $f : I \to \mathbb{R}$ eine (mindestens) n-mal differenzierbare Funktion. Das in Definition 1.16 definierte ***Taylor-Polynom von f vom Grad n mit dem Entwicklungspunkt*** *x_0 (oder* ***um*** *x_0) ist durch die folgende Formel gegeben:*

$$p_n(x) = f(x_0) + f'(x_0)\,(x - x_0) + \frac{f''(x_0)}{2!}\,(x - x_0)^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}\,(x - x_0)^n$$
$$= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}\,(x - x_0)^k. \tag{1.11}$$

Bemerkung 1.21. (rekursive Berechnung der Taylor-Polynome)
Seien die Voraussetzungen und die Notation wie im vorigen Satz 1.20. An der Formel (1.11) sieht man, dass sich $p_1, p_2, \ldots, p_n$ rekursiv berechnen lassen:

$$\begin{aligned}
p_0(x) &= f(x_0),\\
p_1(x) &= p_0(x) + \frac{f'(x_0)}{1!}\,(x - x_0),\\
p_2(x) &= p_1(x) + \frac{f''(x_0)}{2!}\,(x - x_0)^2,\\
p_3(x) &= p_2(x) + \frac{f^{(3)}(x_0)}{3!}\,(x - x_0)^3,\\
&\vdots\\
p_n(x) &= p_{n-1}(x) + \frac{f^{(n)}(x_0)}{n!}\,(x - x_0)^n.
\end{aligned}$$

Zuletzt wollen wir noch an einem Beispiel einen ersten Eindruck gewinnen, wie gut eine Funktion eigentlich durch ein Taylor-Polynom angenähert oder approximiert wird. Im nächsten Teilkapitel wird die Qualität der Annäherung einer Funktion durch seine Taylor-Polynome um x_0 dann genauer untersucht.

Beispiel 1.22. (Qualität der Näherung durch ein Taylor-Polynom)
Die Taylor-Polynome von $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, vom Grad 1, 2 bzw. 3 mit dem Entwicklungspunkt $x_0 = 0$ sind nach Beispielen 1.17, 1.18 und 1.19

$$p_1(x) = 1 + x, \qquad p_2(x) = 1 + x + \frac{1}{2}\,x^2, \qquad p_3(x) = 1 + x + \frac{1}{2}\,x^2 + \frac{1}{6}\,x^3.$$

In der nachfolgenden Tabelle wurden die Werte der Taylor-Polynome p_1, p_2, p_3 und f an verschiedenen Stellen x dicht bei $x_0 = 0$ berechnet. Dabei wurde auf fünf Nachkommastellen gerundet.

x	$p_1(x)$	$p_2(x)$	$p_3(x)$	$f(x) = e^x$
-1,0	0	0,5	0,33333	0,36788
-0,5	0,5	0,625	0,60417	0,60653
-0,1	0,9	0,905	0,90483	0,90484
0	1,0	1,000	1,00000	1,00000
0,1	1,1	1,105	1,10517	1,10517
0,5	1,5	1,625	1,64583	1,64872
1,0	2,0	2,500	2,66667	2,71828

Wir beobachten, das für $x = -0{,}1$ und $x = 0{,}1$ schon das Taylor-Polynom vom Grad 3 eine sehr gute Näherung für $f(x) = e^x$ liefert. Je weiter wir uns mit x von $x_0 = 0$ entfernen, desto schlechter scheint die Näherung $p_3(x)$ von $f(x)$ in der Regel zu werden. – Wir beobachten auch, dass sich die Näherung zu verbessern scheint, wenn man den Grad des Taylor-Polynoms erhöht. ♠

1.3 Fehler bei Näherung durch Taylor-Polynome

Sei I ein offenes Intervall. Der Satz von Taylor (auch Taylorsche Formel genannt) gibt uns an, wie gut eine mindestens $(n+1)$-mal stetig differenzierbare Funktion $f : I \to \mathbb{R}$ durch ihr Taylor-Polynom p_n vom Grad n mit dem Entwicklungspunkt $x_0 \in I$ in einem Punkt $x \in I$ angenähert wird.

Satz 1.23. (Satz von Taylor / Taylorsche Formel)
Seien I ein offenes Intervall, $n \in \mathbb{N}_0$ und $x_0 \in I$, und sei $f : I \to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion. Dann existiert zu jedem $x \in I$ ein Punkt z_x zwischen x_0 und x mit

$$f(x) = \underbrace{\sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k}_{= p_n(x) \text{ (Taylor-Polynom vom Grad } n)} + \underbrace{\frac{1}{(n+1)!} f^{(n+1)}(z_x) (x - x_0)^{n+1}}_{= r_n(x) \text{ (Restglied)}}. \qquad (1.12)$$

*Der Term $r_n(x)$ heißt das **Restglied von** $p_n(x)$. Das Restglied gibt den **Fehler der Näherung** von $f(x)$ durch sein Taylor-Polynom $p_n(x)$ vom Grad n mit dem Entwicklungspunkt x_0 an. (Achtung: Die Stelle z_x in $r_n(x)$ hängt in der Regel von x ab!)*

Auch Satz 1.23 ist eine **Existenzaussage**: Wir bekommen keine Information darüber, wo genau zwischen x und x_0 ein Punkt z_x mit (1.12) auftritt.

Betrachten wir zunächst zwei Beispiele für die Anwendung des Satzes von Taylor.

Beispiel 1.24. (Anwendung des Satzes von Taylor)
Die Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, ist auf ganz $\mathbb{R}$ beliebig oft differenzierbar, und es gilt $f^{(k)}(x) = e^x$ für alle $k \in \mathbb{N}_0$. Nach Beispiel 1.19 (a) ist das Taylor-Polynom von $f(x) = e^x$ vom Grad n mit dem Entwicklungspunkt $x_0 = 0$ durch

$$p_n(x) = 1 + x + \frac{1}{2!}\,x^2 + \ldots + \frac{1}{n!}\,x^n = \sum_{k=0}^{n} \frac{1}{k!}\,x^k$$

gegeben. Das Restglied ist

$$r_n(x) = \frac{1}{(n+1)!}\,f^{(n+1)}(z_x)\,(x-0)^{n+1} = \frac{1}{(n+1)!}\,e^{z_x}\,x^{n+1}$$

mit einem z_x zwischen x und 0.

Der Satz von Taylor für die Annäherung von $f(x) = e^x$ durch sein Taylor-Polynom p_n vom Grad n mit dem Entwicklungspunkt $x_0 = 0$ lautet also:

Zu jedem $x \in \mathbb{R}$ gibt es einen Punkt z_x zwischen x und 0, so dass gilt

$$\begin{aligned} e^x &= \underbrace{1 + x + \frac{1}{2!}\,x^2 + \ldots + \frac{1}{n!}\,x^n}_{=p_n(x)} + \underbrace{\frac{1}{(n+1)!}\,e^{z_x}\,x^{n+1}}_{=r_n(x)} \\ &= \underbrace{\sum_{k=0}^{n} \frac{1}{k!}\,x^k}_{=p_n(x)} + \underbrace{\frac{1}{(n+1)!}\,e^{z_x}\,x^{n+1}}_{=r_n(x)}. \end{aligned} \tag{1.13}$$

Damit erhält man die folgende Fehlerabschätzung für die Qualität der Annäherung von $f(x) = e^x$ durch sein Taylor-Polynom p_n: Durch Umsortieren von (1.13) und Bilden des Absolutbetrags auf beiden Seiten finden wir

$$e^x - \sum_{k=0}^{n} \frac{1}{k!}\,x^k = \frac{1}{(n+1)!}\,e^{z_x}\,x^{n+1}$$

$$\Longrightarrow \qquad \left| e^x - \sum_{k=0}^{n} \frac{1}{k!}\, x^k \right| = \left| \frac{1}{(n+1)!}\, e^{z_x}\, x^{n+1} \right| = \frac{1}{(n+1)!}\, |e^{z_x}|\, |x|^{n+1}.$$

Wir beschränken uns nun auf $x \in [-1;1]$ (also $|x| \leq 1$). Da z_x zwischen 0 und x liegt, folgt dann auch $z_x \in [-1;1]$ und somit $|z_x| \leq 1$. Daraus folgt

$$\left| e^x - p_n(x) \right| = \left| e^x - \sum_{k=0}^{n} \frac{1}{k!} x^k \right| = \frac{1}{(n+1)!} \underbrace{|e^{z_x}|}_{\leq e^{|z_x|} \leq e^1} \underbrace{|x|^{n+1}}_{\leq 1} \leq \frac{e}{(n+1)!},$$

wobei wir genutzt haben, dass $f(x) = e^x$ streng monoton wachsend und positivwertig ist und daher auf $[-1;1]$ in $x = 1$ sein Maximum annimmt. Also erhalten wir die Fehlerabschätzung

$$\left| e^x - p_n(x) \right| = \left| e^x - \sum_{k=0}^{n} \frac{1}{k!} x^k \right| \leq \frac{e}{(n+1)!} \qquad \text{für alle } x \in [-1;1]. \tag{1.14}$$

Frage: Wie groß muss der Grad n des Taylor-Polynoms p_n sein, wenn man $f(x) = e^x$ auf dem Intervall $[-1;1]$ (also für $|x| \leq 1$) durch sein Taylor-Polynom p_n garantiert auf (mindestens) zwei Nachkommastellen genau berechnen möchte?

Antwort: „Auf zwei Nachkommastellen genau“ bedeutet, dass der absolute Fehler (also die Abweichung vom exakten Wert) betraglich kleiner oder gleich $0{,}005 = 5 \cdot 10^{-3}$ ist. In (1.14) sollte also gelten:

$$\frac{e}{(n+1)!} \leq 5 \cdot 10^{-3} \quad \Longleftrightarrow \quad \frac{e}{5 \cdot 10^{-3}} \leq (n+1)! \quad \Longleftrightarrow \quad \underbrace{\frac{2\, e \cdot 10^2}{\doteq 543{,}656}} \leq (n+1)!$$

Das Symbol $\doteq$ bedeutet, dass wir gerundet haben.

Durch Probieren mit $n = 1, 2, 3, 4, 5$ findet man, dass diese Ungleichung zuerst für $n = 5$ mit $(n+1)! = 6! = 720$ erfüllt ist. Also muss der Grad des Taylor-Polynoms $n = 5$ sein, damit $f(x) = e^x$ auf dem Intervall $[-1;1]$ garantiert auf (mindestens) zwei Nachkommastellen genau berechnet wird.

In Abbildung 1.3 sind die Graphen der natürlichen Exponentialfunktion $f(x) = e^x$ und seiner n-ten Taylor-Polynome mit dem Entwicklungspunkt $x_0 = 0$ für $n = 1, 2, 3$ gezeichnet. Man sieht, dass bereits für $n = 3$ das Taylor-Polynom für x dicht bei $x_0 = 0$ die Funktion $f(x) = e^x$ sehr gut annähert. ♠

Beispiel 1.25. (Anwendung des Satzes von Taylor)
Wir wollen alle Taylor-Polynome p_n von $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sin(x)$, mit dem

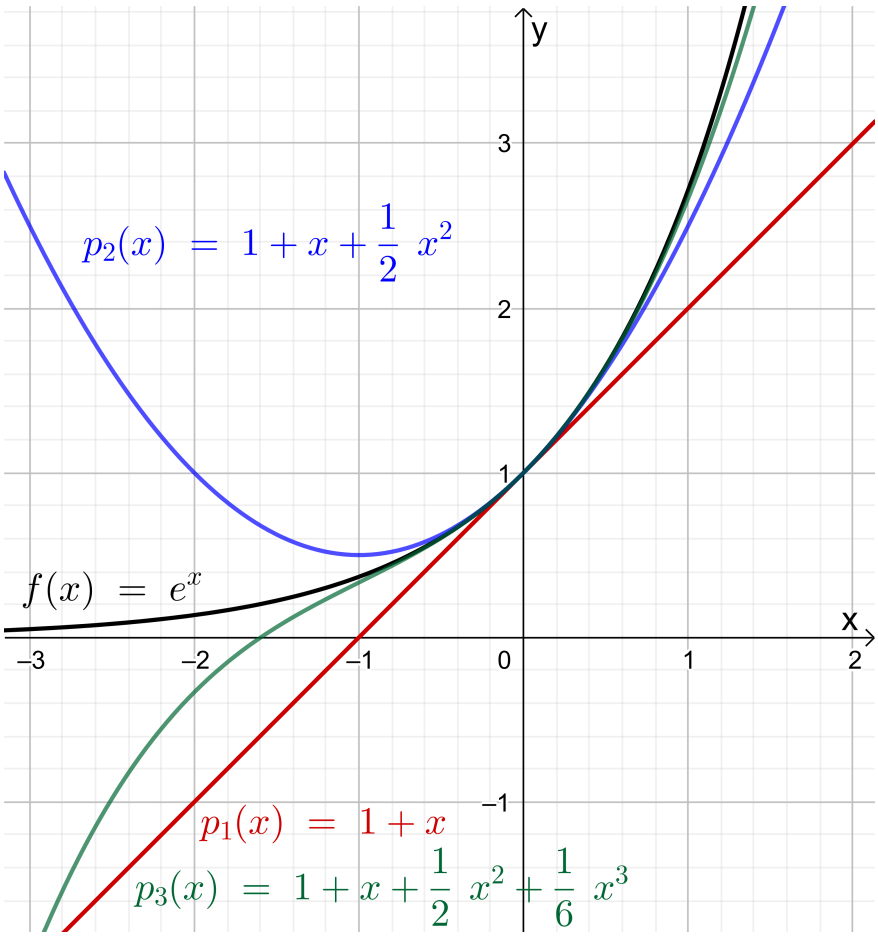


Abb. 1.3: Die Graphen von $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, (schwarz) und seinen n-ten Taylor-Polynomen $p_n(x)$ mit dem Entwicklungspunkt $x_0 = 0$ für $n = 1$ (rot) und $n = 2$ (blau) und $n = 3$ (grün).

Entwicklungspunkt $x_0 = 0$ berechnen und deren Abweichung vom exakten Wert $\sin(x)$ mit dem Satz von Taylor untersuchen. Dazu brauchen wir vorab alle Ableitungen von $f(x) = \sin(x)$.

$$f(x) = \sin(x), \quad f'(x) = \cos(x), \quad f''(x) = -\sin(x), \quad f^{(3)}(x) = -\cos(x),$$
$$f^{(4)}(x) = \sin(x), \quad f^{(5)}(x) = \cos(x), \quad f^{(6)}(x) = -\sin(x), \quad f^{(7)}(x) = -\cos(x),$$

und wir sehen, dass für alle $k \in \mathbb{N}_0$ gilt

$$f^{(4k)}(x) = \sin(x), \qquad f^{(4k+1)}(x) = \cos(x),$$
$$f^{(4k+2)}(x) = -\sin(x), \qquad f^{(4k+3)}(x) = -\cos(x).$$

Damit finden wir mit der rekursiven Berechnung der Taylor-Polynome (siehe Bemerkung 1.21)

$$p_0(x) = \sin(0) = 0,$$

$$p_1(x) = p_0(x) + \frac{f'(0)}{1!}(x-0)^1 = 0 + \cos(0)\,x = x,$$
$$p_2(x) = p_1(x) + \frac{f''(0)}{2!}(x-0)^2 = x + \frac{-\sin(0)}{2!}x^2 = x,$$
$$p_3(x) = p_2(x) + \frac{f^{(3)}(0)}{3!}(x-0)^3 = x + \frac{-\cos(0)}{3!}x^3 = x - \frac{1}{3!}x^3,$$
$$p_4(x) = p_3(x) + \frac{f^{(4)}(0)}{4!}(x-0)^4 = x - \frac{1}{3!}x^3 + \frac{\sin(0)}{4!}x^4 = x - \frac{1}{3!}x^3,$$
$$p_5(x) = p_4(x) + \frac{f^{(5)}(0)}{5!}(x-0)^5 = x - \frac{1}{3!}x^3 + \frac{\cos(0)}{5!}x^5 = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5.$$

Setzt man diesen Prozess fort, so sieht man, dass für das Taylor-Polynom p_{2m+1} vom Grad $n = 2m+1$ mit $m \in \mathbb{N}_0$ und für das Taylor-Polynom p_{2m+2} vom Grad $2m+2$ mit $m \in \mathbb{N}_0$ (jeweils mit dem Entwicklungspunkt $x_0 = 0$) gilt:

$$\begin{aligned} p_{2m+2}(x) = p_{2m+1}(x) &= \sum_{k=0}^{m} \frac{(-1)^k}{(2k+1)!}x^{2k+1} \\ &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 + \ldots + (-1)^m \frac{1}{(2m+1)!}x^{2m+1}. \end{aligned} \tag{1.15}$$

Der Sinus ist eine ungerade Funktion, da $\sin(-x) = -\sin(x)$ für alle $x \in \mathbb{R}$ gilt. Wir beobachten, dass in seinen Taylor-Polynomen nur x^k mit ungeradem k vorkommen. Die Monome x^k mit ungeradem k sind ebenfalls ungerade Funktionen.

In Abbildung 1.4 sind die Graphen von p_1, p_3, p_5 und p_7 zusammen mit dem Graphen der Sinusfunktion $f(x) = \sin(x)$ gezeichnet.

Für die Annäherung von $\sin(x)$ durch $p_{2m+1}(x)$ liefert der Satz vom Taylor die folgende Aussage: Zu jedem $x \in \mathbb{R}$ gibt es ein z_x zwischen 0 und x, so dass gilt

$$\sin(x) = p_{2m+1}(x) + \frac{1}{(2m+2)!}f^{(2m+2)}(z_x)\,(x-0)^{2m+2},$$

wobei $f^{(2m+2)}(z_x) = \sin(z_x)$ für ungerades m und $f^{(2m+2)}(z_x) = -\sin(z_x)$ für gerades m. Umstellen und den Absolutbetrag Anwenden liefert:

$$\sin(x) - p_{2m+1}(x) = \frac{1}{(2m+2)!}f^{(2m+2)}(z_x)\,(x-0)^{2m+2}$$
$$\Longrightarrow \qquad \big|\sin(x) - p_{2m+1}(x)\big| = \frac{1}{(2m+2)!}\,|\sin(z_x)|\,|x|^{2m+2}$$

Wegen $|\sin(x)| \leq 1$ für alle $x \in \mathbb{R}$ folgt

$$\big|\sin(x) - p_{2m+1}(x)\big| = \frac{1}{(2m+2)!}\underbrace{|\sin(z_x)|}_{\leq 1}|x|^{2m+2} \leq \frac{1}{(2m+2)!}|x|^{2m+2}. \tag{1.16}$$

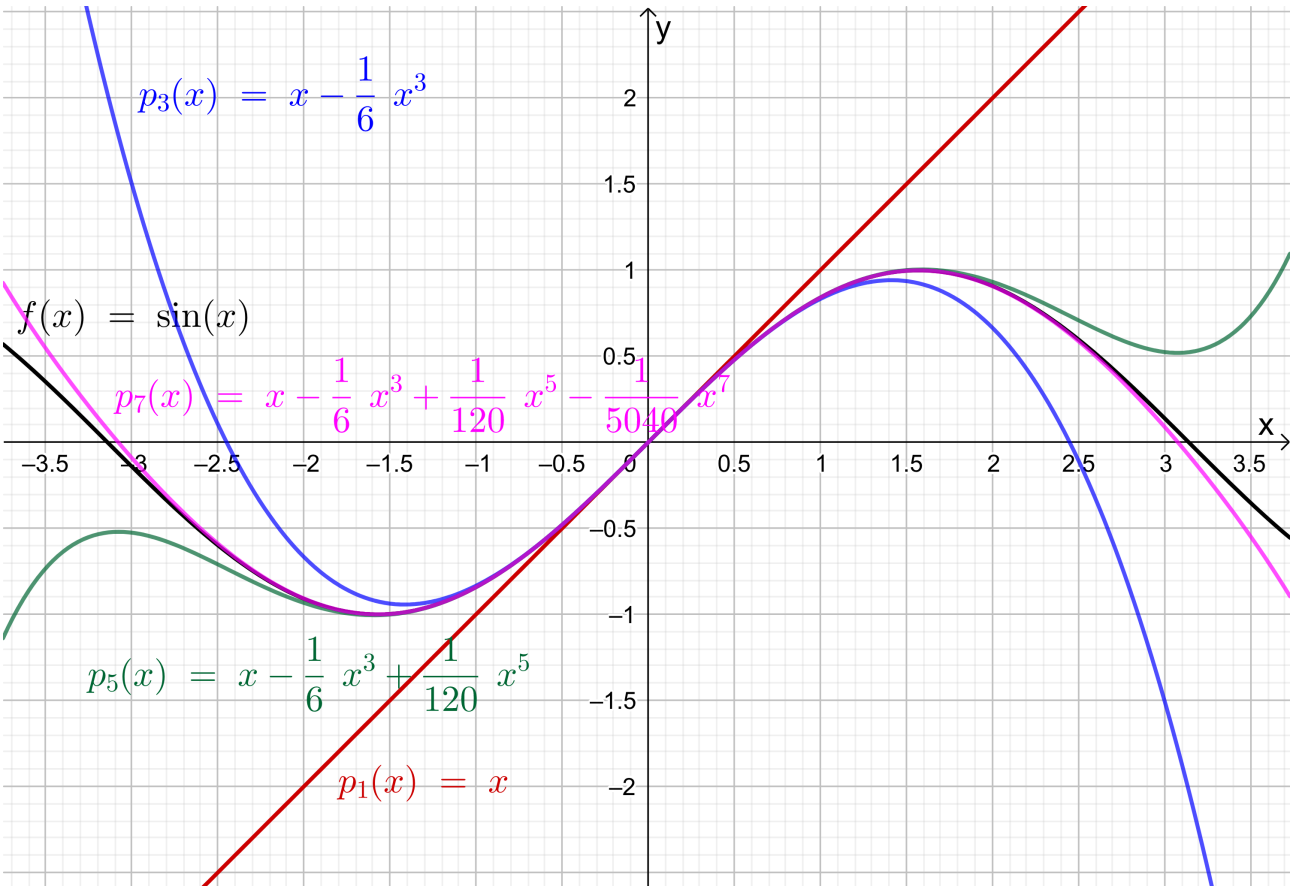


Abb. 1.4: Die Graphen von $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sin(x)$, (schwarz) und seinen Taylor-Polynomen $p_n(x)$ mit dem Entwicklungspunkt $x_0 = 0$ vom Grad $n = 1$ (rot), $n = 3$ (blau), $n = 5$ (grün) und $n = 7$ (violett).

Frage: Wie groß muss man m wählen, um $\sin(x)$ durch sein Taylor-Polynom p_{2m+1} mit dem Entwicklungspunkt $x_0 = 0$ auf dem Intervall $[-2; 2]$ garantiert auf (mindestens) zwei Nachkommastellen genau anzunähern?

Antwort: Für $x \in [-2; 2]$, also $|x| \leq 2$, folgt aus (1.16)

$$\left|\sin(x) - p_{2m+1}(x)\right| \leq \frac{1}{(2m+2)!} \underbrace{|x|^{2m+2}}_{\leq 2^{2m+2}} \leq \frac{2^{2m+2}}{(2m+2)!} \tag{1.17}$$

Annäherung auf zwei Nachkommastellen genau bedeutet, dass der absolute Fehler (also die Abweichung vom exakten Wert) betraglich kleiner oder gleich $0{,}005 = 5 \cdot 10^{-3}$ ist. In (1.17) sollte also gelten:

$$\frac{2^{2m+2}}{(2m+2)!} \leq 5 \cdot 10^{-3} \quad \Longleftrightarrow \quad \frac{1}{5 \cdot 10^{-3}} = 2 \cdot 10^2 = 200 \leq \frac{(2m+2)!}{2^{2m+2}}$$

Durch Probieren mit $m = 1, 2, 3, 4$ findet man, dass diese Ungleichung zuerst für $m = 4$ mit $2m+2 = 10$ erfüllt ist mit $10!/2^{10} \doteq 3543{,}75$ (zum Vergleich für $m = 3$,

also $2m+2=8$, ist $8!/2^8 = 157{,}5$). Diese Abschätzung garantiert, dass p_9 die Funktion $\sin(x)$ auf dem Intervall $[-2;2]$ mindestens auf zwei Nachkommastellen genau annähert.

Setzt man in (1.17) $m=4$ ein, so erhält man

$$\big|\sin(x) - p_9(x)\big| \leq \frac{2^{10}}{10!} \doteq 2{,}82 \cdot 10^{-4} = 0{,}000282 \quad \text{für alle } x \in [-2;2], \tag{1.18}$$

d.h. die Näherung ist sogar bis auf drei Nachkommastellen genau.

Für die Annäherung von $\sin(x)$ durch $p_{2m+2}(x)$ liefert der Satz vom Taylor die folgende Aussage: Zu jedem $x \in \mathbb{R}$ gibt es ein z_x zwischen 0 und x, so dass gilt

$$\sin(x) = p_{2m+2}(x) + \frac{1}{(2m+3)!}\, f^{(2m+3)}(z_x)\,(x-0)^{2m+3}, \tag{1.19}$$

wobei $f^{(2m+3)}(z_x) = \cos(z_x)$ für ungerades m und $f^{(2m+3)}(z_x) = -\cos(z_x)$ für gerades m. Umstellen und den Absolutbetrag Anwenden liefert:

$$\begin{aligned} \sin(x) - p_{2m+2}(x) &= \frac{1}{(2m+3)!}\, f^{(2m+3)}(z_x)\,(x-0)^{2m+3} \\ \Longrightarrow \qquad \big|\sin(x) - p_{2m+2}(x)\big| &= \frac{1}{(2m+3)!}\, |\cos(z_x)|\, |x|^{2m+3} \end{aligned}$$

Wegen $|\cos(x)| \leq 1$ für alle $x \in \mathbb{R}$ folgt

$$\big|\sin(x) - p_{2m+2}(x)\big| = \frac{1}{(2m+3)!} \underbrace{|\cos(z_x)|}_{\leq 1} |x|^{2m+3} \leq \frac{1}{(2m+3)!}\, |x|^{2m+3}. \tag{1.20}$$

Frage: Wie groß muss man m wählen, um $\sin(x)$ durch sein Taylor-Polynom p_{2m+2} mit dem Entwicklungspunkt $x_0 = 0$ auf dem Intervall $[-2;2]$ garantiert auf (mindestens) zwei Nachkommastellen genau anzunähern?

Antwort: Für $x \in [-2;2]$, also $|x| \leq 2$, folgt aus (1.16)

$$\big|\sin(x) - p_{2m+2}(x)\big| \leq \frac{1}{(2m+3)!} \underbrace{|x|^{2m+3}}_{\leq 2^{2m+3}} \leq \frac{2^{2m+3}}{(2m+3)!} \tag{1.21}$$

Annäherung auf zwei Nachkommastellen genau bedeutet, dass der absolute Fehler (also die Abweichung vom exakten Wert) betraglich kleiner oder gleich $0{,}005 = 5 \cdot 10^{-3}$ ist. In (1.21) sollte also gelten:

$$\frac{2^{2m+3}}{(2m+3)!} \leq 5 \cdot 10^{-3} \quad \Longleftrightarrow \quad \frac{1}{5 \cdot 10^{-3}} = 2 \cdot 10^2 = 200 \leq \frac{(2m+3)!}{2^{2m+3}}$$

Durch Probieren mit $m = 1, 2, 3$ findet man, dass diese Ungleichung zuerst für $m = 3$ mit $2m + 3 = 9$ erfüllt ist mit $9!/2^9 = 708{,}75$ (zum Vergleich für $m = 2$, also $2m + 3 = 7$, ist $7!/2^7 = 39{,}375$). Diese Abschätzung garantiert, dass p_8 die Funktion $\sin(x)$ auf $[-2; 2]$ mindestens auf zwei Nachkommastellen genau annähert.

Setzt man in (1.21) $m = 3$ ein, so erhält man

$$\big|\sin(x) - p_8(x)\big| \leq \frac{2^9}{9!} \doteq 1{,}41 \cdot 10^{-3} = 0{,}00141 \qquad \text{für alle } x \in [-2; 2]. \quad (1.22)$$

Da $p_8 = p_7$ gilt, folgt aus (1.22) auch, dass bereits mit dem Taylor-Polynom p_7 vom Grad 7 die Funktion $\sin(x)$ auf $[-2; 2]$ auf zwei Nachkommastellen genau genähert wird. Dieses war aus der Fehlerabschätzung (1.17) nicht ersichtlich. ♠

Die nachfolgenden Spezialfälle des Satzes von Taylor werden in den Natur- und Ingenieurwissenschaften häufig genutzt, um eine komplizierte (einmal, zweimal bzw. dreimal stetig differenzierbare) Funktion durch ein konstantes, ein lineares bzw. ein quadratisches Polynom anzunähern.

Bemerkung 1.26. (Spezialfälle des Satzes von Taylor)
Seien I ein offenes Intervall, $x_0 \in I$ und $f : I \to \mathbb{R}$ dreimal stetig differenzierbar. Dann gilt nach dem Satz von Taylor:

(1) $n = 0$: Zu jedem $x \in I$ gibt es ein z_x zwischen x und x_0, so dass gilt

$$f(x) = f(x_0) + f'(z_x)\,(x - x_0).$$

Dieses ist eine Umformulierung des Mittelwertsatzes der Differentialrechnung (vgl. Satz 1.4).

(2) $n = 1$: Zu jedem $x \in I$ gibt es ein z_x zwischen x und x_0, so dass gilt

$$f(x) = \underbrace{f(x_0) + f'(x_0)\,(x - x_0)}_{\text{Gleichung der Tangente an } f \text{ in } x_0} + \frac{1}{2}\,f''(z_x)\,(x - x_0)^2.$$

(3) $n = 2$: Zu jedem $x \in I$ gibt es ein z_x zwischen x und x_0, so dass gilt

$$f(x) = \underbrace{f(x_0) + f'(x_0)\,(x - x_0) + \frac{1}{2}\,f''(x_0)\,(x - x_0)^2}_{\text{quadratisches Taylor-Polynom}} + \frac{1}{6}\,f^{(3)}(z_x)\,(x - x_0)^3.$$

1.4 Effiziente Auswertung von Polynomen

Möchte man ein Polynom mit dem Computer an sehr vielen Stellen auswerten (z.B. um seinen Graphen zeichnen zu lassen), so sollte dieses **möglichst effizient**, d.h. **mit möglichst wenigen Multiplikationen/Divisionen und Additionen/Subtraktionen**, geschehen. Betrachten wir dazu zunächst ein Beispiel:

$$p(x) = 2 - 3\,x + 4\,x^2 - 5\,x^3 + 6\,x^4 - 7\,x^5$$

Führt man alle Rechenoperationen der Reihe nach aus, so benötigt man

$$5 \text{ Additionen/Subtraktionen} \quad \text{und} \quad 1 + 2 + 3 + 4 + 5 = 15 \text{ Multiplikationen.}$$

Dabei wird $c\,x^k$ mit k Multiplikationen als $c \cdot \underbrace{x \cdot x \cdot \ldots \cdot x}_{k\text{-mal}}$ berechnet.

Effizienter wäre es, wenn man zuerst die Potenzen von x wie folgt berechnet:

$$x^2 = x \cdot x, \qquad x^3 = x \cdot x^2, \qquad x^4 = x \cdot x^3, \qquad x^5 = x \cdot x^4.$$

Dieses erfordert 4 Multiplikationen. Weiter braucht man 5 Multiplikationen um x, x^2, x^3, x^4 bzw. x^5 jeweils mit dem Koeffizienten $-3, 4, -5, 6$ bzw. -7 zu multiplizieren. Die Anzahl der Additionen/Subtraktionen bleibt 5. Mit dieser Vorgehensweise benötigen wir also

$$5 \text{ Additionen/Subtraktionen} \quad \text{und} \quad 4 + 5 = 9 \text{ Multiplikationen.}$$

Noch cleverer ist die folgende „verschachtelte“ Multiplikation

$$p(x) = 2 + x\,(-3 + x\,(4 + x\,(-5 + x\,(6 - 7\,x)))),$$

von deren Richtigkeit man sich leicht durch Ausmultiplizieren überzeugt. Hier benötigen wir nur

$$5 \text{ Additionen/Subtraktionen} \quad \text{und} \quad 5 \text{ Multiplikationen.}$$

Wir formulieren die „verschachtelte“ Multiplikation nun allgemein:

Hilfssatz 1.27. (Horner Schema)

Ein Polynom vom Grad n,

$$p : \mathbb{R} \to \mathbb{R}, \quad p(x) = a_0 + a_1\,x + a_2\,x^2 + \ldots + a_{n-1}\,x^{n-1} + a_n\,x^n,$$

wobei $a_n \neq 0$, wertet man effizient mit dem ***Horner-Schema*** *mit nur n Additionen/Subtraktionen und n Multiplikationen wie folgt aus:*

$$p(x) = a_0 + x\,(a_1 + x\,(a_2 + \ldots + x\,(a_{n-1} + a_n\,x)\ldots)\,)$$

Dabei berechnet man für $x = z$ den Funktionswert $p(z)$ mit einer Folge von Koeffizienten $b_n, b_{n-1}, \ldots, b_1, b_0$ wie folgt:

$$\begin{aligned} b_n &= a_n, \\ b_{n-1} &= a_{n-1} + z\,b_n, \\ b_{n-2} &= a_{n-2} + z\,b_{n-1}, \\ &\vdots \\ b_1 &= a_1 + z\,b_2, \\ b_0 &= a_0 + z\,b_1. \end{aligned}$$

Dann gelten

$$\begin{aligned} p(z) = b_0 \qquad \textit{und} \qquad p(x) &= (x - z)\,q(x) + b_0 \\ \textit{mit} \quad q(x) &= b_1 + b_2\,x + b_3\,x^2 + \ldots + b_n\,x^{n-1}. \end{aligned} \tag{1.23}$$

Die praktische Berechnung von $p(z)$ führt man mit einer Tabelle durch:

Koeffizienten von p	a_n	a_{n-1}	a_{n-2}	$\cdots$	a_1	a_0
$x = z$	–	$z\,b_n$	$z\,b_{n-1}$	$\cdots$	$z\,b_2$	$z\,b_1$
addiere	b_n	b_{n-1}	b_{n-2}	$\cdots$	b_1	b_0

Wichtig: Ist ein Koeffizient a_k des Polynoms p null, so muss man diesen in der Tabelle mit 0 auflisten und darf ihn keinesfalls weglassen!

Die zweite Formel in (1.23) ist nicht offensichtlich. Wir beweisen diese auf einem Übungszettel.

Beispiel 1.28. (Horner-Schema)
Wir berechnen

$$p(x) = 2 - 3\,x + 4\,x^2 - 5\,x^3 + 6\,x^4 - 7\,x^5$$

für $x_1 = 1$, $x_2 = -1$ und $x_3 = 2$ mit dem Horner-Schema.

Koeff. von p	-7	6	-5	4	-3	2
$x=1$	–	$1\cdot(-7)$ $=-7$	$1\cdot(-1)$ $=-1$	$1\cdot(-6)$ $=-6$	$1\cdot(-2)$ $=-2$	$1\cdot(-5)$ $=-5$
addiere	-7	$6+(-7)$ $=-1$	$-5+(-1)$ $=-6$	$4+(-6)$ $=-2$	$-3+(-2)$ $=-5$	$2+(-5)$ $=-3$

Also gelten $p(1)=-3$ und $p(x)=(x-1)\,(-5-2\,x-6\,x^2-x^3-7\,x^4)-3$.

Koeff. von p	-7	6	-5	4	-3	2
$x=-1$	–	$(-1)\cdot(-7)$ $=7$	$(-1)\cdot13$ $=-13$	$(-1)\cdot(-18)$ $=18$	$(-1)\cdot22$ $=-22$	$(-1)\cdot(-25)$ $=25$
addiere	-7	$6+7$ $=13$	$-5+(-13)$ $=-18$	$4+18$ $=22$	$-3+(-22)$ $=-25$	$2+25$ $=27$

Also gelten $p(-1)=27$ und $p(x)=(x+1)\,(-25+22\,x-18\,x^2+13\,x^3-7\,x^4)+27$.

Koeff. von p	-7	6	-5	4	-3	2
$x=2$	–	$2\cdot(-7)$ $=-14$	$2\cdot(-8)$ $=-16$	$2\cdot(-21)$ $=-42$	$2\cdot(-38)$ $=-76$	$2\cdot(-79)$ $=-158$
addiere	-7	$6+(-14)$ $=-8$	$-5+(-16)$ $=-21$	$4+(-42)$ $=-38$	$-3+(-76)$ $=-79$	$2+(-158)$ $=-156$

Also gelten $p(2)=-156$ und $p(x)=(x-2)\,(-79-38\,x-21\,x^2-8\,x^3-7\,x^4)-156$.

Wir üben das Horner-Schema auf einem Übungszettel. ♠

1.5 Eine Anwendung des Taylor-Polynoms

Im letzten Teilkapitel berechnen wir mit Hilfe des Taylor-Polynoms Integrale angenähert. Diese Anwendung macht (neben dem bereits in Teilkapitel 1.2 erwähnten Problem der Funktionsauswertung) deutlich, wieso es sehr nützlich ist, eine Funktion durch ein Polynom anzunähern.

Beispiel 1.29. (Integral angenähert berechnen)
Wir wollen das Integral

$$A = \int_0^1 x\, e^x \,\mathrm{d}x \tag{1.24}$$

mit Hilfe eines geeigneten Taylor-Polynoms angenähert berechnen.

Dieses Integral kann man mit partieller Integration leicht exakt berechnen:

$$\int_0^1 \underbrace{x}_{=f(x)}\ \underbrace{e^x}_{=g'(x)}\,\mathrm{d}x = \Big[\ \underbrace{x\, e^x}_{=f(x)\,g(x)}\ \Big]_{x=0}^{x=1} - \int_0^1 \underbrace{1}_{=f'(x)} \cdot \underbrace{e^x}_{=g(x)}\,\mathrm{d}x$$
$$= \Big[x\, e^x\Big]_{x=0}^{x=1} - \int_0^1 e^x\,\mathrm{d}x = e - 0 - \Big[e^x\Big]_{x=0}^{x=1} = e - \big(e - e^0\big) = e^0 = 1.$$

Aus Beispiel 1.19 wissen wir, dass das Taylor-Polynom p_n von $f(x) = e^x$ vom Grad n mit dem Entwicklungspunkt $x_0 = 0$ durch

$$p_n(x) = 1 + x + \frac{1}{2!}\,x^2 + \frac{1}{3!}\,x^3 + \ldots + \frac{1}{n!}\,x^n = \sum_{k=0}^{n} \frac{1}{k!}\,x^k \tag{1.25}$$

gegeben ist. Der absolute Fehler der Näherung von $f(x) = e^x$ durch $p_n(x)$ wird durch den Satz von Taylor beschrieben (vgl. auch Beispiel 1.24):

Zu jedem $x \in \mathbb{R}$ gibt es ein z_x zwischen x und $x_0 = 0$, so dass gilt

$$e^x = p_n(x) + \underbrace{\frac{e^{z_x}}{(n+1)!}\,x^{n+1}}_{=r_n(x)} = \underbrace{\sum_{k=0}^{n} \frac{1}{k!}\,x^k}_{=p_n(x)} + \underbrace{\frac{e^{z_x}}{(n+1)!}\,x^{n+1}}_{=r_n(x)}. \tag{1.26}$$

(Bei der Darstellung des Restglieds $r_n(x)$ wurde benutzt, dass $f^{(n+1)}(x) = e^x$ für die Exponentialfunktion $f(x) = e^x$ gilt.)

Multipliziert man die Näherung p_n von e^x in (1.25) mit x so erhält man eine Näherung für den Integranden $x\, e^x$ des gegebenen Integrals (1.24)

$$q_n(x) = x\, p_n(x) = x + x^2 + \frac{1}{2!}\,x^3 + \frac{1}{3!}\,x^4 + \ldots + \frac{1}{n!}\,x^{n+1} = \sum_{k=0}^{n} \frac{1}{k!}\,x^{k+1}. \tag{1.27}$$

Die Näherung $p_n(x)$ von e^x sollte für hinreichend großes n in der Nähe von $x_0 = 0$ eine gute Genauigkeit haben. Also können wir auch erwarten, dass $q_n(x) = x\, p_n(x)$ auf dem Integrationsintervall $[0; 1]$ für hinreichend großes n eine gute Näherung

von $x\,e^x$ liefert. Den absoluten Fehler dieser Näherung können wir mit Hilfe von (1.26) angeben:

Zu jedem $x \in \mathbb{R}$ gibt es ein z_x zwischen x und $x_0 = 0$, so dass gilt

$$x\,e^x = \underbrace{x\,p_n(x)}_{=q_n(x)} + \underbrace{\frac{e^{z_x}}{(n+1)!}\,x^{n+2}}_{=x\,r_n(x)} = q_n(x) + \underbrace{\frac{e^{z_x}}{(n+1)!}\,x^{n+2}}_{=x\,r_n(x)}. \tag{1.28}$$

Ersetzen wir den Integranden $x\,e^x$ in dem gegebenen Integral (1.24) durch seine Näherung $q_n(x) = x\,p_n(x)$ (siehe (1.27)) so erhalten wir die folgende Näherung für das Integral:

$$\begin{aligned}
\int_0^1 x\,e^x\,\mathrm{d}x &\approx \int_0^1 x\,p_n(x)\,\mathrm{d}x = \int_0^1 q_n(x)\,\mathrm{d}x \\
&= \int_0^1 \left(x + x^2 + \frac{1}{2!}\,x^3 + \frac{1}{3!}\,x^4 + \ldots + \frac{1}{n!}\,x^{n+1}\right)\,\mathrm{d}x \\
&= \left[\frac{1}{2}\,x^2 + \frac{1}{3}\,x^3 + \frac{1}{2!\,4}\,x^4 + \frac{1}{3!\,5}\,x^5 + \ldots + \frac{1}{n!\,(n+2)}\,x^{n+2}\right]_{x=0}^{x=1} \\
&= \frac{1}{2} + \frac{1}{3} + \frac{1}{2!\,4} + \frac{1}{3!\,5} + \ldots + \frac{1}{n!\,(n+2)} = \sum_{k=0}^{n} \frac{1}{k!\,(k+2)}
\end{aligned}$$

Die Näherung des Integrals A in (1.24) mit dem Wert $A = 1$ ist also

$$A_n = \int_0^1 x\,p_n(x)\,\mathrm{d}x = \sum_{k=0}^{n} \frac{1}{k!\,(k+2)}. \tag{1.29}$$

In Tabelle 1.1 haben wir die Näherungen A_n für $n = 0, 1, 2, \ldots, 8$ berechnet (mit Rundung der Ergebnisse auf sechs Nachkommastellen). Wir beobachten, dass wir bereits bei einem Polynom vom Grad 8 bei Rundung auf sechs Nachkommastellen das „exakte“ Ergebnis erhalten.

Eine theoretische Information über die (mindestens) garantierte Qualität der Näherung des Integrals (1.24) durch A_n erhält man mit Hilfe von (1.28). Integriert man in (1.28) über das Intervall $[0; 1]$ so erhält man:

$$\int_0^1 x\,e^x\,\mathrm{d}x = \underbrace{\int_0^1 q_n(x)\,\mathrm{d}x}_{=A_n} + \underbrace{\int_0^1 \frac{e^{z_x}}{(n+1)!}\,x^{n+2}\,\mathrm{d}x}_{=R_n}$$

$$\iff \qquad \int_0^1 x\,e^x\,\mathrm{d}x - \underbrace{\int_0^1 q_n(x)\,\mathrm{d}x}_{=A_n} = \underbrace{\int_0^1 \frac{e^{z_x}}{(n+1)!}\,x^{n+2}\,\mathrm{d}x}_{=R_n}$$

n	A_n
0	0,500000
1	0,833333
2	0,958333

n	A_n
3	0,991667
4	0,998611
5	0,999802

n	A_n
6	0,999975
7	0,999997
8	1,000000

Tabelle 1.1: Annäherung des Integrals (1.24) mit dem Wert $A = 1$ durch die Näherungen A_n (siehe (1.29)) mit Rundung auf sechs Nachkommastellen.

$$\implies \qquad \left| \int_0^1 x\, e^x \,\mathrm{d}x - \underbrace{\int_0^1 q_n(x)\,\mathrm{d}x}_{=A_n} \right| = \left| \underbrace{\int_0^1 \frac{e^{z_x}}{(n+1)!}\, x^{n+2}\,\mathrm{d}x}_{=R_n} \right| \tag{1.30}$$

Der Ausdruck auf der rechten Seite von (1.30) gibt den Absolutbetrag $|R_n|$ des Fehlers R_n der Näherung des Integrals (1.24) durch A_n an. Wir wollen diesen Fehler nun abschätzen:

$$\begin{aligned} |R_n| &= \left| \int_0^1 \frac{e^{z_x}}{(n+1)!}\, x^{n+2}\,\mathrm{d}x \right| \leq \int_0^1 \left| \frac{e^{z_x}}{(n+1)!}\, x^{n+2} \right| \mathrm{d}x \\ &= \int_0^1 \frac{|e^{z_x}|}{(n+1)!}\, |x|^{n+2}\,\mathrm{d}x = \int_0^1 \frac{e^{z_x}}{(n+1)!}\, x^{n+2}\,\mathrm{d}x \end{aligned} \tag{1.31}$$

Der letzte Schritt folgt dabei, weil der Integrand des Integrals R_n nicht-negativ ist (da $e^t > 0$ für alle $t \in \mathbb{R}$ und $x^{n+2} \geq 0$ für alle $x \in [0;1]$).

Da wir nicht wissen, wie z_x von x abhängt, können wir das Integral in (1.31) nicht direkt berechnen, sondern wir müssen den Integranden im Integral in (1.31) so geeignet nach oben abschätzen, dass die Abhängigkeit von z_x verschwindet.

$$\frac{e^{z_x}}{(n+1)!}\, x^{n+2} \leq \frac{e}{(n+1)!}\, x^{n+2} \qquad \text{für alle } x \in [0;1],\ \text{weil } 0 \leq e^{z_x} \leq e^1. \tag{1.32}$$

Dabei haben wir genutzt, dass e^x streng monoton wachsend ist und dass aus $0 \leq z_x \leq x$ (da z_x zwischen 0 und x liegt) für $x \in [0;1]$ folgt, dass $z_x \in [0;1]$ ist.

Mit (1.32) folgt aus (1.31), dass

$$|R_n| \leq \int_0^1 \frac{e}{(n+1)!}\, x^{n+2}\,\mathrm{d}x = \left[\frac{e}{(n+1)!\,(n+3)}\, x^{n+3} \right]_{x=0}^{x=1} = \frac{e}{(n+1)!\,(n+3)}\,. \tag{1.33}$$

Die Abschätzung (1.33) des Restglieds und (1.30) ergeben also die folgende Feh-

lerabschätzung für die Näherung des Integrals:

$$\Big| \underbrace{\int_0^1 x\, e^x \,\mathrm{d}x}_{=A} - \underbrace{\int_0^1 q_n(x)\,\mathrm{d}x}_{=A_n} \Big| \leq \frac{e}{(n+1)!\,(n+3)}\,. \tag{1.34}$$

Bei einem bis auf sechs Nachkommastellen genauen Ergebnis ist der absolute Fehler betraglich kleiner oder gleich $0{,}5 \cdot 10^{-6}$. In Tabelle 1.1 sehen wir, dass diese Genauigkeit bereits für $n = 8$ erreicht wird.

Frage: Ab welchem Wert für n garantiert uns (1.34), dass die Näherung bis auf sechs Nachkommastellen genau ist? Ab welchem Wert von n gilt also

$$\big|A - A_n\big| \leq \frac{e}{(n+1)!\,(n+3)} \leq 0,5 \cdot 10^{-6}\,? \tag{1.35}$$

Antwort: Umformen liefert:

$$\frac{e}{(n+1)!\,(n+3)} \leq 0{,}5 \cdot 10^{-6} \qquad \Longleftrightarrow \qquad \frac{e}{0{,}5 \cdot 10^{-6}} \leq (n+1)!\,(n+3)$$

$$\Longleftrightarrow \qquad \underbrace{2\,e \cdot 10^6}_{\doteq 5{,}436564 \cdot 10^6} \leq (n+1)!\,(n+3) \tag{1.36}$$

Wir finden $(8+1)!\,(8+3) = 3{,}991680 \cdot 10^6$ und $(9+1)!\,(9+3) = 4{,}35456 \cdot 10^7$, d.h. (1.36) ist erst für $n \geq 9$ erfüllt. Nach der Fehlerabschätzung (1.35) wird also eine (mindestens) bis auf sechs Nachkommastellen genaue Näherung erst ab $n = 9$ garantiert (obwohl diese in der Praxis bereits bei $n = 8$ auftritt). ♠

Betrachten wir noch ein weiteres Beispiel, bei dem das zu berechnende Integral nicht mehr elementar zu berechnen ist.

Beispiel 1.30. (Integral angenähert berechnen)
Wir wollen das folgende Integral angenähert berechnen:

$$A = \int_0^1 \frac{\sin(x)}{x}\,\mathrm{d}x \tag{1.37}$$

Wegen $\lim\limits_{x>0,\, x\to 0} \dfrac{\sin(x)}{x} = 1$ hat der Integrand keine Singularität in $x = 0$ und das Integral ist berechenbar und hat einen endlichen Wert. Dieses Integral lässt sich aber nicht bequem mit einer Substitution oder einer partiellen Integration berechnen! Daher ersetzen wir $\sin(x)$ durch sein Taylor-Polynom vom Grad $n = 2m+2$, vereinfachen den Integranden und berechnen dann das Integral und erhalten so

einen Näherungswert für A. Wir werden auch den absoluten Fehler dieser Näherung untersuchen.

Aus Beispiel 1.25 wissen wir, dass das Taylor-Polynom p_{2m+2} vom Grad $2n+2$ von $\sin(x)$ mit dem Entwicklungspunkt $x_0 = 0$ durch (vgl. (1.15))

$$p_{2m+2}(x) = \sum_{k=0}^{m} \frac{(-1)^k}{(2k+1)!}\, x^{2k+1} = x - \frac{1}{3!}\, x^3 + \frac{1}{5!}\, x^5 - \ldots + \frac{(-1)^m}{(2m+1)!}\, x^{2m+1}$$

gegeben ist. Damit folgt für x dicht bei $x_0 = 0$, dass gilt

$$\begin{aligned} \frac{\sin(x)}{x} \approx \frac{p_{2m+2}(x)}{x} &= \frac{1}{x}\left(x - \frac{1}{3!}\, x^3 + \frac{1}{5!}\, x^5 - \ldots + \frac{(-1)^m}{(2m+1)!}\, x^{2m+1}\right) \\ &= 1 - \frac{1}{3!}\, x^2 + \frac{1}{5!}\, x^4 - \ldots + \frac{(-1)^m}{(2m+1)!}\, x^{2m}. \end{aligned} \tag{1.38}$$

Einsetzen von (1.38) in das Integral (1.37) liefert:

$$\begin{aligned} \int_0^1 \frac{\sin(x)}{x}\, \mathrm{d}x \approx \int_0^1 \frac{p_{2m+2}(x)}{x}\, \mathrm{d}x &= \int_0^1 \left(1 - \frac{1}{3!}\, x^2 + \frac{1}{5!}\, x^4 - \ldots + \frac{(-1)^m}{(2m+1)!}\, x^{2m}\right) \mathrm{d}x \\ &= \left[x - \frac{1}{3!\,3}\, x^3 + \frac{1}{5!\,5}\, x^5 - \ldots + \frac{(-1)^m}{(2m+1)!\,(2m+1)}\, x^{2m+1}\right]_{x=0}^{x=1} \\ &= 1 - \frac{1}{3!\,3} + \frac{1}{5!\,5} - \ldots + \frac{(-1)^m}{(2m+1)!\,(2m+1)} \end{aligned}$$

Wir erhalten also für den Wert des Integrals die folgende Näherung:

$$\begin{aligned} A_{2m+2} = \int_0^1 \frac{p_{2m+2}(x)}{x}\, \mathrm{d}x &= 1 - \frac{1}{3!\,3} + \frac{1}{5!\,5} - \ldots + \frac{(-1)^m}{(2m+1)!\,(2m+1)} \\ &= \sum_{k=0}^{m} \frac{(-1)^k}{(2k+1)!\,(2k+1)} \end{aligned} \tag{1.39}$$

Frage: Was können wir über den absoluten Fehler der Näherung (1.39) aussagen?

Nach dem Satz von Taylor gilt (vgl. (1.19)):

Zu jedem $x \in \mathbb{R}$ gibt es ein z_x zwischen 0 und x, so dass gilt

$$\sin(x) = p_{2m+2}(x) + \frac{1}{(2m+3)!}\, f^{(2m+3)}(z_x)\, x^{2m+3}, \tag{1.40}$$

wobei $f^{(2m+3)}(z_x) = \cos(z_x)$ für ungerades m und $f^{(2m+3)}(z_x) = -\cos(z_x)$ für gerades m ist. Division in (1.40) durch x liefert

$$\frac{\sin(x)}{x} = \frac{p_{2m+2}(x)}{x} + \frac{1}{(2m+3)!}\, f^{(2m+3)}(z_x)\, x^{2m+2}. \tag{1.41}$$

Integration in (1.41) über das Intervall $[0;1]$ und Umstellen liefert:

$$\underbrace{\int_0^1 \frac{\sin(x)}{x}\,\mathrm{d}x}_{=A} = \underbrace{\int_0^1 \frac{p_{2m+2}(x)}{x}\,\mathrm{d}t}_{=A_{2m+2}} + \underbrace{\int_0^1 \frac{1}{(2m+3)!}\, f^{(2m+3)}(z_x)\, x^{2m+2}\,\mathrm{d}x}_{=R_{2m+2}}$$

$$\iff \qquad A - A_{2m+2} = R_{2m+2} = \int_0^1 \frac{1}{(2m+3)!}\, f^{(2m+3)}(z_x)\, x^{2m+2}\,\mathrm{d}x \tag{1.42}$$

Wir nehmen auf beiden Seiten in (1.42) den Absolutbetrag und schätzen ab:

$$\begin{aligned} \big|A - A_{2m+2}\big| = |R_{2m+2}| &= \left| \int_0^1 \frac{1}{(2m+3)!}\, f^{(2m+3)}(z_x)\, x^{2m+2}\,\mathrm{d}x \right| \\ &\le \int_0^1 \frac{1}{(2m+3)!}\, |f^{(2m+3)}(z_x)|\, |x|^{2m+2}\,\mathrm{d}x \\ &= \int_0^1 \frac{1}{(2m+3)!}\, |f^{(2m+3)}(z_x)|\, x^{2m+2}\,\mathrm{d}x, \end{aligned} \tag{1.43}$$

wobei wir in der letzten Umformung genutzt haben, dass $|x| = x$ für alle $x \in [0;1]$ gilt. Wegen $f^{(2m+3)}(z_x) = \cos(z_x)$ für ungerades m und $f^{(2m+3)}(z_x) = -\cos(z_x)$ für gerades m gilt mit $|\cos(t)| \le 1$ für alle $t \in \mathbb{R}$ in (1.43)

$$\begin{aligned} \big|A - A_{2m+2}\big| &\le \int_0^1 \frac{1}{(2m+3)!}\, \underbrace{|\cos(z_x)|}_{\le 1}\, x^{2m+2}\,\mathrm{d}x \le \int_0^1 \frac{1}{(2m+3)!}\, x^{2m+2}\,\mathrm{d}x \\ &= \left[\frac{1}{(2m+3)!\,(2m+3)}\, x^{2m+3} \right]_{x=0}^{x=1} = \frac{1}{(2m+3)!\,(2m+3)} \end{aligned} \tag{1.44}$$

Nach (1.44) gilt für den absoluten Fehler der Näherung A_{2m+2} des durch (1.37) gegebenen Integrals mit Wert A:

$$\big|A - A_{2m+2}\big| \le \frac{1}{(2m+3)!\,(2m+3)} \tag{1.45}$$

Frage: Wie groß muss der Grad $2m+2$ des Taylor-Polynoms p_{2m+2} sein, damit die durch (1.39) gegebene Näherung das Integral mit einem absoluten Fehler von betraglich höchstens $0{,}5 \cdot 10^{-4}$ (also auf mindestens vier Nachkommastellen genau) berechnet wird?

Antwort: Durch Berechnen der rechten Seiten in (1.45) für $m = 1, 2$ findet man, dass für $m = 2$, also $2m + 2 = 6$ zum ersten Mal

$$\left|A - A_{2m+2}\right| \leq \frac{1}{(2 \cdot 2 + 3)!\,(2 \cdot 2 + 3)} = \frac{1}{7!\,7} \doteq 0,283 \cdot 10^{-4}$$

gilt. Also können wir bereits unter Verwendung des Taylor-Polynoms $p_5 = p_6$ von $\sin(x)$ vom Grad 5 mit dem Entwicklungspunkt $x_0 = 0$ den Wert des Integrals (1.37) auf (mindestens) vier Nachkommastellen genau berechnen. ♠

Was wir im vorigen Beispiel gesehen haben, ist exemplarisch für eine Vorgehensweise in der Numerik (oder Numerischen Mathematik):

Wenn man ein kompliziertes Problem (wie das Integral im vorigen Beispiel) **nicht exakt lösen kann, so ersetzt man es durch ein einfacheres Problem, welches das komplizierte Problem angenähert löst.** Im vorigen Beispiel wurde der Integrand durch ein geeignetes Polynom ersetzt, so dass sich das Integral leicht berechnen lies. **Dabei ist es wichtig, dass man angeben kann, wie gut die Lösung des einfacheren Problems die Lösung des komplizierten Problems garantiert (mindestens) annähert.** Im vorigen Beispiel haben wir den absoluten Fehler der Näherung mit Hilfe der Satzes von Taylor untersuchen können.

KAPITEL 2

Fehler und Computer-Arithmetik

2.1 Gleitkommadarstellung

Auf Computern gibt es zwei Möglichkeiten eine Zahl zu speichern: im Integer-Format für ganze Zahlen und allgemeiner im **Gleitkomma-Format** (auch Gleitpunkt-Format genannt) für Zahlen, die keine ganzen Zahlen sind. Auf Computern werden Zahlen im Dualsystem (mit der Basis 2) im Gleitkomma-Format gespeichert, aber der Einfachheit halber besprechen wir hier nur die **(normalisierte) Gleitkommadarstellung im Dezimalsystem**, also mit der Basis 10. Diese ist etwas einfacher zugänglich, weil wir im alltäglichen Leben ständig mit dem Dezimalsystem arbeiten.

Jede reelle Zahl $x \neq 0$ kann **im Dezimalsystem eindeutig** in der folgenden **normalisierten Gleitkommadarstellung** geschrieben werden:

$$\boxed{x = \sigma \cdot \widetilde{x} \cdot 10^e} \tag{2.1}$$

mit dem **Vorzeichen** $\sigma \in \{-1, +1\}$, der **Mantisse** $1 \leq \widetilde{x} < 10$ und dem **Exponenten** $e \in \mathbb{Z}$.

Beispiel 2.1. (Darstellung von Zahlen im Dezimalsystem)
Unter wurden drei reelle Zahl in der eindeutigen Darstellung (2.1) angegeben:

$$134{,}26 = +1{,}3426 \cdot 10^2 \qquad \text{(hier: } \sigma = +1;\ \widetilde{x} = 1{,}3426;\ e = 2\text{)},$$

$$\frac{1}{3} = 0{,}\overline{3} = +3{,}\overline{3} \cdot 10^{-1} \qquad \text{(hier: } \sigma = +1;\ \widetilde{x} = 3{,}\overline{3};\ e = -1\text{)},$$

$$-\frac{8}{500} = -0{,}016 = -1{,}6 \cdot 10^{-2} \qquad \text{(hier: } \sigma = -1;\ \widetilde{x} = 1{,}6;\ e = -2\text{)},$$

wobei $3{,}\overline{3} = 3{,}3333\ldots$ ist. ♠

In einem Computer oder auch einen Taschenrechner kann man in (2.1) nur Mantissen $\widetilde{x}$ und Exponenten e einer endlichen Länge speichern. Dadurch entstehen gegebenenfalls Darstellungsfehler. Wir betrachten dieses an einem Beispiel.

Beispiel 2.2. (4-stellige Gleitkommadarstellung)
In einem Computer soll für den Exponenten in (2.1) gelten $-15 \leq e \leq 15$. Wenn für die Mantisse $\widetilde{x}$ maximal vier Ziffern (also Zahlen aus $\{0; 1; 2; \ldots; 9\}$) gespeichert werden, so erhalten wir eine **4-stellige normalisierte Gleitkommadarstellung** (oder **normalisierte Gleitkommadarstellung mit 4-stelliger Mantisse**): Alle im Computer in 4-stelliger normalisierter Gleitkommadarstellung darstellbaren Zahlen sind dann von der Form

$$\sigma \cdot \underbrace{(a_1, a_2\, a_3\, a_4)}_{=\widetilde{x}} \cdot 10^e$$

mit $e \in \mathbb{Z}$ mit $-15 \leq e \leq 15$ und wobei $\widetilde{x} = (a_1, a_2\, a_3\, a_4)$ die von den angegebenen Ziffern $a_1, a_2, a_3, a_4 \in \{0; 1; 2; 3; 4; 5; 6; 7; 8; 9\}$ gebildete vierstellige Zahl $\widetilde{x}$ mit $1 \leq \widetilde{x} < 10$ ist. Dabei steht das Komma nach der Ziffer a_1, und es gilt $a_1 \neq 0$.

Betrachten wir die Zahlen aus Beispiel 2.1: Die Zahl $-\frac{8}{500}$ lässt sich in der 4-stelligen Gleitkommadarstellung exakt darstellen:

$$-\frac{8}{500} = -0{,}016 = -1{,}600 \cdot 10^{-2} \qquad \text{(hier: } \sigma = -1;\ \widetilde{x} = 1{,}600;\ e = -2\text{)}$$

Die beiden Zahlen 134,26 und $\frac{1}{3}$ lassen sich dagegen nicht exakt in der 4-stelligen Gleitkommadarstellung darstellen, denn sowohl 134,26 als auch $0{,}\overline{3} = 0{,}33333\ldots$ haben mehr als vier Ziffern. Diese Zahlen können daher in 4-stelliger (normalisierter) Gleitkommadarstellung nur angenähert dargestellt werden.

Werden Näherungen dieser Zahlen durch (übliches) **Runden** bestimmt, so erhalten 134,26 bzw. $\frac{1}{3}$ jeweils die 4-stellige normalisierte Gleitkommadarstellung

$$+1{,}343 \cdot 10^2 \qquad \text{bzw.} \qquad +\,3{,}333 \cdot 10^{-1}.$$

Beim **Runden** auf eine k-stellige (normalisierte) Gleitkommadarstellung wird bei einer Zahl mit einer Mantisse der Länge $m > k$ die $(k+1)$-te Ziffer der Mantisse betrachtet. Ist diese eine $5, 6, 7, 8$ oder 9, so wird die k-te Ziffer der Mantisse um 1 erhöht (gegebenenfalls mit Übertrag) und die $(k+1)$-te bis m-te Ziffer der

Mantisse fallen weg. Ist die $(k+1)$-te Ziffer der Mantisse eine $0, 1, 2, 3$ oder 4, so fallen die $(k+1)$-te bis m-te Ziffer der Mantisse weg.

Werden Näherungen dieser Zahlen durch **Abschneiden** bestimmt, so erhalten 134,26 bzw. $\frac{1}{3}$ jeweils die 4-stellige normalisierte Gleitkommadarstellung

$$+1{,}342 \cdot 10^2 \qquad \text{bzw.} \qquad +3{,}333 \cdot 10^{-1}.$$

Beim **Abschneiden** auf eine k-stellige (normalisierte) Gleitkommadarstellung werden also bei einer Zahl mit einer Mantisse der Länge $m > k$ die $m-k$ letzten Ziffern der Mantisse einfach weggelassen (also „abgeschnitten“).

Die größte positive mit dieser 4-stelligen normalisierten Gleitkommadarstellung darstellbare Zahl ist $+9{,}999 \cdot 10^{15}$, und die kleinste positive mit dieser 4-stelligen Gleitkommadarstellung darstellbare Zahl ist $1{,}000 \cdot 10^{-15}$. ♠

2.2 Fehler

Berechnen wir eine quantitative Größe mit einem reellen Zahlenwert (z.B. ein bestimmtes Integral oder einen Funktionswert) nur angenähert, so ist der **absolute Fehler** definiert als:

$$\boxed{\text{(absoluter) Fehler} = \text{exakter Zahlenwert} - \text{angenäherter Zahlenwert}} \tag{2.2}$$

Ist der exakte Zahlenwert ungleich null, so ist der **relative Fehler** definiert als:

$$\boxed{\text{relativer Fehler} = \frac{\text{(absoluter) Fehler}}{\text{exakter Zahlenwert}}} \tag{2.3}$$

Beispiel 2.3. (absoluter und relativer Fehler)
Der exakte Zahlenwert sei die irrationale Zahl $x = \pi = 3{,}14159265\ldots$. Eine bekannte Näherung für $x = \pi$ ist

$$x_N = \frac{22}{7},$$

wobei der Index N in x_N für Näherung (von x) steht. Dann gilt mit Rundung auf 3-stellige Gleitkommadarstellung nach (2.2) und (2.3):

$$\text{absoluter Fehler:} \qquad \text{Fehler}\left(\frac{22}{7}\right) = x - x_N = \pi - \frac{22}{7} \doteq -1{,}26 \cdot 10^{-3},$$

relativer Fehler: $$\text{Rel}\left(\frac{22}{7}\right) = \frac{x - x_N}{x} = \frac{\pi - \frac{22}{7}}{\pi} \doteq -4{,}02 \cdot 10^{-4}.$$

Der relative Fehler von x_N setzt den absoluten Fehler in Beziehung zur Größenordnung von x. ♠

Definition 2.4. (absoluter und relativer Fehler)
Seien $x \in \mathbb{R}$ der exakte Zahlenwert einer quantitativen reellen Größe und $x_N \in \mathbb{R}$ ein Näherungswert für x. Dann gelten:

*(1) Der **absolute Fehler** der Näherung x_N von x ist definiert als*

$$\text{Fehler}(x_N) = x - x_N.$$

*(2) Der **relative Fehler** der Näherung x_N von x ist definiert als*

$$\text{Rel}(x_N) = \frac{\text{Fehler}(x_N)}{x} = \frac{x - x_N}{x}.$$

Betrachten wir ein weiteres Beispiel, das den Unterschied zwischen dem absoluten Fehler und dem relativen Fehler deutlich macht und gut illustriert, wieso der relative Fehler aussagekräftiger ist als der absolute Fehler.

Beispiel 2.5. (absoluter und relativer Fehler)

(a) Betrachten wir die Distanz x zwischen zwei Städten, die exakt $x = 100$ km von einander entfernt sind. Die Näherung dieser Entfernung (z.B. durch eine Messung) sei $x_N = 99$ km. Dann gilt

$$\text{Fehler}(x_N) = x - x_N = 100 \text{ km} - 99 \text{ km} = 1 \text{ km},$$
$$\text{Rel}(x_N) = \frac{\text{Fehler}(x)}{x} = \frac{1 \text{ km}}{100 \text{ km}} = 0{,}01 \mathrel{\widehat{=}} 1\,\%.$$

(b) Betrachten wir die Distanz x zwischen zwei Dörfern, die exakt $x = 2$ km von einander entfernt sind. Die Näherung dieser Entfernung (z.B. durch eine Messung) sei $x_N = 1$ km. Dann gilt

$$\text{Fehler}(x_N) = x - x_N = 2 \text{ km} - 1 \text{ km} = 1 \text{ km},$$
$$\text{Rel}(x_N) = \frac{\text{Fehler}(x)}{x} = \frac{1 \text{ km}}{2 \text{ km}} = 0{,}5 \mathrel{\widehat{=}} 50\,\%.$$

Der absolute Fehler ist in beiden Beispielen jeweils 1 km. Trotzdem ist es auch intuitiv klar, dass der angenäherte Wert für die Entfernung der beiden Städte „viel genauer“ ist als der angenäherte Wert für die Entfernung der beiden Dörfer. Dieses wird durch den relativen Fehler gemessen. Dieser beträgt im Beispiel (a) nur 1 % der Entfernung, aber im Beispiel (b) dagegen 50 % der Entfernung. ♠

Ein weiterer wichtiger Begriff bei der Diskussion des Fehlers einer Näherung sind die signifikanten Ziffern.

Definition 2.6. (signifikante Ziffern)
*Eine Näherung $x_N \in \mathbb{R}$ einer quantitativen reellen Größe $x \in \mathbb{R}$ hat **mindestens** m **signifikante Ziffern**, wenn $|x - x_N|$ kleiner oder gleich 5 Einheiten in der $(m+1)$-ten Ziffer von x ist.*

Beispiel 2.7. (signifikante Ziffern)

(a) Sei $x_N = 0{,}222$ eine Näherung für $x = \frac{2}{9} = 0{,}\overline{2}$. Dann hat die Näherung $x_N = 0{,}222$ drei signifikante Ziffern, denn

$$|x - x_N| = 0{,}000\overline{2},$$

und $0{,}00005 < 0{,}000\overline{2} \leq 0{,}0005 = 5 \cdot 10^{-4}$ und die Ziffer 5 in $0{,}0005 = 5 \cdot 10^{-4}$ steht an der Position der Stelle der $(3+1)$-sten Ziffer von $x = 0{,}\overline{2}$.

(b) Die Zahl $x_N = 31{,}578$ hat als Näherung von $x = 31{,}575$ vier signifikante Ziffern, denn es gilt

$$|x - x_N| = 0{,}003,$$

und $0{,}0005 < 0{,}003 \leq 0{,}005 = 5 \cdot 10^{-3}$ und die Ziffer 5 in $0{,}005 = 5 \cdot 10^{-3}$ steht an der Position der Stelle der $(4+1)$-sten Ziffer von $x = 31{,}575$.

(c) Die Zahl $x_N = 0{,}02138$ hat als Näherung von $x = 0{,}02144$ zwei signifikante Ziffern, denn es gilt

$$|x - x_N| = 0{,}00006,$$

und $0{,}00005 < 0{,}00006 \leq 0{,}0005 = 5 \cdot 10^{-4}$ und die Ziffer 5 in $0{,}0005 = 5 \cdot 10^{-4}$ steht an der Position der Stelle der $(2+1)$-sten Ziffer von $x = 0{,}02144$.

Wir betrachten auf den Übungszetteln weitere Beispiele. ♠

Nur die signifikanten Ziffern einer Näherung sind brauchbar! (Nicht-signifikante Ziffern haben keine Aussagekraft, denn sie können völlig falsch sein.)

Es ist wichtig, sich der vielen möglichen Arten von Fehlern bewusst zu sein.

Bemerkung 2.8. (Arten von Fehlern)
Wir unterscheiden die folgenden Arten von Fehlern:

(1) **Modellfehler:** Wenn man eine physikalische Größe durch ein Modell mit mathematischen Formeln beschreibt, so treten in der Regel Modellfehler auf, weil die Formeln die physikalische Größe nicht exakt sondern nur angenähert beschreiben.

Beispiel: Das einfachste Modell für ein Bevölkerungswachstum ist $N(t) = N_0\, e^{kt}$, wobei $N(t)$ die Bevölkerung zur Zeit t beschreibt, N_0 die Bevölkerung zum Zeitpunkt $t = 0$ ist und k die positive Wachstumskonstante ist. Dieses Modell überschätzt häufig die Bevölkerungszahl für großes t.

(2) **Arithmetische Formelfehler und Programmierfehler**

(3) **Physikalische Messfehler:** Quantitative Daten (z.B. Temperaturen, Längenmaße, Geschwindigkeiten, ...) können nur mit der durch das Messgerät gegebenen Genauigkeit gemessen werden. Weitere Berechnungen mit solchen messfehlerbehafteten quantitativen Daten beinhalten dann ebenfalls den Effekt der Messfehler.

Beispiel: Mit einem Maßband werden die Maße einer rechteckigen Box gemessen. Dieses ergibt: Breite: 21 cm, Länge: 13,7 cm und Höhe: 11,5 cm. Da das Maßband nur eine Genauigkeit von 0,5 mm Messgenauigkeit hat, wissen wir, dass die wahren Maße der rechteckigen Box

$$\text{Breite: } (210 + \epsilon_1) \text{ mm}, \quad \text{Länge: } (137 + \epsilon_2) \text{ mm}, \quad \text{Höhe: } (115 + \epsilon_3) \text{ mm}$$

mit unbekannten Werten $\epsilon_1, \epsilon_2, \epsilon_3 \in \mathbb{R}$ mit $|\epsilon_1|, |\epsilon_2|, |\epsilon_3| \leq 0{,}5$ mm sind.

