Analysis

— Sommersemester 2014 —

Prof. Dr. Karl Stroetmann

26. August 2014

Dieses Skript ist einschließlich der LaTeX-Quellen sowie der in diesem Skript diskutierten Programme unter

https://github.com/karlstroetmann/Analysis

im Netz verfügbar. Das Skript wird laufend überarbeitet. Wenn Sie auf Ihrem Rechner `git` installieren und mein Repository mit Hilfe des Befehls

```
git clone https://github.com/karlstroetmann/Analysis.git
```

klonen, dann können Sie durch Absetzen des Befehls

```
git pull
```

die aktuelle Version meines Skripts aus dem Netz laden.

Inhaltsverzeichnis

1 Einleitung **3**
- 1.1 Überblick über die Vorlesung . . . 3
- 1.2 Ziel der Vorlesung . . . 4
- 1.3 Notation . . . 4
- 1.4 Eine Bitte . . . 4

2 Die reellen Zahlen **5**
- 2.1 Axiomatische Charakterisierung der reellen Zahlen . . . 6
- 2.2 Die formale Konstruktion der reellen Zahlen* . . . 9
- 2.3 Geschichte . . . 17

3 Folgen und Reihen **18**
- 3.1 Folgen . . . 18
- 3.2 Berechnung der Quadrat-Wurzel . . . 29
- 3.3 Reihen . . . 33
- 3.4 Potenz-Reihen . . . 42

4 Stetige und differenzierbare Funktionen **46**
- 4.1 Stetige Funktionen . . . 46
- 4.2 Bestimmung von Nullstellen . . . 50
  - 4.2.1 Die Regula Falsi . . . 54
  - 4.2.2 Das Sekanten-Verfahren . . . 59
  - 4.2.3 Das Illinois-Verfahren . . . 61

5 Differenzierbare Funktionen **63**
- 5.1 Der Begriff der Ableitung . . . 63
- 5.2 Mittelwert-Sätze . . . 71
- 5.3 Monotonie und Konvexität . . . 76
- 5.4 Die Exponential-Funktion . . . 82

6 Anwendungen der Theorie **85**
- 6.1 Taylor-Reihen . . . 85
  - 6.1.1 Der Abbruch-Fehler bei der Taylor-Reihe . . . 87
- 6.2 Beispiele von Taylor-Entwicklungen . . . 89
  - 6.2.1 Berechnung des natürlichen Logarithmus . . . 89
  - 6.2.2 Berechnung des Arcus-Tangens . . . 92
  - 6.2.3 Berechnung von π* . . . 93
- 6.3 Polynom-Interpolation . . . 96
  - 6.3.1 Interpolation nach Newton* . . . 98
  - 6.3.2 Der Interpolations-Fehler . . . 99
- 6.4 Der Banach'sche Fixpunkt-Satz . . . 101
  - 6.4.1 Beschleunigung der Fixpunkt-Iteration . . . 106

6.4.2 Das Newton'sche Verfahren zur Berechnung von Nullstellen 108
6.4.3 Analyse des Newton'schen Verfahrens 110
6.5 Iterative Lösung linearer Gleichungs-Systeme* 113

7 Integral-Rechnung **120**
7.1 Einführung des Integral-Begriffs 120
7.2 Regeln zur Berechnung von Integralen 126
7.2.1 Die Substitutions-Regel 127
7.2.2 Partielle Integration 129
7.2.3 Das Integral von Umkehr-Funktionen* 130
7.2.4 Berechnung der Fläche eines Kreises 131
7.3 Berechnung der Bogenlänge 132
7.4 Uneigentliche Integrale 134
7.5 Numerische Integration* 136
7.5.1 Die Trapez-Regel 136
7.5.2 Die Simpson'sche Regel 139

8 Die Zahlen π und e sind irrational* **142**
8.1 Die Euler'sche Zahl e ist irrational 142
8.2 Die Kreiszahl π ist irrational 144
8.3 Tranzendente Zahlen 147

9 Fourier-Analyse* **150**
9.1 Berechnung der Fourier-Koeffizienten 150
9.2 Konvergenz 154
9.3 Beispiele 154
9.3.1 Fourier-Analyse der Sägezahn-Funktion 156

10 Rundungsfehler* **159**

Kapitel 1

Einleitung

Der vorliegende Text ist ein Fragment eines Vorlesungs-Skriptes für die Analysis-Vorlesung für Informatiker. Ich habe mich bei der Ausarbeitung dieser Vorlesung im wesentlichen auf die folgenden Quellen gestützt:

1. *Analysis* I von Otto Forster [1].
2. *Differential- und Integralrechnung I* von Hans Grauert und Ingo Lieb [2].
3. *Differential and Integral Calculus, Volume 1* von Richard Courant [3].
4. *Advanced Calculus* von Richard Wrede und Murray R. Spiegel [4].

Den Studenten empfehle ich das erste Buch in dieser Liste, denn dieses Buch ist auch in elektronischer Form in unserer Bibliothek vorhanden. Bei dem Buch von Richard Courant ist das Copyright mittlerweile abgelaufen, so dass Sie es im Netz unter

https://ia700700.us.archive.org/34/items/DifferentialIntegralCalculusVolI/Courant-DifferentialIntegralCalculusVolI.pdf

finden können. Schließlich enthält das Buch von Wrede und Spiegel eine Vielzahl gelöster Aufgaben und bietet sich daher besonders zum Üben an.

1.1 Überblick über die Vorlesung

Im Rahmen der Vorlesung werden die folgenden Gebiete behandelt:

1. Im zweiten Kapitel werden die reellen Zahlen mit Hilfe von Dedekind-Schnitten definiert.
2. Das dritte Kapitel führt den Begriff des Grenzwerts für Folgen und Reihen ein.
3. Das vierte Kapitel diskutiert die Begriffe Stetigkeit und Differenzierbarkeit.
4. Das fünfte Kapitel zeigt verschiedene Anwendungen der bis dahin dargestellten Theorie. Insbesondere werden *Taylor-Reihen* diskutiert. Diese können beispielsweise zur Berechnung der trigonometrischen Funktionen verwendet werden. Außerdem diskutieren wir in diesem Kapitel Verfahren zur numerischen Lösung von Gleichungen.
5. Das sechste Kapitel beschäftigt sich mit der Integralrechnung.
6. Im siebten Kapitel zeigen wir, dass π und e keine rationalen Zahlen sind.
7. Im letzten Kapitel diskutieren wir Fourier-Reihen.

1.2 Ziel der Vorlesung

Wir werden im Rahmen der Vorlesung nicht die Zeit haben, alle Aspekte der Analysis zu besprechen. Insbesondere werden wir viele interessante Anwendungen der Analysis in der Informatik nicht diskutieren können. Das ist aber auch gar nicht das Ziel dieser Vorlesung: Mir geht es vor allem darum, Ihnen die Fähigkeit zu vermitteln, sich selbstständig in mathematische Fachliteratur hineinarbeiten zu können. Dazu müssen Sie in der Lage sein, mathematische Beweise sowohl zu verstehen als auch selber entwickeln zu können. Dies ist ein wesentlicher Unterschied zu der Mathematik, an die sich viele von Ihnen auf der Schule gewöhnthaben: Dort werden primär Verfahren vermitteln, mit denen sich spezielle Probleme lösen lassen. Die Kenntnis solcher Verfahren ist allerdings in der Praxis nicht mehr wichtig, denn heutzutage werden solche Verfahren programmiert und daher besteht kein Bedarf mehr dafür, solche Verfahren von Hand anzuwenden. Aus diesem Grund wird in dieser Vorlesung der mathematische Beweis-Begriff im Vordergrund stehen. Die Analysis dient uns dabei als ein Beispiel einer mathematischen Theorie, an Hand derer wir das mathematische Denken üben können.

1.3 Notation

In diesem Skript definieren wir die Menge der natürlichen Zahlen $\mathbb{N}$ über die Formel

$$\mathbb{N} := \{1, 2, 3, \cdots\}.$$

Im Gegensatz zu der Vorlesung über Lineare Algebra im letzten Semester wird die Zahl 0 in diesem Skript also nicht als natürliche Zahl aufgefasst. Weiter definieren wir

$$\mathbb{N}_0 := \{0\} \cup \mathbb{N}.$$

1.4 Eine Bitte

Dieses Skript enthält noch eine Menge Tipp-Fehler. Sollte Ihnen ein Fehler auffallen, so bitte ich um einen Hinweis unter der Adresse

karl.stroetmann@dhbw-mannheim.de.

Wenn Sie mit `github` vertraut sind, können Sie mir auch gerne einen *Pull Request* schicken.

Kapitel 2

Die reellen Zahlen

Bevor wir mit der eigentlichen Analysis beginnen müssen wir klären, was genau reelle Zahlen überhaupt sind. Anschaulich werden reelle Zahlen zur Angabe von Längen benötigt, denn in der Geometrie reicht es nicht, mit den rationalen Zahlen zu arbeiten. Das liegt daran, dass die Diagonale eines Quadrats der Seitenlänge 1 nach dem Satz des Pythagoras die Länge $\sqrt{2}$ hat. Wir hatten im ersten Semester aber gesehen, dass es keine rationale Zahl r gibt, so dass $r^2 = 2$ ist. Folglich reichen die rationalen Zahlen nicht aus, alle in der Geometrie möglichen Längen anzugeben. Es gibt mehrere Wege, die Menge $\mathbb{R}$ der reellen Zahlen so zu konstruieren, so dass die Gleichung

$$r^2 = 2$$

in $\mathbb{R}$ eine Lösung hat. Bevor wir mit dieser Konstruktion beginnen, wollen wir die Menge der reellen Zahlen axiomatisch charakterisieren. Dazu definieren wir den Begriff des

vollständig geordneten Körpers.

Hierbei handelt es sich um ein System von Axiomen, aus dem sich alle weiteren Eigenschaften der reellen Zahlen ableiten lassen. Es lässt sich sogar zeigen, dass diese Axiomatisierung in dem folgenden Sinne vollständig ist: Ist $\mathbb{K}$ ein vollständig angeordneter Körper, so ist $\mathbb{K}$ *isomorph* zu $\mathbb{R}$: Im Klartext heißt dies, dass es eine Abbildung

$$\varphi : \mathbb{K} \to \mathbb{R}$$

gibt, die jedem Element $\alpha \in \mathbb{K}$ genau ein Element $\varphi(\alpha) \in \mathbb{R}$ zuordnet und die mit der Addition und der Multiplikation verträglich ist, für alle $x, y \in \mathbb{K}$ gilt also

$$\varphi(x + y) = \varphi(x) + \varphi(y) \quad \text{und} \quad \varphi(x \cdot y) = \varphi(x) \cdot \varphi(y).$$

Die rein axiomatische Charakterisierung der reellen Zahlen ist philosophisch unbefriedigend, denn dabei bleiben zwei Fragen offen:

1. Gibt es überhaupt eine Struktur $\mathbb{K}$, die den Axiomen eines vollständig geordneten Körpers genügt?

   Solange wir nicht sicher sind, dass diese Frage mit "Ja" beantwortet wird, steht unsere gesamte Theorie auf wackeligen Füßen, denn es könnte dann sein, dass es sich um die Theorie der leeren Menge handelt.

2. Was genau sind reelle Zahlen?

   Diese Frage hat zwar zunächst einen philosophischen Charakter, aber es zeigt sich, dass wir diese Frage zuerst beantworten müssen, bevor wir die erste Frage in Angriff nehmen können.

Wir werden im zweiten Abschnitt zeigen, wie sich die reellen Zahlen aus den rationalen Zahlen mit Hilfe von sogenannten *Dedekind-Schnitten* erzeugen lassen. Da diese Konstruktion jedoch technisch nicht ganz einfach ist, werden wir diesen Abschnitt im Rahmen der Vorlesung nicht besprechen. Die

in diesem Abschnitt präsentierten Details sind zwar eine gute Übung zur Mengenlehre, werden aber für den weiteren Verlauf der Vorlesung nicht benötigt. Der Abschnitt ist daher zum Selbststudium für die Studenten gedacht, die an den Grundlagen der Mathematik interessiert sind.

2.1 Axiomatische Charakterisierung der reellen Zahlen

Wir erinnern zunächst an die im ersten Semester gegebene Definition eines Körpers.

Definition 1 (Körper) Eine Struktur $\mathcal{K} = \langle K, 0, 1, +, \cdot \rangle$ ist ein *Körper*, falls gilt:

1. K ist eine Menge.
2. $0 \in K$ und $1 \in K$.
3. $+$ und $\cdot$ sind binäre Operatoren auf K, wir haben
$$+ : K \times K \to K \quad \text{und} \quad \cdot : K \times K \to K.$$
4. $\langle K, 0, + \rangle$ ist eine kommutative Gruppe.
Im Detail heißt dies, dass die folgenden Axiome gelten:
   (a) $(x + y) + z = x + (y + z)$,
   (b) $x + y = y + x$,
   (c) $0 + x = x$,
   (d) $\exists y \in K : x + y = 0$.
   Dasjenige $y \in K$, für welches $x + y = 0$ gilt, ist dann eindeutig bestimmt und wird mit $-x$ bezeichnet.
5. $\langle K \backslash \{0\}, 1, \cdot \rangle$ ist ebenfalls eine kommutative Gruppe,
Im Detail sind also die folgenden Axiome erfüllt:
   (a) $(x \cdot y) \cdot z = x \cdot (y \cdot z)$,
   (b) $x \cdot y = y \cdot x$,
   (c) $1 \cdot x = x$,
   (d) $x \neq 0 \Rightarrow \exists y \in K : x \cdot y = 1$.
6. Es gilt das Distributiv-Gesetz: Für alle $x, y, z \in K$ haben wir
$$x \cdot (y + z) = x \cdot y + x \cdot z.$$ ◇

Weiter benötigen wir die Definition einer linearen Ordnung, die wir ebenfalls wiederholen.

Definition 2 (Lineare Ordnung) Ein Paar $\langle M, \leq \rangle$ ist eine *lineare Ordnung*, falls gilt:

1. M ist eine Menge,
2. $\leq$ ist eine binäre Relation auf M, es gilt also
$$\leq \subseteq M \times M.$$
3. Zusätzlich gelten die folgenden Axiome:
   (a) $x \leq x$,
   (b) $x \leq y \wedge y \leq x \to x = y$,
   (c) $x \leq y \wedge y \leq z \to x \leq z$,
   (d) $x \leq y \vee y \leq x$. ◇

Die nächste Definition kombiniert die algebraischen Eigenschaften eines Körpers mit den Anordnungs-Axiomen einer linearen Ordnung.

Definition 3 (Geordneter Körper)
Eine Struktur $\mathcal{K} = \langle K, 0, 1, +, \cdot, \leq \rangle$ ist ein *geordneter Körper*, falls

1. $\langle K, 0, 1, +, \cdot, \rangle$ ein Körper und
2. $\langle K, \leq \rangle$

eine lineare Ordnung ist. ◇

Bemerkung: Die Struktur $\langle \mathbb{Q}, 0, 1, +, \cdot, \leq \rangle$ ist ein geordneter Körper, wenn wir die Operationen $+$, $\cdot$ und $\leq$ wie im ersten Semester vorgeführt auf den rationalen Zahlen definieren. Diese Struktur ist allerdings für die Zwecke der Analysis noch nicht ausreichend, da sie in einer noch näher zu spezifizierenden Weise unvollständig ist. Zur Präzisierung dieser Aussage benötigen wir den Begriff einer *vollständigen Ordnung*, die ihrerseits auf dem Begriff des *Supremums* basiert, den wir jetzt einführen.

Definition 4 (Supremum) Es sei $\langle M, \leq \rangle$ eine lineare Ordnung. Eine Menge $A \subseteq M$ ist *nach oben beschränkt*, falls es ein $y \in M$ gibt, so dass gilt:

$$\forall x \in A : x \leq y.$$

Dieses y bezeichnen wir dann als eine *obere Schranke* der Menge A. Ein Element $s \in M$ ist das *Supremum* der Menge A, wenn s die kleinste obere Schranke von A ist, wenn also

$$\forall x \in A : x \leq s \quad \text{und} \quad \forall y \in M : \Big(\big(\forall x \in A : x \leq y \big) \Rightarrow s \leq y \Big)$$

gilt. In diesem Fall schreiben wir

$$s = \sup(A).$$ ◇

Definition 5 (Vollständige Ordnung) Ein Paar $\langle M, \leq \rangle$ bestehend aus einer Menge M und einer Relation $\leq \subseteq M \times M$ ist eine *vollständige* Ordnung genau denn, wenn folgendes gilt:

1. Das Paar $\langle M, \leq \rangle$ ist eine lineare Ordnung.
2. Zu jeder nicht-leeren und nach oben beschränkten Menge $A \subseteq M$ existiert ein Supremum in M.

◇

Bemerkung: Das Paar $\langle \mathbb{Q}, \leq \rangle$ ist keine vollständige Ordnung, denn wenn wir die Menge A als

$$A := \{ r \in \mathbb{Q} \mid r^2 \leq 2 \}$$

definieren, so ist die Menge A nach oben beschränkt, weil die Zahl 2 eine obere Schranke von A ist. Die Menge A hat aber keine kleinste obere Schranke. Anschaulich liegt das daran, dass die kleinste obere Schranke von A die Zahl $\sqrt{2}$ ist, aber aus dem ersten Semester wissen wir, dass $\sqrt{2}$ keine rationale Zahl ist. ◇

Analog zum Begriff des Supremums können wir auch den Begriff des Infimums definieren.

Definition 6 (Infimum) Es sei $\langle M, \leq \rangle$ eine lineare Ordnung. Eine Menge $B \subseteq M$ ist *nach unten beschränkt*, falls es ein $y \in M$ gibt, so dass

$$\forall x \in B : y \leq x$$

gilt. Ein solches y bezeichnen wir als *untere Schranke* von B. Ein Element $i \in M$ ist das *Infimum* einer Menge B, wenn i die größte untere Schranke von B ist, wenn also

$$\forall x \in B : i \leq x \quad \text{und} \quad \forall y \in M : \Big(\big(\forall x \in B : y \leq x \big) \rightarrow y \leq i \Big)$$

gilt. In diesem Fall schreiben wir

$$i = \inf(B).$$ ◇

Aufgabe 1: Es sei $\langle M, \leq \rangle$ eine vollständige lineare Ordnung. Zeigen Sie, dass für jede Teilmenge $B \subseteq M$, die nicht-leer und nach unten beschränkt ist, ein Infimum existiert.

Hinweis: Betrachten Sie die Menge

$$A := \{x \in M \mid \forall y \in B : x \leq y\}$$

der unteren Schranken von B. Diese Menge ist nicht leer, denn nach Voraussetzung ist B nach unten beschränkt. Außerdem ist A nach oben beschränkt, denn jedes Element aus der Menge B ist eine obere Schranke von A. Da nach Voraussetzung das Paar $\langle M, \leq \rangle$ eine vollständige lineare Ordnung ist, besitzt A also ein Supremum. Zeigen Sie, dass dieses Supremum auch das Infimum von B ist. ◇

Wir kommen nun zur zentralen Definition dieses Abschnitts.

Definition 7 (Vollständig angeordneter Körper) Eine Struktur $\mathcal{K} = \langle K, 0, 1, +, \cdot, \leq \rangle$ ist ein *vollständig angeordneter Körper* genau dann, wenn die Struktur $\mathcal{K}$ ein geordneter Körper ist und zusätzlich das Paar $\langle K, \leq \rangle$ eine vollständige Ordnung ist. ◇

Theorem 8 Die Struktur $\langle \mathbb{R}, 0, 1, +, \cdot, \leq \rangle$ der reellen Zahlen ist ein vollständig geordneter Körper. ◇

Einen Beweis diese Behauptung können wir jetzt noch nicht erbringen, denn wir haben bisher nicht definiert, wie die Menge $\mathbb{R}$ der reellen Zahlen konstruiert wird. Diese Konstruktion werden wir im nächsten Abschnitt mit Hilfe der sogenannten *Dedekind-Schnitte* liefern. Für praktische Rechnungen ist diese viel zu schwerfällig. Wir geben daher zum Abschluss dieses Abschnitts noch eine Konstruktion der reellen Zahlen als unendliche Dezimalbrüche an. Dazu betrachten wir eine positive reelle Zahl $x \in \mathbb{R}$ abstrakt als gegeben und untersuchen, wie wir x in der Form

$$x = m + \sum_{k=1}^{\infty} b_k \cdot \frac{1}{10^k}$$

so darstellen können, dass folgendes gilt:

1. $m \in \mathbb{N}$ und
2. $b_k \in \{0, \cdots, 9\}$.

Die Zahl b_k kann damit als die k-te Stelle von x hinter dem Komma interpretiert werden. Die Berechnung der Zahlen m und b_k verläuft für ein gegebenes positives $x \in \mathbb{R}$ wie folgt:

1. $m := \max(\{n \in \mathbb{N} \mid n \leq x\})$.

   m ist also die größte ganze Zahl, die noch kleiner oder gleich x ist.

2. Die Zahlen b_k werden für alle $k \in \mathbb{N}$ induktiv definiert.

   I.A.: $k = 1$. Wir setzen

   $$b_1 := \max\Big(\Big\{z \in \{0, 1, \cdots, 9\} \mid m + z \cdot \frac{1}{10} \leq x\Big\}\Big).$$

   I.A.: $k \mapsto k + 1$. Nach Induktions-Voraussetzung sind die Zahlen b_1, $\cdots$, b_k bereits definiert. Daher können wir b_{k+1} als

   $$b_{k+1} := \max\Big(\Big\{z \in \{0, 1, \cdots, 9\} \mid m + \sum_{i=1}^{k} b_i \cdot \frac{1}{10^i} + z \cdot \frac{1}{10^{k+1}} \leq x\Big\}\Big).$$

   definieren.

Insgesamt gilt mit dieser Konstruktion

$$x = m + \sum_{k=1}^{\infty} b_k \cdot \frac{1}{10^k}.$$

Diese Behauptung können wir allerdings noch nicht beweisen, denn wir haben ja noch gar nicht formal definiert, welchen Wert wir einer unendlichen Reihe der Form

$$\sum_{k=1}^{\infty} b_k \cdot \frac{1}{10^k}$$

zuordnen. Statt dessen können wir sagen, dass

$$x = \sup\Big(\Big\{m + \sum_{i=1}^{k} b_i \cdot \frac{1}{10^i} \mid k \in \mathbb{N}\Big\}\Big)$$

gilt, denn die Menge

$$\Big\{m + \sum_{i=1}^{k} b_i \cdot \frac{1}{10^i} \mid k \in \mathbb{N}\Big\}$$

ist offenbar durch x nach oben beschränkt und es lässt sich zeigen, dass es keine obere Schranke y für diese Menge gibt, die echt kleiner als x ist. Für den Rest der Vorlesung reicht es aus, wenn Sie sich die reellen Zahlen wie oben skizziert als unendliche Dezimalbrüche vorstellen.