(4) **Fehler durch die Darstellung und die arithmetischen Operationen im Computer:** Diese Fehler entstehen durch die Gleitkommadarstellung mit endlicher Mantissenlänge und durch Rundungsfehler oder Abschneidefehler bei Berechnungen mit dem Computer.

(5) **Mathematische Approximationsfehler:** Diese entstehen dadurch, dass eine quantitative reelle Größe mit einem numerischen Verfahren (meist) nur angenähert berechnet werden kann.

Beispiel: Das Integral $A = \int_0^1 e^{-x^2}\, \mathrm{d}x$ kann nicht elementar berechnet werden, weil wir keine Stammfunktion von e^{-x^2} angeben können. Um das Integral angenähert zu berechnen, ersetzt man e^{-x^2} durch die mit Hilfe

des Taylor-Polynoms der Exponentialfunktion gewonnene Näherung

$$e^{-x^2} \approx 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \ldots + (-1)^m \frac{x^{2m}}{m!}$$

und erhält so die folgende Näherung

$$A \approx A_m = \int_0^1 \left(1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \ldots + (-1)^m \frac{x^{2m}}{m!}\right) \mathrm{d}x,$$

für das Integral A. Die Näherung A_m lässt sich nun bequem berechnen, aber sie beinhaltet einen mathematischen Approximationsfehler.

Den Fachbegriff „**Approximation**“ kann man mit „Annäherung“ übersetzen. Das zugehörige Verb „**approximieren**“ bedeutet „annähern“.

Ein großes Problem bei Berechnungen mit einem Computer (oder einem Taschenrechner) ist der **mögliche Verlust signifikanter Stellen** durch **Auslöschung**. Um dieses zu verstehen, betrachten wir ein Beispiel.

Beispiel 2.9. (Verlust signifikanter Stellen durch Auslöschung)
Die Funktion

$$f : [0; \infty[\to \mathbb{R}, \quad f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right),$$

soll für $x = 10^k$ mit $k = 0, 1, 2, 3, 4, 5$ mit einem Taschenrechner (oder einem Computer) berechnet werden, der mit einer 6-stelligen Gleitkommadarstellung arbeitet. Dass mit 6-stelliger Gleitkommadarstellung gerechnet wird, bedeutet, dass wir in jedem Zwischenschritt auf sechs Gleitkommastellen runden. Der für $f(x)$ so erhaltene Näherungswert sei mit $f_N(x)$ bezeichnet.

$$f(1) = 1 \cdot \left(\sqrt{2} - \sqrt{1}\right) \doteq 1 \cdot \big(\underbrace{1{,}41421 - 1}_{=0{,}414210}\big) = 0{,}41421 = f_N(1),$$

$$f(10) = 10 \cdot \left(\sqrt{11} - \sqrt{10}\right) \doteq 10 \cdot \big(\underbrace{3{,}31662 - 3{,}16228}_{=0{,}154340}\big) = 1{,}5434 = f_N(10),$$

$$f(10^2) = 10^2 \cdot \left(\sqrt{101} - \sqrt{100}\right) \doteq 100 \cdot \big(\underbrace{10{,}0499 - 10}_{=0{,}0499000}\big) = 4{,}99 = f_N(10^2),$$

$$f(10^3) = 10^3 \cdot \left(\sqrt{1001} - \sqrt{10^3}\right) \doteq 10^3 \cdot \big(\underbrace{31{,}6386 - 31{,}6228}_{=0{,}0158000}\big) = 15{,}8 = f_N(10^3),$$

$$f(10^4) = 10^4 \cdot \left(\sqrt{10001} - \sqrt{10^4}\right) \doteq 10^4 \cdot \big(\underbrace{100{,}005 - 100}_{=0{,}00500000}\big) = 50 = f_N(10^4),$$

$$f(10^5) = 10^5 \cdot \left(\sqrt{100001} - \sqrt{10^5}\right) \doteq 10^5 \cdot \big(\underbrace{316{,}229 - 316{,}228}_{=0{,}0010000}\big) = 100 = f_N(10^5).$$

Damit erhalten wir also die Ergebnisse in der weiter unten stehenden Tabelle. Die „exakten“ Werte für $f(x)$ wurden jeweils auf 6-stellige Gleitkommadarstellung gerundet.

x	berechnetes $f_N(x)$	exakter Wert für $f(x)$	absoluter Fehler	relativer Fehler
$1 = 10^0$	0,414210	0,414214	0,000004	$9{,}65684 \cdot 10^{-6}$
$10 = 10^1$	1,54340	1,54347	0,00007	$4{,}53524 \cdot 10^{-5}$
$100 = 10^2$	4,99000	4,98756	−0,00244	$-4{,}89217 \cdot 10^{-4}$
$1000 = 10^3$	15,8000	15,8074	0,0074	$4{,}68135 \cdot 10^{-4}$
$10.000 = 10^4$	50,0000	49,9988	−0,0012	$-2{,}40006 \cdot 10^{-5}$
$100.000 = 10^5$	100,000	158,113	58,113	$3{,}67541 \cdot 10^{-1}$

Für $x = 10^5$ ist der absolute Fehler

$$\text{Fehler}\big(f_N(10^5)\big) = f(10^5) - f_N(10^5) = 158{,}113 - 100 = 58{,}113,$$

und der relative Fehler ist

$$\text{Rel}\big(f_N(10^5)\big) = \frac{\text{Fehler}\big(f_N(10^5)\big)}{158{,}113} = \frac{58{,}113}{158{,}113} \doteq 0{,}367541 \mathrel{\widehat{=}} 36{,}8\,\%.$$

Wie kommt dieses eklatant schlechte Ergebnis zustande? – Für großes x sind $\sqrt{x+1}$ und $\sqrt{x}$ fast gleich. Daher führt das Bilden der Differenz $\sqrt{x+1} - \sqrt{x}$ nach der jeweiligen Rundung von $\sqrt{x+1}$ und $\sqrt{x}$ zur Auslöschung vieler signifikanter Ziffern. Dieses kann man gut sehen, wenn man die konkrete Berechnung von $f_N(10^4)$ oder $f_N(10^5)$ mit 6-stelliger Gleitkommadarstellung oberhalb der Tabelle genauer studiert.

Man kann das Problem der Auslöschung signifikanter Ziffern in diesem Fall leicht umgehen, indem man $f(x)$ vorher geeignet umformt und dann mit der neuen Darstellung von $f(x)$ rechnet: Durch Erweitern mit $\sqrt{x+1} + \sqrt{x}$ erhalten wir mit der dritten binomischen Formel:

$$f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right) = x \cdot \frac{\left(\sqrt{x+1} - \sqrt{x}\right)\left(\sqrt{x+1} + \sqrt{x}\right)}{\sqrt{x+1} + \sqrt{x}}$$

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

Berechnet man $f(10^5)$ mit dieser neuen Formel mit 6-stelliger Gleitkommadarstellung, so erhält man

$$f(10^5) = \frac{100000}{\sqrt{100001}+\sqrt{100000}} \doteq \frac{100000}{316{,}229+316{,}228} \doteq \frac{100000}{632{,}457} \doteq 158{,}114,$$

und wir haben nur in der letzten Ziffer eine Abweichung von $f(10^5) = 158{,}113$. Anders ausgedrückt: $f_N(10^5) = 158{,}114$ hat 5 signifikante Ziffern. ♠

KAPITEL 3

Nullstellenberechnung

In diesem Kapitel lernen wir verschiedene numerische Verfahren kennen, um **Nullstellen** einer Funktion f, also Lösungen x der Gleichung $f(x) = 0$, zu berechnen. Das Problem der Nullstellenberechnung tritt bei Anwendungsproblemen häufig als ein Teilproblem auf.

Als Vorbereitung benötigen wir den Zwischenwertsatz für stetige Funktionen.

Satz 3.1. (Zwischenwertsatz)
*Sei $f : [c; d] \to \mathbb{R}$ eine **stetige** Funktion, und seien a und b zwei beliebige Punkte in $[c; d]$ mit der Eigenschaft $c \leq a < b \leq d$. Dann gibt es zu jedem Wert y zwischen $f(a)$ und $f(b)$ einen Punkt $z \in [a; b]$ mit $f(z) = y$.*

Die Formulierung „Wert y zwischen $f(a)$ und $f(b)$“ schließt die Werte $f(a)$ und $f(b)$ mit ein. Gemeint ist also, dass y die Bedingung $f(a) \leq y \leq f(b)$ falls $f(a) \leq f(b)$ bzw. $f(b) \leq y \leq f(a)$ falls $f(b) \leq f(a)$ erfüllt.

Der Zwischenwertsatz ist in Abbildung 3.1 veranschaulicht. Geometrisch bedeutet die Stetigkeit von f auf dem Intervall $[c; d]$, dass man den Graphen von f ohne Absetzen durchzeichnen kann. Da also der Graph von f die Punkte $\big(a; f(a)\big)$ und $\big(b; f(b)\big)$ mit einer durchgehenden Kurve verbindet, muss diese insbesondere auch für jeden Wert y zwischen $f(a)$ und $f(b)$ mindestens einen x-Wert z haben, für den $\big(z; f(z)\big) = (z; y)$, also $f(z) = y$, gilt.

In diesem Kapitel brauchen wir insbesondere den folgenden Sonderfall des Zwischenwertsatzes, bei dem der Funktionswert 0 als Zwischenwert auftritt:

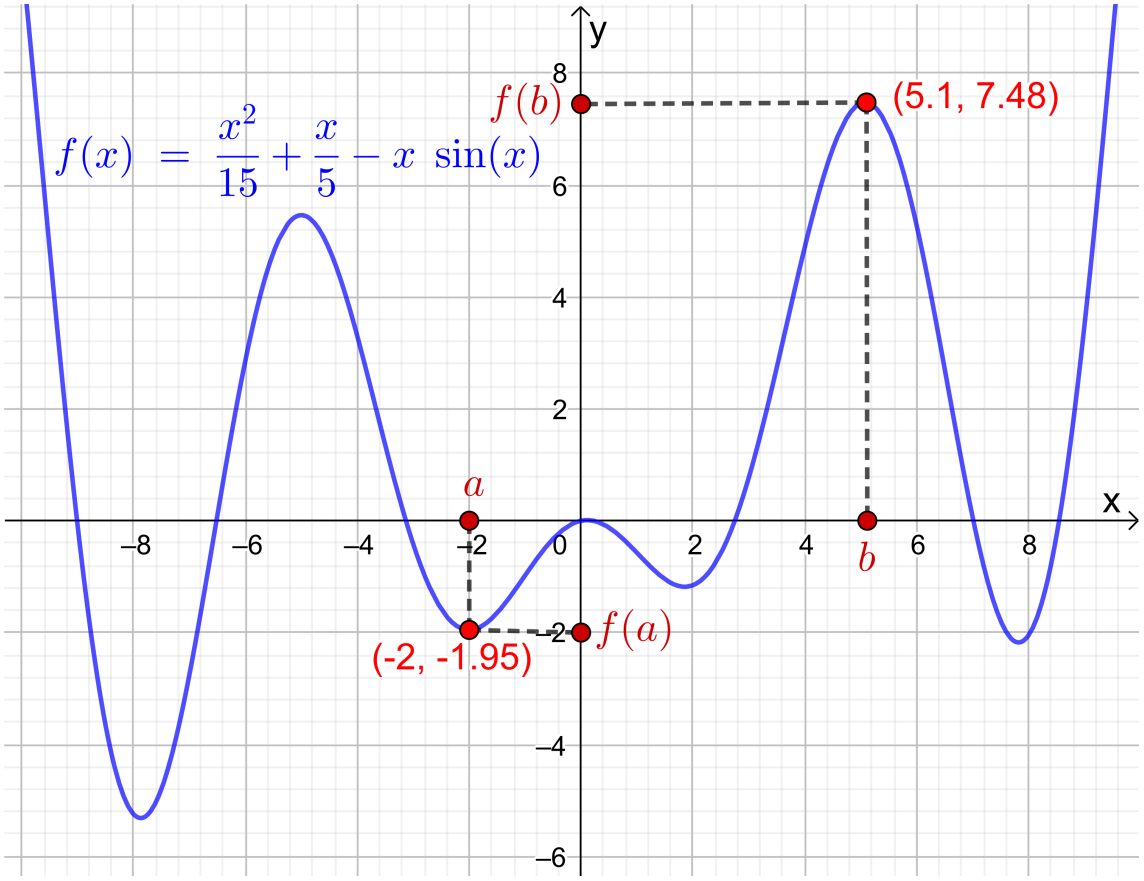


Abb. 3.1: Veranschaulichung des Zwischenwertsatzes: Da f stetig ist, werden im Intervall $[a; b]$ alle Werte zwischen $f(a)$ und $f(b)$ als Funktionswerte angenommen.

Folgerung 3.2. (Existenz einer Nullstelle)
*Sei $f : [c; d] \to \mathbb{R}$ eine **stetige** Funktion, und seien a und b zwei beliebige Punkte in den Intervall $[c; d]$ mit den Eigenschaften $c \leq a < b \leq d$ und*

$$f(a) \leq 0 \leq f(b) \qquad \text{oder} \qquad f(a) \geq 0 \geq f(b).$$

*Dann hat f in $[a; b]$ **mindestens eine Nullstelle**.*

3.1 Bisektionsverfahren

Wir betrachten eine Funktion f, die auf dem Intervall $[a; b]$ definiert und stetig ist. Es gelte weiter

$$f(a) \cdot f(b) < 0,$$

d.h. $f(a)$ und $f(b)$ sind beide ungleich null und haben unterschiedliche Vorzeichen. Dann nimmt f nach dem Zwischenwertsatz in dem Intervall $[a;b]$ mindestens einmal den Wert 0 an. Es gibt also ein $z \in]a;b[$ mit $f(z) = 0$. (Anders ausgedrückt, f wechselt in dem Intervall $]a;b[$ mindestens einmal sein Vorzeichen.)

Frage: Wie kann man eine gute Näherung für eine solche Nullstelle z berechnen?

Die Idee des **Bisektionsverfahrens** zur Nullstellenfindung ist wie folgt: Wir teilen das Intervall in zwei gleich große Teilintervalle auf und behalten das Teilintervall, in dem f garantiert eine Nullstelle hat. Dieses prüfen wir, indem wir die Funktionswerte an den beiden (neuen) Teilintervallenden multiplizieren und ein Teilintervall nehmen, bei dem wir dabei einen nicht-positiven Wert für das Produkt der Funktionswerte erhalten.

Genauer sieht die **Vorgehensweise des Bisektionsverfahrens** wie folgt aus:

Schritt 1: Wir setzen $a_1 = a$ und $b_1 = b$.
(Nach Voraussetzung gilt $f(a_1) \cdot f(b_1) < 0$.)

Wir definieren $c_1 = \dfrac{a_1 + b_1}{2}$. (Dann erhalten wir mit $[a_1; c_1]$ und $[c_1; b_1]$ zwei gleich große Teilintervalle der Länge $(b-a)/2$.)

Ist $f(c_1) \cdot f(b_1) \leq 0$, so setzen wir $a_2 = c_1$ und $b_2 = b_1$.

Andernfalls (also wenn $f(c_1) \cdot f(b_1) > 0$ und damit $f(a_1) \cdot f(c_1) \leq 0$ ist) setzen wir $a_2 = a_1$ und $b_2 = c_1$.

(Wir haben also jetzt ein Teilintervall $[a_2; b_2]$ der Länge $(b-a)/2$ gefunden, in dem garantiert eine Nullstelle von f liegt.)

Schritt 2: (Nach Voraussetzung gilt $f(a_2) \cdot f(b_2) \leq 0$.)

Wir definieren $c_2 = \dfrac{a_2 + b_2}{2}$. (Dann erhalten wir mit $[a_2; c_2]$ und $[c_2; b_2]$ zwei gleich große Teilintervalle der Länge $(b-a)/4$.)

Ist $f(c_2) \cdot f(b_2) \leq 0$, so setzen wir $a_3 = c_2$ und $b_3 = b_2$.

Andernfalls (also wenn $f(c_2) \cdot f(b_2) > 0$ und damit $f(a_2) \cdot f(c_2) \leq 0$ ist) setzen wir $a_3 = a_2$ und $b_3 = c_2$.

(Wir haben also jetzt ein Teilintervall $[a_3; b_3]$ der Länge $(b-a)/4$ gefunden, in dem garantiert eine Nullstelle von f liegt.)

$\vdots$

Schritt n: (Nach Voraussetzung gilt $f(a_n) \cdot f(b_n) \leq 0$.)

Wir definieren $c_n = \dfrac{a_n + b_n}{2}$. (Dann erhalten wir mit $[a_n; c_n]$ und $[c_n; b_n]$ zwei gleich große Teilintervalle der Länge $(b-a)/2^n$.)

Ist $f(c_n) \cdot f(b_n) \leq 0$, so setzen wir $a_{n+1} = c_n$ und $b_{n+1} = b_n$.

Andernfalls (also wenn $f(c_n) \cdot f(b_n) > 0$ und damit $f(a_n) \cdot f(c_n) \leq 0$ ist) setzen wir $a_{n+1} = a_n$ und $b_{n+1} = c_n$.

(Wir haben also jetzt ein Teilintervall $[a_{n+1}; b_{n+1}]$ der Länge $(b-a)/2^n$ gefunden, in dem garantiert eine Nullstelle von f liegt.)

$\vdots$

Abbruchkriterium/Stoppkriterium: Wir wissen nach n Schritten des Bisektionsverfahrens, dass eine Nullstelle von f im Intervall $[a_{n+1}; b_{n+1}]$ der Länge $(b-a)/2^n$ liegt. Wir nehmen $\widetilde{z} = c_n$ als Näherung dieser Nullstelle. Dann gilt $|z - \widetilde{z}| \leq (b-a)/2^n$, weil $z, \widetilde{z} \in [a_{n+1}, b_{n+1}]$, d.h. wir haben eine Näherung $\widetilde{z}$ mit einer maximalen betraglichen Abweichung von $(b-a)/2^n$ vom exakten Wert z bestimmt.

Soll also eine Näherung $\widetilde{z}$ einer Nullstelle z von f mit einer maximalen betraglichen Abweichung $\varepsilon > 0$ vom exakten Wert z gefunden werden (d.h. dass für die Näherung $\widetilde{z}$ der Nullstelle z höchstens $|z - \widetilde{z}| \leq \varepsilon$ gelten soll), so stoppen wir das Verfahren sobald $(b-a)/2^n \leq \varepsilon$ ist, denn dann gilt nach Konstruktion für die Näherung $\widetilde{z} = c_n$

$$|z - \widetilde{z}| = |z - c_n| \leq \frac{b-a}{2^n} \leq \varepsilon,$$

weil z und c_n beide im Intervall $[a_{n+1}; b_{n+1}]$ der Länge $(b-a)/2^n$ liegen.

Wir halten das Bisektionsverfahren nun als Algorithmus fest. Dabei nutzen wir die **Signum-Funktion** (oder **Vorzeichenfunktion**)

$$\operatorname{sgn} : \mathbb{R} \to \mathbb{R}, \qquad \operatorname{sgn}(x) = \begin{cases} -1 & \text{für } x < 0, \\ 0 & \text{für } x = 0, \\ 1 & \text{für } x > 0, \end{cases}$$

um zu vermeiden, dass sehr kleine Werte $f(c_n) \cdot f(b_n)$ vom Computer als 0 interpretiert werden.

Verfahren 3.3. (Bisektionsverfahren)
*Voraussetzungen: Die Funktion f sei **stetig** auf dem Intervall $[a; b]$, und es gelte $f(a) \cdot f(b) < 0$. (Dann hat f mindestens eine Nullstelle z in $]a; b[$.)*

Sei $\varepsilon > 0$ die absolute Fehlerschranke für die Näherung der Nullstelle.

Initialisierung: *Seien $a_1 = a$ und $b_1 = b$.*

Algorithmus: *Für $n = 1, 2, 3, \ldots$ führe die folgenden Schritte aus*

(1) Definiere $c_n = \dfrac{a_n + b_n}{2}$.

(2) Falls $\operatorname{sgn}\big(f(c_n)\big) \cdot \operatorname{sgn}\big(f(b_n)\big) \leq 0$ *ist, setze* $a_{n+1} = c_n$ *und* $b_{n+1} = b_n$. *Andernfalls setze* $a_{n+1} = a_n$ *und* $b_{n+1} = c_n$.

bis $b_{n+1} - a_{n+1} = b_n - c_n \leq \varepsilon$ *gilt.*

(Wegen $b_{n+1} - a_{n+1} = b_n - c_n = \dfrac{b-a}{2^n}$ *bricht der Algorithmus nach endlich vielen Schritten ab.)*

Dann ist $\widetilde{z} = c_n$ *aus dem letzten Schritt des Algorithmus eine Näherung der Nullstelle* z *mit dem betraglichen absoluten Fehler* $|z - \widetilde{z}| \leq \varepsilon$.

Jede Durchführung der Anweisungen (1) und (2) in Verfahren 3.3 nennt man einen **Iterationsschritt**. Bricht der Algorithmus nach n Schritten ab, so sagt man, dass der Algorithmus nach n Iterationsschritten die gewünschte absolute Fehlerschranke erreicht hat.

Betrachten wir ein Beispiel

Beispiel 3.4. (Bisektionsverfahren)

Gesucht ist eine Näherung der größten Nullstelle der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1 = x\,(x^5 - 1) - 1,$$

mit der absoluten Fehlerschranke $\varepsilon = 0{,}001 = 10^{-3}$.

Der Graph der Funktion f ist in Abbildung 3.2 gezeichnet. Als Polynom ist f beliebig oft differenzierbar und insbesondere stetig. Es gilt

$$f(2) = 2\,(2^5 - 1) - 1 = 61 > 0,$$
$$f(1) = 1\,(1^5 - 1) - 1 = -1 < 0.$$

Wegen $f(1) = -1 < 0 < f(2) = 61$ liegt nach dem Zwischenwertsatz eine Nullstelle von f in dem Intervall $[a; b] = [1; 2]$.

Warum liegt im Intervall $[1; 2]$ nur die größte Nullstelle von f und keine weitere Nullstelle von f? Um dieses nachzuweisen, berechnen wir die Ableitung von f:

$$f'(x) = 6\,x^5 - 1$$

n	a_n	b_n	c_n	$b_n - c_n$	$f(c_n)$	$f(b_n)$	$f(b_n) \cdot f(c_n)$
1	1,00000	2,00000	1,50000	0,50000	$8{,}89 \cdot 10^{0}$	$6{,}10 \cdot 10^{1}$	$5{,}42 \cdot 10^{2}$
2	1,00000	1,50000	1,25000	0,25000	$1{,}56 \cdot 10^{0}$	$8{,}89 \cdot 10^{0}$	$1{,}39 \cdot 10^{1}$
3	1,00000	1,25000	1,12500	0,12500	$-9{,}77 \cdot 10^{-2}$	$1{,}56 \cdot 10^{0}$	$-1{,}53 \cdot 10^{-1}$
4	1,12500	1,25000	1,18750	0,06250	$6{,}17 \cdot 10^{-1}$	$1{,}56 \cdot 10^{0}$	$9{,}65 \cdot 10^{-1}$
5	1,12500	1,18750	1,15625	0,03125	$2{,}33 \cdot 10^{-1}$	$6{,}17 \cdot 10^{-1}$	$1{,}44 \cdot 10^{-1}$
6	1,12500	1,15625	1,14063	0,01563	$6{,}16 \cdot 10^{-2}$	$2{,}33 \cdot 10^{-1}$	$1{,}44 \cdot 10^{-2}$
7	1,12500	1,14063	1,13281	0,00781	$-1{,}96 \cdot 10^{-2}$	$6{,}16 \cdot 10^{-2}$	$-1{,}21 \cdot 10^{-3}$
8	1,13281	1,14063	1,13672	0,00391	$2{,}06 \cdot 10^{-2}$	$6{,}16 \cdot 10^{-2}$	$1{,}27 \cdot 10^{-3}$
9	1,13281	1,13672	1,13477	0,00195	$4{,}27 \cdot 10^{-4}$	$2{,}06 \cdot 10^{-2}$	$8{,}80 \cdot 10^{-6}$
10	1,13281	1,13477	1,13379	0,00098	$-9{,}60 \cdot 10^{-3}$	$4{,}27 \cdot 10^{-4}$	$-4{,}10 \cdot 10^{-6}$

Tabelle 3.1: Bisektionsverfahren zur Berechnung der größten Nullstelle der Funktion $f(x) = x^6 - x - 1$ mit den Startwerten $a_1 = 1$ und $b_1 = 2$.

Für die Ableitung gilt

$$f'(x) = 6\,x^5 - 1 \geq 6 \cdot 1^5 - 1 = 5 > 0 \qquad \text{für alle } x \geq 1.$$

Also ist f auf den Intervall $[1; \infty[$ streng monoton wachsend. Daraus folgt, dass f in $[1; \infty[$ höchstens einmal die x-Achse schneidet und damit höchstens eine Nullstelle in $[1; \infty[$ hat. Daher enthält das Intervall $[1; 2]$ nur genau eine Nullstelle von f, und diese ist die einzige Nullstelle von f in $[1; \infty[$ und damit die größte Nullstelle von f.

Wir führen daher das Bisektionsverfahren mit den Startwerten $a_1 = a = 1$ und $b_1 = b = 2$ durch. Die Werte für a_n, b_n, c_n und $b_n - c_n$, sowie $f(c_n)$, $f(b_n)$ und $f(b_n) \cdot f(c_n)$ für $n = 1, 2, \ldots, 10$, sind in Tabelle 3.1 angegeben. Dabei wurden a_n, b_n, c_n auf eine 6-stellige Gleitkommadarstellung gerundet, und $b_n - c_n$ wurde auf 5 Nachkommastellen gerundet angegeben. Die Werte von $f(c_n)$ und $f(b_n)$ bzw. $f(b_n) \cdot f(c_n)$ wurden dabei nur auf eine 3-stellige Gleitkommadarstellung gerundet angegeben, da hier nur die Größenordnung bzw. das Vorzeichen interessant sind.

Wir sehen, dass im 10-ten Schritt gilt

$$|z - c_{10}| \leq b_{10} - c_{10} = 0{,}00098 \leq 0{,}001.$$

Also wird die absolute Fehlerschranke $\varepsilon = 0{,}001 = 10^{-3}$ nach 10 Iterationsschritten erreicht. Wir erhalten als Näherung für die (größte) Nullstelle von f nach 10 Iterationsschritten $c_{10} = 1{,}13379$. (Zum Vergleich: Der auf eine 10-stellige Gleitkommadarstellung gerundete Wert der Nullstelle ist $z \doteq 1{,}134724138$.) ♠

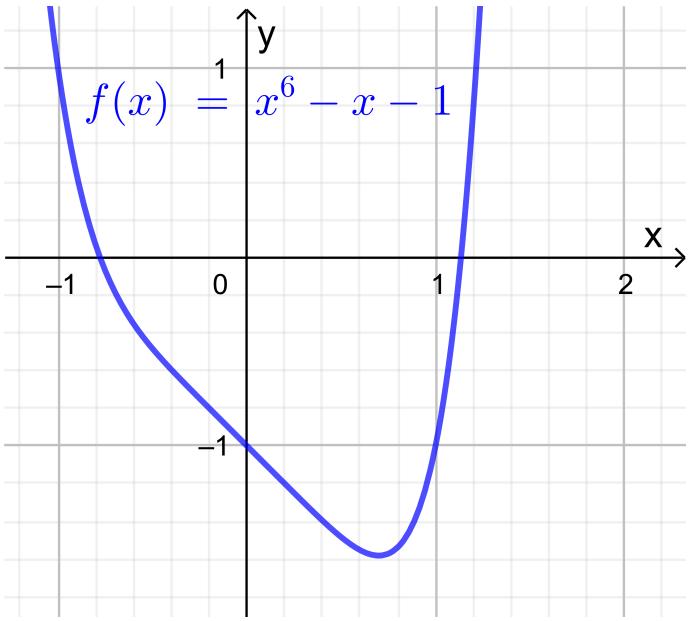


Abb. 3.2: Graph der Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^6 - x - 1$.

Frage: Was kann man allgemein über den **absoluten Fehler** $z - \widetilde{z}$ **der Näherung** $\widetilde{z} = c_n$ **nach** n **Iterationsschritten** aussagen?

Wir starten bei dem Bisektionsverfahren mit dem Intervall $[a; b] = [a_1; b_1]$ der Länge $b - a$. In jedem Iterationsschritt wird die Intervalllänge halbiert:

Nach einem Iterationsschritt erhalten wir also ein Intervall $[a_2; b_2]$ der Länge

$$b_2 - a_2 = \frac{b-a}{2}\,.$$

Nach zwei Iterationsschritten erhalten wir also ein Intervall $[a_3; b_3]$ der Länge

$$b_3 - a_3 = \frac{(b-a)/2}{2} = \frac{b-a}{2^2}\,.$$

$$\vdots$$

Nach n Iterationsschritten erhalten wir also ein Intervall $[a_{n+1}; b_{n+1}]$ der Länge

$$b_{n+1} - a_{n+1} = \frac{(b-a)/2^{n-1}}{2} = \frac{b-a}{2^n}\,,$$

und wir wissen, dass die unbekannte Nullstelle z in diesem Intervall liegt, also $z \in [a_{n+1}; b_{n+1}]$. Daher gilt für die Näherung c_n (welche entweder gleich a_{n+1} oder gleich b_{n+1} ist und damit insbesondere in $[a_{n+1}; b_{n+1}]$ liegt)

$$|z - c_n| \leq b_{n+1} - a_{n+1} = \frac{b-a}{2^n}\,.$$

Wenn die absolute Fehlerschranke ε für die Näherung der Nullstelle z vorgegeben ist, so können wir aus

$$|z - c_n| \leq \frac{b-a}{2^n} \leq \varepsilon$$

berechnen, nach wie vielen Iterationsschritten die absolute Fehlerschranke ε garantiert erreicht wird:

$$\frac{b-a}{2^n} \leq \varepsilon \quad \Big| \cdot 2^n \quad \Longleftrightarrow \quad b - a \leq \varepsilon \cdot 2^n \quad \Big| : \varepsilon \quad \Longleftrightarrow \quad \frac{b-a}{\varepsilon} \leq 2^n = e^{\ln(2) \cdot n}$$

$$\overset{\ln'(x)=\frac{1}{x}>0}{\Longleftrightarrow} \quad \ln\left(\frac{b-a}{\varepsilon}\right) \leq \ln(2) \cdot n \quad \Big| : \ln(2) \quad \overset{\ln(2)>0}{\Longleftrightarrow} \quad \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln(2)} \leq n,$$

wobei wir bei der Umformung von der ersten in die zweite Zeile genutzt haben, dass der natürliche Logarithmus ln streng monoton wachsend ist und das Anwenden von ln daher eine Äquivalenzumformung ist und die Richtung der Ungleichung (wegen des streng monotonen Wachstums von ln) erhalten bleibt.

$$\boxed{\text{Für jedes} \quad n \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln(2)} \quad \text{gilt} \quad |z - c_n| \leq \varepsilon.} \tag{3.1}$$

Beispiel 3.5. (Iterationszahl im Bisektionsverfahren)
Wir betrachten die Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1 = x\,(x^5 - 1) - 1,$$

deren größte Nullstelle in Beispiel 3.4 mit dem Bisektionsverfahren mit den Startwerten $a_1 = a = 1$ und $b_1 = b = 2$ angenähert berechnet wurde. Die absolute Fehlerschranke $\varepsilon = 0{,}001 = 10^{-3}$ wurde nach 10 Iterationsschritten erreicht.

Welche Information liefert uns (3.1) über die Anzahl der Iterationsschritte, nach denen die absolute Fehlerschranke $\varepsilon = 10^{-3}$ spätestens garantiert ist? Nach

$$n \geq \frac{\ln\left(\frac{2-1}{0{,}001}\right)}{\ln(2)} = \frac{\ln(10^3)}{\ln(2)} \doteq 9{,}97,$$

also nach (spätestens) 10 Iterationsschritten, ist die absolute Fehlerschranke $\varepsilon = 0{,}001 = 10^{-3}$ garantiert erreicht. Die theoretischen Überlegungen bestätigen also, was wir bereits in Beispiel 3.4 beobachtet hatten. ♠

3.2 Newton-Verfahren

Das Newton-Verfahren berechnet eine Folge von Näherungen einer Nullstelle einer stetig differenzierbaren Funktion mit Hilfe der Tangenten an den Graphen der Funktion. Das **Newton-Verfahren** ist in Abbildung 3.3 illustriert und funktioniert im Detail wie folgt:

Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion, die in einen Punkt $z \in]a; b[$ eine Nullstelle hat, also $f(z) = 0$. Als **Startwert** für das Newton-Verfahren brauchen wir eine **hinreichend gute Näherung** x_0 für die Nullstelle z. Diese Näherung x_0 kann zum Beispiel mit Hilfe eines Plots des Graphen von f bestimmt werden.

Wir betrachten nun die **Tangente in $\big(x_0; f(x_0)\big)$ an den Graphen und bestimmen deren Schnittpunkt x_1 mit der x-Achse**. Da die Tangente in $\big(x_0; f(x_0)\big)$ an den Graphen dicht bei x_0 eine gute Näherung der Funktion f ist, erwarten wir, dass der Schnittpunkt x_1 dieser Tangente mit der x-Achse eine verbesserte Näherung der Nullstelle z ist.

Die **Tangente in $\big(x_0; f(x_0)\big)$ an den Graphen von** f ist durch das lineare Taylor-Polynom mit dem Entwicklungspunkt x_0 (vgl. Satz 1.20) gegeben:

$$p_1(x) = f(x_0) + f'(x_0)\,(x - x_0)$$

Für die neue (verbesserte) Näherung x_1 der Nullstelle z gilt $p_1(x_1) = 0$, da x_1 als Schnittpunkt des linearen Taylor-Polynoms p_1 mit dem Entwicklungspunkt x_0 mit der x-Achse gegeben ist. Wir suchen also x_1 mit

$$p_1(x_1) = f(x_0) + f'(x_0)\,(x_1 - x_0) = 0, \tag{3.2}$$

wobei wir benötigen, dass die Ableitung $f'(x_0) \neq 0$ ist. (Falls $f'(x_0) = 0$ gilt, ist die Tangente parallel zur x-Achse und schneidet diese nicht, oder die Tangente ist identisch mit der x-Achse, falls auch $f(x_0) = 0$ ist. Im letzteren Fall ist der Startwert x_0 aber bereits eine Nullstelle von f.) Auflösen von (3.2) nach x_1 liefert:

$$\begin{aligned} f(x_0) + f'(x_0)\,(x_1 - x_0) = 0 \quad &\iff \quad f'(x_0)\,(x_1 - x_0) = -f(x_0) \\ \overset{f'(x_0)\neq 0}{\iff} \quad x_1 - x_0 = -\frac{f(x_0)}{f'(x_0)} \quad &\iff \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \end{aligned}$$

Die neue Näherung x_1 für die Nullstelle ist also gegeben durch

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}\,. \tag{3.3}$$

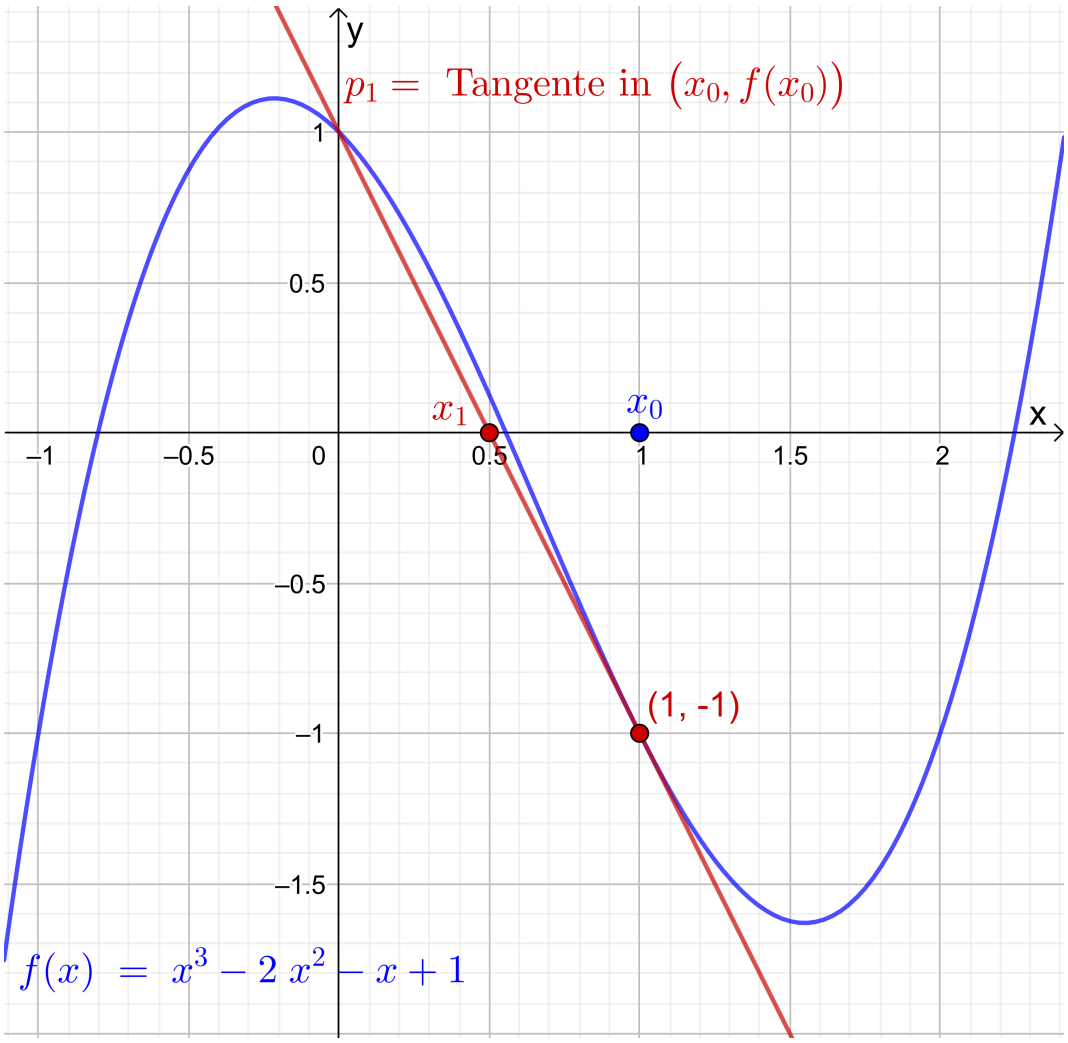


Abb. 3.3: Illustration des Newton-Verfahrens: Für den Startwert $x_0 = 1$ legen wir die Tangente p_1 an den Graphen im Punkt $\big(x_0; f(x_0)\big) = (1; -1)$. Die neue Näherung x_1 für die Nullstelle ist der Schnittpunkt der Tangente mit der x-Achse.

Wir können diese Vorgehensweise wiederholen, wobei wir nun eine verbesserte Näherung von x_1 für die Nullstelle z berechnen wollen. Dazu ersetzen wir x_1 durch die neue Näherung x_2 und den Startwert x_0 durch x_1 und erhalten somit aus (3.3) als Formel für die neue Näherung

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}\,.$$

Setzt man dieses Verfahren fort so erhält man nach $n+1$ Schritten:

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

Wir halten das Newton-Verfahren als Algorithmus fest.

Verfahren 3.6. (Newton-Verfahren)
*Sei $f :]a; b[\to \mathbb{R}$ eine **stetig differenzierbare** Funktion, und sei x_0 eine gute Näherung für eine Nullstelle z von f (d.h. es gilt $f(z) = 0$). Weiter gelte $f'(x) \neq 0$ für alle x dicht bei z. (Insbesondere gilt dann für einen guten Startwert x_0 auch $f'(x_0) \neq 0$.) Das **Newton-Verfahren** berechnet Näherungen für die Nullstelle z mit dem folgenden Iterationsverfahren:*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, 2, \ldots \tag{3.4}$$

Man bezeichnet die im Newton-Verfahren berechneten Näherungen x_n auch als die **Iterierten** (des Newton-Verfahrens). Jede Berechnung einer neuen Näherung x_n ist ein **Iterationsschritt**.

Betrachten wir zunächst ein Beispiel.

Beispiel 3.7. (Newton-Verfahren)
Gesucht ist eine Näherung der größten Nullstelle der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

die bereits in Beispiel 3.4 betrachtet wurde (siehe auch Abbildung 3.2). In Beispiel 3.4 hatten wir nachgewiesen, dass die größte reelle Nullstelle z im Intervall $[1; 2]$ liegt. Wir nehmen daher als Startwert für das Newton-Verfahren $x_0 = 1{,}5$.

Die Ableitung von f ist durch

$$f'(x) = 6\,x^5 - 1$$

gegeben, und somit lautet die Formel für das Newton-Verfahren:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^6 - x_n - 1}{6\,x_n^5 - 1}, \qquad n = 0, 1, 2, \ldots$$

In Tabelle 3.2 wurden die ersten 6 Iterationsschritte berechnet. Außer den Näherungen x_n der Nullstelle z wurden die Funktionswerte $f(x_n)$ angegeben. In der beiden letzten Spalten der Tabelle wurden in der Zeile für x_n der absolute Fehler $z - x_{n-1}$ der vorherigen Näherung x_{n-1} und zusätzlich $x_n - x_{n-1}$ angegeben. Die Iterierten x_n wurden auf eine 9-stellige Gleitkommadarstellung gerundet angegeben. Die Werte für $f(x_n)$, $x_n - x_{n-1}$ und $z - x_{n-1}$ wurden dagegen auf eine 3-stellige Gleitkommadarstellung gerundet, da hier nur die Größenordnung interessant ist.

n	x_n	$f(x_n)$	$x_n - x_{n-1}$	$z - x_{n-1}$
0	1,5	$8{,}89 \cdot 10^{1}$		
1	1,30049088	$2{,}54 \cdot 10^{1}$	$-2{,}00 \cdot 10^{-1}$	$-3{,}65 \cdot 10^{-1}$
2	1,18148042	$5{,}38 \cdot 10^{-1}$	$-1{,}19 \cdot 10^{-1}$	$-1{,}66 \cdot 10^{-1}$
3	1,13945559	$4{,}92 \cdot 10^{-2}$	$-4{,}20 \cdot 10^{-2}$	$-4{,}68 \cdot 10^{-2}$
4	1,13477763	$5{,}50 \cdot 10^{-4}$	$-4{,}68 \cdot 10^{-3}$	$-4{,}73 \cdot 10^{-3}$
5	1,13472415	$7{,}11 \cdot 10^{-8}$	$-5{,}35 \cdot 10^{-5}$	$-5{,}35 \cdot 10^{-5}$
6	1,13472414	$1{,}55 \cdot 10^{-15}$	$-6{,}91 \cdot 10^{-9}$	$-6{,}91 \cdot 10^{-9}$

Tabelle 3.2: Newton-Verfahren zur Berechnung der größten Nullstelle der Funktion $f(x) = x^6 - x - 1$ mit dem Startwert $x_0 = 1{,}5$.

Der auf eine 10-stellige Gleitkommadarstellung gerundete Wert der Nullstelle ist $z \doteq 1{,}134724138$, und wir sehen, dass die Näherung $x_5 \doteq 1{,}13472415$ bereits 8 signifikante Ziffern hat.

Wir werden noch sehen, dass $x_n - x_{n-1}$ eine gute Näherung für den absoluten Fehler $z - x_{n-1}$ liefert. Da bei einer unbekannten Nullstelle z der absolute Fehler $z - x_n$ nicht direkt zu berechnen ist, ist es wichtig einen guten Näherungswert für diesen zu haben, denn nur dann können wir eine Aussage darüber treffen, wann die Näherung x_n die gewünschte absolute Fehlerschranke vermutlich erreicht hat. – An der letzten Spalte lesen wir ab, dass sich der absolute Fehler des Newton-Verfahrens zunächst bei $n = 1, 2, 3$ nur moderat verkleinert, wohingegen er ab $n = 4$ rasant abnimmt. ♠

Was kann man über den **absoluten Fehler** des Newton-Verfahrens aussagen? Der Satz von Taylor (siehe Satz 1.23) liefert uns für die Darstellung von $f(x_n)$ durch das konstante Taylor-Polynom mit dem Entwicklungspunkt $x = z$, dass es ein z_n zwischen x_n und z gibt, so dass gilt

$$f(x_n) = f(z) + f'(z_n)\,(x_n - z).$$

Da $f(z) = 0$ ist folgt

$$f(x_n) = f'(z_n)\,(x_n - z),$$

und Auflösen nach $z - x_n$ liefert:

$$f(x_n) = f'(z_n)\,(x_n - z) \qquad \overset{f'(z_n) \neq 0}{\Longleftrightarrow} \qquad \frac{f(x_n)}{f'(z_n)} = x_n - z$$

$$\iff \qquad -\frac{f(x_n)}{f'(z_n)} = -(x_n - z) \qquad \iff \qquad z - x_n = \frac{-f(x_n)}{f'(z_n)}$$

Ist x_n dicht genug bei z, so folgt für z_n (welches zwischen x_n und z liegt) $z_n \approx x_n$ und somit $f'(z_n) \approx f'(x_n)$. Damit erhalten wir:

$$z - x_n = \frac{-f(x_n)}{f'(z_n)} \approx \frac{-f(x_n)}{f'(x_n)} = \underbrace{x_n - \frac{f(x_n)}{f'(x_n)}}_{= x_{n+1}} - x_n = x_{n+1} - x_n$$

Also gilt, falls x_n dicht genug bei z liegt für den **absoluten Fehler von** x_n:

$$\boxed{\text{Fehler}(x_n) = z - x_n \approx x_{n+1} - x_n} \tag{3.5}$$

Da die Nullstelle z in der Regel nicht bekannt ist, kann man den absoluten Fehler $\text{Fehler}(x_n) = z - x_n$ nicht direkt berechnen. Formel (3.5) ist sehr hilfreich, um $\text{Fehler}(x_n) = z - x_n$ näherungsweise mit $\text{Fehler}(x_n) \approx x_{n+1} - x_n$ zu berechnen.

Beispiel 3.8. (absoluter Fehler des Newton-Verfahrens)
Betrachten wir die Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

deren größte Nullstelle in Beispiel 3.7 mit den Newton-Verfahren angenähert berechnet wurde. In den letzten beiden Spalten von Tabelle 3.2 zu Beispiel 3.7 lesen wir ab für $n \geq 3$ ab:

$$z - x_3 \doteq -4{,}73 \cdot 10^{-3}, \qquad x_4 - x_3 \doteq -4{,}68 \cdot 10^{-3},$$
$$z - x_4 \doteq -5{,}35 \cdot 10^{-5}, \qquad x_5 - x_4 \doteq -5{,}35 \cdot 10^{-5},$$
$$z - x_5 \doteq -6{,}91 \cdot 10^{-9}, \qquad x_6 - x_5 \doteq -6{,}91 \cdot 10^{-9}.$$

Wir sehen, dass die Näherung $\text{Fehler}(x_n) = z - x_n \approx x_{n+1} - x_n$ in (3.5) in diesem Beispiel bereits ab $n = 3$ relativ gute Näherungen für den absoluten Fehler von x_n liefert. ♠

Wir wollen nun das **Verhalten des absoluten Fehlers** des Newton-Verfahrens weiter untersuchen. Dazu setzen wir voraus, dass die Funktion $f :]a; b[\to \mathbb{R}$, deren Nullstelle z wir berechnen wollen, **zweimal stetig differenzierbar** ist und dass für die Ableitung in der Nullstelle z gilt

$$f'(z) \neq 0. \tag{3.6}$$

Die Bedingung (3.6) bedeutet, dass die Tangente an den Graphen vom f im Punkt z nicht parallel zur x-Achse ist. Wegen der Stetigkeit der Ableitung f' folgt daraus, dass $f'(x) \neq 0$ für alle x gilt, die dicht genug bei z liegen.

Wir nutzen nun den Satz von Taylor (siehe Satz 1.23), um $f(z)$ durch das lineare Taylor-Polynom von f mit dem Entwicklungspunkt x_n plus Restglied darzustellen. Nach dem Satz von Taylor gibt es ein z_n zwischen z und x_n, so dass gilt

$$f(z) = f(x_n) + f'(x_n)\,(z - x_n) + \frac{1}{2}\,f''(z_n)\,(z - x_n)^2. \tag{3.7}$$

Da z eine Nullstelle von f ist, gilt $f(z) = 0$, und somit folgt aus (3.7)

$$0 = f(x_n) + f'(x_n)\,(z - x_n) + \frac{1}{2}\,f''(z_n)\,(z - x_n)^2. \tag{3.8}$$

Wir teilen in (3.8) durch $f'(x_n) \neq 0$ und formen weiter um:

$$\begin{aligned}
& 0 = \frac{f(x_n)}{f'(x_n)} + (z - x_n) + \frac{1}{2}\,\frac{f''(z_n)}{f'(x_n)}\,(z - x_n)^2 \\
\iff \qquad & 0 = z - \underbrace{\left(x_n - \frac{f(x_n)}{f'(x_n)}\right)}_{= x_{n+1}} + \frac{f''(z_n)}{2\,f'(x_n)}\,(z - x_n)^2 \\
\iff \qquad & 0 = z - x_{n+1} + \frac{f''(z_n)}{2\,f'(x_n)}\,(z - x_n)^2 \\
\iff \qquad & z - x_{n+1} = -\,\frac{f''(z_n)}{2\,f'(x_n)}\,(z - x_n)^2 = \left[\frac{-f''(z_n)}{2\,f'(x_n)}\right]\,(z - x_n)^2
\end{aligned}$$

Also gilt für den absoluten Fehler von x_{n+1}

$$\text{Fehler}(x_{n+1}) = z - x_{n+1} = \left[\frac{-f''(z_n)}{2\,f'(x_n)}\right]\,\underbrace{(z - x_n)^2}_{= [\text{Fehler}(x_n)]^2}. \tag{3.9}$$

Falls x_n dicht bei z liegt (und damit auch z_n dicht bei z liegt) so gilt in (3.9) angenähert:

$$\boxed{\frac{-f''(z_n)}{2\,f'(x_n)} \approx \frac{-f''(z)}{2\,f'(z)} = M} \tag{3.10}$$

Setzt man (3.10) in (3.9) ein, so erhält man angenähert:

$$\boxed{\text{Fehler}(x_{n+1}) = z - x_{n+1} \approx M\,(z - x_n)^2} \tag{3.11}$$

Multiplikation von (3.11) mit M liefert

$$\boxed{M\,(z - x_{n+1}) \approx M^2\,(z - x_n)^2 = \big[M\,(z - x_n)\big]^2.} \tag{3.12}$$

Sind alle Näherungen x_n, $n = 0, 1, 2, \ldots$ dicht bei z, so dass (3.10) für alle $n = 0, 1, 2, \ldots$ gilt, so folgt durch wiederholte Anwendung von (3.12):

$$\begin{aligned} M\,(z - x_{n+1}) &\approx \big[M\,(z - x_n)\big]^2 \approx \Big[\big[M\,(z - x_{n-1})\big]^2\Big]^2 = \big[M\,(z - x_{n-1})\big]^4 \\ &\approx \Big[\big[M\,(z - x_{n-2})\big]^2\Big]^4 = \big[M\,(z - x_{n-2})\big]^8 = \big[M\,(z - x_{n-2})\big]^{2^3} \\ &\approx \ldots \approx \big[M\,(z - x_1)\big]^{2^n} \approx \big[M\,(z - x_0)\big]^{2^{n+1}} \end{aligned}$$

Also erhalten wir, falls alle x_n, $n = 0, 1, 2, \ldots$, dicht bei z liegen:

$$\boxed{M\,(z - x_n) \approx \big[M\,(z - x_0)\big]^{2^n}, \qquad n = 0, 1, 2, \ldots} \tag{3.13}$$

Aus (3.13) folgt, dass der **absolute Fehler von x_n gegen 0 strebt, wenn gilt**

$$\big|M\,(z - x_0)\big| < 1 \qquad \Longleftrightarrow \qquad \boxed{|z - x_0| < \frac{1}{|M|} = \left|\frac{2\,f'(z)}{-f''(z)}\right|.} \tag{3.14}$$

Wir sehen an (3.14), dass der Startwert x_0 sehr dicht bei z liegen muss, wenn $|M|$ sehr groß ist. Wird der Startwert x_0 also nicht hinreichend dicht bei z gewählt, so dass (3.14) nicht erfüllt ist, so wird das Newton-Verfahren normalerweise nicht gegen die Nullstelle z konvergieren.

Wie man einen guten Startwert x_0 wählt, hängt vom konkreten Beispiel ab: Dieses kann durch Zeichnen des Graphen der Funktion passieren, oder der Startwert ist bei praktischen Anwendungen durch physikalische Überlegungen zu der Problemstellung gegeben. Liegt kein guter Startwert vor, so kann es sinnvoll sein, zunächst ein paar Schritte mit dem Bisektionsverfahren durchzuführen, um einen besseren Startwert zu erhalten.

Wir illustrieren die vorherigen theoretischen Überlegungen an einem Beispiel.

Beispiel 3.9. (absoluter Fehler des Newton-Verfahrens)
Betrachten wir die Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

deren größte Nullstelle in Beispiel 3.7 mit den Newton-Verfahren angenähert berechnet wurde. Die erste und zweite Ableitung dieser Funktion sind

$$f'(x) = 6\,x^5 - 1, \qquad f''(x) = 30\,x^4$$

Wenn x_n dicht bei der Nullstelle $z \doteq 1{,}134724138$ ist, dann erhalten in (3.10) wir (mit Rundung auf eine 3-stellige Gleitkommadarstellung)

$$M = \frac{-f''(z)}{2\,f'(z)} = \frac{-30\,z^4}{2\,(6\,z^5 - 1)} \doteq -2{,}42,$$

also $M \doteq -2{,}42$. Nach (3.11) gilt dann

$$\text{Fehler}(x_{n+1}) = z - x_{n+1} \approx M\,(z - x_n)^2 = -2{,}42\,(z - x_n)^2. \tag{3.15}$$

Betrachten wir beispielsweise $n = 3$ mit $z - x_3 \doteq -4{,}73 \cdot 10^{-3}$ (vgl. Tabelle 3.2). Dann sollte nach (3.15) gelten (mit Rundung auf eine 3-stellige Gleitkommadarstellung)

$$\text{Fehler}(x_4) = z - x_4 \approx -2{,}42\,(z - x_3)^2 \doteq -2{,}42 \cdot \big(-4{,}73 \cdot 10^{-3}\big)^2 \doteq -5{,}41 \cdot 10^{-5}.$$

In Tabelle 3.2 finden wir $z - x_4 \doteq -5{,}53 \cdot 10^{-5}$. Bereits für $n = 4$ liefert (3.15) also eine gute Vorhersage für den absoluten Fehler.

Untersuchen wir noch, ob die Konvergenzbedingung (3.14) an den Startwert in diesem Beispiel erfüllt ist: Wir finden

$$|z - x_0| \doteq |1{,}134724138 - 1{,}5| \doteq 0{,}365 < \frac{1}{|M|} \doteq 0{,}414,$$

d.h. die Konvergenzbedingung war für unseren Startwert $x_0 = 1{,}5$ erfüllt. ♠

Wir halten in einer Bemerkung fest, was wir über die Konvergenz des Newton-Verfahrens gelernt haben:

Bemerkung 3.10. (Konvergenz des Newton-Verfahrens)
Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion mit einer Nullstelle z in $]a; b[$, und es sei $f'(z) \neq 0$. Dann gelten:

(1) Falls der Startwert x_0 des Newton-Verfahrens **„dicht genug" bei** z liegt, **konvergiert** das Newton-Verfahren gegen z, also $\lim\limits_{n\to\infty} x_n = z$.

(2) Falls die Iterierten x_n dicht genug bei z liegen, gilt für den **absoluten Fehler** die Näherung

$$\text{Fehler}(x_n) = z - x_n \approx x_{n+1} - x_n.$$

(3) Falls f sogar zweimal stetig differenzierbar ist und falls die Iterierten x_n dicht genug bei z liegen, gilt für den **absoluten Fehler**

$$\text{Fehler}(x_n) = z - x_n \approx M\,(z - x_{n-1})^2 \quad \text{mit} \quad M = \frac{-f''(z)}{2\,f'(z)}\,.$$

Das Newton-Verfahren **konvergiert normalerweise gegen z, wenn** der Startwert x_0 die folgende Bedingung erfüllt:

$$|z - x_0| < \frac{1}{|M|} = \left|\frac{2\,f'(z)}{-f''(z)}\right|$$

3.3 Sekantenverfahren

Beim **Sekantenverfahren** wird die Näherung der Nullstelle einer stetig differenzierbaren Funktion mit Hilfe von zwei Näherungen und der Sekante durch die beiden zugehörigen Punkte auf dem Graphen bestimmt. Genauer funktioniert dieses wie folgt (siehe auch Abbildung 3.4):

Herleitung des Sekantenverfahrens: Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion mit einer Nullstelle z, also $f(z) = 0$. Seien x_0 und x_1 zwei Näherungswerte für z. Diese können entweder beide auf einer Seite der Nullstelle liegen oder auf gegenüberliegenden Seiten der Nullstelle. Wir legen nun die Sekante durch die beiden Punkte $\big(x_0; f(x_0)\big)$ und $\big(x_1; f(x_1)\big)$ auf dem Graphen von f und bestimmen deren Schnittpunkt x_2 mit der x-Achse. Dieser Schnittpunkt ist die neue Näherung für die Nullstelle z.

Die Gleichung der Sekante ist

$$p(x) = f(x_1) + \frac{f(x_1) - f(x_0)}{x_1 - x_0} \cdot (x - x_1).$$

(Es liegt ein Polynom ersten Grades vor, und in der Tat gelten $p(x_1) = f(x_1)$ und $p(x_0) = f(x_1) + \big(f(x_1) - f(x_0)\big) \cdot (-1) = f(x_0)$.) Wir lösen $p(x_2) = 0$ nach x_2

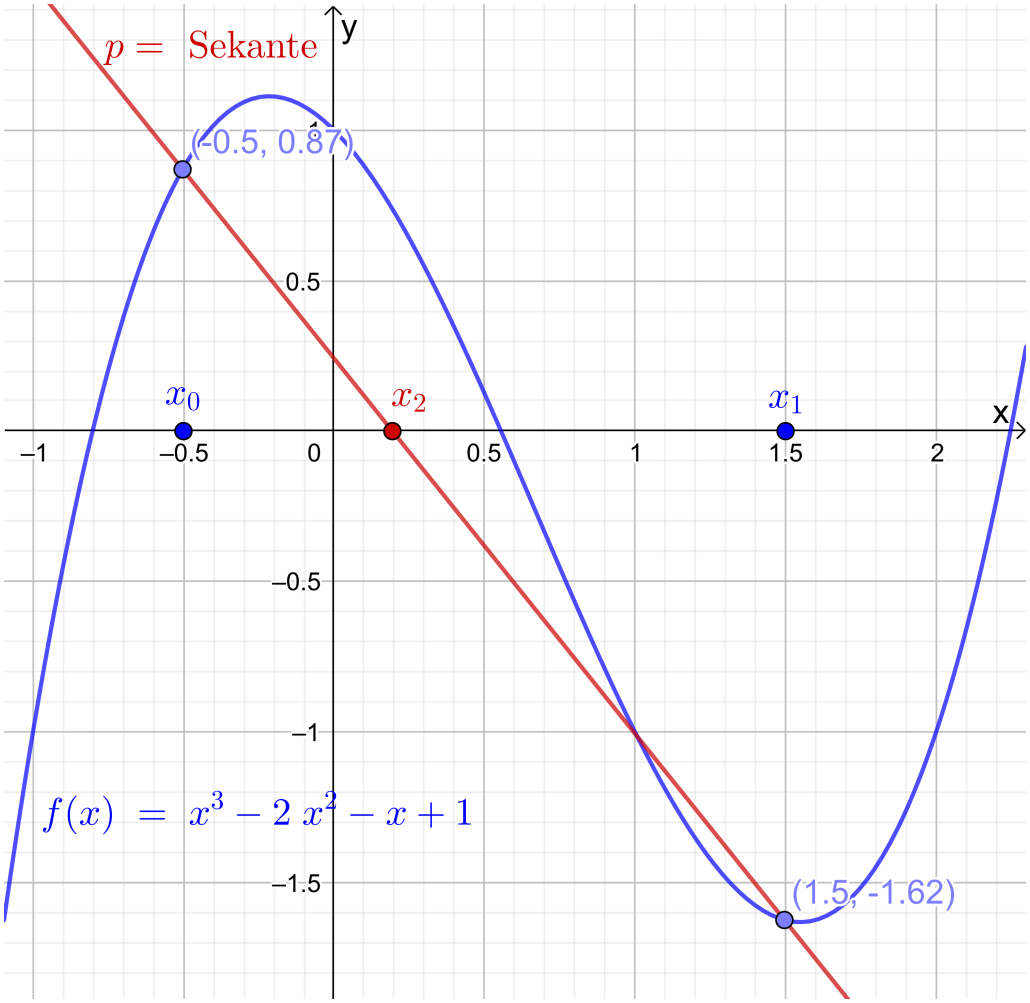


Abb. 3.4: Illustration des Sekantenverfahrens: Für die Startwerte $x_0 = -0{,}5$ und $x_1 = 1{,}5$ legen wir die Sekante p durch $\big(x_0; f(x_0)\big)$ und $\big(x_1; f(x_1)\big)$. Die neue Näherung x_2 für die Nullstelle ist der Schnittpunkt der Sekante mit der x-Achse.

auf, um die Schnittstelle x_2 der Sekante mit der x-Achse zu bestimmen:

$$\begin{aligned} 0 = p(x_2) &= f(x_1) + \frac{f(x_1) - f(x_0)}{x_1 - x_0} \cdot (x_2 - x_1) \\ \iff \quad -f(x_1) &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \cdot (x_2 - x_1) \\ \iff \quad -f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)} &= x_2 - x_1 \\ \iff \quad x_1 - f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)} &= x_2 \end{aligned}$$

Also erhalten wir als neue Näherung für die Nullstelle

$$x_2 = x_1 - f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)} \,. \tag{3.16}$$

Wir wiederholen diese Vorgehensweise mit den beiden Näherungen x_1 und x_2 für die Nullstelle z und erhalten als Schnittpunkt x_3 der Sekante durch $\big(x_1, f(x_1)\big)$ und $\big(x_2, f(x_2)\big)$ mit der x-Achse (ersetze in (3.16) jeweils x_2 durch x_3, x_1 durch x_2 und x_0 durch x_1)

$$x_3 = x_2 - f(x_2) \cdot \frac{x_2 - x_1}{f(x_2) - f(x_1)}\,. \tag{3.17}$$

Wir können diese Vorgehensweise nun mit den Näherungen x_2 und x_3 für die Nullstelle z fortsetzen. Das Wiederholen dieses Prozesses liefert nach $n-1$ Schritten

$$\boxed{x_{n+1} = x_n - f(x_n) \cdot \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\,.} \tag{3.18}$$

(Natürlich sind (3.16) und (3.17) als Sonderfälle von (3.18) für $n = 1$ bzw. $n = 2$ in (3.18) enthalten.) Wir halten das Sekantenverfahren als Algorithmus fest.

Verfahren 3.11. (Sekantenverfahren)
Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion, die in z eine Nullstelle hat, und seien x_0 und x_1 zwei verschiedene (hinreichend gute) Näherungswerte für die Nullstelle z. Das ***Sekantenverfahren*** *berechnet Näherungen für die Nullstelle z mit dem folgenden Iterationsverfahren*

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\,, \qquad n = 1, 2, \ldots. \tag{3.19}$$

Das Sekantenverfahren ist ein **zweistufiges** Iterationsverfahren, weil zur Berechnung der neuen Näherung x_{n+1} zwei Näherungswerte x_n und x_{n-1} benötigt werden. Das Bisektionsverfahren ist ebenfalls ein zweistufiges Iterationsverfahren. Allerdings konvergiert das Sekantenverfahren normalerweise deutlich schneller als das Bisektionsverfahren. – Das Newton-Verfahren ist dagegen ein **einstufiges** Iterationsverfahren, da zur Berechnung der neuen Näherung x_{n+1} nur die Näherung x_n benötigt wird.

Betrachten wir als Beispiel wieder die Funktion, deren größte Nullstelle in den Beispielen 3.4 und 3.7 bereits mit den Bisektionsverfahren bzw. mit dem Newton-Verfahren bestimmt wurde.

Beispiel 3.12. (Sekantenverfahren)
Gesucht ist eine Näherung der größten Nullstelle der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

n	x_n	$f(x_n)$	$x_n - x_{n-1}$	$z - x_{n-1}$
0	2,00000000	$6{,}1 \cdot 10^{1}$		
1	1,00000000	$-1{,}0 \cdot 10^{0}$	$-1{,}0 \cdot 10^{0}$	
2	1,01612903	$-9{,}15 \cdot 10^{-1}$	$1{,}61 \cdot 10^{-2}$	$1{,}35 \cdot 10^{-1}$
3	1,19057777	$6{,}57 \cdot 10^{-1}$	$1{,}74 \cdot 10^{-1}$	$1{,}19 \cdot 10^{-1}$
4	1,11765583	$-1{,}68 \cdot 10^{-1}$	$7{,}29 \cdot 10^{-2}$	$5{,}59 \cdot 10^{-2}$
5	1,13253155	$-2{,}24 \cdot 10^{-2}$	$1{,}49 \cdot 10^{-2}$	$1{,}71 \cdot 10^{-2}$
6	1,13481681	$9{,}54 \cdot 10^{-4}$	$2{,}29 \cdot 10^{-3}$	$2{,}19 \cdot 10^{-3}$
7	1,13472365	$-5{,}07 \cdot 10^{-6}$	$-9{,}32 \cdot 10^{-5}$	$-9{,}27 \cdot 10^{-5}$
8	1,13472414	$-1{,}13 \cdot 10^{-9}$	$4{,}92 \cdot 10^{-7}$	$4{,}92 \cdot 10^{-7}$

Tabelle 3.3: Sekantenverfahren zur Berechnung der größten Nullstelle der Funktion $f(x) = x^6 - x - 1$ mit den Startwerten $x_0 = 2$ und $x_1 = 1$.

die bereits in Beispiel 3.4 und 3.7 betrachtet wurde. In Beispiel 3.4 hatten wir uns überlegt, dass die größte Nullstelle im Intervall $[1; 2]$ liegt. Wir nehmen daher als Startwerte für das Sekantenverfahren $x_0 = 2$ und $x_1 = 1$.

Die Iterierten x_n sind für $n = 0, 1, 2, \ldots, 8$ in Tabelle 3.3 auf eine 9-stellige Gleitkommadarstellung gerundet aufgelistet. Weiter sind in Tabelle 3.3 in der Zeile für x_n der Funktionswert $f(x_n)$, der absolute Fehler $z - x_{n-1}$, sowie $x_n - x_{n-1}$ als Näherung für $z - x_{n-1}$ jeweils auf eine 3-stellige Gleitkommadarstellung gerundet angegeben.

Die gesuchte Nullstelle ist $z \doteq 1{,}134724138$, und wir sehen, dass die Näherung $x_8 \doteq 1{,}13472414$ bereits 8 signifikante Ziffern hat.

An der zweitletzten Spalte beobachten wir (wie auch beim Newton-Verfahren), dass der absolute Fehler zunächst nur langsam abnimmt, aber dann ab $n = 5$ rasant keiner wird. ♠

Die Fehleranalyse des Sekantenverfahrens ist mathematisch komplizierter als die des Newton-Verfahrens, und wir geben daher nur die Ergebnisse an. (Für die Herleitung dieser Ergebnisse siehe beispielsweise [4, Teilkapitel 5.3.1].)

Bemerkung 3.13. (Konvergenz des Sekantenverfahrens)
Sei $f :]a; b[\to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion mit einer Null-

stelle z, also $f(z) = 0$, und sei $f'(z) \neq 0$.

(1) Wenn x_0 und x_1 **dicht genug** bei der Nullstelle z liegen, dann **konvergiert das Sekantenverfahren** gegen z, und es gilt

$$\lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|^r} = \left|\frac{f''(z)}{2\,f'(z)}\right|^{r-1} = |M|^{r-1} \qquad \text{mit} \quad M = \frac{-f''(z)}{2\,f'(z)},$$

wobei $r = (\sqrt{5}+1)/2 \doteq 1{,}62$. Für x_n dicht genug bei z gilt daher

$$|z - x_{n+1}| \approx |M|^{r-1}\,|z - x_n|^r \quad \text{mit} \quad r = (\sqrt{5}+1)/2 \doteq 1{,}62. \tag{3.20}$$

(2) Aus (3.20) folgt durch Multiplizieren mit $|M|$

$$|M|\,|z - x_{n+1}| \approx \big(|M|\,|z - x_n|\big)^r \quad \text{mit} \quad r = (\sqrt{5}+1)/2 \doteq 1{,}62,$$

und wiederholte Anwendung dieser Formel liefert

$$|M|\,|z - x_{n+1}| \approx \big(|M|\,|z - x_1|\big)^{r^n} \quad \text{mit} \quad r = (\sqrt{5}+1)/2 \doteq 1{,}62.$$

Da r^n mit wachsendem n beliebig groß wird, kann man nur erwarten, dass das **Sekantenverfahren konvergiert, wenn für x_1 gilt**

$$|M|\,|z - x_1| < 1 \quad \Longleftrightarrow \quad |z - x_1| < \frac{1}{|M|} = \left|\frac{2\,f'(z)}{-f''(z)}\right|. \tag{3.21}$$

(3) Aus (3.20) folgt für x_n dicht genug bei z eine **Näherung für den absoluten Fehler** von x_{n-1}:

$$\text{Fehler}(x_{n-1}) = z - x_{n-1} \approx x_n - x_{n-1} \tag{3.22}$$

Wir machen uns die Informationen zur Konvergenz an Beispiel 3.12 klar.

Beispiel 3.14. (Sekantenverfahren)
In Beispiel 3.12 wurde die größte Nullstelle $z \doteq 1{,}134724138$ der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

mit dem Sekantenverfahren mit den Startwerten $x_0 = 2$ und $x_1 = 1$ berechnet.

Wir haben die Konstante

$$M = \frac{-f''(z)}{2\,f'(z)} = \frac{-30\,z^4}{2\,(6\,z^5-1)} \doteq -2{,}42 \qquad \Longrightarrow \qquad |M| \doteq 2{,}42.$$

In Tabelle 3.3 lesen wir ab:

$$|z - x_5| \doteq 2{,}19 \cdot 10^{-3} \qquad \text{und} \qquad |z - x_4| \doteq 1{,}71 \cdot 10^{-2}.$$

Daher gilt für die rechte Seite in (3.20) mit $n = 4$

$$|M|^{r-1}\,|z - x_4|^r \doteq (2{,}42)^{0{,}62} \cdot (1{,}71 \cdot 10^{-2})^{1{,}62} \doteq 2{,}37 \cdot 10^{-3}.$$

Die Näherung (3.20) ist also bereits für $n + 1 = 5$ ziemlich gut erfüllt.

Wie sieht es mit der Konvergenzbedingung (3.21) aus? Wir erhalten für $x_1 = 1$

$$|z - x_1| \doteq |1{,}134724138 - 1| \doteq 0{,}135 < \frac{1}{|M|} \doteq 0{,}414,$$

d.h. die Konvergenzbedingung war für unsere Startwerte erfüllt.

Betrachten wir noch (3.22) exemplarisch für $n = 5$: In Tabelle 3.3 lesen wir ab, dass gelten

$$\text{Fehler}(x_5) = z - x_5 \doteq 2{,}19 \cdot 10^{-3} \qquad \text{und} \qquad x_6 - x_5 \doteq 2{,}29 \cdot 10^{-3}.$$

Also liefert $x_6 - x_5$ in der Tat eine gute Näherung für den absoluten Fehler der Iterierten x_5. Für $n = 6$ und $n = 7$ sind die Näherungen $x_n - x_{n-1} \approx z - x_{n-1}$ in Tabelle 3.3 ähnlich gut. ♠

3.4 Vergleich des Newton-Verfahrens und des Sekantenverfahrens

Wir vergleichen das Newton-Verfahren und das Sekantenverfahren.

Beide Verfahren dienen zur Berechnung von Nullstellen stetig differenzierbarer Funktionen f. In beiden Verfahren kommt eine **typische Vorgehensweise der numerischen Analysis** zum Einsatz: **Da das Problem** der Nullstellenberechnung der Funktion f **zu kompliziert ist, ersetzen wir es durch ein einfacheres Problem, welches das ursprüngliche Problem angenähert löst.** Konkret wird die Funktion f durch eine lineare Funktion angenähert, deren Nullstelle nun bestimmt wird. (Beim Newton-Verfahren wird f durch die Gleichung

der Tangente an den Graphen im Punkt $\big(x_n; f(x_n)\big)$ einer gegebenen Näherung x_n der Nullstelle bzw. beim Sekantenverfahren durch die Sekante durch die Punkte $\big(x_{n-1}; f(x_{n-1})\big)$ und $\big(x_n; f(x_n)\big)$ für zwei Näherungen x_{n-1} und x_n der Nullstelle angenähert.) Wiederholung dieser Vorgehensweise führt auf ein **konvergentes Iterationsverfahren, wenn die Startwerte gut genug sind**.

Wo liegen die **Unterschiede** und damit auch die **Vor- und Nachteile** der beiden Verfahren?

	Newton-Verfahren	**Sekantenverfahren**
Typ des Verfahrens	Einschrittverfahren, denn x_{n+1} wird nur unter Nutzung von x_n berechnet.	Zweischrittverfahren, denn x_{n+1} wird unter Nutzung von x_n und x_{n-1} berechnet.
Konvergenz	ähnliche Voraussetzungen an Startwert	ähnliche Voraussetzungen an Startwert
„Geschwindigkeit“ der Konvergenz	schnellere Konvergenz als Sekantenverfahren	langsamere Konvergenz als Newton-Verfahren
Konvergenzordnung (siehe Teilkapitel 3.7)	2	$r = (\sqrt{5}+1)/2 \doteq 1{,}62$
Rechenaufwand zur Berechnung von x_{n+1} und Laufzeit	Zwei neue Funktionsauswertungen $f(x_n)$, $f'(x_n)$ sind erforderlich. f' kann aufwendig zu berechnen sein; daher kann es sein, dass das Sekantenverfahren von der Laufzeit her schneller ist.	Eine neue Funktionsauswertung $f(x_n)$ ist erforderlich (denn $f(x_{n-1})$ wurde bereits im vorigen Schritt berechnet). f' wird nicht benötigt; daher kann die Laufzeit trotz höherer Iterationsschrittzahl kürzer sein als beim Newton-Verfahren.

Es ist wichtig zu beachten, dass die Aussagen bzgl. der „Geschwindigkeit“ der Konvergenz grundsätzlicher Natur sind. Es gibt Beispiele, bei denen das Sekanten-Verfahren schneller konvergiert als das Newton-Verfahren.

3.5 Fixpunktiteration

In diesem Teilkapitel lernen wir eine allgemeine Theorie für Einschrittverfahren kennen. Wir beginnen mit einem Beispiel.

Beispiel 3.15. (Nullstellenproblem als Fixpunktgleichung)
Wir betrachten die Gleichung

$$x^2 - 5 = 0 \quad \Longleftrightarrow \quad x^2 = 5 \quad \Longleftrightarrow \quad 5 - x^2 = 0 \quad \Longleftrightarrow \quad 1 - \frac{x^2}{5} = 0$$

mit der positiven Lösung $x = \sqrt{5} \doteq 2{,}2361$ (und der negativen Lösung $-\sqrt{5}$). Die Gleichung lässt sich wie folgt umformen:

$$x = 5 + x - x^2 \quad \big(\text{addiere } x \text{ zu } 5 - x^2 = 0\big)\,, \tag{3.23}$$

$$x = \frac{5}{x} \quad \big(\text{dividiere } x^2 = 5 \text{ durch } x \neq 0\big)\,, \tag{3.24}$$

$$x = 1 + x - \frac{x^2}{5} \quad \left(\text{addiere } x \text{ zu } 1 - \frac{x^2}{5} = 0\right), \tag{3.25}$$

$$x = \frac{1}{2}\left(x + \frac{5}{x}\right) \quad \left(\text{multipliziere (3.24) mit } \frac{1}{2} \text{ und addiere dann } \frac{1}{2}\,x\right). \tag{3.26}$$

Definieren wir nun die Funktionen

$$g_1 : \mathbb{R} \to \mathbb{R}, \qquad g_1(x) = 5 + x - x^2, \tag{3.27}$$

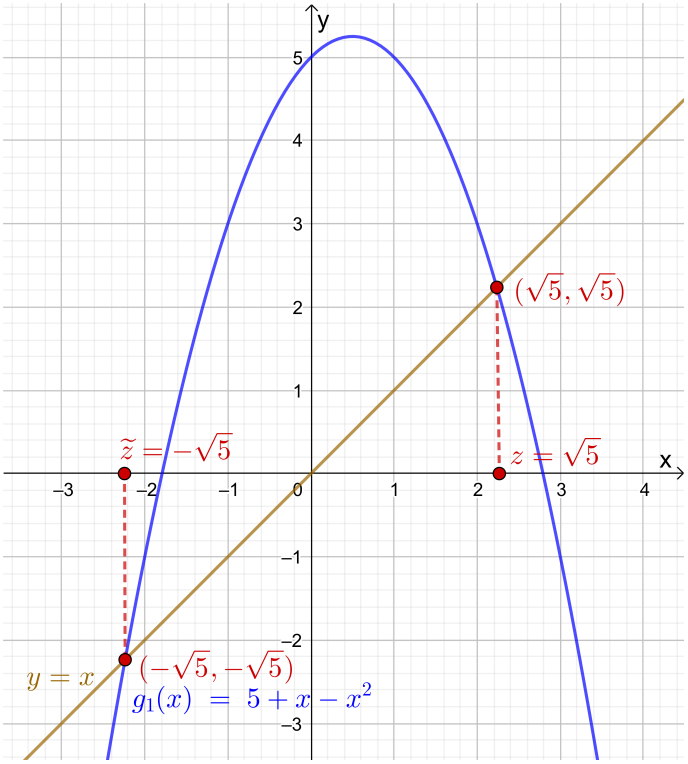
$$g_2 : \mathbb{R} \setminus \{0\} \to \mathbb{R}, \qquad g_2(x) = \frac{5}{x}\,, \tag{3.28}$$

$$g_3 : \mathbb{R} \to \mathbb{R}, \qquad g_3(x) = 1 + x - \frac{x^2}{5}\,, \tag{3.29}$$

$$g_4 : \mathbb{R} \setminus \{0\} \to \mathbb{R}, \qquad g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right), \tag{3.30}$$

so folgt aus (3.23), (3.24), (3.25) und (3.26), dass wir $x^2 - 5 = 0$ für $x \neq 0$ jeweils als die „Fixpunktgleichungen“ $g_k(x) = x$, $k \in \{1, 2, 3, 4\}$, schreiben können.

Die Funktionen g_k, $k \in \{1, 2, 3, 4\}$, mit ihren Fixpunkten $z = \sqrt{5}$ und $\widetilde{z} = -\sqrt{5}$ sind in Abbildungen 3.5 und 3.6 gezeichnet. Man findet die Fixpunkte einer Funktion, indem man die Schnittpunkte des Graphen der Funktion mit der Winkelhalbierenden $y = x$ bestimmt. ♠



Grafische Bestimmung der Fixpunkte: Man findet die Fixpunkte grafisch, indem man die Schnittpunkte des Graphen der Funktion mit der Winkelhalbierenden $y = x$ (in braun gezeichnet) bestimmt.

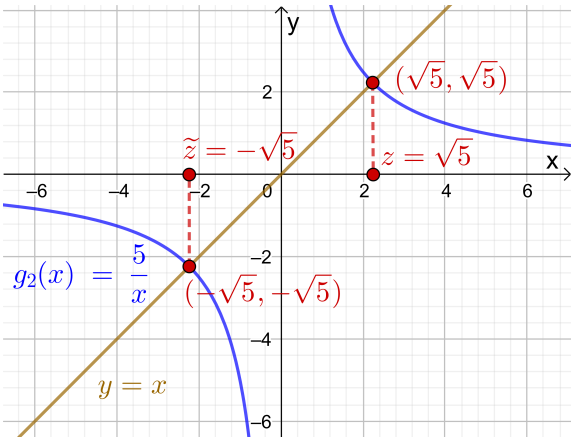


Abb. 3.5: Die Funktionen $g_1(x) = 5 + x - x^2$ und $g_2(x) = \frac{5}{x}$ mit ihren Fixpunkten $z = \sqrt{5}$ und $\widetilde{z} = -\sqrt{5}$.

Definition 3.16. (Fixpunkt und Fixpunktgleichung)
Sei $g : [a; b] \to \mathbb{R}$ eine Funktion.

*(1) Die Gleichung $g(x) = x$ nennt man eine **Fixpunktgleichung**.*

*(2) Ein $z \in [a; b]$ mit der Eigenschaft $g(z) = z$ heißt ein **Fixpunkt von** g.*

Beispiel 3.17. (Fixpunkte und Fixpunktgleichungen)

(a) Die Funktion $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$, hat genau zwei Fixpunkte, denn die Fixpunktgleichung

$$g(x) = x \quad \Longleftrightarrow \quad x^2 = x \quad \Longleftrightarrow \quad 0 = x^2 - x = x\,(x - 1)$$

hat genau die zwei Lösungen $x = 0$ und $x = 1$.

(b) Die Funktion $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2 - x + 2$, hat keinen Fixpunkt, denn die Fixpunktgleichung

$$g(x) = x \quad \Longleftrightarrow \quad x^2 - x + 2 = x \quad \Longleftrightarrow$$
$$0 = x^2 - 2\,x + 2 = (x^2 - 2\,x + 1) + 1 = (x - 1)^2 + 1$$

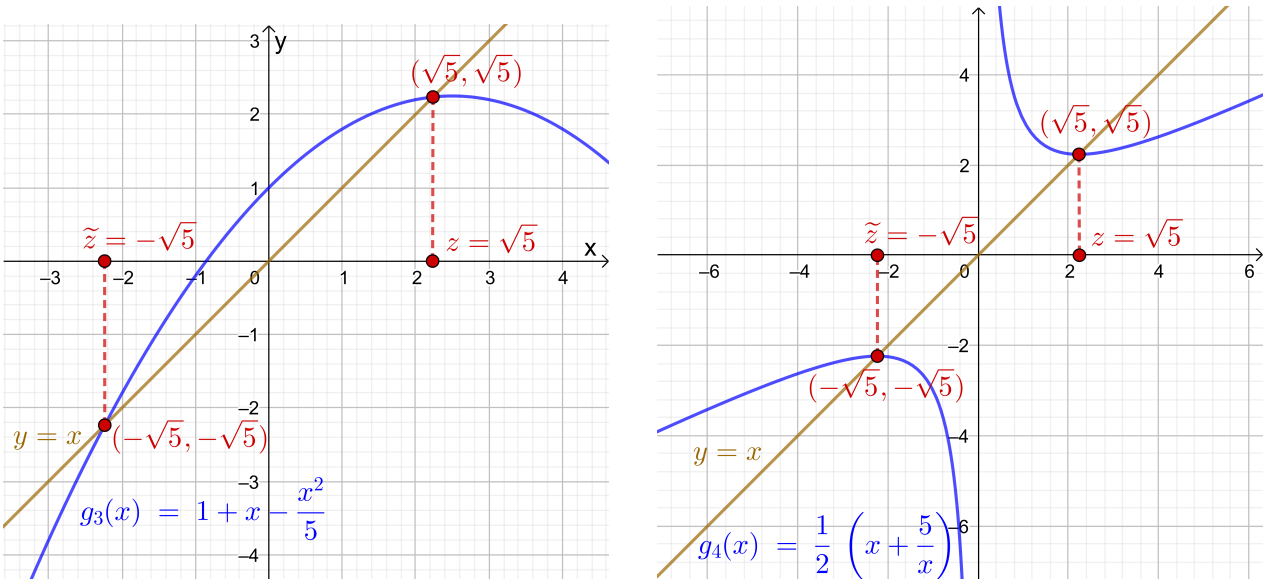


Abb. 3.6: Die Funktionen $g_3(x) = 1 + x - \frac{x^2}{5}$ und $g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right)$ mit ihren Fixpunkten $z = \sqrt{5}$ und $\widetilde{z} = -\sqrt{5}$.

hat keine reelle Lösung, da Quadrate ≥ 0 sind und somit gilt $(x-1)^2+1 \geq 1$.

(c) Die Funktion $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^3 + x$, hat genau einen Fixpunkt, denn die Fixpunktgleichung

$$g(x) = x \qquad \Longleftrightarrow \qquad x^3 + x = x \qquad \Longleftrightarrow \qquad x^3 = 0$$

hat nur die einzige Lösung $x = 0$.

(d) Die Funktionen g_1, g_2, g_3 und g_4 aus Beispiel 3.15 haben alle die zwei Fixpunkte $x = \sqrt{5}$ und $x = -\sqrt{5}$.

Wie wir in Beispiel 3.15 gesehen haben, kann man eine Nullstellengleichung auf verschiedene Weisen in eine Fixpunktgleichung umformen. ♠

Sei g eine **stetige** Funktion, die einen Fixpunkt z besitzt, also $g(z) = z$. Bei einer **Fixpunktiteration** werden ausgehend von einem Startwert x_0, der im Idealfall bereits dicht bei z liegen sollte, Iterierte mit der Iterationsformel

$$\boxed{x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots} \tag{3.31}$$

berechnet. Unter geeigneten Voraussetzungen an g und den Startwert x_0 (die wir noch untersuchen werden) **konvergiert die Fixpunktiteration** (3.31) **gegen den Fixpunkt** z. – Falls die Folge $(x_n)_{n\geq 0}$ der Iterierten der Fixpunktiteration (3.31) konvergiert, also falls es ein β mit $\lim\limits_{n\to\infty} x_n = \beta$ gibt, so muss dieser Wert β

ein Fixpunkt von g sein, denn wegen der Stetigkeit von g folgt

$$\beta = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n) = g\left(\lim_{n\to\infty} x_n\right) = g(\beta).$$

Wir untersuchen das Verfahren der Fixpunktiteration zunächst experimentell für die verschiedenen Fixpunktgleichungen aus Beispiel 3.15, die alle die gleichen zwei Fixpunkte $\sqrt{5}$ und $-\sqrt{5}$ haben. Danach lernen wir die Theorie der Fixpunktiteration kennen.

Beispiel 3.18. (Fixpunktiteration)

Wir versuchen die positive Lösung $z = \sqrt{5} \doteq 2{,}236067977$ der quadratischen Gleichung $x^2 - 5 = 0$ mittels Fixpunktiteration angenähert zu berechnen. Dazu verwenden wir die vier Funktionen g_1, g_2, g_3 und g_4 aus Beispiel 3.15 (siehe (3.27), (3.28), (3.29) und (3.30) in Beispiel 3.15), deren Fixpunkte $z = \sqrt{5}$ und $\widetilde{z} = -\sqrt{5}$ sind. Wir berechnen also die folgenden Fixpunktiterationen:

$$x_{n+1} = 5 + x_n - x_n^2 \qquad \text{für} \quad g_1(x) = 5 + x - x^2, \tag{3.32}$$

$$x_{n+1} = \frac{5}{x_n} \qquad \text{für} \quad g_2(x) = \frac{5}{x}, \tag{3.33}$$

$$x_{n+1} = 1 + x_n - \frac{x_n^2}{5} \qquad \text{für} \quad g_3(x) = 1 + x - \frac{x^2}{5}, \tag{3.34}$$

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{5}{x_n}\right) \qquad \text{für} \quad g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right). \tag{3.35}$$

Als Startwert verwenden wir jeweils $x_0 = 2{,}5$ mit einem relativen Fehler von ungefähr 12 %. Für $k = 1, 2, 3, 4$ sind die Iterierten x_{n+1} der Fixpunktiteration $x_{n+1} = g_k(x_n)$ für $n = 0, 1, 2, \ldots, 5$ jeweils in Tabelle 3.4 aufgelistet. Dabei wurde auf eine 7-stellige Gleitkommadarstellung gerundet.

Basierend auf der Berechnung von x_n für $n = 1, 2, \ldots, 6$ lässt sich vermuten, dass nur die Fixpunktiterationen für g_3 und g_4 gegen $z = \sqrt{5}$ konvergieren werden, denn hier wird im sechsten bzw. dritten Iterationsschritt eine Näherung für den Fixpunkt $z = \sqrt{5}$ mit sechs signifikanten Ziffern erreicht. Bei g_1 ist die Folge der Iterierten vermutlich unbeschränkt und divergiert. Bei g_2 pendelt die Folge der Iterierten immer zwischen 2,5 und 2 hin und her und ist somit ebenfalls divergent.

Mit der Theorie der Fixpunktiteration können wir das beobachtete Verhalten dann auch erklären. ♠

Wir lernen nun die Theorie der Fixpunktiteration kennen.

n	x_n für g_1	x_n für g_2	x_n für g_3	x_n für g_4
0	2,500000	2,5	2,500000	2,500000
1	1,250000	2	2,250000	2,250000
2	4,687500	2,5	2,237500	2,236111
3	−12,28516	2	2,236219	2,236068
4	−158,2102	2,5	2,236084	2,236068
5	−25.183,68	2	2,236070	2,236068
6	−634.243.100	2,5	2,236068	2,236068

Tabelle 3.4: Fixpunktiteration für die Funktionen g_1, g_2, g_3 und g_4 aus Beispiel 3.18 mit dem Startwert $x_0 = 2{,}5$.

Hilfssatz 3.19. (Bedingung für die Existenz eines Fixpunkts)
*Sei $g : [a; b] \to \mathbb{R}$ eine **stetige** Funktion mit der Eigenschaft, dass*

$$a \leq g(x) \leq b \qquad \text{für alle } x \in [a; b]. \tag{3.36}$$

*Dann hat die Gleichung $g(x) = x$ mindestens eine Lösung in dem Intervall $[a; b]$, d.h. die Funktion g hat in $[a; b]$ **mindestens einen Fixpunkt**.*

Beweis von Hilfssatz 3.19: Aus (3.36) folgt:

$$\begin{aligned} & a \leq g(x) \leq b \quad \text{für alle } x \in [a; b] \quad \Big| -x \\ \iff \quad & a - x \leq g(x) - x \leq b - x \quad \text{für alle } x \in [a; b] \quad \Big| \cdot (-1) \\ \iff \quad & x - a \geq x - g(x) \geq x - b \quad \text{für alle } x \in [a; b] \end{aligned} \tag{3.37}$$

(Beachten Sie dabei, dass sich beim Multiplizieren mit (-1) alle Ungleichheitszeichen umkehren.) Wir definieren nun die Funktion

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

Diese ist stetig, weil g stetig ist. Dann gilt wegen (3.37)

$$x - a \geq f(x) \geq x - b \qquad \text{für alle } x \in [a; b],$$

und für $x = a$ folgt $0 = a - a \geq f(a)$ und für $x = b$ folgt $f(b) \geq b - b = 0$. Also gilt $f(a) \leq 0 \leq f(b)$. Nach dem Zwischenwertsatz für stetige Funktionen (siehe

Satz 3.1) gibt es zu jedem y mit $f(a) \leq y \leq f(b)$, also insbesondere zu $y = 0$, einen Punkt $z \in [a; b]$ mit $f(z) = y$. Also gibt es ein $z \in [a; b]$ mit $f(z) = 0$. Wegen

$$0 = f(z) = z - g(z) \qquad \Longleftrightarrow \qquad z = g(z)$$

ist dieses z ein Fixpunkt von g im Intervall $[a; b]$. □

Nach diesen Vorbereitungen können wir den Fixpunktsatz formulieren.

Satz 3.20. (Fixpunktsatz)
Sei $g :]c; d[\to \mathbb{R}$ eine ***stetig differenzierbare Funktion*** *(d.h. g und g' sind stetig auf $]c; d[$). Es sei $[a; b] \subseteq]c; d[$, und g habe die Eigenschaften*

$$a \leq g(x) \leq b \qquad \text{für alle } x \in [a; b] \tag{3.38}$$

und

$$\max_{a \leq x \leq b} |g'(x)| \leq \lambda < 1 \qquad \text{(mit einer Konstante } \lambda\text{)}. \tag{3.39}$$

Dann gelten:

(1) Es gibt genau eine Lösung $z \in [a; b]$ der Gleichung $x = g(x)$, d.h. g hat ***genau einen*** *Fixpunkt z in $[a; b]$.*

(2) Für jeden Startwert $x_0 \in [a; b]$ ***konvergieren die Iterierten***

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots,$$

gegen den Fixpunkt *z.*

(3) ***Fehlerabschätzung:*** *Mit der Konstante λ aus (3.39) gilt*

$$|z - x_n| \leq \frac{\lambda^n}{1 - \lambda} |x_1 - x_0| \qquad \text{für alle } n = 1, 2, \ldots. \tag{3.40}$$

(4) ***Näherung für den Fehler:*** *Es gilt*

$$\lim_{n \to \infty} \frac{z - x_{n+1}}{z - x_n} = g'(z). \tag{3.41}$$

Daraus folgt für x_n dicht bei z

$$z - x_{n+1} \approx g'(z)\,(z - x_n). \tag{3.42}$$

Da der Beweis von Satz 3.20 instruktiv ist und eine tiefere Einsicht in das Verfahren der Fixpunktiteration bietet, schauen wir uns den Beweis an.

Beweis von Satz 3.20: Als Vorbereitung für den Beweis leiten wir eine wichtige Abschätzung her: Nach dem Mittelwertsatz der Differentialrechnung (siehe Satz 1.4) gilt für zwei beliebige verschiedene Punkte $u, w \in [a; b]$

$$\frac{g(u) - g(w)}{u - w} = g'(c) \qquad \text{mit einem Punkt } c \text{ zwischen } u \text{ und } w \qquad \Longleftrightarrow$$
$$g(u) - g(w) = g'(c)\,(u - w) \qquad \text{mit einem Punkt } c \text{ zwischen } u \text{ und } w.$$

Durch Anwenden des Absolutbetrags erhalten wir

$$|g(u) - g(w)| = |g'(c)|\,|u - w|.$$

Aus Voraussetzung (3.39) folgt $|g'(c)| \leq \lambda$ und wir erhalten somit

$$|g(u) - g(w)| = |g'(c)|\,|u - w| \leq \lambda\,|u - w|.$$

Da $u, w \in [a; b]$ beliebig waren (und da offenbar auch $|g(u) - g(w)| \leq \lambda\,|u - w|$ für $u = w$ gilt), erhalten wir

$$\boxed{|g(u) - g(w)| \leq \lambda\,|u - w| \qquad \text{für alle } u, w \in [a; b].} \tag{3.43}$$

Da die **positive Konstante λ aus** (3.39) **kleiner als 1** ist, nennt man (3.43) auch eine **Kontraktionseigenschaft der Funktion** g: $g(u)$ und $g(w)$ haben einen geringeren Abstand als u und w, g.h. g kontrahiert (Erklärung: „kontrahieren“ bedeutet „zusammenziehen“).

(1) Die Voraussetzung (3.38) im Satz 3.20 gestattet uns Hilfssatz 3.19 anzuwenden. Dieses liefert und die Existenz mindestens eines Fixpunkts $z \in [a; b]$.

Wir müssen nun noch zeigen, dass g genau einen Fixpunkt in $[a; b]$ hat. Dazu geben wir einen Widerspruchsbeweis: Wir nehmen an, z und $\widetilde{z}$ aus $[a; b]$ seien zwei verschiedene Fixpunkte von g. Aus (3.43) folgt dann mit $u = z$ und $w = \widetilde{z}$

$$|g(z) - g(\widetilde{z})| \leq \lambda\,|z - \widetilde{z}|.$$

Da z und $\widetilde{z}$ Fixpunkte von g sind, gelten aber $g(z) = z$ und $g(\widetilde{z}) = \widetilde{z}$. Also erhalten wir

$$|z - \widetilde{z}| \overset{\substack{g(z)=z,\\ g(\widetilde{z})=\widetilde{z}\\ \downarrow}}{=} |g(z) - g(\widetilde{z})| \leq \lambda\,|z - \widetilde{z}| \overset{\substack{\lambda<1\\ \downarrow}}{<} |z - \widetilde{z}|,$$

wobei wir im letzten Schritt $z \neq \widetilde{z}$ und damit $|z - \widetilde{z}| > 0$ genutzt haben. Die Gleichung $|z - \widetilde{z}| < |z - \widetilde{z}|$ ist aber ein Widerspruch. $\lightning$. Also war die Annahme, dass g (mindestens) zwei verschiedene Fixpunkte in $[a; b]$ hat, falsch, und wir haben somit gezeigt, dass g nur einen Fixpunkt in $[a; b]$ hat.

Im Folgenden bezeichnen wir den einzigen Fixpunkt von g in $[a; b]$ mit z.

(2) Aus (3.38) folgt, dass für jeden Startwert $x_0 \in [a;b]$ alle Iterierten $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, im Intervall $[a;b]$ liegen. (Erklärung: In der Tat folgt aus (3.38) $a \leq x_1 = g(x_0) \leq b$ da $x_0 \in [a;b]$ ist. Für $x_2 = g(x_1)$ folgt analog $a \leq x_2 = g(x_1) \leq b$, da $x_1 \in [a;b]$ ist. Setzt man den Prozess fort, so sieht man, dass alle $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, in $[a;b]$ liegen.)

Um die Konvergenz zu zeigen, betrachten wir den absoluten Fehler von x_{n+1}

$$\text{Fehler}(x_{n+1}) = z - x_{n+1} = g(z) - g(x_n),$$

wobei wir im zweiten Schritt $z = g(z)$ (da z der Fixpunkt von g ist) und $x_{n+1} = g(x_n)$ genutzt haben. Wir wenden den Absolutbetrag an und nutzen anschließend die Kontraktionseigenschaft (3.43) mit $u = z$ und $w = x_n$ aus:

$$|z - x_{n+1}| = |g(z) - g(x_n)| \leq \lambda\, |z - x_n|. \tag{3.44}$$

Wir wiederholen den Prozess für den absoluten Fehler von x_n und erhalten

$$|z - x_n| = |g(z) - g(x_{n-1})| \leq \lambda\, |z - x_{n-1}|. \tag{3.45}$$

Analog erhalten wir

$$\begin{aligned} |z - x_{n-1}| &= |g(z) - g(x_{n-2})| \leq \lambda\, |z - x_{n-2}|, \\ &\vdots \\ |z - x_2| &= |g(z) - g(x_1)| \leq \lambda\, |z - x_1|, \\ |z - x_1| &= |g(z) - g(x_0)| \leq \lambda\, |z - x_0|. \end{aligned} \tag{3.46}$$

Nacheinander Anwenden der Ungleichungen aus (3.44), (3.45) und (3.46) liefert nun

$$|z - x_{n+1}| \leq \lambda\, |z - x_n| \leq \lambda^2\, |z - x_{n-1}| \leq \ldots \leq \lambda^{n+1}\, |z - x_0|. \tag{3.47}$$

Ersetzen wir in (3.47) $n + 1$ durch n so erhalten wir

$$|z - x_n| \leq \lambda^n\, |z - x_0|, \qquad n = 0, 1, 2, \ldots. \tag{3.48}$$

Da $0 < \lambda < 1$ ist, gilt $\lim_{n\to\infty} \lambda^n = 0$ und somit strebt die rechte Seite in (3.48) für $n \to \infty$ gegen 0. Daraus folgt aber:

$$\lim_{n\to\infty} |z - x_n| = 0 \qquad \Longleftrightarrow \qquad \lim_{n\to\infty} x_n = z.$$

Also konvergiert die Fixpunktiteration $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, für jeden Startwert $x_0 \in [a;b]$ gegen den einzigen Fixpunkt z.

(3) Mit der Dreiecksungleichung gilt

$$|z - x_0| = \big|(z - x_1) + (x_1 - x_0)\big| \leq |z - x_1| + |x_1 - x_0|.$$

Wir wenden nun für $|z - x_1|$ die letzte Abschätzung aus (3.46) an und erhalten

$$|z - x_0| \leq |z - x_1| + |x_1 - x_0| \leq \lambda\,|z - x_0| + |x_1 - x_0|.$$

Umsortieren liefert:

$$\begin{aligned} & |z - x_0| \leq \lambda\,|z - x_0| + |x_1 - x_0| \quad \Big| \; -\lambda\,|z - x_0| \\ \iff \qquad & (1 - \lambda)\,|z - x_0| \leq |x_1 - x_0| \quad \Big| \; :(1 - \lambda) \\ \iff \qquad & |z - x_0| \leq \frac{1}{1 - \lambda}\,|x_1 - x_0| \end{aligned} \tag{3.49}$$

Einsetzen von (3.49) in (3.48) liefert dann (3.40):

$$|z - x_n| \leq \lambda^n\,|z - x_0| \leq \frac{\lambda^n}{1 - \lambda}\,|x_1 - x_0|, \qquad n = 1, 2, \ldots.$$

(4) Wegen $z = g(z)$ und $x_{n+1} = g(x_n)$ gilt

$$\frac{z - x_{n+1}}{z - x_n} = \frac{g(z) - g(x_n)}{z - x_n} = \frac{g(x_n) - g(z)}{x_n - z}\,.$$

Da die Folge x_n gegen den Fixpunkt z konvergiert und da g stetig differenzierbar ist, folgt aus der Definition der Ableitung als Grenzwert des Differenzenquotienten

$$\lim_{n\to\infty} \frac{z - x_{n+1}}{z - x_n} = \lim_{n\to\infty} \frac{g(x_n) - g(z)}{x_n - z} = g'(z).$$

Das beweist (3.41). – Liegt x_n dicht genug bei z, so gilt also angenähert:

$$\frac{z - x_{n+1}}{z - x_n} \approx g'(z) \qquad \iff \qquad z - x_{n+1} \approx g'(z)\,(z - x_n)$$

Damit ist der Beweis abgeschlossen. □

Da die Bedingung (3.38) in der Praxis nicht immer leicht zu überprüfen ist bzw. da es nicht immer einfach ist, ein passendes Intervall $[a; b]$ mit der Eigenschaft (3.38) anzugeben, benötigen wir noch den folgenden Satz.

Satz 3.21. (Kriterium für Konvergenz der Fixpunktiteration)

Sei $g :]c; d[\to \mathbb{R}$ stetig differenzierbar, und g habe einen Fixpunkt z im Intervall $]c; d[$. Dann gelten die folgenden Aussagen:

(1) Wenn $|g'(z)| < 1$ ist, dann gibt es ein Intervall $[a; b] \subseteq]c; d[$ mit $z \in [a; b]$, für welches die Voraussetzungen (3.38) und (3.39) gelten und somit auch alle Schlussfolgerungen aus Satz 3.20 erfüllt sind.

(2) Wenn $|g'(z)| > 1$ ist, dann konvergiert die Fixpunktiteration $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, nicht gegen den Fixpunkt z oder sie erreicht den Fixpunkt bereits nach endlich vielen Schritten.

(3) Ist $|g'(z)| = 1$, so können wir keine Aussage treffen. (Falls die Fixpunktiteration den Fixpunkt nicht in endlich vielen Schritten erreicht und in diesem Fall konvergieren sollte, wird die Konvergenz so langsam sein, dass das Verfahren nicht praktisch anwendbar ist.)

Mit Hilfe von Satz 3.20 und Satz 3.21 können wir nun auch das Verhalten der Fixpunktiterationen in Beispiel 3.18 erklären.

Beispiel 3.22. (Konvergenz bzw. Divergenz der Fixpunktiterationen)
In Beispiel 3.18 wurden die folgenden Fixpunktiterationen betrachtet:

$$x_{n+1} = 5 + x_n - x_n^2 \qquad \text{für} \quad g_1(x) = 5 + x - x^2,$$

$$x_{n+1} = \frac{5}{x_n} \qquad \text{für} \quad g_2(x) = \frac{5}{x},$$

$$x_{n+1} = 1 + x_n - \frac{x_n^2}{5} \qquad \text{für} \quad g_3(x) = 1 + x - \frac{x^2}{5},$$

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{5}{x_n}\right) \qquad \text{für} \quad g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right).$$

Die Funktionen g_1, g_2, g_3 und g_4 haben dabei alle die beiden Fixpunkte $z = \sqrt{5}$ und $\widetilde{z} = -\sqrt{5}$. In Beispiel 3.18 schienen diese Fixpunktiterationen mit dem Startwert $x_0 = 2{,}5$ für die Funktionen g_1 und g_2 zu divergieren. Dagegen schienen die Fixpunktiterationen mit dem Startwert $x_0 = 2{,}5$ für die Funktionen g_3 und g_4 gegen $z = \sqrt{5}$ zu konvergieren. Wir können dieses Verhalten nun mit Hilfe von Satz 3.21 erklären:

(a) Für $g_1(x) = 5 + x - x^2$ erhalten wir $g_1'(x) = 1 - 2\,x$. Also gilt

$$g_1'(z) = g_1'(\sqrt{5}) = 1 - 2\sqrt{5} \doteq -3{,}4721 \quad \Longrightarrow \quad |g_1'(z)| \doteq 3{,}4721 > 1,$$

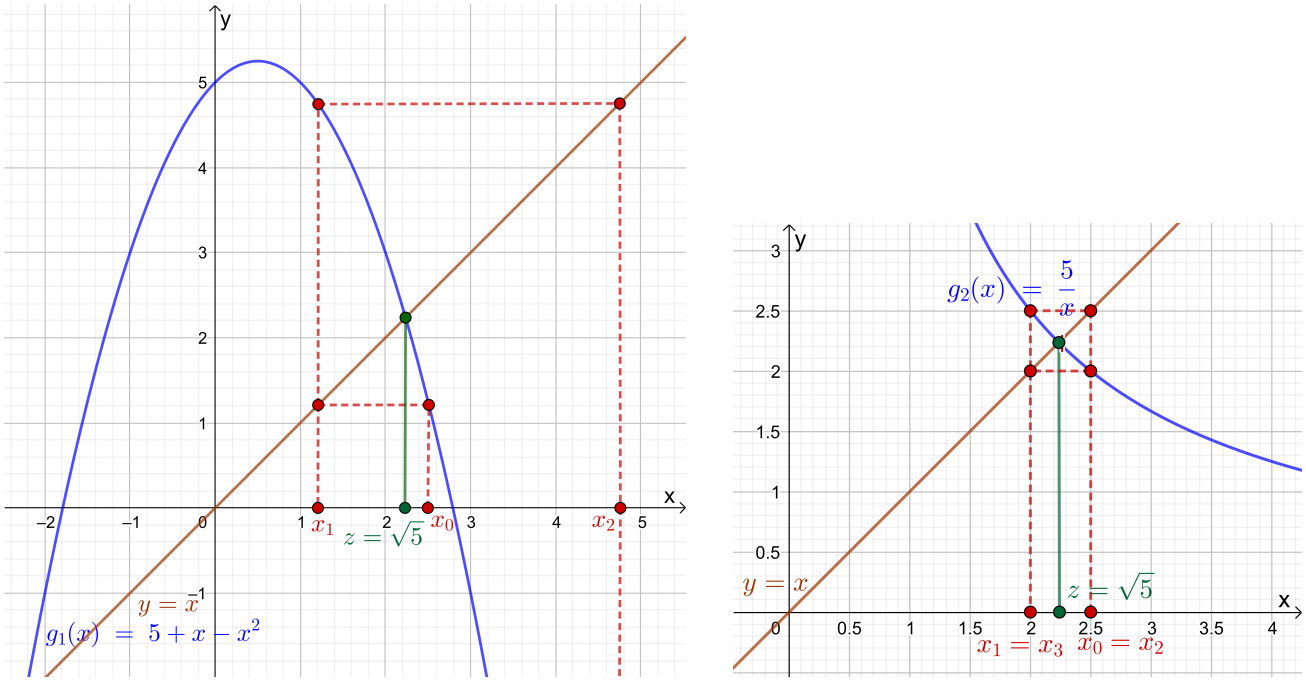


Abb. 3.7: Die Fixpunktiterationen für die Funktionen $g_1(x) = 5 + x - x^2$ und $g_2(x) = \frac{5}{x}$ mit ihrem Fixpunkt $z = \sqrt{5}$ (in grün): Man sieht im linken Bild, wie die Iterierten immer weiter vom Fixpunkt weglaufen, und im rechten Bild sieht man gut, wie die Iterierten zwischen den Werten 2 und 2,5 hin- und her pendeln.

und nach Satz 3.21 (2) wird die Fixpunktiteration für keinen Startwert gegen den Fixpunkt $z = \sqrt{5}$ konvergieren.

(b) Für $g_2(x) = \dfrac{5}{x} = 5\,x^{-1}$ erhalten wir $g_2'(x) = -5\,x^{-2} = \dfrac{-5}{x^2}$. Also gilt

$$g_2'(z) = g_2'(\sqrt{5}) = \frac{-5}{(\sqrt{5})^2} = -1 \quad \Longrightarrow \quad |g_2'(z)| = |-1| = 1,$$

und nach Satz 3.21 (3) können wir keine Aussage treffen. Wir beobachten aber, dass die Iterierten nicht konvergieren, sondern abwechselnd die Werte 2,5 und 2,0 annehmen. (Man kann sich leicht überlegen, dass sich dieser Prozess fortsetzt, wenn wir x_n für $n = 7, 8, \ldots,$ berechnen.)

(c) Für $g_3(x) = 1 + x - \dfrac{x^2}{5}$ erhalten wir $g_3'(x) = 1 - \dfrac{2\,x}{5}$. Also gilt

$$g_3'(z) = g_3'(\sqrt{5}) = 1 - \frac{2\sqrt{5}}{5} \doteq 0{,}1056 \quad \Longrightarrow \quad |g_3'(z)| \doteq 0{,}1056 < 1,$$

und nach Satz 3.21 (1) sollte die Fixpunktiteration für einen Startwert, der dicht genug bei $z = \sqrt{5}$ liegt, konvergieren. Dieses ist für den Startwert $x_0 = 2{,}5$ offenbar der Fall.

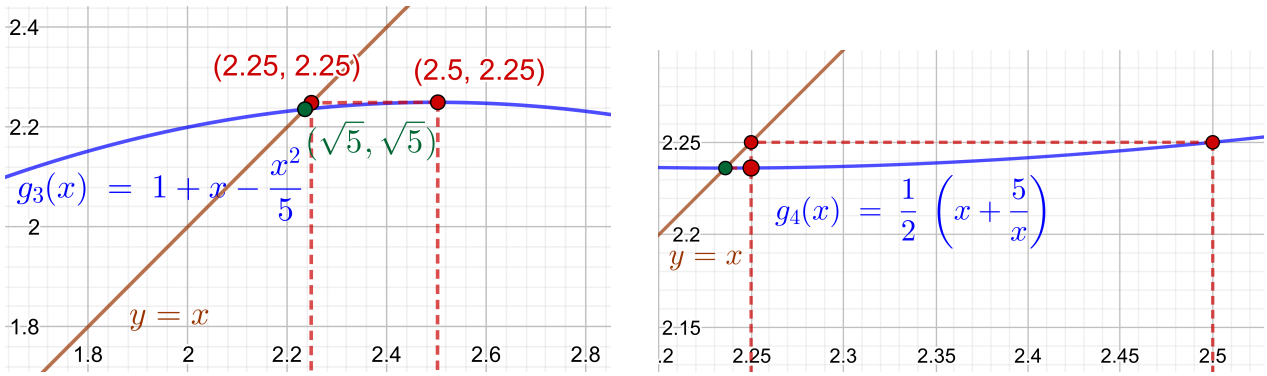


Abb. 3.8: Die Fixpunktiterationen für die Funktionen $g_3(x) = 1 + x - \frac{x^2}{5}$ und $g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right)$ mit ihrem Fixpunkt $z = \sqrt{5}$ (in grün): Wegen des guten Startwerts $x_0 = 2{,}5$ und der schnellen Konvergenz lässt sich hier nur ein Iterationsschritt sichtbar machen und wir mussten schon dafür stark hineinzoomen.

(d) Für $g_4(x) = \dfrac{1}{2}\left(x + \dfrac{5}{x}\right) = \dfrac{1}{2}\,x + \dfrac{5}{2}\,x^{-1}$ erhalten wir die Ableitung

$$g_4'(x) = \frac{1}{2} - \frac{5}{2}\,x^{-2} = \frac{1}{2} - \frac{5}{2}\,\frac{1}{x^2}\,.$$

Also gilt

$$g_4'(z) = g_4'(\sqrt{5}) = \frac{1}{2} - \frac{5}{2}\,\frac{1}{(\sqrt{5})^2} = 0 \quad \Longrightarrow \quad |g_4'(z)| = 0 < 1,$$

und nach Satz 3.21 (1) sollte die Fixpunktiteration für einen Startwert, der dicht genug bei $z = \sqrt{5}$ liegt, konvergieren. Dieses ist für den Startwert $x_0 = 2{,}5$ offenbar auch der Fall.

Wir können also in drei der vier Beispiele die Konvergenz bzw. Divergenz der Fixpunktiteration mit Hilfe von Satz 3.21 erklären.

Die Konvergenz bzw. Divergenz ist in Abbildungen 3.7 und 3.8 jeweils grafisch mit einem „Spinnwebdiagramm“ (Erklärung s.u.) veranschaulicht. ♠

Beispiel 3.23. („Spinnwebdiagramm“ der Fixpunktiteration)
Um die Konvergenz einer Fixpunktiteration besser grafisch zu veranschaulichen, als dieses in Abbildung 3.8 möglich ist, betrachten wir die Fixpunktgleichung

$$x = \sqrt{x}.$$

Diese hat als einzige positive Lösung $z = 1$ (und weiter die Lösung $\widetilde{z} = 0$). Um den

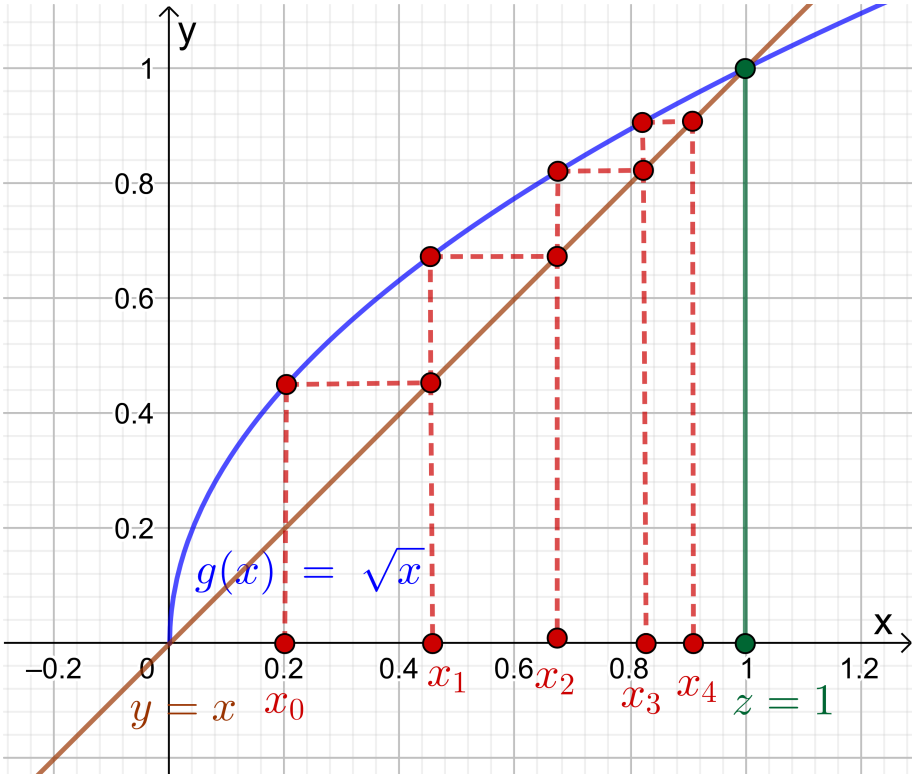


Abb. 3.9: Indem wir den Schnittpunkt der horizontalen Geraden durch $\big(x_0; f(x_0)\big)$ mit mit $x_0 = 0{,}2$ mit der Winkelhalbierenden $y = x$ bestimmen und von dort ein Lot auf die x-Achse fällen, erhalten wir $x_1 = f(x_0)$ auf der x-Achse abgetragen. Nun wiederholen wir diesen Prozess mit x_1 anstelle von x_0, um x_2 zu bestimmen. Die Iterierten x_n, $n = 0, 1, 2, 3, 4$, sind in der Grafik eingezeichnet. Die Konvergenz gegen den Fixpunkt $z = 1$ ist gut sichtbar. Man erhält ein „Spinnwebdiagramm“.

Fixpunkt mit der Fixpunktiteration zu berechnen, verwenden wir die Funktion

$$g : [0; \infty[\to \mathbb{R}, \qquad g(x) = \sqrt{x},$$

und erhalten die Fixpunktiteration

$$x_{n+1} = g(x_n) = \sqrt{x_n}, \qquad n = 0, 1, 2, \ldots,$$

die in Abbildung 3.9 für den Startwert $x_0 = 0{,}2$ mit einem „Spinnwebdiagramm“ veranschaulicht ist. ♠

Bemerkung 3.24. (Praktische Anwendung der Fixpunktiteration)
Satz 3.20 wird selten direkt angewendet, da es schwierig ist, ein Intervall $[a; b]$ mit (3.38) zu finden. **Statt dessen nutzt man in der Regel Satz 3.21 wie folgt:** Man schreibt die zu lösende Gleichung $f(x) = 0$ so in eine Fixpunktgleichung $x = g(x)$ um, dass für alle x in der Nähe des Fixpunkts

gilt $|g'(x)| \leq \lambda < 1$ mit einer positiven Konstante $\lambda < 1$. Wenn man nun die Fixpunktiteration mit einem hinreichend guten Näherungswert x_0 für den Fixpunkt startet, kann man erwarten, dass die Fixpunktiteration gegen den Fixpunkt z konvergiert.

Zum Abschluss des Teilkapitels erklären wir noch den Beweis von Satz 3.21.

Beweis von Satz 3.21:

(1) Da g' stetig ist folgt aus $|g'(z)| < 1$, dass es ein kleines Intervall $[z-\varepsilon; z+\varepsilon]$ mit $\varepsilon > 0$ um den Fixpunkt z gibt mit $|g'(x)| \leq \lambda < 1$ für alle $x \in [z-\varepsilon; z+\varepsilon]$ mit einer positiven Konstante $\lambda < 1$. Damit ist die Bedingung (3.39) aus Satz 3.20 erfüllt.

Nach dem Mittelwertsatz (siehe Satz 1.4) gibt es zu allen $u, w \in [z-\varepsilon; z+\varepsilon]$ mit $u \neq w$ ein y zwischen u und w mit

$$\frac{g(u)-g(w)}{u-w} = g'(y) \qquad \Longleftrightarrow \qquad g(u)-g(w) = g'(y)\,(u-w)$$

$$\Longrightarrow \qquad |g(u)-g(w)| = |g'(y)|\,|u-w| \leq \lambda\,|u-w|, \qquad (3.50)$$

wobei wir genutzt haben, dass y ebenfalls in $[z-\varepsilon; z+\varepsilon]$ liegt und dass damit $|g'(y)| \leq \lambda$ gilt. Die Ungleichung auf der rechten Seite von (3.50) ist auch für $u = w$ wahr (dann sind beide Seiten der Ungleichung 0). Also gilt

$$|g(u)-g(w)| \leq \lambda\,|u-w| \qquad \text{für alle } u, w \in [z-\varepsilon; z+\varepsilon]. \qquad (3.51)$$

Insbesondere folgt aus (3.51) für $u = z$

$$|z-g(w)| = |g(z)-g(w)| \leq \lambda\,|z-w| \leq \lambda\,\varepsilon \qquad \text{für alle } w \in [z-\varepsilon; z+\varepsilon],$$

wobei wir genutzt haben, dass $g(z) = z$ (da z ein Fixpunkt ist) und dass $|z-w| \leq \varepsilon$ (da $w \in [z-\varepsilon; z+\varepsilon]$) gelten. Also finden wir

$$|z-g(w)| \leq \lambda\,|z-w| \leq \lambda\,\varepsilon \overset{\lambda<1}{\underset{\downarrow}{<}} \varepsilon \qquad \text{für alle } w \in [z-\varepsilon; z+\varepsilon]$$

$$\Longrightarrow \qquad g(w) \in [z-\varepsilon; z+\varepsilon] \qquad \text{für alle } w \in [z-\varepsilon; z+\varepsilon],$$

d.h. die Funktionswerte von g liegen für w aus $[a; b] = [z-\varepsilon; z+\varepsilon]$ wieder in $[a; b] = [z-\varepsilon; z+\varepsilon]$. Damit ist die Bedingung (3.38) aus Satz 3.20 erfüllt.

Da alle Voraussetzungen aus Satz 3.20 erfüllt sind, können wie Satz 3.20 anwenden. Dieser liefert uns insbesondere, dass die Fixpunktiteration $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, für jeden Startwert x_0 aus dem Intervall $[a; b] = [z-\varepsilon; z+\varepsilon]$ konvergiert.

(2) Da g' stetig ist folgt aus $|g'(z)| > 1$, dass es ein kleines Intervall $[z-\varepsilon; z+\varepsilon]$ um den Fixpunkt z gibt mit $|g'(x)| \geq \lambda$ für alle $x \in [z-\varepsilon; z+\varepsilon]$ mit einer positiven Konstante λ mit $|g'(z)| \geq \lambda > 1$.

Nach dem Mittelwertsatz (siehe Satz 1.4) gibt es zu allen $u, w \in [z-\varepsilon; z+\varepsilon]$ mit $u \neq w$ ein y zwischen u und w mit

$$\frac{g(u)-g(w)}{u-w} = g'(y) \qquad \Longleftrightarrow \qquad g(u)-g(w) = g'(y)\,(u-w)$$

$$\Longrightarrow \qquad |g(u)-g(w)| = |g'(y)|\,|u-w| \geq \lambda\,|u-w|,$$

wobei wir genutzt haben, dass y ebenfalls in $[z-\varepsilon; z+\varepsilon]$ liegt und damit $|g'(y)| \geq \lambda$ gilt. Für $u = w$ gilt $|g(u)-g(w)| = 0 \geq \lambda\,0 = \lambda\,|u-w|$. Also gilt

$$|g(u)-g(w)| \geq \lambda\,|u-w| \qquad \text{für alle } u, w \in [z-\varepsilon; z+\varepsilon]. \tag{3.52}$$

Insbesondere folgt aus (3.52) für $u = z$ und $w = x_n \in [z-\varepsilon; z+\varepsilon]$ mit $x_n \neq z$

$$|z-x_{n+1}| = |g(z)-g(x_n)| \geq \lambda\,|z-x_n| \overset{\lambda>1}{>} |z-x_n|, \tag{3.53}$$

wobei wir genutzt haben, dass $g(z) = z$ gilt (da z ein Fixpunkt ist), dass $g(x_n) = x_{n+1}$ ist und dass $\lambda > 1$ ist. Aus (3.53) folgt, dass die Iterierte x_{n+1} einen größeren Abstand zu z hat als x_n. Bei Konvergenz von x_n (und damit von x_{n+1}) gegen z müsste $|z-x_{n+1}|$ gegen 0 streben. Also kann die Fixpunktiteration mit einem Startwert $x_0 \in [z-\varepsilon; z+\varepsilon] \setminus \{z\}$ nicht gegen z konvergieren, solange die Iterierten x_n in $[z-\varepsilon; z+\varepsilon]$ liegen. – Gilt $x_0 = z$, so liegt der Fixpunkt bereits vor. – Liegt der Startwert x_0 nicht in $[z-\varepsilon; z+\varepsilon]$ oder tritt eine Iterierte x_n auf, die nicht in $[z-\varepsilon; z+\varepsilon]$ liegt, so kann die Fixpunktiteration nur konvergieren, wenn für x_0 bzw. eine Iterierte $x_n \notin [z-\varepsilon; z+\varepsilon]$ gilt $x_1 = g(x_0) = z$ bzw. $x_{n+1} = g(x_n) = z$. (Dann wird der Fixpunkt nach endlich vielen Schritten erreicht.) Andernfalls müsste bei Konvergenz der Fixpunktiteration $x_{n+1} = g(x_n)$ gegen z irgendwann für ein n gelten $x_n \in [z-\varepsilon; z+\varepsilon] \setminus \{z\}$, und dann können wir wie in (3.53) argumentieren, d.h. es kann keine Konvergenz vorliegen. (*Beispiel zum Fall der Konvergenz in endlich vielen Schritten:* Die Funktion $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$, hat in $z_1 = 0$ und in $z_2 = 1$ jeweils einen Fixpunkt. Im Fixpunkt $z_2 = 1$ gilt $|g'(z_2)| = |2\,z_2| = |2 \cdot 1| = 2 > 1$. Auf dem Intervall $\left[\frac{2}{3}; \frac{4}{3}\right]$ gilt dann $|g'(x)| = |2\,x| \geq 2 \cdot \frac{2}{3} = \frac{4}{3} > 1$. Trotzdem gilt für den Startwert $x_0 = -1$, der nicht in $\left[\frac{2}{3}; \frac{4}{3}\right]$ liegt, dass $x_1 = g(-1) = (-1)^2 = 1 = z_2$ ist, d.h. die Fixpunktiteration erreicht den Fixpunkt $z_2 = 1$ in einem Schritt.)

(3) Dieses kann man zeigen, indem man geeignete Beispiele von Funktionen g mit Ableitung $|g'(z)| = 1$ untersucht. □

3.6 Aitkens Fehlerabschätzung und Extrapolationsformel*

Dieses Teilkapitel wird im Sommersemester 2024 nicht behandelt und ist damit auch nicht klausurrelevant.

Wir wollen nun untersuchen, welche Informationen wir über den Fehler der Fixpunktiteration aus dem Fixpunktsatz (siehe Satz 3.20) gewinnen können. Zunächst liefert uns (3.40) die folgende Abschätzung für den absoluten Fehler:

$$|z - x_n| \leq \frac{\lambda^n}{1-\lambda} |x_1 - x_0| \qquad \text{für alle } n = 1, 2, \ldots.$$

Auf der rechten Seite erscheint aber die Konstante λ, welche durch (3.39) als

$$\lambda = \max_{a \leq x \leq b} |g'(x)| < 1$$

definiert ist. Da in der Praxis aber oft das Intervall $[a; b]$ nicht bekannt ist (vgl. Bemerkung 3.24), kann man λ auch nicht exakt bestimmen. Wir brauchen also eine passende und berechenbare Näherung für dem für den absoluten Fehler.

Wir beginnen mit dem Grenzwert (siehe Formel (3.41) in Satz 3.20)

$$\lim_{n \to \infty} \frac{z - x_n}{z - x_{n-1}} = g'(z).$$

Dieser bedeutet, dass bei einer konvergenten Fixpunktiteration die Folge der Quotienten

$$\boxed{r_n = \frac{z - x_n}{z - x_{n-1}}, \qquad n = 1, 2, \ldots} \tag{3.54}$$

gegen $\lambda_z = g'(z)$ (also gegen die Ableitung im Fixpunkt z) strebt. Da wir den Fixpunkt z aber in der Regel auch nicht kennen, brauchen wir eine Näherung für r_n. Eine Näherung für r_n ist durch

$$\boxed{\lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}, \qquad n = 1, 2, \ldots} \tag{3.55}$$

gegeben, denn wegen $x_n = g(x_{n-1})$ und $x_{n-1} = g(x_{n-2})$ gilt

$$\lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} = \frac{g(x_{n-1}) - g(x_{n-2})}{x_{n-1} - x_{n-2}},$$

*Dieses Teilkapitel wird im Sommersemester 2024 nicht behandelt und ist damit auch nicht klausurrelevant.

und nach dem Mittelwertsatz (siehe Satz 1.4) gibt es ein c_n zwischen x_{n-1} und x_{n-2}, so dass gilt

$$\lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} = \frac{g(x_{n-1}) - g(x_{n-2})}{x_{n-1} - x_{n-2}} = g'(c_n).$$

Wenn die Iterierten x_n der Fixpunktiteration gegen den Fixpunkt z streben, dann muss auch c_n gegen z streben. Damit strebt dann $g'(c_n)$ gegen $\lambda_z = g'(z)$. Also folgt für eine konvergente Fixpunktiteration

$$\boxed{\lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} \to g'(z) = \lambda_z \qquad \text{für } n \to \infty.}$$

Als Nächstes nutzen wir die Näherung (3.42) aus Satz 3.20 für x_n dicht bei z

$$z - x_n \approx g'(z)\,(z - x_{n-1}) = \lambda_z\,(z - x_{n-1}). \tag{3.56}$$

Da wir den Fixpunkt in der Regel nicht kennen und die Formel so nicht direkt zur Näherung des absoluten Fehlers nutzen können, lösen wir (3.56) nach z auf:

$$z - x_n \approx \lambda_z\,(z - x_{n-1}) \qquad \Longleftrightarrow \qquad z - x_n \approx \lambda_z\,z - \lambda_z\,x_{n-1} \;\Big|\; -\lambda_z\,z + x_n$$

$$\Longleftrightarrow \qquad z - \lambda_z\,z \approx x_n - \lambda_z\,x_{n-1} \qquad \Longleftrightarrow \qquad (1 - \lambda_z)\,z \approx x_n - \lambda_z\,x_{n-1}$$

$$\Longleftrightarrow \qquad (1 - \lambda_z)\,z \approx \big((1 - \lambda_z) + \lambda_z\big)\,x_n - \lambda_z\,x_{n-1}$$

$$\Longleftrightarrow \qquad (1 - \lambda_z)\,z \approx (1 - \lambda_z)\,x_n + \lambda_z\,(x_n - x_{n-1}) \;\Big|\; :(1 - \lambda_z)$$

$$\Longleftrightarrow \qquad \boxed{z \approx x_n + \frac{\lambda_z}{1 - \lambda_z}\,(x_n - x_{n-1}) \quad \Longleftrightarrow \quad z - x_n \approx \frac{\lambda_z}{1 - \lambda_z}\,(x_n - x_{n-1})} \tag{3.57}$$

Da wir $\lambda_z = g'(z)$ in der Regel nicht kennen, nutzen wir in (3.57) die Näherung λ_n aus (3.55) für $\lambda_z = g'(z)$ und erhalten

$$\boxed{z \approx x_n + \frac{\lambda_n}{1 - \lambda_n}\,(x_n - x_{n-1})} \tag{3.58}$$

$$\boxed{z - x_n \approx \frac{\lambda_n}{1 - \lambda_n}\,(x_n - x_{n-1})} \tag{3.59}$$

Die Formel (3.58) heißt **Aitkens Extrapolationsformel**, und sie ermöglicht es aus x_n, x_{n-1} und λ_n eine verbesserte Näherung für z zu berechnen.

n	x_n	$z - x_n$	r_n	$x_n - x_{n-1}$	λ_n
0	2,50000000	$-2{,}64 \cdot 10^{-1}$			
1	2,25000000	$-1{,}39 \cdot 10^{-2}$	0,0528	$-2{,}50 \cdot 10^{-1}$	
2	2,23750000	$-1{,}43 \cdot 10^{-3}$	0,1028	$-1{,}25 \cdot 10^{-2}$	0,0500
3	2,23621875	$-1{,}51 \cdot 10^{-4}$	0,1053	$-1{,}28 \cdot 10^{-3}$	0,1025
4	2,23608389	$-1{,}59 \cdot 10^{-5}$	0,1055	$-1{,}35 \cdot 10^{-4}$	0,1053
5	2,23606966	$-1{,}68 \cdot 10^{-6}$	0,1056	$-1{,}42 \cdot 10^{-5}$	0,1055
6	2,23606815	$-1{,}77 \cdot 10^{-7}$	0,1056	$-1{,}50 \cdot 10^{-6}$	0,1056
7	2,23606800	$-1{,}87 \cdot 10^{-8}$	0,1056	$-1{,}59 \cdot 10^{-7}$	0,1056

n	x_n	Aitkens Fehlerabschätzung	Aitkens Extrapolationsformel	Fehler von Aitkens Extrapolationsformel
0	2,50000000			
1	2,25000000			
2	2,23750000	$-6{,}58 \cdot 10^{-4}$	2,23684211	$-7{,}74 \cdot 10^{-4}$
3	2,23621875	$-1{,}46 \cdot 10^{-4}$	2,23607242	$-4{,}45 \cdot 10^{-6}$
4	2,23608389	$-1{,}59 \cdot 10^{-5}$	2,23606803	$-4{,}83 \cdot 10^{-8}$
5	2,23606966	$-1{,}68 \cdot 10^{-6}$	2,23606798	$-5{,}36 \cdot 10^{-10}$
6	2,23606815	$-1{,}77 \cdot 10^{-7}$	2,23606798	$-5{,}98 \cdot 10^{-12}$
7	2,23606800	$-1{,}87 \cdot 10^{-8}$	2,23606798	$-6{,}66 \cdot 10^{-14}$

Tabelle 3.5: Aitkens Fehlerabschätzung und Aitkens Extrapolationsformel für die Fixpunktiteration für $g(x) = 1 + x - \frac{x^2}{5}$ mit dem Startwert $x_0 = 2$

Die Formel (3.59) gibt eine Näherung für den absoluten Fehler an und wird **Aitkens Fehlerabschätzung** genannt.

Betrachten wir dazu ein Beispiel.

Beispiel 3.25. (Aitkens Fehlerabschätzung und Extrapolationsformel)
In Beispielen 3.18 und 3.22 wurde die Fixpunktiteration

$$x_{n+1} = 1 + x_n - \frac{x_n^2}{5} \qquad \text{für} \quad g(x) = 1 + x - \frac{x^2}{5}$$

betrachtet, die für den Startwert $x_0 = 2{,}5$ gegen den Fixpunkt $z = \sqrt{5}$ konver-

giert. Für diese gilt mit

$$g'(x) = 1 - \frac{2\,x}{5} \quad \Longrightarrow \quad \lambda_z = g'(z) = g'(\sqrt{5}) = 1 - \frac{2\,\sqrt{5}}{5} = 1 - \frac{2}{\sqrt{5}} \doteq 0{,}1056.$$

In der oberen Tabelle in Tabelle 3.5 wurden mit dem Startwert $x_0 = 2{,}5$ für $n = 0, 1, \ldots, 7$ neben x_n auch der absolute Fehler $z - x_n$, $x_n - x_{n-1}$ und die Näherungen $r_n = \frac{z-x_n}{z-x_{n-1}}$ und $\lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}$ für $g'(z) = g'(\sqrt{5}) \doteq 0{,}1056$ angegeben. Dabei wurden x_n und die mit Aitkens Extrapolationsformel berechneten verbesserten Näherungen auf eine 9-stellige Gleitkommadarstellung gerundet angegeben; und r_n und λ_n wurden auf eine 4-stellige Gleitkommadarstellung gerundet angegeben. Die verschiedenen absoluten Fehler bzw. Abschätzungen derselben, sowie $x_n - x_{n-1}$, wurden alle auf eine 3-stellige Gleitkommadarstellung gerundet angegeben.

Insbesondere sehen wir in der oberen Tabelle in Tabelle 3.5, dass r_n bereits für $n = 5$ die Ableitung $g'(z)$ bei Rundung auf eine 4-stellige Gleitkommadarstellung „exakt“ berechnet. Weiter beobachten wir, dass λ_n ab $n = 4$ eine ähnlich gute Näherung für $g'(z) = g'(\sqrt{5}) \doteq 0{,}1056$ liefert wie r_n.

In der unteren Tabelle in Tabelle 3.5 in sind zusätzlich Aitkens Fehlerabschätzung, Aitkens Extrapolationsformel und deren absoluter Fehler angegeben:

Vergleichen wir den absoluten Fehler $z - x_n$ in der oberen Tabelle in Tabelle 3.5 mit Aitkens Fehlerabschätzung, so sehen wir, dass Aitkens Fehlerabschätzung bereits ab $n = 3$ eine gute Näherung für den absoluten Fehler bildet.

Betrachten wir nun die mit Aitkens Extrapolationsformel berechneten Näherungen für den Fixpunkt, so sehen wir, dass diese eine deutliche Verbesserung liefern. Beispielsweise gilt $z - x_4 \doteq -1{,}59 \cdot 10^{-5}$, und die mit x_4, x_3 und λ_4 mittels Aitkens Extrapolationsformel (3.58) berechnete Näherung hat nur einen absoluten Fehler von $-4{,}83 \cdot 10^{-8}$. Bei der Fixpunktiteration wird ein betraglich höchstens genauso großer absoluter Fehler erst mit x_7 erreicht. ♠

Wir halten die vor dem Beispiel hergeleiteten Näherungen als Bemerkung fest.

Bemerkung 3.26. (Fehlerabschätzung und Extrapolationsformel)
Sei $g :]c; d[\to \mathbb{R}$ eine **stetig differenzierbare** Funktion, $[a; b] \subseteq]c; d[$, und g habe einem Fixpunkt $z \in [a; b]$ (also $g(z) = z$). Es sei $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, eine Fixpunktiteration mit einem Startwert $x_0 \in [a; b]$, die gegen den Fixpunkt z konvergiert. Dann gelten:

(1) $\lambda_n = \dfrac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}$ strebt für $n \to \infty$ gegen $\lambda_z = g'(z)$ und liefert somit

(meist schon für relativ kleine n) gute Näherungen für $\lambda_z = g'(z)$.

(2) **Aitkens Fehlerabschätzung** liefert eine Näherung für den absoluten Fehler $z - x_n$ der Iterierten x_n:

$$z - x_n \approx \frac{\lambda_n}{1-\lambda_n}\,(x_n - x_{n-1}) \qquad \text{mit} \quad \lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}$$

(3) Mit **Aitkens Extrapolationsformel**

$$z \approx x_n + \frac{\lambda_n}{1-\lambda_n}\,(x_n - x_{n-1}) \qquad \text{mit} \quad \lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}$$

kann man aus x_n, x_{n-1} und x_{n-2} eine (verglichen mit x_n) deutlich verbesserte Näherung für z berechnen.

3.7 Konvergenzordnung

Wir lernen nun noch die Idee der **Konvergenzordnung** eines Iterationsverfahrens kennen: Für eine konvergente Fixpunktiteration gilt nach (3.42) in Satz 3.20 für Iterierte x_n dicht bei dem Fixpunkt z

$$z - x_{n+1} \approx g'(z)\,(z - x_n) \qquad \text{mit} \qquad |g'(z)| < 1. \tag{3.60}$$

Bei einem konvergenten Newton-Verfahren gilt für x_n dicht genug bei den Nullstelle z nach Bemerkung 3.10 (3)

$$z - x_{n+1} \approx M\,(z - x_n)^2 \qquad \text{mit} \quad M = \frac{-f''(z)}{2\,f'(z)}\,. \tag{3.61}$$

Bei einem konvergenten Sekantenverfahren gilt (nach Bemerkung 3.13 (1)) für x_n dicht genug bei den Nullstelle z

$$|z - x_{n+1}| \approx |M|^{r-1}\,|z - x_n|^r \quad \text{mit} \quad M = \frac{-f''(z)}{2\,f'(z)}\,, \quad r = (\sqrt{5}+1)/2 \doteq 1{,}62. \tag{3.62}$$

Aus (3.60), (3.61) bzw. (3.62) folgt unter geeigneten Voraussetzungen an die Funktion g bzw. f jeweils, dass es eine Konstante $c > 0$ gibt, so dass

$$|z - x_{n+1}| \le c\,|z - x_n|^p \qquad \text{für alle } n = 0, 1, 2, \ldots \tag{3.63}$$

gilt mit den Werten

- $p = 1$ und einer Konstanten $c < 1$ bei der Fixpunktiteration,
- $p = 2$ bei dem Newton-Verfahren und
- $p = r = (\sqrt{5} + 1)/2 \doteq 1{,}62$ beim Sekantenverfahren.

Definition 3.27. (Konvergenzordnung)
Sei $(x_n)_{n\geq 0}$ eine Folge reeller Zahlen, die gegen z konvergiert.

(1) Wenn es eine Zahl $p > 1$ und eine positive Konstante c gibt, so dass

$$|z - x_{n+1}| \leq c\,|z - x_n|^p \qquad \text{für alle } n = 0, 1, 2, \ldots$$

*gilt, so hat die Folge $(x_n)_{n\geq 0}$ die **Konvergenzordnung** p. (Ist $p = 2$, so spricht man auch von **quadratischer Konvergenz**.)*

(2) Wenn es eine positive Konstante $c < 1$ gibt, so dass

$$|z - x_{n+1}| \leq c\,|z - x_n| \qquad \text{für alle } n = 0, 1, 2, \ldots$$

*gilt, so hat die Folge $(x_n)_{n\geq 0}$ die **Konvergenzordnung 1**. (Man spricht auch von **linearer Konvergenz**.)*

Nach unseren vorigen Überlegungen gilt also:

Bemerkung 3.28. (Konvergenzordnung der Iterationsverfahren)

(1) Die Fixpunktiteration hat die Konvergenzordnung $p = 1$.

(2) Das Newton-Verfahren hat die Konvergenzordnung $p = 2$.

(3) Das Sekantenverfahren hat die Konvergenzordnung $p = \frac{\sqrt{5}+1}{2} \doteq 1{,}62$.

Je größer die Konvergenzordnung p ist, desto schneller konvergiert ein Iterationsverfahren in der Regel. Das Newton-Verfahren hat also unter den besprochenen Verfahren die beste Konvergenzordnung.

KAPITEL 4

Interpolation und Approximation

In diesem Kapitel beschäftigen wir uns mit Polynominterpolation. Wir haben in Kapitel 1 bereits ein Beispiel der Approximation („Approximation“ bedeutet Annäherung) einer Funktion durch ein Polynom kennengelernt, nämlich die Annäherung einer $(n+1)$-mal stetig differenzierbaren Funktion durch seine Taylor-Polynome p_k mit dem Entwicklungspunkt x_0 vom Grad $k \leq n$. Diese Näherungen waren meist dicht bei dem Entwicklungspunkt selbst für einen relativ kleinen Polynomgrad k schon recht gut. Allerdings ist das Taylor-Polynom keine effiziente Näherung; in der Regel lässt sich ein Polynom kleineren Grades finden, das mindestens eine ebenso gute Näherung liefert.

In diesem Kapitel interessieren wir uns für die **Interpolation** einer Funktion durch ein Polynom passenden Grades. Wir lernen in Teilkapitel 4.1 zunächst, was Interpolation überhaupt bedeutet. Dann werden wir uns in Teilkapitel 4.2 insbesondere für Interpolation mit Polynomen interessieren und die **Interpolationsformel von Lagrange** kennenlernen. Diese liefert uns eine Darstellung des interpolierenden Polynoms, und sie ist leicht zu verstehen und ist für theoretische Überlegungen nützlich. Für praktische Anwendungen ist die Interpolationsformel von Lagrange aber ineffizient. Daher führen wir in Teilkapitel 4.3 sogenannte **dividierte Differenzen** ein und lernen darauf aufbauend die **Interpolationsformel von Newton** kennen. Diese eignet sich gut zur effizienten Berechnung von Interpolationspolynomen. In Teilkapitel 4.4 untersuchen wir schließlich, wie gut das aus $n+1$ Datenpunkten einer hinreichend oft differenzierbaren Funktion f berechnete interpolierende Polynom P_n vom Grad $\leq n$ diese Funktion f eigentlich approximiert.

Interpolation mit Polynomen und allgemeineren Funktionenklassen spielt in vie-

len Anwendungsproblemen eine wichtige Rolle. Neben der typischen Anwendung, einen Datensatz $(t_i; y_i)$, $i = 0, 1, 2, \ldots, n$, (wobei z.B. y_i die Temperatur zum Zeitpunkt t_i ist,) durch eine Funktion darzustellen, spielt Interpolation auch bei der Konstruktion numerischer Integrationsformeln eine wichtige Rolle.

4.1 Interpolation

Wie beginnen damit, dass wir zunächst erklären, was der Begriff „Interpolation“ überhaupt bedeutet.