2.2 Die formale Konstruktion der reellen Zahlen*

Bis jetzt haben wir so getan, als wüssten wir schon, was reelle Zahlen sind und haben Ihre Eigenschaften in Form von Axiomen angegeben. Was noch fehlt ist nun die Konstruktion der reellen Zahlen mit Hilfe der Mengenlehre.

Die zentrale Idee bei der Konstruktion der reellen Zahlen ist die Beobachtung, dass eine reelle Zahl x durch die Menge M_1 aller rationalen Zahlen kleiner als x und die Menge M_2 der rationalen Zahlen größer-gleich x bereits vollständig festgelegt wird. Definieren wir für eine reelle Zahl x

$$M_1 := \{q \in \mathbb{Q} \mid q < x\} \quad \text{und} \quad M_2 := \{q \in \mathbb{Q} \mid x \leq q\},$$

so liegt x gerade zwischen M_1 und M_2. Falls die Menge M_2 kein Minimum hat, so haben die rationalen Zahlen zwischen M_1 und M_2 eine Lücke. Die Idee ist nun, die reellen Zahlen gerade als diese Lücken zu definieren um dadurch sicherzustellen, dass es bei den reellen Zahlen selber keine Lücken mehr gibt. Versuchen wir den Begriff eines *Lücke* zu präzisieren, so kommen wir zur nun folgende Definition eines Dedekind'schen-Schnittes.

Definition 9 (Dedekind-Schnitt)
Ein Paar $\langle M_1, M_2\rangle$ heißt *Dedekind-Schnitt* (Richard Dedekind, 1831-1916) falls folgendes gilt:

1. $M_1 \subseteq \mathbb{Q}, \quad M_2 \subseteq \mathbb{Q}$.
2. $M_1 \neq \emptyset, \quad M_2 \neq \emptyset$.
3. $\forall x_1 \in M_1 : \forall x_2 \in M_2 : x_1 < x_2$.

   Diese Bedingung besagt, dass alle Elemente aus M_1 kleiner als alle Elemente aus M_2 sind. Diese Bedingung bezeichnen wir als die *Trennungs-Eigenschaft*.
4. $M_1 \cup M_2 = \mathbb{Q}$.
5. M_1 hat kein Maximum.

   Da alle Elemente aus M_1 kleiner als alle Elemente von M_2 sind und da darüber hinaus $M_2 \neq \emptyset$ ist, ist M_1 sicher nach oben beschränkt. Aber wenn für ein y

$$\forall x \in M_1 : x \leq y$$

gilt, dann darf y eben kein Element von M_1 sein. Als Formel schreibt sich das als

$$\forall y \in Q : \big((\forall x \in M_1 : x \leq y) \rightarrow y \notin M_1\big).$$ ◇

Beispiel: Definieren wir

$$M_1 := \{x \in \mathbb{Q} \mid x \leq 0 \vee x^2 \leq 2\} \quad \text{und} \quad M_2 := \{x \in \mathbb{Q} \mid x > 0 \wedge x^2 > 2\},$$

so enthält M_1 alle die Zahlen, die kleiner oder gleich $\sqrt{2}$ sind, während M_2 alle Zahlen enthält, die größer als $\sqrt{2}$ sind. Das Paar $\langle M_1, M_2 \rangle$ ist dann ein Dedekind-Schnitt. Intuitiv spezifiziert dieser Dedekind-Schnitt die Zahl $\sqrt{2}$. ◇

Das Beispiel legt nahe, die Menge der reellen Zahlen formal als die Menge aller Dedekind-Schnitte zu definieren

$$\boxed{\mathbb{R} := \big\{\langle M_1, M_2 \rangle \in 2^{\mathbb{Q}} \times 2^{\mathbb{Q}} \mid \langle M_1, M_2 \rangle \text{ ist ein Dedekind-Schnitt}\big\}.}$$

Nach dieser Definition müssen wir nun zeigen, wie sich auf der so definierten Menge der reellen Zahlen die arithmetischen Operationen Addition, Subtraktion, Multiplikation und Division definieren lassen und wie die Relation $\leq$ für zwei Dedekind-Schnitte festgelegt werden kann. Zusätzlich müssen wir nachweisen, dass wir mit diesen Definitionen einen vollständig angeordneten Körper konstruieren.

Bei einem Dedekind-Schnitt $\langle M_1, M_2 \rangle$ ist die Menge M_2 durch die Angabe von M_1 bereits vollständig festgelegt, denn aus der Gleichung $M_1 \cup M_2 = \mathbb{Q}$ folgt sofort $M_2 = \mathbb{Q} \backslash M_1$. Die Frage ist nun, welche Eigenschaften eine Menge M haben muss, damit umgekehrt das Paar $\langle M, \mathbb{Q} \backslash M \rangle$ ein Dedekind-Schnitt ist. Die Antwort auf diese Frage wird in der nun folgenden Definition einer *Dedekind-Menge* gegeben.

Definition 10 (Dedekind-Menge) Eine Menge $M \subseteq \mathbb{Q}$ ist eine *Dedekind-Menge* genau dann, wenn die folgenden Bedingungen erfüllt sind.

1. $M \neq \emptyset$,
2. $M \neq \mathbb{Q}$,
3. $\forall x, y \in \mathbb{Q} : \big(y < x \wedge x \in M \rightarrow y \in M\big)$.

   Die letzte Bedingung besagt, dass M *nach unten abgeschlossen* ist: Wenn eine Zahl x in M liegt, dann liegt auch jede Zahl, die kleiner als x ist, in M.
4. Die Menge M hat kein Maximum, es gibt also kein $m \in M$, so dass

   $$x \leq m \quad \text{für alle } x \in M \text{ gilt.}$$

   Diese Bedingung können wir auch etwas anders formulieren: Wenn $x \in M$ ist, dann finden wir immer ein $y \in M$, dass noch größer als x ist, denn sonst wäre x ja das Maximum von M. Formal können wir das als

   $$\forall x \in M : \exists y \in M : x < y$$

   schreiben.

Aufgabe 2: Zeigen Sie, dass eine Menge $M \subseteq \mathbb{Q}$ genau dann eine Dedekind-Menge ist, wenn das Paar $\langle M, \mathbb{Q} \backslash M \rangle$ ein Dedekind-Schnitt ist. ◇

Lösung: Da es sich bei der zu beweisenden Aussage um eine Äquivalenz-Aussage handelt, zerfällt der Beweis in zwei Teile.

"⇒": Zunächst nehmen wir an, dass $M \subseteq \mathbb{Q}$ eine Dedekind-Menge ist. Wir haben zu zeigen, dass dann $\langle M, \mathbb{Q}\backslash M\rangle$ ein Dedekind-Schnitt ist. Von den zu überprüfenden Eigenschaften ist nur die Trennungs-Eigenschaft nicht offensichtlich. Sei als $x \in M$ und $y \in \mathbb{Q}\backslash M$. Wir haben zu zeigen, dass dann

$$x < y$$

gilt. Wir führen diesen Nachweis indirekt und nehmen an, dass $y \leq x$. Da M nach unten abgeschlossen ist, folgt daraus aber $y \in M$, was im Widerspruch zu $y \in \mathbb{Q}\backslash M$ steht. Dieser Widerspruch zeigt, dass $x < y$ ist und das war zum Nachweis der Trennungs-Eigenschaft zu zeigen.

"⇐": Nun nehmen wir an, dass $\langle M, \mathbb{Q}\backslash M\rangle$ ein Dedekind-Schnitt ist und zeigen, dass dann M eine Dedekind-Menge sein muss. Von den zu überprüfenden Eigenschaften ist nur Tatsache, dass M nach unten abgeschlossen ist, nicht offensichtlich. Sei also $x \in M$ und $y < x$. Wir haben zu zeigen, dass dann y ebenfalls ein Element von M ist. Wir führen diesen Nachweis indirekt und nehmen $y \in \mathbb{Q}\backslash M$ an. Aufgrund der Trennungs-Eigenschaft des Dedekind-Schnitts $\langle M, \mathbb{Q}\backslash M\rangle$ muss dann

$$x < y$$

gelten, was im Widerspruch zu $y < x$ steht. Dieser Widerspruch zeigt, dass $y \in M$ gilt und das war zu zeigen. □

Die letzte Aufgabe hat gezeigt, dass Dedekind-Schnitte und Dedekind-Mengen zu einander äquivalent sind. Daher werden wir im Folgenden mit Dedekind-Mengen arbeiten, denn das macht die Notation einfacher. Wir definieren dazu $\mathcal{D}$ als die Menge aller rationalen Dedekind-Mengen, wir setzen also

$$\mathcal{D} := \big\{M \in 2^{\mathbb{Q}} \mid M \text{ is Dedekind-Menge}\big\}.$$

Wir werden später die Menge $\mathbb{R}$ der reellen Zahlen als diese Menge $\mathcal{D}$ definieren. Die nächste Aufgabe zeigt, wie wir auf der Menge $\mathcal{D}$ eine lineare Ordnung erzielen können.

Aufgabe 3: Auf der Menge $\mathcal{D}$ definieren wir eine binäre Relation $\leq$ durch die Festsetzung

$$A \leq B \stackrel{\text{def}}{\iff} A \subseteq B \quad \text{für alle } A, B \in \mathcal{D}.$$

Zeigen Sie, dass die so definierte Relation $\leq$ eine lineare Ordnung auf der Menge $\mathcal{D}$ ist.

Lösung: Es ist zu zeigen, dass die Relation $\leq$ reflexiv, anti-symmetrisch und transitiv ist und dass außerdem die Linearitäts-Eigenschaft

$$A \leq B \vee B \leq A \quad \text{für alle Dedekind-Mengen } A, B \in \mathcal{D}$$

gilt. Die Reflexivität, Anti-Symmetrie und Transitivität der Relation $\leq$ folgen sofort aus der Reflexivität, Anti-Symmetrie und Transitivität der Teilmengen-Relation $\subseteq$. Es bleibt, den Nachweis der Linearitäts-Eigenschaft zu führen. Seien also $A, B \in \mathcal{D}$ gegeben. Falls $A = B$ ist, gilt sowohl $A \subseteq B$ als auch $B \subseteq A$, woraus sofort $A \leq B$ und $B \leq B$ folgt. Wir nehmen daher an, dass $A \neq B$ ist. Dann gibt es zwei Möglichkeiten:

1. Fall: Es existiert ein $x \in \mathbb{Q}$ mit $x \in A$ und $x \notin B$.

   Wir zeigen, dass dann $B \subseteq A$, also $B \leq A$ gilt. Zum Nachweis der Beziehung $B \subseteq A$ nehmen wir an, dass $y \in B$ ist und müssen $y \in A$ zeigen.

   Wir behaupten, dass $y < x$ ist und führen den Beweis dieser Behauptung indirekt, nehmen also $x \leq y$ an. Da die Dedekind-Menge B nach unten abgeschlossen ist und $y \in B$ ist, würde daraus

   $$x \in B$$

   folgen, was im Widerspruch zu der in diesem Fall gemachten Annahme $x \notin B$ steht. Also

haben wir

$$y < x.$$

Da die Menge A als Dedekind-Menge nach unten abgeschlossen ist und $x \in A$ ist, folgt

$$y \in A,$$

so dass wir $B \subseteq A$ gezeigt haben.

2. Fall: Es existiert ein $x \in \mathbb{Q}$ mit $x \in B$ und $x \notin A$.

   Dieser Fall ist offenbar analog zum ersten Fall. □

Aufgabe 4: Zeigen Sie, dass jede nicht-leere und in $\mathcal{D}$ nach oben beschränkte Menge $\mathcal{M} \subseteq \mathcal{D}$ ein Supremum $S \in \mathcal{D}$ hat.

Hinweis: Sie können das Supremum von $\mathcal{M}$ als die Vereinigung aller Mengen aus $\mathcal{M}$ definieren. Es gilt also

$$\sup(\mathcal{M}) = \bigcup \mathcal{M} := \{x \in \mathbb{Q} \mid \exists A \in \mathcal{M} : x \in A\}.$$ ◇

Definition 11 (Addition von Dedekind-Mengen) Es seien A und B Dedekind-Mengen. Dann definieren wir die Summe $A + B$ wie folgt:

$$A + B := \{x + y \mid x \in A \wedge y \in B\}.$$ ◇

Aufgabe 5: Es seien $A, B \in \mathcal{D}$. Zeigen Sie, dass dann auch $A + B \in \mathcal{D}$ ist. ◇

Lösung: Wir zeigen, dass $A + B$ eine Dedekind-Menge ist.

1. Wir zeigen $A + B \neq \{\}$.

   Da A eine Dedekind-Menge ist, gibt es ein Element $a \in A$ und da B ebenfalls eine Dedekind-Menge ist, gibt es auch ein Element $b \in B$. Nach Definition von $A+B$ folgt dann $a+b \in A+B$ und damit gilt $A + B \neq \{\}$.

2. Wir zeigen $A + B \neq \mathbb{Q}$.

   Da A und B Dedekind-Mengen sind, gilt $A \neq \mathbb{Q}$ und $B \neq \mathbb{Q}$. Also gibt es $x, y \in \mathbb{Q}$ mit $x \notin A$ und $y \notin B$. Wir zeigen, dass dann $x + y \notin A + B$ ist und führen diesen Nachweis indirekt. Wir nehmen also an, dass

   $$x + y \in A + B$$

   gilt. Nach Definition der Menge $A + B$ gibt es dann ein $a \in A$ und ein $b \in B$, so dass

   $$x + y = a + b$$

   ist. Aus $x \notin A$ und $a \in A$ folgt, dass

   $$a < x$$

   ist, denn da A eine Dedekind-Menge ist, würde aus $x \leq a$ sofort $x \in A$ folgern. Weil B eine Dedekind-Menge ist, gilt dann auch

   $$b < y.$$

   Addieren wir diese beiden Ungleichungen, so erhalten wir

   $$a + b < x + y,$$

   was im Widerspruch zu der Gleichung $x + y = a + b$ steht.

3. Wir zeigen, dass die Menge $A + B$ nach unten abgeschlossen ist.

   Sei also $x \in A + B$ und $y < x$. Nach Definition von $A + B$ gibt es dann ein $a \in A$ und ein $b \in B$, so dass $x = a + b$ gilt. Wir definieren

   $$c := x - y, \quad u := a - \frac{1}{2} \cdot c \quad \text{und} \quad v := b - \frac{1}{2} \cdot c.$$

   Aus $y < x$ folgt zunächst $c > 0$ und daher gilt $u < a$ und $v < b$. Da $a \in A$ ist und die Menge A als Dedekind-Menge nach unten abgeschlossen ist, folgt $u \in A$. Analog sehen wir, dass auch $v \in B$ ist. Insgesamt folgt dann

   $$u + v \in A + B.$$

   Wir haben aber

   $$\begin{array}{rcll} u + v & = & a - \frac{1}{2} \cdot c + b - \frac{1}{2} \cdot c & \\ & = & a + b - c & \\ & = & a + b - (x - y) & \text{denn } c = x - y \\ & = & x - (x - y) & \text{denn } x = a + b \\ & = & y & \end{array}$$

   Wegen $u + v \in A + B$ haben wir insgesamt $y \in A + B$ nachgewiesen, was zu zeigen war.

4. Wir zeigen, dass die Menge $A + B$ kein Maximum enthält.

   Wir führen den Beweis indirekt und nehmen an, dass ein $m \in A + B$ existiert, so dass $m = \max(A + B)$ gilt. Nach Definition der Menge $A + B$ gibt es dann ein $a \in A$ und ein $b \in B$ so dass $m = a + b$ ist. Sei nun $u \in A$. Wir wollen zeigen, dass $u \leq a$ ist. Wäre $u > a$, dann würde auch

   $$u + b > a + b$$

   gelten, und da $u + b \in A + B$ ist, könnte m dann nicht das Maximum der Menge $A + B$ sein. Also gilt $u \leq a$. Dann ist aber a das Maximum der Menge A und außerdem in A enthalten. Dies ist ein Widerspruch zu der Voraussetzung, dass A eine Dedekind-Menge ist. □

Aufgabe 6: Zeigen Sie, dass die Menge

$$O := \{x \in \mathbb{Q} \mid x < 0\}$$

eine Dedekind-Menge ist und zeigen Sie weiter, dass die Struktur $\langle \mathcal{D}, 0, + \rangle$ eine kommutative Gruppe ist. ◇

Lösung: Wir zeigen zunächst, dass O eine Dedekind-Menge ist und weisen dazu die einzelnen Eigenschaften einer Dedekind-Menge nach.

1. $O \neq \{\}$, denn es gilt $-1 \in O$.
2. $O \neq \mathbb{Q}$, denn es gilt $1 \notin O$.
3. Die Menge O ist nach unten abgeschlossen.

   Sei $x \in O$ und $y < x$. Nach Definition von O haben wir $x < 0$ und aus $y < x$ und $x < 0$ folgt $y < 0$, also gilt nach Definition der Menge O auch $y \in O$.
4. Die Menge O enthält kein Maximum, denn falls m das Maximum der Menge O wäre, dann wäre $m < 0$ und daraus folgt sofort $\frac{1}{2} \cdot m < 0$. Damit wäre dann nach Definition der Menge O auch

   $$\frac{1}{2} \cdot m \in O.$$

Da andererseits aber

$$m < \frac{1}{2} \cdot m$$

ist, kann dann m nicht das Maximum der Menge O sein. Folglich hat die Menge O kein Maximum.

Als nächstes zeigen wir, dass die Menge O das links-neutrale Element bezüglich der Addition von Dedekind-Mengen ist, wir zeigen also, dass

$$O + A = A$$

gilt. Wir spalten den Nachweis dieser Mengen-Gleichheit in den Nachweis zweier Inklusionen auf.

1. "$\subseteq$": Es sei $u \in O + A$. Wir müssen $u \in A$ zeigen.

   Nach Definition von $O + A$ existiert ein $o \in O$ und ein $a \in A$ mit $u = o + a$. Aus $o \in O$ folgt $o < 0$. Also haben wir

   $$u < a$$

   und da A als Dedekind-Menge nach unten abgeschlossen ist, folgt $u \in A$.

2. "$\supseteq$": Sei nun $a \in A$. Zu zeigen ist $a \in O + A$.

   Da die Menge A eine Dedekind-Menge ist, kann a nicht das Maximum der Menge A sein. Folglich gibt es ein $b \in A$, dass größer als a ist, wir haben also

   $$a < b.$$

   Wir definieren $u := a - b$. Aus $a < b$ folgt dann $u < 0$ und damit gilt $u \in O$. Damit haben wir

   $$u + b \in O + A.$$

   Andererseits gilt

   $$u + b = (a - b) + b = a,$$

   so dass wir insgesamt $a \in O + A$ gezeigt haben.

Die Tatsache, dass für die Addition von Dedekind-Mengen sowohl das Kommutativ-Gesetz als auch das Assoziativ-Gesetz gilt, folgt unmittelbar aus der Kommutativität und der Assoziativität der Addition rationaler Zahlen.

Als nächstes geben wir für eine Dedekind-Menge A das additive Inverse $-A$ an:

$$-A := \{x \in \mathbb{Q} \mid \exists r \in \mathbb{Q} : r > 0 \wedge -x - r \notin A\}.$$

Die Menge $-A$ enthält also die rationalen Zahlen x für die $-x$ so groß ist, dass für ein geeignetes $r > 0$ die Zahl $-x - r$ kein Element von A mehr ist. Als erstes zeigen wir, dass $-A$ eine Dedekind-Menge ist.

1. Wir zeigen $-A \neq \emptyset$.

   Da A eine Dedekind-Menge ist, ist $A \neq \mathbb{Q}$. Daher gibt es ein $y \in \mathbb{Q}$ mit $y \notin A$. Wir definieren

   $$x := -(y + 1) \quad \text{und} \quad r := 1.$$

   Dann gilt offenbar

   $$-x - r = y + 1 - 1 = y \notin A$$

   und nach Definition der Menge $-A$ folgt $x \in -A$. Also haben wir $-A \neq \emptyset$ gezeigt.

2. Wir zeigen $-A \neq \mathbb{Q}$.

   Da A eine Dedekind-Menge ist, gilt $A \neq \emptyset$. Also gibt es ein $y \in A$. Wir definieren

$$x := -y.$$

Für beliebige $r \in \mathbb{Q}$ mit $r > 0$ gilt dann

$$-x - r = y - r < y$$

und das $y \in A$ ist und A als Dedekind-Menge nach unten abgeschlossen ist, folgt daraus

$$-x - r \in A \quad \text{für alle } r \in \mathbb{Q} \text{ mit } r > 0.$$

Nach der Definition von $-A$ folgt nun, dass x kein Element von $-A$ ist. Also gilt $-A \neq \mathbb{Q}$.

3. Wir zeigen, dass die Menge $-A$ nach unten abgeschlossen ist.

   Sei als $x \in -A$ und $y < x$. Wir müssen zeigen, dass dann auch $y \in -A$ ist. Aus der Voraussetzung $x \in -A$ folgt, dass es ein $r \in \mathbb{Q}$ mit $r > 0$ gibt, so dass

   $$-x - r \notin A$$

   ist. Aus $y < x$ folgt $-y > -x$ und damit gilt auch

   $$-x - r < -y - r.$$

   Wir zeigen, dass $-y - r \notin A$ ist und führen diesen Nachweis indirekt: Wäre $-y - r \in A$, so folgt aus der Ungleichung $-x - r < -y - r$ und der Tatsache, dass A als Dedekind-Menge nach unten abgeschlossen ist, dass $-x - r \in A$ wäre, was falsch ist. Also folgt $-y - r \notin A$ und nach Definition von $-A$ folgern wir $y \in -A$. Damit haben wir gezeigt, dass $-A$ nach unten abgeschlossen ist.

4. Wir zeigen, dass $-A$ kein Maximum hat.

   Wir führen diesen Nachweis indirekt und nehmen an, dass $x = \max(-A)$ ist. Insbesondere ist x dann ein Element von $-A$ und daher gibt es dann ein $r \in \mathbb{Q}$ mit $r > 0$ und $-x - r \notin A$. Die letzte Formel können wir auch als

   $$-x - \tfrac{1}{2} \cdot r - \tfrac{1}{2} \cdot r \notin A$$

   schreiben, woraus wir folgern können, dass

   $$x + \tfrac{1}{2} \cdot r \in -A$$

   ist. Da andererseits $x < x + \frac{1}{2} \cdot r$ ist, kann dann aber x nicht das Maximum von $-A$ sein. Dieser Widerspruch zeigt, dass die Menge $-A$ kein Maximum hat.