Problemstellung 4.1. (Interpolationsproblem)
*Sei $n \in \mathbb{N}_0$. Gegeben sind $n+1$ Datenpunkte $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ in der $(x; y)$-Ebene mit **paarweise verschiedenen** $x_0, x_1, \ldots, x_n$ (d.h. $x_i \neq x_j$, wenn $i \neq j$ ist). Gesucht ist eine (geeignete) Funktion g, deren Graph durch diese Punkte geht, also welche $g(x_i) = y_i$ für $i = 0, 1, 2, \ldots, n$ erfüllt. Man sagt dann, die Funktion g **interpoliert** die Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, und nennt g die **Interpolierende** (oder **Interpolante**) dieser Datenpunkte.*

In der Regel kommen die Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, vom Graphen einer (unbekannten) Funktion f, also $y_i = f(x_i)$, $i = 0, 1, 2, \ldots, n$, oder von einem Anwendungsproblem, bei dem die y_i als Funktionswerte an der Punkten x_i aufgefasst werden können. Hier sind einige Beispiele.

Beispiel 4.2. (Datensätze für Interpolation)

(a) $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ mit $y_i =$ Temperatur zum Zeitpunkt x_i.

(b) $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ mit $y_i = \cos(x_i)$, $i = 0, 1, 2, \ldots, n$.

(c) $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ mit $x_i =$ „Körpergröße gerundet auf ganze cm“ und $y_i =$ „durchschnittliches Gewicht in der Bevölkerung der BRD bei der Körpergröße x_i“.

Überlegen Sie sich einige weitere Beispiele aus dem täglichen Leben. ♠

Betrachten wir nun ein Beispiel zur Interpolation.

Beispiel 4.3. (Interpolationsprobleme)
Gegeben seien die Datenpunkte $(x_0; y_0) = (0; 1)$ und $(x_1; y_1) = \big(\ln(2); -2\big)$.

(a) Gesucht ist ein lineares Polynom $y(x) = a\,x + b$ (mit passend zu wählenden Konstanten a und b), das die gegebenen Datenpunkte interpoliert.

Um das Interpolationsproblem zu lösen, nutzt man die Interpolationsbedingungen $y(x_k) = y_k$ für $k = 0, 1$ aus:

$$1 = y(0) = a \cdot 0 + b \qquad \Longrightarrow \qquad b = 1,$$

$$-2 = y\big(\ln(2)\big) = a \cdot \ln(2) + b \qquad \Longrightarrow \qquad a = \frac{-2-b}{\ln(2)} \overset{b=1}{=} \frac{-3}{\ln(2)}$$

Also ist das interpolierende lineare Polynom

$$y(x) = \frac{-3}{\ln(2)} \cdot x + 1.$$

(b) Gesucht ist eine Funktion der Form $y(x) = a\,e^x + b\,e^{2x}$ (mit passend zu wählenden Konstanten a und b), die die gegebenen Datenpunkte interpoliert.

Um das Interpolationsproblem zu lösen, nutzt man die Interpolationsbedingungen $y(x_k) = y_k$ für $k = 0, 1$ aus:

$$1 = y(0) = a\,e^0 + b\,e^{2\cdot 0} = a + b \qquad \Longleftrightarrow \qquad a + b = 1,$$

$$-2 = y\big(\ln(2)\big) = a\,\underbrace{e^{\ln(2)}}_{=2} + b\,e^{2\ln(2)} = a \cdot 2 + b \cdot \underbrace{(e^{\ln(2)})^2}_{=2^2=4} = 2\,a + 4\,b$$

$$\Longleftrightarrow \qquad 2\,a + 4\,b = -2 \qquad \Longleftrightarrow \qquad a + 2\,b = -1.$$

Wir erhalten also die beiden Gleichungen:

$$\begin{aligned} a + b &= 1 \qquad \text{(I)} \\ a + 2\,b &= -1 \qquad \text{(II)} \end{aligned}$$

Wir ziehen die Gleichung (I) von Gleichung (II) ab und erhalten:

$$a + 2\,b - (a + b) = -1 - 1 \qquad \Longleftrightarrow \qquad b = -2$$

Einsetzen von $b = -2$ in (I) liefert:

$$a + b = 1 \qquad \overset{b=-2}{\Longleftrightarrow} \qquad a + (-2) = 1 \qquad \Longleftrightarrow \qquad a = 3$$

Also ist die Interpolierende

$$y(x) = 3\,e^x - 2\,e^{2x}.$$

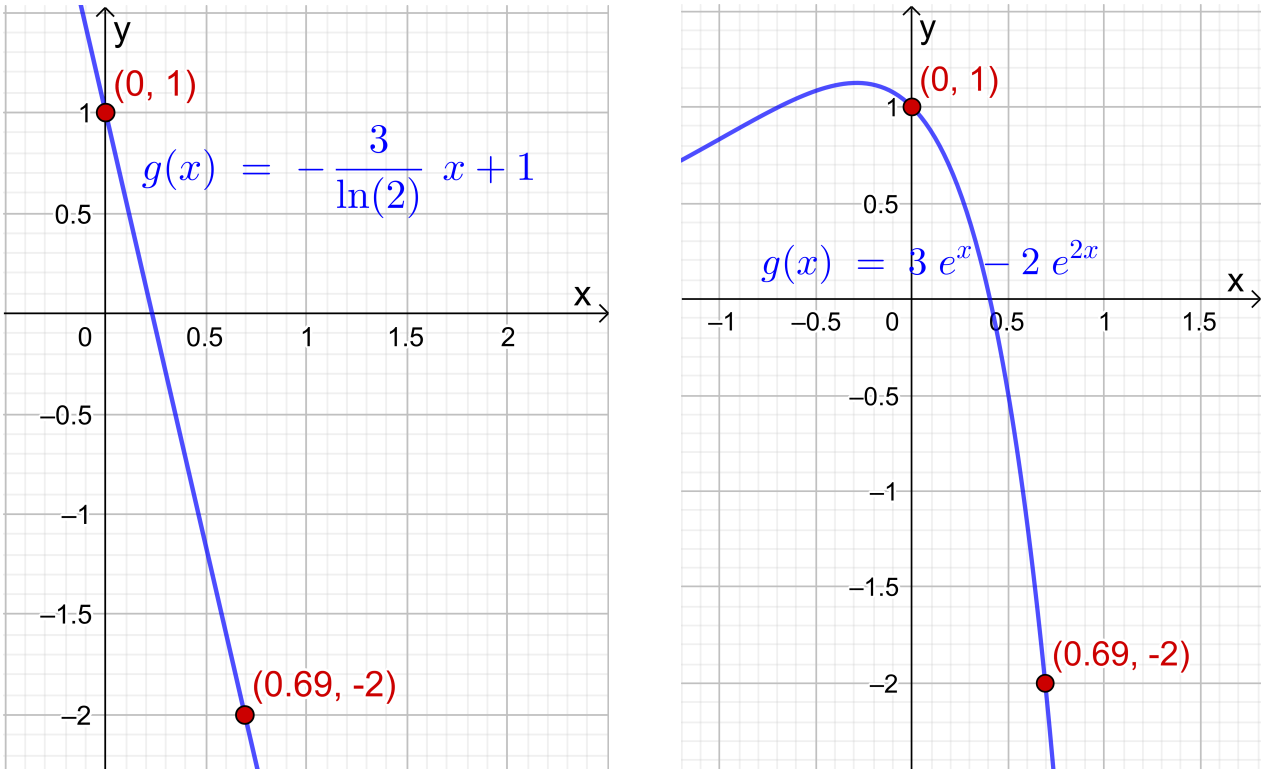


Abb. 4.1: Die Datenpunkte $(0;1)$ und $\big(\ln(2);-2\big) \doteq (0{,}69;-2)$ und die Interpolierenden aus Beispiel 4.3.

Die Graphen der Interpolierenden aus Teil (a) und (b) sind in Abbildung 4.1 gezeichnet. Natürlich sehen diese (abgesehen davon, dass beide Funktionen die Datenpunkte interpolieren) völlig unterschiedlich aus.

Wir beobachten: In jedem der beiden Interpolationsprobleme hatten wir genauso viele Konstanten wie Gleichungen. ♠

4.2 Polynominterpolation

Im Folgenden interessieren wir uns nur für Interpolation durch Polynome. Da ein Polynom vom Grad $\leq n$ (also in $\mathbb{P}_n$) von der Form

$$p(x) = a_0 + a_1\,x + a_2\,x^2 + \ldots + a_n\,x^n$$

mit $n+1$ Konstanten $a_0, a_1, a_2, \ldots, a_n \in \mathbb{R}$ ist, vermuten wir, dass man $n+1$ Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, in der $(x;y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$ (also $x_i \neq x_j$, wenn $i \neq j$) durch ein Polynom vom Grad $\leq n$ interpolieren kann. Wir werden dieses nun genauer untersuchen.

In Verallgemeinerung von Beispiel 4.3 (a) kann man zeigen, dass das **lineare Interpolationspolynom** $y(x) = a\,x + b$ für die zwei Datenpunkte $(x_0; y_0)$ und $(x_1; y_1)$ mit $x_0 \neq x_1$ von der folgenden Form ist:

$$\boxed{P_1(x) = \frac{y_1 - y_0}{x_1 - x_0}\,x + \frac{y_0\,x_1 - x_0\,y_1}{x_1 - x_0}} \tag{4.1}$$

Wir werden diese Formel auf einem Übungsblatt selber herleiten. Dort zeigen wir auch, dass sich (4.1) in die folgende Form bringen lässt:

$$\boxed{P_1(x) = y_0\,L_0(x) + y_1\,L_1(x) \quad \text{mit} \quad L_0(x) = \frac{x - x_1}{x_0 - x_1}\,, \quad L_1(x) = \frac{x - x_0}{x_1 - x_0}} \tag{4.2}$$

Die Funktionen L_0 und L_1 heißen die **Lagrange-Polynome vom Grad** 1 und haben die folgenden interessanten Eigenschaften:

$$L_0(x_0) = \frac{x_0 - x_1}{x_0 - x_1} = 1, \qquad L_0(x_1) = \frac{x_1 - x_1}{x_0 - x_1} = 0,$$
$$L_1(x_0) = \frac{x_0 - x_0}{x_1 - x_0} = 0, \qquad L_1(x_1) = \frac{x_1 - x_0}{x_1 - x_0} = 1.$$

Damit folgt aus $P_1(x) = y_0\,L_0(x) + y_1\,L_1(x)$ direkt

$$P_1(x_0) = y_0\,L_0(x_0) + y_1\,L_1(x_0) = y_0 \cdot 1 + y_1 \cdot 0 = y_0,$$
$$P_1(x_1) = y_0\,L_0(x_1) + y_1\,L_1(x_1) = y_0 \cdot 0 + y_1 \cdot 1 = y_1.$$

Es ist anschaulich klar, dass es zu gegebenen Datenpunkten $(x_0; y_0)$ und $(x_1; y_1)$ mit $x_0 \neq x_1$ **genau ein** lineares Interpolationspolynom gibt, da man durch zwei verschiedene Punkte in der $(x; y)$-Ebene nur genau eine Gerade legen kann.

Betrachten wir ein Beispiel.

Beispiel 4.4. (lineares Interpolationspolynom)
Gesucht ist das lineare Interpolationspolynom zu den Datenpunkten $(x_0; y_0) = (1; 1)$ und $(x_1; y_1) = (4; 2)$. Nach (4.1) ist das lineare Interpolationspolynom

$$P_1(x) = \frac{2 - 1}{4 - 1}\,x + \frac{1 \cdot 4 - 1 \cdot 2}{4 - 1} = \frac{1}{3}\,x + \frac{2}{3}\,.$$

Mit (4.2) können wir das lineare Interpolationspolynom auch wie folgt schreiben:

$$P_1(x) = 1 \cdot \frac{x - 4}{1 - 4} + 2 \cdot \frac{x - 1}{4 - 1} = \frac{x - 4}{-3} + 2 \cdot \frac{x - 1}{3}$$

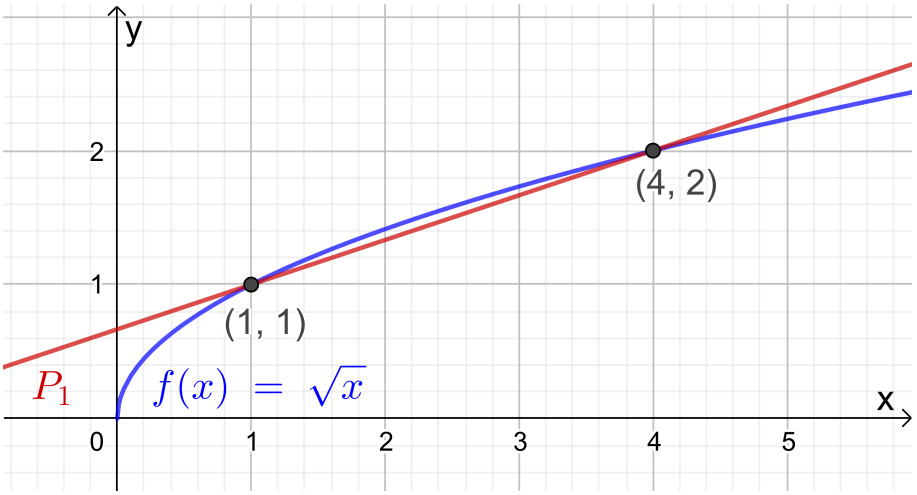


Abb. 4.2: Die Funktion $f : [0; \infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und ihr lineares Interpolationspolynom P_1 für die Datenpunkte $(1; 1)$ und $(4; 2)$.

Die gegebenen Datenpunkte stammen von der Quadratwurzelfunktion

$$f : [0; \infty[\to \mathbb{R}, \qquad f(x) = \sqrt{x},$$

denn es gilt

$$\big(1; f(1)\big) = \big(1; \sqrt{1}\big) = (1; 1) \qquad \text{und} \qquad \big(4; f(4)\big) = \big(4; \sqrt{4}\big) = (4; 2).$$

In Abbildung 4.2 ist die Quadratwurzelfunktion zusammen mit dem linearen Interpolationspolynom P_1 gezeichnet. Wir sehen, dass das lineare Interpolationspolynom P_1 die Funktion $f(x) = \sqrt{x}$ auf dem Intervall $\left[\frac{1}{2}; \frac{9}{2}\right] = [0{,}5;\ 4{,}5]$ gar nicht so schlecht annähert, aber außerhalb dieses Intervalls liefert das lineare Interpolationspolynom nur eine schlechte Näherung. So erhalten wir beispielsweise

$$P_1(3) = \frac{1}{3} \cdot 3 + \frac{2}{3} = \frac{5}{3} \doteq 1{,}6667 \qquad \text{und} \qquad \sqrt{3} \doteq 1,7321,$$

d.h. für $x = 3$ haben wir einen relativen Fehler von

$$\mathrm{Rel}\big(P_1(3)\big) = \frac{\sqrt{3} - P_1(3)}{\sqrt{3}} = \frac{\sqrt{3} - \frac{5}{3}}{\sqrt{3}} \doteq 0{,}038,$$

also einen relativen Fehler von knapp 4 %. ♠

Wir wollen nun den Fall der Interpolation von drei Datenpunkten $(x_0; y_0)$, $(x_1; y_1)$ und $(x_2; y_2)$ in der $(x; y)$-Ebene mit $x_0 \neq x_1$, $x_0 \neq x_2$ und $x_1 \neq x_2$ mit einem (höchstens) quadratischen Polynom betrachten. Das **quadratische Interpolationspolynom** P_2 ist durch die folgende Formel gegeben:

$$P_2(x) = y_0\,L_0(x) + y_1\,L_1(x) + y_2\,L_2(x) \quad \text{mit} \quad L_0(x) = \frac{(x-x_1)\,(x-x_2)}{(x_0-x_1)\,(x_0-x_2)}\,,$$
$$L_1(x) = \frac{(x-x_0)\,(x-x_2)}{(x_1-x_0)\,(x_1-x_2)}\,, \quad L_2(x) = \frac{(x-x_0)\,(x-x_1)}{(x_2-x_0)\,(x_2-x_1)}\,. \tag{4.3}$$

Die Funktionen L_0, L_1, L_2 heißen die **Lagrange-Polynome vom Grad** 2, und (4.3) wird als die **Interpolationsformel von Lagrange** bezeichnet.

Warum ist die Formel (4.3) **richtig?** Wir bemerken zunächst durch Inspektion, dass die Funktionen L_0, L_1, L_2 jeweils Polynome vom Grad 2 sind. Also muss P_2, definiert durch (4.3), ein Polynom vom Grad ≤ 2 sein. Weiter überprüft man leicht (dieses ist eine Übungsaufgabe), dass gilt

$$L_j(x_i) = \delta_{i,j} \qquad \text{für alle } i,j = 0,1,2, \tag{4.4}$$

wobei $\delta_{i,j}$ das **Kronecker-Delta** ist, welches wie folgt definiert ist:

$$\delta_{i,j} = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j. \end{cases} \tag{4.5}$$

Dann folgt mit (4.4)

$$P_2(x_0) = y_0\,L_0(x_0) + y_1\,L_1(x_0) + y_2\,L_2(x_0) = 1\cdot y_0 + 0\cdot y_1 + 0\cdot y_2 = y_0,$$
$$P_2(x_1) = y_0\,L_0(x_1) + y_1\,L_1(x_1) + y_2\,L_2(x_1) = 0\cdot y_0 + 1\cdot y_1 + 0\cdot y_2 = y_1,$$
$$P_2(x_2) = y_0\,L_0(x_2) + y_1\,L_1(x_2) + y_2\,L_2(x_2) = 0\cdot y_0 + 0\cdot y_1 + 1\cdot y_2 = y_2,$$

und wir sehen, dass P_2 die Datenpunkte $(x_0; y_0)$, $(x_1; y_1)$ und $(x_2; y_2)$ interpoliert.

Betrachten wir ein Beispiel.

Beispiel 4.5. (quadratisches Interpolationspolynom)
Gesucht ist ein (höchstens) quadratisches Polynom, das die Datenpunkte

$$(x_0; y_0) = (1; 1), \qquad (x_1; y_1) = (4; 2) \qquad \text{und} \qquad (x_2; y_2) = (2{,}89; 1{,}7)$$

interpoliert. Es gilt $(x_i; y_i) = \left(x_i; \sqrt{x_i}\right)$ für $i = 0,1,2$, d.h. die Datenpunkte stammen von der Quadratwurzelfunktion $f :]0; \infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$.

Nach (4.3) ist ein (höchstens) quadratische Interpolationspolynom gegeben durch

$$P_2(x) = 1\cdot\frac{(x-4)\,(x-2{,}89)}{(1-4)\,(1-2{,}89)} + 2\cdot\frac{(x-1)\,(x-2{,}89)}{(4-1)\,(4-2{,}89)} + 1{,}7\cdot\frac{(x-1)\,(x-4)}{(2{,}89-1)\,(2{,}89-4)}\,.$$

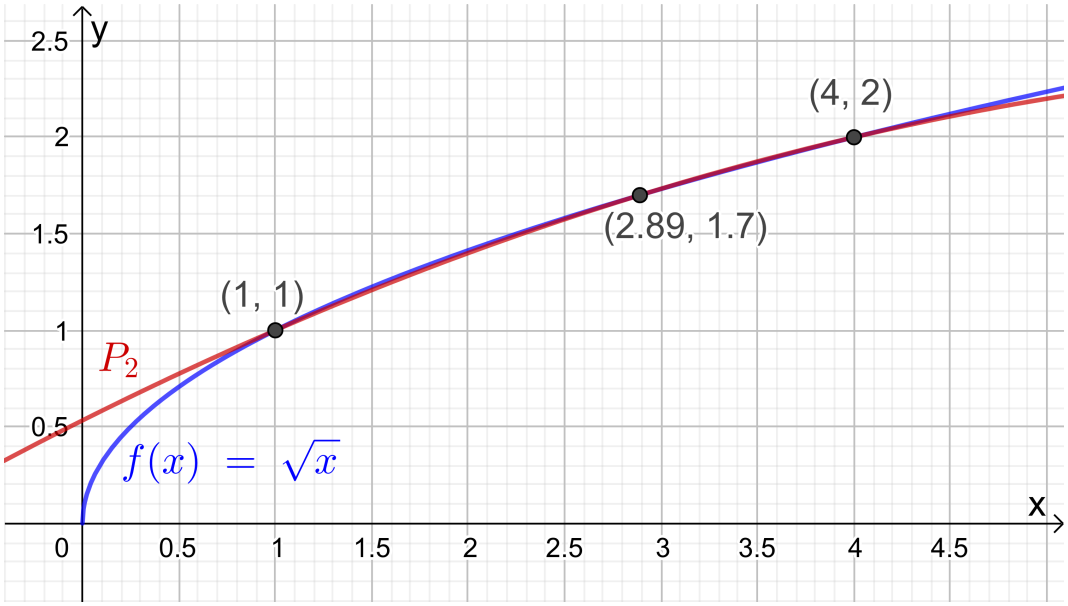


Abb. 4.3: Die Funktion $f : [0; \infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und ihr quadratisches Interpolationspolynom P_2 für die Datenpunkte $(1; 1)$, $(4; 2)$ und $(2{,}89; 1{,}7)$.

In Abbildung 4.3 ist die Quadratwurzelfunktion zusammen mit dem (höchstens) quadratischen Interpolationspolynom gezeichnet. Verglichen mit dem linearen Interpolationspolynom P_1 (siehe Beispiel 4.4 und Abbildung 4.2) hat sich die Näherung von $f(x) = \sqrt{x}$ deutlich verbessert. Mit dem bloßen Auge sieht man auf dem Intervall $[0{,}8;\ 4{,}4]$ in der Veranschaulichung der Graphen in Abbildung 4.3 fast keinen Unterschied mehr zwischen $f(x) = \sqrt{x}$ und P_2. Nun gilt in $x = 3$

$$P_2(3) = 1 \cdot \frac{(3-4)\,(3-2{,}89)}{(1-4)\,(1-2{,}89)} + 2 \cdot \frac{(3-1)\,(3-2{,}89)}{(4-1)\,(4-2{,}89)} + 1{,}7 \cdot \frac{(3-1)\,(3-4)}{(2{,}89-1)\,(2{,}89-4)}$$
$$\doteq 1{,}7334,$$

d.h. für $x = 3$ haben wir einen relativen Fehler von

$$\mathrm{Rel}\big(P_2(3)\big) = \frac{\sqrt{3} - P_2(3)}{\sqrt{3}} \doteq 0{,}00078,$$

also einen relativen Fehler von knapp unter 0,08 %. (Natürlich liegt 3 dicht bei $2,98$, und für einen anderen Punkt wie $x = 2$ hätten wir vermutlich eine etwas schlechtere Näherung bekommen.) ♠

Wir gehen nun noch einen Schritt weiter und wollen $n + 1$ Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, in der $(x; y)$-Ebene mit paarweise verschiedenen Punkten x_i, $i = 0, 1, 2, \ldots, n$, durch ein Polynom $P_n \in \mathbb{P}_n$ (also vom Grad $\leq n$) interpolieren.

(Dass die Datenpunkte „paarweise verschieden“ sind, bedeutet, dass $x_i \neq x_j$ gilt, wenn $i \neq j$ ist.) In Analogie zu (4.2) und (4.3) bekommt man dann den folgenden Satz, den wir auch beweisen werden:

Satz 4.6. (Interpolationsformel von Lagrange)
Seien $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, genau $n+1$ Datenpunkte mit paarweise verschiedenen $x_0, x_1, x_2, \ldots, x_n$ (d.h. $x_i \neq x_j$ wenn $i \neq j$). Dann ist

$$P_n(x) = y_0\, L_0(x) + y_1\, L_1(x) + \ldots + y_n\, L_n(x) = \sum_{i=0}^{n} y_i\, L_i(x) \tag{4.6}$$

*mit den **Lagrange-Polynomen vom Grad** n (bzgl. $x_0, x_1, \ldots, x_n$)*

$$L_i(x) = \frac{(x - x_0) \cdot \ldots \cdot (x - x_{i-1}) \cdot (x - x_{i+1}) \cdot \ldots \cdot (x - x_n)}{(x_i - x_0) \cdot \ldots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)}\,, \tag{4.7}$$

*$i = 0, 1, 2, \ldots, n$, ein **Polynom in $\mathbb{P}_n$ (also vom Grad $\leq n$), das die Datenpunkte interpoliert**. Die Darstellung (4.6) eines interpolierenden Polynoms in $\mathbb{P}_n$ nennt man die **Interpolationsformel von Lagrange**.*

Schauen wir uns die Formel (4.7) für die Lagrange-Polynome genauer an: Im Zähler von L_i steht ein Polynom vom exakten Grad n, bei dem alle Linearfaktoren $(x - x_j)$, $j = 0, 1, 2, \ldots, n$, mit Ausnahme von $(x - x_i)$ multipliziert werden. Im Nenner steht eine Konstante, die man durch Multiplikation der $(x_i - x_j)$, $j = 0, 1, 2, \ldots, n$, mit Ausnahme von $(x_i - x_i)$ erhält. Nach Formel (4.7) gilt also

$$\begin{aligned}
L_0(x) &= \frac{(x - x_1) \cdot \ldots \cdot (x - x_n)}{(x_0 - x_1) \cdot \ldots \cdot (x_0 - x_n)}\,,\\
L_1(x) &= \frac{(x - x_0) \cdot (x - x_2) \cdot \ldots \cdot (x - x_n)}{(x_1 - x_0) \cdot (x_1 - x_2) \cdot \ldots \cdot (x_1 - x_n)}\,,\\
L_2(x) &= \frac{(x - x_0) \cdot (x - x_1) \cdot (x - x_3) \cdot \ldots \cdot (x - x_n)}{(x_2 - x_0) \cdot (x_2 - x_1) \cdot (x_2 - x_3) \cdot \ldots \cdot (x_2 - x_n)}\,,\\
&\vdots\\
L_{n-1}(x) &= \frac{(x - x_0) \cdot (x - x_1) \cdot \ldots \cdot (x - x_{n-2}) \cdot (x - x_n)}{(x_{n-1} - x_0) \cdot (x_{n-1} - x_1) \cdot \ldots \cdot (x_{n-1} - x_{n-2}) \cdot (x_{n-1} - x_n)}\,,\\
L_n(x) &= \frac{(x - x_0) \cdot (x - x_1) \cdot \ldots \cdot (x - x_{n-1})}{(x_n - x_0) \cdot (x_n - x_1) \cdot \ldots \cdot (x_n - x_{n-1})}\,.
\end{aligned}$$

Die Formeln (4.2) und (4.3) sind jeweils der Sonderfall von (4.6) und (4.7) für den

Fall $n = 1$ bzw. $n = 2$. (Inspizieren Sie bitte (4.6) und (4.7) noch einmal, um sich dieses klar zu machen.)

Wir werden im Beweis von Satz 4.6 noch zeigen, dass die in (4.7) definierten Lagrange-Polynome vom Grad n die Eigenschaft

$$L_j(x_i) = \delta_{i,j} \qquad \text{für alle } i, j = 0, 1, 2, \ldots, n \tag{4.8}$$

haben, wobei $\delta_{i,j}$ das in (4.5) definierte Kronecker-Delta ist. Mit (4.8) zeigt man dann leicht, dass (4.6) die Datenpunkte interpoliert.

Betrachten wir zunächst ein Beispiel.

Beispiel 4.7. (Interpolationspolynom vom Grad ≤ 4)
Wir wollen die Cosinusfunktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \cos(x)$, durch ein Interpolationspolynom in $\mathbb{P}_4$ (also vom Grad ≤ 4) bzgl. der nachfolgenden fünf Datenpunkte (mit paarweise verschiedenen x_i, $i = 0, 1, \ldots, 4$) interpolieren. Die Variable von $f(x) = \cos(x)$ ist dabei im Bogenmaß angegeben (vgl. Anhang A.8).

$$\begin{aligned}
(x_0; y_0) &= \left(-\tfrac{\pi}{2}; \cos\left(-\tfrac{\pi}{2}\right)\right) = \left(-\tfrac{\pi}{2}; 0\right), \\
(x_1; y_1) &= \left(-\tfrac{\pi}{4}; \cos\left(-\tfrac{\pi}{4}\right)\right) = \left(-\tfrac{\pi}{4}; \tfrac{\sqrt{2}}{2}\right), \\
(x_2; y_2) &= (0; \cos(0)) = (0; 1), \\
(x_3, y_3) &= \left(\tfrac{\pi}{4}; \cos\left(\tfrac{\pi}{4}\right)\right) = \left(\tfrac{\pi}{4}; \tfrac{\sqrt{2}}{2}\right), \\
(x_4, y_4) &= \left(\tfrac{\pi}{2}; \cos\left(\tfrac{\pi}{2}\right)\right) = \left(\tfrac{\pi}{2}; 0\right).
\end{aligned}$$

Wir stellen zuerst die Lagrange-Polynome von Grad 4 auf:

$$\begin{aligned}
L_0(x) &= \frac{\left(x + \frac{\pi}{4}\right) \cdot (x - 0) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right)}{\left(-\frac{\pi}{2} + \frac{\pi}{4}\right) \cdot \left(-\frac{\pi}{2} - 0\right) \cdot \left(-\frac{\pi}{2} - \frac{\pi}{4}\right) \cdot \left(-\frac{\pi}{2} - \frac{\pi}{2}\right)} \\
&= \frac{32}{3\,\pi^4} \cdot \left(x + \frac{\pi}{4}\right) \cdot x \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right), \\
L_1(x) &= \frac{\left(x + \frac{\pi}{2}\right) \cdot (x - 0) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right)}{\left(-\frac{\pi}{4} + \frac{\pi}{2}\right) \cdot \left(-\frac{\pi}{4} - 0\right) \cdot \left(-\frac{\pi}{4} - \frac{\pi}{4}\right) \cdot \left(-\frac{\pi}{4} - \frac{\pi}{2}\right)} \\
&= -\frac{128}{3\,\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot x \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right), \\
L_2(x) &= \frac{\left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right)}{\left(0 + \frac{\pi}{4}\right) \cdot \left(0 + \frac{\pi}{2}\right) \cdot \left(0 - \frac{\pi}{4}\right) \cdot \left(0 - \frac{\pi}{2}\right)} \\
&= \frac{64}{\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right),
\end{aligned}$$

$$
\begin{aligned}
L_3(x) &= \frac{\left(x+\frac{\pi}{2}\right)\cdot\left(x+\frac{\pi}{4}\right)\cdot(x-0)\cdot\left(x-\frac{\pi}{2}\right)}{\left(\frac{\pi}{4}+\frac{\pi}{2}\right)\cdot\left(\frac{\pi}{4}+\frac{\pi}{4}\right)\cdot\left(\frac{\pi}{4}-0\right)\cdot\left(\frac{\pi}{4}-\frac{\pi}{2}\right)} \\
&= -\frac{128}{3\,\pi^4}\cdot\left(x+\frac{\pi}{2}\right)\cdot\left(x+\frac{\pi}{4}\right)\cdot x\cdot\left(x-\frac{\pi}{2}\right), \\
L_4(x) &= \frac{\left(x+\frac{\pi}{2}\right)\cdot\left(x+\frac{\pi}{4}\right)\cdot(x-0)\cdot\left(x-\frac{\pi}{4}\right)}{\left(\frac{\pi}{2}+\frac{\pi}{2}\right)\cdot\left(\frac{\pi}{2}+\frac{\pi}{4}\right)\cdot\left(\frac{\pi}{2}-0\right)\cdot\left(\frac{\pi}{2}-\frac{\pi}{4}\right)} \\
&= \frac{32}{3\,\pi^4}\cdot\left(x+\frac{\pi}{2}\right)\cdot\left(x+\frac{\pi}{4}\right)\cdot x\cdot\left(x-\frac{\pi}{4}\right).
\end{aligned}
$$

Mit diesen ist das Interpolationspolynom vom Grad ≤ 4 gegeben durch

$$
\begin{aligned}
P_4(x) &= 0\cdot L_0(x)+\frac{\sqrt{2}}{2}\cdot L_1(x)+1\cdot L_2(x)+\frac{\sqrt{2}}{2}\cdot L_3(x)+0\cdot L_4(x) \\
&= \frac{\sqrt{2}}{2}\,L_1(x)+L_2(x)+\frac{\sqrt{2}}{2}\,L_3(x) = L_2(x)+\frac{\sqrt{2}}{2}\,\big(L_1(x)+L_3(x)\big).
\end{aligned}
$$

Wir können P_4 in diesem konkreten Beispiel noch weiter vereinfachen:

$$
\begin{aligned}
L_1(x)+L_3(x) &= -\frac{128}{3\,\pi^4}\cdot\left(x+\frac{\pi}{2}\right)\cdot x\cdot\left(x-\frac{\pi}{4}\right)\cdot\left(x-\frac{\pi}{2}\right) \\
&\quad -\frac{128}{3\,\pi^4}\cdot\left(x+\frac{\pi}{2}\right)\cdot\left(x+\frac{\pi}{4}\right)\cdot x\cdot\left(x-\frac{\pi}{2}\right) \\
&= -\frac{128}{3\,\pi^4}\cdot\left(x+\frac{\pi}{2}\right)\cdot x\cdot\left(x-\frac{\pi}{2}\right)\cdot\underbrace{\left[\left(x-\frac{\pi}{4}\right)+\left(x+\frac{\pi}{4}\right)\right]}_{=2\,x} \\
&= -\frac{256}{3\,\pi^4}\cdot x^2\cdot\left(x+\frac{\pi}{2}\right)\cdot\left(x-\frac{\pi}{2}\right) = -\frac{256}{3\,\pi^4}\cdot x^2\cdot\left(x^2-\frac{\pi^2}{4}\right),
\end{aligned}
$$

wobei wir im letzten Schritt die dritte binomische Formel verwendet haben. Ebenfalls mit der dritten binomischen Formel folgt

$$
\begin{aligned}
L_2(x) &= \frac{64}{\pi^4}\cdot\left(x+\frac{\pi}{2}\right)\cdot\left(x+\frac{\pi}{4}\right)\cdot\left(x-\frac{\pi}{4}\right)\cdot\left(x-\frac{\pi}{2}\right) \\
&= \frac{64}{\pi^4}\cdot\left(x+\frac{\pi}{2}\right)\cdot\left(x-\frac{\pi}{2}\right)\cdot\left(x+\frac{\pi}{4}\right)\cdot\left(x-\frac{\pi}{4}\right) \\
&= \frac{64}{\pi^4}\cdot\left(x^2-\frac{\pi^4}{4}\right)\cdot\left(x^2-\frac{\pi^2}{16}\right).
\end{aligned}
$$

Also erhalten wir für das Interpolationspolynom

$$
P_4(x) = \frac{64}{\pi^4}\cdot\left(x^2-\frac{\pi^4}{4}\right)\cdot\left(x^2-\frac{\pi^2}{16}\right)-\frac{128\sqrt{2}}{3\,\pi^4}\cdot x^2\cdot\left(x^2-\frac{\pi^2}{4}\right).
$$

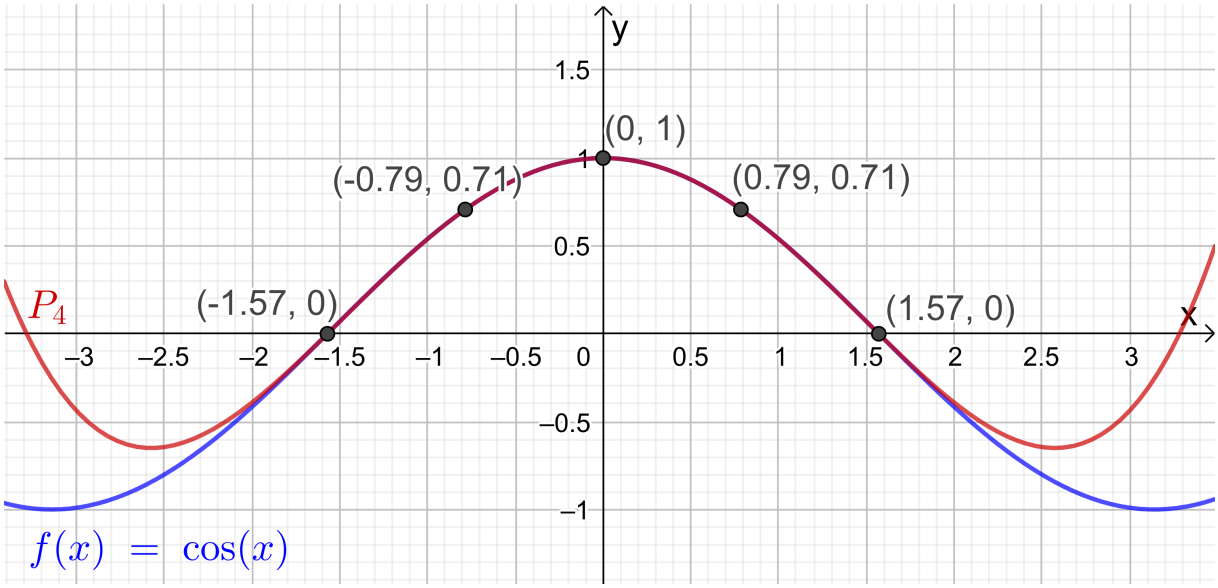


Abb. 4.4: $\cos(x)$ und sein Interpolationspolynom P_4.

Die Funktion $f(x) = \cos(x)$ und das Interpolationspolynom P_4 vom Grad ≤ 4 sind in Abbildung 4.4 gezeichnet. ♠

Wir beweisen nun die Interpolationsformel von Lagrange.

Beweis von Satz 4.6: Wir untersuchen zunächst die Lagrange-Polynome (4.7): Der Nenner von L_i, $i = 0, 1, 2, \ldots, n$, ist jeweils eine Konstante. Im Zähler von L_i, $i = 0, 1, 2, \ldots, n$, gibt es jeweils n lineare Faktoren; also muss L_i ein Polynom vom exakten Grad n sein. An der Formel (4.6), also an

$$P_n(x) = y_0\, L_0(x) + y_1\, L_1(x) + y_2\, L_2(x) + \ldots + y_n\, L_n(x),$$

sieht man dann direkt, dass P_n dann auch ein Polynom vom Grad $\leq n$ sein muss, also in $\mathbb{P}_n$ liegt.

Es gilt für die Lagrange-Polynome

$$L_i(x_k) = \frac{(x_k - x_0) \cdot \ldots \cdot (x_k - x_{i-1}) \cdot (x_k - x_{i+1}) \cdot \ldots \cdot (x_k - x_n)}{(x_i - x_0) \cdot \ldots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)} = 0$$
$$\text{für } i, k = 0, 1, 2, \ldots, n \text{ mit } i \neq k,$$

weil im Zähler dann ein Term, nämlich $(x_k - x_k)$, gleich 0 ist; und es gilt

$$L_i(x_i) = \frac{(x_i - x_0) \cdot \ldots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)}{(x_i - x_0) \cdot \ldots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)} = 1$$
$$\text{für } i = 0, 1, 2, \ldots, n.$$

Damit folgt also

$$L_i(x_k) = \delta_{i,k} \qquad \text{für alle } i,k = 0,1,2,\ldots,n,$$

wobei $\delta_{i,k}$ das in (4.5) definierte Kronecker-Delta ist. Daraus folgt direkt

$$P_n(x_k) = y_0\, L_0(x_k) + y_1\, L_1(x_k) + \ldots + y_n\, L_n(x_k) = y_k \quad \text{für } k = 0,1,2,\ldots,n,$$

denn nur für $i = k$ gilt $L_i(x_k) = 1$, und für alle i mit $i \neq k$ gilt $L_i(x_k) = 0$. Also interpoliert des Polynom P_n die Datenpunkte $(x_i; y_i)$, $i = 0,1,2,\ldots,n$. □

Frage: Gibt es **mehr als ein Polynom in $\mathbb{P}_n$ (also vom Grad $\leq n$), welches gegebene Datenpunkte** $(x_i; y_i)$, $i = 0,1,2,\ldots,n$, in der $(x;y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$ **interpoliert?** Hierüber gibt der nachfolgende Satz Auskunft.

Satz 4.8. (Existenz und Eindeutigkeit des Interpolationspolynoms)
*Seien $(x_i; y_i)$, $i = 0,1,2,\ldots,n$, genau $n+1$ Datenpunkte in der $(x;y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$ (d.h. $x_i \neq x_j$ wenn $i \neq j$). Dann existiert **genau ein Polynom P_n in $\mathbb{P}_n$ (also vom Grad $\leq n$)** mit*

$$P_n(x_i) = y_i \qquad \textit{für alle } i = 0,1,2,\ldots,n, \tag{4.9}$$

*d.h. das Interpolationspolynom P_n von Grad $\leq n$ ist **eindeutig bestimmt**.*

Beweis von Satz 4.8: Die Interpolationsformel von Lagrange (siehe Satz 4.6) liefert uns, dass mindestens ein Polynom $P_n \in \mathbb{P}_n$ vom Grad $\leq n$ (nämlich das durch (4.6) und (4.7) gegebene Polynom) existiert, welches die $n+1$ Datenpunkte $(x_i; y_i)$, $i = 0,1,2,\ldots,n$, interpoliert. Wir müssen also nur noch zeigen, dass dieses Interpolationspolynom eindeutig bestimmt ist.

Dazu nehmen wir an, dass P_n und Q_n zwei Polynome in $\mathbb{P}_n$ (also vom Grad $\leq n$) seien, für die jeweils gilt $P_n(x_i) = y_i$, $i = 0,1,2,\ldots,n$, bzw. $Q_n(x_i) = y_i$, $i = 0,1,2,\ldots,n$. Die Differenzfunktion $F_n(x) = P_n(x) - Q_n(x)$ ist dann ebenfalls ein Polynom vom Grad $\leq n$ (d.h. sie liegt in $\mathbb{P}_n$), und sie erfüllt

$$F_n(x_i) = P_n(x_i) - Q_n(x_i) = y_i - y_i = 0 \qquad \text{für alle } i = 0,1,2,\ldots,n.$$

Damit hat das Polynom $F_n \in \mathbb{P}_n$ vom Grad $\leq n$ aber $n+1$ Nullstellen. Daraus folgt nach Satz 1.14, dass F_n das Nullpolynom ist, also

$$F_n(x) = P_n(x) - Q_n(x) = 0 \quad \text{für alle } x \in \mathbb{R}$$

$$\iff \qquad P_n(x) = Q_n(x) \quad \text{für alle } x \in \mathbb{R}.$$

Also sind P_n und Q_n gleich, und das interpolierende Polynom in $\mathbb{P}_n$ ist damit eindeutig bestimmt. □

4.3 Dividierte Differenzen und die Interpolationsformel von Newton

Die Berechnung des interpolierenden Polynoms $P_n \in \mathbb{P}_N$ mit der Interpolationsformel von Lagrange ist relativ aufwendig, denn man muss die $n+1$ Lagrange-Polynome $L_0, L_1, \ldots, L_n$ auswerten, was jeweils $2n-1$ Multiplikationen/Divisionen und zusätzlich $2n-2$ Additionen/Subtraktionen erfordert. In der Tat sind die Lagrange-Polynome und die Interpolationsformel von Lagrange vor allem für theoretische Untersuchungen wichtig. Für die praktische Berechnung des interpolierenden Polynoms $P_n \in \mathbb{P}_n$ verwendet man die **Interpolationsformel von Newton**, die wir in diesem Teilkapitel kennenlernen werden. Als Vorbereitung dafür benötigen wir aber zunächst **dividierte Differenzen**.

Definition 4.9. (dividierte Differenzen)
Sei $n \in \mathbb{N}$. Sei $f : [a;b] \to \mathbb{R}$ eine Funktion, und seien $x_0, x_1, \ldots, x_n$ genau $n+1$ verschiedene Punkte in $[a;b]$. Dann definieren wir rekursiv :

(1) $f[x_i, x_{i+1}] = \dfrac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$

heißt eine ***dividierte Differenz erster Ordnung*** *von f.*

(2) $f[x_{i-1}, x_i, x_{i+1}] = \dfrac{f[x_i, x_{i+1}] - f[x_{i-1}, x_i]}{x_{i+1} - x_{i-1}}$

heißt eine ***dividierte Differenz zweiter Ordnung*** *von f.*

$\vdots$

(n) $f[x_0, x_1, \ldots, x_{n-1}, x_n] = \dfrac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}$

heißt eine ***dividierte Differenz n-ter Ordnung*** *von f.*

Beispiel 4.10. (dividierte Differenzen)
Sei $f : [0;\infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und seien $x_0 = 1$, $x_1 = 4$, $x_2 = 2{,}89$. Die

zugehörigen Funktionswerte sind dann

$$f(x_0) = f(1) = \sqrt{1} = 1,$$
$$f(x_1) = f(4) = \sqrt{4} = 2,$$
$$f(x_2) = f(2{,}89) = \sqrt{2{,}89} = 1{,}7,$$

und die dividierten Differenzen erster und zweiter Ordnung lauten

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{2-1}{4-1} = \frac{1}{3} \doteq 0{,}3333,$$
$$f[x_1, x_2] = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{1{,}7-2}{2{,}89-4} = \frac{-0{,}3}{-1{,}11} = \frac{30}{111} \doteq 0{,}2703,$$
$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{\frac{30}{111} - \frac{1}{3}}{2{,}89 - 1} = \frac{\frac{90-111}{333}}{1{,}89} = \frac{\frac{-21}{333}}{1{,}89}$$
$$= \frac{\frac{-7}{111}}{1{,}89} = -\frac{700}{111 \cdot 189} = -\frac{100}{111 \cdot 27} = -\frac{100}{2997} \doteq -0{,}03337.$$

Mit nur drei Punkten können wir keine dividierten Differenzen höherer Ordnung bilden. ♠

Welche Eigenschaften haben die dividierten Differenzen? Damit befasst sich der nächste Satz.

Satz 4.11. (Eigenschaften der dividierten Differenzen)
Seien $f :]c; d[\to \mathbb{R}$ eine Funktion und $[a; b] \subseteq]c; d[$, und seien $x_0, x_1, \ldots, x_n$ genau $n+1$ verschiedene Punkte in $[a; b]$. Dann gelten:

(1) Verändert man in einer dividierten Differenz die Reihenfolge der Punkte (d.h. permutiert man die Punkte), so ändert sich der Wert der dividierten Differenz nicht.

(2) Aus dem Mittelwertsatz der Differentialrechnung (siehe Satz 1.4) folgt: Ist f auf $]c; d[$ stetig differenzierbar, so gibt es einen Punkt z zwischen x_i und x_{i+1}, so dass gilt

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = f'(z).$$

(3) Ist f auf $]c;d[$ n-mal stetig differenzierbar, so gibt es einen Punkt z im kleinsten Intervall, das alle $x_0, x_1, \ldots, x_n$ enthält, so dass gilt

$$f[x_0, x_1, \ldots, x_{n-1}, x_n] = \frac{1}{n!} f^{(n)}(z).$$

Wir werden Satz 4.11 (1) auf einem Übungszettel für dividierte Differenzen erster und zweiter Ordnung beweisen.

Nach dieser Vorbereitung können wir schließlich die Interpolationsformel von Newton einführen.

Satz 4.12. (Interpolationsformel von Newton)
*Sei $n \in \mathbb{N}_0$. Sei $f : [a;b] \to \mathbb{R}$ eine Funktion, und seien $x_0, x_1, \ldots, x_n$ genau $n+1$ verschiedene Punkte in $[a;b]$. Seien $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, n$, die zugehörigen Datenpunkte in der $(x;y)$-Ebene für die Interpolation der Funktion f. Die **(jeweils eindeutig bestimmten) interpolierenden Polynome** $P_j \in \mathbb{P}_j$ **(also** $\mathrm{Grad}(P_j) \leq j$**) mit** $j = 0, 1, 2, \ldots, n$, die jeweils die Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, j$, interpolieren, können mit Hilfe der dividierten Differenzen von f wie folgt berechnet werden:*

$$\begin{aligned}
P_0(x) &= f(x_0) \\
P_1(x) &= f(x_0) + (x - x_0)\, f[x_0, x_1], \\
P_2(x) &= f(x_0) + (x - x_0)\, f[x_0, x_1] + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x_2], \\
P_3(x) &= f(x_0) + (x - x_0)\, f[x_0, x_1] + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x_2] \\
&\quad + (x - x_0)\,(x - x_1)\,(x - x_2)\, f[x_0, x_1, x_2, x_3], \\
&\vdots \\
P_n(x) &= f(x_0) + (x - x_0)\, f[x_0, x_1] + \ldots \\
&\quad + (x - x_0) \cdot \ldots \cdot (x - x_{n-2})\, f[x_0, x_1, \ldots, x_{n-1}] \\
&\quad + (x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_{n-1})\, f[x_0, x_1, \ldots, x_n].
\end{aligned}$$

Warum ist die Interpolationsformel von Newton so nützlich? Es gilt für jedes $j = 0, 1, 2, \ldots, n-1$ die Rekursionsformel

$$\boxed{P_{j+1}(x) = P_j(x) + (x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_{j+1}],} \tag{4.10}$$

denn nach Satz 4.12 gilt

$$\begin{aligned} P_{j+1}(x) &= \underbrace{f(x_0) + \ldots + (x - x_0) \cdot \ldots \cdot (x - x_{j-1})\, f[x_0, x_1, \ldots, x_j]}_{= P_j(x)} \\ &\quad + (x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_j, x_{j+1}] \\ &= P_j(x) + (x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_n, x_{j+1}]. \end{aligned}$$

Die Formel (4.10) besagt das Folgende: Wollen wir die Interpolation verbessern und fügen daher einen neuen Datenpunkt $\big(x_{j+1}; f(x_{j+1})\big)$ zu $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, j$, hinzu, so brauchen wir das Interpolationspolynom nicht neu zu berechnen, sondern müssen nur zu $P_j(x)$ einfach den Term

$$(x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_j, x_{j+1}]$$

addieren, um $P_{j+1}(x)$ zu erhalten.

Wir werden die Interpolationsformel von Newton auf einem Übungszettel für die Fälle $n = 1$ und $n = 2$ beweisen.

Betrachten wir zunächst ein Beispiel.

Beispiel 4.13. (Interpolationsformel von Newton)
Sei $f : [0; \infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und seien $x_0 = 1$, $x_1 = 4$ und $x_2 = 2{,}89$. Die zugehörigen Funktionswerte sind dann $f(x_0) = 1$, $f(x_1) = 2$, $f(x_2) = 1{,}7$. Nach Beispiel 4.10 sind die dividierten Differenzen erster und zweiter Ordnung

$$f[x_0, x_1] = \frac{1}{3}\,, \qquad f[x_1, x_2] = \frac{30}{111}\,, \qquad f[x_0, x_1, x_2] = -\frac{100}{2997}\,.$$

Das konstante Interpolationspolynom P_0 des Datenpunkts $(1; 1)$, das lineare Interpolationspolynom P_1 der Datenpunkte $(1; 1)$ und $(4; 2)$ bzw. das quadratische Interpolationspolynom P_2 der Datenpunkte $(1; 1)$, $(4; 2)$ und $(2{,}89; 1{,}7)$ sind also nach der Interpolationsformel von Newton gegeben durch

$$\begin{aligned} P_0(x) &= f(x_0) = 1, \\ P_1(x) &= P_0(x) + (x - x_0)\, f[x_0, x_1] = 1 + \frac{1}{3}\,(x - 1) = \frac{1}{3}\,x + \frac{2}{3}\,, \qquad (4.11) \\ P_2(x) &= P_1(x) + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x_2] \\ &= 1 + \frac{1}{3}\,(x - 1) - \frac{100}{2997}\,(x - 1)\,(x - 4). \qquad (4.12) \end{aligned}$$

Natürlich ist (4.11) identisch mit der Formel für das lineare Interpolationspolynom, die wir in Beispiel 4.4 mit der Interpolationsformel von Lagrange erhalten

haben. Ebenso liefert (4.12) das gleiche quadratische Interpolationspolynom, welches in Beispiel 4.5 berechnet wurde; allerdings ist dieses (ohne Vereinfachungen) nicht offensichtlich. ♠

Was passiert in Satz 4.12 und Definition 4.9, wenn die Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, nur als Punkte in der Ebene gegeben sind, die eine unbekannte Funktion (angenähert) beschreiben, und nicht als Punkte auf dem Graphen einer bekannten Funktion f? Natürlich funktioniert die Berechnung der Interpolierenden ganz analog. Dieses wird in der nächsten Bemerkung erklärt.

Bemerkung 4.14. (dividierte Differenzen für $(x_i; y_i)$, $i = 0, 1, \ldots, n$)
Sei $n \in \mathbb{N}_0$, und seien $(x_i; y_i)$, $i = 0, 1, \ldots, n$, genau $n+1$ Datenpunkte in der $(x; y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$.

(1) Die **dividierten Differenzen** werden wie folgt berechnet:

$$\begin{aligned}
y[x_i, x_{i+1}] &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \qquad i = 0, 1, \ldots, n-1, \\
y[x_{i-1}, x_i, x_{i+1}] &= \frac{y[x_i, x_{i+1}] - y[x_{i-1}, x_i]}{x_{i+1} - x_{i-1}}, \qquad i = 1, \ldots, n-1, \\
&\vdots \\
y[x_0, x_1, \ldots, x_{n-1}, x_n] &= \frac{y[x_1, \ldots, x_n] - y[x_0, \ldots, x_{n-1}]}{x_n - x_0}.
\end{aligned}$$

(2) Die **interpolierenden Polynome** $P_j \in \mathbb{P}_j$ (also $\mathrm{Grad}(P_j) \leq j$) mit $j = 0, 1, 2, \ldots, n$, die jeweils die Datenpunkte $\big(x_i; y_i\big)$, $i = 0, 1, \ldots, j$, interpolieren, werden wie folgt berechnet:

$$\begin{aligned}
P_0(x) &= y_0 \\
P_1(x) &= y_0 + (x - x_0)\, y[x_0, x_1], \\
P_2(x) &= y_0 + (x - x_0)\, y[x_0, x_1] + (x - x_0)\,(x - x_1)\, y[x_0, x_1, x_2], \\
P_3(x) &= y_0 + (x - x_0)\, y[x_0, x_1] + (x - x_0)\,(x - x_1)\, y[x_0, x_1, x_2] \\
&\quad + (x - x_0)\,(x - x_1)\,(x - x_2)\, y[x_0, x_1, x_2, x_3], \\
&\vdots \\
P_n(x) &= y_0 + (x - x_0)\, y[x_0, x_1] + \ldots \\
&\quad + (x - x_0) \cdot \ldots \cdot (x - x_{n-2})\, y[x_0, x_1, \ldots, x_{n-1}] \\
&\quad + (x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_{n-1})\, y[x_0, x_1, \ldots, x_n].
\end{aligned}$$

(3) Auch hier gilt für jedes $j = 0, 1, \ldots, n-1$ die **rekursive Beziehung**

$$P_{j+1}(x) = P_j(x) + (x - x_0) \cdot \ldots \cdot (x - x_j)\, y[x_0, x_1, \ldots, x_{j+1}].$$

4.4 Der Fehler der Polynominterpolation

Sei f eine $(n+1)$-mal stetig differenzierbare Funktion, und seien $x_0, x_1, \ldots, x_n \in \mathbb{R}$ paarweise verschieden. Wir interessieren uns nun dafür, **wie gut das interpolierende Polynom** $P_n \in \mathbb{P}_n$ der Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, 2, \ldots, n$, **die Funktion** f auf einen geeigneten Intervall, welches $x_0, x_1, \ldots, x_n$ enthält, **approximiert d.h. annähert**. Zunächst lernen wir den folgenden Satz kennen.

Satz 4.15. (Interpolationsfehler)
*Sei $f :]c; d[\to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion, sei $[a; b] \subseteq]c; d[$, und seien $x_0, x_1, \ldots, x_n \in [a; b]$ paarweise verschieden. Sei $P_n \in \mathbb{P}_n$ das **interpolierende Polynom der Datenpunkte** $\big(x_i; f(x_i)\big)$, $i = 0, 1, 2, \ldots, n$. Dann gibt es zu jedem $x \in [a; b]$ einen Punkt c_x zwischen dem Minimum und dem Maximum von $x_0, x_1, \ldots, x_n$ und x, so dass der **absolute Fehler der Näherung** $P_n(x)$ **des Funktionswertes** $f(x)$ durch*

$$f(x) - P_n(x) = \frac{(x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_n)}{(n+1)!}\, f^{(n+1)}(c_x) \qquad (4.13)$$

gegeben ist.

Betrachten wir zwei Beispiele, um uns klar zu machen, was (4.13) bedeutet.

Beispiel 4.16. (Interpolationsfehler bei linearer Interpolation)
Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und seien $x_0, x_1 \in [0; 1]$ mit $0 \leq x_0 < x_1 \leq 1$. Nach (4.13) gibt es dann zu jedem $x \in [0; 1]$ ein c_x zwischen dem Minimum und dem Maximum von x_0, x_1 und x mit

$$e^x - P_1(x) = \frac{(x - x_0)\,(x - x_1)}{2!}\, f''(c_x) = \frac{(x - x_0)\,(x - x_1)}{2}\, e^{c_x}, \qquad (4.14)$$

wobei wir genutzt haben, dass $f'(x) = e^x$ und $f''(x) = e^x$ ist, und wobei $P_1 \in \mathbb{P}_1$ das lineare Interpolationspolynom bzgl. der Datenpunkte $(x_0; e^{x_0}), (x_1; e^{x_1})$ ist.

Um den Fehler (4.14) im Folgenden weiter zu untersuchen, nehmen wir ab jetzt an, dass $x_0 < x < x_1$ gilt. Dann folgt mit $x - x_1 = -(x_1 - x)$ aus (4.14), dass

$$e^x - P_1(x) = -\frac{(x - x_0)\,(x_1 - x)}{2}\,e^{c_x}, \tag{4.15}$$

wobei c_x nun in $[x_0; x_1]$ liegen muss (da $x_0 = \min\{x_0, x_1, x\}$, $x_1 = \max\{x_0, x_1, x\}$). Wegen $x - x_0 > 0$ und $x_1 - x > 0$ und $e^{c_x} > 0$ sehen wir an (4.15), dass der Interpolationsfehler immer negativ ist. Falls $[x_0; x_1]$ ein sehr kleines Intervall ist, so ist e^x dort annähernd konstant. Der Fehler (4.15) verhält sich dann annähernd wie ein quadratisches Polynom.

Wir schätzen den Fehler (4.15) nun weiter ab. Wenn wir den Absolutbetrag auf beiden Seiten von (4.15) anwenden, erhalten wir

$$|e^x - P_1(x)| = \left| -\frac{(x - x_0)\,(x_1 - x)}{2}\,e^{c_x} \right| = \frac{(x - x_0)\,(x_1 - x)}{2}\,e^{c_x}. \tag{4.16}$$

Wegen dem streng monoton wachsenden Wachstumsverhalten der Exponentialfunktion folgt aus $x_0 \le c_x \le x_1$, dass $e^{x_0} \le e^{c_x} \le e^{x_1}$. Somit folgt aus (4.16)

$$\frac{(x - x_0)\,(x_1 - x)}{2}\,e^{x_0} \le |e^x - P_1(x)| \le \frac{(x - x_0)\,(x_1 - x)}{2}\,e^{x_1}. \tag{4.17}$$

Die untere und die obere Schranke für den Fehler in (4.17) hängen immer noch von x ab. Um aus (4.17) eine Abschätzung herzuleiten, die von x unabhängig ist, nutzen wir, dass gilt

$$\max_{x \in [x_0; x_1]} \frac{(x - x_0)\,(x_1 - x)}{2} = \frac{h^2}{8} \qquad \text{mit} \qquad h = x_1 - x_0. \tag{4.18}$$

Dieses folgt daraus, dass $(x - x_0)\,(x_1 - x)$ eine nach unten geöffnete Parabel mir den Nullstellen x_0, x_1 ist. Das globale Maximum in $[x_0; x_1]$ muss dann in der Mitte zwischen seinen Nullstellen, also bei $\frac{1}{2}\,(x_1 + x_0)$ auftreten. Damit erhalten wir für $x = \frac{1}{2}\,(x_1 + x_0)$ also $x - x_0 = \frac{1}{2}\,(x_1 + x_0) - x_0 = \frac{1}{2}\,(x_1 - x_0) = \frac{h}{2}$ und $x_1 - x = x_1 - \frac{1}{2}\,(x_1 + x_0) = \frac{1}{2}\,(x_1 - x_0) = \frac{h}{2}$, woraus (4.18) direkt folgt.

Nutzt man (4.18) für die obere Schranke in (4.17) aus, so gilt mit $h = x_1 - x_0$

$$|e^x - P_1(x)| \le \left(\max_{x \in [x_0; x_1]} \frac{(x - x_0)\,(x_1 - x)}{2} \right) e^{x_1} \le \frac{h^2}{8}\,e^{x_1}. \tag{4.19}$$

Da wir nur $x_0, x_1 \in [0; 1]$ betrachten, können wir e^{x_1} weiter durch $e^{x_1} \le e^1 = e$ abschätzen, und es folgt mit $h = x_1 - x_0$

$$|e^x - P_1(x)| \le \frac{e}{8}\,h^2 \qquad \text{für alle } x \text{ mit } 0 \le x_0 \le x \le x_1 \le 1. \tag{4.20}$$

Betrachten wir ein Zahlenbeispiel: Seien $x_0 = 0{,}82$ und $x_1 = 0{,}83$. Dann sind (mit Rundung auf eine 7-stellige Gleitkommadarstellung)

$$f(x_0) = e^{0{,}82} \doteq 2{,}270500, \qquad f(x_1) = e^{0{,}83} \doteq 2{,}293319,$$

und die lineare Interpolierende ist (nach der Interpolationsformel von Lagrange)

$$\begin{aligned} P_1(x) &\doteq 2{,}270500 \cdot \frac{x - 0{,}83}{0{,}082 - 0{,}83} + 2{,}293319 \cdot \frac{x - 0{,}82}{0{,}083 - 0{,}82} \\ &= \frac{2{,}270500 \cdot (0{,}83 - x) + 2{,}293319 \cdot (x - 0{,}82)}{0{,}01}. \end{aligned}$$

Für $x = 0{,}826$ finden wir die Näherung

$$P_1(0{,}826) \doteq 2{,}284191.$$

Der wahre Wert ist

$$f(0{,}826) = e^{0{,}826} \doteq 2{,}284164,$$

und der Interpolationsfehler ist somit

$$e^{0{,}826} - P_1(0{,}826) \doteq 2{,}284164 - 2{,}284191 = -0{,}0000274 = -2{,}74 \cdot 10^{-5}. \tag{4.21}$$

Laut (4.20) sollte mit $h = x_1 - x_0 = 0{,}83 - 0{,}82 = 0{,}01$ gelten

$$|e^{0{,}826} - P_1(0{,}826)| \leq \frac{e}{8} \cdot (0{,}01)^2 \doteq 0{,}0000340 = 3{,}40 \cdot 10^{-5}, \tag{4.22}$$

und wir sehen, dass der Interpolationsfehler (4.21) sehr wohl innerhalb der durch (4.22) gegebenen absoluten Fehlerschranke liegt. ♠

Beispiel 4.17. (Interpolationsfehler bei quadratischer Interpolation)
Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und seien $x_0, x_1, x_2 \in [0; 1]$ paarweise verschieden. Nach (4.13) gibt es dann zu jedem $x \in [0; 1]$ ein c_x zwischen dem Minimum und dem Maximum von x_0, x_1, x_2 und x mit

$$\begin{aligned} e^x - P_2(x) &= \frac{(x - x_0)\,(x - x_1)\,(x - x_2)}{3!}\, f'''(c_x) \\ &= \frac{(x - x_0)\,(x - x_1)\,(x - x_2)}{6}\, e^{c_x}, \end{aligned} \tag{4.23}$$

wobei wir genutzt haben, dass $f'(x) = e^x$, $f'(x) = e^x$, $f'''(x) = e^x$ ist, und wobei $P_2 \in \mathbb{P}_2$ das (höchstens) quadratische Interpolationspolynom bzgl. der Datenpunkte $(x_0; e^{x_0})$, $(x_1; e^{x_1})$, $(x_2; e^{x_2})$ ist.

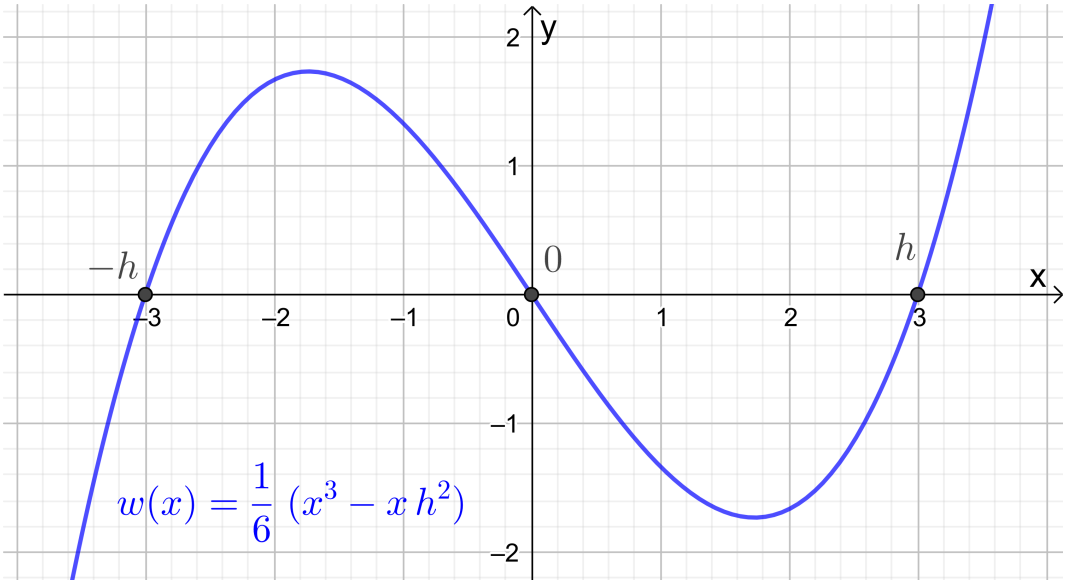


Abb. 4.5: Die Funktion $w(x) = \frac{(x+h)\,x\,(x-h)}{6} = \frac{x^3 - x\,h^2}{6}$ mit $h = 3$.

Wir nehmen nun an, dass $0 \leq x_0 < x_1 < x_2 \leq 1$ gilt und dass die drei Punkte x_0, x_1, x_2 jeweils gleiche Abstände zum Nachbarpunkt haben, also $h = x_2 - x_1 = x_1 - x_0$. Erfüllt x nun $0 \leq x_0 < x < x_2 \leq 1$, dann liefert das Anwenden des Absolutbetrags in (4.23)

$$\begin{aligned} |e^x - P_2(x)| &= \left| \frac{(x - x_0)\,(x - x_1)\,(x - x_2)}{6}\, e^{c_x} \right| \\ &\leq \left| \frac{(x - x_0)\,(x - x_1)\,(x - x_2)}{6} \right| e, \end{aligned} \tag{4.24}$$

wobei wir im zweiten Schritt $0 < e^{c_x} \leq e^1 = e$ genutzt haben. (Dieses folgt, weil c_x in $[x_0; x_2]$ und damit in $[0; 1]$ liegt und damit $e^{c_x} \leq e^1 = e$ gilt, da die Exponentialfunktion streng monoton wachsend ist.)

Um den Fehler weiter abzuschätzen, nutzen wir, dass gilt

$$\max_{x \in [x_0; x_2]} \left| \frac{(x - x_0)\,(x - x_1)\,(x - x_2)}{6} \right| = \frac{h^3}{9\sqrt{3}}\,. \tag{4.25}$$

Diese Abschätzung folgt, indem man zunächst x durch $x + x_1$ ersetzt, also

$$\begin{aligned} w(x) &= \frac{(x + x_1 - x_0)\,(x + x_1 - x_1)\,(x + x_1 - x_2)}{6} \\ &= \frac{(x + h)\,x\,(x - h)}{6} = \frac{1}{6}\,x\,(x^2 - h^2) = \frac{1}{6}\,(x^3 - x\,h^2)\,, \end{aligned}$$

wobei wir $x_1 - x_0 = x_2 - x_1 = h$ und danach die dritte binomische Formel $(x+h)(x-h) = x^2 - h^2$ genutzt haben. Das Intervall $[x_0; x_2]$ geht dann über in das Intervall $[x_0 - x_1; x_2 - x_1] = [-h; h]$. Mit

$$0 = w'(x) = \frac{3\,x^2 - h^2}{6} = \frac{1}{2}\left(x^2 - \frac{h^2}{3}\right) = \frac{1}{2}\left(x + \frac{h}{\sqrt{3}}\right)\left(x - \frac{h}{\sqrt{3}}\right)$$

sehen wir unter Berücksichtigung der Vorzeichenwechsel der Ableitung w', dass w auf $[-h; h]$ in $-\frac{h}{\sqrt{3}}$ sein globales Maximum und in $\frac{h}{\sqrt{3}}$ sein globales Minimum annimmt (siehe auch den Graphen in Abbildung 4.5). (Eigentlich folgt aus den Vorzeichenwechseln der Ableitung nur, dass lokale Extrema vorliegen, aber zusammen mit $w(-h) = w(h) = 0$ kann man folgern, dass dieses auch globale Extrema sind.) Also folgt

$$\begin{aligned}\max_{x\in[x_0;x_2]} \left|\frac{(x-x_0)(x-x_1)(x-x_2)}{6}\right| &= \max_{x\in[-h;h]} \left|\frac{1}{6}\,x\,(x^2-h^2)\right| \\ = \left|\frac{1}{6}\left(\pm\frac{h}{\sqrt{3}}\right)\left(\left(\pm\frac{h}{\sqrt{3}}\right)^2 - h^2\right)\right| &= \frac{1}{6}\left|\pm\frac{h}{\sqrt{3}}\left(-\frac{2\,h^2}{3}\right)\right| = \frac{1}{9\sqrt{3}}\,h^3.\end{aligned}$$

Wenden wir (4.25) in (4.24) an, so erhalten wir für $x \in [x_0; x_2]$

$$|e^x - P_2(x)| \le \left|\frac{(x-x_0)(x-x_1)(x-x_2)}{6}\right| e \le \frac{e}{9\sqrt{3}} \cdot h^3 \doteq 0{,}174 \cdot h^3. \quad (4.26)$$

Ist beispielsweise $h = 0{,}01$, dann gilt für $x \in [x_0; x_2] \subseteq [0; 1]$ mit $x_2 - x_1 = x_1 - x_0 = h = 0{,}01$

$$|e^x - P_2(x)| \le 0{,}174 \cdot (0{,}01)^3 = 1{,}74 \cdot 10^{-7}. \quad (4.27)$$

Vergleichen wir mit der Abschätzung (4.22) für den Interpolationsfehler der linearen Interpolierenden mit $x_1 - x_0 = h = 0{,}01$, so sehen wir, dass sich in (4.27) der Interpolationsfehler bei dem quadratischen Interpolationspolynom ungefähr um den Faktor $0{,}5 \cdot 10^{-2}$ verkleinert hat. ♠

Wir wollen nun noch eine weitere Darstellung für den Interpolationsfehler mit Hilfe der Interpolationsformel von Newton herleiten. Nach der Interpolationsformel von Newton ist das Polynom $P_{n+1} \in \mathbb{P}_{n+1}$ (also $\mathrm{Grad}(P_{n+1}) \le n+1$), welches $f : [a; b] \to \mathbb{R}$ an den paarweise verschiedenen Punkten $x_0, x_1, \ldots, x_{n+1} \in [a; b]$ interpoliert (also welches die Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, n+1$, interpoliert), durch

$$P_{n+1}(x) = P_n(x) + (x - x_0) \cdot \ldots \cdot (x - x_n)\, f[x_0, x_1, \ldots, x_n, x_{n+1}] \quad (4.28)$$

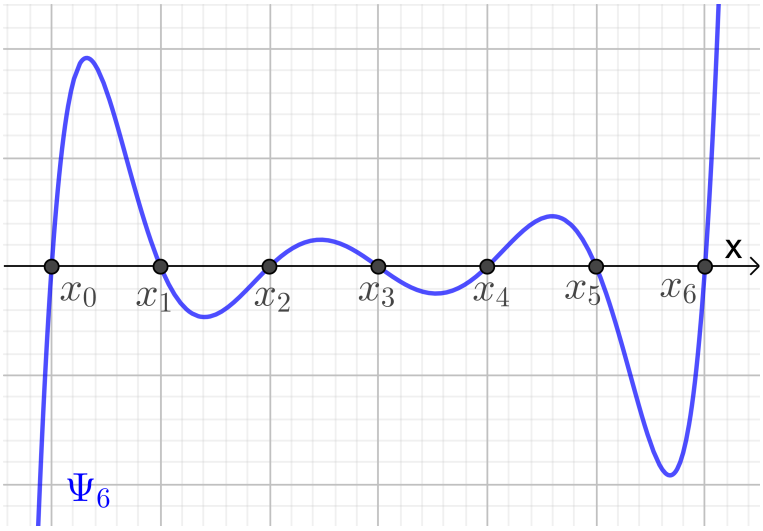


Abb. 4.6: Die Funktion Ψ_6 (siehe (4.32)) für $n+1 = 7$ Punkte $x_0 < x_1 < \ldots < x_6$ mit gleichen Abständen.

gegeben. Da P_{n+1} die Funktion f in x_{n+1} interpoliert gilt $f(x_{n+1}) = P_{n+1}(x_{n+1})$. Setzen wir in (4.28) $x = x_{n+1}$, so bekommen wir mit $f(x_{n+1}) = P_{n+1}(x_{n+1})$

$$f(x_{n+1}) = P_n(x_{n+1}) + (x_{n+1} - x_0) \cdot \ldots \cdot (x_{n+1} - x_n)\, f[x_0, x_1, \ldots, x_n, x_{n+1}]. \tag{4.29}$$

Wir wollen nun x_{n+1} als Variable behandeln und benennen es daher in t um. Dann wird (4.29) zu

$$\begin{aligned} f(t) = P_n(t) + (t - x_0) \cdot \ldots \cdot (t - x_n)\, f[x_0, x_1, \ldots, x_n, t] \quad &\Longleftrightarrow \\ f(t) - P_n(t) = (t - x_0) \cdot \ldots \cdot (t - x_n)\, f[x_0, x_1, \ldots, x_n, t]. \quad &(4.30) \end{aligned}$$

Die Formel (4.30) liefert eine Darstellung des Interpolationsfehlers von $P_n(t)$. Also gilt für $t \notin \{x_0, x_1, \ldots, x_n\}$

$$\boxed{f(t) - P_n(t) = (t - x_0) \cdot \ldots \cdot (t - x_n)\, f[x_0, x_1, \ldots, x_n, t].} \tag{4.31}$$

Man kann die Definition der dividierten Differenzen so verallgemeinern, dass auch gleiche Punkte zugelassen sind, und (4.31) gilt dann für alle $t \in [a; b]$. Natürlich werden die rechte und die linke Seite in (4.31) jeweils null, wenn $t = x_i$ für ein $i \in \{0; 1; 2; \ldots; n\}$ ist.

Sowohl in der Darstellung (4.13) in Satz 4.15 als auch in der Darstellung (4.31) des Interpolationsfehlers tritt das Polynom

$$\Psi_n(x) = (x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_n) \tag{4.32}$$

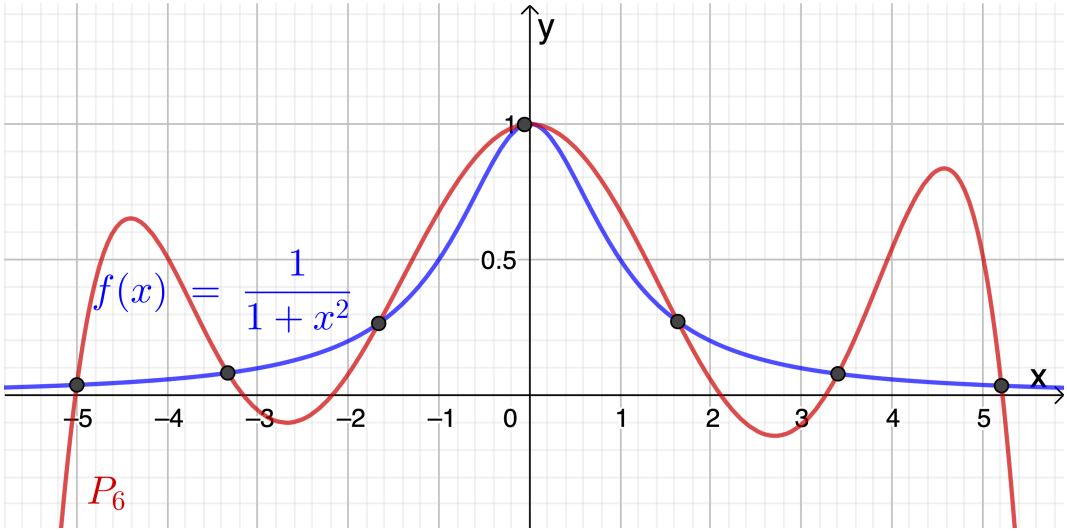


Abb. 4.7: Interpolierendes Polynom P_6 von $f(x) = \frac{1}{1+x^2}$ für Datenpunkte vom Graphen von f mit $-5 = x_0 < x_1 < \ldots < x_6 = 5$ mit gleichen Abstand.

vom Grad $n+1$ auf. Dieses ist der wichtigste Faktor, wenn wir den Fehler betrachten. In Abbildung 4.6 haben wir Ψ_6 für paarweise verschiedene Punkte $x_0 < x_1 < \ldots < x_6$ gezeichnet, die jeweils den gleichen Abstand haben. Man sieht direkt, dass $\Psi_6(x)$ und damit auch der Interpolationsfehler massiv größer werden, wenn man sich den Enden des Intervalls $[x_0; x_6]$ nähert. Wir vermuten daher, dass Interpolationspunkte mit gleichen Abständen vermutlich keine besonders gute Wahl sind.

Zum Abschluss betrachten wir noch ein Beispiel, an dem man sieht, dass das interpolierende Polynom P_n einer Funktion f bzgl. der Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, n$, (mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$) mit wachsendem n nicht immer gegen die Funktion f strebt.

Beispiel 4.18. (interpolierendes Polynom strebt nicht gegen f)
Gegeben sei die rationale Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \frac{1}{1+x^2}\,.$$

Wir betrachten das interpolierende Polynom $P_n \in \mathbb{P}_n$ bzgl. der Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, n$, wobei die $-5 = x_0 < x_1 < \ldots < x_n = 5$ gleiche Abstände haben, also

$$x_i = -5 + i\,h, \qquad i = 0, 1, \ldots, n, \qquad \text{wobei} \quad h = \frac{10}{n}\,.$$

Dann strebt $P_n(x)$ in vielen Punkten $x \in [-5; 5]$ mit wachsendem n nicht gegen

$f(x)$. Insbesondere für x mit $|x| > 4$ (also $x < -4$ oder $x > 4$) ist die Annäherung von $f(x)$ durch $P_n(x)$ sehr schlecht. In Abbildung 4.7 ist dieses für P_6 illustriert. Mit wachsendem n wird dieser Effekt noch schlimmer. ♠

KAPITEL 5

Numerische Integration

In diesem Kapitel lernen wir Verfahren zur numerischen Berechnung von Integralen, also zur **numerischen Integration** oder **Quadratur** kennen. Wir haben bereits in Teilkapitel 1.5 Integrale numerisch berechnet, und dabei wurde der Integrand durch ein geeignetes Taylor-Polynom des Integranden ersetzt. Dieses liefert aber in der Regel eine deutlich schlechtere angenäherte Berechnung des Integrals als wir sie mit den in diesem Kapitel besprochenen Verfahren bekommen.

Warum benötigt man numerische Integrationsverfahren? Viele Integrale, wie z.B.

$$\int_0^1 e^{x^2}\,\mathrm{d}x \qquad \text{oder} \qquad \int_0^\pi x^\pi \sin\left(\sqrt{x}\right)\mathrm{d}x$$

sind nicht (mit Hilfe des Hauptsatzes der Differential- und Integralrechnung) elementar berechenbar, weil die Integranden keine (elementar berechenbare) Stammfunktion haben. In diesem Fall braucht man eine numerische Integrationsformel, um das Integral angenähert zu berechnen.

5.1 Trapezregel

Die Grundidee zur Konstruktion **numerischer Integrationsformeln** oder **Quadraturformeln** ist, den **stetigen Integranden** f im bestimmten Integral

$$I(f) = \int_a^b f(x)\,\mathrm{d}x \tag{5.1}$$

durch eine geeignete Approximation zu ersetzen, die leicht exakt zu integrieren ist. Wir beginnen in diesem Teilkapitel mit der Herleitung der einfachsten elementaren Quadraturformel, nämlich der Trapezregel.

Um die Trapezregel zu bekommen, ersetzen wir den stetigen Integranden f in (5.1) durch sein lineares Interpolationspolynom $P_1 \in \mathbb{P}_1$ bzgl. der Datenpunkte $\big(a; f(a)\big)$ und $\big(b; f(b)\big)$ von f an den Intervallenden:

$$P_1(x) = f(a)\,\frac{x-b}{a-b} + f(b)\,\frac{x-a}{b-a} \tag{5.2}$$

Einsetzen von (5.2) in (5.1) und anschließendes Berechnen des Integrals über das Interpolationspolynom P_1 liefern

$$\begin{aligned}
\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b P_1(x)\,\mathrm{d}x &= \int_a^b \left(f(a)\,\frac{x-b}{a-b} + f(b)\,\frac{x-a}{b-a} \right) \mathrm{d}x \\
&= \left[f(a)\,\frac{(x-b)^2}{2\,(a-b)} + f(b)\,\frac{(x-a)^2}{2\,(b-a)} \right]_{x=a}^{x=b} \\
&= \left[0 + f(b)\,\frac{b-a}{2} \right] - \left[f(a)\,\frac{a-b}{2} + 0 \right] \\
&= f(b)\,\frac{b-a}{2} + f(a)\,\frac{-(a-b)}{2} = \frac{f(a)+f(b)}{2}\,(b-a).
\end{aligned}$$

Wir erhalten also die sogenannte **Trapezregel**

$$\boxed{T_1(f) = (b-a)\,\frac{f(a)+f(b)}{2}\,,}$$

die nun (ohne vorherige Berechnung des linearen Interpolationspolynoms) direkt bei Kenntnis der Funktionswerte $f(a)$ und $f(b)$ angewendet werden kann.

Verfahren 5.1. (Trapezregel)
Sei $f : [a; b] \to \mathbb{R}$ eine stetige Funktion. Die ***Trapezregel***

$$T_1(f) = (b-a)\,\frac{f(a)+f(b)}{2} \tag{5.3}$$

liefert eine Nährung für das Integral $I(f) = \displaystyle\int_a^b f(x)\,\mathrm{d}x$.

Die Trapezregel hat eine geometrische Anschauung, der sie Ihren Namen verdankt (siehe Abbildung 5.1): Die Fläche zwischen dem Graphen des linearen Interpolationspolynoms $P_1(x)$ und der x-Achse von $x = a$ bis $x = b$ bildet ein Trapez.

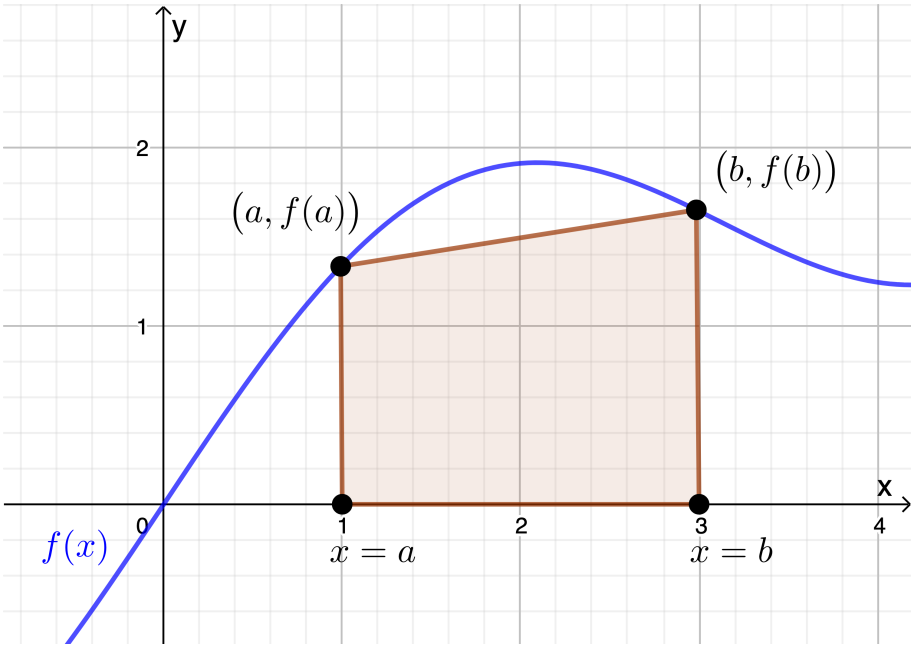


Abb. 5.1: Veranschaulichung der Trapezregel T_1 aus Verfahren 5.1.

Beispiel 5.2. (Trapezregel)
Wir wollen das Integral

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x$$

mit der Trapezregel angenähert berechnen:

$$T_1(f) = (1-0)\,\frac{f(0)+f(1)}{2} = \frac{1}{2}\left(\frac{1}{1+0} + \frac{1}{1+1}\right) = \frac{3}{4} = 0{,}75.$$

Wir berechnen das Integral direkt (mit Rundung des Ergebnisses auf 10-stellige Gleitkommadarstellung)

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \Big[\ln(1+x)\Big]_{x=0}^{x=1} = \ln(2) - \ln(1) = \ln(2) \doteq 0{,}6931471806.$$

Der absolute Fehler der Näherung ist (mit Rundung auf 3-stellige Gleitkommadarstellung) $I(f) - T_1(f) = \ln(2) - \frac{3}{4} \doteq -0{,}0569$. ♠

Hilfssatz 5.3. (Trapezregel ist exakt für Polynome vom Grad ≤ 1)
Für jedes Polynom $p_1(x) = c\,x + d$ mit Konstanten $c, d \in \mathbb{R}$ gilt für die durch (5.3) definierte Trapezregel $T_1(p_1) = I(p_1)$, d.h. die Trapezregel ***integriert alle Polynome von Grad ≤ 1 (also in $\mathbb{P}_1$)*** *exakt.*

Beweis von Hilfssatz 5.3: Dieses zeigt man, indem man $T_1(p_1)$ und $I(p_1)$ konkret berechnet und umformt, bis man sieht, dass sie gleich sind. – Alternativ kann man dieses auch mit Hilfe der Eigenschaften des linearen Interpolationspolynoms begründen. – Wir führen den Beweis auf einem Übungsblatt durch. □

Wie kann man mit Hilfe der Trapezregel eine bessere numerische Integrationsformel bekommen? Man könnte das **Integrationsintervall** $[a;b]$ **in mehrere gleichlange Teilintervalle zerlegen** und dann **auf jedem Teilintervall die Trapezregel nutzen**. Betrachten wir dieses zunächst für ein Beispiel.

Beispiel 5.4. („zusammengesetzte" Trapezregel)
Wir wollen das Integral

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \ln(2) \doteq 0{,}6931471806$$

in zwei Teilintegrale über $\left[0;\frac{1}{2}\right]$ und $\left[\frac{1}{2};1\right]$ zerlegen und diese dann jeweils mit der Trapezregel angenähert berechnen:

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \int_0^{1/2} \frac{1}{1+x}\,\mathrm{d}x + \int_{1/2}^1 \frac{1}{1+x}\,\mathrm{d}x,$$

$$\begin{aligned} T_2(f) &= \left(\frac{1}{2}-0\right)\frac{f(0)+f\left(\frac{1}{2}\right)}{2} + \left(1-\frac{1}{2}\right)\frac{f\left(\frac{1}{2}\right)+f(1)}{2} \\ &= \frac{1}{2}\cdot\frac{1+\frac{2}{3}}{2} + \frac{1}{2}\cdot\frac{\frac{2}{3}+\frac{1}{2}}{2} = \frac{1}{4}\cdot\frac{5}{3} + \frac{1}{4}\cdot\frac{7}{6} = \frac{10}{24}+\frac{7}{24} = \frac{17}{24} \doteq 0{,}70833, \end{aligned}$$

wobei das Ergebnis auf 5-stellige Gleitkommadarstellung gerundet wurde. Der absolute Fehler dieser Näherung ist mit Rundung auf 3-stellige Gleitkommadarstellung $I(f) - T_2(f) = \ln(2) - \frac{17}{24} \doteq -0{,}0152$. Gegenüber dem Ergebnis $T_1(f)$ aus Beispiel 5.2 beträgt der absolute Fehler von $T_2(f)$ ungefähr nur $\frac{1}{4}$ des absoluten Fehlers von $T_1(f)$. ♠

Wir wollen die Idee aus dem letzten Beispiel nun verallgemeinern und das **Integrationsintervall** $[a;b]$ **in** n **gleich große Teilintervalle unterteilen und auf jedem Teilintervall die Trapezregel anwenden**: Sei also

$$h = \frac{b-a}{n}\,.$$

Dann erhält man mit

$$x_k = a + k\,h, \qquad k = 0, 1, 2, \ldots, n,$$

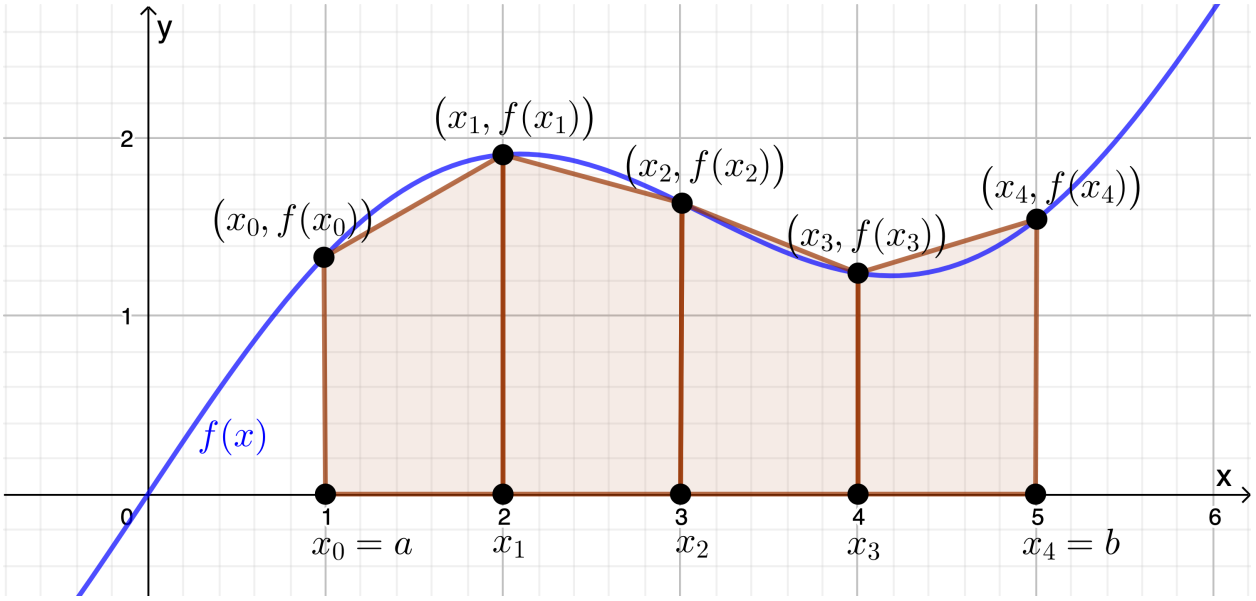


Abb. 5.2: Veranschaulichung der Trapezregel T_n für numerische Integration aus Verfahren 5.5 (hier mit einer Unterteilung in $n = 4$ gleich lange Teilintervalle).

durch $[x_{k-1}; x_k]$, $k = 1, 2, \ldots, n$, eine Unterteilung von $[a; b]$ in n gleich große Teilintervalle der Länge h. Das Integral (5.1) ist dann (mit $x_0 = a$ und $x_n = b$) entsprechend die Summe der Integrale über diese Teilintervalle

$$\begin{aligned} I(f) &= \int_a^b f(x)\,\mathrm{d}x = \int_{x_0}^{x_n} f(x)\,\mathrm{d}x \\ &= \int_{x_0}^{x_1} f(x)\,\mathrm{d}x + \int_{x_1}^{x_2} f(x)\,\mathrm{d}x + \ldots + \int_{x_{n-1}}^{x_n} f(x)\,\mathrm{d}x. \end{aligned}$$

Wir nutzen nun auf jedem dieser Teilintervalle der Länge h die Trapezregel (5.3) und erhalten damit

$$\begin{aligned} I(h) &\approx h\,\frac{f(x_0) + f(x_1)}{2} + h\,\frac{f(x_1) + f(x_2)}{2} + \ldots + h\,\frac{f(x_{n-1}) + f(x_n)}{2} \\ &= \frac{h}{2}\left[\big(f(x_0) + f(x_1)\big) + \big(f(x_1) + f(x_2)\big) + \ldots + \big(f(x_{n-1}) + f(x_n)\big)\right] \\ &= h\left[\frac{1}{2}\,f(x_0) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}\,f(x_n)\right]. \end{aligned}$$

Dieses ist die **Trapezregel für numerische Integration**, und wir halten diese als Verfahren fest. Die Trapezregel für numerische Integration ist in Abbildung 5.2 mit $n = 4$ illustriert.

Verfahren 5.5. (Trapezregel für numerische Integration)
*Sei $f : [a;b] \to \mathbb{R}$ eine stetige Funktion. Für $n \in \mathbb{N}$ sei $h = \frac{1}{n}(b-a)$ und die **Knoten(punkte)** seien $x_k = a + k\,h$, $k = 0, 1, 2, \ldots, n$. Dann liefert die **Trapezregel für numerische Integration***

$$T_n(f) = h\left[\frac{1}{2}f(x_0) + \sum_{k=1}^{n-1} f(x_k) + \frac{1}{2}f(x_n)\right] \tag{5.4}$$
$$= h\left[\frac{1}{2}f(x_0) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(x_n)\right]$$

eine Näherung für das Integral $I(f) = \int_a^b f(x)\,\mathrm{d}x$.

Der Index n von $T_n(f)$ steht für die Anzahl der gleichlangen Teilintervalle, in die $[a;b]$ unterteilt wurde. Natürlich ist die „einfache Trapezregel“ in Verfahren 5.1 der Sonderfall $n = 1$ von Verfahren 5.5.

Aus Hilfssatz 5.3 folgt direkt, dass die Trapezregel T_n für numerische Integration für alle Polynome vom Grad ≤ 1 exakt ist.

Satz 5.6. (Trapezregel für numerische Integration ist exakt auf $\mathbb{P}_1$)
*Für jedes Polynom $p_1 \in \mathbb{P}_1$, also $p_1(x) = c\,x + d$ mit Konstanten $c, d \in \mathbb{R}$, gilt für die durch (5.4) definierte Trapezregel für numerische Integration $T_n(p_1) = I(p_1)$, d.h. die Trapezregel T_n für numerische Integration **integriert alle Polynome in $\mathbb{P}_1$ (also von Grad ≤ 1) exakt**.*

Beweis von Satz 5.6 Sei $p_1 \in \mathbb{P}_1$ ein beliebiges Polynom vom Grad ≤ 1. Dann gilt nach der Konstruktion der Trapezregel T_n

$$T_n(p_1) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} P_{1,k}(x)\,\mathrm{d}x$$
$$= \int_{x_0}^{x_1} P_{1,1}(x)\,\mathrm{d}x + \int_{x_1}^{x_2} P_{1,2}(x)\,\mathrm{d}x + \ldots + \int_{x_{n-1}}^{x_n} P_{1,n}(x)\,\mathrm{d}x,$$

wobei für $k = 1, 2, \ldots, n$, das Polynom $P_{1,k} \in \mathbb{P}_1$ jeweils das lineare Interpolationspolynom von p_1 bzgl. der beiden Datenpunkte $\big(x_{k-1}; p_1(x_{k-1})\big)$ und $\big(x_k, p_1(x_k)\big)$ ist. Da $P_{1,k} \in \mathbb{P}_1$ für $k = 1, 2, \ldots, n$, jeweils die zwei Interpolationsbedingungen $P_{1,k}(x_{k-1}) = p_1(x_{k-1})$ und $P_{1,k}(x_k) = p_1(x_k)$ erfüllt und da p_1 ebenfalls ein

n	$h = \frac{b-a}{n} = \frac{1}{n}$	$T_n(f)$	$I(f) - T_n(f)$	$\frac{\lvert I(f) - T_{n/2}(f)\rvert}{\lvert I(f) - T_n(f)\rvert}$
$2 = 2^1$	0,5	0,731370252	$1{,}55 \cdot 10^{-2}$	
$4 = 2^2$	0,25	0,742984098	$3{,}84 \cdot 10^{-3}$	4,02
$8 = 2^3$	0,125	0,745865615	$9{,}59 \cdot 10^{-4}$	4,01
$16 = 2^4$	0,0625	0,746584597	$2{,}40 \cdot 10^{-4}$	4,00
$32 = 2^5$	0,03125	0,746764255	$5{,}99 \cdot 10^{-5}$	4,00
$64 = 2^6$	0,015625	0,746809164	$1{,}50 \cdot 10^{-5}$	4,00
$128 = 2^7$	0,0078125	0,746820391	$3{,}74 \cdot 10^{-6}$	4,00

Tabelle 5.1: Trapezregel $T_n(f)$ mit $f(x) = e^{-x^2}$ und $[a; b] = [0; 1]$ zur Berechnung des Integrals $\int_0^1 e^{-x^2}\,\mathrm{d}x$.

Polynom vom Grad ≤ 1 ist, welches die Interpolationsbedingungen trivialerweise erfüllt, folgt aus der Eindeutigkeit des linearen Interpolationspolynoms, dass $P_{1,k} = p_1$ gelten muss. Also folgt

$$T_n(p_1) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} P_{1,k}(x)\,\mathrm{d}x = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} p_1(x)\,\mathrm{d}x = \int_a^b p_1(x)\,\mathrm{d}x = I(p_1),$$

und die Trapezregel T_n integriert p_1 offensichtlich exakt. □

Betrachten wir ein Beispiel.

Beispiel 5.7. (Trapezregel für numerische Integration)
Wir berechnen das Integral

$$\int_0^1 e^{-x^2}\,\mathrm{d}x \doteq 0{,}746824132812427$$

für $n = 2^j$ mit $j = 1, 2, \ldots, 7$ mit der Trapezregel für numerische Integration T_n. Es sind also $[a; b] = [0; 1]$, $f(x) = e^{-x^2}$ und $h = (1-0)/n = 1/n$, und die Knoten sind dann $x_k = a + k\,h = 0 + k \cdot \frac{1}{n} = \frac{k}{n}$, $k = 0, 1, 2, \ldots, n$. Wir müssen also für $n = 2^j$ mit $j = 1, 2, \ldots, 7$ jeweils die folgende gewichtete Summe berechnen:

$$T_n(f) = \frac{1}{n}\left[\frac{e^{-0^2}}{2} + \sum_{k=1}^{n-1} e^{-(k/n)^2} + \frac{e^{-1^2}}{2}\right] = \frac{1}{n}\left[\frac{1}{2} + \sum_{k=1}^{n-1} e^{-(k/n)^2} + \frac{e^{-1}}{2}\right]$$

Die Ergebnisse sind in Tabelle 5.1 auf eine 9-stellige Gleitkommadarstellung gerundet angegeben. Weiter haben wir den absoluten Fehler $I(f) - T_n(f)$ auf eine 3-stellige Gleitkommadarstellung gerundet aufgelistet, sowie den Quotienten $|I(f) - T_{n/2}(f)|/|I(f) - T_n(f)|$.

An den Quotienten $|I(f) - T_{n/2}(f)|/|I(f) - T_n(f)|$ sehen wir, dass gilt

$$\frac{|I(f) - T_{n/2}(f)|}{|I(f) - T_n(f)|} \approx 4 \qquad \Longleftrightarrow \qquad \frac{1}{4}\,|I(f) - T_{n/2}(f)| \approx |I(f) - T_n(f)|.$$

Dieses bedeutet, dass bei einer Verdoppelung der Anzahl der Teilintervalle und damit einer Halbierung des Abstands der Knotenpunkte x_k, $k = 0, 1, \ldots, n$, (und ungefähr einer Verdoppelung der Anzahl der Knotenpunkte) der Betrag des absoluten numerischen Integrationsfehler $|I(f) - T_n(f)|$ ungefähr $1/4$ des Betrags des absoluten numerischen Integrationsfehler $|I(f) - T_{n/2}(f)|$ mit dem doppelten Abstand $2h$ ist. Daher vermuten wir, dass sich der Betrag des absoluten numerischen Integrationsfehler $|I(f) - T_n(f)|$ möglicherweise wie $C\,h^2$ mit einer Konstante $C > 0$ verhält, denn: Gelten

$$|I(f) - T_n(f)| \approx C\,h^2 \quad \text{und entsprechend} \quad |I(f) - T_{n/2}(f)| \approx C\,(2h)^2 = 4\,C\,h^2,$$

so folgt angenähert

$$\frac{|I(f) - T_{n/2}(f)|}{|I(f) - T_n(f)|} \approx \frac{4\,C\,h^2}{C\,h^2} = 4,$$

wie wir es in Tabelle 5.1 beobachtet haben. ♠

Wie der nachfolgende Satz zeigt, ist das im vorigen Beispiel vermutete Konvergenzverhalten $|I(f) - T_n(f)| \approx C\,h^2$ (mit einer Konstante $C > 0$) richtig, bzw. es liegt mindestens ein solches Konvergenzverhalten vor.

Satz 5.8. (Konvergenz der Trapezregel für numerische Integration)
Sei $n \in \mathbb{N}$, und sei $f : [a; b] \to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion. Dann gilt für den ***Betrag des absoluten Fehlers*** $E_n^T(f) = |I(f) - T_n(f)|$ *der Näherung des Integrals*

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

durch die in Verfahren 5.5 definierte ***Trapezregel für numerische Integration*** $T_n(f)$ *die Fehlerabschätzung*

$$E_n^T(f) = |I(f) - T_n(f)| \leq \frac{b-a}{12}\,h^2 \max_{x \in [a;b]} |f''(x)|. \tag{5.5}$$

Warum ist Satz 5.8 extrem nützlich? Satz 5.8 erlaubt es uns, mit Hilfe von (5.5) auszurechnen, ab welchem Wert für n der absolute Fehler $I(f)-T_n(f)$ betraglich garantiert höchstens so groß wie eine vorgegebene absolute Fehlerschranke ε ist. Wir demonstrieren dieses an einem Beispiel.

Beispiel 5.9. (Fehlerabschätzung für die Trapezregel für num. Int.)
Gesucht ist ein (möglichst kleiner) Wert für n, ab dem $T_n(f)$ als Näherung für

$$I(f) = \int_0^1 e^{-x^2}\,\mathrm{d}x \doteq 0{,}746824132812427 \qquad \text{mit} \quad f : \mathbb{R} \to \mathbb{R}, \quad f(x) = e^{-x^2},$$

einen absoluten Fehler hat, der betraglich garantiert höchstens $\varepsilon = 10^{-3}$ ist.

Mit $[a;b] = [0;1]$ und $h = (b-a)/n = (1-0)/n = \frac{1}{n}$ und

$$\begin{aligned} f'(x) &= -2\,x\,e^{-x^2}, \\ f''(x) &= -2\,e^{-x^2} + (-2\,x)^2\,e^{-x^2} = \big(4\,x^2 - 2\big)\,e^{-x^2}, \end{aligned}$$

gilt nach (5.5) in Satz 5.8 die Fehlerabschätzung

$$E_n^T(f) = |I(f) - T_n(f)| \leq \frac{1-0}{12}\left(\frac{1}{n}\right)^2 \max_{x\in[0;1]} \Big|\big(4\,x^2-2\big)\,e^{-x^2}\Big|. \tag{5.6}$$

Um das Maximum von $|f''(x)|$ auf $[0;1]$ zu bestimmen, berechnen wir f''':

$$\begin{aligned} f'''(x) &= 8\,x\,e^{-x^2} + \big(4\,x^2-2\big)\,(-2\,x)\,e^{-x^2} = \big(-8\,x^3 + 12\,x\big)\,e^{-x^2} \\ &= -8\,x\left(x^2 - \frac{3}{2}\right) = -8\,x\left(x - \frac{\sqrt{3}}{\sqrt{2}}\right)\left(x + \frac{\sqrt{3}}{\sqrt{2}}\right) \end{aligned}$$

Für $x \in [0;1]$ ist $-8\,x \leq 0$, $x - \frac{\sqrt{3}}{\sqrt{2}} < 0$ und $x + \frac{\sqrt{3}}{\sqrt{2}} > 0$. Also ist $f'''(x) \geq 0$ für alle $x \in [0;1]$, und wir sehen, dass f'' monoton wachsend ist. Somit gilt für $x \in [0;1]$, dass $f''(0) \leq f''(x) \leq f''(1)$ ist, und es folgt

$$\begin{aligned} \max_{x\in[0;1]} \Big|\big(4\,x^2-2\big)\,e^{-x^2}\Big| &= \max_{x\in[0;1]} |f''(x)| = \max\big\{|f''(0)|, |f''(1)|\big\} \\ &= \max\big\{|-2\,e^0|; |2\,e^{-1}|\big\} = \max\big\{2; 2\,e^{-1}\big\} = 2. \end{aligned}$$

Einsetzen in (5.6) liefert

$$E_n^T(f) = |I(f) - T_n(f)| \leq \frac{1-0}{12}\left(\frac{1}{n}\right)^2 \cdot 2 = \frac{1}{6}\cdot\frac{1}{n^2}. \tag{5.7}$$

Wir suchen nun $n \in \mathbb{N}$, so dass gilt

$$E_n^T(f) = |I(f) - T_n(f)| \leq \frac{1}{6} \cdot \frac{1}{n^2} \leq \varepsilon = 10^{-3}. \tag{5.8}$$

Auflösen der Bedingung $\frac{1}{6} \cdot \frac{1}{n^2} \leq \varepsilon = 10^{-3}$ aus (5.8) liefert

$$\frac{1}{6} \cdot \frac{1}{n^2} \leq 10^{-3} \qquad \Longleftrightarrow \qquad \frac{10^3}{6} \leq n^2 \qquad \Longleftrightarrow \qquad \underbrace{\sqrt{\frac{10^3}{6}}}_{\doteq 12{,}9} \leq n,$$

wobei wir im letzten Schritt genutzt haben, dass die Quadratwurzel streng monoton wachsend ist und deshalb die $\leq$ Beziehung erhalten bleibt. Also ist für $n \geq 13$ garantiert $E_n^T(f) = |I(f) - T_n(f)| \leq \varepsilon = 10^{-3}$. Ein Vergleich mit Tabelle 5.1, in der die Näherungen $T_n(f)$ für $n = 2^j$, $j = 1, 2, \ldots, 7$, für das obige Integral $I(f)$ zusammen mit ihren absoluten Fehlern angegeben wurden, zeigt, dass bereits für $n = 8$ gilt $E_n^T(f) = |I(f) - T_n(f)| \leq \varepsilon = 10^{-3}$. Die Abschätzung (5.7) liefert hier also keine gute Prognose. ♠

Für mathematisch Interessierte zeigen wir noch den Beweis von Satz 5.8.

Beweis von Satz 5.8: Wir erinnern uns, dass wir die Trapezregel $T_1(f)$ zu Beginn des Kapitels als das Integral über das lineare Interpolationspolynom P_1 bzgl. der Datenpunkte $\big(a; f(a)\big)$, $\big(b; f(b)\big)$ eingeführt haben. Entsprechend gilt für das Teilintervall $[x_{k-1}; x_k]$, $k \in \{1, 2, \ldots, n\}$, bei der Trapezregel für numerische Integration $T_n(f)$ also

$$\int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x \approx \int_{x_{k-1}}^{x_k} P_{1,k}(x)\,\mathrm{d}x \qquad \text{mit dem linearen Interpolationspolynom}$$

$$P_{1,k}(x) = f(x_{k-1})\,\frac{x - x_k}{x_{k-1} - x_k} + f(x_k)\,\frac{x - x_{k-1}}{x_k - x_{k-1}} \quad \text{bzgl.} \quad \begin{cases} \big(x_{k-1}; f(x_{k-1})\big), \\ \big(x_k; f(x_k)\big), \end{cases}$$

wobei das Integral auf der rechten Seite gerade der Beitrag zu $T_n(f)$ durch die Trapezregel auf dem Teilintervall $[x_{k-1}; x_k]$ ist. Es gilt mit der Dreiecksungleichung

$$\begin{aligned} E_n^T(f) = |I(f) - T_n(f)| &= \left| \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x - \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} P_{1,k}(x)\,\mathrm{d}x \right| \\ &= \left| \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \big[f(x) - P_{1,k}(x)\big]\,\mathrm{d}x \right| \leq \sum_{k=1}^{n} \left| \int_{x_{k-1}}^{x_k} \big[f(x) - P_{1,k}(x)\big]\,\mathrm{d}x \right| \\ &\leq \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \big|f(x) - P_{1,k}(x)\big|\,\mathrm{d}x. \end{aligned} \tag{5.9}$$

Die Integrale in der dritten Zeile von (5.9) sind alle von der Form

$$\int_{\alpha}^{\alpha+h} \big|f(x) - P_1(x)\big| \, \mathrm{d}x \quad \text{mit} \quad P_1(x) = f(\alpha)\,\frac{(\alpha+h)-x}{h} + f(\alpha+h)\,\frac{x-\alpha}{h}\,.$$

Nach Satz 4.15 gilt (da f zweimal stetig differenzierbar ist) mit einem (von $x \in [\alpha;\alpha+h]$ abhängigen) Punkt c_x in $[\alpha;\alpha+h]$

$$\begin{aligned} |f(x) - P_1(x)| &= \left| \frac{(x-\alpha)\big(x-(\alpha+h)\big)}{2!}\, f''(c_x) \right| \\ &= \frac{|f''(c_x)|}{2} \underbrace{(x-\alpha)}_{\geq 0} \underbrace{\big((\alpha+h)-x\big)}_{\geq 0} \\ &\leq \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)| \right) (x-\alpha)\big((\alpha+h)-x\big). \end{aligned} \tag{5.10}$$

Wir formen nun $(x-\alpha)\big((\alpha+h)-x\big)$ geeignet um:

$$\begin{aligned} (x-\alpha)\big((\alpha+h)-x\big) &= \left(x - \left(\alpha+\frac{h}{2}\right) + \frac{h}{2}\right) \left(\left(\alpha+\frac{h}{2}\right) - x + \frac{h}{2}\right) \\ &= \left(\frac{h}{2} + \left[x - \left(\alpha+\frac{h}{2}\right)\right]\right) \left(\frac{h}{2} - \left[x - \left(\alpha+\frac{h}{2}\right)\right]\right) \\ &= \left(\frac{h}{2}\right)^2 - \left[x - \left(\alpha+\frac{h}{2}\right)\right]^2 = \frac{h^2}{4} - \left[x - \left(\alpha+\frac{h}{2}\right)\right]^2, \end{aligned} \tag{5.11}$$

Einsetzen von (5.11) in (5.10) und Integrieren über $[\alpha;\alpha+h]$ ergibt:

$$\begin{aligned} &\int_{\alpha}^{\alpha+h} \big|f(x) - P_1(x)\big| \, \mathrm{d}x \\ &\leq \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)| \right) \int_{\alpha}^{\alpha+h} \left(\frac{h^2}{4} - \left[x - \left(\alpha+\frac{h}{2}\right)\right]^2 \right) \mathrm{d}x \\ &= \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)| \right) \left(\left[\frac{h^2}{4}\, x\right]_{x=\alpha}^{x=\alpha+h} - \left[\frac{1}{3}\left(x - \left(\alpha+\frac{h}{2}\right)\right)^3\right]_{x=\alpha}^{x=\alpha+h} \right) \\ &= \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)| \right) \left(\left[\frac{h^2}{4}\big((\alpha+h)-\alpha\big)\right] - \left[\frac{1}{3}\left(\frac{h}{2}\right)^3 - \frac{1}{3}\left(-\frac{h}{2}\right)^3\right] \right) \\ &= \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)| \right) \left(\frac{1}{4}\,h^3 - \frac{1}{12}\,h^3 \right) = \frac{1}{12}\,h^3 \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)| \right) \end{aligned} \tag{5.12}$$

Anwenden der Abschätzung (5.12) auf jedes der Integrale in (5.9) liefert

$$E_n^T(f) \leq \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \big|f(x) - P_{1,k}(x)\big| \,\mathrm{d}x \leq \sum_{k=1}^{n} \frac{1}{12}\, h^3 \left(\max_{t\in[x_{k-1};x_k]} |f''(t)| \right)$$
$$\leq \frac{1}{12}\, h^2 \left(\max_{t\in[a;b]} |f''(t)| \right) \underbrace{\sum_{k=1}^{n} h}_{\substack{=n\,h \\ =b-a}} = \frac{b-a}{12}\, h^2 \left(\max_{t\in[a;b]} |f''(t)| \right),$$

womit (5.5) bewiesen ist. □

5.2 Simpson-Regel

Um eine elementare, aber bessere, Quadraturformel als die Trapezregel für numerische Integration zu bekommen, gehen wir analog zur Herleitung der Trapezregel vor, aber ersetzen den Integranden durch ein (höchstens) quadratisches Interpolationspolynom aus $\mathbb{P}_2$. Genauer ersetzen wir in

$$I(f) = \int_a^b f(x)\,\mathrm{d}x \tag{5.13}$$

den stetigen Integranden f durch das (höchstens) quadratische Interpolationspolynom $P_2 \in \mathbb{P}_2$ bzgl. der Datenpunkte $\big(a; f(a)\big)$, $\big(c; f(c)\big)$, $\big(b; f(b)\big)$, wobei $c = (a+b)/2$ der Punkt genau in der Mitte zwischen a und b ist. Das quadratische Interpolationspolynom $P_2 \in \mathbb{P}_2$ hat dann die folgende Form:

$$P_2(x) = f(a)\,\frac{(x-b)\,(x-c)}{(a-b)\,(a-c)} + f(c)\,\frac{(x-a)\,(x-b)}{(c-a)\,(c-b)} + f(b)\,\frac{(x-a)\,(x-c)}{(b-a)\,(b-c)}$$

Also erhalten wir die folgende Näherungsformel für das Integral

$$I(f) \approx \int_a^b P_2(x)\,\mathrm{d}x = f(a) \int_a^b \frac{(x-b)\,(x-c)}{(a-b)\,(a-c)}\,\mathrm{d}x$$
$$+ f(c) \int_a^b \frac{(x-a)\,(x-b)}{(c-a)\,(c-b)}\,\mathrm{d}x + f(b) \int_a^b \frac{(x-a)\,(x-c)}{(b-a)\,(b-c)}\,\mathrm{d}x. \tag{5.14}$$

Die drei Integrale lassen sich exakt berechnen. Zuvor ist es aber zweckmäßig $h = (b-a)/2$ einzuführen. Dann gelten $h = c - a = b - c$ und $b - a = 2\,h$, und somit folgt auch $c = a + h$, $b = c + h = a + 2\,h$. Dann führen wir in jedem der drei Integrale die Substitution $t = x - a \iff x = t + a$, $\mathrm{d}t = \mathrm{d}x$, durch,

damit in dem Ergebnis nicht mehr a, b, c, sondern nur h auftaucht. Wir erhalten mit dieser Vorgehensweise mit $x-b = t+a-b = t-2\,h$, $x-c = t+a-c = t-h$ und $x-a = t$ jeweils

$$\int_a^b \frac{(x-b)\,(x-c)}{(a-b)\,(a-c)}\,\mathrm{d}x = \frac{1}{2\,h^2}\int_0^{2h} (t-2h)\,(t-h)\,\mathrm{d}t$$

$$= \frac{1}{2\,h^2}\int_0^{2h} \left(t^2 - 3\,h\,t + 2\,h^2\right)\mathrm{d}t = \frac{1}{2\,h^2}\left[\frac{1}{3}\,t^3 - \frac{3}{2}\,h\,t^2 + 2\,h^2\,t\right]_{t=0}^{t=2h}$$

$$= \frac{1}{2\,h^2}\left(\left[\frac{8}{3}\,h^3 - 6\,h^3 + 4\,h^3\right] - 0\right) = \frac{1}{2\,h^2}\cdot\frac{8-18+12}{3}\,h^3 = \frac{1}{2}\cdot\frac{2}{3}\,h = \frac{h}{3}\,,$$

$$\int_a^b \frac{(x-a)\,(x-b)}{(c-a)\,(c-b)}\,\mathrm{d}x = \frac{-1}{h^2}\int_0^{2h} t\,(t-2h)\,\mathrm{d}t = \frac{-1}{h^2}\int_0^{2h} \left(t^2 - 2\,h\,t\right)\mathrm{d}t$$

$$= \frac{-1}{h^2}\left[\frac{1}{3}\,t^3 - h\,t^2\right]_{t=0}^{t=2h} = \frac{-1}{h^2}\left(\left[\frac{8}{3}\,h^3 - 4\,h^3\right] - 0\right) = \frac{-1}{h^2}\cdot\frac{-4}{3}\,h^3 = \frac{4\,h}{3}\,,$$

$$\int_a^b \frac{(x-a)\,(x-c)}{(b-a)\,(b-c)}\,\mathrm{d}x = \frac{1}{2\,h^2}\int_0^{2h} t\,(t-h)\,\mathrm{d}t = \frac{1}{2\,h^2}\int_0^{2h} \left(t^2 - h\,t\right)\mathrm{d}t$$

$$= \frac{1}{2\,h^2}\left[\frac{1}{3}\,t^3 - \frac{1}{2}\,h\,t^2\right]_{t=0}^{t=2h} = \frac{1}{2\,h^2}\left(\left[\frac{8}{3}\,h^3 - 2\,h^3\right] - 0\right) = \frac{1}{2\,h^2}\cdot\frac{2}{3}\,h^3 = \frac{h}{3}\,.$$

Einsetzen der Werte für diese Integrale in (5.14) liefert die **Simpson-Regel**

$$I(f) \approx f(a)\cdot\frac{h}{3} + f(c)\cdot\frac{4\,h}{3} + f(b)\cdot\frac{h}{3} = \frac{h}{3}\left[f(a) + 4\,f(c) + f(b)\right]$$

$$= \boxed{\frac{h}{3}\left[f(a) + 4\,f\left(\frac{a+b}{2}\right) + f(b)\right] = S_2(f),} \tag{5.15}$$

wobei wir in der zweiten Zeile von (5.15) noch $c = (a+b)/2$ eingesetzt haben.

Wir halten die Simpson-Regel als Verfahren fest:

Verfahren 5.10. (Simpson-Regel)
Seien $f : [a;b] \to \mathbb{R}$ eine stetige Funktion und $h = \frac{b-a}{2}$. Die ***Simpson-Regel***

$$S_2(f) = \frac{h}{3}\left[f(a) + 4\,f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{5.16}$$

liefert eine Nährung für das Integral $I(f) = \displaystyle\int_a^b f(x)\,\mathrm{d}x$.

Betrachten wir zunächst ein Beispiel.

Beispiel 5.11. (Simpson-Regel)
Wir wollen das Integral

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \ln(2) \doteq 0{,}6931471806$$

(welches bereits in Beispiel 5.2 mit der Trapezregel und in Beispiel 5.4 mit der „zusammengesetzten" Trapezregel berechnet wurde) mit der Simpson-Regel angenähert berechnen: Es gilt $h = (1-0)/2 = \frac{1}{2}$ und damit $h/3 = 1/6$ und

$$S_2(f) = \frac{1}{6}\left[f(0) + 4\,f\left(\frac{1}{2}\right) + f(1)\right] = \frac{1}{6}\left[\frac{1}{1+0} + 4\cdot\frac{1}{1+\frac{1}{2}} + \frac{1}{1+1}\right]$$

$$= \frac{1}{6}\left[1 + \frac{8}{3} + \frac{1}{2}\right] = \frac{25}{36} \doteq 0{,}69444.$$

Der exakte Wert des Integrals ist (mit Rundung auf 10-stellige Gleitkommadarstellung) $I(f) = \ln(2) \doteq 0{,}6931471806$, und der absolute Fehler der Simpson-Regel ist auf eine 3-stellige Gleitkommadarstellung gerundet somit $I(f) - S_2(f) = \ln(2) - \frac{25}{36} \doteq -0{,}00130$. Wir sehen, dass der absolute Fehler deutlich kleiner ist als in Beispielen 5.2 und 5.4. Dabei ist der Vergleich mit Beispiel 5.4 eher angebracht, weil dort auch drei Knoten verwendet werden wie in der Simpson-Regel. Aber selbst verglichen mit Beispiel 5.4 ist der Fehler ungefähr um einen Faktor 10 kleiner. ♠

Analog zu Hilfssatz 5.3 gilt der folgende Hilfssatz für die Simpson-Regel. Dabei erscheint es zunächst verblüffend, dass die Simpson-Regel sogar alle Polynome vom Grad ≤ 3 und nicht nur alle Polynome vom Grad ≤ 2 exakt integriert.

Hilfssatz 5.12. (Simpson-Regel ist exakt für Polynome in $\mathbb{P}_3$)
Für jedes Polynom $p_3 \in \mathbb{P}_3$, also $p_3(x) = c_3\,x^3 + c_2\,x^2 + c_1\,x + c_0$ mit den Konstanten $c_0, c_1, c_2, c_3 \in \mathbb{R}$, gilt für die durch (5.16) definierte Simpson-Regel $S_2(p_3) = I(p_3)$, d.h. die Simpson-Regel ***integriert alle Polynome in $\mathbb{P}_3$ (also von Grad ≤ 3) exakt****.*

Beweis von Hilfssatz 5.12: Da die Polynome $q_0(x) = 1$, $q_1(x) = x - a$, $q_2(x) = (x-a)^2$ und $q_3(x) = (x-a)^3$ eine Basis des Polynomraums $\mathbb{P}_3$ der Polynome

vom Grad ≤ 3 bilden, kann man jedes Polynom p_3 im $\mathbb{P}_3$ in der Form

$$\begin{aligned} p_3(x) &= c_0\, q_0(x) + c_1\, q_1(x) + c_2\, q_2(x) + c_3\, q_3(x) \\ &= c_0 + c_1\,(x-a) + c_2\,(x-a)^2 + c_3\,(x-a)^3 \end{aligned}$$

schreiben. Wegen der Linearität des Integrals und der Simpson-Regel, gelten

$$\begin{aligned} I(p_3) &= c_0\, I(q_0) + c_1\, I(q_1) + c_2\, I(q_2) + c_3\, I(q_3), \\ S_2(p_3) &= c_0\, S_2(q_0) + c_1\, S_2(q_1) + c_2\, S_2(q_2) + c_3\, S_2(q_3), \end{aligned}$$

so dass es reicht, zu zeigen, dass $I(q_\ell) = S_2(q_\ell)$ für $\ell = 0, 1, 2, 3$. Wir erhalten

$$\begin{aligned} I(q_\ell) &= \int_a^b (x-a)^\ell \,\mathrm{d}x = \left[\frac{1}{\ell+1}\,(x-a)^{\ell+1}\right]_{x=a}^{x=b} \\ &= \frac{1}{\ell+1}\,(b-a)^{\ell+1} - 0 = \frac{1}{\ell+1}\,(2h)^{\ell+1}, \qquad \ell \in \mathbb{N}_0, \end{aligned} \tag{5.17}$$

wobei wir im letzten Schritt $b - a = 2h$ genutzt haben. Anderseits liefert die Simpson-Regel mit $b - a = 2\,h$ und $\frac{a+b}{2} - a = \frac{b-a}{2} = h$

$$\begin{aligned} S_2(q_\ell) &= \frac{h}{3}\left[(a-a)^\ell + 4\left(\frac{a+b}{2} - a\right)^\ell + (b-a)^\ell\right] \\ &= \frac{h}{3}\left[0^\ell + 4\,h^\ell + (2\,h)^\ell\right] = \begin{cases} \dfrac{4+2^\ell}{3}\,h^{\ell+1} & \text{für } \ell \in \mathbb{N}, \\ \dfrac{h\,(1+4+1)}{3} = 2\,h & \text{für } \ell = 0. \end{cases} \end{aligned} \tag{5.18}$$

Also finden wir durch den Vergleich von (5.17) und (5.18)

$$\begin{aligned} I(q_0) &= \frac{1}{1}(2h)^1 = 2\,h, & S_2(q_0) &= 2\,h = I(q_0), \\ I(q_1) &= \frac{1}{2}\,(2h)^2 = 2\,h^2, & S_2(q_1) &= \frac{4+2^1}{3}\,h^2 = 2\,h^2 = I(q_1), \\ I(q_2) &= \frac{1}{3}\,(2h)^3 = \frac{8}{3}\,h^3, & S_2(q_2) &= \frac{4+2^2}{3}\,h^3 = \frac{8}{3}\,h^3 = I(q_2), \\ I(q_3) &= \frac{1}{4}\,(2h)^4 = 4\,h^4, & S_2(q_3) &= \frac{4+2^3}{3}\,h^4 = 4\,h^4 = I(q_3), \end{aligned}$$

d.h. die Simpson-Regel ist in der Tat für Polynome in $\mathbb{P}_3$ exakt. □

Analog zur Vorgehensweise bei der Trapezregel wollen wir nun mit der Simpson-Regel eine „zusammengesetzte“ Simpson-Regel bauen, indem wir das **Intervall**

$[a;b]$ in n gleich große Teilintervalle zerlegen und auf jedem Teilintervall die Simpson-Regel nutzen: Sei also

$$h = \frac{b-a}{2n}\,.$$

Dann erhält man mit

$$x_k = a + k\,h, \qquad k = 0, 1, 2, \ldots, 2n,$$

durch $[x_{k-2}; x_k]$, $k = 2, 4, \ldots, 2n$, eine Unterteilung von $[a;b]$ in n gleich große Teilintervalle der Länge $2h$. Das Integral (5.13) ist dann entsprechend die Summe der Integrale über diese Teilintervalle, also

$$\begin{aligned} I(f) &= \int_a^b f(x)\,\mathrm{d}x = \int_{x_0}^{x_{2n}} f(x)\,\mathrm{d}x \\ &= \int_{x_0}^{x_2} f(x)\,\mathrm{d}x + \int_{x_2}^{x_4} f(x)\,\mathrm{d}x + \ldots + \int_{x_{2n-2}}^{x_{2n}} f(x)\,\mathrm{d}x. \end{aligned}$$

Wir nutzen nun auf jedem dieser Teilintervalle der Länge $2h$ die Simpson-Regel (5.16) und erhalten damit

$$\begin{aligned} I(f) &\approx \frac{h}{3}\left[f(x_0) + 4\,f(x_1) + f(x_2)\right] + \frac{h}{3}\left[f(x_2) + 4\,f(x_3) + f(x_4)\right] \\ &\quad + \ldots + \frac{h}{3}\left[f(x_{2n-2}) + 4\,f(x_{2n-1}) + f(x_{2n})\right] \\ &= \frac{h}{3}\left[f(x_0) + 4\,f(x_1) + 2\,f(x_2) + 4\,f(x_3) + 2\,f(x_4)\right. \\ &\quad \left. + \ldots + 2\,f(x_{2n-2}) + 4\,f(x_{2n-1}) + f(x_{2n})\right] \\ &= \boxed{\frac{h}{3}\left[f(x_0) + 2\sum_{k=1}^{n-1} f(x_{2k}) + 4\sum_{k=1}^{n} f(x_{2k-1}) + f(x_{2n})\right] = S_{2n}(f)} \end{aligned}$$

Dieses ist die **Simpson-Regel für numerische Integration**, und wir halten diese als Verfahren fest. Die Simpson-Regel für numerische Integration ist seit zwei Jahrhunderten eine der beliebtesten elementaren Integrationsregeln.

Verfahren 5.13. (Simpson-Regel für numerische Integration)
*Sei $f : [a;b] \to \mathbb{R}$ eine stetige Funktion. Für $n \in \mathbb{N}$ sei $h = \frac{1}{2n}(b-a)$, und die **Knoten(punkte)** seien $x_k = a + k\,h$, $k = 0, 1, 2, \ldots, 2n$. Dann liefert die **Simpson-Regel für numerische Integration***

$$S_{2n}(f) = \frac{h}{3}\left[f(x_0) + 2\sum_{k=1}^{n-1} f(x_{2k}) + 4\sum_{k=1}^{n} f(x_{2k-1}) + f(x_{2n})\right] \qquad (5.19)$$

$$= \frac{h}{3}\left[f(x_0) + 4\,f(x_1) + 2\,f(x_2) + 4\,f(x_3) + 2\,f(x_4) + \ldots + 2\,f(x_{2n-2}) + 4\,f(x_{2n-1}) + f(x_{2n})\right]$$

eine Näherung für das Integral $I(f) = \displaystyle\int_a^b f(x)\,\mathrm{d}x$.

Betrachten wir zunächst ein Beispiel.

Beispiel 5.14. (Simpson-Regel für numerische Integration)
Wir berechnen das Integral

$$\int_0^1 e^{-x^2}\,\mathrm{d}x \doteq 0{,}746824132812427$$

(welches bereits in Beispiel 5.7 mit der Trapezregel für numerische Integration T_n berechnet wurde) für $2n = 2^j$ mit $j = 1, 2, \ldots, 7$ mit der Simpson-Regel für numerische Integration S_{2n}: Es sind also $[a; b] = [0; 1]$, $f(x) = e^{-x^2}$, $h = \frac{1-0}{2n} = \frac{1}{2n}$, und die Knoten sind dann $x_k = a + k\,h = 0 + k \cdot \frac{1}{2n} = \frac{k}{2n}$, $k = 0, 1, 2, \ldots, 2n$. Wir müssen also für $2n = 2^j$ mit $j = 1, 2, \ldots, 7$ jeweils die folgende gewichtete Summe berechnen:

$$S_{2n}(f) = \frac{1}{6n}\left[e^{-0^2} + 2\sum_{k=1}^{n-1} e^{-((2k)/(2n))^2} + 4\sum_{k=1}^{n} e^{-((2k-1)/(2n))^2} + e^{-1^2}\right]$$

$$= \frac{1}{6n}\left[1 + 2\sum_{k=1}^{n-1} e^{-(k/n)^2} + 4\sum_{k=1}^{n} e^{-((k-\frac{1}{2})/n)^2} + e^{-1}\right]$$

Die Ergebnisse sind in Tabelle 5.2 auf eine 11-stellige Gleitkommadarstellung gerundet angegeben. Weiter haben wir den absoluten Fehler $I(f) - S_{2n}(f)$ und den Quotienten $|I(f) - S_n(f)|/|I(f) - S_{2n}(f)|$ auf eine 3-stellige Gleitkommadarstellung gerundet aufgelistet. – Ein Vergleich der Tabelle 5.2 mit Tabelle 5.1 zeigt, dass die Simpson-Regel S_{2n} eine deutlich bessere Näherung liefert als die Trapezregel T_{2n} mit der gleichen Knotenzahl.

An den Quotienten $|I(f) - S_n(f)|/|I(f) - S_{2n}(f)|$ sehen wir, dass gilt für $n \geq 4$ gilt

$$\frac{|I(f) - S_n(f)|}{|I(f) - S_{2n}(f)|} \approx 16 \qquad \Longleftrightarrow \qquad \frac{1}{16}\,|I(f) - S_n(f)| \approx |I(f) - S_{2n}(f)|.$$

Dies bedeutet, dass bei einer Verdoppelung der Anzahl der Teilintervalle und damit einer Halbierung des Abstands der Knotenpunkte x_k, $k = 0, 1, \ldots, 2n$,

$2n$	$h = \frac{b-a}{2n} = \frac{1}{2n}$	$S_{2n}(f)$	$I(f) - S_{2n}(f)$	$\frac{\lvert I(f) - S_n(f)\rvert}{\lvert I(f) - S_{2n}(f)\rvert}$
$2 = 2^1$	0,5	0,74718042891	$-3{,}56 \cdot 10^{-4}$	
$4 = 2^2$	0,25	0,74685537979	$-3{,}12 \cdot 10^{-5}$	11,4
$8 = 2^3$	0,125	0,74682612053	$-1{,}99 \cdot 10^{-6}$	15,7
$16 = 2^4$	0,0625	0,74682425744	$-1{,}25 \cdot 10^{-7}$	15,9
$32 = 2^5$	0,03125	0,74682414061	$-7{,}79 \cdot 10^{-9}$	16,0
$64 = 2^6$	0,015625	0,74682413330	$-4{,}87 \cdot 10^{-10}$	16,0
$128 = 2^7$	0,0078125	0,74682413284	$-3{,}04 \cdot 10^{-11}$	16,0

Tabelle 5.2: Simpson-Regel $S_{2n}(f)$ mit $f(x) = e^{-x^2}$ und $[a;b] = [0;1]$ zur Berechnung des Integrals $\int_0^1 e^{-x^2}\,\mathrm{d}x$.

(und ungefähr einer Verdoppelung der Anzahl der Knotenpunkte) der Betrag des absoluten numerischen Integrationsfehler $|I(f) - S_{2n}(f)|$ ungefähr $1/16$ des Betrags des absoluten numerischen Integrationsfehler $|I(f) - S_n(f)|$ mit dem doppelten Abstand $2h$ ist. Daher vermuten wir, dass sich der Betrag des absoluten numerischen Integrationsfehler $|I(f) - S_{2n}(f)|$ möglicherweise wie $C\,h^4$ mit einer Konstante $C > 0$ verhält, denn: Gelten

$$|I(f) - S_{2n}(f)| \approx C\,h^4 \quad \text{und entsprechend} \quad |I(f) - S_n(f)| \approx C\,(2h)^4 = 16\,C\,h^4,$$

so folgt angenähert

$$\frac{|I(f) - S_n(f)|}{|I(f) - S_{2n}(f)|} \approx \frac{16\,C\,h^4}{C\,h^4} = 16,$$

wie wir es in Tabelle 5.2 beobachtet haben. ♠

Der nachfolgende Satz zeigt, dass das im vorigen Beispiel vermutete Konvergenzverhalten $|I(f) - S_{2n}(f)| \approx C\,h^4$ (mit einer Konstante $C > 0$) richtig ist, bzw. es liegt mindestens ein solches Konvergenzverhalten vor.

Satz 5.15. (Konvergenz der Simpson-Regel für num. Integration)
Sei $n \in \mathbb{N}$, und sei $f : [a;b] \to \mathbb{R}$ eine viermal stetig differenzierbare Funktion. Dann gilt für den ***Betrag des absoluten Fehlers*** $E_{2n}^S(f) = |I(f) - S_{2n}(f)|$ *der Näherung des Integrals*

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

durch die in Verfahren 5.13 definierte ***Simpson-Regel für numerische Integration*** $S_{2n}(f)$ *die Fehlerabschätzung*

$$E_{2n}^S(f) = |I(f) - S_{2n}(f)| \leq \frac{b-a}{180}\, h^4 \max_{x\in[a;b]} |f^{(4)}(x)|.$$

Man kann Satz 5.15 analog zu der Vorgehensweise in Beispiel 5.9 für Satz 5.8 bei der Trapezregel nutzen, um einen Wert für n zu berechnen, ab dem $E_{2n}^S(f) = |I(f) - S_{2n}(f)|$ garantiert kleiner oder gleich einer gegebenen absoluten Fehlerschranke ε ist. Wir wenden Satz 5.15 in Übungsaufgaben auf diese Art an.

Aus Hilfssatz 5.12 folgt schließlich noch, dass die Simpson-Regel für numerische Integration auf $\mathbb{P}_3$ exakt ist, indem man Hilfssatz 5.12 für jedes der Teilintervalle ausnutzt.

Satz 5.16. (Simpson-Regel für num. Integration ist exakt auf $\mathbb{P}_3$)
Für jedes Polynom $p_3(x) = c_3\,x^3 + c_2\,x^2 + c_1\,x + c_0$ *in* $\mathbb{P}_3$ *mit Konstanten* $c_0, c_1,\ c_2, c_3 \in \mathbb{R}$ *gilt für die durch* (5.19) *definierte Simpson-Regel für numerische Integration* $S_{2n}(p_3) = I(p_3)$*, d.h. die Simpson-Regel für numerische Integration* ***integriert Polynome in*** $\mathbb{P}_3$ ***(also von Grad*** ≤ 3***) exakt****.*

5.3 Gauß Quadratur*

Bei der Konstruktion der Trapezregel und der Simpson-Regel für die numerische Integration des Integrals

$$I(f) = \int_a^b f(x)\,\mathrm{d}x \qquad \text{mit stetigem} \qquad f : [a;b] \to \mathbb{R} \tag{5.20}$$

wurde der Integrand in (5.20) auf jedem Teilintervall jeweils durch das interpolierende Polynom (bzgl. äquidistanter Knotenpunkte) vom Grad 1 bzw. 2 ersetzt. Die so erhaltenen Integrationsformeln hatten die Eigenschaft, dass sie alle Polynome in $\mathbb{P}_1$ (bei der Trapezregel) bzw. alle Polynome in $\mathbb{P}_3$ (bei der Simpson-Regel) exakt integrierten. Nun wollen wir numerische Integrationsformeln finden, die **alle Polynome bis zum einem möglichst hohen Grad exakt integrieren**.

*Dieses Teilkapitel ist nicht klausurrelevant.

n	$\varrho_n(f)$	n	$\varrho_n(f)$
1	$5{,}30 \cdot 10^{-2}$	6	$7{,}82 \cdot 10^{-6}$
2	$1{,}79 \cdot 10^{-2}$	7	$4{,}62 \cdot 10^{-7}$
3	$6{,}63 \cdot 10^{-4}$	8	$9{,}64 \cdot 10^{-8}$
4	$4{,}63 \cdot 10^{-4}$	9	$8{,}05 \cdot 10^{-9}$
5	$1{,}62 \cdot 10^{-5}$	10	$9{,}16 \cdot 10^{-10}$

Tabelle 5.3: Fehler der Minimax-Approximation von $f : [0;1] \to \mathbb{R}$, $f(x) = e^{-x^2}$.

Warum ist eine solche Vorgehensweise sinnvoll? Stetige Funktion lassen sich sehr gut durch Polynome approximieren, wie das nachfolgende Beispiel zeigt. Daher können wir hoffen, dass eine Integrationsformel, die Polynome bis zu einem hinreichend hohen Grad exakt integriert, auch stetige Funktionen gut integriert.

Genauer gilt für eine stetige Funktion $f : [a;b] \to \mathbb{R}$, dass es ein **eindeutig bestimmtes Polynom** $q_n \in \mathbb{P}_n$ (also vom Grad $\leq n$) gibt, so dass gilt

$$\max_{x\in[a;b]} |f(x) - q_n(x)| = \min_{p\in\mathbb{P}_n} \left(\max_{x\in[a;b]} |f(x) - p(x)| \right) = \varrho_n(f). \tag{5.21}$$

Bei $\varrho_n(f)$ in (5.21) handelt es sich um die maximale betragliche Abweichung von q_n von f, wobei q_n das Polynom in $\mathbb{P}_n$ ist, für welches diese betragliche Abweichung am kleinsten ist. Man nennt q_n in (5.21) auch die **Minimax-Approximation von f in $\mathbb{P}_n$** und $\varrho_n(f)$ in (5.21) den **Minimax-Fehler** (der Minimax-Approximation von f in $\mathbb{P}_n$). Im nachfolgenden Beispiel wurde $\varrho_n(f)$ für eine konkrete Funktion berechnet, und wir sehen, dass $\varrho_n(f)$ rapide gegen null strebt.

Beispiel 5.17. (Approximation einer stetigen Funktion in $\mathbb{P}_n$)
Sei $f : [0;1] \to \mathbb{R}$, $f(x) = e^{-x^2}$. In Tabelle 5.3 wurde der Minimax-Fehler (5.21) auf eine 3-stellige Gleitkommadarstellung gerundet angegeben. Wir sehen, dass dieser rapide gegen null strebt. Allerdings nimmt der Minimax-Fehler nicht mit einer gleichmäßigen Rate ab. ♠

Kommen wir nach dieser Motivation zu unserer Ausgangsfragestellung zurück: Wir möchten also numerische Integrationsformeln konstruieren, die alle Polynome in $\mathbb{P}_m$ mit einen niedrigen bis moderaten Grad m exakt integrieren. Wir beschränken uns bei der nachfolgenden Diskussion auf das Intervall $[a;b] = [-1;1]$. Der Transfer auf andere Intervalle ist unkompliziert und wird am Ende besprochen.

Wir wollen also für stetige Funktionen $f : [-1; 1] \to \mathbb{R}$ das Integral

$$I(f) = \int_{-1}^{1} f(x)\,\mathrm{d}x \tag{5.22}$$

mit einer **numerischen Integrationsformel** oder **Quadraturformel**

$$Q_n(f) = \sum_{j=1}^{n} w_j\, f(x_j), \tag{5.23}$$

welche alle Polynome bis zu einem möglichst hohen Grad m exakt über $[-1; 1]$ integriert, angenähert berechnen. In (5.23) nennen wir die n paarweise verschiedenen Punkte $x_1, x_2, \ldots, x_n$ die **Knoten(punkte)** der Quadraturformel Q_n, und die Zahlen $w_1, w_2, \ldots, w_n$ nennen wir die zugehörigen **Gewichte**. Üblicherweise verlangt man auch noch, dass die **Gewichte** $w_1, w_2, \ldots, w_n$ **positiv** sind. Es gibt allerdings auch numerische Integrationsformeln mit positiven und negativen Gewichten.

Die Quadraturformel (5.23) hat also $2n$ Parameter, nämlich $x_1, x_2, \ldots, x_n$ und $w_1, w_2, \ldots, w_n$, die wir passend wählen können, um zum erreichen, dass (5.23) Polynome bis zu einem möglichst hohen Grad m exakt integriert. Der Raum $\mathbb{P}_m$ der Polynome vom Grad $\leq m$ hat die Dimension $\dim(\mathbb{P}_m) = m + 1$, denn die $m + 1$ Monome $p_\ell(x) = x^\ell$, $\ell = 0, 1, \ldots, m$, bilden eine Basis für $\mathbb{P}_m$, d.h. jedes Polynom p in $\mathbb{P}_m$ lässt sich als

$$p(x) = a_0 + a_1\, x + a_2\, x^2 + \ldots + a_m\, x^m$$

mit $m + 1$ geeigneten Koeffizienten $a_0, a_1, a_2, \ldots, a_m \in \mathbb{R}$ darstellen. Daher vermuten wir, dass es mit einer Quadraturformel (5.23) bei passender Wahl der insgesamt $2n = \dim(\mathbb{P}_{2n-1})$ Parameter (n Knoten(punkte) und n Gewichte) möglich sein sollte, alle Polynome in $\mathbb{P}_{2n-1}$, also vom Grad $\leq 2n - 1$, exakt zu integrieren. Formeln mit dieser Eigenschaft nennt man **Gauß Quadraturformeln**. Wir halten dieses direkt für den Fall eines beliebigen Intervalls $[a; b]$ als Definition fest.

Definition 5.18. (Gauß Quadraturformel/Gauß Quadratur)
Sei Q_n eine Quadraturformel

$$Q_n(f) = \sum_{j=1}^{n} w_j\, f(x_j), \qquad f \in \mathcal{C}\big([a; b]\big),$$

*mit den **n (verschiedenen) Knoten(punkten)** $x_1, x_2, \ldots, x_n$ und den **n zugehörigen Gewichten** $w_1, w_2, \ldots, w_n$, als Näherung des Integrals*

$$I(f) = \int_{a}^{b} f(x)\,\mathrm{d}x, \qquad f \in \mathcal{C}\big([a; b]\big).$$

Wenn gilt $Q_n(p) = I(p)$ *für alle* $p \in \mathbb{P}_{2n-1}$, *also*

$$\sum_{j=1}^{n} w_j \, p(x_j) = \int_a^b p(x) \, \mathrm{d}x \qquad \text{für alle } p \in \mathbb{P}_{2n-1},$$

dann nennen wir Q_n *eine* ***Gauß Quadraturformel*** *oder* ***Gauß Quadratur***.