Als nächstes zeigen wir, dass für jede Dedekind-Menge A die Gleichung

$$(-A) + A = O$$

gilt. Wir spalten den Nachweis dieser Mengengleicheit in zwei Teile auf.

1. "$\subseteq$": Es sei $x + y \in -A + A$, also $x \in -A$ und $y \in A$. Wir haben zu zeigen, dass $x + y \in O$ ist.

   Wegen $x \in -A$ gibt es nach Definition der Menge $-A$ ein $r \in \mathbb{Q}$ mit $r > 0$, so dass $-x - r \notin A$ ist. Da $y \in A$ ist, muss $y < -x - r$ gelten. Daraus folgt

   $$x + y < -r < 0$$

   und damit gilt $x + y \in O$.

2. "$\supseteq$": Es sei nun $o \in O$, also $o < 0$. Wir müssen ein $x \in -A$ und ein $y \in A$ finden, so dass $o = x + y$ gilt.

   Wir definieren

   $$r := -\tfrac{1}{2} \cdot o.$$

   Da $o < 0$ ist, folgt $r > 0$ und außerdem gilt $r \in \mathbb{Q}$. Wir definieren die Menge M als

$$M := \{n \in \mathbb{Z} \mid n \cdot r \in A\}$$

Da $A \neq \mathbb{Q}$ ist, gibt es ein $z \in \mathbb{Q}$ so dass $z \notin A$ ist. Für die Zahlen $n \in \mathbb{Z}$, für die $n \cdot r > z$ ist, folgt dann $n \notin M$. Folglich ist die Menge M nach oben beschränkt und hat daher ein Maximum. Wir definieren

$$\widehat{n} := \max(M).$$

Dann gilt $\widehat{n} + 1 \notin M$, also

$$(\widehat{n} + 1) \cdot r \notin A.$$

Wir definieren jetzt

$$x := \widehat{n} \cdot r \quad \text{und} \quad y := -(\widehat{n} + 2) \cdot r.$$

Nach Definition von $\widehat{n}$ und M gilt dann $x \in A$ und aus $(\widehat{n} + 1) \cdot r \notin A$ folgt

$$-y - r = (\widehat{n} + 2) \cdot r - r = (\widehat{n} + 1) \cdot r \notin A,$$

so dass $y \in -A$ ist. Außerdem gilt

$$x + y = \widehat{n} \cdot r - (\widehat{n} + 2) \cdot r = -2 \cdot r = o.$$

Damit haben wir $O \subseteq A + -A$ gezeigt. □

Aufgabe 7: Überlegen Sie, wie sich auf der Menge $\mathcal{D}$ eine Multiplikation definieren lässt, so dass $\mathcal{D}$ mit dieser Multiplikation und der oben definierten Addition ein Körper wird.

Lösung: Wir nennen eine Dedekind-Menge A positiv, wenn $0 \in A$ gilt. Für zwei positive Dedekind-Mengen A und B lässt sich die Multiplikation $A \cdot B$ als

$$A \cdot B := \{x \cdot y \mid x \in A \wedge y \in B \wedge x > 0 \wedge y > 0\} \cup \{z \in \mathbb{Q} \mid z \leq 0\}$$

definieren. Wir zeigen, dass die so definerte Menge $A \cdot B$ eine Dedekind-Menge ist. Dazu weisen wir die einzelnen Eigenschaften getrennt nach.

1. Wir zeigen $A \cdot B \neq \{\}$.

   Nach Definition von $A \cdot B$ gilt $0 \in A \cdot B$. Daraus folgt sofort $A \cdot B \neq \{\}$.

2. Wir zeigen $A \cdot B \neq \mathbb{Q}$.

   Da A und B als Dedekind-Mengen von der Menge $\mathbb{Q}$ verschieden sind, gibt es $u, v \in \mathbb{Q}$ mit $u \notin A$ und $v \notin B$. Wir definieren $w := \max(u, v)$. Dann gilt

   $$(\forall x \in A : x < w) \wedge (\forall y \in B : y < w)$$

   Daraus folgt sofort, dass für alle $x \in A$ und $y \in B$ die Ungleichung

   $$x \cdot y < w \cdot w$$

   gilt. Das heißt aber $w^2 \notin A \cdot B$.

3. Wir zeigen, dass $A \cdot B$ nach unten abgeschlossen ist.

   Es sei $x \cdot y \in A \cdot B$ und $z \in \mathbb{Q}$ mit $z < x \cdot y$. Wir müssen $z \in A \cdot B$ zeigen. Wir führen eine Fall-Unterscheidung danach durch, ob $z > 0$ ist.

   (a) Fall: $z > 0$. Dann definieren wir

   $$\alpha := \frac{z}{x \cdot y}$$

   Aus $z < x \cdot y$ folgt $\alpha < 1$. Wir setzen $u := \alpha \cdot x$ und folglich gilt $u < x$. Da A nach unten abgeschlossen ist, folgt $u \in A$. Damit haben wir insgesamt $u \cdot y \in A \cdot B$. Es gilt aber

   $$u \cdot y = \alpha \cdot x \cdot y = \frac{z}{x \cdot y} \cdot x \cdot y = z,$$

so dass wir insgesamt $z \in A \cdot B$ gezeigt haben.

(b) Fall: $z \leq 0$. Dann folgt unmittelbar aus der Definition von $A \cdot B$, dass $z \in A \cdot B$ ist.

4. Wir zeigen, dass $A \cdot B$ kein Maximum hat.

   Wir führen den Nachweis indirekt und nehmen an, dass die Menge $A \cdot B$ eine Maximum c hat. Es gilt dann

   $$c \in A \cdot B \quad \text{und} \quad \forall z \in A \cdot B : z \leq c.$$

   Nach Definition von $A \cdot B$ gibt es dann ein $a \in A$ und ein $b \in B$ mit $c = a \cdot b$. Wir zeigen, dass dann a das Maximum der Menge A ist. Sei also $u \in A$. Dann gilt

   $$u \cdot b \in A \cdot B \quad \text{und folglich gilt} \quad u \cdot b \leq c = a \cdot b.$$

   Teilen wir die letzte Ungleichung durch b so folgt

   $$u \leq a$$

   und damit wäre a das Maximum der Menge A. Das ist eine Widerspruch zu der Tatsache, dass A eine Dedekind-Menge ist.

Bisher haben wir das Produkt $A \cdot B$ nur für den Fall definiert, dass A und B beide positiv sind. Falls A oder B gleich O ist, definieren wir das Produkt als O:

$$A \cdot O := O \cdot B := O$$

Falls A weder positiv noch gleich O ist, sagen wir, dass A *negativ* ist. In einem solchen Fall ist $-A$ positiv. Falls A oder B negativ ist, lautet die Definition wie folgt:

2. Fall: A ist positiv, aber B negativ. Dann ist $-B$ positiv und wir können

   $$A \cdot B := -\big(A \cdot (-B)\big)$$

   definieren.

3. Fall: B ist positiv, aber A ist negativ. Dann setzen wir

   $$A \cdot B := -\big((-A) \cdot B\big).$$

4. Fall: A und B sind negativ. Wir definieren

   $$A \cdot B := (-A) \cdot (-B).$$

Nun müssten wir noch nachweisen, dass für die so definierte Multiplikation zusammen mit der oben definierten Addition die Körper-Axiome gelten. Aus Zeitgründen verzichten wir darauf.

Literatur-Hinweise
In dem Buch "*Grundlagen der Analysis*" von Edmund Landau [5] wird die oben skizzierte Konstruktion der reellen Zahlen im Detail beschrieben. Auch das Buch "*Principles of Mathematical Analysis*" von Walter Rudin [6] diskutiert die Konstruktion der reellen Zahlen mit Hilfe von Dedekind-Mengen ausführlich.

2.3 Geschichte

Die Konstruktion der reellen Zahlen mit Hilfe von Schnitten geht auf Richard Dedekind zurück, der die nach ihm benannten Schnitte in dem Buch Stetigkeit und irrationale Zahlen [7], das im Jahre 1872 erschienen ist, beschrieben hat. Damit war erstmals eine formale Definition des Begriffs der reellen Zahlen gefunden worden. Diese Definition war eine der wichtigsten Fortschritte im Bereich der mathematischen Grundlagenforschung des 19. Jahrhunderts, denn sie ermöglichte es, die Analysis auf ein solides Fundament zu stellen.

Kapitel 3

Folgen und Reihen

Die Begriffe *Folgen* und *Reihen* sowie der Begriff des *Grenzwerts* bilden die Grundlage, auf der die Analysis aufgebaut ist. Da Reihen nichts anderes sind als spezielle Folgen, beginnen wir unsere Diskussion mit den Folgen.

3.1 Folgen

Definition 12 (Folge) Eine Funktion $f : \mathbb{N} \to \mathbb{R}$ bezeichnen wir als eine *reellwertige Folge*. Eine Funktion $f : \mathbb{N} \to \mathbb{C}$ bezeichnen wir als eine *komplexwertige Folge*. ◇

Ist die Funktion f ein Folge, so schreiben wir dies kürzer als $\big(f(n)\big)_{n\in\mathbb{N}}$ oder $\big(f_n\big)_{n\in\mathbb{N}}$ oder noch kürzer als $\big(f_n\big)_n$.

Beispiele:

1. Die Funktion $a : \mathbb{N} \to \mathbb{R}$, die durch $a(n) = \dfrac{1}{n}$ definiert ist, schreiben wir als die Folge $\Big(\dfrac{1}{n}\Big)_{n\in\mathbb{N}}$.
2. Die Funktion $a : \mathbb{N} \to \mathbb{R}$, die durch $a(n) = (-1)^n$ definiert ist, schreiben wir als die Folge $\big((-1)^n\big)_{n\in\mathbb{N}}$.
3. Die Funktion $a : \mathbb{N} \to \mathbb{R}$, die durch $a(n) = n$ definiert ist, schreiben wir als die Folge $\big(n\big)_{n\in\mathbb{N}}$.

Folgen können auch induktiv definiert werden. Um die Gleichung $x = \cos(x)$ zu lösen, können wir eine Folge $(x_n)_{n\in\mathbb{N}}$ induktiv wie folgt definieren:

1. Induktions-Anfang: $n = 1$. Wir setzen
$$x_1 := 0.$$
2. Induktions-Schritt: $n \mapsto n + 1$. Nach Induktions-Voraussetzung ist x_n bereits definiert. Wir definieren x_{n+1} als
$$x_{n+1} := \cos(x_n).$$ ◇

Wir können die ersten 40 Glieder dieser Folge mit dem in Abbildung 3.1 gezeigten SetlX-Programm berechnen. Wir erhalten dann die in der Tabelle 3.1 auf Seite 19 gezeigten Ergebnisse. Bei näherer Betrachtung der Ergebnisse stellen wir fest, dass die Folge $\big(x_n\big)_{n\in\mathbb{N}}$ in einem gewissen Sinne gegen einen festen *Grenzwert* strebt. Diese Beobachtung wollen wir in der folgenden Definition präzisieren. Vorab bezeichnen wir die Menge der positiven reellen Zahlen mit $\mathbb{R}_+$, es gilt also

$$\mathbb{R}_+ = \big\{x \in \mathbb{R} \mid x > 0\big\}.$$

```
solve := procedure(k) {
    x    := [];  // x[n+1] stores x_{n}
    x[1] := 0.0;
    for (n in [1 .. k]) {
        x[n+1] := cos(x[n]);
        print("x_{$n$} = $x[n+1]$");
    }
};
```

Abbildung 3.1: Berechnung der durch $x_0 = 0$ und $x_{n+1} = \cos(x_n)$ definierten Folge.

n	x_n	n	x_n	n	x_n	n	x_n
0	0.000000	10	0.731404	20	0.738938	30	0.739082
1	1.000000	11	0.744237	21	0.739184	31	0.739087
2	0.540302	12	0.735605	22	0.739018	32	0.739084
3	0.857553	13	0.741425	23	0.739130	33	0.739086
4	0.654290	14	0.737507	24	0.739055	34	0.739085
5	0.793480	15	0.740147	25	0.739106	35	0.739086
6	0.701369	16	0.738369	26	0.739071	36	0.739085
7	0.763960	17	0.739567	27	0.739094	37	0.739085
8	0.722102	18	0.738760	28	0.739079	38	0.739085
9	0.750418	19	0.739304	29	0.739089	39	0.739085

Tabelle 3.1: Die ersten 40 Glieder der durch $x_0 = 0$ und $x_{n+1} = \cos(x_n)$ definierten Folge.

Definition 13 (Grenzwert) Eine Folge $(a_n)_{n\in\mathbb{N}}$ *konvergiert* gegen den *Grenzwert* g, falls gilt:

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall n \in \mathbb{N} : n \geq K \rightarrow |a_n - g| < \varepsilon.$$

In diesem Fall schreiben wir

$$\lim_{n\to\infty} a_n = g. \qquad \diamond$$

Anschaulich besagt diese Definition, dass fast alle Glieder a_n der Folge $(a_n)_{n\in\mathbb{N}}$ einen beliebig kleinen Abstand zu dem Grenzwert g haben. Für die oben induktiv definierte Folge x_n können wir den Nachweis der Konvergenz erst in einem späteren Kapitel antreten. Wir betrachten statt dessen ein einfacheres Beispiel und beweisen, dass

$$\lim_{n\to\infty} \frac{1}{n} = 0$$

gilt.

Beweis: Für jedes $\varepsilon > 0$ müssen wir eine Zahl K angeben, so dass für alle natürlichen Zahlen n, die größer-gleich K sind, die Abschätzung

$$\left| \frac{1}{n} - 0 \right| < \varepsilon$$

gilt. Wir definieren $K := \dfrac{1}{\varepsilon} + 1$. Damit ist K wohldefiniert, denn da ε positiv ist, gilt sicher auch $\varepsilon \neq 0$. Nun benutzen wir die Voraussetzung $n \geq K$ für $K = \dfrac{1}{\varepsilon} + 1$:

$$
\begin{aligned}
& n \geq \frac{1}{\varepsilon} + 1 \\
\Rightarrow \quad & n > \frac{1}{\varepsilon} && | \cdot \varepsilon \\
\Rightarrow \quad & n \cdot \varepsilon > 1 && | \cdot \frac{1}{n} \\
\Rightarrow \quad & \varepsilon > \frac{1}{n}
\end{aligned}
$$

Da andererseits $0 < \frac{1}{n}$ gilt, haben wir insgesamt für alle $n > K$

$$
\begin{aligned}
& 0 < \frac{1}{n} < \varepsilon \\
\Rightarrow \quad & \left| \frac{1}{n} \right| < \varepsilon \\
\Rightarrow \quad & \left| \frac{1}{n} - 0 \right| < \varepsilon
\end{aligned}
$$

gezeigt und damit ist der Beweis abgeschlossen. □

Aufgabe 8:

(a) Beweisen Sie unter Rückgriff auf die Definition des Grenzwert-Begriffs, dass

$$\lim_{n \to \infty} \frac{1}{2^n} = 0$$

gilt.

(b) Beweisen Sie unter Rückgriff auf die Definition des Grenzwert-Begriffs, dass

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} = 0$$

gilt. ◇

Wir formulieren und beweisen einige unmittelbare Folgerungen aus der obigen Definition des Grenzwerts.

Satz 14 (Eindeutigkeit des Grenzwerts) Konvergiert die Folge $(a_n)_{n \in \mathbb{N}}$ sowohl gegen den Grenzwert g_1 als auch gegen den Grenzwert g_2, so gilt $g_1 = g_2$.

Beweis: Wir führen den Beweis indirekt und nehmen an, dass $g_1 \neq g_2$ ist. Dann definieren wir $\varepsilon = \frac{1}{2} \cdot |g_2 - g_1|$ und aus der Annahme $g_1 \neq g_2$ folgt $\varepsilon > 0$. Aus der Voraussetzung, dass $(a_n)_{n \in \mathbb{N}}$ gegen g_1 konvergiert folgt, dass es ein K_1 gibt, so dass gilt:

$$\forall n \in \mathbb{N} : n \geq K_1 \rightarrow |a_n - g_1| < \varepsilon$$

Analog folgt aus der Voraussetzung, dass $(a_n)_{n \in \mathbb{N}}$ gegen g_2 konvergiert, dass es ein K_2 gibt, so dass gilt:

$$\forall n \in \mathbb{N} : n \geq K_2 \rightarrow |a_n - g_2| < \varepsilon$$

Wir setzen $K := \max(K_1, K_2)$. Alle $n \in \mathbb{N}$, die größer-gleich K sind, sind dann sowohl größer-gleich K_1 als auch größer-gleich K_2. Unter Benutzung der *Dreiecksungleichung*[1] erhalten wir für alle $n \geq K$ die folgende Kette von Ungleichungen:

[1] Sind $a, b \in \mathbb{R}$, so gilt $|a + b| \leq |a| + |b|$. Diese Ungleichung trägt den Namen *Dreiecksungleichung*.

$$\begin{aligned} 2 \cdot \varepsilon &= |g_2 - g_1| \\ &= |(g_2 - a_n) + (a_n - g_1)| \\ &\leq |g_2 - a_n| + |a_n - g_1| \qquad \text{(Dreiecksungleichung)} \\ &< \varepsilon + \varepsilon \\ &= 2 \cdot \varepsilon \end{aligned}$$

Aus dieser Ungleichungs-Kette würde aber $2 \cdot \varepsilon < 2 \cdot \varepsilon$ folgen und dass ist ein Widerspruch. Somit ist die Annahme $g_1 \neq g_2$ falsch und es muss $g_1 = g_2$ gelten. □

Bemerkung: Die Schreibweise $\lim\limits_{n \to \infty} a_n = g$ wird durch den letzten Satz im Nachhinein gerechtfertigt.

Aufgabe 9: Zeigen Sie, dass die Folge $\left((-1)^n\right)_{n \in \mathbb{N}}$ nicht konvergent ist. ◇

Lösung: Wir führen den Beweis indirekt und nehmen an, dass die Folge $\left((-1)^n\right)_{n \in \mathbb{N}}$ konvergiert. Bezeichnen wir diesen Grenzwert mit s, so gilt also

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall n \in \mathbb{N} : n \geq K \rightarrow \left|(-1)^n - s\right| < \varepsilon$$

Daher gibt es für $\varepsilon = 1$ eine Zahl K, so dass

$$\forall n \in \mathbb{N} : n \geq K \rightarrow \left|(-1)^n - s\right| < 1$$

gilt. Da aus $n \geq K$ sicher auch $2 \cdot n \geq K$ und $2 \cdot n + 1 \geq K$ folgt, hätten wir dann für $n \geq K$ die beiden folgenden Ungleichungen:

$$\left|(-1)^{2 \cdot n} - s\right| < 1 \quad \text{und} \quad \left|(-1)^{2 \cdot n + 1} - s\right| < 1$$

Wegen $(-1)^{2 \cdot n} = 1$ und $(-1)^{2 \cdot n + 1} = -1$ haben wir also

$$\left|1 - s\right| < 1 \quad \text{und} \quad \left|-1 - s\right| < 1. \tag{3.1}$$

Wegen $-1 - s = (-1) \cdot (1 + s)$ und $|a \cdot b| = |a| \cdot |b|$ können wir die letzte Ungleichung noch vereinfachen zu

$$\left|1 + s\right| < 1. \tag{3.2}$$

Aus den beiden Ungleichungen $|1 - s| < 1$ und $|1 + s| < 1$ erhalten wir nun einen Widerspruch:

$$\begin{aligned} 2 &= \left|1 + 1\right| \\ &= \left|(1 - s) + (s + 1)\right| \\ &\leq \left|1 - s\right| + \left|1 + s\right| \qquad \text{(Dreiecksungleichung)} \\ &< 1 + 1 \qquad \text{wegen der Ungleichungen (3.1) und (3.2)} \\ &= 2 \end{aligned}$$

Fassen wir diese Ungleichungs-Kette zusammen, so haben die (offensichtlich falsche) Ungleichung $2 < 2$ abgeleitet. Damit haben wir aus der Annahme, dass die Folge gegen den Grenzwert s konvergiert, einen Widerspruch hergeleitet. □

Definition 15 (beschränkte Folgen) Eine Folge $\left(a_n\right)_{n \in \mathbb{N}}$ ist *beschränkt*, falls es eine *Schranke* S gibt, so dass

$$\forall n \in \mathbb{N} : \left|a_n\right| \leq S$$

gilt. ◇

Die Folge $\left((-1)^n\right)_{n \in \mathbb{N}}$ ist durch die Schranke $S = 1$ beschränkt, denn offenbar gilt

$$\left|(-1)^n\right| = 1 \leq 1,$$

aber die Folge $\big(n\big)_{n\in\mathbb{N}}$ ist nicht beschränkt, denn sonst gäbe es eine Zahl S, so dass für alle natürlichen Zahlen n die Ungleichung $n \leq S$ gilt. Da es beliebig große natürliche Zahlen gibt, kann dies nicht sein.

Satz 16 (Beschränktheit konvergenter Folgen) Jede konvergente Folge ist beschränkt.

Beweis: Es sei $\big(a_n\big)_{n\in\mathbb{N}}$ eine konvergente Folge und es gelte

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

Dann gibt es für beliebige $\varepsilon > 0$ ein K, so dass gilt

$$\forall n \in \mathbb{N} : n \geq K \to \big|a_n - g\big| < \varepsilon.$$

Wir können also für $\varepsilon = 1$ ein K finden, so dass

$$\forall n \in \mathbb{N} : n \geq K \to \big|a_n - g\big| < 1$$

gilt. Wir können ohne Einschränkung der Allgemeinheit davon ausgehen, dass K eine natürliche Zahl ist, denn wenn K keine natürliche Zahl ist, können wir K einfach durch die erste natürliche Zahl ersetzen, die größer als K ist. Dann definieren wir

$$S := \max\big\{|a_0|, |a_1|, \cdots, |a_K|, 1 + |g|\big\}.$$

Wir behaupten, dass S eine Schranke für die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ ist, wir zeigen also, dass für alle $n \in \mathbb{N}$ gilt:

$$|a_n| \leq S$$

Um diese Ungleichung nachzuweisen, führen wir eine Fall-Unterscheidung durch:

1. Fall: $n \leq K$. Dann gilt offenbar

   $$|a_n| \in \big\{|a_0|, |a_1|, \cdots, |a_K|, 1 + |g|\big\}.$$

   und daraus folgt sofort

   $$|a_n| \leq \max\big\{|a_0|, |a_1|, \cdots, |a_K|, 1 + |g|\big\} = S.$$

2. Fall: $n > K$. Dann haben wir

   $$\begin{array}{lcll} |a_n| & = & |a_n - g + g| & \\ & \leq & |a_n - g| + |g| & \text{(Dreiecksungleichung)} \\ & < & 1 + |g| & \text{wegen } n > K \\ & \leq & S. & \end{array}$$

□

Aus den letzten beiden Sätzen folgt nun sofort, dass die Folge $\big(n\big)_{n\in\mathbb{N}}$ nicht konvergiert, denn diese Folge ist noch nicht einmal beschränkt.