Wenn wir prüfen möchten, ob eine Integrationsformel (5.23) alle Polynome in $\mathbb{P}_m$ exakt über $[-1; 1]$ integriert, reicht es dieses für die Monome $p_\ell(x) = x^\ell$, $\ell = 0, 1, \ldots, m$, zu überprüfen, denn für ein beliebiges Polynom $p \in \mathbb{P}_m$ gilt

$$p(x) = a_0 + a_1 \, x + a_2 \, x^2 + \ldots + a_m \, x^m = \sum_{\ell=0}^{m} a_\ell \, x^\ell,$$

und damit folgen

$$I(p) = \int_{-1}^{1} p(x) \, \mathrm{d}x = \int_{-1}^{1} \left(\sum_{\ell=0}^{m} a_\ell \, x^\ell \right) \mathrm{d}x = \sum_{\ell=0}^{m} a_\ell \left(\int_{-1}^{1} x^\ell \, \mathrm{d}x \right) = \sum_{\ell=0}^{m} a_\ell \, I(x^\ell) \tag{5.24}$$

und

$$Q_n(p) = \sum_{j=1}^{n} w_j \left(\sum_{\ell=0}^{m} a_\ell \, x_j^\ell \right) = \sum_{\ell=0}^{m} a_\ell \left(\sum_{j=1}^{n} w_j x_j^\ell \right) = \sum_{\ell=0}^{m} a_\ell \, Q_n(x^\ell). \tag{5.25}$$

Gilt also $Q_n(x^\ell) = I(x^\ell)$ für alle $\ell = 0, 1, 2, \ldots, m$, so folgt aus (5.24) und (5.25) sofort, dass für alle $p \in \mathbb{P}_m$ ebenfalls $Q_n(p) = I(p)$ gilt.

Wir studieren das Problem der Konstruktion von Gauß Quadraturformeln zunächst für $n = 1$ und $n = 2$ separat.

Beispiel 5.19. (Gauß Quadratur mit $n = 1$ Knoten)
Wir betrachten also

$$\int_{-1}^{1} f(x) \, \mathrm{d}x \approx w_1 \, f(x_1) = Q_1(f), \tag{5.26}$$

wobei das Gewicht w_1 und der Knoten(punkt) x_1 so gewählt werden sollen, dass die durch die rechte Seite von (5.26) gegebene Integrationsformel Q_1 alle Polynome bis zu einem möglichst hohen Grad exakt über $[-1; 1]$ integriert.

Wir verlangen also zunächst in (5.26) Gleichheit für das konstante Polynom $p_0(x) = 1$, d.h. es soll gelten:

$$Q_1(p_0) = w_1\, p_0(x_1) = w_1 \cdot 1 = w_1 \overset{!}{=} \int_{-1}^{1} p_0(x)\,\mathrm{d}x = \int_{-1}^{1} 1\,\mathrm{d}x$$
$$= \Big[x\Big]_{x=-1}^{x=1} = 1 - (-1) = 2 \qquad \Longrightarrow \qquad w_1 = 2 \tag{5.27}$$

Weiter verlangen wir in (5.26) Gleichheit für das lineare Polynom $p_1(x) = x$, d.h. es soll gelten:

$$Q_1(p_1) = w_1\, p_1(x_1) = w_1 \cdot x_1 \overset{!}{=} \int_{-1}^{1} p_1(x)\,\mathrm{d}x = \int_{-1}^{1} x\,\mathrm{d}x = \left[\frac{x^2}{2}\right]_{x=-1}^{x=1}$$
$$= \frac{1^2}{2} - \frac{(-1)^2}{2} = 0 \quad \Longrightarrow \quad w_1\, x_1 = 0 \quad \overset{w_1=2}{\Longrightarrow} \quad 2\, x_1 = 0 \quad \Longrightarrow \quad x_1 = 0 \tag{5.28}$$

Mit $w_1 = 2$ und $x_1 = 0$ (aus (5.27) und (5.28)) erhalten wir also die Gauß Quadraturformel

$$\boxed{Q_1(f) = 2\, f(0),}$$

welche alle Polynome in $\mathbb{P}_1$ (also vom Grad $\leq 1 = 2 \cdot 1 - 1$) exakt über das Intervall $[-1; 1]$ integriert. ♠

Beispiel 5.20. (Gauß Quadratur mit $n = 2$ Knoten)
Wir betrachten also

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx w_1\, f(x_1) + w_2\, f(x_2) = Q_2(f), \tag{5.29}$$

wobei die Gewichte w_1 und w_2 und die Knoten(punkte) x_1 und x_2 so gewählt werden sollen, dass die durch die rechte Seite von (5.29) gegebene Integrationsformel Q_2 alle Polynome bis zu einem möglichst hohen Grad exakt über $[-1; 1]$ integriert. – Wir verlangen also zunächst in (5.29) Gleichheit für das konstante Polynom $p_0(x) = 1$, d.h. es soll gelten:

$$Q_2(p_0) = w_1\, p_0(x_1) + w_2\, p_0(x_2) = w_1 \cdot 1 + w_2 \cdot 1 = w_1 + w_2 \overset{!}{=} \int_{-1}^{1} p_0(x)\,\mathrm{d}x$$
$$= \int_{-1}^{1} 1\,\mathrm{d}x = \Big[x\Big]_{x=-1}^{x=1} = 1 - (-1) = 2 \quad \Longrightarrow \quad w_1 + w_2 = 2 \tag{5.30}$$

Weiter verlangen wir in (5.29) Gleichheit für das lineare Polynom $p_1(x) = x$, d.h. es soll gelten:

$$Q_2(p_1) = w_1\, p_1(x_1) + w_2\, p_1(x_2) = w_1 \cdot x_1 + w_2 \cdot x_2 \stackrel{!}{=} \int_{-1}^{1} p_1(x)\,\mathrm{d}x = \int_{-1}^{1} x\,\mathrm{d}x$$
$$= \left[\frac{x^2}{2}\right]_{x=-1}^{x=1} = \frac{1^2}{2} - \frac{(-1)^2}{2} = 0 \quad \Longrightarrow \quad w_1\, x_1 + w_2\, x_2 = 0 \qquad (5.31)$$

Weiter verlangen wir in (5.29) Gleichheit für das quadratische Polynom $p_2(x) = x^2$, d.h. es soll gelten:

$$Q_2(p_2) = w_1\, p_2(x_1) + w_2\, p_2(x_2) = w_1 \cdot x_1^2 + w_2 \cdot x_2^2 \stackrel{!}{=} \int_{-1}^{1} p_2(x)\,\mathrm{d}x = \int_{-1}^{1} x^2\,\mathrm{d}x$$
$$= \left[\frac{x^3}{3}\right]_{x=-1}^{x=1} = \frac{1^3}{3} - \frac{(-1)^3}{3} = \frac{2}{3} \quad \Longrightarrow \quad w_1\, x_1^2 + w_2\, x_2^2 = \frac{2}{3} \qquad (5.32)$$

Weiter verlangen wir in (5.29) Gleichheit für das kubische Polynom $p_3(x) = x^3$, d.h. es soll gelten:

$$Q_2(p_3) = w_1\, p_3(x_1) + w_2\, p_3(x_2) = w_1 \cdot x_1^3 + w_2 \cdot x_2^3 \stackrel{!}{=} \int_{-1}^{1} p_3(x)\,\mathrm{d}x = \int_{-1}^{1} x^3\,\mathrm{d}x$$
$$= \left[\frac{1}{4}\,x^4\right]_{x=-1}^{x=1} = \frac{1^4}{4} - \frac{(-1)^4}{4} = 0 \quad \Longrightarrow \quad w_1\, x_1^3 + w_2\, x_3^2 = 0 \qquad (5.33)$$

Aus (5.30), (5.31), (5.32) und (5.33) erhalten wir die folgenden vier (teilweise nicht-linearen) Gleichungen in den vier Unbekannten w_1, w_2, x_1, x_2:

$$\begin{aligned} w_1 + w_2 &= 2, \\ w_1\, x_1 + w_2\, x_2 &= 0, \\ w_1\, x_1^2 + w_2\, x_2^2 &= \frac{2}{3}, \\ w_1\, x_1^3 + w_2\, x_2^3 &= 0. \end{aligned}$$

Dieses ist ein **nicht-lineares System von vier Gleichungen in vier Unbekannten**. Man kann zeigen, dass dieses die folgenden beiden Lösungen hat

$$\left(w_1 = w_2 = 1,\ x_1 = -\tfrac{\sqrt{3}}{3},\ x_2 = \tfrac{\sqrt{3}}{3}\right), \quad \left(w_1 = w_2 = 1,\ x_1 = \tfrac{\sqrt{3}}{3},\ x_2 = -\tfrac{\sqrt{3}}{3}\right),$$

welche bis auf eine Umnummerierung der Knoten identisch sind. Wir erhalten also die Gauß Quadraturformel

$$\boxed{Q_2(f) = 1 \cdot f\left(-\tfrac{\sqrt{3}}{3}\right) + 1 \cdot f\left(\tfrac{\sqrt{3}}{3}\right) = f\left(-\tfrac{\sqrt{3}}{3}\right) + f\left(\tfrac{\sqrt{3}}{3}\right),}$$

welche alle Polynome in $\mathbb{P}_3$ (also vom Grad $\leq 3 = 2 \cdot 2 - 1$) exakt über das Intervall $[-1; 1]$ integriert. ♠

Betrachten wir ein konkretes Beispiel.

Beispiel 5.21. (Gauß Quadratur mit Q_2)
Wir berechnen das Integral

$$I(f) = \int_{-1}^{1} e^x \, \mathrm{d}x = \Big[e^x\Big]_{x=-1}^{x=1} = e - e^{-1} \doteq 2{,}3504024$$

mit der in Beispiel 5.20 hergeleiteten Gauß Quadraturformel Q_2 angenähert:

$$Q_2(f) = e^{-\sqrt{3}/3} + e^{\sqrt{3}/3} \doteq 2{,}3426961.$$

Dabei wurde jeweils auf eine 8-stellige Gleitkommadarstellung gerundet. Der absolute Fehler und der relative Fehler sind auf eine 3-stellige Gleitkommadarstellung gerundet $I(f) - Q_2(f) \doteq 0{,}00771$ bzw. $\big[I(f) - Q_2(f)\big]/I(f) \doteq 0{,}00328$. Unter Berücksichtigung der Tatsache, dass nur zwei Knoten(punkte) verwendet wurden, ist der relative Fehler ziemlich klein. ♠

Wie sieht es in Verallgemeinerung von Beispielen 5.19 und 5.20 mit der **Konstruktion von Gauß Quadraturformeln Q_n mit $n \in \mathbb{N}$ (die also auf $\mathbb{P}_{2n-1}$ exakt sind)** aus?

Wir wollen also für stetige Funktionen $f : [-1; 1] \to \mathbb{R}$ das Integral

$$I(f) = \int_{-1}^{1} f(x) \, \mathrm{d}x$$

mit einer numerischen Integrationsformel

$$Q_n(f) = \sum_{j=1}^{n} w_j \, f(x_j) \tag{5.34}$$

angenähert berechnen, und die Formel (5.34) soll Polynome bis zu einem möglichst hohen Grad m exakt über $[-1; 1]$ integrieren. Die Quadraturformel (5.34) hat n Knoten(punkte) $x_1, x_2, \ldots, x_n$ und n zugehörige Gewichte $w_1, w_2, \ldots, w_n$. Wir können also insgesamt $2n$ Parameter wählen und erwarten daher, dass es möglich sein sollte, die $2n$ Monome $q_\ell(x) = x^\ell$, $\ell = 0, 1, 2, \ldots, 2n - 1$, exakt über $[-1; 1]$ zu integrieren. Damit würden dann von Q_n auch auch alle Polynome in $\mathbb{P}_{2n-1}$ exakt über $[-1; 1]$ integriert.

n	j	w_j	x_j
1	1	2,0000000000	0,0000000000
2	1	1,0000000000	−0,5773502692
	2	1,0000000000	0,5773502692
3	1	0,5555555556	−0,7745966692
	2	0,8888888889	0,0000000000
	3	0,5555555556	0,7745966692
4	1	0,3478548451	−0,8611363116
	2	0,6521451549	−0,3399810436
	3	0,6521451549	0,3399810436
	4	0,3478548451	0,8611363116
5	1	0,2369268851	−0,9061798459
	2	0,4786286705	−0.5384693101
	3	0,5688888889	0,0000000000
	4	0,4786286705	0,5384693101
	5	0,2369268851	0,9061798459
6	1	0,1713244924	−0,9324695142
	2	0,3607615730	−0,6612093865
	3	0,4679139346	−0,2386191861
6	4	0,4679139346	0,2386191861
	5	0,3607615730	0,6612093865
	6	0,1713244924	0,9324695142
7	1	0,1294849662	−0,9491079123
	2	0,2797053915	−0,7415311856
	3	0,3818300505	−0,4058451514
	4	0,4179591837	0,0000000000
	6	0,3818300505	0,4058451514
	6	0,2797053915	0,7415311856
	7	0,1294849662	0,9491079123
8	1	0,1012285363	−0,9602898565
	2	0,2223810345	−0,7966664774
	3	0,3137066459	−0,5255324099
	4	0,3626837834	−0,1834346425
	5	0,3626837834	0,1834346425
	6	0,3137066459	0,5255324099
	7	0,2223810345	0,7966664774
	8	0,1012285363	0,9602898565

Tabelle 5.4: Gewichte und Knoten der Gauß Quadraturformeln Q_n, $n = 1, \ldots, 8$, (mit einer Rundung auf eine Gleitkommadarstellung mit 10-stelliger Mantisse) zur numerischen Integration über $[-1; 1]$ mit einem Exaktheitsgrad von $2n - 1$.

Aus den Bedingungen $Q_n(x^\ell) = I(x^\ell)$ für alle $\ell = 0, 1, 2, \ldots, 2n-1$ erhalten wir ein **System mit** den folgenden $2n$ **nicht-linearen Gleichungen**:

$$\begin{aligned} w_1 + w_2 + \ldots + w_n &= 2, \\ w_1\, x_1 + w_2\, x_2 + \ldots + w_n\, x_n &= 0, \\ w_1\, x_1^2 + w_2\, x_2^2 + \ldots + w_n\, x_n^2 &= \tfrac{2}{3}\,, \\ w_1\, x_1^3 + w_2\, x_2^3 + \ldots + w_n\, x_n^3 &= 0, \\ \vdots \qquad \vdots \quad \vdots & \end{aligned}$$

$$w_1\, x_1^{2n-2} + w_2\, x_2^{2n-2} + \ldots + w_n\, x_n^{2n-2} = \tfrac{2}{2n-1}\,,$$
$$w_1\, x_1^{2n-1} + w_2\, x_2^{2n-1} + \ldots + w_n\, x_n^{2n-1} = 0. \qquad (5.35)$$

Es ist in der Tat möglich, die n Knoten(punkte) $x_1, x_2, \ldots, x_n$ und n zugehörige Gewichte $w_1, w_2, \ldots, w_n$ so zu wählen, dass die $2n$ nicht-linearen Gleichungen in (5.35) alle erfüllt sind. Die Lösung dieses nicht-linearen Gleichungssystems ist allerdings (vor allem für größere Werte von n) eine sehr herausfordernde Aufgabe. Günstigerweise liegen diese Knoten und Gewichte aber in Tabellen vor, und in Tabelle 5.4 sind die Knoten und Gewichte der Gauß Quadraturformeln Q_n (mit einer Rundung auf eine Gleitkommadarstellung mit 10-stelliger Mantisse) für $n = 1, 2, \ldots, 8$ angegeben.

Betrachten wir ein Beispiel.

Beispiel 5.22. (Gauß Quadratur)
Wir berechnen das Integral

$$I(f) = \int_{-1}^{1} e^x \,\mathrm{d}x = e - e^{-1} \doteq 2{,}350402387$$

aus Beispiel 5.21 nun mit den Gauß Quadraturformeln $Q_n(f)$ für $n = 1, 2, \ldots, 5$. Die Ergebnisse sind in Tabelle 5.5 auf eine 10-stellige Gleitkommadarstellung gerundet zusammen mit den auf eine 3-stellige Gleitkommadarstellung gerundeten absoluten Fehlern angegeben. Wir sehen, das wir bereits mit fünf Knoten und fünf zugehörigen Gewichten eine sehr gute Näherung erreichen. ♠

Transfer für andere Integrationsintervalle: Wie berechnet man ein Integral

$$I(f) = \int_a^b f(x)\,\mathrm{d}x \qquad \text{mit stetigem} \qquad f : [a;b] \to \mathbb{R}, \qquad (5.36)$$

bei dem $[a;b] \neq [-1;1]$ ist? Hier hilft die **Substitutionsregel:** Mit der affin linearen Funktion

$$x = x(t) = \frac{b + a + t\,(b-a)}{2}\,, \qquad t \in [-1;1], \qquad (5.37)$$

wird das Intervall $[-1;1]$ auf das Intervall $[a;b]$ abgebildet, denn $x(-1) = a$ und $x(1) = b$. Mit der Substitution $x = x(t)$ in (5.37) mit

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{b-a}{2} \qquad \Longleftrightarrow \qquad \mathrm{d}x = \frac{b-a}{2}\,\mathrm{d}t$$

n	$Q_n(f)$	$I(f) - Q_n(t)$
1	2,0000000000	$3{,}50 \cdot 10^{-1}$
2	2,3426960879	$7{,}71 \cdot 10^{-3}$
3	2,3503369288	$6{,}55 \cdot 10^{-5}$
4	2,3504020921	$2{,}95 \cdot 10^{-7}$
5	2,3504023866	$7{,}08 \cdot 10^{-10}$

Tabelle 5.5: Gauß Quadratur $Q_n(f)$ zur Berechnung von $I(f) = \int_{-1}^{1} e^x \,\mathrm{d}x$.

folgt aus (5.36)

$$I(f) = \int_a^b f(x)\,\mathrm{d}x = \int_{-1}^{1} f\left(\frac{b+a+t\,(b-a)}{2}\right)\frac{b-a}{2}\,\mathrm{d}t$$

$$= \frac{b-a}{2}\int_{-1}^{1} \underbrace{f\left(\frac{b+a+t\,(b-a)}{2}\right)}_{=\widetilde{f}(t)}\,\mathrm{d}t = \frac{b-a}{2}\int_{-1}^{1}\widetilde{f}(t)\,\mathrm{d}t \qquad (5.38)$$

mit der neuen Funktion

$$\widetilde{f}(t) = f\big(x(t)\big) = f\left(\frac{b+a+t\,(b-a)}{2}\right), \qquad t \in [-1;1]. \qquad (5.39)$$

Also können wir mittels

$$\boxed{\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2}\int_{-1}^{1}\widetilde{f}(t)\,\mathrm{d}t \qquad \text{mit} \quad \widetilde{f}(t) = f\left(\frac{b+a+t\,(b-a)}{2}\right)} \qquad (5.40)$$

Integrale über $[a;b]$ mit den eben eingeführten Gauß Quadraturformeln für numerische Integration über $[-1;1]$ berechnen. Nutzt man die Gauß Quadraturformel Q_n, um das Integral auf rechten Seite von (5.40) zu berechnen, so werden alle $f \in \mathbb{P}_n$ exakt integriert, denn durch die Transformation (5.39) wird ein Polynom $f \in \mathbb{P}_n$ wieder auf ein Polynom $\widetilde{f} \in \mathbb{P}_n$ abgebildet. Man erhält dann die folgende **Quadraturformel für eine Gauß Quadratur für das Intervall** $[a;b]$:

$$\boxed{\widetilde{Q}_n(f) = \frac{b-a}{2}\sum_{k=1}^{n} w_k\, f\big(x(t_k)\big),} \qquad (5.41)$$

n	$Q_n(f)$	$I(f) - Q_n(t)$
1	0,7788007831	$-3{,}20 \cdot 10^{-2}$
2	0,7465946883	$2{,}29 \cdot 10^{-4}$
3	0,7468145842	$9{,}55 \cdot 10^{-6}$
4	0,7468244681	$-3{,}35 \cdot 10^{-7}$
5	0,7468241268	$6{,}01 \cdot 10^{-9}$
6	0,7468241329	$-7{,}61 \cdot 10^{-11}$

Tabelle 5.6: Gauß Quadratur $Q_n(f)$ mit $f(x) = e^{-x^2}$ und $[a;b] = [0;1]$ zur Berechnung von $I(f) = \int_0^1 e^{-x^2}\,\mathrm{d}x$.

wobei t_k, $k = 1, 2, \ldots, n$, die Knoten und w_k, $k = 1, 2, \ldots, n$, die zugehörigen Gewichte der Gauß Quadratur Q_n für $[-1;1]$ sind. Die Gauß Quadratur (5.41) für das Intervall $[a;b]$ hat dann also die Knoten $x_k = x(t_k)$, $k = 1, 2, \ldots, n$, und die zugehörigen Gewichte $\frac{b-a}{2}\, w_k$, $k = 1, 2, \ldots, n$.

Betrachten wir ein Beispiel dazu.

Beispiel 5.23. (Gauß Quadraturformel)
Wir berechnen das Integral

$$\int_0^1 e^{-x^2}\,\mathrm{d}x \doteq 0{,}746824132812427$$

(welches bereits in Beispiel 5.7 mit der Trapezregel für numerische Integration und in Beispiel 5.14 mit der Simpson-Regel für numerische Integration berechnet wurde) nun mit den Gauß Quadraturformeln Q_n. Hier ist $[a;b] = [0;1]$, und somit benötigen wir die affin lineare Transformation

$$x(t) = \frac{1 + 0 + t\,(1-0)}{2} = \frac{1+t}{2}\,.$$

Mit (5.40) erhalten wir

$$\int_0^1 e^{-x^2}\,\mathrm{d}x = \frac{1}{2}\int_{-1}^1 \widetilde{f}(t)\,\mathrm{d}t \quad \text{mit} \quad \widetilde{f}(t) = \exp\Big(-\big(x(t)\big)^2\Big) = \exp\Big(-\big(\tfrac{1+t}{2}\big)^2\Big)\,.$$

Also lautet die transformierte Formel für die Gauß Quadratur

$$Q_n(f) = \frac{1}{2}\sum_{k=1}^n w_k\, \exp\left(-\left(\frac{1+x_k}{2}\right)^2\right),$$

wobei $x_1, x_2, \ldots, x_n$ die Knoten(punkte) und $w_1, w_2, \ldots, w_n$ die zugehörigen Gewichte der Gauß Quadraturformel für $[-1; 1]$ sind.

Die Ergebnisse sind in Tabelle 5.6 auf eine 10-stellige Gleitkommadarstellung gerundet angegeben, und die absoluten Fehler wurden auf eine 3-stellige Gleitkommadarstellung gerundet angegeben. Mit Q_6, also mit nur $n = 6$ Knoten, erreichen wir ein ausgezeichnetes Ergebnis mit neun signifikanten Ziffern. Ein vergleichbares Ergebnis wird von der Simpson-Regel für numerische Integration erst für ein $64 < 2n \leq 128$ erreicht. Wir sehen, wie effizient die Gauß Quadratur (verglichen mit der populären Simpson-Regel für numerische Integration) ist. ♠

Zuletzt sollen noch zwei relevante Dinge kurz besprochen werden:

Der erste Aspekt betrifft die **Berechnung der Knoten der Gauß Quadraturformeln**: Es ist nicht nötig, die Knoten $x_1, x_2, \ldots, x_n$ über die $2n$ nicht-linearen Gleichungen in (5.35) zu bestimmen, denn die Knoten von Q_n sind genau die n verschiedenen reellen Nullstellen des Legendre Polynoms P_n (und diese liegen auch alle im Intervall $[-1; 1]$). Die **Legendre Polynome** P_n, $n \in \mathbb{N}_0$, sind ein **System orthogonaler Polynome** mit den folgenden Eigenschaften:

(1) Für jedes $n \in \mathbb{N}_0$ gilt: P_n hat genau den Grad n.

(2) $\int_{-1}^{1} P_n(t)\, P_m(t)\, \mathrm{d}t = 0$ für alle $m, n \in \mathbb{N}_0$ mit $m \neq n$.

(3) $P_n(1) = 1$ für alle $n \in \mathbb{N}_0$.

Der zweite Aspekt betrifft eine Verallgemeinerung der Theorie: In manchen Anwendungen treten **Integrale mit einer Gewichtsfunktion** auf:

$$I(f) = \int_a^b f(x)\, w(x)\, \mathrm{d}t \qquad \text{mit stetigem} \qquad f : [a; b] \to \mathbb{R}, \tag{5.42}$$

wobei die **Gewichtsfunktion** $w : [a; b] \to \mathbb{R}$ nur nicht-negative Werte annimmt und höchstens an einzelnen Stellen in $[a; b]$ den Wert null annimmt. Ein Beispiel für ein Integral mit einer Gewichtsfunktion ist

$$\int_{-1}^{1} f(t)\, (1 - t^2)\, \mathrm{d}t \qquad \text{mit der Gewichtsfunktion} \quad w(t) = 1 - t^2.$$

In einem Anwendungsproblem könnte die Gewichtsfunktion w beispielsweise eine Dichtefunktion (Massendichte, Ladungsdichte) sein. Für Integrale (5.42) kann man dann auch Gauß Quadraturformeln konstruieren, bei denen der Effekt der Gewichtsfunktion direkt durch die Wahl der Knoten(punkte) und der zugehörigen Gewichte berücksichtigt ist.

ANHANG A

Grundlagen aus der Mittel- und Oberstufe

A.1 Mengen und Mengenoperationen

Wir wiederholen kurz die Mengennotation, sowie Mengenrelationen und Mengenoperationen.

Definition A.1. (Menge nach Georg Cantor)
*Eine **Menge** ist eine Zusammenfassung von bestimmten wohlunterschiedenen Objekten unserer Anschauung und unseres Denkens (welche die Elemente der Menge genannt werden) zu einem Ganzen.*

Beispiel A.2. (Zahlenmengen)

$\mathbb{N}$	Menge der natürlichen Zahlen: $1, 2, 3, \ldots$
$\mathbb{N}_0$	Menge der natürlichen Zahlen mit null: $0, 1, 2, 3, \ldots$
$\mathbb{Z}$	Menge der ganzen Zahlen: $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$
$\mathbb{Q}$	Menge der rationalen Zahlen (oder Brüche)
$\mathbb{R}$	Menge der reellen Zahlen

Diese Zahlenmengen werden hier als aus der bekannt vorausgesetzt. ♠

Beispiel A.3. (Mengen)

(a) $A = \{1; 2; 3; 7; 9; -16; \pi\}$

(b) $B = \{\alpha; \beta; \gamma\}$

(c) $C =$ Menge der Studierenden der Uni Paderborn im Sommersemester 2020

(d) $D = \{0; \alpha; 17\}$

(e) $E = \{\mathbb{N}; \mathbb{Z}; \mathbb{Q}; \mathbb{R}\}$

Mengen können also ganz unterschiedliche Objekte enthalten. ♠

Notation A.4. (Mengenklammen und Elementsymbol)
Man schreibt Mengen mit geschweiften Klammern „{“ und „}“, sogenannten **Mengenklammern**. Weiter schreiben wir:

- „$a \in M$“ für „a ist ein Element der Menge M“ oder kurz „a ist in M“.
- „$a \notin M$“ für „a ist kein Element der Menge M“ oder kurz „a ist nicht in M“.

Beispiel: Es gilt $0 \in \mathbb{N}_0$ und $0 \notin \mathbb{N}$. Es gilt $\sqrt{2} \in \mathbb{R}$ aber $\sqrt{2} \notin \mathbb{Q}$.

Man nennt „$\in$“ das **Elementsymbol**.

Definition A.5. (Gleichheit von Mengen und leere Menge)

*(1) Zwei Mengen A, B heißen **gleich** (in Zeichen: $A = B$), wenn sie dieselben Elemente enthalten. Sind zwei Mengen A, B **nicht gleich**, so schreiben wir in Zeichen $A \neq B$.*

*(2) Die Menge, die keine Elemente enthält, heißt die **leere Menge** und wird mit $\emptyset$ (oder auch $\{\}$) bezeichnet.*

Betrachten wir noch ein Beispiel, um uns klar zu machen, wie wir Mengen darstellen.

Bemerkung A.6. (Darstellung von Mengen)
Es sei M die Menge, deren Elemente die Zahlen $1, 2$ und 3 sind. Dann lässt sich M mathematisch beschreiben durch:

(1) Aufzählung ihrer Elemente:

$$M = \{1; 2; 3\}$$

aber es gilt z.B. auch

$$M = \{2; 1; 3\}, \qquad M = \{3; 2; 1\}, \qquad M = \{1; 3; 2\},$$

$$M = \{2; 3; 1\}, \qquad M = \{3; 1; 2\}, \qquad M = \{1; 2; 3; 1\}.$$

Die Reihenfolge, in der die Elemente aufgelistet werden, spielt keine Rolle. Ebenso ändert eine mehrfache Auflistung von Elementen die Menge nicht.

(2) Angabe einer Auswahleigenschaft:

$$M = \{x \in \mathbb{N} \ : \ x < 4\} = \{x \in \mathbb{Z} \ : \ 1 \leq x \leq 3\}$$

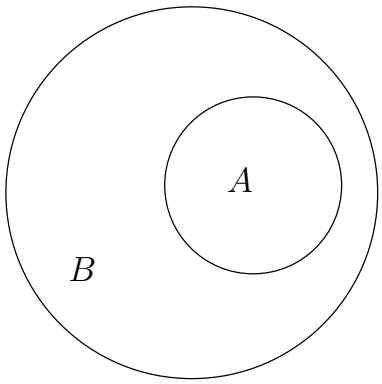
Dabei bedeutet „:“ in der obigen Zeile „für die gilt“ oder „so dass“.

Wir führen nun Relationen für Mengen ein.

Definition A.7. (Teilmenge und Obermenge)

*(1) Eine Menge A heißt eine **Teilmenge** einer Menge B (in Zeichen: $A \subseteq B$), wenn jedes Element von A auch in B liegt. B wird dann auch als eine **Obermenge** von A bezeichnet (in Zeichen: $B \supseteq A$).*

*(2) Ist $A \subseteq B$ und $A \neq B$, so heißt A eine **echte Teilmenge** von B (in Zeichen: $A \subsetneq B$). Ist $B \supseteq A$ und $B \neq A$, so heißt B eine **echte Obermenge** von A (in Zeichen: $B \supsetneq A$).*



Wir sehen: Wenn A eine **echte** Teilmenge von B ist, so ist B eine Obermenge von A, und B enthält **mindestens ein** Element, welches nicht in A ist.

Die Zeichnung neben Definition A.7 ist ein **Euler-Venn-Diagramm**, mit dem man Mengen veranschaulichen kann: Mengen werden als kreisförmige Gebilde dargestellt, und alles (also alle Elemente) innerhalb des kreisförmigen Gebildes gehören zu der jeweiligen Menge.

Beispiel A.8. (Teilmengen und Obermengen)

(a) Seien $A = \{1; 3\}$ und $B = \{1; 2; 3\}$. Dann gilt $A \subseteq B$ und $B \supseteq A$. Es gilt sogar $A \subsetneq B$ und $B \supsetneq A$.

(b) Es gilt $\mathbb{N} \subsetneq \mathbb{N}_0 \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}$. Natürlich gilt auch $\mathbb{N} \subseteq \mathbb{N}_0 \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$.

(c) Sei $C = \big\{1; 2; \{3; 4\}\big\}$. Dann gilt z.B. $\{1; 2\} \subseteq C$ und $\big\{\{3; 4\}\big\} \subseteq C$ und $\big\{1; \{3; 4\}\big\} \subseteq C$, aber $\{1; 2; 3\}$ ist keine Teilmenge. Es gilt auch nicht $\{3; 4\} \subseteq C$, denn die Menge $\{3; 4\}$ ist ein Element von C (und keine Teilmenge). – Genauer hat C die drei Elemente 1, 2 und $\{3; 4\}$. Jede Teilmenge

von C ist entweder die leere Menge oder eine Menge, welche ein, mehrere oder alle Elemente von C (als Elemente) enthält.

(d) Die leere Menge $\emptyset$ ist eine Teilmenge jeder Menge.

Überlegen Sie sich weitere Beispiele. ♠

Welche Teilmengen-Beziehungen gelten, wenn zwei Mengen gleich sind?

Bemerkung A.9. (gleiche Mengen)
Es gilt $A = B$ genau dann, wenn $A \subseteq B$ und $B \subseteq A$ ist. In Zeichen:

$$A = B \qquad \Longleftrightarrow \qquad \big(A \subseteq B \text{ und } \ B \subseteq A\big) \tag{A.1}$$

Notation A.10. (Pfeile und Doppelpfeile)

(1) Ein **Folgerungspfeil** „$\Longrightarrow$“ bedeutet „daraus folgt“. Also bedeutet

$$x = 2 \qquad \Longrightarrow \qquad x^2 = 4, \tag{A.2}$$

dass aus $x = 2$ die Gleichung $x^2 = 4$ folgt. Dieses ist das gleiche wie

$$x^2 = 4 \qquad \Longleftarrow \qquad x = 2.$$

Eine Aussage mit einem Folgerungspfeil nennt man eine **Implikation**.

(2) Der in (A.1) verwendete Doppelpfeil „$\Longleftrightarrow$“, genannt **Äquivalenzpfeil**, steht für „genau dann, wenn“ und bedeutet das gleiche, wie wenn wir die Formel zweimal hinschreiben, wobei „$\Longleftrightarrow$“ einmal durch „$\Longrightarrow$“ und einmal durch „$\Longleftarrow$“ ersetzt wird. Also bedeutet (A.1), dass die **beiden** folgenden Aussagen gelten:

$$A = B \qquad \Longrightarrow \qquad \big(A \subseteq B \text{ und } \ B \subseteq A\big),$$
$$\big(A \subseteq B \text{ und } \ B \subseteq A\big) \qquad \Longrightarrow \qquad A = B.$$

Eine Aussage mit einem Äquivalenzpfeil nennt man eine **Äquivalenz** (oder eine **Äquivalenzaussage**).

(3) Es ist **nicht egal**, ob man bei einer Reihe von mathematischen Überlegungen „$\Longrightarrow$“ oder „$\Longleftrightarrow$“ schreibt! In (A.2) ist es falsch, wenn wir schreiben

$$x = 2 \qquad \Longleftrightarrow \qquad x^2 = 4,$$

denn die Aussage

$$x^2 = 4 \qquad \Longrightarrow \qquad x = 2$$

ist falsch! Richtig wäre

$$x^2 = 4 \qquad \Longrightarrow \qquad \big(x = 2 \ \text{ oder } \ x = -2\big),$$

und es gilt sogar

$$x^2 = 4 \qquad \Longleftrightarrow \qquad \big(x = 2 \ \text{ oder } \ x = -2\big).$$

Zuletzt halten wir die Notation für Intervalle fest.

Definition A.11. (Intervalle)
Seien $a, b \in \mathbb{R}$ mit $a < b$. Die ***beschränkten Intervalle*** *sind:*

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

Die ***unbeschränkten Intervalle*** *sind:*

$$\begin{aligned}
[a;\infty[&= \{x \in \mathbb{R} \ : \ a \leq x\},\\
]a;\infty[&= \{x \in \mathbb{R} \ : \ a < x\},\\
]-\infty;b] &= \{x \in \mathbb{R} \ : \ x \leq b\},\\
]-\infty;b[&= \{x \in \mathbb{R} \ : \ x < b\},\\
]-\infty;\infty[&= \mathbb{R}.
\end{aligned}$$

Dabei steht das Symbol „∞“ für „unendlich“. Bei ∞ bzw. $-\infty$ steht immer die nach außen geöffnete Klammer, da weder ∞ noch $-\infty$ reelle Zahlen sind und daher nicht zum Intervall gehören.

Beispiel A.12. (Intervalle)

(a) $]-1;3] = \big\{x \in \mathbb{R} \ : \ -1 < x \leq 3\big\}$

Das Intervall ist nicht zu verwechseln mit der Menge $\{-1;3\}$ welche die Zahlen -1 und 3 als Elemente enthält.

(b) $]-\infty;0[= \big\{x \in \mathbb{R} \ : \ x < 0\big\}$ enthält alle negativen reellen Zahlen.

Dagegen ist $\{-\infty;0\}$ die Menge mit den zwei Elementen $-\infty$ und 0.

Es ist wichtig bei der Notation genau hinzuschauen, ob Intervallklammern oder Mengenklammern dastehen. ♠

Bemerkung A.13. (alternative Intervallnotation)

Es gibt auch eine alternative Intervallnotation für die offenen und halboffenen Intervalle: Für $a, b \in \mathbb{R}$ mit $a < b$ schreibt man

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

Die abgeschlossenen Intervalle schreibt man wie in Definition A.11.

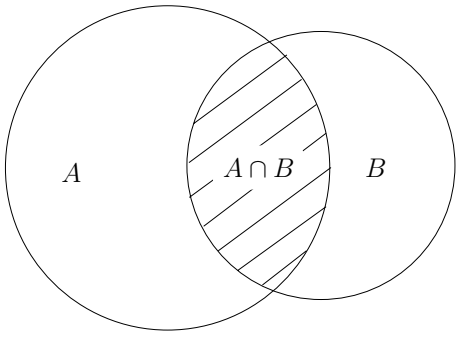
Nun lernen wir die folgenden Mengenoperationen kennen: das Schneiden und das Vereinigen von Mengen, sowie das Bilden einer Differenzmenge. Alle Mengenoperationen werden durch Euler-Venn-Diagramme veranschaulicht.

Definition A.14. (Durchschnitt von Mengen und disjunkte Mengen)

(1) Der ***Durchschnitt*** $A \cap B$ *der Mengen* A, B *ist die Menge*

$$A \cap B = \{x \ : \ x \in A \text{ und } x \in B\}.$$

(2) Ist $A \cap B = \emptyset$*, so heißen* A, B ***disjunkt****.*



Der Durchschnitt zweier Mengen wird auch ***Schnittmenge*** *genannt.*

Es gilt immer $A \cap B = B \cap A$.

Beispiel A.15. (Durchschnitt von Mengen)

(a) $\{1;2;3\} \cap \{3;4\} = \{3\}$

(b) $\mathbb{Z} \cap [0;\infty[= \mathbb{N}_0$

(c) $[2;7] \cap]3;11[=]3;7]$

(d) Die natürlichen Zahlen $\mathbb{N}$ und das Intervall $[-1;0]$ sind disjunkt, denn $\mathbb{N} \cap [-1;0] = \emptyset$.

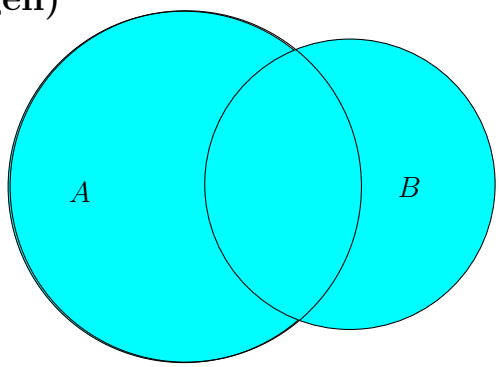
Überlegen Sie sich weitere Beispiele. ♠

Definition A.16. (Vereinigung von Mengen)

*Die **Vereinigung** $A \cup B$ der Mengen A, B ist die Menge*

$$A \cup B = \{x \ : \ x \in A \text{ oder } x \in B\}.$$

Mit „oder“ ist das „einschließende oder“ gemeint (und nicht „entweder ... oder“).



Also: $x \in A \cup B$ liegt in A oder in B oder auch in beiden Mengen A und B.

Es gilt immer $A \cup B = B \cup A$.

Beispiel A.17. (Vereinigung von Mengen)

(a) $\{1; 2; 3\} \cup \{3; 4\} = \{1; 2; 3; 4\}$

(b) $]-\infty; 0[\cup \{0\} \cup]0; \infty[= \mathbb{R}$

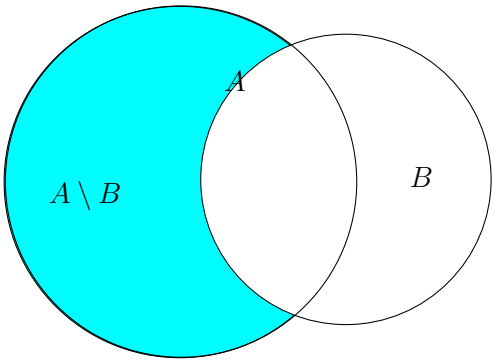
Überlegen Sie sich weitere Beispiele. ♠

Definition A.18. (Differenz von Mengen)

*(1) Die **Differenz** $A \setminus B$ („A ohne B“) der Mengen A, B ist die Menge*

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

*(2) Ist $B \subseteq A$, so heißt $A \setminus B$ auch das **Komplement von B in A**.*



Beispiel A.19. (Differenz von Mengen)

(a) $\{1; 2; 3\} \setminus \{3; 4\} = \{1; 2\}, \quad \{3; 4\} \setminus \{1; 2; 3\} = \{4\}$

(b) $\mathbb{Z} \setminus \mathbb{N}_0 = \{-1; -2; -3; \ldots\} = \{-n \ : \ n \in \mathbb{N}\}$

Wir sehen, dass im Allgemeinen gilt: $A \setminus B \neq B \setminus A$ ♠

Bemerkung A.20. (Ausführen mehrerer Mengenoperationen)
Wendet man mehrere Mengenoperationen an, so ist es ganz wichtig, dass **Klammern gesetzt sind**, damit klar ist, in welcher Reihenfolge die Mengenoperationen auszuführen sind! Beispielsweise gilt im Allgemeinen (d.h. bis auf mögliche Sonderfälle)

$$A \setminus (B \cup C) \neq (A \setminus B) \cup C,$$
$$A \cap (B \cup C) \neq (A \cap B) \cup C,$$

wie man sich leicht an den folgenden Beispielen klar macht:

Seien $A = \{1; 2; 3; 4\}$, $B = \{-1; 0; 1; 2\}$ und $C = \{3; 4; 5; 6\}$. Dann gilt

$$\begin{aligned} A \setminus (B \cup C) &= \{1; 2; 3; 4\} \setminus \{-1; 0; 1; 2; 3; 4; 5; 6\} = \emptyset \\ &\neq (A \setminus B) \cup C = \{3; 4\} \cup \{3; 4; 5; 6\} = \{3; 4; 5; 6\}, \end{aligned}$$

$$\begin{aligned} A \cap (B \cup C) &= \{1; 2; 3; 4\} \cap \{-1; 0; 1; 2; 3; 4; 5; 6\} = \{1; 2; 3; 4\} \\ &\neq (A \cap B) \cup C = \{1; 2\} \cup \{3; 4; 5; 6\} = \{1; 2; 3; 4; 5; 6\}. \end{aligned}$$

Es gibt Ausnahmen, bei denen man keine Klammen setzen muss, nämlich wenn alle Mengenoperationen $\cup$ oder wenn alle Mengenoperationen $\cap$ sind:

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

Weil die Klammersetzung hier keine Rolle spielt, lässt man sie (wie jeweils im Ausdruck ganz rechts) in der Regel weg.

A.2 Rechnen mit reellen Zahlen

In diesem Teilkapitel seien a, b, c reelle Zahlen.

Es gelten die folgenden Rechenregeln für die **Addition** reeller Zahlen:

(1) **Assoziativgesetz**: $(a + b) + c = a + (b + c)$
(2) **Kommutativgesetz**: $a + b = b + a$

Wegen des Assoziativgesetzes der Addition spielt es keine Rolle, in welcher Reihenfolge wir die Additionen ausführen. Wir dürfen daher die Klammern auch einfach weglassen, also:

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

Es gelten die folgenden Rechenregeln für die **Multiplikation** reeller Zahlen:

(1) **Assoziativgesetz**: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$

(2) **Kommutativgesetz**: $a \cdot b = b \cdot a$

Wegen des Assoziativgesetzes der Multiplikation spielt es keine Rolle, in welcher Reihenfolge wir die Multiplikationen ausführen. Wir dürfen daher die Klammern auch einfach weglassen, also:

$$(a \cdot b) \cdot c = a \cdot (b \cdot c) = a \cdot b \cdot c$$

Weiter gelten die beiden **Distributivgesetze**:

$$(a + b) \cdot c = a \cdot c + b \cdot c,$$
$$a \cdot (b + c) = a \cdot b + a \cdot c.$$

Generell gilt **„Punkt(rechnung) vor Strich(rechnung)“**, d.h. ist ein Ausdruck mit Multiplikationen oder Divisionen (also Punktrechnungen) und Additionen oder Subtraktionen (also Strichrechnungen) gegeben, so müssen die Punktrechnungen zuerst ausgeführt werden, wenn es nicht durch Klammersetzung anders vorgegeben ist.

Beispiel A.21. („Punkt vor Strich“, Klammersetzung)

$$13 + 2 \cdot 4 + 7 = 13 + 8 + 7 = 28,$$
$$(13 + 2) \cdot 4 + 7 = 15 \cdot 4 + 7 = 60 + 7 = 67.$$

Der Unterschied durch die geänderte Reihenfolge der Ausführung der Rechenoperationen ist deutlich im Ergebnis sichtbar. Die Klammersetzung spielt eine entscheidende Rolle! ♠

Auch wenn man nur Additionen und Subtraktionen hat, spielt die Klammersetzung eine Rolle, denn es gelten:

$$a - (b + c) = a - b - c,$$
$$a - (b - c) = a - b + c,$$
$$a + (b - c) = a + b - c.$$

Dieses folgt aus den obigen Gesetzen, indem man die Subtraktion mittels

$$a - b = a + (-1) \cdot b$$

als Addition auffasst:

$$\begin{aligned} a-(b+c) &= a+(-1)\cdot(b+c) = a+\big((-b)+(-c)\big) \\ &= a+(-b)+(-c) = a-b-c, \end{aligned}$$

$$\begin{aligned} a-(b-c) &= a+(-1)\cdot\big(b+(-c)\big) = a+\big((-b)+c\big) \\ &= a+(-b)+c = a-b+c, \end{aligned}$$

$$a+(b-c) = a+\big((b+(-c)\big) = a+b+(-c) = a+b-c.$$

A.3 Bruchrechnung

In der Unter- und Mittelstufe lernt man die rationalen Zahlen $\mathbb{Q}$, also die Menge aller Zahlen der Form

$$\frac{m}{n} \qquad \text{mit} \qquad m,n\in\mathbb{Z}, \quad \text{wobei} \quad n\neq 0,$$

kennen. Solche Zahlen nennen wir **Brüche**. Später werden in der Schule auch „Brüche“ betrachtet, deren Zähler und Nenner nicht mehr in $\mathbb{Z}$ sondern beliebige reelle Zahlen sind, wobei der Nenner natürlich nach wie vor ungleich Null sein muss, also z.B.

$$\frac{\pi}{2}, \qquad \frac{\sqrt{2}}{2} \qquad \text{oder} \qquad \frac{e}{\sqrt{7}}.$$

Für das Rechnen mit Brüchen gelten die Rechenregeln aus dem nachfolgenden Hilfssatz.

Hilfssatz A.22. (Rechenregeln der Bruchrechnung)

Folgende Rechenregeln gelten für reelle Zahlen a, b, c und d:

(*i*) *Erweitern und Kürzen:* $\dfrac{a}{b} = \dfrac{a\cdot c}{b\cdot c}$

(*ii*) *Addition von Brüchen mit gleichem Nenner:* $\dfrac{a}{b}+\dfrac{c}{b} = \dfrac{a+c}{b}$

(*iii*) *Addition allgemeiner Brüche:* $\dfrac{a}{b}+\dfrac{c}{d} = \dfrac{a\cdot d+c\cdot b}{b\cdot d}$

(*iv.a*) *Multiplikation von Brüchen:* $\dfrac{a}{b}\cdot\dfrac{c}{d} = \dfrac{a\cdot c}{b\cdot d}$

(*iv.b*) *Multiplikation mit* $a \in \mathbb{R}$*:* $\quad a \cdot \dfrac{b}{c} = \dfrac{a}{1} \cdot \dfrac{b}{c} = \dfrac{a \cdot b}{c}$

(*v*) *Kehrwert:* $\quad \dfrac{\frac{1}{c}}{\frac{d}{}} = 1 : \dfrac{c}{d} = 1 \cdot \dfrac{d}{c} = \dfrac{d}{c}$

(*vi*) *Doppelbruch:* $\quad \dfrac{\frac{a}{b}}{\frac{c}{d}} = \dfrac{a}{b} : \dfrac{c}{d} = \dfrac{a}{b} \cdot \dfrac{d}{c} = \dfrac{a \cdot d}{b \cdot c}$

Dabei gilt: ***Kein Nenner darf dabei gleich Null sein!*** *Bei Doppelbrüchen setzt man den Hauptbruchstrich auf Höhe des Gleichheitszeichens.*

Betrachten wir einige Zahlenbeispiele für die Rechenregeln aus Hilfssatz A.22.

Beispiel A.23. (Rechenregeln der Bruchrechnung)

(a) $\dfrac{28}{32} = \dfrac{7 \cdot 4}{8 \cdot 4} = \dfrac{7}{8}$

(b) $\dfrac{2}{5} + \dfrac{1}{5} = \dfrac{2+1}{5} = \dfrac{3}{5}$

(c) $\dfrac{1}{3} + \dfrac{1}{2} = \dfrac{2 \cdot 1}{2 \cdot 3} + \dfrac{3 \cdot 1}{3 \cdot 2} = \dfrac{2}{6} + \dfrac{3}{6} = \dfrac{5}{6}$

(d) $\dfrac{33}{28} \cdot \dfrac{4}{11} = \dfrac{33 \cdot 4}{28 \cdot 11} = \dfrac{3 \cdot 11 \cdot 4}{7 \cdot 4 \cdot 11} = \dfrac{3}{7}$

(e) $3 \cdot \dfrac{7}{9} = \dfrac{3 \cdot 7}{9} = \dfrac{3 \cdot 7}{3 \cdot 3} = \dfrac{7}{3}$

(f) $\dfrac{\frac{1}{8}}{\frac{}{33}} = \dfrac{33}{8}$

(g) $\dfrac{\frac{2}{21}}{\frac{4}{7}} = \dfrac{2}{21} \cdot \dfrac{7}{4} = \dfrac{2 \cdot 7}{21 \cdot 4} = \dfrac{2 \cdot 7}{3 \cdot 7 \cdot 2 \cdot 2} = \dfrac{1}{6}$

Überlegen Sie sich weitere Beispiele. ♠

Wichtig ist zu beachten, dass die folgenden **falschen** Regeln **nicht** gelten:

$$\frac{a+b}{a} \neq \frac{1+b}{a}, \qquad \text{sondern korrekt ist} \qquad \frac{a+b}{a} = \frac{a}{a} + \frac{b}{a} = 1 + \frac{b}{a},$$

$$\frac{a}{a+b} \neq \frac{1}{1+b},$$

$$\frac{a+b}{c+d} \neq \frac{a}{c} + \frac{b}{d}.$$

A.4 Rechnen mit Ungleichungen

Eine Ungleichung ist eine Formel mit reellen Zahlen und/oder reellwertigen Variablen/Parametern (also Buchstaben), in der ein Ungleichheitszeichen vorkommt, also $<$, $\leq$, $>$ oder $\geq$ (oder $\neq$, wobei dieser Fall selten von Interesse ist).

Beispiel A.24. (Ungleichungen)

(a) $2 < 3$ und $5 \leq 4$ und $2 \leq 3$ sind alles Ungleichungen. Dabei sind $2 < 3$ und $2 \leq 3$ wahre Ungleichungen, wogegen die Ungleichung $5 \leq 4$ falsch ist.

(b) $a < b$ und $x^2 + 2\,x - 4 \geq 0$ sind ebenfalls Ungleichungen. Diese sind wahr für geeignete Wahlen von a und b bzw. x.

(c) Die Ungleichung $x^2 \geq 0$ ist immer wahr, d.h. egal wie man $x \in \mathbb{R}$ wählt (weil Quadrate immer positiv sind), und die Ungleichung $y^2 < 0$ für kein $y \in \mathbb{R}$ wahr.

Ungleichungen werden uns oft in diesem Kurs begegnen. ♠

Beinhaltet eine Ungleichung Variablen/Parameter (also Buchstaben), so ist das Ziel in der Regel herauszufinden für welche Zahlenwerte der Variablen/Parameter die Ungleichung erfüllt ist. Wir wollen die Ungleichung also auflösen. Dazu müssen wir die Regeln für das Rechnen mit Ungleichungen beherrschen.

Hilfssatz A.25. (Regeln für das Rechnen mit Ungleichungen)

(1) Für alle $a, b, c \in \mathbb{R}$ gelten:

$$a < b \quad \Longleftrightarrow \quad a + c < b + c$$
$$a > b \quad \Longleftrightarrow \quad a + c > b + c$$

$$a \leq b \quad \Longleftrightarrow \quad a + c \leq b + c$$
$$a \geq b \quad \Longleftrightarrow \quad a + c \geq b + c$$

In Worten: Man darf auf beiden Seiten einer Ungleichung die gleiche reelle Zahl (bzw. den gleichen reellwertigen Term) addieren.

(2) Für alle $a, b \in \mathbb{R}$ und alle $c > 0$ gelten:

$$a < b \quad \Longleftrightarrow \quad c \cdot a < c \cdot b$$
$$a > b \quad \Longleftrightarrow \quad c \cdot a > c \cdot b$$
$$a \leq b \quad \Longleftrightarrow \quad c \cdot a \leq c \cdot b$$
$$a \geq b \quad \Longleftrightarrow \quad c \cdot a \geq c \cdot b$$

In Worten: Man darf auf beiden Seiten einer Ungleichung mit der gleichen positiven reellen Zahl (bzw. mit dem gleichen positiven reellwertigen Term) multiplizieren.

(3) Für alle $a, b \in \mathbb{R}$ und alle $c < 0$ gelten:

$$a < b \quad \Longleftrightarrow \quad c \cdot a > c \cdot b$$
$$a > b \quad \Longleftrightarrow \quad c \cdot a < c \cdot b$$
$$a \leq b \quad \Longleftrightarrow \quad c \cdot a \geq c \cdot b$$
$$a \geq b \quad \Longleftrightarrow \quad c \cdot a \leq c \cdot b$$

In Worten: Multipliziert man beide Seiten einer Ungleichung mit der gleichen negativen reellen Zahl (bzw. mit dem gleichen negativen reellwertigen Term), so kehrt sich das Ungleichheitszeichen um.

Achtung: Merken Sie sich Hilfssatz A.25 (3) gut: Wenn man eine Ungleichung mit einer **negativen** reellen Zahl (bzw. einem **negativen** reellwertigen Term) multipliziert, so **kehrt sich das Ungleichheitszeichen um!**

Betrachten wir einige elementare Beispiele.

Beispiel A.26. (Rechnen mit Ungleichungen)

(a) $7 > 5 \quad \Longleftrightarrow \quad 7 + 3 > 5 + 3 \quad \Longleftrightarrow \quad 10 > 8$

(b) Subtraktion von $a \in \mathbb{R}$ realisieren wir als Addition von $-a$, also z.B.:

$$7 \geq 5 \quad \Longleftrightarrow \quad 7 + (-5) \geq 5 + (-5) \quad \Longleftrightarrow \quad 2 \geq 0$$

(c) $-\frac{1}{2} < -\frac{1}{3} \iff -\frac{1}{2} \cdot 6 < -\frac{1}{3} \cdot 6 \iff -\frac{6}{2} < -\frac{6}{3} \iff -3 < -2$

(d) $-1 < 2 \iff (-1) \cdot (-1) > 2 \cdot (-1) \iff 1 > -2$

(e) Division durch eine reelle Zahl $a \neq 0$ realisieren wir als Multiplikation mit $1/a$, also z.B. für Division durch 2 und danach Division durch 3:

$$2 \leq 3 \iff 2 \cdot \frac{1}{2} \leq 3 \cdot \frac{1}{2} \iff 1 \leq \frac{3}{2}$$

$$\iff 1 \cdot \frac{1}{3} \leq \frac{3}{2} \cdot \frac{1}{3} \iff \frac{1}{3} \leq \frac{1}{2}$$

Überlegen Sie sich weitere Beispiele. ♠

Betrachten wir einige anspruchsvollere Beispiele.

Beispiel A.27. (Lösen von Ungleichungen)

(a) Welche $x \in \mathbb{R}$ erfüllen die Ungleichung

$$x^2 \leq 2\,x + 3?$$

Lösung: Wir formen um:

$$x^2 \leq 2\,x+3 \;\Big|\; -2\,x+1 \iff x^2-2\,x+1 \leq 4 \iff (x-1)^2 \leq 4,$$

wobei wir im letzten Schritt die zweite binomische Formel auf der linken Seite angewendet haben. Da Quadrate immer nicht-negativ sind, kann die Ungleichung nur gelten, wenn gilt

$$(x-1)^2 \leq 4 \iff -2 \leq x-1 \leq 2 \;\Big|\; +1 \iff -1 \leq x \leq 3.$$

Also ist die Lösungsmenge $\mathbb{L} = \big\{x \in \mathbb{R} \;:\; -1 \leq x \leq 3\big\} = [-1;3]$.

Wir hätten die Ungleichung auch wie folgt lösen können:

$$x^2 \leq 2\,x + 3 \;\Big|\; -2\,x-3 \iff x^2 - 2\,x - 3 \leq 0$$

$$\iff x^2 - 2\,x + 1 - 4 \leq 0 \overset{\text{2. binom. Formel}}{\iff} (x-1)^2 - 2^2 \leq 0$$

$$\overset{\text{3. binom. Formel}}{\iff} (x-1-2)\,(x-1-2) \leq 0 \iff (x-3)\,(x+1) \leq 0$$

Wann ist die linke Seite ≤ 0?

- Wenn die linke Seite $= 0$ ist, also für $x = 3$ oder $x = -1$, oder
- wenn $x - 3 < 0$ und $x + 1 > 0$, also wenn $x < 3$ und $x > -1$, d.h. wenn $-1 < x < 3$, oder
- wenn $x - 3 > 0$ und $x + 1 < 0$, also wenn $x > 3$ und $x < -1$ ist.

Im letzten Fall gibt es aber keine x, die die beiden Ungleichungen $x > 3$ und $x < -1$ erfüllen. Aus den ersten zwei Fällen entnehmen wir, dass die Lösungsmenge $\mathbb{L} = \{x \in \mathbb{R} \ : \ -1 \leq x \leq 3\} = [-1; 3]$ ist.

(b) Welche $x \in \mathbb{R}$ erfüllen die Ungleichung

$$\frac{2\,x - 3}{x - 3} \geq 4\,? \tag{A.3}$$

Lösung: Die Zahl $x = 3$ muss vorab ausgeschlossen werden, weil sonst auf der linken Seite durch null dividiert wird!

Im Folgenden unterscheiden wir zwei Fälle:

- $x < 3$ bzw. $x - 3 < 0$,
- $x > 3$ bzw. $x - 3 > 0$.

Wir suchen die Lösungen der Ungleichungen für jeden Fall separat.

- *Fall $x < 3$:* Multiplikation der Ungleichung (A.3) mit $x - 3 < 0$ ergibt („$\geq$“ wird umgekehrt):

$$\begin{aligned} & (x-3) \cdot \frac{2\,x-3}{x-3} \leq 4\,(x-3) \\ \iff \quad & 2\,x - 3 \leq 4\,x - 12 \ \Big| \ + 12 - 2\,x \\ \iff \quad & 9 \leq 2\,x \quad \iff \quad \frac{9}{2} \leq x\,. \end{aligned}$$

  Die Ungleichung $9/2 \leq x$ steht aber im Widerspruch zu $x < 3$ (da $3 < 9/2 = 4{,}5$). Es gibt also keine Lösung der Ungleichung (A.3) mit $x < 3$, d.h. die Lösungsmenge in diesem Fall ist $\mathbb{L}_1 = \emptyset$.

- *Fall $x > 3$:* Multiplikation der Ungleichung (A.3) mit $x - 3 > 0$ ergibt („$\geq$“ bleibt erhalten):

$$\begin{aligned} & (x-3) \cdot \frac{2\,x-3}{x-3} \geq 4\,(x-3) \\ \iff \quad & 2\,x - 3 \geq 4\,x - 12 \ \Big| \ + 12 - 2\,x \\ \iff \quad & 9 \geq 2\,x \quad \iff \quad \frac{9}{2} \geq x\,. \end{aligned}$$

Alle x mit $x > 3$ und $x \leq 9/2$, also alle x mit $3 < x \leq 9/2$, erfüllen die Ungleichung (A.3), d.h. die Lösungsmenge in diesem Fall ist $\mathbb{L}_2 = \left]3; \frac{9}{2}\right]$.

Die Lösungsmenge von (A.3) ist also

$$\mathbb{L} = \mathbb{L}_1 \cup \mathbb{L}_2 = \emptyset \cup \left]3; \frac{9}{2}\right] = \left]3; \frac{9}{2}\right].$$

(c) Für welche $x \in \mathbb{R}$ gilt die folgende Ungleichung?

$$\frac{2\,x+1}{x-1} < 1 \tag{A.4}$$

Lösung: Zunächst müssen wir $x - 1 = 0$, also $x = 1$, ausschließen, da durch null teilen verboten ist. Wir wollen nun beide Seiten der Gleichung (A.4) mit $(x-1)$ multiplizieren. Dabei müssen wir aber das Vorzeichen von $(x-1)$ berücksichtigen, da sich für $x - 1 < 0$ das „<“ in (A.4) in ein „>“ umkehrt. Wir nehmen also eine Fallunterscheidung vor:

- *Fall 1:* Für $x - 1 > 0$, also für $x > 1$, erhalten wir nach der Multiplikation von (A.4) mit $(x - 1)$

$$2\,x + 1 < x - 1 \quad \Big| \; -x - 1 \qquad \Longleftrightarrow \qquad x < -2.$$

  Da $x < -2$ mit $x > 1$ nicht vereinbar ist, gibt es in diesem Fall keine Lösungen, also $\mathbb{L}_1 = \emptyset$.
- *Fall 2:* Für $x - 1 < 0$, also für $x < 1$, erhalten wir nach der Multiplikation von (A.4) mit $(x - 1)$

$$2\,x + 1 > x - 1 \quad \Big| \; -x - 1 \qquad \Longleftrightarrow \qquad x > -2.$$

  Wir erhalten also die beiden Bedingung $x > -2$, also $-2 < x$, und $x < 1$ an x, d.h. $-2 < x < 1$. Dieser Fall liefert $\mathbb{L}_2 =]-2; 1[$.

Fazit: Die Lösungsmenge der Ungleichung (A.4) ist das offene Intervall

$$\mathbb{L} = \mathbb{L}_1 \cup \mathbb{L}_2 =]-2; 1[.$$

Wir sehen, dass die Fallunterscheidungen ein wesentlicher Teil des Lösungsprozesses sind. ♠

A.5 Der Absolutbetrag

Als Nächstes lernen wir den Betrag (oder Absolutbetrag) kennen.

Definition A.28. (Betrags/Absolutbetrag)
*Für $x \in \mathbb{R}$, ist der **Betrag** (oder der **Absolutbetrag**) $|x|$ definiert durch*

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

Anschauung: $|x|$ misst den Abstand (auf der Zahlengeraden) von x zum Nullpunkt 0.

Man sollte die Eigenschaften des Betrags kennen, die im nächsten Hilfssatz zusammengestellt sind.

Hilfssatz A.29. (Eigenschaften des Betrags)

(1) $|x| \geq 0$ *für alle* $x \in \mathbb{R}$.

(2) $|x| = 0$ *gilt in* $\mathbb{R}$ *genau dann, wenn* $x = 0$ *ist.*

(3) $|\lambda \cdot x| = |\lambda| \cdot |x|$ *für alle* $\lambda, x \in \mathbb{R}$

(4) $|x + y| \leq |x| + |y|$ *für alle* $x, y \in \mathbb{R}$ *(**Dreiecksungleichung**).*

Aus diesen grundlegenden Eigenschaften folgen die abgeleiteten weiteren Eigenschaften in nächsten Hilfssatz.

Hilfssatz A.30. (weitere Eigenschaften des Betrags)
Weitere Eigenschaften des (Absolut-)Betrags $|\cdot|$ sind:

(5) $|-x| = |x|$ *für alle* $x \in \mathbb{R}$.

(6) $x^2 = |x|^2$ *für alle* $x \in \mathbb{R}$.

(7) $-|x| \leq x \leq |x|$ *für alle* $x \in \mathbb{R}$.

(8) Für $c \geq 0$ gilt: $|x| < c \iff -c < x < c$

(9) Für $c \geq 0$ gilt: $|x| \leq c \iff -c \leq x \leq c$

Hilfssatz A.30 (8) und (9) sind beim Auflösen von Ungleichungen wichtig, wie wir

in den nachfolgenden Beispielen noch sehen werden.

Aufgrund der „stückweisen“ Definition von $|x|$ **können Gleichungen und Ungleichungen in denen $|x|$ vorkommt, am besten durch Fallunterscheidung aufgelöst werden!** Betrachten wir dazu ein paar Beispiele.

Beispiel A.31. (Gleichungen und Ungleichungen mit Absolutbeträgen)

(a) Für welche $x \in \mathbb{R}$ gilt die Ungleichung $|x-3| \leq 1$?

Anschaulich besagt $|x-3| \leq 1$, dass der Abstand von x zu 3 kleiner oder gleich 1 ist. Durch Einzeichnen auf dem Zahlenstrahl erhält man als Lösungsmenge das abgeschlossene Intervall

$$\mathbb{L} = [2;4] = \big\{x \in \mathbb{R} \ : \ 2 \leq x \leq 4\big\}.$$

Lösung: Wir wollen unser anschaulich ermitteltes Ergebnis nun durch mathematische Berechnungen nachweisen:

Wegen Hilfssatz A.30 (9) gilt

$$|x-3| \leq 1 \quad \Longleftrightarrow \quad -1 \leq x-3 \leq 1 \ \Big| \ +3 \quad \Longleftrightarrow \quad 2 \leq x \leq 4,$$

d.h. die Lösungsmenge ist $\mathbb{L} = \big\{x \in \mathbb{R} \ : \ 2 \leq x \leq 4\big\} = [2;4]$.

Alternative Lösung: Falls man Hilfssatz A.30 (9) nicht im Kopf hat, kann man auch direkt mit der Definition des Betrags mit einer Fallunterscheidung arbeiten:

- *Fall 1:* Gilt $x-3 \geq 0$, also $x \geq 3$, dann ist $|x-3| = x-3$. Also wird $|x-3| \leq 1$ zu

$$x-3 \leq 1 \ \Big| \ +3 \quad \Longleftrightarrow \quad x \leq 4.$$

Die Ungleichungen $x \geq 3$ und $x \leq 4$ liefern als Lösungsmenge das Intervall

$$\mathbb{L}_1 = \big\{x \in \mathbb{R} \ : \ 3 \leq x \leq 4\big\} = [3;4].$$

- *Fall 2:* Gilt $x-3 < 0$, also $x < 3$, dann ist $|x-3| = -(x-3) = 3-x$, und die Gleichung $|x-3| \leq 1$ wird

$$3-x \leq 1 \ \Big| \ +x-1 \quad \Longleftrightarrow \quad 2 \leq x.$$

Die Ungleichungen $x < 3$ und $2 \leq x$ liefern als Lösungsmenge das Intervall

$$\mathbb{L}_2 = \big\{x \in \mathbb{R} \ : \ 2 \leq x < 3\big\} = [2;3[\,.$$

Fazit: Vereinigen wir die Lösungen aus Fall 1 und Fall 2, so erhalten wir als Lösungsmenge das Intervall

$$\mathbb{L} = \mathbb{L}_1 \cup L_2 = [3;4] \cup [2;3[= [2;4] = \big\{x \in \mathbb{R} \ : \ 2 \leq x \leq 4\big\}.$$

(b) Was sind die Lösungen der Gleichung $|x-2| = |2\,x|$?

Lösung: Wir treffen drei Fallunterscheidungen: $x \geq 2$, $0 \leq x < 2$ und $x < 0$.

- *Fall 1:* Für $x \geq 2 \iff x - 2 \geq 0$ ist auch $2\,x \geq 0$, und es gilt:

$$|x-2| = |2\,x| \quad \iff \quad x - 2 = 2\,x \ \Big| \ -x \quad \iff \quad -2 = x.$$

  Da für $x \geq 2$ der Fall $x = -2$ nicht auftreten kann, haben wir hier keine Lösung, also $\mathbb{L}_1 = \emptyset$.
- *Fall 2:* Für $x < 2$ und $x \geq 0$, also für $0 \leq x < 2$ ist $x - 2 < 0$ und $2\,x \geq 0$ und damit $|x-2| = -(x-2) = 2 - x$ und $|2\,x| = 2\,x$. Somit gilt

$$|x-2| = |2\,x| \quad \iff \quad 2 - x = 2\,x \ \Big| \ +x$$
$$\iff \quad 2 = 3\,x \ \Big| \ :3 \quad \iff \quad x = \frac{2}{3}\,.$$

  Da $x = 2/3$ die Bedingung $0 \leq x < 2$ erfüllt, ist $x = 2/3$ eine Lösung, also $\mathbb{L}_2 = \big\{\frac{2}{3}\big\}$.
- *Fall 3:* Sei nun $x < 0$. Dann ist $x - 2 < 0$ und $2\,x < 0$ und damit $|x-2| = -(x-2) = 2 - x$ und $|2\,x| = -2\,x$. Somit finden wir

$$|x-2| = |2\,x| \quad \iff \quad 2 - x = -2\,x \ \Big| \ +x$$
$$\iff \quad 2 = -x \ \Big| \ \cdot(-1) \quad \iff \quad -2 = x.$$

  Da $x = -2$ die Bedingung $x < 0$ erfüllt, ist $x = -2$ ebenfalls eine Lösung, also $\mathbb{L}_3 = \{-2\}$.

Fazit: Wir haben mit den drei betrachteten Fällen alle reellen Zahlen x abgedeckt. Daher ist die Lösungsmenge von $|x-2| = |2\,x|$

$$\mathbb{L} = \mathbb{L}_1 \cup \mathbb{L}_2 \cup \mathbb{L}_3 = \left\{-2; \frac{2}{3}\right\}.$$

(c) Wir wollen alle Punkte $(x;y)$ in der Ebene finden, für die gilt

$$|x| + |y| \leq 1.$$

Dazu müssen wir die folgenden vier Fälle betrachten:

- *Fall 1:* $x \geq 0$ und $y \geq 0$ (d.h. wir befinden uns im 1. Quadranten): Dann gelten $|x| = x$ und $|y| = y$ und somit

$$|x| + |y| = x + y \leq 1,$$

und wir finden durch Auflösen nach y die Ungleichung $y \leq 1 - x$. Hier erhalten wir die Teillösungsmenge

$$\mathbb{L}_1 = \big\{(x; y) \ : \ x \geq 0 \text{ und } 0 \leq y \leq 1 - x\big\}.$$

- *Fall 2:* $x < 0$ und $y \geq 0$ (d.h. wir befinden uns im 2. Quadranten): Dann gelten $|x| = -x$ und $|y| = y$ und somit

$$|x| + |y| = -x + y \leq 1,$$

und wir finden durch Auflösen nach y die Ungleichung $y \leq 1 + x$. Hier erhalten wir die Teillösungsmenge

$$\mathbb{L}_2 = \big\{(x; y) \ : \ x < 0 \text{ und } 0 \leq y \leq 1 + x\big\}.$$

- *Fall 3:* $x < 0$ und $y < 0$ (d.h. wir befinden uns im 3. Quadranten): Dann gelten $|x| = -x$ und $|y| = -y$ und somit

$$|x| + |y| = -x - y \leq 1,$$

und wir finden durch Auflösen nach y die Ungleichung $y \geq -x - 1$. Hier erhalten wir die Teillösungsmenge

$$\mathbb{L}_3 = \big\{(x; y) \ : \ x < 0 \text{ und } -x - 1 \leq y < 0\big\}.$$

- *Fall 4:* $x \geq 0$ und $y < 0$ (d.h. wir befinden uns im 4. Quadranten): Dann gelten $|x| = x$ und $|y| = -y$ und somit

$$|x| + |y| = x - y \leq 1,$$

und wir finden durch Auflösen nach y die Ungleichung $y \geq x - 1$. Hier erhalten wir die Teillösungsmenge

$$\mathbb{L}_4 = \big\{(x; y) \ : \ x \geq 0 \text{ und } x - 1 \leq y < 0\big\}.$$

Fazit: Die Lösungsmenge ist also

$$\begin{aligned}\mathbb{L} &= \mathbb{L}_1 \cup \mathbb{L}_2 \cup \mathbb{L}_3 \cup \mathbb{L}_4 \\ &= \left\{(x; y) \ : \ \begin{array}{l} \big(x \geq 0 \text{ und } x - 1 \leq y \leq -x + 1\big) \text{ oder} \\ \big(x < 0 \text{ und } -x - 1 \leq y \leq x + 1\big) \end{array} \right\}\end{aligned}$$

Die Lösungsmenge aus (c) ist in Abbildung A.1 grafisch dargestellt. ♠

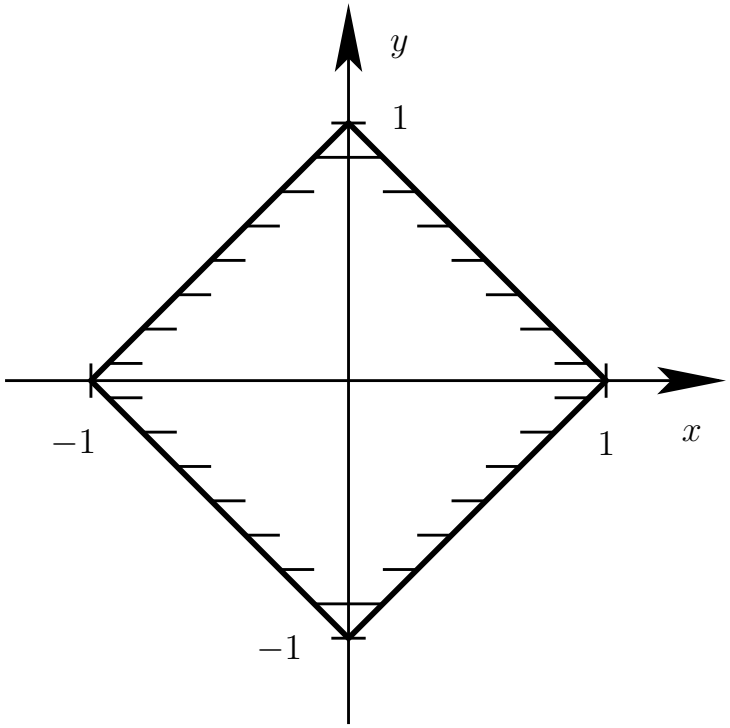


Abb. A.1: Graphische Darstellung der Menge aller Punkte $(x; y)$ mit $|x| + |y| \leq 1$ in der $(x; y)$-Ebene.