Satz 17 (Summe konvergenter Folgen) Sind $\big(a_n\big)_{n\in\mathbb{N}}$ und $\big(b_n\big)_{n\in\mathbb{N}}$ zwei Folgen, so dass

$$\lim_{n\to\infty} a_n = a \quad \wedge \quad \lim_{n\to\infty} b_n = b$$

gilt, dann konvergiert die Folge $\big(a_n + b_n\big)_{n\in\mathbb{N}}$ gegen den Grenzwert $a + b$, in Zeichen:

$$\lim_{n\to\infty} \big(a_n + b_n\big) = \Big(\lim_{n\to\infty} a_n\Big) + \Big(\lim_{n\to\infty} b_n\Big).$$

Beweis: Es sei $\varepsilon > 0$ fest vorgegeben. Wir suchen ein K, so dass

$$\forall n \in \mathbb{N} : n \geq K \to \big|\big(a_n + b_n\big) - (a + b)\big| < \varepsilon$$

gilt. Nach Voraussetzung gibt es für beliebige $\varepsilon' > 0$ ein K_1 und ein K_2, so dass

$$\forall n \in \mathbb{N} : n \geq K_1 \to \big|a_n - a\big| < \varepsilon' \quad \text{und} \quad \forall n \in \mathbb{N} : n \geq K_2 \to \big|b_n - b\big| < \varepsilon'$$

gilt. Wir setzen nun $\varepsilon' := \dfrac{1}{2} \cdot \varepsilon$. Dann gibt es also K_1 und K_2, so dass

$$\forall n \in \mathbb{N} : n \geq K_1 \to \big|a_n - a\big| < \frac{1}{2} \cdot \varepsilon \quad \text{und} \quad \forall n \in \mathbb{N} : n \geq K_2 \to \big|b_n - b\big| < \frac{1}{2} \cdot \varepsilon$$

gilt. Wir definieren $K := \max(K_1, K_2)$. Damit gilt dann für alle $n \geq K$:

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

Damit ist die Behauptung gezeigt. □

Aufgabe 10: Zeigen Sie: Sind $\big(a_n\big)_{n\in\mathbb{N}}$ und $\big(b_n\big)_{n\in\mathbb{N}}$ zwei Folgen, so dass

$$\lim_{n\to\infty} a_n = a \quad \wedge \quad \lim_{n\to\infty} b_n = b$$

gilt, dann konvergiert die Folge $\big(a_n - b_n\big)_{n\in\mathbb{N}}$ gegen den Grenzwert $a - b$, in Zeichen:

$$\lim_{n\to\infty} \big(a_n - b_n\big) = \Big(\lim_{n\to\infty} a_n\Big) - \Big(\lim_{n\to\infty} b_n\Big).$$ ◇

Satz 18 (Produkt konvergenter Folgen) Sind $\big(a_n\big)_{n\in\mathbb{N}}$ und $\big(b_n\big)_{n\in\mathbb{N}}$ zwei Folgen, so dass

$$\lim_{n\to\infty} a_n = a \quad \wedge \quad \lim_{n\to\infty} b_n = b$$

gilt, dann konvergiert die Folge $\big(a_n \cdot b_n\big)_{n\in\mathbb{N}}$ gegen den Grenzwert $a \cdot b$, in Zeichen:

$$\lim_{n\to\infty} \big(a_n \cdot b_n\big) = \Big(\lim_{n\to\infty} a_n\Big) \cdot \Big(\lim_{n\to\infty} b_n\Big).$$

Beweis: Es sei $\varepsilon > 0$ fest vorgegeben. Wir suchen ein K, so dass

$$\forall n \in \mathbb{N} : n \geq K \to \big|\big(a_n \cdot b_n\big) - (a \cdot b)\big| < \varepsilon$$

gilt. Da die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ konvergent ist, ist diese Folge auch beschränkt, es gibt also eine Zahl S, so dass

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

gilt. Nach Voraussetzung gibt es für beliebige $\varepsilon_1 > 0$ ein K_1 und für beliebige $\varepsilon_2 > 0$ ein K_2, so dass

$$\forall n \in \mathbb{N} : n \geq K_1 \to \big|a_n - a\big| < \varepsilon_1 \quad \text{und} \quad \forall n \in \mathbb{N} : n \geq K_2 \to \big|b_n - b\big| < \varepsilon_2$$

gilt. Wir setzen nun $\varepsilon_1 := \dfrac{\varepsilon}{2 \cdot (|b| + 1)}$ und $\varepsilon_2 := \dfrac{\varepsilon}{2 \cdot S}$. Dann gibt es also K_1 und K_2, so dass

$$\forall n \in \mathbb{N} : n \geq K_1 \to \big|a_n - a\big| < \frac{\varepsilon}{2 \cdot (|b| + 1)} \quad \text{und} \quad \forall n \in \mathbb{N} : n \geq K_2 \to \big|b_n - b\big| < \frac{\varepsilon}{2 \cdot S}$$

gilt. Wir definieren $K := \max(K_1, K_2)$. Damit gilt dann für alle $n \geq K$:

$$\begin{aligned}
\big|a_n \cdot b_n - a \cdot b\big| &= \big|\big(a_n \cdot b_n - a_n \cdot b\big) + \big(a_n \cdot b - a \cdot b\big)\big| \\
&\leq \big|\big(a_n \cdot b_n - a_n \cdot b\big)\big| + \big|\big(a_n \cdot b - a \cdot b\big)\big| \quad \text{(Dreiecksungleichung)} \\
&= \big|a_n\big| \cdot \big|b_n - b\big| + \big|a_n - a\big| \cdot |b| \\
&\leq S \cdot \big|b_n - b\big| + \big|a_n - a\big| \cdot (|b| + 1) \\
&< S \cdot \frac{\varepsilon}{2 \cdot S} + \frac{\varepsilon}{2 \cdot (|b| + 1)} \cdot (|b| + 1) \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&= \varepsilon.
\end{aligned}$$

Damit ist die Behauptung gezeigt. □

Aufgabe 11: Zeigen Sie: *Sind* $\big(a_n\big)_{n\in\mathbb{N}}$ *und* $\big(b_n\big)_{n\in\mathbb{N}}$ *zwei Folgen, so dass*

$$\lim_{n\to\infty} a_n = a \quad \wedge \quad \lim_{n\to\infty} b_n = b$$

gilt und gilt $b_n \neq 0$ *für alle* $n \in \mathbb{N}$*, sowie* $b \neq 0$*, so konvergiert die Folge* $\big(a_n/b_n\big)_{n\in\mathbb{N}}$ *gegen den Grenzwert* a/b*, in Zeichen:*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\left(\lim\limits_{n\to\infty} a_n\right)}{\left(\lim\limits_{n\to\infty} b_n\right)} = \frac{a}{b}.$$ ◇

Lösung: Zunächst können wir das Problem vereinfachen, wenn wir die Folge $\big(a_n/b_n\big)_{n\in\mathbb{N}}$ als Folge von Produkten schreiben:

$$\left(\frac{a_n}{b_n}\right)_{n\in\mathbb{N}} = \big(a_n\big)_{n\in\mathbb{N}} \cdot \left(\frac{1}{b_n}\right)_{n\in\mathbb{N}}$$

Falls wir zeigen können, dass

$$\lim_{n\to\infty} \frac{1}{b_n} = \frac{1}{b}$$

gilt, dann folgt die Behauptung aus dem Satz über das Produkt konvergenter Folgen. Bei unserer Suche nach einem Beweis starten wir damit, dass wir die Behauptung noch einmal hinschreiben:

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall n \in \mathbb{N} : n \geq K \rightarrow \left|\frac{1}{b_n} - \frac{1}{b}\right| < \varepsilon \tag{3.3}$$

Wir müssen also für alle $\varepsilon > 0$ ein K finden, so dass für alle natürlichen Zahlen $n > K$ die Ungleichung

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| < \varepsilon \tag{3.4}$$

gilt. Irgendwie müssen wir die Voraussetzung, dass die Folge $\big(b_n\big)_{n\in\mathbb{N}}$ gegen b konvergiert, ausnutzen. Diese Voraussetzung lautet ausgeschrieben

$$\forall \varepsilon' \in \mathbb{R}_+ : \exists K' \in \mathbb{R} : \forall n \in \mathbb{N} : n \geq K' \rightarrow \big|b_n - b\big| < \varepsilon' \tag{3.5}$$

Wir zeigen zunächst eine Abschätzung für die Beträge $|b_n|$, die wir später brauchen. Hier hilft uns die Voraussetzung, dass $b \neq 0$ ist. Setzen wir in Ungleichung (3.5) für ε' den Wert $\frac{1}{2} \cdot |b|$ ein, so erhalten wir eine Zahl K_1, so dass für alle natürlichen Zahlen $n \geq K_1$

$$\big|b_n - b\big| < \frac{1}{2} \cdot |b|$$

gilt. Damit folgt:

$$\begin{aligned}
& |b| &=& |b - b_n + b_n| \\
\Rightarrow\quad & |b| &\leq& |b - b_n| + |b_n| \\
\Rightarrow\quad & |b| &<& \frac{1}{2} \cdot |b| + |b_n| \\
\Rightarrow\quad & \frac{1}{2} \cdot |b| &<& |b_n| \\
\Rightarrow\quad & \frac{2}{|b|} &>& \frac{1}{|b_n|}
\end{aligned}$$

Damit wissen wir also, dass für alle $n > K_1$ die Ungleichung

$$\frac{1}{2} \cdot |b| < |b_n|$$

gilt. Um nun für ein gegebenes $\varepsilon > 0$ die Ungleichung (3.4) zu zeigen, setzen wir in der Voraussetzung (3.5) $\varepsilon' = \frac{1}{2} \cdot |b|^2 \cdot \varepsilon$ und erhalten ein K_2, so dass für alle $n > K_2$ die Ungleichung

$$|b - b_n| < \frac{1}{2} \cdot |b|^2 \cdot \varepsilon \tag{3.6}$$

gilt. Setzen wir $K := \max(K_1, K_2)$, so erhalten wir für alle $n > K$ die folgende Ungleichungs-Kette:

$$\begin{aligned}
\left| \frac{1}{b_n} - \frac{1}{b} \right| &= \left| \frac{b - b_n}{b \cdot b_n} \right| \\
&= \frac{1}{|b| \cdot |b_n|} \cdot |b - b_n| \\
&< \frac{2}{|b| \cdot |b|} \cdot |b - b_n| \qquad \text{wegen } \frac{2}{|b|} > \frac{1}{|b_n|} \\
&< \frac{2}{|b| \cdot |b|} \cdot \frac{1}{2} \cdot |b|^2 \cdot \varepsilon \quad \text{wegen (3.6)} \\
&= \varepsilon
\end{aligned}$$

Damit haben wir für $n \geq K$ die Ungleichung $\left| \frac{1}{b_n} - \frac{1}{b} \right| < \varepsilon$ hergeleitet und der Beweis ist abgeschlossen. □

Die bisher bewiesenen Sätzen können wir benutzen um die Grenzwerte von Folgen zu berechnen. Wir geben ein Beispiel:

$$\begin{aligned}
\lim_{n\to\infty} \frac{n}{n+1} &= \lim_{n\to\infty} \frac{1}{1 + \frac{1}{n}} \\
&= \frac{\lim\limits_{n\to\infty} 1}{\lim\limits_{n\to\infty} 1 + \frac{1}{n}} \\
&= \frac{1}{\lim\limits_{n\to\infty} 1 + \lim\limits_{n\to\infty} \frac{1}{n}} \\
&= \frac{1}{1+0} \\
&= 1
\end{aligned}$$

◇

Satz 19 Sind $\big(a_n\big)_{n\in\mathbb{N}}$ und $\big(b_n\big)_{n\in\mathbb{N}}$ zwei konvergente Folgen, so dass

$$\forall n \in \mathbb{N} : a_n \leq b_n$$

gilt, dann gilt auch

$$\lim_{n\to\infty} a_n \leq \lim_{n\to\infty} b_n.$$

Aufgabe 12: Beweisen Sie den letzten Satz. ◇

Definition 20 (monoton) Eine Folge $\big(a_n\big)_{n\in\mathbb{N}}$ ist *monoton steigend* falls

$$\forall n \in \mathbb{N} : a_n \leq a_{n+1}$$

gilt. Analog heißt eine Folge *monoton fallend* falls

$$\forall n \in \mathbb{N} : a_n \geq a_{n+1}.$$ ◇

Ein Beispiel für eine monoton fallende Folge ist die Folge $\left(\frac{1}{n}\right)_{n\in\mathbb{N}}$, denn es gilt

$$\begin{array}{rrcl} & n+1 & \geq & n \qquad | \cdot \frac{1}{n} \\ \Rightarrow & \frac{n+1}{n} & \geq & 1 \qquad | \cdot \frac{1}{n+1} \\ \Rightarrow & \frac{1}{n} & \geq & \frac{1}{n+1} \end{array}$$

Satz 21 Ist die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ monoton fallend und beschränkt, so ist die Folge auch konvergent.

Beweis: Wir definieren zunächst die Menge M als als die Menge aller unteren Schranken der Folge $\big(a_n\big)_{n\in\mathbb{N}}$

$$M := \big\{x \in \mathbb{Q} \mid \forall n \in \mathbb{N} : x \leq a_n\big\}.$$

Weil wir vorausgesetzt haben, dass die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ beschränkt ist, ist die Menge M sicher nicht leer. Außerdem ist die Menge M nach oben beschränkt, eine obere Schranke ist das Folgenglied a_1. Folglich hat die Menge M ein Supremum und wir können daher

$$s := \sup(M)$$

definieren. Wir zeigen, dass

$$\lim_{n\to\infty} a_n = s$$

gilt. Sei also $\varepsilon > 0$ gegeben. Da

$$s + \varepsilon > s$$

ist und s als das Supremum der Menge M definiert ist, können wir folgern, dass $s + \varepsilon \notin M$ ist. Nach Definition der Menge M gibt es dann eine Zahl $\widehat{n} \in \mathbb{N}$, so dass

$$a_{\widehat{n}} < s + \varepsilon \text{ ist.}$$

Da die Folge monoton fallend ist, gilt dann auch

$$a_n < s + \varepsilon \quad \text{für alle } n \geq \widehat{n}.$$

Andererseits ist $s - \frac{1}{2} \cdot \varepsilon < s$, so dass $s - \frac{1}{2} \cdot \varepsilon$ sicher ein Element der Menge M ist und damit dann auch eine untere Schranke der Folge $(a_n)_{n\in\mathbb{N}}$. Folglich gilt für alle $n \in \mathbb{N}$

$$s - \varepsilon < s - \tfrac{1}{2} \cdot \varepsilon \leq a_n.$$

Damit haben wir insgesamt

$$s - \varepsilon < a_n < s + \varepsilon \quad \text{für alle } n \geq \widehat{n}$$

und dies können wir auch als

$$|a_n - s| < \varepsilon \quad \text{für alle } n \geq \widehat{n}$$

schreiben. Nach Definition des Grenzwerts haben wir damit die Behauptung gezeigt. □

Aufgabe 13: Die Folge $\left(a_n\right)_{n\in\mathbb{N}}$ sei monoton steigend und beschränkt. Zeigen Sie, dass der Grenzwert

$$\lim_{n\to\infty} a_n$$

existiert. ◇

Aufgabe 14: Es seien $a, b \in \mathbb{R}$ und es gelte $a < b$. Zeigen Sie, dass dann auch

$$a < \frac{1}{2} \cdot (a + b) \quad \text{und} \quad \frac{1}{2} \cdot (a + b) < b$$

gilt. ◇

Definition 22 (Cauchy-Folge)
Eine Folge $\left(a_n\right)_{n\in\mathbb{N}}$ heißt *Cauchy-Folge* (Augustin-Louis Cauchy, 1789-1857), falls gilt:

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall m, n \in \mathbb{N} : m \geq K \wedge n \geq K \rightarrow \left|a_m - a_n\right| < \varepsilon. \qquad \diamond$$

In einer Cauchy-Folge $\left(a_n\right)_{n\in\mathbb{N}}$ liegen also die einzelnen Folgenglieder a_n mit wachsendem n immer dichter zusammen. Wir werden sehen, dass eine Folge genau dann konvergent ist, wenn die Folge eine Cauchy-Folge ist.

Satz 23 Jede konvergente Folge $\left(a_n\right)_{n\in\mathbb{N}}$ ist eine Cauchy-Folge.

Beweis: Es sei $a := \lim\limits_{n\to\infty} a_n$. Sei $\varepsilon > 0$ gegeben. Aufgrund der Konvergenz der Folge $\left(a_n\right)_{n\in\mathbb{N}}$ gibt es dann ein K, so dass

$$\forall n \in \mathbb{N} : n \geq K \rightarrow \left|a_n - a\right| < \frac{\varepsilon}{2}$$

gilt. Damit gilt für alle $m, n \in \mathbb{N}$ mit $m \geq K$ und $n \geq K$ die folgende Abschätzung:

$$\begin{aligned} |a_m - a_n| &= \left|(a_m - a) + (a - a_n)\right| \\ &\leq \left|a_m - a\right| + \left|a - a_n\right| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon \end{aligned}$$

Damit ist gezeigt, dass $\left(a_n\right)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist. □

Satz 24 Jede Cauchy-Folge ist beschränkt.

Beweis: Wenn $\left(a_n\right)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist, dann finden wir eine Zahl K, so dass für alle natürlichen Zahlen m, n, die größer-gleich K sind, die Ungleichung

$$\left|a_n - a_m\right| < 1$$

gilt. Sei nun h eine natürliche Zahl, die größer als K ist. Wir definieren

$$S := \max\left\{|a_1|, |a_2|, \cdots, |a_h|, 1 + |a_h|\right\}$$

und zeigen, dass S eine Schranke der Cauchy-Folge $\left(a_n\right)_{n\in\mathbb{N}}$ ist, wir zeigen also

$$\forall n \in \mathbb{N} : |a_n| \leq S.$$

Falls $n \leq h$ ist, ist diese Ungleichung evident. Für alle $n > h$ haben wir die folgende Abschätzung:

$$\begin{aligned} \big|a_n\big| &= \big|a_n - a_h + a_h\big| \\ &\leq \big|a_n - a_h\big| + \big|a_h\big| \\ &< 1 + \big|a_h\big| \\ &\leq S. \end{aligned}$$

Damit ist der Beweis abgeschlossen. □

Aufgabe 15: In dem gleich folgenden Beweis der Tatsache, dass jede Cauchy-Folge konvergent ist, werden wir zwei Eigenschaften des Supremums einer Menge M benutzen, die zwar offensichtlich sind, die wir aber auch formal beweisen sollten. Nehmen Sie an, dass Folgendes gilt:

1. $M \subseteq \mathbb{R}$ ist nach unten abgeschlossen, es gilt also
$$y < x \wedge x \in M \rightarrow y \in M,$$
2. $s = \sup(M)$ und
3. $\varepsilon > 0$.

Beweisen Sie, dass dann

$$s - \varepsilon \in M \quad \text{und} \quad s + \varepsilon \notin M$$

gilt. ◇

Theorem 25 Jede Cauchy-Folge ist konvergent.

Beweis: Der Beweis verläuft ähnlich wie der Nachweis, dass eine monotone und beschränkte Folge konvergent ist und zerfällt in zwei Teile:

1. Zunächst definieren wir eine Menge M, die nicht leer und nach oben beschränkt ist und definieren s als das Supremum dieser Menge.
2. Anschließend zeigen wir, dass die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ gegen s konvergiert.

Wir definieren die Menge M wie folgt:

$$M := \big\{x \in \mathbb{R} \mid \exists K \in \mathbb{N} : \forall n \in \mathbb{N} : n \geq K \rightarrow x \leq a_n\big\}.$$

Anschaulich ist M die Menge aller unteren Grenzen für die Mehrheit der Folgenglieder: Ist $x \in M$, so müssen von einem bestimmten Index K an alle weiteren Folgenglieder a_n durch x nach unten abgeschätzt werden. Wir nennen M daher die Menge der *unteren Majoritäts-Schranken* der Folge $(a_n)_{n\in\mathbb{N}}$, denn jedes Element aus M ist eine untere Schranke für die Mehrheit der Folgenglieder.

Es ist klar, dass M nach unten abgeschlossen ist, es gilt

$$y < x \wedge x \in M \rightarrow y \in M,$$

denn wenn x eine untere Schranke der Mehrheit aller Folgenglieder ist, dann ist sicher jede Zahl y die kleiner als x ist, ebenfalls eine untere Schranke der Mehrheit der Folgenglieder. Wir werden diese Eigenschaft später benötigen.

Außerdem ist die Menge M nach oben beschränkt, denn die Folge $(a_n)_{n\in\mathbb{N}}$ ist durch S nach oben beschränkt. Die Beschränktheit der Cauchy-Folge impliziert, dass die Menge M nicht leer ist, denn wenn für alle $n \in \mathbb{N}$ die Ungleichung $|a_n| \leq S$ gilt, dann gilt insbesondere $-S \leq a_n$ und daraus folgt sofort $-S \in M$. Als nicht-leere und nach oben beschränkte Menge hat M ein Supremum. Wir definieren

$$s := \sup(M)$$

und zeigen, dass

$$\lim_{n\to\infty} a_n = s$$

gilt. Sei $\varepsilon > 0$ gegeben. Wir suchen eine Zahl K, so dass für alle natürlichen Zahlen $n \geq K$ die Ungleichung

$$\big|a_n - s\big| < \varepsilon$$

gilt. Wir betrachten zunächst die Zahl $s - \frac{\varepsilon}{2}$. Wegen $s - \frac{\varepsilon}{2} < s$ und $s = \sup(M)$ folgt $s - \frac{\varepsilon}{2} \in M$, denn M ist nach unten abgeschlossen. Damit existiert dann nach Definition der Menge M als Menge der unteren Majoritäts-Schranken eine Zahl K_1, so dass für alle $n \in \mathbb{N}$ mit $n \geq K_1$ die Ungleichung

$$s - \frac{\varepsilon}{2} \leq a_n \tag{3.7}$$

gilt. Da die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist, gibt es eine Zahl K_2, so dass für alle $m, n \in \mathbb{N}$ mit $m > K_2$ und $n > K_2$ die Ungleichung

$$\big|a_n - a_m\big| < \frac{\varepsilon}{2} \tag{3.8}$$

gilt. Wir setzen nun $K = \max(K_1, K_2)$ und betrachten die Zahl $s + \frac{\varepsilon}{2}$, die wegen $s < s + \frac{\varepsilon}{2}$ sicher kein Element von M mehr ist, denn sonst wäre s nicht das Supremum von M. Nach Definition von M finden wir dann eine natürliche Zahl m, die größer als K ist, so dass

$$a_m < s + \frac{\varepsilon}{2} \tag{3.9}$$

gilt. Für diese Zahl m gilt sicher auch die Ungleichung (3.7), so dass wir insgesamt

$$s - \frac{\varepsilon}{2} \leq a_m < s + \frac{\varepsilon}{2}$$

haben. Daraus folgt sofort

$$\big|a_m - s\big| \leq \frac{\varepsilon}{2}. \tag{3.10}$$

Aufgrund der Ungleichung (3.8) haben wir jetzt für alle natürlichen Zahlen $n > K$ die folgende Kette von Ungleichungen:

$$\begin{aligned}
\big|a_n - s\big| &= \big|(a_n - a_m) + (a_m - s)\big| \\
&\leq \big|(a_n - a_m)\big| + \big|(a_m - s)\big| \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&= \varepsilon
\end{aligned}$$

Damit ist der Beweis abgeschlossen. □

3.2 Berechnung der Quadrat-Wurzel

Wir präsentieren nun eine Anwendung der bisher entwickelte Theorie und zeigen, wie die Quadrat-Wurzel einer reellen Zahl berechnet werden kann. Es sei eine reelle Zahl $a > 0$ gegeben. Gesucht ist eine reelle Zahl $b > 0$, so dass $b^2 = a$ ist. Unsere Idee ist es, die Zahl b iterativ als Lösung einer Fixpunkt-Gleichung zu berechnen. Wir definieren eine Folge b_n induktiv wie folgt:

I.A.: $n = 1$.

$$b_1 := \begin{cases} 1 & \text{falls } a \leq 1, \\ a & \text{sonst.} \end{cases}$$

I.S.: $n \mapsto n+1$.

$$b_{n+1} := \frac{1}{2} \cdot \left(b_n + \frac{a}{b_n} \right).$$

Um diese Definition zu verstehen, nehmen wir zunächst an, dass der Grenzwert dieser Folge existiert und den Wert $b \neq 0$ hat. Dann gilt

$$\begin{aligned} b &= \lim_{n\to\infty} b_n \\ &= \lim_{n\to\infty} b_{n+1} \\ &= \lim_{n\to\infty} \frac{1}{2} \cdot \left(b_n + \frac{a}{b_n} \right) \\ &= \frac{1}{2} \cdot \left(\lim_{n\to\infty} b_n + \frac{a}{\lim\limits_{n\to\infty} b_n} \right) \\ &= \frac{1}{2} \cdot \left(b + \frac{a}{b} \right) \end{aligned}$$

Damit ist b also eine Lösung der Gleichung $b = \frac{1}{2} \cdot \left(b + \frac{a}{b} \right)$. Wir formen diese Gleichung um:

$$\begin{aligned} & b = \frac{1}{2} \cdot \left(b + \frac{a}{b} \right) && \mid \cdot 2 \\ \Leftrightarrow \quad & 2 \cdot b = b + \frac{a}{b} && \mid -b \\ \Leftrightarrow \quad & b = \frac{a}{b} && \mid \cdot b \\ \Leftrightarrow \quad & b^2 = a && \mid \sqrt{\ } \\ \Leftrightarrow \quad & b = \sqrt{a} \end{aligned}$$

Falls die oben definierte Folge $\left(b_n\right)_{n\in\mathbb{N}}$ einen Grenzwert hat, dann ist dieser Grenzwert also die Wurzel der Zahl a. Wir werden die Konvergenz der Folge nachweisen indem wir zeigen, dass die Folge $\left(b\right)_{n\in\mathbb{N}}$ einerseits monoton fallend und andererseits nach unten beschränkt ist. Dazu betrachten wir zunächst die Differenz $b_{n+1}^2 - a$:

$$\begin{aligned} b_{n+1}^2 - a &= \frac{1}{4} \cdot \left(b_n + \frac{a}{b_n} \right)^2 - a \\ &= \frac{1}{4} \cdot \left(b_n^2 + 2 \cdot a + \frac{a^2}{b_n^2} \right) - a \\ &= \frac{1}{4} \cdot \left(b_n^2 - 2 \cdot a + \frac{a^2}{b_n^2} \right) \\ &= \frac{1}{4} \cdot \left(b_n - \frac{a}{b_n} \right)^2 \\ &\geq 0, \end{aligned}$$

denn das Quadrat einer reellen Zahl ist immer größer-gleich Null. Addieren wir auf beiden Seiten der Ungleichung

$$b_{n+1}^2 - a \geq 0$$

die Zahl a, so haben wir

$$b_{n+1}^2 \geq a \quad \text{und damit auch} \quad b_{n+1} \geq \sqrt{a} \quad \text{für alle } n \in \mathbb{N}$$

gezeigt. Nach unserer Definition der Folge $\left(b_n\right)_{n\in\mathbb{N}}$ gilt diese Ungleichung auch für den ersten Wert

$n = 1$, so dass wir also insgesamt die Ungleichung

$$b_n^2 \geq a \quad \text{und} \quad b_n \geq \sqrt{a} \quad \text{für alle } n \in \mathbb{N}$$

gezeigt haben. Daraus folgt, dass $\sqrt{a}$ eine untere Schranke der Folge $\big(b_n\big)_{n\in\mathbb{N}}$ ist. Dividieren wir die erste Ungleichung durch b_n, so folgt

$$b_n \geq \frac{a}{b_n}.$$

Die Zahl $\frac{1}{2} \cdot \left(b_n + \frac{a}{b_n}\right)$ ist der arithmetische Mittelwert der Zahlen b_n und $\frac{a}{b_n}$ und muss daher zwischen diesen beiden Zahlen liegen:

$$b_n \geq \frac{1}{2} \cdot \left(b_n + \frac{a}{b_n}\right) \geq \frac{a}{b_n}.$$

Dieser Mittelwert ist aber gerade b_{n+1}, es gilt also

$$b_n \geq b_{n+1} \geq \frac{a}{b_n}.$$

Dies zeigt, dass die Folge $\big(b_n\big)_{n\in\mathbb{N}}$ monoton fallend ist und da wir oben gesehen haben, dass die Folge durch $\sqrt{a}$ nach unten beschränkt ist, konvergiert die Folge. Wir hatten oben schon gezeigt, dass der Grenzwert dieser Folge dann den Wert $\sqrt{a}$ haben muss, es gilt also

$$\lim_{n\to\infty} b_n = \sqrt{a}$$

Abbildung 3.2 auf Seite 31 zeigt die Definition einer Prozedur `mySqrt`() in SetlX, die die ersten 9 Glieder der Folge berechnet und dann jeweils mit Hilfe der Funktion `nDecimalPlaces`() die ersten 100 Stellen der Werte ausgibt.

Die von diesem Programm berechnete Ausgabe ist in Abbildung 3.3 gezeigt. Sie können sehen, dass die Folge sehr schnell konvergiert. b_2 stimmt auf 2 Stellen hinter dem Komma mit dem Ergebnis überein, bei b_3 sind es bereits 5 Stellen, bei b_4 sind es 11 Stellen, bei b_5 sind es 23 Stellen, bei b_6 sind es 47 Stellen, bei b_7 haben wir 96 Stellen und ab dem Folgeglied b_8 ändern sich die ersten 100 Stellen hinter dem Komma nicht mehr.

In modernen Mikroprozessoren wird übrigens eine verfeinerte Version des in diesem Abschnitt beschriebenen Verfahrens eingesetzt. Die Verfeinerung besteht im wesentlichen darin, dass zunächst ein guter Startwert b_1 in einer Tabelle nachgeschlagen wird, die restlichen Folgeglieder werden dann in der Tat über die Rekursionsformel $b_{n+1} = \frac{1}{2} \cdot \left(b_n + \frac{a}{b_n}\right)$ berechnet.

```
1  mySqrt := procedure(a) {
2      if (a <= 1) {
3          b := 1;
4      } else {
5          b := a;
6      }
7      for (n in [1 .. 9]) {
8          b := 1/2 * (b + a/b);
9          print("$n$: $nDecimalPlaces(b, 100)$");
10     }
11     return b;
12 };
```

Abbildung 3.2: Ein *SetlX*-Programm zur iterativen Berechnung der Quadrat-Wurzel.

```
1: 1.5000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
2: 1.4166666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666
3: 1.4142156862745098039215686274509803921568627450980392156862745098039215686274509803921568627450980392156862745098039215686274509803921568627450980392
4: 1.4142135623746899106262955788901349101165596221157440445849050192000543718353892683589900431576443402
5: 1.4142135623730950488016896235025302436149819257761974284982894986231958242289236217849418367358303565
6: 1.4142135623730950488016887242096980785696718753772340015610131331132652556303399785317871612507104752
7: 1.4142135623730950488016887242096980785696718753769480731766797379907324784621070388503875343276416016
8: 1.4142135623730950488016887242096980785696718753769480731766797379907324784621070388503875343276415727
9: 1.4142135623730950488016887242096980785696718753769480731766797379907324784621070388503875343276415727
```

Abbildung 3.3: Berechnung der Quadrat-Wurzel mit Hilfe der Folge $b_{n+1} = \frac{1}{2} \cdot (b_n + \frac{a}{b_n})$.

Aufgabe 16:

(a) Es seien $a, b \in \mathbb{R}$. Die Folge $(a_n)_{n\in\mathbb{N}}$ werde durch Induktion wie folgt definiert:

(a) $a_1 := a$,

(b) $a_2 := b$,

(c) $a_{n+2} := \frac{1}{2} \cdot (a_n + a_{n+1})$.

Zeigen Sie, dass die Folge $(a_n)_{n\in\mathbb{N}}$ konvergiert und berechnen Sie den Grenzwert $\lim\limits_{n\to\infty} a_n$ in Abhängigkeit von den Startwerten a und b.

Hinweis: Die Gleichung für a_{n+2} ist eine Rekurrenz-Gleichung, die Sie über den Ansatz $a_n = \lambda^n$ lösen können. Sie werden dabei für λ zwei mögliche Werte λ_1 und λ_2 finden, die Sie für die Lösung in der Form

$$a_n = c_1 \cdot \lambda_1^n + c_2 \cdot \lambda_2^n$$

linear kombinieren müssen. Die Koeffizienten c_1 und c_2 können Sie aus den Gleichungen für a_1 und a_2 bestimmen.

(b) Es seien $a, b \in \mathbb{R}$ und zusätzlich gelte $a > 0$ und $b > 0$. Die Folge $(a_n)_{n\in\mathbb{N}}$ werde durch Induktion wie folgt definiert:

(a) $a_1 := a$,

(b) $a_2 := b$,

(c) $a_{n+2} := \sqrt{a_n \cdot a_{n+1}}$.

Zeigen Sie, dass die Folge $(a_n)_{n\in\mathbb{N}}$ konvergiert und berechnen Sie den Grenzwert $\lim\limits_{n\to\infty} a_n$ in Abhängigkeit von den Startwerten a und b. ◇

Bemerkung: Für zwei positive Zahlen a und b wird die Zahl c für die

$$c := \sqrt{a \cdot b}$$

gilt, als das *geometrische Mittel* von a und b bezeichnet.

(c) Es seien $a, b \in \mathbb{R}$ und zusätzlich gelte $a > 0$ und $b > 0$. Die Folge $(a_n)_{n\in\mathbb{N}}$ werde durch Induktion wie folgt definiert:

(a) $a_1 := a$,

(b) $a_2 := b$,

(c) $\frac{1}{a_{n+2}} := \frac{1}{2} \cdot \left(\frac{1}{a_n} + \frac{1}{a_{n+1}} \right)$.

Zeigen Sie, dass die Folge $(a_n)_{n\in\mathbb{N}}$ konvergiert und berechnen Sie den Grenzwert $\lim\limits_{n\to\infty} a_n$ in Abhängigkeit von den Startwerten a und b. ◇

Bemerkung: Für zwei positive Zahlen a und b wird die Zahl c für die

$$\frac{1}{c} := \frac{1}{2}\cdot\left(\frac{1}{a}+\frac{1}{b}\right)$$

gilt, als das *harmonische Mittel* von a und b bezeichnet.

3.3 Reihen

Definition 26 (Reihe) Ist $\left(a_n\right)_{n\in\mathbb{N}}$ eine Folge, so definieren wir die Folge der *Partial-Summen* $\left(s_n\right)_{n\in\mathbb{N}}$ durch die Festsetzung

$$s_n := \sum_{i=1}^{n} a_i.$$

Diese Folge bezeichnen wir auch als unendliche *Reihe*. Die Folge $\left(a_n\right)_{n\in\mathbb{N}}$ bezeichnen wir als die der Reihe $\left(\sum_{i=0}^{n} a_i\right)_{n\in\mathbb{N}}$ *zugrunde liegende Folge*. Falls die Folge der Partial-Summen konvergiert, so schreiben wir den Grenzwert als

$$\sum_{i=1}^{\infty} a_i := \lim_{n\to\infty}\sum_{i=1}^{n} a_i.$$ ◇

Gelegentlich treten in der Praxis Folgen $\left(a_n\right)_{n\in\mathbb{N}}$ auf, für welche die Folgenglieder a_i erst ab einem Index $k>1$ definiert sind. Um auch aus solchen Folge bequem Reihen bilden zu können, definieren wir in einem solchen Fall die Partial-Summen s_n durch

$$s_n = \sum_{i=k}^{n} a_i,$$

wobei wir vereinbaren, dass $\sum_{i=k}^{n} a_i = 0$ ist, falls $k>n$ ist.

Satz 27 (Bernoullische Ungleichung) Es sei $x\in\mathbb{R}$, $n\in\mathbb{N}_0$ und es gelte $x\geq -1$. Dann gilt

$$(1+x)^n \geq 1+n\cdot x.$$

Diese Ungleichung wird als *Bernoullische Ungleichung* (Jakob Bernoulli, 1655-1705) bezeichnet.

Beweis: Wir beweisen die Ungleichung durch vollständige Induktion für alle $n\in\mathbb{N}_0$.

I.A.: $n=0$. Es gilt

$$(1+x)^0 = 1 \geq 1 = 1+0\cdot x. \quad \checkmark$$

I.S.: $n\mapsto n+1$. Nach Induktions-Voraussetzung gilt

$$(1+x)^n \geq 1+n\cdot x \tag{3.11}$$

Da $x\geq -1$ ist, folgt $1+x\geq 0$, so dass wir die Ungleichung 3.11 mit $1+x$ multiplizieren können. Dann erhalten wir die folgende Ungleichungs-Kette

$$\begin{aligned}(1+x)^{n+1} &\geq (1+n\cdot x)\cdot(1+x)\\ &= 1+(n+1)\cdot x+n\cdot x^2\\ &\geq 1+(n+1)\cdot x\end{aligned}$$

Also haben wir insgesamt

$$(1+x)^{n+1} \geq 1 + (n+1) \cdot x$$

gezeigt und das ist die Behauptung für $n+1$. ✓ □

Satz 28 Es sei $q \in \mathbb{R}$ mit $|q| < 1$. Dann gilt

$$\lim_{n\to\infty} q^n = 0.$$

Beweis: Wir nehmen zunächst an, dass q positiv ist. Aus $q < 1$ folgt dann

$$1 < \frac{1}{q} \quad \text{und damit} \quad 0 < \frac{1}{q} - 1.$$

Wir definieren nun

$$x := \frac{1}{q} - 1.$$

Mit Hilfe der Bernoullischen Ungleichung sehen wir nun, dass Folgendes gilt:

$$\begin{aligned} \frac{1}{q^n} &= \left(1 + \left(\frac{1}{q} - 1\right)\right)^n \\ &\geq 1 + n \cdot \left(\frac{1}{q} - 1\right) \\ &= 1 + n \cdot x. \end{aligned}$$

Durch Invertierung dieser Ungleichung erhalten wir

$$q^n \leq \frac{1}{1 + n \cdot x}$$

Ist nun ein $\varepsilon > 0$ gegeben, so definieren wir

$$K := \left(\frac{1}{\varepsilon} - 1\right) \cdot \frac{1}{x} + 1.$$

Dann gilt für alle $n \geq K$:

$$\begin{aligned} & \left(\frac{1}{\varepsilon} - 1\right) \cdot \frac{1}{x} + 1 \leq n \\ \Rightarrow \quad & \left(\frac{1}{\varepsilon} - 1\right) \cdot \frac{1}{x} < n \\ \Rightarrow \quad & \left(\frac{1}{\varepsilon} - 1\right) < n \cdot x \\ \Rightarrow \quad & \frac{1}{\varepsilon} < 1 + n \cdot x \\ \Rightarrow \quad & \frac{1}{1 + n \cdot x} < \varepsilon \end{aligned}$$

Insgesamt haben wir nun für alle $n \geq K$ gezeigt, dass

$$0 < q^n \leq \frac{1}{1 + n \cdot x} < \varepsilon$$

gilt, also haben wir für $n \geq K$

$$|q^n| < \varepsilon.$$

Für $q = 0$ ist diese Ungleichung offenbar auch gültig und wenn q negativ ist, gilt $-q > 0$, so dass die Ungleichung für $-q$ gilt:

$$|(-q)^n| < \varepsilon.$$

Wegen $|(-q)^n| = |q^n|$ folgt daraus also, dass für alle q die Ungleichung

$$|q^n| < \varepsilon \quad \text{für } n \geq K$$

gültig ist und damit ist die Behauptung bewiesen. □

Wir präsentieren nun einige Beispiele für konvergente Reihen:

1. Wir betrachten die Folge $\left(\frac{1}{n \cdot (n+1)}\right)_{n \in \mathbb{N}}$. Für die Partial-Summen zeigen wir durch Induktion über n, dass

$$\sum_{i=1}^{n} \frac{1}{i \cdot (i+1)} = 1 - \frac{1}{n+1} \tag{3.12}$$

gilt.

   (a) (Induktions-Anfang) $n = 1$: Einerseits haben wir für $n = 1$

$$\sum_{i=1}^{n} \frac{1}{i \cdot (i+1)} = \sum_{i=1}^{1} \frac{1}{i \cdot (i+1)} = \frac{1}{1 \cdot (1+1)} = \frac{1}{2},$$

   andererseits gilt

$$1 - \frac{1}{n+1} = 1 - \frac{1}{1+1} = 1 - \frac{1}{2} = \frac{1}{2}. \checkmark$$

   (b) (Induktions-Schritt) $n \mapsto n+1$:

$$\begin{aligned}
\sum_{i=1}^{n+1} \frac{1}{i \cdot (i+1)} &= \sum_{i=1}^{n} \frac{1}{i \cdot (i+1)} + \frac{1}{(n+1) \cdot (n+2)} \\
&\overset{IV}{=} 1 - \frac{1}{(n+1)} + \frac{1}{(n+1) \cdot (n+2)} \\
&= 1 - \frac{n+2-1}{(n+1) \cdot (n+2)} \\
&= 1 - \frac{n+1}{(n+1) \cdot (n+2)} \\
&= 1 - \frac{1}{n+2} \quad \checkmark
\end{aligned}$$

   Damit haben wir Gleichung (3.12) durch vollständige Induktion nachgewiesen. Aus Gleichung (3.12) folgt nun

$$\sum_{i=1}^{\infty} \frac{1}{i \cdot (i+1)} = \lim_{n \to \infty} \sum_{i=1}^{n} \frac{1}{i \cdot (i+1)} = \lim_{n \to \infty} \left(1 - \frac{1}{n+1}\right) = 1.$$

2. Wir betrachten die Folge $\left(q^n\right)_{n \in \mathbb{N}}$ für eine Zahl $q \in \mathbb{R}$. Für die Partial-Summen gilt

$$s_n = \sum_{i=0}^{n} q^i.$$

   Wir betrachten den Ausdruck $(1-q) \cdot s_n$:

$$\begin{aligned}
(1-q)\cdot s_n &= (1-q)\cdot \sum_{i=0}^{n} q^i \\
&= \sum_{i=0}^{n} q^i \;-\; q\cdot \sum_{i=0}^{n} q^i \\
&= \sum_{i=0}^{n} q^i \;-\; \sum_{i=0}^{n} q^{i+1} \\
&= \sum_{i=0}^{n} q^i \;-\; \sum_{i=1}^{n+1} q^i \\
&= \left(q^0 + \sum_{i=1}^{n} q^i\right) \;-\; \left(\sum_{i=1}^{n} q^i + q^{n+1}\right) \\
&= q^0 - q^{n+1} \\
&= 1 - q^{n+1}
\end{aligned}$$

Es gilt also

$$(1-q)\cdot \sum_{i=0}^{n} q^i = 1 - q^{n+1}$$

Dividieren wir diese Gleichung durch $(1-q)$, so erhalten wir für die Partial-Summen den Ausdruck

$$\sum_{i=0}^{n} q^i = \frac{1-q^{n+1}}{1-q}.$$

Falls $|q| < 1$ ist, konvergiert die Folge $\left(q^n\right)_{n\in\mathbb{N}}$ gegen 0. Damit gilt für $|q| < 1$

$$\sum_{i=0}^{\infty} q^i = \frac{1}{1-q}.$$

Die Reihe $\left(\sum_{i=0}^{n} q^i\right)_{n\in\mathbb{N}}$ wird als *geometrische Reihe* bezeichnet. ◇

Definition 29 (Alternierende Reihe) Hat eine Reihe die Form

$$\left(\sum_{i=1}^{n} (-1)^i \cdot a_i\right)_{n\in\mathbb{N}}$$

und gilt entweder

$$\forall i \in \mathbb{N} : a_i \geq 0 \quad \text{oder} \quad \forall i \in \mathbb{N} : a_i \leq 0$$

so sprechen wir von einer *alternierenden Reihe*. ◇

Beispiel: Die Reihe

$$\left(\sum_{i=1}^{n} \frac{(-1)^i}{i}\right)_{n\in\mathbb{N}}$$

ist eine alternierende Reihe. Wir werden später sehen, dass diese Reihe gegen den Wert $-\ln(2)$ konvergiert. ◇

Definition 30 (Null-Folge) Die Folge $\left(a_n\right)_{n\in\mathbb{N}}$ ist eine *Null-Folge* wenn gilt:

$$\lim_{n\to\infty} a_n = 0.$$ ◇

Satz 31 (Leibniz-Kriterium, (Gottfried Wilhelm Leibniz, 1646-1716))
Wenn die Folge $\left(a_n\right)_{n\in\mathbb{N}}$ eine monoton fallende Null-Folge ist, dann konvergiert die alternierende Reihe

$$\left(\sum_{i=1}^{n}(-1)^i\cdot a_i\right)_{n\in\mathbb{N}.}$$

Beweis: Die Partial-Summen s_n sind durch

$$s_n = \sum_{i=1}^{n}(-1)^i\cdot a_i$$

definiert. Wir betrachten zunächst die Folge der Partial-Summen mit geraden Indizes, wir betrachten also die Folge $\left(s_{2\cdot n}\right)_{n\in\mathbb{N}}$ und zeigen, dass diese Folge monoton fallend ist. Es gilt

$$s_{2\cdot(n+1)} = s_{2\cdot n} + (-1)^{2\cdot n+1}\cdot a_{2\cdot n+1} + (-1)^{2\cdot n+2}\cdot a_{2\cdot n+2} = s_{2\cdot n} - a_{2\cdot n+1} + a_{2\cdot n+2}. \tag{3.13}$$

Daraus folgt

$$\begin{array}{rrcl} & s_{2\cdot(n+1)} & \leq & s_{2\cdot n} \\ \Leftrightarrow & s_{2\cdot n} - a_{2\cdot n+1} + a_{2\cdot n+2} & \leq & s_{2\cdot n} \\ \Leftrightarrow & a_{2\cdot n+2} & \leq & a_{2\cdot n+1} \end{array}$$

Die letzte Ungleichung ist aber nichts anderes als die Monotonie der Folge $\left(a_n\right)_{n\in\mathbb{N}}$.