A.6 Potenzen und Wurzeln

Wir wiederholen nun, wie eine Potenz a^r mit **Basis** $a \in]0; \infty[$ und **Exponent** $r \in \mathbb{Z}$ definiert ist. Danach erlauben wir auch $r \in \mathbb{Q}$ und betrachten dabei den Begriff der n-ten Wurzel.

Definition A.32. (Potenzen mit ganzzahligem Exponenten)
Wir definieren

$$a^0 = 1 \qquad \text{für alle } a \in \mathbb{R} \setminus \{0\},$$

und für ***positive ganze Zahlen****, also $n \in \mathbb{N}$, ist a^n definiert durch*

$$a^n = \underbrace{a \cdot a \cdot \ldots \cdot a}_{n\text{-mal}} \qquad \text{für alle } a \in \mathbb{R}.$$

Ist n eine ***negative ganze Zahl****, also $n \in \mathbb{Z} \setminus \mathbb{N}_0$, so ist $n = -m$ mit $m \in \mathbb{N}$, und wir definieren*

$$a^n = a^{-m} = \frac{1}{a^m} = \frac{1}{\underbrace{a \cdot a \cdot \ldots \cdot a}_{m\text{-mal}}} \qquad \text{für alle } a \in \mathbb{R} \setminus \{0\}.$$

Insbesondere gilt

$$a^{-1} = \frac{1}{a}\,.$$

Beispiel A.33. (Potenzen reeller Zahlen mit Exponenten in $\mathbb{Z}$)

(a) $2^3 = 2 \cdot 2 \cdot 2 = 8$

(b) $10^4 = 10 \cdot 10 \cdot 10 \cdot 10 = 10.000$

(c) $2^{-1} = \frac{1}{2} = 0{,}5$

(d) $10^{-2} = \frac{1}{10^2} = \frac{1}{10 \cdot 10} = \frac{1}{100} = 0{,}01$

(e) $3^{-3} = \frac{1}{3^3} = \frac{1}{3 \cdot 3 \cdot 3} = \frac{1}{27}$

(f) $(-2)^3 = (-2) \cdot (-2) \cdot (-2) = -8$

(g) $(-4)^{-2} = \frac{1}{(-4)^2} = \frac{1}{(-4) \cdot (-4)} = \frac{1}{16} = 0{,}0625$

Überlegen Sie sich selber weitere Beispiele. ♠

Hilfssatz A.34. (Regeln für das Rechnen Exponenten in $\mathbb{Z}$)
Seien $a, b \in \mathbb{R} \setminus \{0\}$, und seien n und m in $\mathbb{Z}$. Dann gelten

$$a^{n \cdot m} = (a^n)^m = (a^m)^n \tag{A.5}$$

und

$$a^{n+m} = a^n\, a^m \qquad \text{und} \qquad a^{n-m} = a^n\, a^{-m} = \frac{a^n}{a^m}\,. \tag{A.6}$$

Weiter gelten

$$(a \cdot b)^n = a^n\, b^n \qquad \text{und} \qquad \left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}\,. \tag{A.7}$$

Beispiel A.35. (Regeln für das Rechnen mit Exponenten in $\mathbb{Z}$)

(a) $(10^4)^2 = 10^{4 \cdot 2} = 10^8 = 100.000.000$

(b) $2^4 \cdot 2^6 = 2^{4+6} = 2^{10} = 1024$

(c) $17^{-5} \cdot 17^4 = 17^{-5+4} = 17^{-1} = \dfrac{1}{17} \approx 0{,}05882$

(d) $\left(\dfrac{1}{2}\right)^{13} \cdot 2^{13} = \left(\dfrac{1}{2} \cdot 2\right)^{13} = 1^{13} = 1$

(e) $2^{-3} \cdot 3^{-3} = (2 \cdot 3)^{-3} = 6^{-3} = \dfrac{1}{6^3} = \dfrac{1}{216} \approx 0{,}0046296$

Überlegen Sie sich selber weitere Beispiele. ♠

Wir beweisen nun Hilfssatz A.34 teilweise, weil dieses unser Verständnis der Rechenregeln erhöht.

Beweis von Hilfssatz A.34: Wir geben den Beweis nur für den Fall $n > 0$ und $m > 0$. Die Fälle $n < 0$ oder $m < 0$ können analog bewiesen werden, aber sie sind etwas aufwendiger.

$$(a^n)^m = \underbrace{a^n \cdot a^n \cdot \ldots \cdot a^n}_{m\text{-mal}} = \underbrace{a \cdot a \cdot \ldots \cdot a}_{(n \cdot m)\text{-mal}} = \underbrace{a^m \cdot a^m \cdot \ldots \cdot a^m}_{n\text{-mal}} = (a^m)^n$$

und

$$a^{n+m} = \underbrace{a \cdot a \cdot \ldots \cdot a}_{(n+m)\text{-mal}} = \big(\underbrace{a \cdot a \cdot \ldots \cdot a}_{n\text{-mal}}\big) \cdot \big(\underbrace{a \cdot a \cdot \ldots \cdot a}_{m\text{-mal}}\big) = a^n \cdot a^m,$$

$$a^{n-m} = \underbrace{a \cdot a \cdot \ldots \cdot a}_{(n-m)\text{-mal}} = \big(\underbrace{a \cdot a \cdot \ldots \cdot a}_{n\text{-mal}}\big) \cdot \frac{1}{\underbrace{a \cdot a \cdot \ldots \cdot a}_{m\text{-mal}}} = a^n \cdot a^{-m}.$$

Damit haben wir die Gleichungen (A.5) und (A.6) für $m > 0$ und $n > 0$ bewiesen. Weiter gilt für $n > 0$

$$(a \cdot b)^n = \underbrace{(a \cdot b) \cdot (a \cdot b) \cdot \ldots \cdot (a \cdot b)}_{n\text{-mal}} = \big(\underbrace{a \cdot a \cdot \ldots \cdot a}_{n\text{-mal}}\big) \cdot \big(\underbrace{b \cdot b \cdot \ldots \cdot b}_{n\text{-mal}}\big) = a^n \cdot b^n,$$

$$\left(\frac{a}{b}\right)^n = \underbrace{\frac{a}{b} \cdot \frac{a}{b} \cdot \ldots \cdot \frac{a}{b}}_{n\text{-mal}} = \underbrace{\frac{a \cdot a \cdot \ldots \cdot a}{b \cdot b \cdot \ldots \cdot b}}_{\text{jeweils } n\text{-mal}} = \frac{a^n}{b^n},$$

und wir haben (A.7) für $n > 0$ ebenfalls bewiesen. □

Als Nächstes wollen wir Potenzen mit rationalem Exponenten definieren. Dazu benötigen wir als Vorbereitung die n-te Wurzel.

Definition A.36. (n-te Wurzel einer nicht-negativen Zahl)
*Sei $a \in \mathbb{R}$ eine **nicht-negative** reelle Zahl, und sei $n \in \mathbb{N}$ eine natürliche Zahl. Dann ist die n-te **Wurzel** $a^{1/n} = \sqrt[n]{a}$ als die **nicht-negative** Zahl b definiert, für die gilt $b^n = a$.*

Wir bemerken, dass wir für $n = 2$ insbesondere die **Quadratwurzel** erhalten: Für $a \in \mathbb{R}$ mit $a \geq 0$ ist $\sqrt{a}$ die nicht-negative reelle Zahl, für die gilt $(\sqrt{a})^2 = a$.

Beispiel A.37. (n-te Wurzeln von $a > 0$)

(a) $1000^{1/3} = 10$, weil $10^3 = 1000$

(b) $2^{1/2} = \sqrt{2}$, da $(\sqrt{2})^2 = \sqrt{2} \cdot \sqrt{2} = 2$

(c) $81^{1/4} = 3$, weil $3^4 = 81$

(d) $8^{1/3} = 2$, denn $2^3 = 8$

(e) $a^{1/2} = \sqrt{a}$, weil $(\sqrt{a})^2 = \sqrt{a} \cdot \sqrt{a} = a$

(f) $0^{1/7} = 0$, da $0^7 = 0$.

Überlegen Sie sich selber weitere Beispiele. ♠

Analog zu (A.5) und (A.7) in Hilfssatz A.34 können wir auch Regeln für das Rechnen mit n-ten Wurzeln herleiten.

Hilfssatz A.38. (Rechenregeln für n-te Wurzeln)
Seien $a, b \in \mathbb{R}$ nicht-negative reelle Zahlen und $m, n \in \mathbb{N}$. Dann gelten

$$a^{1/(n \cdot m)} = (a^{1/n})^{1/m} = (a^{1/m})^{1/n},$$

und

$$(a \cdot b)^{1/n} = a^{1/n}\, b^{1/n}.$$

Man kann Hilfssatz A.38 mit Hilfe der Definition der n-ten Wurzel und unter Ausnutzung von Rechenregeln A.5 und A.7 beweisen. Hilfssatz A.38 ist nützlich, um n-te Wurzeln zu berechnen und zu vereinfachen. Wir betrachten einige Beispiele.

Beispiel A.39. (Rechenregeln für n-te Wurzeln)

(a) $8^{1/6} = 8^{1/(2\cdot 3)} = (8^{1/3})^{1/2}$, und wegen $2^3 = 8$ gilt

$$8^{1/6} = (8^{1/3})^{1/2} = 2^{1/2} = \sqrt{2}.$$

(b) $6561^{1/8} = 6561^{1/(2\cdot 4)} = (6561^{1/2})^{1/4}$, und wegen $81^2 = 6561$ gilt

$$6561^{1/8} = (6561^{1/2})^{1/4} = 81^{1/4} = (81^{1/2})^{1/2} = 9^{1/2} = 3,$$

wobei wir $9^2 = 81$ und $3^2 = 9$ verwendet haben.

(c) $24^{1/3} = (3\cdot 8)^{1/3} = 3^{1/3}\, 8^{1/3} = 3^{1/3}\cdot 2 = 2\cdot 3^{1/3}$, wobei wir $2^3 = 8$ ausgenutzt haben.

Überlegen Sie sich selber weitere Beispiele. ♠

Mit Hilfe der Potenzen mit ganzzahligem Exponenten und mit der n-ten Wurzel können wir nun Potenzen mit rationalem Exponenten einführen.

Definition A.40. (Potenzen mit rationalem Exponenten)
Sie a eine positive reelle Zahl, und seien $m \in \mathbb{Z}$ und $n \in \mathbb{N}$. Dann ist $a^{m/n}$ definiert durch

$$a^{m/n} = \left(a^{1/n}\right)^m = (a^m)^{1/n}.$$

Beispiel A.41. (Potenzen mit rationalem Exponenten)

(a) $2^{-1/2} = (2^{1/2})^{-1} = (\sqrt{2})^{-1} = 1/\sqrt{2}$.

(b) $9^{3/2} = (9^{1/2})^3 = (\sqrt{9})^3 = 3^3 = 27$, wobei wir $3^2 = 9$ verwendet haben.

(c) $1000^{-4/3} = (1000^{1/3})^{-4} = 10^{-4} = 1/10^4 = 0{,}0001$, wobei wir $10^3 = 1000$ benutzt haben.

(d) $(\sqrt{8})^{-2/3} = \left((\sqrt{8})^{-2}\right)^{1/3} = \left((8^{1/2})^{-2}\right)^{1/3} = (8^{-1})^{1/3} = (8^{1/3})^{-1} = 2^{-1} = 1/2$, wobei wir $2^3 = 8$ ausgenutzt haben.

Überlegen Sie sich selber weitere Beispiele. ♠

Aus Hilfssätzen A.34 und A.38 kann man den folgenden Hilfssatz herleiten.

Hilfssatz A.42. (Rechnen mit Potenzen mit Exponenten in $\mathbb{Q}$)
Seien $a, b \in \mathbb{R}$ positive reelle Zahlen, und seinen $m, k \in \mathbb{Z}$ und $n, \ell \in \mathbb{N}$. Dann gilt

$$a^{\frac{mk}{n\ell}} = a^{\frac{m}{n}\cdot\frac{k}{\ell}} = (a^{m/n})^{k/\ell} = (a^{k/\ell})^{m/n}.$$

Weiter gelten

$$a^{\frac{m}{n}+\frac{k}{\ell}} = a^{m/n}\, a^{k/\ell} \qquad \textit{und} \qquad a^{\frac{m}{n}-\frac{k}{\ell}} = a^{m/n}\, a^{-k/\ell} = \frac{a^{m/n}}{a^{k/\ell}}$$

und

$$(a \cdot b)^{m/n} = a^{m/n}\, b^{m/n} \qquad \textit{und} \qquad \left(\frac{a}{b}\right)^{m/n} = \frac{a^{m/n}}{b^{m/n}}\,.$$

Beispiel A.43. (Rechnen mit Potenzen mit rationalen Exponenten)
In diesem Beispiel wollen wir die Rechenregeln aus Hilfssatz A.42 anwenden, um zu vereinfachen:

(a) $2^{1/3} \cdot 2^{2/3} = 2^{\frac{1}{3}+\frac{2}{3}} = 2^1 = 2$.

(b) $50^{3/2} = (2{\cdot}25)^{3/2} = 2^{3/2}{\cdot}25^{3/2} = 2^{1+1/2}{\cdot}(25^{1/2})^3 = 2{\cdot}2^{1/2}{\cdot}5^3 = 2{\cdot}\sqrt{2}{\cdot}125 = 250 \cdot \sqrt{2}$, wobei wir $5^2 = 25$ benutzt haben.

(c) $8^{5/6} = 8^{\frac{1}{2}+\frac{1}{3}} = 8^{1/2} \cdot 8^{1/3} = (4 \cdot 2)^{1/2} \cdot 2 = 4^{1/2} \cdot 2^{1/2} \cdot 2 = 2 \cdot 2^{1/2} \cdot 2 = 4 \cdot \sqrt{2}$, wobei wir $2^2 = 4$ und $2^3 = 8$ ausgenutzt haben.

Überlegen Sie sich selber weitere Beispiele. ♠

A.7 Lösen quadratischer Gleichungen und binomischer Satz

Zum Lösen von quadratischen Gleichungen sind die binomischen Formeln nützlich.

Hilfssatz A.44. (binomische Formeln)
Seien $a, b \in \mathbb{R}$. Dann gelten:

(1) ***Erste binomische Formel:***

$$(a+b)^2 = a^2 + 2\,a\,b + b^2 \qquad \Longleftrightarrow \qquad a^2 + 2\,a\,b + b^2 = (a+b)^2.$$

(2) ***Zweite binomische Formel:***

$$(a-b)^2 = a^2 - 2\,a\,b + b^2 \qquad \Longleftrightarrow \qquad a^2 - 2\,a\,b + b^2 = (a-b)^2.$$

(3) ***Dritte binomische Formel:***

$$(a+b)\cdot(a-b) = a^2 - b^2 \qquad \Longleftrightarrow \qquad a^2 - b^2 = (a+b)\cdot(a-b).$$

Wir wiederholen zunächst die Definition einer quadratischen Gleichung.

Definition A.45. (quadratische Gleichung)

Eine Gleichung der Form

$$a\,x^2 + b\,x + c = 0$$

mit $a, b, c \in \mathbb{R}$ *und* $a \neq 0$ *nennt man eine* ***quadratische Gleichung****. Indem man durch* $a \neq 0$ *teilt erhält man die* ***Standardform*** *der quadratischen Gleichung:*

$$x^2 + \frac{b}{a}\,x + \frac{c}{a} = 0 \qquad \Longleftrightarrow \qquad x^2 + p\,x + q = 0 \quad \textit{mit} \;\; p = \frac{b}{a}\,,\; q = \frac{c}{a}\,.$$

Beispiel A.46. (quadratische Gleichungen)

(a) $2\,x^2 - 12\,x + 16 = 0$ ist eine quadratische Gleichung. Ihre Standardform ist

$$x^2 - 6\,x + 8 = 0.$$

(b) $x^2 - 4 = 0$ ist eine quadratische Gleichung; sie befindet sich bereits in Standardform.

(c) $2\,x - 7 = 0$ ist **keine** quadratische Gleichung, sondern eine lineare Gleichung.

(d) $x^3 + 5\,x^2 + x = 0$ ist ebenfalls **keine** quadratische Gleichung, denn hier tritt eine dritte Potenz von x auf.

Überlegen Sie sich selber weitere Beispiele quadratischer Gleichungen. ♠

Ab jetzt betrachten wir nur noch quadratische Gleichungen in Standardform.

Lösungsmethode A.47. (Lösen mit quadratischer Ergänzung)

*Durch eine **quadratische Ergänzung** wird der Term $x^2+p\,x+q$ als Summe eines Quadrats und einem konstanten Terms dargestellt:*

$$x^2+p\,x+q=\underbrace{x^2+2\,\frac{p}{2}\,x+\left(\frac{p}{2}\right)^2}_{=\left(x+\frac{p}{2}\right)^2}-\left(\frac{p}{2}\right)^2+q=\left(x+\frac{p}{2}\right)^2-\frac{p^2-4\,q}{4}\,,$$

wobei wir im letzten Schritt die erste binomische Formel (siehe Hilfssatz A.44 (1)) verwendet haben. Dabei gilt $\left(x+\frac{p}{2}\right)^2\geq 0$, weil Quadrate in den reellen Zahlen immer nicht-negativ sind. Es können nun drei Fälle auftreten, die vom Vorzeichne von $d=p^2-4\,q$ abhängen:

(1) Fall $d=p^2-4\,q=0$: *Ist $p^2-4\,q=0$ so vereinfacht sich die quadratische Gleichung zu*

$$x^2+p\,x+q=\left(x+\frac{p}{2}\right)^2,$$

*und wir lesen ab, dass die **eine (doppelte) reelle Lösung** $x=-\frac{p}{2}$ ist.*

(2) Fall $d=p^2-4\,q>0$: *Ist $p^2-4\,q>0$ so können wir die dritte binomische Formel (siehe Hilfssatz A.44 (3)) anwenden um die Lösungen zu bestimmen:*

$$x^2+p\,x+q=\left(x+\frac{p}{2}\right)^2-\underbrace{\frac{p^2-4\,q}{4}}_{>0}=\left(x+\frac{p}{2}\right)^2-\left(\frac{\sqrt{p^2-4\,q}}{2}\right)^2$$

$$=\left(x+\frac{p}{2}-\frac{\sqrt{p^2-4\,q}}{2}\right)\left(x+\frac{p}{2}+\frac{\sqrt{p^2-4\,q}}{2}\right).$$

*Wir lesen die **zwei verschiedenen reellen Lösungen** ab:*

$$x_1=-\frac{p}{2}+\frac{\sqrt{p^2-4\,q}}{2}\qquad x_2=-\frac{p}{2}-\frac{\sqrt{p^2-4\,q}}{2}\,.$$

(3) Fall $d=p^2-4\,q<0$: *Ist $p^2-4\,q<0$, so ist $-\frac{p^2-4\,q}{4}>0$ und es gilt*

$$x^2+p\,x+q=\underbrace{\left(x+\frac{p}{2}\right)^2}_{\geq 0}+\underbrace{\left(-\frac{p^2-4\,q}{4}\right)}_{>0}>0,$$

*und die quadratische Gleichung hat **keine reellen Lösungen**.*

Es lohnt sich nicht, die Formeln aus Lösungsmethode A.47 auswendig zu lernen. Wenn man die Vorgehensweise verstanden hat, dann kann man sie mit Hilfe der binomischen Formeln an jedem Beispiel direkt durchführen.

Beispiel A.48. (quadratische Ergänzung)

(a) $x^2 - 6\,x + 8 = 0$

Wir führen die quadratische Ergänzung durch:

$$x^2 - 6\,x + 8 = x^2 + 2\,(-3)\,x + (-3)^2 - (-3)^2 + 8 = (x-3)^2 - 1.$$

mit der dritten binomischen Formel (siehe Hilfssatz A.44 (3)) erhalten wir also

$$x^2 - 6\,x + 8 = (x-3)^2 - 1^2 = (x-3-1)\,(x-3+1) = (x-4)\,(x-2).$$

Also hat die Gleichung $x^2 - 6\,x + 8 = 0$ die beiden verschiedenen reellen Lösungen $x_1 = 4$ und $x_2 = 2$.

(b) $x^2 + 4\,x + 4 = 0$

Wir führen die quadratische Ergänzung durch:

$$x^2 - 4\,x + 4 = x^2 + 2\,(-2)\,x + (-2)^2 = (x-2)^2,$$

und wir lesen die (doppelte) reelle Lösung $x = 2$ ab.

(c) $x^2 + 6\,x + 10 = 0$

Wir führen die quadratische Ergänzung durch:

$$x^2 + 6\,x + 10 = x^2 + 2\,(3\,x) + 3^2 - 3^2 + 10 = (x+3)^2 + 1,$$

und wir lesen ab, dass $x^2 + 6\,x + 10 = 0$ keine reellen Lösungen hat.

Überlegen Sie sich selber weitere Beispiele quadratischer Gleichungen und lösen Sie diese. ♠

An der Lösungsmethode A.47 kann man die q-p-Formel zum Lösen einer quadratischen Gleichung direkt ablesen:

Bemerkung A.49. (p-q-Formel)

Setzt man in Lösungsmethode A.47 $a = p$ und $b = q$, so erhält man

$$0 = x^2 + p\,x + q = \left(x + \frac{p}{2}\right)^2 - \frac{p^2 - 4\,q}{4} \qquad \text{mit } p, q \in \mathbb{R},$$

und list an den drei Fällen ab: Die quadratische Gleichung ist in den reellen Zahlen nur dann lösbar, wenn $d = p^2 - 4\,q \geq 0$ ist. Die Lösungen sind dann

$$x_1 = -\frac{p}{2} + \frac{\sqrt{p^2 - 4\,q}}{2} \qquad \text{und} \qquad x_2 = -\frac{p}{2} - \frac{\sqrt{p^2 - 4\,q}}{2}\,.$$

Dabei erhalten wir für $d = p^2 - 4\,q =$ nur eine (doppelte) reelle Lösung. Für $d = p^2 - 4\,q < 0$ hat die quadratische Gleichung keine Lösungen in $\mathbb{R}$.

Ein nützliches Hilfsmittel zum Lösen von manchen quadratischen Gleichungen ist der Wurzelsatz von Vieta.

Hilfssatz A.50. (Wurzelsatz von Vieta)
Sei $x^2 + p\,x + q = 0$ mit $p, q \in \mathbb{R}$ und $p \neq 0$ eine quadratische Gleichung, für die gilt $d = p^2 - 4\,q \geq 0$ ist. Dann besitzt die quadratische Gleichung zwei reelle (nicht notwendigerweise verschiedene) Lösungen x_1 und x_2, für die gilt

$$x_1 + x_2 = -p \qquad \textit{und} \qquad x_1\,x_2 = q \tag{A.8}$$

und

$$x^2 + p\,x + q = (x - x_1)\,(x - x_2). \tag{A.9}$$

Sieht man durch Inspizieren der Gleichung Lösungen x_1 und x_2 von (A.8)*, so hat man die Lösungen der quadratischen Gleichung gefunden.*

Beweis von Hilfssatz A.50: Formel (A.9) ist gerade die Faktorisierung der quadratischen Gleichung. Da per Annahme x_1 und x_2 Lösungen sind, muss (A.9) gelten. Man verifiziert die Formeln in (A.8) durch Ausmultiplizieren und Sortieren der rechten Seite in (A.9):

$$\begin{aligned} x^2 + p\,x + q = (x - x_1)\,(x - x_2) &= x^2 - x_1\,x - x\,x_2 + x_1\,x_2 \\ &= x^2 - (x_1 + x_2)\,x + x_1\,x_2, \end{aligned}$$

Da die Koeffizienten der gleichen Potenzen von x auf der linken und der rechten Seite übereinstimmen müssen, liest man ab, dass $p = -(x_1 + x_2)$ und $q = x_1\,x_2$. gelten muss. □

Betrachten wir ein Beispiel zum Wurzelsatz von Vieta.

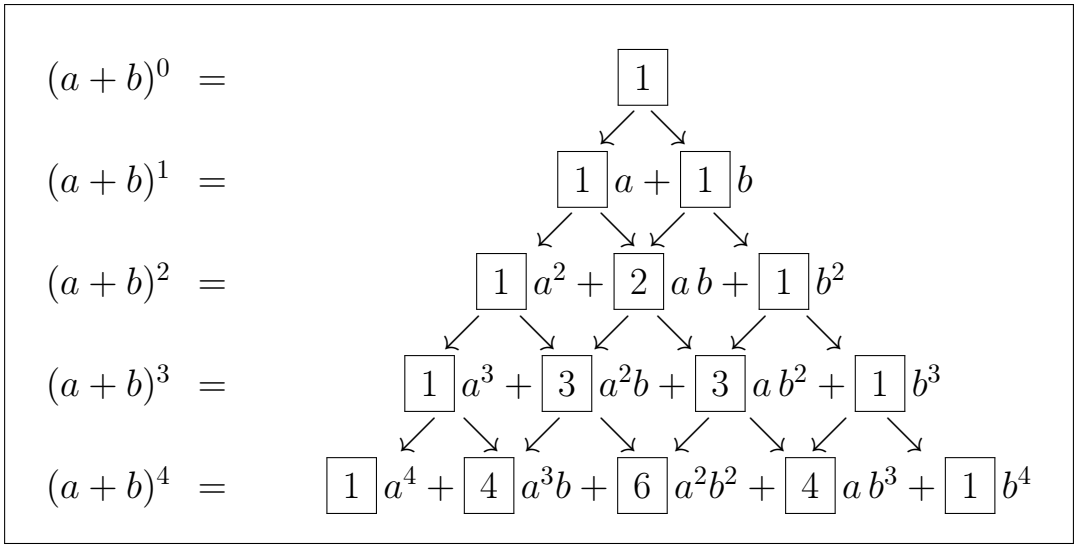


Abb. A.2: Das Pascalsche Dreieck: Durch Addition der Koeffizienten in der vorigen Zeile, von denen ein Pfeil auf den neuen Koeffizienten weist, erhält man jeweils den Wert des neuen Koeffizienten.

Beispiel A.51. (Wurzelsatz von Vieta)
Wir lösen die quadratische Gleichung

$$3\,x^2 + 27\,x + 60 = 0 \tag{A.10}$$

mit dem Wurzelsatz von Vieta. Dazu teilen wir erst durch 3, um die Standardform zu bekommen.

$$3\,x^2 + 27\,x + 60 = 0 \;\Big|\; :3 \qquad \Longleftrightarrow \qquad x^2 + 9\,x + 20 = 0.$$

Wir haben also $p = 9$ und $q = 20$. Wegen

$$d = p^2 - = 4\,q = 9^2 - 4 \cdot 20 = 81 - 80 = 1 > 0$$

hat die quadratische Gleichung zwei reelle Lösungen. Nun gilt für $x_1 = -5$ und $x_2 = -4$, dass

$$x_1 + x_2 = -5 - 4 = -9 = -p \qquad \text{und} \qquad x_1\,x_2 = (-5)\,(-4) = 20 = q.$$

Daher folgt nach dem Wurzelsatz von Vieta

$$x^2 + 9\,x + 20 = \big(x - (-5)\big)\,\big(x - (-4)\big) = (x+5)\,(x+4) = 0,$$

und die beiden reellen Lösungen von (A.10) sind $x_1 = -5$ und $x_2 = -4$. ♠

In Verallgemeinerung der ersten binomischen Formel gelten das Pascalsche Dreieck und der binomische Satz zur Berechnung vom $(a+b)^n$ mit beliebigem $n \in \mathbb{N}_0$.

Multipliziert man nun $(a+b)^n$ für $n \geq 2$ aus, so findet man, dass die Koeffizienten der Potenzen $a^n,\ a^{n-1}\,b,\ a^{n-2}\,b^2,\ \ldots,\ a\,b^{n-1},\ b^n$ aus den Koeffizienten der Potenzen $a^{n-1},\ a^{n-2}\,b,\ a^{n-3}\,b^2,\ \ldots,\ a\,b^{n-2},\ b^{n-1}$ in der ausmultiplizierten Darstellung von $(a+b)^{n-1}$ **mit Hilfe des Pascalschen Dreiecks rekursiv berechnet** werden können. Dieses ist in Abbildung A.2 illustriert.

Das Pascalsche Dreieck hat einen entscheidenden Nachteil. Um die Koeffizienten für das Ausmultiplizieren von $(a+b)^n$ zu bekommen, müssen wir vorher alle Koeffizienten für das Ausmultiplizieren von $(a+b)^m$ mit $m = 1, 2, \ldots, n-1$ berechnen. Der binomische Satz umgeht dieses Problem und liefert eine direkte Formel für die Koeffizienten. Zur Vorbereitung müssen wir zunächst Fakultäten und Binomialkoeffizienten einführen.

Definition A.52. (Fakultät)
Die natürliche Zahl

$$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n \qquad \textit{für } n \in \mathbb{N},$$
$$0! = 1 \qquad \textit{für } n = 0,$$

heißt ***Fakultät von*** **n** *oder* **n-*Fakultät***.

Es gilt die Rekursionsformel:

$$(n+1)! = \underbrace{1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n}_{=n!} \cdot (n+1) = n! \cdot (n+1) \qquad \text{für alle } n \in \mathbb{N}_0.$$

Beispiel A.53. (Fakultäten)
Wir haben

$$0! = 1, \quad 1! = 1, \quad 2! = 2 \cdot 1! = 2, \quad 3! = 3 \cdot 2! = 6, \quad 4! = 4 \cdot 3! = 24.$$

Berechnen Sie 5! und 6! zur Übung. ♠

Definition A.54. (Binomialkoeffizient)
Sei $n \in \mathbb{N}_0$ und $k \in \mathbb{N}_0$ mit $0 \leq k \leq n$. Der ***Binomialkoeffizient*** $\binom{n}{k}$ *ist*

definiert als

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!}. \tag{A.11}$$

In Worten sagt man für $\binom{n}{k}$ *„(Binomialkoeffizient)* n *über* k*“.*

Die zweite Darstellung von $\binom{n}{k}$ folgt, indem man $(n-k)!$ kürzt.

Nun können wir den binomischen Satz für das Ausmultiplizieren (und für die umgekehrte Richtung, das Faktorisieren) von $(a+b)^n$ formulieren.

Satz A.55. (binomischer Satz)
Seien $a, b \in \mathbb{R}$ *und* $n \in \mathbb{N}_0$*. Dann gilt:*

$$\begin{aligned}(a+b)^n &= \binom{n}{0} a^n b^0 + \binom{n}{1} a^{n-1} b^1 + \binom{n}{2} a^{n-2} b^2 + \ldots + \binom{n}{n} a^0 b^n \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k.\end{aligned}$$

Als Spezialfall erhalten wir für $n = 2$ *die erste binomische Formel. Für* $n = 2$ *und* $b = -d$ *erhalten wir die zweite binomische Formel.*

Zur Illustration leiten wir die erste und die zweite binomische Formel aus dem binomischen Satz ab.

Beispiel A.56. (erste und zweite binomische Formel)
Für $n = 2$ liest sich der binomische Satz wie folgt:

$$\begin{aligned}(a+b)^2 &= \binom{2}{0} a^2 b^0 + \binom{2}{1} a\, b + \binom{2}{2} a^0 b^2 \\ &= \binom{2}{0} a^2 + \binom{2}{1} a\, b + \binom{2}{2} b^2, \end{aligned} \tag{A.12}$$

und wir haben

$$\binom{2}{0} = \frac{2!}{0! \cdot (2-0)!} = \frac{2!}{0! \cdot 2!} = \frac{2}{1 \cdot 2} = 1,$$

$$\binom{2}{1} = \frac{2!}{1! \cdot (2-1)!} = \frac{2!}{1! \cdot 1!} = \frac{2}{1 \cdot 1} = 2,$$

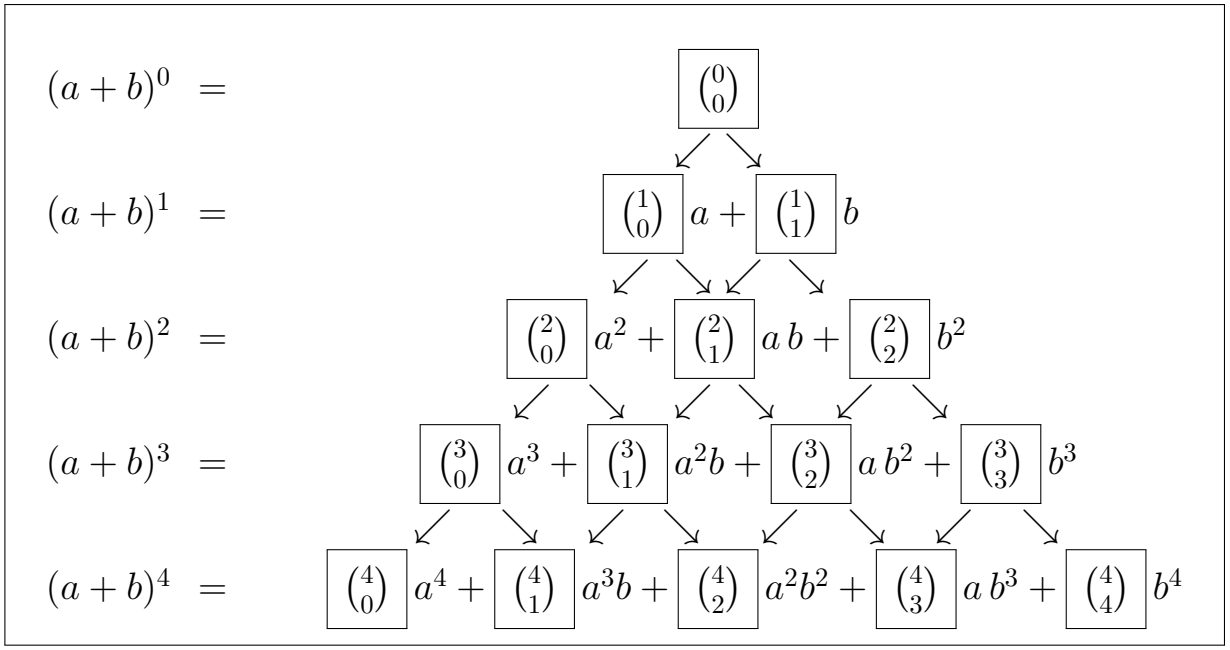


Abb. A.3: Illustration von (A.13) am Pascalschen Dreieck.

$$\binom{2}{2} = \frac{2!}{2! \cdot (2-2)!} = \frac{2!}{2! \cdot 0!} = \frac{2}{2 \cdot 1} = 1.$$

Einsetzen in (A.12) ergibt die erste binomische Formel:

$$(a+b)^2 = \binom{2}{0} a^2 + \binom{2}{1} a\,b + \binom{2}{2} b^2 = 1\,a^2 + 2\,a\,b + 1\,b^2 = a^2 + 2\,a\,b + b^2.$$

Setzen wir in der letzten Formel nun $b = -d$, so finden wir

$$(a-d)^2 = a^2 + 2\,a\,(-d) + (-d)^2 = a^2 - 2\,a\,d + d^2,$$

und wir haben auch die zweite binomische Formel hergeleitet. ♠

Es gelten die folgenden Identitäten für die Fakultäten.

Hilfssatz A.57. (Rechenregeln für Binomialkoeffizienten)
Für alle $n \in \mathbb{N}$ mit $n \geq 2$ und $k \in \mathbb{N}_0$ mit $k \leq n$ gelten folgende Identitäten:

$$\binom{n}{0} = \binom{n}{n} = 1,$$

$$\binom{n}{1} = \binom{n}{n-1} = n,$$

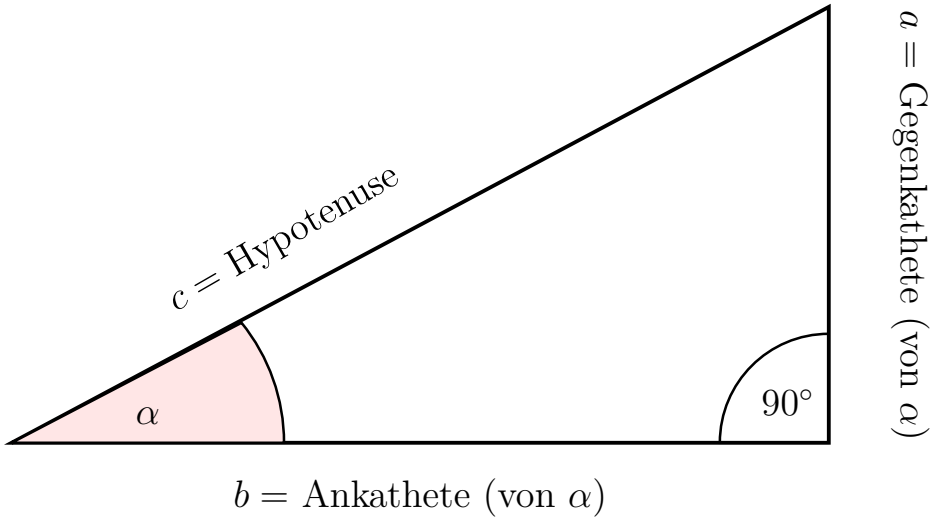


Abb. A.4: Die Definition von Sinus und Cosinus am rechtwinkligen Dreieck: $\sin(\alpha) = a/c$ und $\cos(\alpha) = b/c$.

sowie

$$\binom{n}{k} = \binom{n}{n-k},$$

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}, \qquad \textit{wobei } k < n. \tag{A.13}$$

Wir bemerken, dass (A.13) gerade die Formel ist, welche die Berechnung der Koeffizienten im Pascalschen Dreieck beschreibt (siehe Abbildung A.3).

A.8 Sinus und Cosinus als Kreisfunktionen

Wir beginnen unsere Einführung der trigonometrischen Funktionen mit der Wiederholung der Definition von Sinus und Cosinus am rechtwinkligen Dreieck.

Definition A.58. (Sinus und Cosinus im rechtwinkligen Dreieck)
Für Winkel α *mit* $0° < \alpha < 90°$ *sind* $\sin(\alpha)$ *(„**Sinus von** α“) und* $\cos(\alpha)$

*(**„Cosinus von** α“) im rechtwinkligen Dreieck wie folgt definiert:*

$$\sin(\alpha) = \frac{a}{c} = \frac{\text{Gegenkathete (von } \alpha)}{\text{Hypotenuse}},$$
$$\cos(\alpha) = \frac{b}{c} = \frac{\text{Ankathete (von } \alpha)}{\text{Hypotenuse}}.$$

*Die **Ankathete** (von α), die **Gegenkathete** (von α) und die **Hypotenuse**, sowie die Bezeichnungen der Dreiecksseiten sind in Abbildung A.4 illustriert.*

Für Berechnungen an nicht-winkligen Dreiecken sind auch der Sinussatz und der Cosinussatz wichtig.

Hilfssatz A.59. (Sinussatz und Cosinussatz)
In beliebigen Dreiecken gelten:

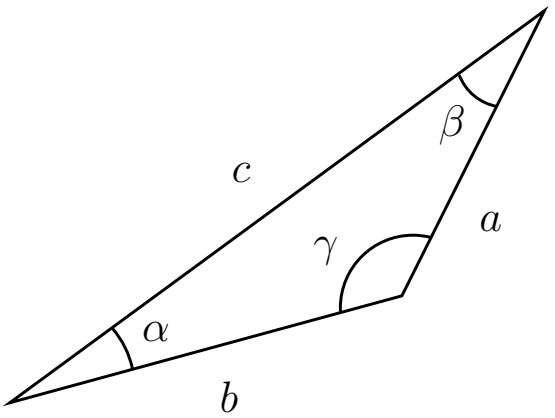
(1) ***Sinussatz:***

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

(2) ***Cosinussatz:***

$$c^2 = a^2 + b^2 - 2 \cdot a \cdot b \cdot \cos(\gamma)$$

Dabei sind die Bezeichnungen der Winkel und der Seiten in der nebenstehenden Skizze festgelegt.



Wir bemerken, dass für rechtwinklige Dreiecke der **Satz des Pythagoras** gilt:

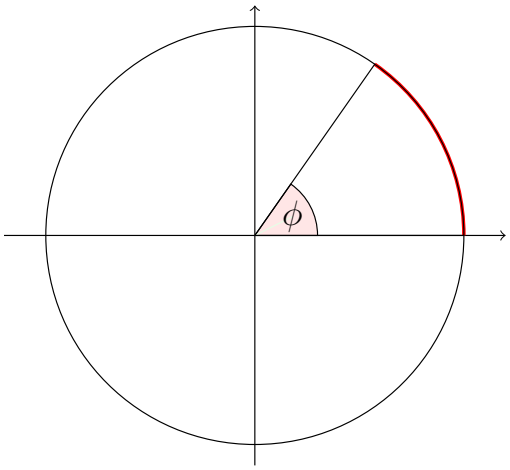
$$\big[\text{Gegenkathete (von } \alpha)\big]^2 + \big[\text{Ankathete (von } \alpha)\big]^2 = \big[\text{Hypotenuse}\big]^2$$

oder in der Beschriftung der Abbildung A.4

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

Wir wollen nun den Sinus und den Cosinus für beliebige Winkel definieren, indem wir Sinus und Cosinus als **trigonometrische Funktionen am Einheitskreis** einführen. Es ist dabei üblich, die Variable einer trigonometrischen Funktion nicht in Grad sondern im Bogenmaß anzugeben, welches wir daher zuerst einführen.

Definition A.60. (Bogenmaß)
*Das **Bogenmaß** b zu dem Winkel ϕ (gemessen in Grad) ist die Länge des Kreisbogens am **Einheitskreis** mit Radius $r = 1$ zu diesem Winkel ϕ (siehe Skizze rechts). Nach der Formel für den Kreisumfang $2\pi\, r = 2\pi$ hat der Kreisbogen zum Winkel $360°$ die Länge 2π Damit gilt die Gleichheit*



$$\frac{\phi}{360°} = \frac{b}{2\pi},$$

mit der wir zwischen Gradmaß und Bogenmaß umrechnen können:

$$b = \frac{2\pi}{360°} \cdot \phi \qquad \text{und} \qquad \phi = \frac{360°}{2\pi} \cdot b.$$

In der Tabelle A.1 ist die Umrechnung für das Gradmaß und das Bogenmaß für einige der wichtigsten Winkel aufgelistet. Sie sollten die Umrechnung zumindest für die in der Tabelle aufgeführten Winkel im Kopf haben.

Gradmaß	0	30	45	60	90	180	270	360	ϕ
Bogenmaß	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$	2π	$\frac{2\pi}{360}\,\phi$

Tabelle A.1: Umrechnung zwischen Gradmaß und Bogenmaß.

Nachdem wir das Bogenmaß eingeführt haben, können wir nun die Sinus- und die Cosinusfunktion am Einheitskreis definieren.

Definition A.61. (Sinusfunktion und Cosinusfunktion)
*Der **Einheitskreis** ist der Kreis in der $(x;y)$-Ebene mit Zentrum im Ursprung $(0;0)$ und mit Radius $r = 1$. Es seien $(x;y)$ die Koordinaten des Punktes P auf dem Einheitskreis, für den der Winkel gegen den Uhrzeigersinn von der positiven x-Achse aus gerade ϕ (im Bogenmaß) beträgt (siehe Abbildung A.5).*

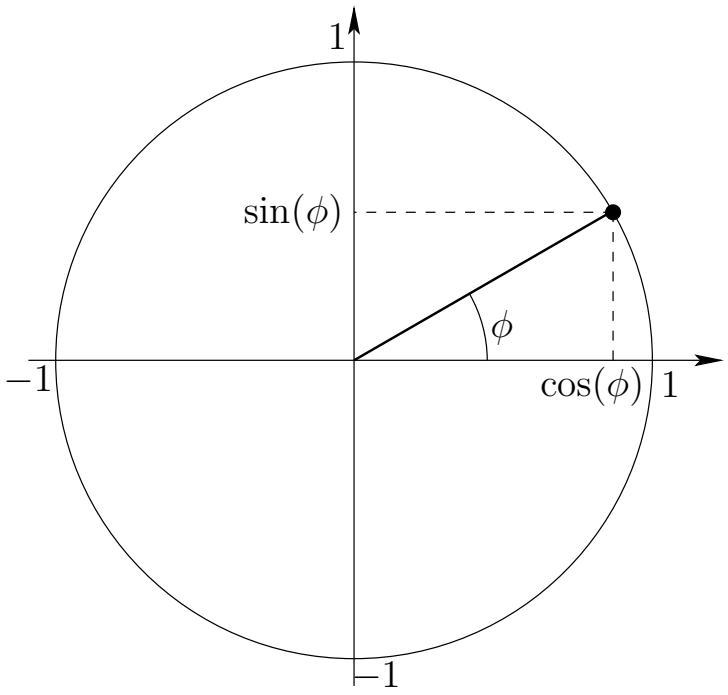


Abb. A.5: Definition von Sinus und Cosinus am Einheitskreis.

*Dann definieren wir den **Sinus** und den **Cosinus** durch:*

$$\sin(\phi) = y \qquad \textit{und} \qquad \cos(\phi) = x. \tag{A.14}$$

Dadurch sind $\sin(\phi)$ *und* $\cos(\phi)$ *für Winkel* $\phi \in [0; 2\pi[$ *erklärt. Für andere Werte* $\phi \in \mathbb{R}$ *definieren wir*

$$\sin(\phi) = \sin(\phi - 2k\pi) \qquad \textit{und} \qquad \cos(\phi) = \cos(\phi - 2k\pi), \tag{A.15}$$

wobei $k \in \mathbb{Z}$ *so gewählt ist, dass* $\phi - 2k\pi \in [0; 2\pi[$ *gilt.*

Durch (A.14) in Definition A.61 sind $\sin(\phi)$ und $\cos(\phi)$ für alle $\phi \in [0, 2\pi[$ definiert, d.h. wir haben zunächst jeweils eine Funktion auf dem Intervall $[0, 2\pi[$. Mit (A.15) werden die Sinusfunktion und die Cosinusfunktion durch sogenannte 2π**-periodische Fortsetzung** von $\sin(\phi)$ bzw. $\cos(\phi)$ von dem Intervall $[0, 2\pi[$ auf ganz $\mathbb{R}$ fortgesetzt.

In Abbildung A.6 haben wir die Graphen der Sinusfunktion und der Cosinusfunktion geometrisch veranschaulicht.

In der Tabelle A.2 sind die Werte von $\sin(x)$ und $\cos(x)$ für einige wichtige Winkel aufgelistet. Diese sollte man im Kopf haben.

x in Bogenmaß	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$	$\frac{5\pi}{6}$	π	$\frac{3\pi}{2}$	2π
x in Gradmaß	0	30	45	60	90	120	135	150	180	270	360
$\sin(x)$	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0	-1	0
$\cos(x)$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{3}}{2}$	-1	0	1

Tabelle A.2: Einige wichtige Werte der Sinus- bzw. der Cosinusfunktion.

Mit der Beobachtung, dass $0 = \frac{\sqrt{0}}{2}$, $\frac{1}{2} = \frac{\sqrt{1}}{2}$, $1 = \frac{\sqrt{4}}{2}$ sieht man, dass die Werte von $\sin(x)$ und $\cos(x)$ in Tabelle A.2 von der Form

$$\pm \frac{\sqrt{k}}{2}, \qquad k = 0, 1, 2, 3, 4,$$

sind, und kann sich das Muster leicht merken.

Beispiel A.62. (Berechnung der Werte von Sinus und Cosinus)
Man kann die Werte in Tabelle A.2 einfach mittels der Definition von Sinus und Cosinus über das Dreieck am Einheitskreis ablesen bzw. mit elementargeometrischen Überlegungen berechnen.

(a) Man sieht am Einheitskreis für den Winkel $x = 0$ direkt, dass

$$\sin(0) = 0 \qquad \text{und} \qquad \cos(0) = 1.$$

(b) Man sieht am Einheitskreis für den Winkel $x = \pi/2$ (also 90°) direkt, dass

$$\sin\left(\frac{\pi}{2}\right) = 1 \qquad \text{und} \qquad \cos\left(\frac{\pi}{2}\right) = 0.$$

(c) Für den Winkel $x = \pi/4$ (also 45°) haben wir ein gleichschenkliges Dreieck mit Hypotenuse der Länge 1, wie in dem linken Bild in Abbildung A.7 eingezeichnet. Nach dem Satz von Pythagoras gilt dann für die Länge $a = \cos(\pi/4) = \sin(\pi/4)$ der beiden gleichlangen Katheten des Dreiecks

$$a^2 + a^2 = 1 \qquad \Longleftrightarrow \qquad 2\,a^2 = 1 \qquad \Longleftrightarrow \qquad a^2 = \frac{1}{2}$$

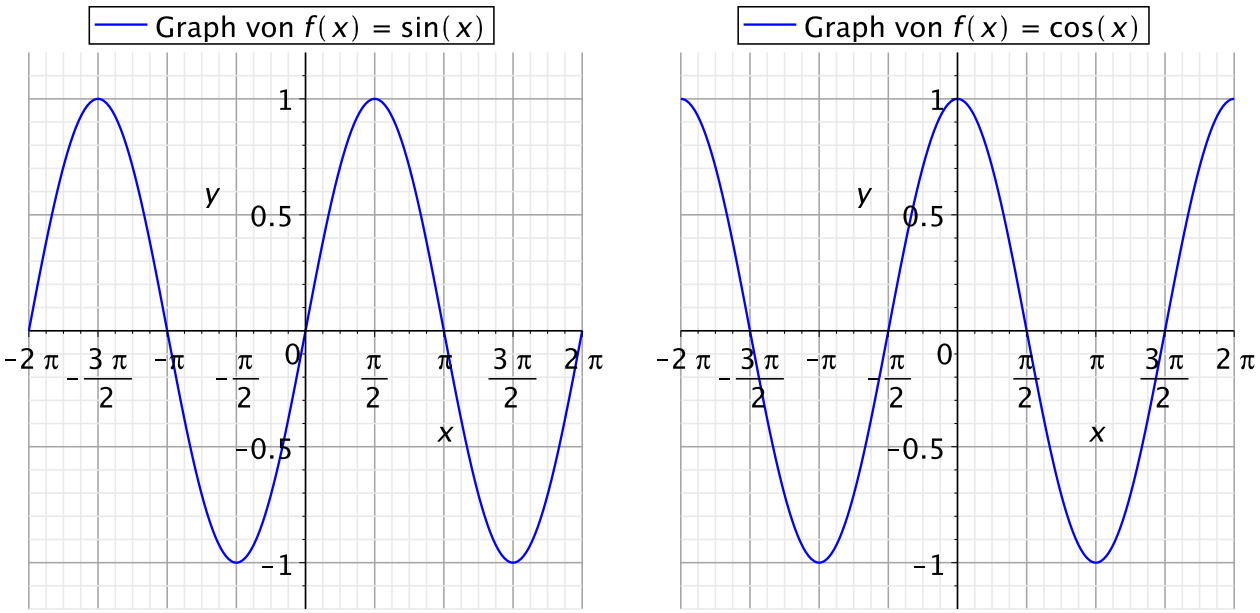


Abb. A.6: Veranschaulichung der Graphen der Sinusfunktion (linkes Bild) und der Cosinusfunktion (rechtes Bild).

$$\overset{a\geq 0}{\Longleftrightarrow} \qquad a = \sqrt{\frac{1}{2}} = \frac{1}{\sqrt{2}}\,.$$

Also finden wir

$$\sin\left(\frac{\pi}{4}\right) = \cos\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}\,.$$

(d) Zur Bestimmung von $\sin(x)$ und $\cos(x)$ für $x = \pi/6$ (also 30°) drehen wir das Dreieck am Einheitskreis und ergänzen eine gespiegelte Kopie des Dreiecks, so dass wir mit beiden Dreiecken zusammen ein gleichseitiges Dreieck erhalten, dessen Höhe $h = \cos(\pi/6)$ und dessen halbe Grundseite $\sin(\pi/6)$ ist (siehe das rechte Bild in Abbildung A.7). Wir können dann direkt ablesen, dass gilt $\sin(\pi/6) = 1/2$, und nach dem Satz des Pythagoras finden wir

$$1 = \left[\sin\left(\frac{\pi}{6}\right)\right]^2 + \left[\cos\left(\frac{\pi}{6}\right)\right]^2$$

$$\Longrightarrow \qquad \left[\cos\left(\frac{\pi}{6}\right)\right]^2 = 1 - \left[\sin\left(\frac{\pi}{6}\right)\right]^2 = 1 - \left[\frac{1}{2}\right]^2 = 1 - \frac{1}{4} = \frac{3}{4}$$

$$\Longrightarrow \qquad \cos\left(\frac{\pi}{6}\right) = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{2}\,.$$

Wir finden also

$$\sin\left(\frac{\pi}{6}\right) = \frac{1}{2} \qquad \text{und} \qquad \cos\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{2}\,.$$

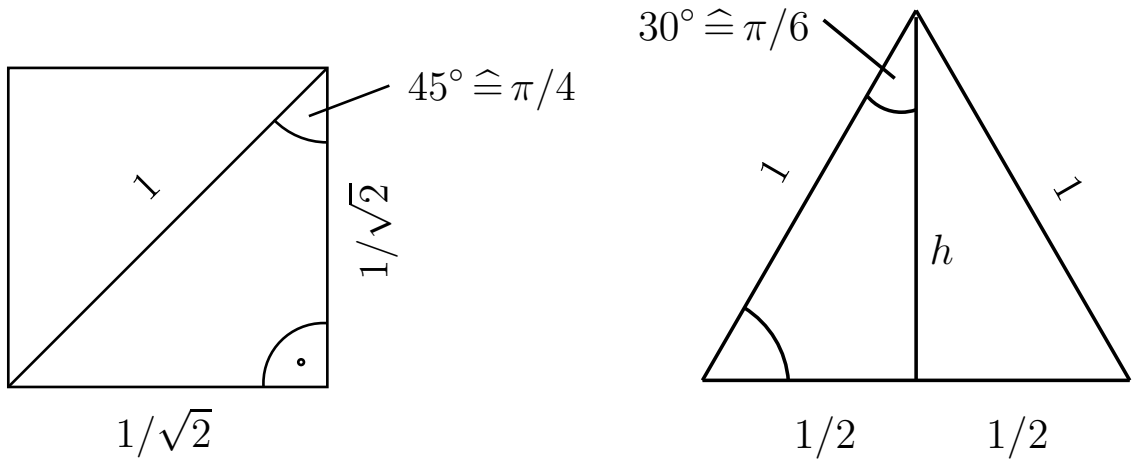


Abb. A.7: Skizzen zur Bestimmung von $\sin(x)$ und $\cos(x)$ für $x = \pi/4$ (linkes Bild) und $x = \pi/6$ (rechtes Bild).

Auf die Werte von Sinus und Cosinus in beispielsweise $x = 3\pi/4$ kommt man mit den bereits bekannten Werten und Symmetrieüberlegungen für die Dreiecke am Einheitskreis. ♠

A.9 Summen

Hier erklären wir die Summen-Notation und das Rechnen mit Summen.

Definition A.63. (Summen-Notation)
*Seien $m, n \in \mathbb{Z}$ mit $m \leq n$. Die Summe von $x_m, x_{m+1}, x_{m+2}, \ldots, x_n \in \mathbb{R}$ schreibt man mit dem **Summenzeichen**:*

$$\sum_{k=m}^{n} x_k = x_m + x_{m+1} + x_{m+2} + \ldots + x_n. \tag{A.16}$$

*Wir nennen k den **Summationsindex**, und der kleinste Wert des Summationsindexes (also m in (A.16)) wird als **untere Grenze** des Summationsindexes und der größte Wert des Summationsindexes (also n in (A.16)) wird also **obere Grenze** des Summationsindexes bezeichnet. Der Summationsindex ist frei wählbar und hat keine Bedeutung für den Wert der Summe, d.h.*

$$\sum_{k=m}^{n} x_k = \sum_{j=m}^{n} x_j.$$

Eine Summe, deren obere Grenze des Summationsindexes kleiner ist als deren untere Grenze, wird ***leere Summe*** *genannt. Wir definieren die leere Summe als Summe ohne Summanden und setzen formal*

$$\sum_{k=m}^{n} x_k = 0 \qquad \text{für } n < m.$$

Verdeutlichen wir uns die Summennotation an zwei Beispielen.

Beispiel A.64. (Summen-Notation)

(a) Seien $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, ..., $x_k = k$, ..., $x_n = n$. Dann gelten:

$$\sum_{k=0}^{n} x_k = \sum_{k=0}^{n} k = 1 + 2 + \ldots + n,$$

$$\sum_{k=0}^{4} x_k = \sum_{k=0}^{4} k = 0 + 1 + 2 + 3 + 4 = 10,$$

$$\sum_{k=2}^{5} x_k = \sum_{k=2}^{5} k = 2 + 3 + 4 + 5 = 14.$$

(b) Sei $x_k = k^2$ für alle $k \in \mathbb{N}_0$. Dann gelten:

$$\sum_{k=1}^{5} x_k = \sum_{k=1}^{5} k^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 1 + 4 + 9 + 16 + 25 = 55,$$

$$\sum_{k=10}^{10} x_k = x_{10} = 10^2 = 100.$$

Überlegen Sie sich selber weitere Beispiele. ♠

Indem man die Summen in dem nachfolgenden Hilfssatz ausschreibt, erhält man die folgenden Rechenregeln für Summen.

Hilfssatz A.65. (Rechenregeln für Summen)
Es seien $m, n \in \mathbb{Z}$ mit $m \leq n$. Dann gelten die folgenden Rechenregeln für

Summen:

$$\sum_{k=m}^{n} x_k + \sum_{k=m}^{n} y_k = \sum_{k=m}^{n} (x_k + y_k), \tag{A.17}$$

$$\sum_{k=m}^{n} x_k - \sum_{k=m}^{n} y_k = \sum_{k=m}^{n} (x_k - y_k), \tag{A.18}$$

$$\sum_{k=m}^{n} c\, x_k = c \sum_{k=m}^{n} x_k \qquad \textit{für alle } c \in \mathbb{R},$$

$$\sum_{k=m}^{p} x_k + \sum_{k=p+1}^{n} x_k = \sum_{k=m}^{n} x_k, \qquad \textit{wenn } p \in \mathbb{Z} \textit{ mit } m \leq p < n. \tag{A.19}$$

Formel (A.19) besagt, dass wir die Summe in zwei Teilsummen zerlegen können. Man kann bei einer Summe auch den Summationsindex um $p \in \mathbb{N}$ nach rechts bzw. links verschieben:

$$\sum_{k=m}^{n} x_k = \sum_{\ell=m+p}^{n+p} x_{\ell-p} \qquad (\textit{Indexverschiebung nach rechts}), \tag{A.20}$$

$$\sum_{k=m}^{n} x_k = \sum_{\ell=m-p}^{n-p} x_{\ell+p} \qquad (\textit{Indexverschiebung nach links}). \tag{A.21}$$

Erklärung zu (A.20) und (A.21): Formal werden die Indexverschiebungen (A.20) bzw. (A.21) durchgeführt, indem man den neuen Summationsindex $\ell = k + p$ (Indexverschiebung nach rechts) bzw. $\ell = k - p$ (Indexverschiebung nach links) einführt und damit $k = \ell - p$ (Indexverschiebung nach rechts) bzw. $k = \ell + p$ (Indexverschiebung nach links) erhält und entsprechend ersetzt. In (A.20) erhält man für den neuen Summationsindex $\ell = k + p$ die neue untere bzw. obere Grenze $m + p$ bzw. $n + p$, und der Index k in x_k wird durch $k = \ell - p$ ersetzt. Bei (A.21) geht man analog vor.

Betrachten wir zwei Beispiele, in denen die Rechenregeln für Summen aus Hilfssatz A.65 angewendet werden.

Beispiel A.66. (Rechnen mit Summen)

$$\sum_{k=1}^{n} k - \sum_{k=1}^{n} (k-1) = \sum_{k=1}^{n} \big[k - (k-1)\big] = \sum_{k=1}^{n} 1 = n.$$

Wie man sieht, ist die Berechnung durch die Regel (A.18) für die Subtraktion von Summen erheblich vereinfacht worden. ♠

Beispiel A.67. (Rechnen mit Summen)
Beim Berechnen von

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

bemerken wir zuerst, dass die Terme hinter dem jeweiligen Summenzeichen durch das Ersetzen von k durch $k+1$ ineinander überführt werden können. Daher führen wir in der zweiten Summe die Indexverschiebung $\ell = k + 1$ (vgl. (A.20)) durch und erhalten die neue untere Grenze $1 + 1 = 2$ bzw. die neue obere Grenze $n + 1$. Anschließend benennen wir ℓ wieder in k um.

$$\sum_{k=1}^{n} k^2 - \sum_{k=1}^{n} (k+1)^2 = \sum_{k=1}^{n} k^2 - \sum_{\ell=2}^{n+1} \ell^2 = \sum_{k=1}^{n} k^2 - \sum_{k=2}^{n+1} k^2.$$

Der Unterschied zwischen den beiden Summen besteht nun nur noch in den Grenzen für den Summationsindex. In der ersten Summe wird über $k = 1, 2, \ldots, n$ summiert, und in der zweiten Summe wird über $k = 2, \ldots, n, n+1$ summiert. Intuitiv ist damit klar, dass bei der Subtraktion beiden Summen genau der erste Term der ersten Summe und der letzte Term der zweiten Summe übrig bleiben. Wir nutzen (A.19), um aus der ersten Summe den Term für $k = 1$ und aus der zweiten Summe den Term mit $k = n + 1$ herauszuziehen, und erhalten

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

Insgesamt erhalten wir also

$$\sum_{k=1}^{n} k^2 - \sum_{k=1}^{n} (k+1)^2 = -n^2 - 2\,n.$$

Überlegen Sie sich selber weitere Beispiele. ♠

ANHANG B

Berechnung von Integralen

In diesem Anhang sind einige wichtige Resultate und Regeln zur Berechnung von Integralen zusammengestellt.

B.1 Geometrische Anschauung des Integrals

Seien $I = [a; b]$ ein **abgeschlossenes** Intervall und $f : [a; b] \to [0; \infty[$ eine **stetige** Funktion. Gesucht ist der **Flächeninhalt** des Bereichs, der von dem Graphen von f und der Geraden $y = 0$ (also der x-Achse) sowie den Senkrechten durch $x = a$ und $x = b$ berandet wird. Wir werden diesen Flächeninhalt als

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

bezeichnen und so das **bestimmte Integral von f über** $[a; b]$ (d.h. von $x = a$ bis $x = b$) definieren.

Um diesen Flächeninhalt zu berechnen, geht man wie folgt vor (vgl. Abbildung B.1):

(1) Man zerlegt $[a; b]$ in n Teilintervalle $I_k = [x_{k-1}; x_k]$, $k = 1, 2, \ldots, n$, wobei

$$a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b.$$

(2) In jedem Teilintervall I_k wählt man eine Stützstelle $\xi_k \in I_k$.

(3) Man betrachtet die Rechtecke R_k mit Grundlinie I_k und Höhe $f(\xi_k)$. Der Flächeninhalt von R_k ist dann

$$|R_k| = f(\xi_k)\,(x_k - x_{k-1}).$$

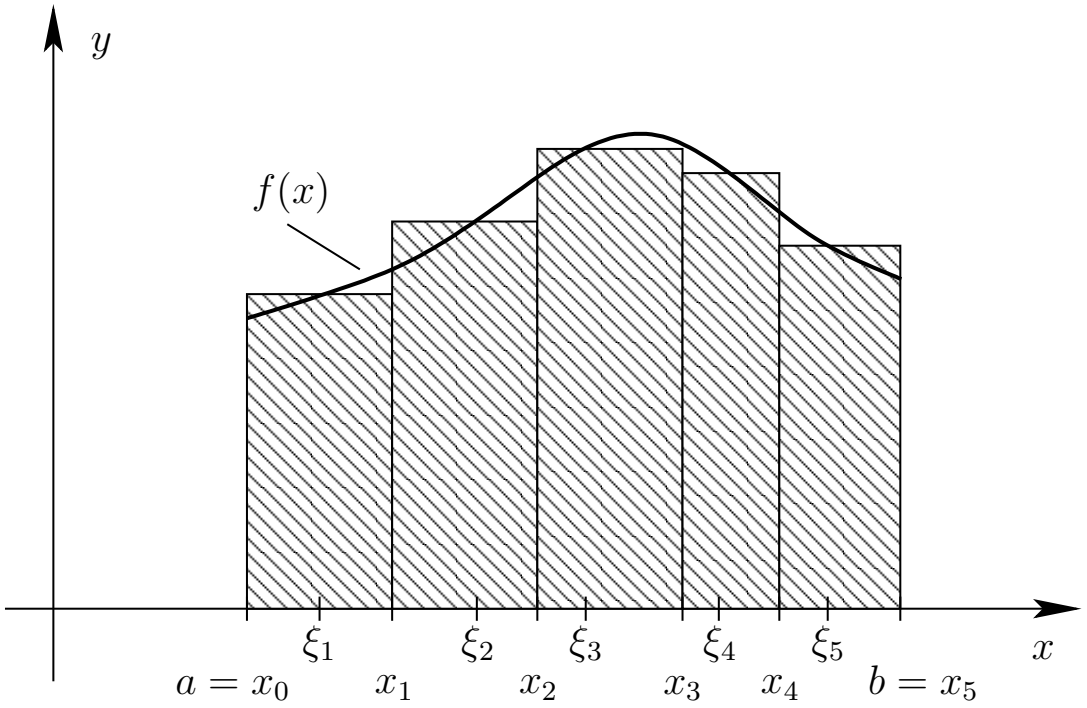


Abb. B.1: Der Flächeninhalt zwischen der Funktion f und der x-Achse von $x = a$ bis $x = b$ wird mit geeigneten Rechtecken abgedeckt. Die Summe der Flächeninhalte dieser Rechtecke ergibt eine Näherung für den Flächeninhalt zwischen dem Graphen der Funktion f und der x-Achse von $x = a$ bis $x = b$, also für den Wert des Integrals.

(4) Man addiert die Flächeninhalte auf:

$$S = \sum_{k=1}^{n} |R_k| = \sum_{k=1}^{n} f(\xi_k)\,(x_k - x_{k-1}).$$

Dieses ergibt eine Näherung des gesuchten Flächeninhalts A.

Wenn man immer mehr Teilintervalle nimmt, so dass die maximale Breite dieser Teilintervalle, also

$$\max_{1\leq k\leq n} \big(x_k - x_{k-1}\big),$$

immer schmaler wird, so wird die Näherung für den Flächeninhalt immer besser. Führt man nun einen Grenzübergang (für $n \to \infty$) durch, bei dem die maximale Breite dieser Teilintervalle gegen null strebt, so erhält man den exakten Wert A des gesuchten Flächeninhaltes.

Was passiert, wenn nicht alle Funktionswerte der stetigen Funktion f größer oder gleich null sind?

Für stetige Funktionen $f : [a; b] \to]-\infty; 0]$ können wir analog vorgehen: Wir

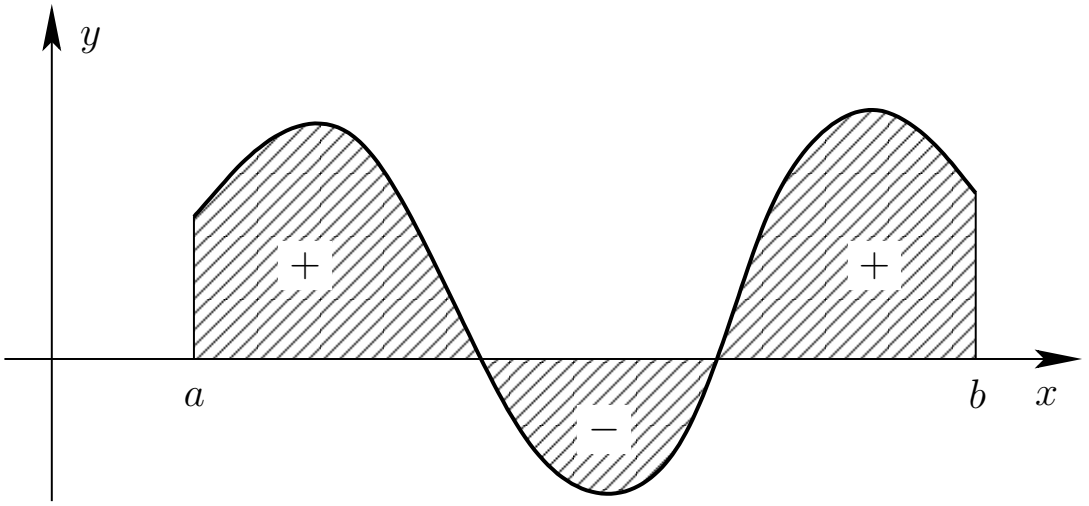


Abb. B.2: Geometrische Interpretation des Integrals $\int_a^b f(x)\,\mathrm{d}x$ als Flächeninhalt.

suchen den Flächeninhalt des Bereiches, der durch den Graphen von f, die x-Achse und die senkrechten Geraden $x = a$ und $x = b$ begrenzt wird. Allerdings definieren wir $\int_a^b f(x)\,\mathrm{d}x$ als -1 mal diesen Flächeninhalt, da die Fläche unterhalb der x-Achse liegt.

Für stetige Funktionen $f : [a; b] \to \mathbb{R}$ mit beliebigen (also möglicherweise positiven und negativen) Funktionswerten weisen wir den Flächeninhalten der Flächenstücke zwischen Graphen und x-Achse (von $x = a$ bis $x = b$) **oberhalb** der x-Achse ein positives Vorzeichen zu und den Flächeninhalten der Flächenstücke zwischen Graphen und x-Achse (von $x = a$ bis $x = b$) **unterhalb** der x-Achse ein negatives Vorzeichen zu (vgl. Abbildung B.2). Dann summieren wir diese „Flächeninhalte mit Vorzeichen“ auf und erhalten so $\int_a^b f(x)\,\mathrm{d}x$.

Das Integral wird nur für sogenannte beschränkte Funktionen eingeführt. Entgegen unserer Anschauung wird die Stetigkeit der zu integrierenden Funktion nicht vorausgesetzt.

Definition B.1. (beschränkte Funktion)
*Seien $D \subseteq \mathbb{R}$ und $f : D \to \mathbb{R}$ eine Funktion. f heißt **beschränkt**, wenn es eine Schranke $S \geq 0$ gibt, so dass $|f(x)| \leq S$ für alle $x \in D$ ist.*

Beispiel B.2. (beschränkte Funktionen)

(a) Die Funktion $\sin : \mathbb{R} \to \mathbb{R}$ ist beschränkt, denn es gilt $|\sin(x)| \leq 1$ für alle $x \in \mathbb{R}$. Eine Schranke ist hier also $S = 1$.

(b) Die Funktion $f :]0;\infty[\to \mathbb{R}$, $f(x) = 1/x$, ist nicht beschränkt (also unbeschränkt), denn $f(x) = 1/x$ wird beliebig groß, wenn man sich von rechts dem unteren Intervallende 0 nähert.