Als nächstes zeigen wir durch vollständige Induktion, dass die Folge der Partial-Summen mit geraden Indizes nach unten beschränkt ist, genauer gilt

$$s_{2\cdot n} \geq -a_1.$$

Um dies nachzuweisen, zeigen wir durch vollständige Induktion, dass für alle $n\in\mathbb{N}_0$ gilt:

$$s_{2\cdot n+1} \geq -a_1.$$

I.A.: $n = 0$.

$$s_{2\cdot 0+1} = s_1 = -a_1 \geq -a_1.$$

I.S.: $n \mapsto n+1$

$$\begin{array}{rcll} s_{2\cdot(n+1)+1} & = & s_{2\cdot n+1} + a_{2\cdot n+2} - a_{2\cdot n+3} & \\ & \geq & -a_1 + a_{2\cdot n+2} - a_{2\cdot n+3} & \text{nach Induktions-Voraussetzung} \\ & \geq & -a_1 & \text{wegen } a_{2\cdot n+2} \geq a_{2\cdot n+3}. \end{array}$$

Nun gilt für $n\in\mathbb{N}$

$$s_{2\cdot n} = s_{2\cdot n-1} + a_{2\cdot n} \geq s_{2\cdot n-1} \geq -a_1.$$

Da wir nun gezeigt haben, dass die Folge $\left(s_{2\cdot n}\right)_{n\in\mathbb{N}}$ sowohl monoton fallend als auch nach unten beschränkt ist, folgt aus Satz 21, dass diese Folge konvergent ist. Der Grenzwert dieser Folge sei s:

$$s := \lim_{n\to\infty} s_{2\cdot n}.$$

Dann konvergiert auch die Folge $\left(s_n\right)_{n\in\mathbb{N}}$ gegen s. Dies sehen wir wie folgt: Sei $\varepsilon > 0$ gegeben. Weil $\left(s_{2\cdot n}\right)_{n\in\mathbb{N}}$ gegen s konvergiert gibt es eine Zahl K_1, so dass für alle $n \geq K_1$ die Ungleichung

$$\left|s_{2\cdot n} - s\right| < \frac{1}{2}\cdot\varepsilon \tag{3.14}$$

erfüllt ist. Weil $\left(a_n\right)_{n\in\mathbb{N}}$ eine Null-Folge ist gibt es außerdem eine Zahl K_2, so dass für alle $n \geq K_2$ die Ungleichung

$$\left|a_n - 0\right| < \frac{1}{2}\cdot\varepsilon \tag{3.15}$$

gilt. Wir setzen $K := \max(2 \cdot K_1 + 1, K_2)$ und zeigen, dass für alle $n \geq K$ die Ungleichung

$$\left|s_n - s\right| < \varepsilon$$

gilt. Wir erbringen diesen Nachweis über eine Fall-Unterscheidung:

1. n ist gerade, also gilt $n = 2 \cdot m$.

$$\begin{aligned} \left|s_n - s\right| &= \left|s_{2 \cdot m} - s\right| \\ &< \frac{1}{2} \cdot \varepsilon \\ &< \varepsilon, \end{aligned}$$

   denn aus $n = 2 \cdot m$ und $n \geq K$ folgt $m \geq K_1$.

2. n ist ungerade, also gilt $n = 2 \cdot m + 1$.

$$\begin{aligned} \left|s_n - s\right| &= \left|s_{2 \cdot m+1} - s\right| \\ &= \left|s_{2 \cdot m+1} - s_{2 \cdot m} + s_{2 \cdot m} - s\right| \\ &\leq \left|s_{2 \cdot m+1} - s_{2 \cdot m}\right| + \left|s_{2 \cdot m} - s\right| \\ &< \left|a_{2 \cdot m+1}\right| + \frac{1}{2} \cdot \varepsilon \\ &< \frac{1}{2} \cdot \varepsilon + \frac{1}{2} \cdot \varepsilon \\ &= \varepsilon, \end{aligned}$$

   denn aus $n = 2 \cdot m + 1$ und $n \geq K$ folgt $m \geq K_1$ und $n \geq K_2$.

Damit ist der Beweis abgeschlossen. □

Satz 32 (Cauchy'sches Konvergenz-Kriterium für Reihen)
Die Reihe $\left(\sum_{i=0}^{n} a_i\right)_{n \in \mathbb{N}}$ ist genau dann konvergent, wenn es für alle $\varepsilon > 0$ eine Zahl K gibt, so dass gilt:

$$\forall n, l \in \mathbb{N} : n \geq K \rightarrow \left| \sum_{i=n+1}^{n+l} a_i \right| < \varepsilon.$$

Beweis: Nach den Sätzen aus dem Abschnitt über Folgen ist die Folge $\left(s_n\right)_{n \in \mathbb{N}}$ der durch

$$s_n = \sum_{i=0}^{n} a_i$$

definierten Partial-Summen genau dann konvergent, wenn $\left(s_n\right)_{n \in \mathbb{N}}$ eine Cauchy-Folge ist, wenn also gilt:

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall m, n \in \mathbb{N} : m \geq K \wedge n \geq K \rightarrow \left|s_m - s_n\right| < \varepsilon.$$

In der letzten Formel können wir ohne Einschränkung der Allgemeinheit annehmen, dass $n \leq m$ gilt. Dann ist $m = n + l$ für eine natürliche Zahl l. Setzen wir hier die Definition der Partial-Summen ein, so erhalten wir

$$\begin{aligned} |s_m - s_n| &= |s_{n+l} - s_n| \\ &= \left| \sum_{i=1}^{n+l} a_i - \sum_{i=1}^{n} a_i \right| \\ &= \left| \sum_{i=n+1}^{n+l} a_i \right| \end{aligned}$$

und damit ist klar, dass die Ungleichung des Satzes äquivalent dazu ist, dass die Folge der Partial-Summen eine Cauchy-Folge ist. □

Korollar 33
Wenn die Reihe $\left(\sum_{i=1}^{n} a_i\right)_{n\in\mathbb{N}}$ konvergent ist, dann ist die Folge $(a_n)_{n\in\mathbb{N}}$ eine Null-Folge.

Beweis: Nach dem Cauchy'schen Konvergenz-Kriterium gilt

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall n, l \in \mathbb{N} : n \geq K \rightarrow \left| \sum_{i=n+1}^{n+l} a_i \right| < \varepsilon.$$

Setzen wir hier $l = 1$ so haben wir insbesondere

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall n \in \mathbb{N} : n \geq K \rightarrow \left| \sum_{i=n+1}^{n+1} a_i \right| < \varepsilon.$$

Wegen

$$\left| \sum_{i=n+1}^{n+1} a_i \right| = |a_{n+1}|$$

folgt also

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall n \in \mathbb{N} : n \geq K \rightarrow |a_{n+1}| < \varepsilon.$$

Diese Formel drückt aus, dass $(a_n)_{n\in\mathbb{N}}$ eine Null-Folge ist. □

Mit Hilfe des letzten Satzes können wir zeigen, dass die *harmonische Reihe*

$$\left(\sum_{i=1}^{n} \frac{1}{i}\right)_{n\in\mathbb{N}}$$

divergiert. Wäre diese Reihe konvergent, so gäbe es nach dem Cauchy'schen Konvergenz-Kriterium eine Zahl K, so dass für alle $n \geq K$ und alle l die Ungleichung

$$\sum_{i=n+1}^{n+l} \frac{1}{i} < \frac{1}{2}$$

gilt. Insbesondere würde diese Ungleichung dann für $l = n$ gelten. Für beliebige n gilt aber die folgende Abschätzung:

$$\sum_{i=n+1}^{n+n} \frac{1}{i} \geq \sum_{i=n+1}^{2\cdot n} \frac{1}{2\cdot n} = n \cdot \frac{1}{2\cdot n} = \frac{1}{2}$$

Damit erfüllt die harmonische Reihe das Cauchy'sche Konvergenz-Kriterium nicht.

Satz 34 (Majoranten-Kriterium) Für die Folgen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ gelte:

1. $\forall n \in \mathbb{N} : 0 \leq a_n \leq b_n$.
2. Der Grenzwert $\sum_{i=1}^{\infty} b_i$ existiert.

Dann existiert auch der Grenzwert $\sum_{i=1}^{\infty} a_i$.

Beweis: Es gilt

$$\sum_{i=1}^{n} a_i \leq \sum_{i=1}^{n} b_i \leq \sum_{i=1}^{\infty} b_i.$$

Also ist die Folge $\left(\sum_{i=1}^{n} a_i\right)_{n\in\mathbb{N}}$ monoton wachsend und beschränkt und damit konvergent. □

Bemerkung: Oft wird im Majoranten-Kriterium die Voraussetzung

$$\forall n \in \mathbb{N} : 0 \leq a_n \leq b_n$$

abgeschwächt zu

$$\forall n \in \mathbb{N} : n \geq K \to 0 \leq a_n \leq b_n.$$

Hierbei ist K dann eine geeignet gewählte Schranke. Die Gültigkeit dieser Form des Majoranten-Kriteriums folgt aus der Tatsache, dass das Abändern endlich vieler Glieder einer Reihe für die Frage, ob eine Reihe konvergent ist, unbedeutend ist.

Aufgabe 17: Zeigen Sie mit dem Majoranten-Kriterium, dass die Reihe $\left(\sum_{i=1}^{n} \frac{1}{i^2}\right)_{n\in\mathbb{N}}$ konvergiert.

Lösung: Es gilt

$$\begin{aligned} & i+1 \geq i && \mid \cdot(i+1) \\ \Rightarrow \quad & (i+1)^2 \geq i \cdot (i+1) && \mid \frac{1}{\cdot} \\ \Rightarrow \quad & \frac{1}{(i+1)^2} \leq \frac{1}{i \cdot (i+1)} \end{aligned}$$

Damit ist die Reihe $\left(\sum_{i=1}^{n} \frac{1}{i \cdot (i+1)}\right)_{n\in\mathbb{N}}$ eine konvergente Majorante der Reihe $\left(\sum_{i=1}^{n} \frac{1}{(i+1)^2}\right)_{n\in\mathbb{N}}$.
Wegen

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{1}{1^2} + \sum_{i=1}^{\infty} \frac{1}{(i+1)^2}$$

folgt die Konvergenz aus dem Majoranten-Kriterium.

Bemerkung: Wir werden später zeigen, dass

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6} \quad \text{gilt.}$$

Satz 35 (Minoranten-Kriterium) Für die Folgen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ gelte:

1. $\forall n \in \mathbb{N} : 0 \leq a_n \leq b_n$.
2. Der Grenzwert $\sum_{i=1}^{\infty} a_i$ existiert nicht.

Dann existiert auch der Grenzwert $\sum_{i=1}^{\infty} b_i$ nicht.

Beweis: Wir führen den Beweis indirekt und nehmen an, dass $\sum_{i=1}^{\infty} b_i$ existiert. Nach dem Majoranten-Kriterium müsste dann auch der Grenzwert $\sum_{i=1}^{\infty} a_i$ existieren und dass steht im Widerspruch zur Voraussetzung. □

Aufgabe 18: Zeigen Sie, dass die Reihe

$$\left(\sum_{i=1}^{n} \frac{1}{\sqrt{i}}\right)_{n\in\mathbb{N}}$$

nicht konvergiert.

Lösung: Wir benutzen das Minoranten-Kriterium und zeigen, dass die Reihe $\left(\sum_{i=1}^{n} \frac{1}{i}\right)_{n\in\mathbb{N}}$ eine divergente Minorante der Reihe $\left(\sum_{i=1}^{n} \frac{1}{\sqrt{i}}\right)_{n\in\mathbb{N}}$ ist:

$$\begin{array}{rrcll} & \frac{1}{i} & \leq & \frac{1}{\sqrt{i}} & \mid \cdot \frac{1}{\cdot} \\ \Leftrightarrow & i & \geq & \sqrt{i} & \mid \cdot^2 \\ \Leftrightarrow & i^2 & \geq & i & \mid \cdot \frac{1}{i} \\ \Leftrightarrow & i & \geq & 1 & \mid \end{array}$$

Da die letzte Ungleichung offenbar für alle $n \in \mathbb{N}$ wahr ist, ist der Beweis abgeschlossen. □

Satz 36 (Quotienten-Kriterium) Es sei $(a_n)_{n\in\mathbb{N}}$ eine Folge und $q \in \mathbb{R}$ eine Zahl, so dass gilt:

1. $0 \leq q < 1$
2. $\forall n \in \mathbb{N} : 0 \leq a_n$
3. $\forall n \in \mathbb{N} : a_{n+1} \leq q \cdot a_n$

Dann konvergiert die Reihe $\left(\sum_{i=1}^{n} a_i\right)_{n\in\mathbb{N}}$.

Beweis: Wir zeigen, dass die geometrische Reihe $\left(\sum_{i=1}^{n} a_0 \cdot q^i\right)_{n\in\mathbb{N}}$ eine konvergente Majorante der Reihe $\left(\sum_{i=1}^{n} a_i\right)_{n\in\mathbb{N}}$ ist. Dazu zeigen wir durch Induktion über n, dass folgendes gilt:

$$\forall n \in \mathbb{N} : a_n \leq a_1 \cdot q^{n-1}$$

I.A. : $n = 1$. Wegen $q^0 = 1$ gilt trivialerweise

$$a_1 \leq a_1 \cdot q^0.$$

I.S. : $n \mapsto n + 1$. Es gilt:

$$\begin{array}{rcll} a_{n+1} & \leq & q \cdot a_n & \text{nach Voraussetzung} \\ & \leq & q \cdot a_1 \cdot q^{n-1} & \text{nach Induktions-Voraussetzung} \\ & = & a_1 \cdot q^n. & \end{array}$$

□

Bemerkung: Beim Quotienten-Kriterium sind eigentlich nur die Beträge der Folgenglieder $|a_n|$ wichtig, denn es lässt sich folgende Verschärfung des Quotienten-Kriteriums zeigen: Ist $(a_n)_{n\in\mathbb{N}}$ eine Folge, $q \in \mathbb{R}$ und $K \in \mathbb{R}$, so dass

$$0 \leq q < 1 \quad \wedge \quad \forall n \in \mathbb{N} : n \geq K \rightarrow \left|\frac{a_{n+1}}{a_n}\right| \leq q$$

gilt. Dann konvergiert die Reihe $\left(\sum_{i=1}^{n} a_i\right)_{n\in\mathbb{N}}$.

Beispiel: Wir zeigen mit Hilfe des Quotienten-Kriteriums, dass die Reihe $\left(\sum_{i=1}^{n} \frac{z^i}{i!}\right)_{n\in\mathbb{N}}$ für alle $z \in \mathbb{C}$ konvergiert. Für $z = 0$ ist die Konvergenz der Reihe trivial und sonst betrachten wir den Quotienten a_{n+1}/a_n für diese Reihe, setzen $K = 2 \cdot |z|$ und $q = \frac{1}{2}$ und zeigen, dass das Quotienten-Kriterium erfüllt ist, denn für alle $n \geq K$ gilt:

$$\left|\frac{a_{n+1}}{a_n}\right| = \frac{\frac{|z^{n+1}|}{(n+1)!}}{\frac{|z^n|}{n!}} = \frac{|z^{n+1}| \cdot n!}{|z^n| \cdot (n+1)!} = \frac{|z|}{n+1} \leq \frac{|z|}{K} = \frac{|z|}{2 \cdot |z|} = \frac{1}{2}.$$

Satz 37 (Wurzel-Kriterium) Es sei $\left(a_n\right)_{n\in\mathbb{N}}$ eine Folge und $q \in \mathbb{R}$ eine Zahl, so dass

1. $0 \leq q < 1$
2. $\forall n \in \mathbb{N} : 0 \leq a_n$
3. $\forall n \in \mathbb{N} : n > 0 \rightarrow \sqrt[n]{a_n} \leq q$

gilt. Dann konvergiert die Reihe $\left(\sum_{i=1}^{n} a_i\right)_{n\in\mathbb{N}}$.

Beweis: Auch hier können wir den Nachweis der Konvergenz dadurch führen indem wir zeigen, dass die geometrische Reihe $\left(\sum_{i=1}^{n} q^i\right)_{n\in\mathbb{N}}$ eine konvergente Majorante ist: Für $n > 0$ gilt

$$a_n \leq q^n \;\Leftrightarrow\; \sqrt[n]{a_n} \leq q.$$ □

Bemerkung: Beim Wurzel-Kriterium sind eigentlich nur die Beträge der Folgenglieder $|a_n|$ wichtig, denn es lässt sich folgende Verschärfung des Wurzel-Kriteriums zeigen: Ist $\left(a_n\right)_{n\in\mathbb{N}}$ eine Folge, $q \in \mathbb{R}$ und $K \in \mathbb{R}$, so dass

$$0 \leq q < 1 \quad \wedge \quad \forall n \in \mathbb{N} : n \geq K \rightarrow \sqrt[n]{|a_n|} \leq q$$

gilt. Dann konvergiert die Reihe $\left(\sum_{i=1}^{n} a_i\right)_{n\in\mathbb{N}}$.

Beispiel: Wir zeigen mit dem Wurzel-Kriterium, dass die Reihe $\left(\sum_{i=1}^{n} \frac{1}{i!}\right)_{n\in\mathbb{N}}$ konvergiert. Wir setzen $K = 4$ und $q = \frac{1}{2}$. Zunächst können Sie mit vollständiger Induktion leicht zeigen, dass für alle natürlichen Zahlen $n \geq 4$ die Ungleichung $n! \geq 2^n$ gilt. Damit haben wir für $n \geq 4$:

$$\sqrt[n]{\frac{1}{n!}} \leq \frac{1}{2} \quad \Leftrightarrow \quad \frac{1}{n!} \leq \left(\frac{1}{2}\right)^n \quad \Leftrightarrow \quad n! \geq 2^n.$$

3.4 Potenz-Reihen

Es bezeichne x eine Variable und $\left(a_n\right)_{n\in\mathbb{N}}$ sei eine Folge von Zahlen. Dann bezeichnen wir den Ausdruck

$$\sum_{n=0}^{\infty} a_n \cdot x^n$$

als *formale Potenz-Reihe*. Wichtig ist hier, dass x keine feste Zahl ist, sondern eine Variable, für die wir später reelle (oder auch komplexe) Zahlen einsetzen. Wenn wir in einer Potenz-Reihe für x eine feste Zahl einsetzen, wird aus der Potenz-Reihe eine gewöhnliche Reihe. Der Begriff der Potenz-Reihen kann als eine Verallgemeinerung des Begriffs des Polynoms aufgefasst werden.

Beispiele:

1. $\sum\limits_{n=0}^{\infty} \frac{x^n}{n!}$ ist eine formale Potenz-Reihe. Wir haben oben mit Hilfe des Quotienten-Kriteriums gezeigt, dass diese Reihe für beliebige reelle Zahlen konvergiert.
2. $\sum\limits_{n=1}^{\infty} \frac{x^n}{n}$ ist eine formale Potenz-Reihe. Setzen wir für x den Wert 1 ein, so erhalten wir die divergente harmonische Reihe. Für $x = -1$ erhalten wir eine alternierende Reihe, die nach dem Leibniz-Kriterium konvergent ist.

In der Theorie der Potenz-Reihen ist die Frage entscheidend, welche Zahlen wir für die Variable x einsetzen können, so dass die Reihe konvergiert. Diese Frage wird durch die folgenden Sätze beantwortet.

Satz 38 Wenn die Potenz-Reihe

$$\sum_{n=0}^{\infty} a_n \cdot x^n$$

für einen Wert $u \in \mathbb{C}$ konvergiert, dann konvergiert die Reihe auch für alle $v \in \mathbb{C}$, für die $|v| < |u|$ ist.

Beweis: Da die Reihe $\sum_{n=0}^{\infty} a_n \cdot u^n$ konvergiert, folgt aus dem Korollar zum Cauchy'schen Konvergenz-Kriterium, dass die Folge $\left(a_n \cdot u^n\right)_{n\in\mathbb{N}}$ eine Null-Folge ist. Also gibt es eine Zahl K, so dass für alle $n \geq K$ die Ungleichung

$$|a_n \cdot u^n| \leq 1$$

gilt. Wir definieren

$$q := \left|\frac{v}{u}\right|.$$

Aus $|v| < |u|$ folgt $q < 1$. Dann haben wir für alle $n \geq K$ die folgende Abschätzung:

$$|a_n \cdot v^n| = |a_n \cdot u^n| \cdot \left|\frac{v^n}{u^n}\right| = |a_n \cdot u^n| \cdot q^n \leq q^n.$$

Diese Abschätzung zeigt, dass die geometrische Reihe eine konvergente Majorante der Reihe $a_n \cdot v^n$ ist. Damit folgt die Konvergenz der Reihe $\sum_{n=0}^{\infty} a_n \cdot v^n$ aus dem Majoranten-Kriterium. □

Satz 39 Wenn die Potenz-Reihe $\sum_{n=0}^{\infty} a_n \cdot x^n$ für einen Wert $u \in \mathbb{C}$ divergiert, dann divergiert die Reihe auch für alle $v \in \mathbb{C}$, für die $|u| < |v|$ ist.