Insbesondere ist jede stetige Funktion auf einem abgeschlossenen Intervall beschränkt, denn sie hat dort ein Maximum und ein Minimum. Genauer: Ist eine Funktion $f : [a;b] \to \mathbb{R}$ stetig, so gilt

$$m = \min_{t\in[a;b]} f(t) \leq f(x) \leq \max_{t\in[a;b]} f(t) = M \qquad \text{für alle } x \in [a;b]$$

und damit $|f(x)| \leq \max\big\{|m|;|M|\big\} = S$ für alle $x \in [a;b]$. ♠

B.2 Elementare Rechenregeln für Integrale

Im Folgenden sei die **Menge aller über ein Intervall** $[a;b]$ **integrierbaren Funktionen** (also die Menge aller Funktionen, für die

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

existiert) mit $\mathcal{R}\big([a;b]\big)$ bezeichnet.

Das Integral hat die folgenden elementaren Eigenschaften.

Satz B.3. (Eigenschaften des Integrals)
Seien $f : [a;b] \to \mathbb{R}$ und $g : [a;b] \to \mathbb{R}$ beschränkte Funktionen.

(1) Ist $a < c < b$, so gilt

$$f \in \mathcal{R}\big([a;b]\big) \qquad \Longleftrightarrow \qquad f \in \mathcal{R}\big([a;c]\big) \quad \text{und} \quad f \in \mathcal{R}\big([c;b]\big).$$

In diesem Fall gilt: $$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x.$$

(2) Ist $f \in \mathcal{R}\big([a;b]\big)$ und $\alpha \in \mathbb{R}$, so ist $\alpha f \in \mathcal{R}\big([a;b]\big)$ und

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

(3) Sind $f, g \in \mathcal{R}\big([a;b]\big)$, so ist auch $f + g \in \mathcal{R}\big([a;b]\big)$ und es gilt

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

(4) Sind $f, g \in \mathcal{R}\big([a;b]\big)$ und gilt $f(x) \leq g(x)$ für alle $x \in [a;b]$, so ist

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

(5) Ist $f \in \mathcal{R}\big([a;b]\big)$, so ist auch $|f| \in \mathcal{R}\big([a;b]\big)$, und es gilt

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

B.3 Hauptsatz der Integralrechnung

Der Hauptsatz der Differentialrechnung stellt den Zusammenhang zwischen Integral und Ableitung her. Zunächst benötigen wir den Begriff einer Stammfunktion.

Definition B.4. (Stammfunktion)
*Seien I ein Intervall, $f : I \to \mathbb{R}$ und $F : I \to \mathbb{R}$. Falls F in I differenzierbar ist und $F' = f$ gilt, so heißt F eine **Stammfunktion** von f.*

Betrachten wir ein paar Beispiele von Stammfunktionen.

Beispiel B.5. (Stammfunktionen)

(a) Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Dann ist $F : \mathbb{R} \to \mathbb{R}$, $F(x) = x^3/3$, eine Stammfunktion von f, denn es gilt

$$F'(x) = \frac{1}{3} \cdot 3\,x^2 = x^2 = f(x) \qquad \text{für alle } x \in \mathbb{R}.$$

$F(x) = x^3/3$ ist aber nicht die einzige Stammfunktion von $f(x) = x^2$, denn z.B. sind

$$G : \mathbb{R} \to \mathbb{R}, \quad G(x) = \frac{1}{3}\,x^3 + 5, \quad \text{und} \quad H : \mathbb{R} \to \mathbb{R}, \quad H(x) = \frac{1}{3}\,x^3 - e,$$

ebenfalls Stammfunktionen von $f(x) = x^2$.

(b) Sei $g : \mathbb{R} \to \mathbb{R}$, $g(x) = e^x$. Dann ist jede Funktion der Form $G : \mathbb{R} \to \mathbb{R}$, $G(x) = e^x + c$, mit einer beliebigen Konstante c eine Stammfunktion von g,

denn

$$G'(x) = \big(e^x + c\big)' = e^x = g(x) \qquad \text{für alle } x \in \mathbb{R}.$$

(c) Sei $h :]0;\infty[\to \mathbb{R}$, $h(x) = 1/x$. Dann ist jede Funktion $H :]0;\infty[\to \mathbb{R}$, $H(x) = \ln(x) + c$, mit einer beliebigen Konstante c eine Stammfunktion von h, denn

$$H'(x) = \big(\ln(x) + c\big)' = \frac{1}{x} = h(x) \qquad \text{für alle } x \in]0;\infty[\,.$$

Ist F eine Stammfunktion von f, so ist auch $F + c$ für jede Konstante c eine Stammfunktion von f, denn $(F + c)' = F' + 0 = f$. ♠

Nach dieser Vorbereitung können wir den Hauptsatz der Differential- und Integralrechnung formulieren.

Satz B.6. (Hauptsatz der Differential- und Integralrechnung)
Sei $f : [a;b] \to \mathbb{R}$ stetig. Dann gelten:

(1) Die Funktion

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

*ist eine **Stammfunktion** von f.*

(2) Ist umgekehrt F irgendeine Stammfunktion von f, so gibt es eine Konstante $c \in \mathbb{R}$ mit

$$F(x) = c + \int_a^x f(t)\,\mathrm{d}t$$

und es gilt

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a) = \Big[F(x)\Big]_{x=a}^{x=b}. \tag{B.1}$$

Warum ist der Hauptsatz der Differential- und Integralrechnung so wichtig? Mit unserem Wissen über Ableitungen können wir Stammfunktionen bestimmen; und mit (B.1) können wir Integrale bequem berechnen, wenn wir eine Stammfunktion F des Integranden f kennen.

Bemerkung B.7. (Stammfunktion und unbestimmtes Integral)
Ist F eine Stammfunktion von f, so schreiben wir auch

$$\int f(x)\,\mathrm{d}x$$

für $F(x)$. $\int f(x)\,\mathrm{d}x$ ist bis auf eine additive Konstante eindeutig bestimmt. $\int f(x)\,\mathrm{d}x$ heißt auch das **unbestimmte Integral** von f im Gegensatz zu einem **bestimmten Integral** von f

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

Betrachten wir einige Beispiele für unbestimmte Integrale.

Beispiel B.8. (unbestimmte Integrale)

(a) Für $a \in \mathbb{R} \setminus \{-1\}$ gilt

$$\int x^a\,\mathrm{d}x = \frac{1}{a+1}\,x^{a+1} + c,$$

da

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{a+1}\,x^{a+1}\right) = \frac{1}{a+1}\,(a+1)\,x^{a+1-1} = x^a.$$

(b) Für $x \neq 0$ gilt $\displaystyle\int \frac{1}{x}\,\mathrm{d}x = \ln\big(|x|\big) + c$, denn

$$\ln\big(|x|\big) = \begin{cases} \ln(x) & \text{für } x > 0, \\ \ln(-x) & \text{für } x < 0 \end{cases}$$

$$\Longrightarrow \qquad \frac{\mathrm{d}}{\mathrm{d}x}\ln\big(|x|\big) = \left\{ \begin{array}{ll} \dfrac{1}{x} & \text{für } x > 0 \\ -\dfrac{1}{-x} = \dfrac{1}{x} & \text{für } x < 0 \end{array} \right\} = \frac{1}{x}\,.$$

(c) $\displaystyle\int \exp(x)\,\mathrm{d}x = \exp(x) + c$

(d) $\displaystyle\int \sin(x)\,\mathrm{d}x = -\cos(x) + c$

Definitionsmenge D	Funktion $f(x)$	Stammfunktion $F(x)$
$\mathbb{R}$	$a =$ Konstante	$a\,x + c$
$\mathbb{R}$ bzw. $\mathbb{R} \setminus \{0\}$	x^n mit $n \in \mathbb{N}$ bzw. $n \in \mathbb{Z} \setminus \big(\mathbb{N}_0 \cup \{-1\}\big)$	$\dfrac{1}{n+1}\,x^{n+1} + c$
$]0;\infty[$	x^r mit $r \in \mathbb{R} \setminus \{-1\}$	$\dfrac{1}{r+1}\,x^{r+1} + c$
$\mathbb{R}$	e^x	$e^x + c$
$\mathbb{R} \setminus \{0\}$	$\dfrac{1}{x}$	$\ln\big(\lvert x\rvert\big) + c$
$\mathbb{R}$	$\sin(x)$	$-\cos(x) + c$
$\mathbb{R}$	$\cos(x)$	$\sin(x) + c$

Tabelle B.1: Wichtige Stammfunktionen.

(e) $\displaystyle\int \cos(x)\,\mathrm{d}x = \sin(x) + c$

Überlegen Sie sich selber weitere Beispiele. ♠

Betrachten wir auch einige Beispiele für bestimmte Integrale.

Beispiel B.9. (bestimmte Integrale)

(a) $\displaystyle\int_0^{2\pi} \cos(x)\,\mathrm{d}x = \Big[\sin(x)\Big]_{x=0}^{x=2\pi} = \sin(2\pi) - \sin(0) = 0 - 0 = 0$

(b) $\displaystyle\int_1^2 \frac{1}{x}\,\mathrm{d}x = \Big[\ln\big(\lvert x\rvert\big)\Big]_{x=1}^{x=2} = \ln(2) - \ln(1) = \ln(2)$

(c) $\displaystyle\int_{-e}^{-1} \frac{1}{x}\,\mathrm{d}x = \Big[\ln\big(\lvert x\rvert\big)\Big]_{x=-e}^{x=-1} = \ln\big(\lvert -1\rvert\big) - \ln\big(\lvert -e\rvert\big)$

$$= \ln(1) - \ln(e) = 0 - 1 = -1$$

(d) $\displaystyle\int_0^{\pi} \sin(x)\,\mathrm{d}x = \Big[-\cos(x)\Big]_{x=0}^{x=\pi} = -\cos(\pi) + \cos(0) = -(-1) + 1 = 2$

(e) $\displaystyle\int_{-3}^{3} x^3 \,\mathrm{d}x = \left[\frac{1}{4}\,x^4\right]_{x=-3}^{x=3} = \frac{1}{4}\,3^4 - \frac{1}{4}\,(-3)^4 = 0$

Überlegen Sie sich selber weitere Beispiele. ♠

Die in Tabelle B.1 aufgelisteten Stammfunktionen sollten Sie kennen.

B.4 Partielle Integration

Die Methode der partiellen Integration beruht auf der **Produktregel der Differentiation** (siehe Satz 1.1 (3)): Sind $u : [a;b] \to \mathbb{R}$ und $v : [a;b] \to \mathbb{R}$ stetig differenzierbar, so gilt

$$\big(u(x)\,v(x)\big)' = u'(x)\,v(x) + u(x)\,v'(x). \tag{B.2}$$

Da alle auftretenden Funktion u, v, u' und v' auf $[a;b]$ stetig sind können wir das Integral von (B.2) über $[a;b]$ berechnen und erhalten

$$\underbrace{\int_a^b \big(u(x)\,v(x)\big)'\,\mathrm{d}x}_{=\left[u(x)\,v(x)\right]_{x=a}^{x=b}} = \int_a^b \big(u'(x)\,v(x) + u(x)\,v'(x)\big)\,\mathrm{d}x$$

$$\iff \qquad \Big[u(x)\,v(x)\Big]_{x=a}^{x=b} = \int_a^b u'(x)\,v(x)\,\mathrm{d}x + \int_a^b u(x)\,v'(x)\,\mathrm{d}x$$

$$\iff \qquad \Big[u(x)\,v(x)\Big]_{x=a}^{x=b} - \int_a^b u(x)\,v'(x)\,\mathrm{d}x = \int_a^b u'(x)\,v(x)\,\mathrm{d}x.$$

Die Formel in der letzten Zeile bezeichnet man als die Methode der partiellen Integration. Wir halten dieses als Satz fest.

Satz B.10. (Methode der partiellen Integration)
Sind $u : [a;b] \to \mathbb{R}$ und $v : [a;b] \to \mathbb{R}$ stetig differenzierbar, so gilt:

$$\int_a^b u'(x)\,v(x)\,\mathrm{d}x = \Big[u(x)\,v(x)\Big]_{x=a}^{x=b} - \int_a^b u(x)\,v'(x)\,\mathrm{d}x.$$

Bemerkung B.11. (Partielle Integration unbestimmter Integrale)

$$\int u'(x)\,v(x)\,\mathrm{d}x = u(x)\,v(x) - \int u(x)\,v'(x)\,\mathrm{d}x.$$

Betrachten wir einige Beispiele.

Beispiel B.12. (partielle Integration: bestimmte Integrale)
Bestimmte Integrale können wir mit zwei Varianten berechnen (siehe unten): entweder direkt oder indem wir zunächst das zugehörige unbestimmte Integral berechnen und erst danach die Grenzen einsetzen.

(a) *Variante 1*: Direkte Berechnung von $\displaystyle\int_{-1}^{3} x\,e^x\,\mathrm{d}x$

$$\begin{aligned}\int_{-1}^{3} \underbrace{x}_{=v(x)}\ \underbrace{e^x}_{=u'(x)}\,\mathrm{d}t &= \Big[\underbrace{x}_{=v(x)}\ \underbrace{e^x}_{=u(x)}\Big]_{x=-1}^{x=3} - \int_{-1}^{3} \underbrace{1}_{=v'(x)}\cdot\underbrace{e^x}_{=u(x)}\,\mathrm{d}x\\ &= 3\,e^3 - (-1)\,e^{-1} - \int_{-1}^{3} e^x\,\mathrm{d}x = 3\,e^3 + e^{-1} - \Big[e^x\Big]_{x=-1}^{x=3}\\ &= 3\,e^3 + e^{-1} - \left(e^3 - e^{-1}\right) = 2\,e^3 + 2\,e^{-1}\end{aligned}$$

Variante 2: Berechnung von $\displaystyle\int_{-1}^{3} x\,e^x\,\mathrm{d}x$ mit dem unbestimmtem Integral

$$\begin{aligned}\int \underbrace{x}_{=v(x)}\ \underbrace{e^x}_{=u'(x)}\,\mathrm{d}t &= \underbrace{x}_{=v(x)}\ \underbrace{e^x}_{=u(x)} - \int \underbrace{1}_{=v'(x)}\cdot\underbrace{e^x}_{=u(x)}\,\mathrm{d}x\\ &= x\,e^x - \int e^x\,\mathrm{d}x = x\,e^x - e^x + c = (x-1)\,e^x + c\end{aligned}$$

und somit

$$\int_{-1}^{3} x\,e^x\,\mathrm{d}x = \Big[(x-1)\,e^x\Big]_{x=-1}^{x=3} = 2\,e^3 - (-2)\,e^{-1} = 2\,e^3 + 2\,e^{-1}.$$

Wir dürfen bei der Berechnung des bestimmten Integrals die Integrationskonstante weglassen, denn in (B.1) in Satz B.6 kann jede beliebige Stammfunktion gewählt werden (also auch die mit Konstante $c = 0$).

(b) *Variante 1:* Direkte Berechnung von $\int_0^\pi t\,\sin(t)\,\mathrm{d}t$

$$\int_0^\pi \underbrace{t}_{=v(t)}\,\underbrace{\sin(t)}_{=u'(t)}\,\mathrm{d}t = \Big[\underbrace{t}_{=v(t)}\,\underbrace{\big(-\cos(t)\big)}_{=u(t)}\Big]_{t=0}^{t=\pi} - \int_0^\pi \underbrace{1}_{=v'(t)}\cdot\underbrace{\big(-\cos(t)\big)}_{=u(t)}\,\mathrm{d}t$$

$$= \pi\cdot\big(-\cos(\pi)\big) - 0\cdot\big(-\cos(0)\big) + \int_0^\pi \cos(t)\,\mathrm{d}t$$

$$= \pi - 0 + \Big[\sin(t)\Big]_{t=0}^{t=\pi} = \pi + \sin(\pi) - \sin(0) = \pi + 0 - 0 = \pi$$

Variante 2: Berechnung von $\int_0^\pi t\,\sin(t)\,\mathrm{d}t$ mit dem unbestimmtem Integral

$$\int \underbrace{t}_{=v(t)}\,\underbrace{\sin(t)}_{=u'(t)}\,\mathrm{d}t = \underbrace{t}_{=v(t)}\,\underbrace{\big(-\cos(t)\big)}_{=u(t)} - \int \underbrace{1}_{=v'(t)}\cdot\underbrace{\big(-\cos(t)\big)}_{=u(t)}\,\mathrm{d}t$$

$$= -t\,\cos(t) + \int \cos(t)\,\mathrm{d}t = -t\,\cos(t) + \sin(t) + c$$

und somit

$$\int_0^\pi t\,\sin(t)\,\mathrm{d}t = \Big[-t\,\cos(t) + \sin(t)\Big]_{t=0}^{t=\pi}$$

$$= \big[-\pi\,\cos(\pi) + \sin(\pi)\big] - \big[-0 + \sin(0)\big] = \pi.$$

Variante 2 hat den Vorteil, dass man zunächst eine Stammfunktion des Integranden bestimmt, deren Korrektheit durch Ableiten leicht zu überprüfen ist. ♠

Beispiel B.13. (partielle Integration: unbestimmte Integrale)

(a) Berechnung von $\int \ln(x)\,\mathrm{d}x$ für $x > 0$:

$$\int \ln(x)\,\mathrm{d}x = \int \underbrace{1}_{=u'(x)}\cdot\underbrace{\ln(x)}_{=v(x)}\,\mathrm{d}x = \underbrace{x}_{=u(x)}\,\underbrace{\ln(x)}_{=v(x)} - \int \underbrace{x}_{=u(x)}\,\underbrace{\frac{1}{x}}_{=v'(x)}\,\mathrm{d}x$$

$$= x\,\ln(x) - \int 1\,\mathrm{d}x = x\,\ln(x) - x + c.$$

(b) Berechnung von $\int x^2\, e^x \,\mathrm{d}x$: Bei diesem Integral muss man partielle Integration zweimal hintereinander anwenden.

$$\begin{aligned}
\int \underbrace{x^2}_{=v(x)}\ \underbrace{e^x}_{=u'(x)}\,\mathrm{d}x &= \underbrace{x^2}_{=v(x)}\ \underbrace{e^x}_{=u(x)} - \int \underbrace{2\,x}_{=v'(x)}\ \underbrace{e^x}_{=u(x)}\,\mathrm{d}x \\
&= x^2\, e^x - 2 \int \underbrace{x}_{=\widetilde{v}(x)}\ \underbrace{e^x}_{=\widetilde{u}'(x)}\,\mathrm{d}x \\
&= x^2\, e^x - 2 \left(\underbrace{x}_{=\widetilde{v}(x)}\ \underbrace{e^x}_{=\widetilde{u}(x)} - \int \underbrace{1}_{=\widetilde{v}'(x)} \cdot \underbrace{e^x}_{=\widetilde{u}(x)}\,\mathrm{d}x \right) \\
&= x^2\, e^x - 2\left(x\, e^x - e^x + c\right) \\
&= \left(x^2 - 2\,x + 2\right) e^x + \widetilde{c} \qquad \text{mit } \widetilde{c} = -2\,c.
\end{aligned}$$

(c) Berechnung von $\int t\, \ln(t)\,\mathrm{d}t$ für $t > 0$ mit zwei Varianten:

Variante 1:

$$\begin{aligned}
\int \underbrace{t}_{=u'(t)}\ \underbrace{\ln(t)}_{=v(t)}\,\mathrm{d}t &= \underbrace{\frac{1}{2}\,t^2}_{=u(t)}\ \underbrace{\ln(t)}_{=v(t)} - \int \underbrace{\frac{1}{2}\,t^2}_{=u(t)}\ \underbrace{\frac{1}{t}}_{=v'(t)}\,\mathrm{d}t \\
&= \frac{1}{2}\,t^2\, \ln(t) - \frac{1}{2}\int t\,\mathrm{d}t \\
&= \frac{1}{2}\,t^2\, \ln(t) - \frac{1}{4}\,t^2 + c.
\end{aligned}$$

Für *Variante 2* nutzen wir, dass wir bereits wissen (vgl. Teil (a)), dass $\big(t\, \ln(t) - t\big)' = \ln(t)$ gilt.

$$\begin{aligned}
\int \underbrace{t}_{=v(t)}\ \underbrace{\ln(t)}_{=u'(t)}\,\mathrm{d}t &= \underbrace{t}_{=v(t)}\ \underbrace{\big(t\, \ln(t) - t\big)}_{=u(t)} - \int \underbrace{1}_{=v'(t)}\ \underbrace{\big(t\, \ln(t) - t\big)}_{=u(t)}\,\mathrm{d}t \\
&= t^2\, \ln(t) - t^2 - \int t\, \ln(t)\,\mathrm{d}t + \int t\,\mathrm{d}t \\
&= t^2\, \ln(t) - t^2 - \int t\, \ln(t)\,\mathrm{d}t + \frac{1}{2}\,t^2 \\
&= t^2\, \ln(t) - \frac{1}{2}\,t^2 - \int t\, \ln(t)\,\mathrm{d}t,
\end{aligned}$$

also

$$\int t\, \ln(t)\,\mathrm{d}t = t^2\, \ln(t) - \frac{1}{2}\,t^2 - \int t\, \ln(t)\,\mathrm{d}t \quad \Big| \; + \int t\, \ln(t)\,\mathrm{d}t$$

$$\Longrightarrow \qquad 2\int t\,\ln(t)\,\mathrm{d}t = t^2\,\ln(t) - \frac{1}{2}\,t^2 + \widetilde{c} \tag{B.3}$$

$$\Longrightarrow \qquad \int t\,\ln(t)\,\mathrm{d}t = \frac{1}{2}\,t^2\,\ln(t) - \frac{1}{4}\,t^2 + c \qquad \text{mit } c = \frac{\widetilde{c}}{2}\,.$$

Da wir keine weiteren Integrale auswerten, müssen wir im Schritt (B.3) die Integrationskonstante $\widetilde{c}$ ergänzen.

Vergessen Sie bei unbestimmten Integralen nicht die Integrationskonstante. ♠

Bemerkung B.14. (Praxistipps für partielle Integration)

(1) Integrale der Form

$$\int x\,f(x)\,\mathrm{d}x$$

lassen sich durch partielle Integration lösen, sofern $\int f(x)\,\mathrm{d}x$ bekannt ist. Man wählt dann $u'(x) = f(x)$ und $v(x) = x$. Die **Ausnahme** von dieser Regel ist

$$\int x\,\ln(x)\,\mathrm{d}x.$$

Hier ist es günstiger, $u'(x) = x$ und $v(x) = \ln(x)$ zu wählen (siehe dazu Beispiel B.13 (c)).

(2) Integrale der Form

$$\int x^k\,f(x)\,\mathrm{d}x \qquad \text{mit } k \in \mathbb{N}$$

lassen sich manchmal durch (mehrfache) partielle Integration berechnen (siehe Beispiel B.13 (b)).

B.5 Die Substitutionsregel

Sei F eine Stammfunktion von f. Die Kettenregel (siehe Satz 1.2) liefert

$$(F \circ u)'(x) = F'\big(u(x)\big)\,u'(x) = f\big(u(x)\big)\,u'(x).$$

Integration über $[a;b]$ auf beiden Seiten liefert

$$\int_a^b f\big(u(x)\big)\,u'(x)\,\mathrm{d}x = \int_a^b (F \circ u)'(x)\,\mathrm{d}x = \Big[(F \circ u)'(x)\Big]_{x=a}^{x=b}$$

$$= F\big(u(b)\big) - F\big(u(a)\big) = \Big[F(t)\Big]_{t=u(a)}^{t=u(b)} = \int_{u(a)}^{u(b)} f(t)\,\mathrm{d}t.$$

Nun können wir wir die Substitutionsregel formulieren.

Satz B.15. (Substitutionsregel)
Sei $u : [a;b] \to \mathbb{R}$ stetig differenzierbar, und die Bildmenge von u, $u\big([a;b]\big) = \{u(x) \ : \ x \in [a;b]\}$, erfülle $u\big([a;b]\big) \subseteq [c,d]$. Ist $f : [c,d] \to \mathbb{R}$ stetig, so gilt:

$$\int_a^b f\big(u(x)\big)\,u'(x)\,\mathrm{d}x = \int_{u(a)}^{u(b)} f(t)\,\mathrm{d}t. \tag{B.4}$$

Bemerkung B.16. (Substitutionsregel für unbestimmte Integrale)

$$\int f\big(u(x)\big)\,u'(x)\,\mathrm{d}x = \Big[\int f(t)\,\mathrm{d}t\Big]_{t=u(x)} \tag{B.5}$$

Achtung: Rücksubstitution nicht vergessen! Es ist ganz wichtig, dass man nach dem Berechnen von $\int f(t)\,\mathrm{d}t$ wieder $t = u(x)$ einsetzt!

Betrachten wir einige Beispiele.

Beispiel B.17. (Substitutionsregel)

(a) Berechnung von $\displaystyle\int_0^2 e^{-x^2}\,x\,\mathrm{d}x$: Wir setzen

$$u = u(x) = -x^2 \qquad \Longrightarrow \qquad \frac{\mathrm{d}u}{\mathrm{d}x} = -2\,x \qquad \Longrightarrow \qquad -\frac{1}{2}\,\mathrm{d}u = x\,\mathrm{d}x$$

mit den neuen Integralgrenzen

$$u(0) = -0^2 = 0 \qquad \text{und} \qquad u(2) = -2^2 = -4$$

und erhalten mit dieser Substitution

$$\int_0^2 e^{-x^2}\,x\,\mathrm{d}x = \int_0^{-4} e^u \left(-\frac{1}{2}\right) \mathrm{d}u = \Big[-\frac{1}{2}\,e^u\Big]_{u=0}^{u=-4}$$
$$= -\frac{1}{2}\,e^{-4} - \left(-\frac{1}{2}\right) e^0 = \frac{1}{2}\,\big(1 - e^{-4}\big).$$

(b) Berechnung von $\int \sin^3(x)\,\cos(x)\,\mathrm{d}x$: Wir setzen

$$u = u(x) = \sin(x) \quad \Longrightarrow \quad \frac{\mathrm{d}u}{\mathrm{d}x} = \cos(x) \quad \Longrightarrow \quad \mathrm{d}u = \cos(x)\,\mathrm{d}x$$

und erhalten mit dieser Substitution

$$\int \sin^3(x)\,\cos(x)\,\mathrm{d}x = \left[\int u^3\,\mathrm{d}u\right]_{u=\sin(x)}$$
$$= \left[\frac{1}{4}\,u^4 + c\right]_{u=\sin(x)} = \frac{1}{4}\,\sin^4(x) + c.$$

Denken Sie bei unbestimmten Integralen an die Integrationskonstante. ♠

Bemerkung B.18. (Anwendung der Substitutionsregel)
In der Praxis wird die Substitutionsregel oft „von rechts nach links“ angewendet, d.h. wir ersetzen auf der rechten Seite von (B.5) (bzw. von (B.4)) $t = u(x)$ mit einer **injektiven** (also umkehrbaren) Funktion u und erhalten mit

$$\frac{\mathrm{d}t}{\mathrm{d}x} = u'(x) \qquad \Longleftrightarrow \qquad \mathrm{d}t = u'(x)\,\mathrm{d}x$$

somit

$$\int f(t)\,\mathrm{d}t = \left[\int f\big(u(x)\big)\,u'(x)\,\mathrm{d}x\right]_{x=u^{-1}(t)}, \tag{B.6}$$

falls f stetig und u stetig differenzierbar ist. Man beachte, dass die **Injektivität (also Umkehrbarkeit)** von u **erforderlich** ist, damit man im letzten Schritt nach der Berechnung des Integrals die Substitution $t = u(x)$ durch $x = u^{-1}(t)$ mit Hilfe der Umkehrfunktion u^{-1} von u wieder „rückgängig machen“ kann. Für bestimmte Integrale erhalten wir analog zu (B.6)

$$\int_c^d f(t)\,\mathrm{d}t = \int_{u^{-1}(c)}^{u^{-1}(d)} f\big(u(x)\big)\,u'(x)\,\mathrm{d}x. \tag{B.7}$$

Betrachten wir ein Beispiel, in dem wir eine „Rückwärts-Substitution“ benutzen.

Beispiel B.19. (Substitutionsregel)

$$\int_1^e \frac{1}{t\,\big(1+\ln(t)\big)}\,\mathrm{d}t$$

Wir wählen die Substitution $t = e^x$ ($\Leftrightarrow\ x = \ln(t)$), also

$$t = e^x \qquad \Longrightarrow \qquad \frac{\mathrm{d}t}{\mathrm{d}x} = e^x \qquad \Longrightarrow \qquad \mathrm{d}t = e^x\,\mathrm{d}x$$

und erhalten mit den neuen Grenzen $\ln(1) = 0$ und $\ln(e) = 1$

$$\int_1^e \frac{1}{t\,\big(1+\ln(t)\big)}\,\mathrm{d}t = \int_0^1 \frac{1}{e^x\,\big(1+\ln(e^x)\big)}\,e^x\,\mathrm{d}x = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x.$$

Mit der weiteren Substitution

$$y = 1 + x \qquad \Longrightarrow \qquad \frac{\mathrm{d}y}{\mathrm{d}x} = 1 \qquad \Longrightarrow \qquad \mathrm{d}y = \mathrm{d}x$$

folgt mit den neuen Grenzen $y(0) = 1$ und $y(1) = 2$

$$\begin{aligned}\int_1^e \frac{1}{t\,\big(1+\ln(t)\big)}\,\mathrm{d}t = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \int_1^2 \frac{1}{y}\,\mathrm{d}y &= \Big[\ln\big(|y|\big)\Big]_{y=1}^{y=2} \\ &= \Big[\ln(y)\Big]_{y=1}^{y=2} = \ln(2) - \ln(1) = \ln(2).\end{aligned}$$

Es gilt also $\displaystyle\int_1^e \frac{1}{t\,\big(1+\ln(t)\big)}\,\mathrm{d}t = \ln(2).$ ♠

In der nächsten Bemerkung halten wir zwei Standardsubstitutionen fest.

Bemerkung B.20. (zwei Standardsubstitutionen)

(1) Seien f stetig und $\lambda, \mu \in \mathbb{R}$ mit $\lambda \neq 0$. Dann gilt:

$$\int f(\lambda\,x + \mu)\,\mathrm{d}x = \left[\int f(u)\,\frac{1}{\lambda}\,\mathrm{d}u\right]_{u=\lambda\,x+\mu} = \left[\frac{1}{\lambda}\int f(u)\,\mathrm{d}u\right]_{u=\lambda\,x+\mu}.$$

Erklärung: Dieses folgt mit der Substitution

$$u = \lambda\,x + \mu \qquad \Longrightarrow \qquad \frac{\mathrm{d}u}{\mathrm{d}x} = \lambda \qquad \Longrightarrow \qquad \frac{1}{\lambda}\,\mathrm{d}u = \mathrm{d}x.$$

(2) Sei $f \in \mathcal{C}^1(I)$ mit $f(x) \neq 0$ in I. Dann gilt:

$$\int \frac{f'(x)}{f(x)}\,\mathrm{d}x = \left[\int \frac{1}{u}\,\mathrm{d}u\right]_{u=f(x)} = \Big[\ln\big(|u|\big) + c\Big]_{u=f(x)} = \ln\big(|f(x)|\big) + c.$$

Erklärung: Dieses folgt mit der Substitution

$$u = f(x) \qquad \Longrightarrow \qquad \frac{\mathrm{d}u}{\mathrm{d}x} = f'(x) \qquad \Longrightarrow \qquad \mathrm{d}u = f'(x)\,\mathrm{d}x.$$

Beispiel B.21. (Standardsubstitutionen)

(a) $\displaystyle\int \cos(3\,x-5)\,\mathrm{d}x = \frac{1}{3}\,\sin(3\,x-5)+c$

(b) $\displaystyle\int \frac{2\,x}{x^2+1}\,\mathrm{d}x = \ln\left(|x^2+1|\right)+c = \ln(x^2+1)+c$

Überlegen Sie sich weitere Beispiele. ♠

ANHANG C

Mathematische Aussagen und Beweistechniken

In diesem Anhang lernen wir in Teilkapitel C.1, wie man mathematische Aussagen (also z.B. die Sätze, Hilfssätze und Definitionen in diesem Skript) liest und richtig versteht. Dazu gehört, dass wir uns den Unterschied zwischen einer Implikation („wenn dann“-Aussage) und einer Äquivalenz („genau dann wenn“-Aussage) klar machen. Im nachfolgenden Teilkapitel C.2 werden wir uns dann mit verschiedenen Beweistechniken beschäftigen. Im Teilkapitel C.3 lernen wir schließlich das Prinzip der vollständigen Induktion kennen.

C.1 Implikationen und Äquivalenzen

In diesem Teilkapitel lernen wir die zwei grundlegenden mathematischen Aussagetypen Implikation („wenn dann“-Aussage) und Äquivalenz („genau dann wenn“-Aussage) kennen und studieren diese an verschiedenen Beispielen. Natürlich finden Sie überall in diesem Skript weitere Beispiele für solche mathematischen Aussagen; genau genommen ist jeder Satz und Hilfssatz ein Beispiel einer Implikation oder eine Äquivalenz. Definitionen sind als Äquivalenzen zu lesen; auch wenn in der Formulierung meist nur ein „wenn“ (und kein „genau dann wenn“) steht. Die hier gewählten Beispiele sind mit Absicht besonders einfach, damit ihr Inhalt keine Schwierigkeiten bereitet und auch, weil wir diese beweisen wollen.

Definition C.1. (Implikation/„wenn dann“-Aussage)
*Seien A und B zwei Aussagen. Die **Implikation** (oder **„wenn dann“-Aussage**) „$A \Rightarrow B$“ bedeutet „Aus A folgt B.“ oder gleichbedeutend „Wenn A gilt, dann gilt auch B.“ oder gleichbedeutend „A impliziert B.“ Dabei können wir Aussage A als die **Voraussetzung** für die **Behauptung** der Gültigkeit der Aussage B auffassen.*

Betrachten wir zunächst ein einfaches Beispiel einer Implikation.

Beispiel C.2. (Implikation/„wenn dann“-Aussage)
„Sei $n \in \mathbb{N}$. Wenn n eine gerade Zahl ist, dann ist n^2 eine gerade Zahl.“

Diese Aussage können wir auch wie folgt formulieren:

„Sei $n \in \mathbb{N}$. Aus der Aussage, n ist eine gerade Zahl, folgt, dass n^2 eine gerade Zahl ist.“

oder kürzer

„Sei $n \in \mathbb{N}$. Dann gilt: n ist eine gerade Zahl. $\Longrightarrow$ n^2 ist eine gerade Zahl.“

Hier ist „Sei $n \in \mathbb{N}$.“ die allgemeine Voraussetzung (für beide Aussagen). Die Aussage A ist „n ist eine gerade Zahl.“ und die Aussage B ist „n^2 ist eine gerade Zahl.“.

Diese Aussage ist wahr. Wir beweisen sie mit einem *direkten Beweis*:

$n \in \mathbb{N}$ ist gerade, wenn n durch 2 teilbar ist (mit Ergebnis in $\mathbb{N}$), also wenn gilt $n/2 = m \in \mathbb{N}$ oder gleichwertig $n = 2\,m$ mit $m \in \mathbb{N}$. Dann ist $n^2 = (2\,m)^2 = 2 \cdot (2\,m^2)$ und $2\,m^2 \in \mathbb{N}$, d.h. n^2 ist durch zwei teilbar. Also ist $n^2 \in \mathbb{N}$ ebenfalls gerade. ♠

Beispiel C.3. (Implikation/„wenn dann“-Aussage)
„Das Produkt einer geraden und einer ungeraden natürlichen Zahl ist eine gerade natürliche Zahl.“

Zunächst müssen wir diese Aussage sauber als Implikationen formulieren. Wir haben die folgende Voraussetzung (Aussage A): „$n \in \mathbb{N}$ ist eine gerade Zahl und $m \in \mathbb{N}$ ist eine ungerade Zahl.“ Die Behauptung (Aussage B) ist dann: „Das Produkt $n \cdot m \in \mathbb{N}$ ist eine gerade Zahl.“ Also haben wir die folgende Implikation:

„Wenn $n \in \mathbb{N}$ eine gerade Zahl und $m \in \mathbb{N}$ eine ungerade Zahl ist, dann ist das

Produkt $n \cdot m \in \mathbb{N}$ eine gerade Zahl."

bzw.

„Aus der Aussage, $n \in \mathbb{N}$ ist eine gerade Zahl und $m \in \mathbb{N}$ ist eine ungerade Zahl, folgt, dass das Produkt $n \cdot m \in \mathbb{N}$ eine gerade Zahl ist."

oder kürzer:

„$n \in \mathbb{N}$ ist eine gerade Zahl, und $m \in \mathbb{N}$ ist eine ungerade Zahl. $\Longrightarrow$ $n \cdot m \in \mathbb{N}$ ist eine gerade Zahl."

Wir wollen diese Aussage nun mit einem *direkten Beweis* beweisen:

Da n gerade ist, ist n durch zwei teilbar, d.h. $n/2 = p$ mit $p \in \mathbb{N}$. Also gilt $n = 2\,p$ mit $p \in \mathbb{N}$. Daraus folgt $n \cdot m = (2\,p)\,m = 2\,(p \cdot m)$ mit $p \cdot m \in \mathbb{N}$. Also ist $n \cdot m$ durch zwei teilbar und somit gerade.

Wir können den *direkten Beweis* auch mit Implikationspfeilen hinschreiben:

$n \in \mathbb{N}$ sei gerade und m sei ungerade.
$\Longrightarrow$ n ist durch 2 teilbar.
$\Longrightarrow$ $\dfrac{n}{2} = p$ mit $p \in \mathbb{N}$
$\Longrightarrow$ $n = 2\,p$ mit $p \in \mathbb{N}$
$\Longrightarrow$ $n \cdot m = (2\,p)\,m = 2\,(p \cdot m)$ und $p \cdot m \in \mathbb{N}$
$\Longrightarrow$ $n \cdot m$ ist durch 2 teilbar.
$\Longrightarrow$ $n \cdot m$ ist gerade.

Wir bemerken, dass wir in dem Beweis gar nicht verwendet haben, dass m ungerade ist. Dieses liegt daran, dass $n \cdot m$ auch dann gerade ist, wenn n und m beide gerade sind! ♠

Wir lernen nun den wichtigen Begriff der Äquivalenz kennen.

Definition C.4. (Äquivalenz/„genau dann wenn"-Aussage)
*Zwei Aussagen A, B sind **äquivalent** (in Zeichen „$A \Leftrightarrow B$") wenn die Implikationen „$A \Rightarrow B$", „$B \Rightarrow A$" beide gelten. Man bezeichnet „$A \Leftrightarrow B$" als **Äquivalenz** (oder **Äquivalenzaussage**), und wir sagen „Aussage A gilt **genau dann, wenn** Aussage B gilt." oder gleichbedeutend „Aussage A und Aussage B sind **äquivalent**.".*

Betrachten wir zwei Beispiele für Äquivalenzaussagen.

Beispiel C.5. (Äquivalenz)
Die Äquivalenzaussage

$$n^2 = 4 \quad \Longleftrightarrow \quad \big(n = 2 \text{ oder } n = -2\big),$$

oder in Worten

„Die Zahl n^2 hat genau dann den Wert 4, wenn $n = 2$ oder $n = -2$ gilt."

bedeutet:

$$n^2 = 4 \quad \Longrightarrow \quad \big(n = 2 \text{ oder } n = -2\big), \tag{C.1}$$

$$\big(n = 2 \text{ oder } n = -2\big) \quad \Longrightarrow \quad n^2 = 4. \tag{C.2}$$

Um diese Aussage mit einem direkten Beweis nachzuweisen, müssen wir also beide Implikationen beweisen.

Beweis von (C.1): Sei also $n^2 = 4$. Dann ist $n = 2 = \sqrt{4}$ eine Lösung der Gleichung $n^2 = 4$. Weiter gilt aber auch $(-2)^2 = 4$. Damit sind $n_1 = 2$ und $n_2 = -2$ beides Lösungen von $n^2 = 4$. Eine quadratische Gleichung hat aber aber maximal zwei verschiedene Lösungen. Also haben wir mit $n_1 = 2$ und $n_2 = -2$ alle Lösungen von $n^2 = 4$ gefunden.

Beweis von (C.2): Für $n = 2$ finden wir $n^2 = 2^2 = 4$, und für $n = -2$ finden wir $n^2 = (-2)^2 = 4$. Also gilt in beiden Fällen $n^2 = 4$. ♠

Bemerkung C.6. (Implikationen sind oft keine Äquivalenzen!)
Nicht alle Aussagen sind Äquivalenzen!

Beispiel: Wir haben in Abwandlung des vorigen Beispiels sehr wohl

$$n = 2 \quad \Longrightarrow \quad n^2 = 4,$$

aber aus $n^2 = 4$ folgt nicht $n = 2$ (sondern „$n = 2$ oder $n = -2$").

Beispiel C.7. (Äquivalenz)
„Seien $m, n \in \mathbb{N}$. Dann gilt $m < n$ genau dann, wenn $m^2 < n^2$ ist."

oder gleichbedeutend aber kürzer:

„Seien $m, n \in \mathbb{N}$. Dann gilt: $m < n \iff m^2 < n^2$.

Wir müssen also die folgenden beiden Aussagen zeigen:

$$\text{Seien } m, n \in \mathbb{N}. \text{ Dann gilt: } m < n \Longrightarrow m^2 < n^2. \tag{C.3}$$

$$\text{Seien } m, n \in \mathbb{N}. \text{ Dann gilt: } m^2 < n^2 \Longrightarrow m < n. \tag{C.4}$$

Direkter Beweis von (C.3)*:* Seien $m, n \in \mathbb{N}$ mit $m < n$ beliebig. Wegen $m, n \in \mathbb{N}$ gilt $m > 0$ und $n > 0$. Damit folgt

$$m^2 = \underbrace{m}_{>0} \cdot \underbrace{m}_{0<m<n} < \underbrace{m}_{0<m<n} \cdot \underbrace{n}_{>0} < n \cdot n = n^2. \tag{C.5}$$

Also folgt $m^2 < n^2$.

Direkter Beweis von (C.4)*:* Seien $m, n \in \mathbb{N}$ mit $m^2 < n^2$ beliebig. Es muss gelten $m < n$ oder $m \geq n$ (mehr Fälle gibt es nicht).

Für $m < n$ finden wir mit der Ungleichungskette (C.5), dass $m^2 < n^2$ gilt.

Für $m \geq n$ folgt wegen $m > 0$ und $n > 0$ (da $m, n \in \mathbb{N}$), dass gilt

$$m^2 = \underbrace{m}_{>0} \cdot \underbrace{m}_{0<n\leq m} \geq \underbrace{m}_{0<n\leq m} \cdot \underbrace{n}_{>0} \geq n \cdot n = n^2,$$

d.h. es gilt $m^2 \geq n^2$. Also kann für $m \geq n$ die Aussage $m^2 < n^2$ nicht gelten.

Aus beiden Falluntersuchungen zusammen sieht man nun, dass aus $m^2 < n^2$ nur $m < n$ folgt. ♠

C.2 Beweistechniken

In diesem Teilkapitel lernen wir die klassischen Beweistechniken kennen, von denen uns schon einige im vorigen Teilkapitel in den verschiedenen Beispielen begegnet sind. Die einzige Beweistechnik, die wir hier nicht behandeln, ist die vollständige Induktion. Diese wird in Teilkapitel C.3 ausführlich besprochen.

Beweistechnik C.8. (direkter Beweis)
Beweist man eine Implikation $A \Rightarrow B$ in der Mathematik mit einem ***direkten Beweis****, so führt man einige Beweisschritte/Implikationen nacheinander aus, bis man von A nach B kommt: $A \Rightarrow C_1 \Rightarrow C_2 \Rightarrow \ldots \Rightarrow C_n \Rightarrow B$. Dabei stellen $C_1, C_2, \ldots, C_n$ Aussagen dar, die als Zwischenergebnisse nach den einzelnen Beweisschritten erreicht werden.*

Beispiel C.9. (direkter Beweis)
Die Aussage

„Sei $n \in \mathbb{N}$. Wenn n gerade ist, dann ist die Funktion $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^n$, eine gerade Funktion."

beweist man mit einem *direkten Beweis* wie folgt: (Zur Erinnerung: Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ ist gerade, wenn $f(x) = f(-x)$ für alle $x \in \mathbb{R}$ gilt.)

Direkter Beweis: Sei $n \in \mathbb{N}$ gerade. Dann existiert $m \in \mathbb{N}$ so dass $n = 2\,m$, und $f(x) = x^n = x^{2m} = (x^2)^m$. Somit gilt für alle $x \in \mathbb{R}$

$$f(-x) = (-x)^n = \big((-x)^2\big)^m = \big(\underbrace{(-1)^2}_{=1}\, x^2\big)^m = (x^2)^m = x^n = f(x),$$

d.h. f ist gerade. ♠

Eine häufig nützliche Beweistechnik ist der Beweis durch Widerspruch. In manchen Situationen ist der **Beweis durch Widerspruch sehr viel einfacher zu führen** als ein direkter Beweis.

Beweistechnik C.10. (Beweis durch Widerspruch)
Eine mathematische Aussage

„Wenn die Aussage A gilt, dann gilt die Aussage B."

oder

„Unter gewissen Voraussetzungen gilt die Aussage B"

*können wir wie folgt mit einem sogenannten **Widerspruchsbeweis** beweisen: Wir nehmen an, dass die Aussage A bzw. die Voraussetzungen gelten. Dann nehmen wir an, **dass die Aussage** B **nicht gilt**, d.h. wir nehmen an, **dass die Verneinung (oder Negation) der Aussage** B gilt. Wenn wir hieraus einen **Widerspruch** zu bereits bekannten Aussagen oder zu den Voraussetzungen herleiten können, dann war unsere Annahme, dass die Verneinung (oder Negation) der Aussage B gilt, falsch. Also muss die Aussage B gelten.*

Betrachten wir zunächst ein einfaches Beispiel, um uns klar zu machen, wie ein Widerspruchsbeweis funktioniert.

Beispiel C.11. (Beweis durch Widerspruch)
Wir wollen die folgende Aussage mit einem Beweis durch Widerspruch beweisen:

„Sei $n \in \mathbb{N}$. Dann gilt: Wenn n gerade ist, dann ist n^2 gerade."

Als Voraussetzung bzw. Aussage A haben wir dann, dass $n \in \mathbb{N}$ gerade ist, und als Behauptung bzw. Aussage B haben wir, dass n^2 gerade ist. Für den Widerspruchsbeweis nehmen wir an, dass die Voraussetzung wahr ist, aber dass die Behauptung falsch ist, d.h. dass ihre Negation wahr ist.

Beweis durch Widerspruch: Sei also $n \in \mathbb{N}$ gerade, und es gelte $n^2 \in \mathbb{N}$ ist nicht gerade. Dann ist n^2 nicht durch 2 teilbar. Daraus folgt, dass n nicht durch 2 teilbar ist (denn ansonsten wäre n^2 auch durch 2 teilbar). Also ist n **nicht gerade** $\lightning$, und wir haben einen **Widerspruch** (denn per Annahme war n gerade). – Da wir einen Widerspruch gefunden haben, folgt, dass die Annahme, dass n^2 nicht gerade ist, falsch war. Also muss n^2 gerade sein.

Das Symbol $\lightning$ schreibt man häufig dort hin, wo der Widerspruch auftritt. ♠

Betrachten wir noch ein aufwendigeres Beispiel. Aus der Schule wissen Sie, das $\sqrt{2}$ keine rationale Zahl sondern eine irrationale Zahl ist (d.h. $\sqrt{2}$ ist eine reelle Zahl, die man nicht als einen Bruch schreiben kann). Dieses wollen wir nun beweisen.

Beispiel C.12. (Beweis durch Widerspruch)
Wir wollen die folgende Aussage beweisen:

„Die Zahl $\sqrt{2}$ ist nicht in $\mathbb{Q}$."

Wir formulieren dieses besser (aber äquivalent) als:

„Sei x die nicht-negative reelle Zahl mit $x^2 = 2$. Dann ist x nicht in $\mathbb{Q}$."

Hierbei haben wir benutzt, das die Quadratwurzel $\sqrt{2}$ gerade als die nicht-negative Zahl x in $\mathbb{R}$ mit $x^2 = 2$ definiert ist.

Hier ist also die Voraussetzung (Aussage A) „Sie x die nicht-negative reelle Zahl mit $x^2 = 2$.", und die Behauptung (Aussage B) ist „x ist nicht in $\mathbb{Q}$."

Wir wollen einen *Widerspruchsbeweis* geben. Also nehmen wir an, dass die Voraussetzung (Aussage A) gilt, aber die Behauptung falsch ist, also dass die Negation der Behauptung (also die Aussage $\neg B$) gilt:

Widerspruchsbeweis: Sei x die nicht-negative Zahl in $\mathbb{R}$ mit $x^2 = 2$. Wir nehmen an, dass x in $\mathbb{Q}$ liegt. Dann gibt es Zahlen $p \in \mathbb{N}$ und $q \in \mathbb{N}$ mit

$$x = \frac{p}{q}\,. \tag{C.6}$$

Wir dürfen annehmen, dass wir in dem Bruch $x = p/q$ den Zähler p und Nenner q nicht mehr kürzen können, also dass p und q keine gemeinsamen Teiler haben.

Durch Quadrieren auf beiden Seiten vom (C.6) erhalten wir

$$\underbrace{x^2}_{=2} = \left(\frac{p}{q}\right)^2 = \frac{p^2}{q^2} \quad \Longrightarrow \quad 2 = \frac{p^2}{q^2} \quad \Longrightarrow \quad 2\,q^2 = p^2 \quad \Longrightarrow \quad p^2 = 2\,q^2.$$

Aus $p^2 = 2\,q^2$ folgt, dass p^2 durch 2 teilbar ist, denn $p^2/2 = q^2 \in \mathbb{N}$ (da $q \in \mathbb{N}$). Dann ist auch p durch 2 teilbar (denn wäre p nicht durch 2 teilbar, so wäre auch p^2 nicht durch 2 teilbar $\lightning$). Also gilt $p/2 = m$ mit $m \in \mathbb{N}$, d.h. $p = 2\,m$ mit $m \in \mathbb{N}$.

Einsetzen von $p = 2\,m$ in $p^2 = 2\,q^2$ liefert nun

$$(2\,m)^2 = 2\,q^2 \quad \Longrightarrow \quad 2\,(2\,m^2) = 2\,q^2 \quad \Longrightarrow \quad 2\,m^2 = q^2. \quad \Longrightarrow \quad q^2 = 2\,m^2.$$

Also ist (mit der gleichen Argumentation wie oben) q^2 ebenfalls durch 2 teilbar. Dann ist auch q durch 2 teilbar (denn wäre q nicht durch 2 teilbar, so wäre auch q^2 nicht durch 2 teilbar $\lightning$). Also gilt $q/2 = n$ mit $n \in \mathbb{N}$, d.h. $q = 2\,n$ mit $n \in \mathbb{N}$.

Wir haben also gefunden, dass sowohl p also auch q durch 2 teilbar sind, also $p = 2\,m$ und $q = 2\,n$ mit $m, n \in \mathbb{N}$. Damit finden wir

$$x = \frac{p}{q} = \frac{2\,m}{2\,n} = \frac{m}{n}\,,$$

und dieses steht im **Widerspruch** zu unserer Annahme, dass der Zähler p und Nenner q in $x = p/q$ keine gemeinsamen Teiler hatten. $\lightning$

Da wir einen Widerspruch hergeleitet haben, war unsere Annahme, dass $x = \sqrt{2}$ rational ist falsch. Also haben wir gezeigt, dass $x = \sqrt{2}$ irrational ist, also $x = \sqrt{2} \notin \mathbb{Q}$. ♠

Aussagen, die für eine ganze Klasse von Objekten (also alle Objekte der Klasse) gelten sollen, (sogenannte „Allaussagen“) kann man durch ein Gegenbeispiel widerlegen, wenn diese falsch sind.

Beweistechnik C.13. (Widerlegen von „Allaussagen“ durch Angeben eines Gegenbeispiels)
Will man eine Aussage A der Gestalt „Für alle x aus der Menge M gelten die Eigenschaften $E_1, E_2, \ldots, E_n$.“ ***widerlegen****, so reicht es* ***ein Gegenbeispiel****, d.h. ein $x \in M$, für dass $E_1, E_2, \ldots, E_n$ nicht alle gelten, zu finden und nachzuweisen, dass für dieses mindestens eine der Eigenschaften $E_1, E_1, \ldots, E_n$ verletzt ist.*

Betrachten wir zwei Beispiele.

Beispiel C.14. (Widerlegen von „Allaussagen“ durch Gegenbeispiel)
Wir wollen zeigen, dass die Aussage „Alle Schafe in England sind weiß.“ falsch ist. Dazu reicht es, wenn wir ein Schaf in England finden, das nicht weiß (sondern beispielsweise braun, schwarz oder gescheckt) ist. ♠

Beispiel C.15. (Widerlegen von „Allaussagen“ durch Gegenbeispiel)
Betrachten wir die folgende Allaussage:

„Alle Polynomfunktionen vom Grad ≤ 2 sind gerade Funktionen.“

Hier ist die betrachte Menge M die Menge aller Polynomfunktionen vom Grad ≤ 2, also

$$M = \big\{p : \mathbb{R} \to \mathbb{R},\ p(x) = a_2\, x^2 + a_1\, x + a_0\ :\ a_0, a_1, a_2 \in \mathbb{R}\big\}.$$

Beweis: Um die Aussage „Alle Polynomfunktionen vom Grad ≤ 2 sind gerade.“ zu widerlegen, reicht es eine Polynomfunktion vom Grad ≤ 2 zu finden, die nicht gerade ist. Betrachte hierzu $p : \mathbb{R} \to \mathbb{R}$, $p(x) = x$. Dann ist $p(-x) = -x = -p(x)$ für alle $x \in \mathbb{R}$ und für $x \neq 0$ gilt $p(-x) = -x \neq x = p(x)$, d.h. p ist eine ungerade und keine gerade Funktion. Da wir ein Gegenbeispiel gefunden haben, war die Aussage falsch. ♠

Existenzaussagen kann man dagegen beweisen, indem man ein Objekt mit den gesuchten Eigenschaften findet.

Beweistechnik C.16. (Beweisen von Existenzaussagen durch Angeben eines Beispiels)
Will man eine Aussage der Form „Es existiert ein Objekt x mit bestimmten Eigenschaften.“ beweisen, so reicht es ***ein Beispiel*** *für ein solches Objekt x zu finden und nachzuweisen, dass dieses die gewünschten Eigenschaften hat.*

Beispiel C.17. (Beweisen von Existenzaussagen durch ein Beispiel)
Um die Aussage

„Es gibt Funktionen $f : \mathbb{R} \to \mathbb{R}$, die sowohl gerade als auch ungerade sind.“

nachzuweisen reicht es, das Beispiel der Nullfunktion anzugeben und nachzuweisen, dass diese die gewünschten Eigenschaften hat.

Beweis: Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 0$. Dann gilt $f(x) = 0 = f(-x)$ für jedes $x \in \mathbb{R}$, d.h. f ist gerade. Weiter gilt auch $f(-x) = 0 = -0 = -f(x)$ für jedes $x \in \mathbb{R}$, d.h. f ist ungerade. ♠

Will man einen Beweis führen, in dem viele Implikationen zu zeigen sind, so kann man natürlich für jede einzelne Implikation eine andere Beweismethode wählen.

C.3 Beweis durch vollständige Induktion

Wir formulieren das Prinzip der vollständigen Induktion für Aussagen, die für alle ganzen Zahlen $n \geq n_0$ mit einem (festen) $n_0 \in \mathbb{Z}$ gelten sollen.

Beweistechnik C.18. (vollständige Induktion – Version I)
Sei $n_0 \in \mathbb{Z}$, und für jedes $n \in \mathbb{Z}$ mit $n \geq n_0$ sei $A(n)$ eine (von n abhängige) Aussage. Angenommen man kann die folgenden beiden Dinge beweisen:

(i) $A(n_0)$ ist wahr.

(ii) Die Implikation „Wenn $A(n)$ wahr ist, dann ist auch $A(n+1)$ wahr.“ ist für jedes $n \in \mathbb{Z}$ mit $n \geq n_0$ wahr.

Dann ist $A(n)$ für alle $n \in \mathbb{Z}$ mit $n \geq n_0$ wahr.

Praktische Umsetzung: In der Praxis geht man bei der Anwendung des Beweisprinzips der vollständigen Induktion wie folgt vor: Nachdem man $A(n)$ und n_0 identifiziert hat, führt man den Beweis in den folgenden zwei Schritten durch:

(i) **Induktionsanfang (IA):** Die Aussage $A(n)$ wird für $n = n_0$ bewiesen (oft durch eine direkte Rechnung).

(ii) **Induktionsschritt (IS):** Für beliebiges $n \geq n_0$ wird unter Benutzung der Aussage $A(n)$ die Aussage $A(n+1)$ bewiesen. $A(n)$ wird dabei als **Induktionsvoraussetzung (IV)** bezeichnet. Die Stelle im Beweis, an der diese eingeht, wird sollte mit „(IV)“ gekennzeichnet sein (um darauf hinzuweisen, dass hier die Induktionsvoraussetzung genutzt wurde).

Für das Induktionsverfahren ist es **unerlässlich**, dass Sie **sowohl den Induktionsanfang als auch den Induktionsschritt** beweisen. Allein sagt keiner dieser Beweisschritte etwas über die Gültigkeit der Aussage $A(n)$ für alle $n \in \mathbb{Z}$ mit $n \geq n_0$ aus.

Betrachten wir zunächst ein einfaches Beispiel, an dem wir das Prinzip der voll-

ständigen Induktion anwenden.

Beispiel C.19. (Zahlen von 1 bis n aufsummieren)

Für alle $n \in \mathbb{N}$ gilt: $$1 + 2 + 3 + \ldots + n = \sum_{k=1}^{n} k = \frac{n\,(n+1)}{2}\,. \tag{C.7}$$

Beweis mit vollständiger Induktion: Wir haben $n_0 = 1$ und die Aussage

$$A(n): \qquad 1 + 2 + 3 + \ldots + n = \sum_{k=1}^{n} k = \frac{n\,(n+1)}{2}$$

Induktionsanfang (IA) $n = 1$: $A(1)$ ist wahr, denn $\displaystyle\sum_{k=1}^{1} k = 1 = \frac{1\,(1+1)}{2}\,.$

Induktionsvoraussetzung (IV): Sei $n \in \mathbb{N}$ fest. Es gelte $A(n)$.

Induktionsschritt (IS) $n \curvearrowright n+1$: Wir müssen zeigen:

$$A(n+1): \quad 1 + 2 + 3 + \ldots + n + (n+1) = \sum_{k=1}^{n+1} k = \frac{(n+1)\,(n+2)}{2} \tag{C.8}$$

Dazu starten wir mit der linken Seite von (C.8) und formen diese unter Ausnutzung der Induktionsvoraussetzung (IV) so lange geeignet um, bis wir die rechte Seite von (C.8) erhalten:

$$1 + 2 + 3 + \ldots + n + (n+1) = \underbrace{\big(1 + 2 + 3 + \ldots + n\big)}_{= \frac{n(n+1)}{2} \text{ nach (IV)}} + (n+1)$$

$$\overset{\text{(IV)}}{=} \frac{n\,(n+1)}{2} + (n+1) = \frac{n\,(n+1) + 2\,(n+1)}{2} = \frac{(n+2)\,(n+1)}{2}\,,$$

oder mit Summenschreibweise

$$\sum_{k=1}^{n+1} k = \underbrace{\sum_{k=1}^{n} k}_{= \frac{n(n+1)}{2} \text{nach (IV)}} + (n+1) \overset{\text{(IV)}}{=} \frac{n\,(n+1)}{2} + (n+1)$$

$$= \frac{n\,(n+1) + 2\,(n+1)}{2} = \frac{(n+2)\,(n+1)}{2}\,.$$

Damit haben wir die Aussage $A(n+1)$ bewiesen.

Nach dem Prinzip der vollständigen Induktion gilt $A(n)$ für alle $n \in \mathbb{N}$. ♠

Bemerkung C.20. (Warum funktioniert das Induktionsprinzip?)

- Wir beweisen, dass $A(n_0)$ wahr ist (Induktionsanfang).
- Dann beweisen wir im Induktionsschritt für beliebiges $n \in \mathbb{N}$, dass aus „$A(n)$ ist wahr." folgt „$A(n+1)$ ist wahr.".
- Mit dem Induktionsschritt können wir für $n = n_0$ aus der Gültigkeit von $A(n_0)$ (Induktionsanfang) die Gültigkeit von $A(n_0 + 1)$ schlussfolgern. Anschließend können wir mit dem Induktionsschritt aus der Gültigkeit von $A(n_0 + 1)$ die Gültigkeit von $A(n_0 + 2)$ schlussfolgern usw.. So erhalten wir die Gültigkeit der Aussage $A(n)$ für alle $n \in \mathbb{Z}$ mit $n \geq n_0$.

Wir formulieren eine zweite Variante des Induktionsprinzips, die natürlich zu der ersten Variante äquivalent ist.

Beweistechnik C.21. (vollständige Induktion – Version II)
Sei $n_0 \in \mathbb{Z}$, und für jedes $n \in \mathbb{Z}$ mit $n \geq n_0$ sei $A(n)$ eine (von n abhängige) Aussage. Angenommen man kann die folgenden beiden Dinge beweisen:

(i) *$A(n_0)$ ist wahr.*

(ii) *Die Implikation „Wenn $A(k)$ für alle $k = n_0, n_0+1, \ldots, n$ wahr ist, dann ist auch $A(n+1)$ wahr." ist für jedes $n \in \mathbb{Z}$ mit $n \geq n_0$ wahr.*

Dann ist $A(n)$ für alle $n \in \mathbb{Z}$ mit $n \geq n_0$ wahr.

Gelegentlich ist diese zweite Variante, der vollständigen Induktion nützlich, weil man im Induktionsschritt die Gültigkeit der Aussage $A(k)$ nicht nur für $k = n$ sondern auch für $k = n - 1$ (und gegebenenfalls weitere $k \leq n$) nutzen möchte.

Bemerkung C.22. (Varianten des Induktionsprinzips)
Alternativ hätten wir auch den folgenden Induktionsschritt (IS) durchführen können, der zum selben Ergebnis führt:

Version I: (ii) Die Implikation „Wenn $A(n-1)$ wahr ist, dann ist auch $A(n)$ wahr." ist für jedes $n \in \mathbb{Z}$ mit $n > n_0$ wahr.

Version II: (ii) Die Implikation „Wenn $A(k)$ für alle $k = n_0, n_0+1, \ldots, n-1$ wahr ist, dann ist auch $A(n)$ wahr." ist für jedes $n \in \mathbb{Z}$ mit $n > n_0$ wahr.

Man beachte in beiden Fällen die **echte** Größerrelation $n > n_0$ **statt** $n \geq n_0$.

Beweise durch vollständige Induktion gehören zu den grundlegenden Techniken. In Lehrbüchern werden Aussagen, die man mit vollständiger Induktion zeigen kann, oft ohne Nachweis oder Begründung verwendet.

Das Prinzip der vollständigen Induktion gibt uns leider keine Hilfsmittel, um gültige Sätze zu formulieren. Um das Induktionsprinzip zu nutzen, müssen Sie bereits wissen, was Sie beweisen wollen!

Betrachten wir noch ein Beispiel.

Beispiel C.23. (Summe ungerader natürlicher Zahlen)

$$\text{Für alle } n \in \mathbb{N} \text{ gilt:} \quad 1+3+5+\ldots+(2n-1) = \sum_{k=1}^{n}(2k-1) = n^2. \tag{C.9}$$

Beweis mit vollständiger Induktion: Wir haben $n_0 = 1$ und die Aussage

$$A(n): \qquad 1+3+5+\ldots+(2n-1) = \sum_{k=1}^{n}(2k-1) = n^2$$

Induktionsanfang (IA) $n = 1$*:* $A(1)$ ist wahr, denn

$$\sum_{k=1}^{1}(2k-1) = 2 \cdot 1 - 1 = 1 = 1^2.$$

Induktionsvoraussetzung (IV): Sei $n \in \mathbb{N}$ fest. Es gelte $A(n)$.

Induktionsschritt (IS) $n \curvearrowright n+1$*:* Wir müssen zeigen:

$$A(n+1): \quad 1+3+\ldots+(2n-1)+(2n+1) = \sum_{k=1}^{n+1}(2k-1) = (n+1)^2, \tag{C.10}$$

wobei wir auf der linken Seite genutzt haben, dass $2(n+1)-1 = 2n+1$ ist.

Wir starten mit der linken Seite in (C.10) und formen diese unter Ausnutzung der Induktionsvoraussetzung um, bis wir die rechte Seite von (C.10) erhalten:

$$1+3+\ldots+(2n-1)+(2n+1) = \underbrace{\big(1+3+\ldots+(2n-1)\big)}_{=n^2 \text{ nach (IV)}} + (2n+1)$$

$$\overset{\text{(IV)}}{=} n^2 + (2n+1) = n^2 + 2n + 1 = (n+1)^2,$$

oder mit Summenschreibweise

$$\sum_{k=1}^{n+1}(2k-1) = \underbrace{\sum_{k=1}^{n}(2k-1)}_{=n^2 \text{ nach (IV)}} + \big(2(n+1)-1\big)$$

$$\overset{\text{(IV)}}{=} n^2 + (2n+1) = n^2 + 2n + 1 = (n+1)^2.$$

Damit haben wir $A(n+1)$ bewiesen.

Nach dem Prinzip der vollständigen Induktion gilt $A(n)$ für alle $n \in \mathbb{N}$. ♠

Bemerkung C.24. (typische Probleme bei Induktionsbeweisen)
Beim Erlernen von Induktionsbeweisen treten oft die folgenden **Probleme** auf, bzw. die folgenden Dinge wurden **nicht** beachtet:

- Anfangs ist es oft ein Problem, herauszufinden, was Sie im Induktionsschritt eigentlich zeigen wollen. Es hilft, sich die im Induktionsschritt zu beweisende Aussage als Erinnerung hinzuschreiben („zu zeigen: ...“).
- Wird die im Induktionsschritt zu zeigende Aussage $A(n+1)$ notiert, erhält man diese aus $A(n)$, indem man in $A(n)$ **überall** n durch $n+1$ ersetzt. (Ersetzt man in $A(n)$ nicht überall sondern nur an einigen Stellen n durch $n+1$, so erhält man nicht $A(n+1)$ sondern eine andere (meistens falsche) Aussage.)
- Beachten Sie, dass Sie im Induktionsschritt (IS) $n \curvearrowright n+1$ **nicht** zeigen müssen, dass die Aussage $A(n)$ für n gilt. Dieses ist die Induktionsvoraussetzung (IV). Es hilft, wann man sich diese gesondert notiert.

Betrachten wir noch ein Beispiel, in dem eine Ungleichung mit vollständiger Induktion bewiesen wird. In diesem Beispiel kommen Fakultäten vor:

n-Fakultät, in Zeichen $n!$, ist für $n \in \mathbb{N}_0$ definiert durch

$$0! = 1; \qquad n! = 1 \cdot 2 \cdot \ldots \cdot n \quad \text{für alle } n \in \mathbb{N}.$$

Beispiel C.25. (Ungleichung)
Für alle $n \in \mathbb{N}$ mit $n \geq 4$ gilt: $n! > 2^n$

Beweis mit vollständiger Induktion: Wir haben $n_0 = 4$ und die Aussage

$$A(n): \qquad n! > 2^n$$

Induktionsanfang (IA) $n = 4$*:* $A(4)$ ist wahr, denn $4! = 24 > 16 = 2^4$.

Induktionsvoraussetzung (IV): Sei $n \in \mathbb{N}$ mit $n \geq 4$ fest. Es gelte $A(n)$.

Induktionsschritt (IS) $n \curvearrowright n+1$: Wir müssen zeigen: $(n+1)! > 2^{n+1}$

Dazu starten wir auf der linken Seite von $(n+1)! > 2^{n+1}$ und formen um bzw. schätzen nach unten ab, bis wir die rechte Seite von $(n+1)! > 2^{n+1}$ erhalten:

$$(n+1)! = \underbrace{1 \cdot 2 \cdot \ldots n}_{=n!} \cdot (n+1) = \underbrace{n!}_{>2^n} (n+1) \overset{\text{(IV)}}{>} 2^n \underbrace{(n+1)}_{>2 \text{ weil } n \geq 4} > 2^n \cdot 2 = 2^{n+1}.$$

Damit haben wir $A(n+1)$ bewiesen.

Nach dem Prinzip der vollständigen Induktion gilt die Aussage $A(n)$ für alle $n \in \mathbb{N}$ mit $n \geq 4$. ♠

Als Letztes beweisen wir eine Teilbarkeitsaussage mit vollständiger Induktion.

Beispiel C.26. (Teilbarkeitsaussage)
Für jedes $n \in \mathbb{N}_0$ ist $n^3 + 3\,n^2 + 2\,n$ durch 6 teilbar.

Beweis mit vollständiger Induktion: Wir haben $n_0 = 0$ und die Aussage

$$A(n): \qquad n^3 + 3\,n^2 + 2\,n \text{ ist durch 6 teilbar.}$$

Induktionsanfang (IA) $n = 0$*:* $A(0)$ ist wahr, denn $0^3 + 3 \cdot 0^2 + 2 \cdot 0 = 0$ ist durch 6 teilbar.

Induktionsvoraussetzung (IV): Sei $n \in \mathbb{N}_0$ mit $n \geq 0$ fest. Es gelte $A(n)$.

Induktionsschritt (IS) $n \curvearrowright n+1$:

Wir müssen zeigen: $(n+1)^3 + 3\,(n+1)^2 + 2\,(n+1)$ ist durch 6 teilbar.

Dazu formen wir $(n+1)^3 + 3\,(n+1)^2 + 2\,(n+1)$ geeignet um und nutzen die Induktionsvoraussetzung aus:

$$\begin{aligned}
&(n+1)^3 + 3\,(n+1)^2 + 2\,(n+1) \\
&\quad = (n^3 + 3\,n^2 + 3\,n + 1) + 3\,(n^2 + 2\,n + 1) + 2\,(n+1) \\
&\quad = (n^3 + 3\,n^2 + 2\,n) + 3\,n^2 + 9\,n + 6 \\
&\quad = (n^3 + 3\,n^2 + 2\,n) + 3\,(n^2 + 3\,n + 2) \\
&\quad = (n^3 + 3\,n^2 + 2\,n) + 3\,(n+1)\,(n+2),
\end{aligned}$$

wobei man die Faktorisierung $n^2 + 3\,n + 2 = (n + 1)\,(n + 2)$ mit dem Satz von Viéta direkt ablesen kann oder diese alternativ mit quadratischer Ergänzung und den binomischen Formeln berechnen kann:

$$n^2 + 3\,n + 2 = n^2 + 3\,n + \frac{9}{4} - \frac{1}{4} = \left(n + \frac{3}{2}\right)^2 - \left(\frac{1}{2}\right)^2$$
$$= \left(n + \frac{3}{2} - \frac{1}{2}\right)\left(n + \frac{3}{2} + \frac{1}{2}\right) = (n + 1)\,(n + 2).$$

Wir haben also

$$(n + 1)^3 + 3\,(n + 1)^2 + 2\,(n + 1) = (n^3 + 3\,n^2 + 2\,n) + 3\,(n + 1)\,(n + 2).$$

Nach der Induktionsvoraussetzung (IV) ist $n^3 + 3\,n^2 + 2\,n$ durch 6 teilbar.

$3\,(n+1)(n+2)$ ist durch 3 teilbar. Weil $n+1$ und $n+2$ zwei aufeinanderfolgende ganze Zahlen sind, muss eine der beiden Zahlen gerade und somit durch 2 teilbar sein. Also ist $3\,(n+1)(n+2)$ auch durch 2 teilbar. Daraus folgt, dass $3\,(n+1)(n+2)$ durch 2 und 3 und somit durch 6 teilbar ist.

Als Summe zweier durch 6 teilbarer Zahlen ist $(n + 1)^3 + 3\,(n + 1)^2 + 2\,(n + 1)$ durch 6 teilbar. Damit haben wir $A(n + 1)$ bewiesen.

Nach dem Prinzip der vollständigen Induktion gilt $A(n)$ für alle $n \in \mathbb{N}_0$. ♠