Beweis: Würde die Reihe $\sum_{n=0}^{\infty} a_n \cdot v^n$ konvergieren, dann müsste nach Satz 38 auch die Reihe $\sum_{n=0}^{\infty} a_n \cdot u^n$ konvergieren. □

Die letzten beiden Sätze ermöglichen es nun, den Begriff *Konvergenz-Radius* zu definieren. Es sei

$$\sum_{n=0}^{\infty} a_n \cdot x^n$$

eine formale Potenz-Reihe. Wenn diese Reihe für alle $x \in \mathbb{C}$ konvergiert, dann sagen wir, dass der Konvergenz-Radius den Wert ∞ hat. Andernfalls definieren wir den Konvergenz-Radius als

$$R := \sup\left\{|u| \;\middle|\; u \in \mathbb{C} \wedge \sum_{n=0}^{\infty} a_n \cdot u^n \text{ konvergiert}\right\}.$$

Aus den letzten beiden Sätzen folgt dann:

1. $\forall z \in \mathbb{C} : |z| < R \rightarrow \sum\limits_{n=0}^{\infty} a_n \cdot z^n$ konvergiert.
2. $\forall z \in \mathbb{C} : |z| > R \rightarrow \sum\limits_{n=0}^{\infty} a_n \cdot z^n$ divergiert.

In der Gauß'schen Zahlen-Ebene ist die Menge $\{z \in \mathbb{C} \mid |z| < R\}$ das Innere eines Kreises mit dem Radius R um den Nullpunkt. Der folgende Satz gibt uns eine effektive Möglichkeit, den Konvergenz-Radius zu berechnen.

Satz 40 Es sei

$$\sum_{n=0}^{\infty} a_n \cdot z^n$$

eine formale Potenz-Reihe und die Folge

$$\left(\frac{|a_n|}{|a_{n+1}|}\right)_{n\in\mathbb{N}}$$

konvergiere. Dann ist der Konvergenz-Radius R durch folgende Formel gegeben:

$$R = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|.$$

Beweis: Es sei $u \in \mathbb{C}$ mit $|u| < R$. In diesem Fall müssen wir zeigen, dass die Reihe

$$\sum_{n=0}^{\infty} a_n \cdot u^n$$

konvergiert. Wir werden diesen Nachweis mit Hilfe des Quotienten-Kriteriums erbringen. Wir setzen

$$q := \frac{|u|}{\frac{1}{2}\cdot(R+|u|)}$$

und zeigen, dass $q < 1$ ist:

$$\begin{array}{ll} & q < 1 \\ \Leftrightarrow & \dfrac{|u|}{\frac{1}{2}\cdot(R+|u|)} < 1 \\ \Leftrightarrow & 2\cdot|u| < R + |u| \\ \Leftrightarrow & |u| < R \end{array}$$

und die letzte Ungleichung ist nach Wahl von u wahr. Da wir nur Äquivalenzumformungen benutzt haben, ist damit auch die Formel $q < 1$ wahr. Wir zeigen weiter, dass für alle hinreichend großen n die Ungleichung

$$\frac{|a_{n+1}\cdot u^{n+1}|}{|a_n \cdot u^n|} \leq q$$

erfüllt ist. Um diesen Beweis zu führen, müssen wir etwas ausholen. Zunächst folgt aus

$$R = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|,$$

dass es für beliebige $\varepsilon > 0$ eine Zahl K gibt, so dass für alle $n \geq K$ die Ungleichung

$$\left|\left|\frac{a_n}{a_{n+1}}\right| - R\right| < \varepsilon$$

gilt. Wir setzen $\varepsilon := \frac{1}{2}(R - |u|)$. Wir zeigen, dass dann für alle $n \geq K$ die Ungleichung $\left|\frac{a_n}{a_{n+1}}\right| > \frac{1}{2}(R+|u|)$ gilt:

$$
\begin{aligned}
& \left| \, |\frac{a_n}{a_{n+1}}| + \big(R - |\frac{a_n}{a_{n+1}}|\big) \, \right| = |R| = R \\
\Rightarrow \quad & |\frac{a_n}{a_{n+1}}| + \big| \, \big(R - |\frac{a_n}{a_{n+1}}|\big) \, \big| \geq R \\
\Rightarrow \quad & |\frac{a_n}{a_{n+1}}| + \varepsilon > R \\
\Rightarrow \quad & |\frac{a_n}{a_{n+1}}| + \frac{1}{2}(R - |u|) > R \\
\Rightarrow \quad & |\frac{a_n}{a_{n+1}}| > R - \frac{1}{2}(R - |u|) \\
\Rightarrow \quad & |\frac{a_n}{a_{n+1}}| > \frac{1}{2}(R + |u|).
\end{aligned}
$$

Jetzt können wir zeigen, dass die Reihe $\sum_{n=0}^{\infty} a_n \cdot u^n$ das Quotienten-Kriterium erfüllt, denn für alle $n \geq K$ gilt:

$$\left| \frac{a_{n+1} \cdot u^{n+1}}{a_n \cdot u^n} \right| = \left| \frac{a_{n+1}}{a_n} \right| \cdot |u| = \frac{|u|}{\left| \frac{a_n}{a_{n+1}} \right|} < \frac{|u|}{\frac{1}{2}(R + |u|)} = q$$

Um den Beweis abzuschließen müssen wir noch zeigen, die Reihe $\sum_{n=0}^{\infty} a_n \cdot u^n$ divergiert wenn $R < |u|$ ist. Dies folgt aus der Tatsache, dass die Folge $\big(a_n \cdot u^n\big)_{n \in \mathbb{N}}$ für $|u| > R$ keine Null-Folge ist. Die Details bleiben dem Leser überlassen. □

Bemerkung: Der obige Satz bleibt auch richtig, wenn

$$\lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right| = \infty$$

ist, denn dann ist die Potenz-Reihe $\sum_{n=0}^{\infty} a_n \cdot u^n$ für alle $u \in \mathbb{C}$ konvergent.

Beispiel: Die Potenz-Reihe $\sum_{n=1}^{\infty} \frac{x^n}{n}$ hat den Konvergenz-Radius $R = 1$, denn es gilt

$$\lim_{n \to \infty} \left| \frac{\frac{1}{n}}{\frac{1}{n+1}} \right| = \lim_{n \to \infty} \frac{n+1}{n} = 1 + \lim_{n \to \infty} \frac{1}{n} = 1 + 0 = 1. \qquad \diamond$$

Satz 41 (Hadamard) Es sei $\sum_{n=0}^{\infty} a_n \cdot x^n$ eine Potenz-Reihe und die Folge $\left(\sqrt[n]{|a_n|} \right)_{n \in \mathbb{N}}$ konvergiere. Dann gilt

$$\frac{1}{R} = \lim_{n \to \infty} \sqrt[n]{|a_n|}.$$

Bemerkung: Setzen wir $\frac{1}{\infty} = 0$, so bleibt die Formel

$$\frac{1}{R} = \lim_{n \to \infty} \sqrt[n]{|a_n|}.$$

auch in dem Fall $\lim_{n \to \infty} \sqrt[n]{|a_n|} = 0$ richtig, denn dann gilt $R = \infty$.

Beispiel: Die Potenz-Reihe $\sum_{n=1}^{\infty} \frac{x^n}{n^n}$ hat den Konvergenz-Radius $R = \infty$, denn es gilt

$$\lim_{n \to \infty} \sqrt[n]{\frac{1}{n^n}} = \lim_{n \to \infty} \frac{1}{n} = 0.$$

Kapitel 4

Stetige und differenzierbare Funktionen

4.1 Stetige Funktionen

Wir wollen in diesem Abschnitt präzisieren, wann eine Funktion so *glatt* ist, dass wir sie zeichnen können, ohne dabei den Stift absetzen zu müssen. Für eine glatte Funktion soll außerdem der *Zwischenwert-Satz* gelten. Der Zwischenwert-Satz besagt, dass für eine glatte Funktion

$$f : \mathbb{R} \to \mathbb{R},$$

für die es $a, b \in \mathbb{R}$ gibt mit $f(a) < 0$ und $f(b) > 0$, auch ein $c \in \mathbb{R}$ existiert, so dass $f(c) = 0$ ist. Anschaulich ist dieser Satz klar: Ist beispielsweise $a < b$ und zeichne ich die Funktion f in dem Intervall $[a, b]$, so muss der Graph der Funktion an irgendeiner Stelle des Intervalls $[a, b]$ die x-Achse schneiden. Voraussetzung dafür, dass dies tatsächlich so ist, ist aber die Forderung, dass die Funktion f in einem gewissen Sinne *glatt* ist, denn wenn wir beispielsweise die Funktion

$$g : \mathbb{R} \to \mathbb{R}$$

betrachten, die durch

$$g(x) := \begin{cases} -1 & \text{falls } x < 0 \\ +1 & \text{falls } x \geq 0 \end{cases}$$

definiert ist, so haben wir zwar $g(-1) = -1$ und $g(1) = 1$, aber es gibt kein $x \in [-1, 1]$, für das $g(x) = 0$ wäre. Das liegt daran, dass die Funktion eben nicht *glatt* ist, denn die Funktion hat an der Stelle $x = 0$ einen Sprung. Unser Ziel in diesem Abschnitt ist es, zunächst exakt zu definieren, was wir unter einer glatten Funktion verstehen wollen. In der Mathematik wird an Stelle des Attributs *glatt* der Begriff der *Stetigkeit* verwendet.[1] Um diesen Begriff einführen zu können, bedarf es einer Reihe von zusätzlichen Definitionen, die nun folgen.

Es sei $(x_n)_{n\in\mathbb{N}}$ eine Folge und $D \subseteq \mathbb{R}$. Wir sagen, dass *die Folge* $(x_n)_{n\in\mathbb{N}}$ *in* D *liegt*, wenn für alle $n \in \mathbb{N}$ das Folgenglied $x_n \in D$ ist.

[1] Gelegentlich wird eine Funktion als *glatt* bezeichnet, wenn die Funktion unendlich oft differenzierbar ist. Das ist eine wesentlich schärfere Forderung als der Begriff der Stetigkeit. In diesem Skript verwende ich den Begriff *glatt* aber synonym mit dem Begriff *stetig*.

Definition 42 (Grenzwert) Es sei $D \subseteq \mathbb{R}$ und $f : D \to \mathbb{R}$. Weiter sei $\widehat{x} \in \mathbb{R}$ und $\lambda \in \mathbb{R}$. Außerdem gebe es mindestens eine Folge $(x_n)_{n\in\mathbb{N}}$, die gegen $\widehat{x}$ konvergiert. Dann ist λ der *Grenzwert* der Funktion f *im Punkt* $\widehat{x}$, wenn für jede in D liegende Folge $(x_n)_{n\in\mathbb{N}}$ gilt:

$$\lim_{n\to\infty} x_n = \widehat{x} \;\Rightarrow\; \lim_{n\to\infty} f(x_n) = \lambda.$$

In diesem Fall schreiben wir

$$\lim_{x\to\widehat{x}} f(x) = \lambda. \qquad \diamond$$

Bemerkung: In der obigen Definition ist nicht gefordert, dass $\widehat{x}$ ein Element des Definitionsbereichs D ist. In vielen interessanten Fällen ist dies auch nicht der Fall, beispielsweise werden wir später zeigen, dass

$$\lim_{x\to 0} \frac{\sin(x)}{x} = 1$$

gilt. Die Funktion $x \mapsto \dfrac{\sin(x)}{x}$ ist für $x = 0$ nicht definiert, trotzdem existiert der Grenzwert. $\diamond$

Definition 43 (Stetigkeit) Es sei $D \subseteq \mathbb{R}$ und $f : D \to \mathbb{R}$. Weiter sei $\widehat{x} \in D$. Dann ist die Funktion f *stetig im Punkt* $\widehat{x}$, wenn gilt:

$$\lim_{x\to\widehat{x}} f(x) = f(\widehat{x}). \qquad \diamond$$

Aus den beiden letzten Definitionen folgt, dass eine stetige Funktion mit dem Prozess der Grenzwert-Bildung vertauschbar ist. Für eine konvergente Folge $(x_n)_{n\in\mathbb{N}}$ und eine stetige Funktion f gilt

$$f\Big(\lim_{n\to\infty} x_n\Big) = \lim_{n\to\infty} f(x_n).$$

Beispiele:

1. Es sei $c \in \mathbb{R}$. Dann ist die konstante Funktion $f : \mathbb{R} \to \mathbb{R}$, die durch $f(x) := c$ definiert ist, in jedem Punkt $\widehat{x} \in \mathbb{R}$ stetig, denn für jede beliebige Folge $(x_n)_{n\in\mathbb{N}}$ gilt

   $$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} c = c.$$

2. Die identische Funktion $id : \mathbb{R} \to \mathbb{R}$, die durch $id(x) = x$ definiert ist, ist in jedem Punkt $\widehat{x} \in \mathbb{R}$ stetig, denn wenn $(x_n)_{n\in\mathbb{N}}$ eine Folge ist, so dass

   $$\lim_{n\to\infty} x_n = \widehat{x}$$

   gilt, dann folgt sofort

   $$\lim_{n\to\infty} id(x_n) = \lim_{n\to\infty} x_n = \widehat{x}.$$

3. Die Funktion $f : \mathbb{R} \to \mathbb{R}$, die durch $f(x) = x^2$ definiert ist, ist in jedem Punkt stetig, denn falls $(x_n)_{n\in\mathbb{N}}$ eine Folge ist, so dass

   $$\lim_{n\to\infty} x_n = \widehat{x}$$

   gilt, dann folgt nach dem Satz über den Grenzwert einer Folge von Produkten

   $$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_n^2 = \Big(\lim_{n\to\infty} x_n\Big) \cdot \Big(\lim_{n\to\infty} x_n\Big) = \widehat{x} \cdot \widehat{x} = \widehat{x}^2.$$

4. Das letzte Beispiel lässt sich verallgemeinern: Die Funktionen $f : \mathbb{R} \to \mathbb{R}$ und $g : \mathbb{R} \to \mathbb{R}$ seien im Punkt $\widehat{x}$ stetig. Dann ist auch die Funktion $h : \mathbb{R} \to \mathbb{R}$ die durch $h(x) := f(x) \cdot g(x)$ definiert ist, stetig. Denn sei $(x_n)_{n\in\mathbb{N}}$ eine Folge, die gegen $\widehat{x}$ konvergiert. Dann gilt

$$\begin{aligned}
\lim_{n\to\infty} h(x_n) &= \lim_{n\to\infty} f(x_n)\cdot g(x_n) && \text{Definition von } h\\
&= \Big(\lim_{n\to\infty} f(x_n)\Big)\cdot\Big(\lim_{n\to\infty} g(x_n)\Big) && \text{Grenzwert von Produkten}\\
&= f\Big(\lim_{n\to\infty} x_n\Big)\cdot g\Big(\lim_{n\to\infty} x_n\Big) && f \text{ und } g \text{ sind stetig}\\
&= f\big(\widehat{x}\big)\cdot g\big(\widehat{x}\big) && \lim_{n\to\infty} x_n = \widehat{x}\\
&= h\big(\widehat{x}\big) && \text{Definition von } h
\end{aligned}$$

5. Mit einer zum letzten Fall analogen Argumentation können wir leicht einsehen, dass alle Funktionen, die ausgehend von den konstanten Funktionen $x \mapsto c$ und der identischen Funktion $x \mapsto x$ mit Hilfe der elementaren Rechen-Operationen "+", "−", "·" und "/" gebildet werden können, stetig sind. Solche Funktionen werden als *rationale Funktionen* bezeichnet. Ein Beispiel für eine solche Funktion ist

   $$x \mapsto \frac{x^3 - 2\cdot x + 1}{x^2 - 1}.$$

   Diese Funktion ist für alle $x \in \mathbb{R}\backslash\{1,-1\}$ definiert und ist nach der obigen Argumentation stetig.

6. Die Funktion $\mathit{sign} : \mathbb{R} \to \mathbb{R}$ sei durch

   $$\mathit{sign}(x) = \begin{cases} +1 & \text{falls } x > 0,\\ 0 & \text{falls } x = 0,\\ -1 & \text{falls } x < 0. \end{cases}$$

   definiert. Diese Funktion ist im Punkt 0 nicht stetig, denn für die Folge $\big(\frac{1}{n}\big)_{n\in\mathbb{N}}$ gilt

   $$\lim_{n\to\infty} \frac{1}{n} = 0, \quad \text{aber} \quad \lim_{n\to\infty} \mathit{sign}\left(\frac{1}{n}\right) = \lim_{n\to\infty} 1 = 1 \neq 0 = \mathit{sign}(0).$$

   Anschaulich ist die Funktion $\mathit{sign} : \mathbb{R} \to \mathbb{R}$ im Punkt 0 nicht stetig, weil sie an dieser Stelle einen Sprung hat. ◇

Definition 44 (Uneigentliche Konvergenz) Wir sagen, dass eine Folge $\big(x_n\big)_{n\in\mathbb{N}}$ gegen Unendlich konvergiert und schreiben

$$\lim_{n\to\infty} x_n = \infty$$

wenn gilt:

$$\forall c \in \mathbb{R} : \exists K \in \mathbb{N} : \forall n \in \mathbb{N} : n > K \rightarrow x_n > c. \qquad \diamond$$

Beispiel: Für die Folge $\big(n\big)_{n\in\mathbb{N}}$ gilt offenbar

$$\lim_{n\to\infty} n = \infty. \qquad \diamond$$

Definition 45 Es sei $f : \mathbb{R} \to \mathbb{R}$ eine Funktion und $\lambda \in \mathbb{R}$. Gilt für jede Folge $\big(x_n\big)_{n\in\mathbb{N}}$

$$\lim_{n\to\infty} x_n = \infty \;\Rightarrow\; \lim_{n\to\infty} f(x_n) = \lambda,$$

dann schreiben wir

$$\lim_{x\to\infty} f(x) = \lambda. \qquad \diamond$$

Beispiel: Es gilt

$$\lim_{x\to\infty} \frac{1}{x} = 0.$$

Beweis: Es sei eine Folge $(x_n)_{n\in\mathbb{N}}$ gegeben, so dass

$$\lim_{n\to\infty} x_n = \infty$$

gilt. Nach Definition der uneigentlichen Konvergenz gegen ∞ gilt dann

$$\forall c \in \mathbb{R} : \exists K \in \mathbb{N} : \forall n \in \mathbb{N} : n > K \to x_n > c \tag{4.1}$$

Wir müssen zeigen, dass gilt:

$$\lim_{n\to\infty} \frac{1}{x_n} = 0.$$

Dies ist nach der Definition des Grenzwerts einer Folge äquivalent zu der Formel

$$\forall \varepsilon \in \mathbb{R}_+ : \exists K \in \mathbb{R} : \forall n \in \mathbb{N} : n > K \to \left|\frac{1}{x_n}\right| < \varepsilon \tag{4.2}$$

Um diese Formel nachzuweisen, nehmen wir an, dass eine Zahl $\varepsilon > 0$ gegeben ist. Wir müssen dann ein K finden, so dass für alle natürlichen Zahlen n, die größer als K sind, die Ungleichung

$$\left|\frac{1}{x_n}\right| < \varepsilon$$

gilt. Dies gelingt uns mit Hilfe der Formel (4.1), denn wenn wir in dieser Formel $c := \frac{1}{\varepsilon}$ definieren, dann finden wir eine Zahl K, so dass für alle natürlichen Zahlen n, die größer als K sind, die Ungleichung

$$x_n > c, \quad \text{also } x_n > \frac{1}{\varepsilon}$$

gilt. Invertieren wir nun diese Ungleichung, so folgt

$$\frac{1}{x_n} < \varepsilon$$

und da andererseits aus $x_n > \frac{1}{\varepsilon}$ und $\varepsilon > 0$ auch $x_n > 0$ und damit $\frac{1}{x_n} > 0$ folgt, haben wir insgesamt

$$\left|\frac{1}{x_n}\right| < \varepsilon$$

für alle $n > K$ gezeigt. □

Es gibt eine alternative Definition der Stetigkeit, die zu der oben gegebenen Definition äquivalent ist. Diese Definition trägt den Namen *ε-δ-Definition der Stetigkeit* und Funktionen, die nach dieser Definition stetig sind, heißen ε-δ-stetig.

Definition 46 (ε-δ-Stetigkeit) Eine Funktion $f : D \to \mathbb{R}$ ist *ε-δ-stetig* im Punkt $\widehat{x}$, wenn gilt:

$$\forall \varepsilon \in \mathbb{R}_+ : \exists \delta \in \mathbb{R}_+ : \forall x \in \mathbb{R} : |x - \widehat{x}| < \delta \to |f(x) - f(\widehat{x})| < \varepsilon. \qquad \diamond$$

Aufgabe 19:

(a) Zeigen Sie, dass jede Funktion, die ε-δ-stetig ist, auch stetig ist.

(b) Zeigen Sie, dass jede stetige Funktion auch ε-δ-stetig ist. ◇

Definition 47 (Allgemeine Stetigkeit)
Eine Funktion $f : D \to \mathbb{R}$ heißt *stetig* genau dann, wenn die Funktion f für alle $\widehat{x} \in D$ stetig ist. ◇

Aufgabe 20: Es sei $f : \mathbb{R} \to \mathbb{R}$. Geben Sie eine sinnvolle Definition für

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

Bemerkung: Wir haben den Begriff des Grenzwerts einer Funktion mit Hilfe von Folgen definiert. Es gibt eine dazu äquivalente ε-δ-Definition des Grenzwerts. Bei dieser Definition sagen wir, dass eine Funktion

$$f : D \to \mathbb{R}$$

an der Stelle $\bar{x}$ den Grenzwert λ hat, wenn

$$\forall \varepsilon \in \mathbb{R}_+ : \exists \delta \in \mathbb{R}_+ : \forall x \in D : |x - \bar{x}| < \delta \to |f(x) - \lambda| < \varepsilon$$

gilt. Genau wie die ε-δ Definition der Stetigkeit äquivalent ist zu dem Stetigkeits-Begriff, den wir mit Hilfe von Folgen definiert haben, ist auch die ε-δ-Definition des Grenzwerts äquivalent zu der Definition des Grenzwerts, die wir früher mit Hilfe von Folgen gegeben haben. Der Beweis dieser Behauptung ist Gegenstand der folgenden Aufgabe.

Aufgabe 21: Es sei

$$f : D \to \mathbb{R}$$

eine reellwertige Funktion. Beweisen Sie die beiden folgenden Behauptungen:

(a) Falls die Formel

$$\forall \varepsilon \in \mathbb{R}_+ : \exists \delta \in \mathbb{R}_+ : \forall x \in D : |x - \bar{x}| < \delta \to |f(x) - \lambda| < \varepsilon$$

richtig ist, dann gilt $\lim\limits_{x \to \bar{x}} f(x) = \lambda$.

(b) Falls $\lim\limits_{x \to \bar{x}} f(x) = \lambda$ gilt, dann gilt auch

$$\forall \varepsilon \in \mathbb{R}_+ : \exists \delta \in \mathbb{R}_+ : \forall x \in D : |x - \bar{x}| < \delta \to |f(x) - \lambda| < \varepsilon.$$ ◇

4.2 Bestimmung von Nullstellen

In der Praxis tritt häufig die Frage auf, ob eine Funktion in einem bestimmten Intervall eine Nullstelle hat. Zusätzlich werden Verfahren benötigt, mit denen eine solche Nullstelle gegebenenfalls berechnet werden kann.

Satz 48 (Zwischenwert-Satz) Die Funktion $f : [a, b] \to \mathbb{R}$ sei stetig. Weiter sei $f(a) \leq 0$ und $f(b) \geq 0$. Dann gibt es ein $x_0 \in [a, b]$, so dass $f(x_0) = 0$ ist.

Beweis: Wir geben ein Verfahren an, mit dem eine Nullstelle berechnet werden kann und weisen dann nach, dass der von diesem Verfahren gelieferte Wert tatsächlich eine Nullstelle der Funktion ist. Das Verfahren, dass wir vorstellen werden, wird in der Literatur als *Verfahren der Intervall-Halbierung* oder auch als *Bisektions-Verfahren* bezeichnet. Das Verfahren folgt dem Paradigma "*Teile und Herrsche*". Im Englischen werden solche Verfahren als "*divide and conquer algorithms*" bezeichnet. Beim Bisektions-Verfahren definieren wir induktiv zwei Folgen $\big(a_n\big)_{n \in \mathbb{N}}$ und $\big(b_n\big)_{n \in \mathbb{N}}$ wie folgt:

I.A.: $n = 1$.

$a_1 := a, \quad b_1 := b.$

I.S.: $n \mapsto n + 1$

Zunächst definieren wir c_n als das arithmetische Mittel von a_n und b_n:

$$c_n := \frac{1}{2} \cdot (a_n + b_n).$$

Dann definieren wir a_{n+1} und b_{n+1} simultan durch Fall-Unterscheidung:

$$\langle a_{n+1}, b_{n+1} \rangle := \begin{cases} \langle a_n, c_n \rangle & \text{falls} \quad f(c_n) > 0 \\ \langle c_n, b_n \rangle & \text{falls} \quad f(c_n) \leq 0. \end{cases}$$

Aus dieser Definition folgt sofort per Induktion, dass für alle $n \in \mathbb{N}$ gilt:

1. $f(a_n) \leq 0$.
2. $f(b_n) \geq 0$.
3. $a_n \leq a_{n+1}, \quad$ die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ ist also monoton steigend.
4. $b_n \geq b_{n+1}, \quad$ die Folge $\big(b_n\big)_{n\in\mathbb{N}}$ ist also monoton fallend.
5. $a_n \leq b_n$.
6. $b_n - a_n = \left(\frac{1}{2}\right)^{n-1} \cdot (b - a)$.

Wir führen hier nur den Nachweis der letzten Behauptung vor, denn diese Behauptung ist am wenigsten offensichtlich.

I.A.: $n = 1$. Es gilt

$$b_1 - a_1 = b - a = \left(\frac{1}{2}\right)^{1-1} \cdot (b - a).$$

I.S.: $n \mapsto n + 1$.

Es ist $c_n = \frac{1}{2} \cdot (a_n + b_n)$. Wir führen eine Fall-Unterscheidung nach dem Vorzeichen von $f(c_n)$ durch.

(a) $f(c_n) > 0$. Dann gilt $a_{n+1} = a_n$ und $b_{n+1} = c_n = \frac{1}{2} \cdot (a_n + b_n)$. Also haben wir

$$\begin{aligned} b_{n+1} - a_{n+1} &= \frac{1}{2} \cdot (a_n + b_n) - a_n \\ &= \frac{1}{2} \cdot (b_n - a_n) \\ &\overset{IV}{=} \frac{1}{2} \cdot \left(\frac{1}{2}\right)^{n-1} \cdot (b - a) \\ &= \left(\frac{1}{2}\right)^{(n+1)-1} \cdot (b - a) \ \surd \end{aligned}$$

(b) $f(c_n) \leq 0$. Jetzt gilt $a_{n+1} = c_n = \frac{1}{2} \cdot (a_n + b_n)$ und $b_{n+1} = b_n$. Also haben wir

$$\begin{aligned} b_{n+1} - a_{n+1} &= b_n - \frac{1}{2} \cdot (a_n + b_n) \\ &= \frac{1}{2} \cdot (b_n - a_n) \\ &\overset{IV}{=} \frac{1}{2} \cdot \left(\frac{1}{2}\right)^{n-1} \cdot (b - a) \\ &= \left(\frac{1}{2}\right)^{(n+1)-1} \cdot (b - a) \quad \surd \end{aligned}$$

Damit ist die Behauptung in beiden Fällen bewiesen.

Aus den Behauptungen 3., 4., und 5. folgt, dass die Folge $\big(b_n\big)_{n\in\mathbb{N}}$ durch a nach unten beschränkt ist, denn es gilt

$a = a_1 \leq \cdots \leq a_{n-1} \leq a_n \leq b_n$, also gilt $a \leq b_n$ für alle $n \in \mathbb{N}$

Da die Folge $\big(b_n\big)_{n\in\mathbb{N}}$ monoton fallend und nach unten beschränkt ist, muss diese Folge nach Satz 21 auch konvergent sein. Wir definieren

$$\widehat{b} := \lim_{n\to\infty} b_n.$$

In analoger Weise sehen wir, dass die Folge $\big(a_n\big)_{n\in\mathbb{N}}$ konvergent ist und definieren

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

Als nächstes weisen wir nach, dass $\widehat{a} = \widehat{b}$ ist. Dazu betrachten wir die Differenz der Grenzwerte:

$$\begin{aligned} \widehat{b} - \widehat{a} &= \left(\lim_{n\to\infty} b_n\right) - \left(\lim_{n\to\infty} a_n\right) \\ &= \lim_{n\to\infty} b_n - a_n \\ &= \lim_{n\to\infty} \left(\frac{1}{2}\right)^{n-1} \cdot (b - a) \\ &= (b - a) \cdot \lim_{n\to\infty} \left(\frac{1}{2}\right)^{n-1} = 0. \end{aligned}$$

Da die Funktion f stetig ist, gilt

$$\lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right) = f(\widehat{a}).$$

Weil $f(a_n) \leq 0$ ist für alle $n \in \mathbb{N}$ folgt dann sofort

$$f(\widehat{a}) \leq 0.$$

Genauso folgt aus der Stetigkeit von f, dass

$$\lim_{n\to\infty} f(b_n) = f\left(\lim_{n\to\infty} b_n\right) = f\big(\widehat{b}\,\big)$$

gilt. Aus $\forall n \in \mathbb{N} : f(b_n) \geq 0$ folgt dann sofort

$$f(\widehat{b}) \geq 0.$$

Da $\widehat{a} = \widehat{b}$ gilt, haben wir natürlich auch $f\big(\widehat{a}\big) = f\big(\widehat{b}\,\big)$. Dann haben wir aber sowohl

$$f\big(\widehat{a}\big) \leq 0 \quad \text{als auch} \quad f\big(\widehat{a}\big) \geq 0$$

und das funktioniert nur, wenn $f\big(\widehat{a}\big) = 0$ ist. Damit haben wir eine Nullstelle von f in dem Intervall $[a, b]$ gefunden. □

```
findZero := procedure(f, a, b, n) {
    assert(a < b, "a has to be less than b");
    assert(f(a) < 0 && 0 < f(b), "we need f($a$) < 0 and f($b$) > 0");
    [ fa, fb ] := [ f(a), f(b) ];
    for (k in [1 .. n]) {
        c := 1/2 * (a + b); fc := f(c);
        if (fc < 0) {
            a := c; fa := fc;
        } else {
            b := c; fb := fc;
        }
    }
    return 1/2 * (a + b);
};
```

Abbildung 4.1: Implementierung des Bisektions-Verfahrens in SetlX.

Das im Beweis des letzten Satzes beschriebene Intervall-Halbierungs-Verfahren lässt sich ohne große Mühe implementieren. Abbildung 4.1 zeigt eine solche Implementierung in der Sprache SetlX. Die Funktion `findZero` erhält vier Argumente:

1. `f` ist die Funktion, deren Nullstelle bestimmt werden soll,
2. `a` ist die linke Intervall-Grenze,
3. `b` ist die rechte Intervall-Grenze und
4. `n` ist die Anzahl der Iterationen, die durchgeführt werden soll.

Zu Beginn testen wir, ob erstens die linke Intervall-Grenze `a` kleiner als die rechte Intervall-Grenze `b` ist und zweitens ob $\texttt{f}(\texttt{a}) < \texttt{f}(\texttt{b})$ ist, denn sonst sind die Voraussetzungen des Zwischenwert-Satzes nicht erfüllt und das Bisektions-Verfahren lässt sich nicht anwenden.

Die Implementierung setzt den oben skizzierten Algorithmus unmittelbar um. Da wir immer nur die beiden letzten Werte der Folgen $(a_n)_n$ und $(b_n)_n$ benötigen, ist es nicht notwendig, die Folgen zu speichern. Es reicht, die Werte a_n und b_n in den Variablen `a` und `b` abzulegen. Wir haben bei der Implementierung außerdem geachtet, dass die Funktion `f` nicht an derselben Stelle mehrfach berechnet wird. Wir erreichen dies, indem wir den Funktionswert, den die Funktion `f` an der Stelle `a` annimmt, in der Variablen `fa` abspeichern. Genauso wird der Funktionswert der Funktion `f` an der Stelle `b` in der Variablen `fb` abgelegt.

Wenn wir dieses Verfahren einsetzen wollen um in einem vorgegeben Intervall nach einer Nullstelle zu suchen, so können wir im Voraus berechnen, wie viele Iterationen zur Erzielung einer geforderten Genauigkeit benötigt werden: Soll die Nullstelle mit einer Genauigkeit von ε bestimmt werden, so muss die Zahl n der Iterationen so gewählt werden, dass

$$\left(\frac{1}{2}\right)^n \cdot (b-a) \leq \varepsilon$$

gilt. Um n zu bestimmen, logarithmieren wir diese Ungleichung und erhalten:

$$
\begin{aligned}
& n \cdot \ln\Big(\frac{1}{2}\Big) + \ln(b-a) \le \ln(\varepsilon) \\
\Leftrightarrow \quad & n \cdot \ln\Big(\frac{1}{2}\Big) \le \ln(\varepsilon) - \ln(b-a) \\
\Leftrightarrow \quad & n \cdot \ln\Big(\frac{1}{2}\Big) \le \ln\left(\frac{\varepsilon}{b-a}\right) \\
\Leftrightarrow \quad & -n \cdot \ln(2) \le \ln\left(\frac{\varepsilon}{b-a}\right) \\
\Leftrightarrow \quad & n \ge -\frac{1}{\ln(2)} \cdot \ln\Big(\frac{\varepsilon}{b-a}\Big) \\
\Leftrightarrow \quad & n \ge \frac{1}{\ln(2)} \cdot \ln\Big(\frac{b-a}{\varepsilon}\Big)
\end{aligned}
$$

Wollen wir beispielsweise die Nullstelle der Funktion $x \mapsto x - \cos(x)$ im Intervall $[0,1]$ auf eine Genauigkeit von $\varepsilon = 10^{-9}$ bestimmen, so finden wir

$$n \ge \frac{\ln\big(10^9\big)}{\ln(2)} = 9 \cdot \frac{\ln(10)}{\ln(2)} \approx 29.89735286,$$

Damit ist klar, dass wir 30 Iterationen des Verfahrens benötigen um die geforderte Genauigkeit zu erreichen. Tabelle 4.1 zeigt die Werte, die a_n und b_n bei der Lösung der Gleichung $x - \cos(x) = 0$ beim Intervall-Halbierungs-Verfahren annehmen. Nach 30 Iterationen weichen die Intervall-Grenzen a_n und b_n um weniger als 10^{-9} voneinander ab.

4.2.1 Die Regula Falsi

Beim Bisektions-Verfahren wird das Interval in jedem Schritt in zwei gleich große Teile zerteilt, denn wir bestimmen den Mittelpunkt des Intervalls $[a,b]$ nach der Formel

$$c = \frac{1}{2} \cdot (a+b).$$

Bei dieser Formel werden die Beträge der Funktionswerte von f an den Stellen a und b überhaupt nicht berücksichtigt. Es liegt nahe, die Beträge der Funktionswerte in die Formel mit einfließen zu lassen, denn wenn beispielsweise $|f(a)|$ wesentlich kleiner $|f(b)|$ ist, dann ist zu vermuten, dass die Nullstelle von f näher an a als an b liegt.

Betrachten wir beispielsweise die Tabelle 4.1, so sehen wir, dass in dem 24-ten Iterations-Schritt die Funktion $x \mapsto x - \cos(x)$ an der rechten Intervall-Grenze b_n den Wert $\approx 7.7 \cdot 10^{-9}$ hat, während die Funktion an der linken Intervall-Grenze a_n den Wert $-9.2 \cdot 10^{-8}$ hat. Der Betrag dieses Wertes ist mehr als 10 mal so groß wie der Wert an der rechten Intervall-Grenze. Folglich liegt es nahe zu vermuten, dass die Nullstelle näher an der rechten Intervall-Grenze liegt als an der linken. Die weitere Berechnung bestätigt diese Vermutung auch, denn die rechte Intervall-Grenze ändert sich bei den nächsten drei Iterationen nicht. Wie können wir diese Beobachtung ausnutzen? Anstatt in der Formel $c_n = \frac{1}{2} \cdot (a_n + b_n)$ die Punkte a und b unabhängig von den Funktionswerten gleich stark zu gewichten, könnten wir eine Intervall-Grenze dann stärker gewichten, wenn der Funktionswert dort kleiner ist, weil wir dann vermuten würden, dass dieser Punkt schon näher an der Nullstelle liegt. Eine naheliegende Idee ist daher, die Punkte a und b mit den Beträgen der reziproken Funktionswerte zu gewichten, denn die werden um so größer, je kleiner der Funktionswert ist. Dieser Ansatz führt auf die Formel

$$c = \frac{\dfrac{1}{|f(a)|} \cdot a + \dfrac{1}{|f(b)|} \cdot b}{\dfrac{1}{|f(a)|} + \dfrac{1}{|f(b)|}} = \frac{|f(b)| \cdot a + |f(a)| \cdot b}{|f(a)| + |f(b)|}$$

Wir erhalten dieselbe Formel, wenn wir c dadurch bestimmen, dass wir eine Gerade durch die

n	a_n	b_n	$f(a_n)$	$f(b_n)$
0:	0.000000000	1.000000000	-1.00000000e+00	4.59697694e-01
1:	0.500000000	1.000000000	-3.77582562e-01	4.59697694e-01
2:	0.500000000	0.750000000	-3.77582562e-01	1.83111311e-02
3:	0.625000000	0.750000000	-1.85963120e-01	1.83111311e-02
4:	0.687500000	0.750000000	-8.53349462e-02	1.83111311e-02
5:	0.718750000	0.750000000	-3.38793724e-02	1.83111311e-02
6:	0.734375000	0.750000000	-7.87472546e-03	1.83111311e-02
7:	0.734375000	0.742187500	-7.87472546e-03	5.19571174e-03
8:	0.738281250	0.742187500	-1.34514975e-03	5.19571174e-03
9:	0.738281250	0.740234375	-1.34514975e-03	1.92387278e-03
10:	0.738281250	0.739257813	-1.34514975e-03	2.89009147e-04
11:	0.738769531	0.739257813	-5.28158434e-04	2.89009147e-04
12:	0.739013672	0.739257813	-1.19596671e-04	2.89009147e-04
13:	0.739013672	0.739135742	-1.19596671e-04	8.47007314e-05
14:	0.739074707	0.739135742	-1.74493466e-05	8.47007314e-05
15:	0.739074707	0.739105225	-1.74493466e-05	3.36253482e-05
16:	0.739074707	0.739089966	-1.74493466e-05	8.08791474e-06
17:	0.739082336	0.739089966	-4.68073746e-06	8.08791474e-06
18:	0.739082336	0.739086151	-4.68073746e-06	1.70358327e-06
19:	0.739084244	0.739086151	-1.48857844e-06	1.70358327e-06
20:	0.739084244	0.739085197	-1.48857844e-06	1.07502077e-07
21:	0.739084721	0.739085197	-6.90538266e-07	1.07502077e-07
22:	0.739084959	0.739085197	-2.91518116e-07	1.07502077e-07
23:	0.739085078	0.739085197	-9.20080247e-08	1.07502077e-07
24:	0.739085078	0.739085138	-9.20080247e-08	7.74702466e-09
25:	0.739085108	0.739085138	-4.21305004e-08	7.74702466e-09
26:	0.739085123	0.739085138	-1.71917379e-08	7.74702466e-09
27:	0.739085130	0.739085138	-4.72235666e-09	7.74702466e-09
28:	0.739085130	0.739085134	-4.72235666e-09	1.51233399e-09
29:	0.739085132	0.739085134	-1.60501133e-09	1.51233399e-09
30:	0.739085133	0.739085134	-4.63386709e-11	1.51233399e-09

Tabelle 4.1: Die ersten 30 Schritte des Bisektions-Verfahrens zur Lösung von $x - \cos(x) = 0$.

Punkte $\langle a, f(a)\rangle$ und $\langle b, f(b)\rangle$ legen und c als den Punkt festsetzen, bei dem diese Gerade die x-Achse scheidet. Die Gleichung für eine Gerade $g(x)$ hat die Form

$$g(x) = \alpha \cdot x + \beta.$$

Setzen wir hier für x den Wert a und für $g(x)$ den Wert $f(a)$ ein, so erhalten wir die Gleichung

$$f(a) = \alpha \cdot a + \beta. \tag{4.3}$$

Analog erhalten wir die Gleichung

$$f(b) = \alpha \cdot b + \beta \tag{4.4}$$

wenn wir für x den Wert b und für $g(x)$ den Wert $f(b)$ einsetzen. Subtrahieren wir die beiden Gleichungen voneinander, so verschwindet die Unbekannte β und wir haben

$$f(b) - f(a) = \alpha \cdot (b - a), \quad \text{also} \quad \alpha = \frac{f(b) - f(a)}{b - a}.$$

Setzen wir diesen Wert für α in die Gleichung 4.3 ein, so ergibt sich

$$f(a) = \frac{f(b) - f(a)}{b - a} \cdot a + \beta.$$

Wir lösen diese Gleichung nach β auf und erhalten

$$\beta = \frac{f(a) \cdot (b - a) - \big(f(b) - f(a)\big) \cdot a}{b - a} = \frac{f(a) \cdot b - f(b) \cdot a}{b - a}.$$

Wir bestimmen c aus der Forderung, dass $g(c) = 0$ ist, also

$$0 = \alpha \cdot c + \beta$$

$$\Leftrightarrow \quad c = -\frac{\beta}{\alpha}$$

Setzen wir hier die eben berechneten Werte für α und β ein, so erhalten wir

$$c = -\frac{\dfrac{f(a) \cdot b - f(b) \cdot a}{b - a}}{\dfrac{f(b) - f(a)}{b - a}} = \frac{f(b) \cdot a - f(a) \cdot b}{f(b) - f(a)}$$

Falls nun $f(a) < 0$ und $f(b) > 0$ ist, gilt $-f(a) = |f(a)|$ und $f(b) = |f(b)|$. Setzen wir diese Werte in die obige Gleichung ein, so erhalten wir

$$c = \frac{|f(b)| \cdot a + |f(a)| \cdot b}{|f(a)| + |f(b)|}$$

und das ist die gleiche Formel, die wir auch oben schon abgeleitet hatten. Abbildung 4.2 zeigt die graphische Bestimmung von c als Schnittpunkt der Geraden mit der x-Achse.

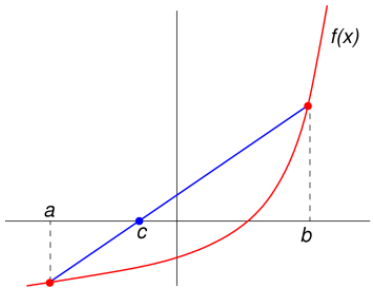

Abbildung 4.2: Die Regula-Falsi zur Nullstellen-Bestimmung.

Das Verfahren, das mit dieser Formel arbeitet, ist unter dem Namen *Regula Falsi* bekannt und sieht genauso aus wie das Bisektions-Verfahren, nur dass wir für c jetzt die oben abgeleitete Formel verwenden:

I.A.: $n = 1$.

$a_1 := a, \quad b_1 := b.$

I.S.: $n \mapsto n + 1$

$$c_n := \frac{|f(b_n)| \cdot a_n + |f(a_n)| \cdot b_n}{|f(a_n)| + |f(b_n)|}.$$

Dann definieren wir a_{n+1} und b_{n+1} durch Fall-Unterscheidung:

$$\langle a_{n+1}, b_{n+1} \rangle := \begin{cases} \langle a_n, c_n \rangle & \text{falls} \quad f(c_n) > 0 \\ \langle c_n, b_n \rangle & \text{falls} \quad f(c_n) \leq 0. \end{cases}$$

Ähnlich wie beim Beweis des Zwischenwert-Satzes lässt sich zeigen, dass die Folge $\big(a_n\big)_{n \in \mathbb{N}}$ monoton steigend ist, während die Folge $\big(b_n\big)_{n \in \mathbb{N}}$ monoton fallend ist. Da die Folgen überdies beschränkt

sind, denn a_n ist immer kleiner als b und b_n ist immer größer als a, konvergieren beide Folgen. Allerdings ist nicht garantiert, dass a_n und b_n gegen den gleichen Grenzwert konvergieren! Es lässt sich lediglich zeigen, dass entweder a_n oder b_n gegen eine Nullstelle der Funktion f konvergiert. Um das Verfahren experimentell untersuchen zu können, implementieren wir es. Abbildung 4.3 zeigt die Implementierung der Methode *findZero*(). Diese Implementierung ist weitgehend analog zu der Implementierung des Bisektions-Verfahrens. Es gibt eigentlich nur zwei wesentliche Unterschiede:

1. In Zeile 6 berechnen wir `c` nun nach der Formel
$$c := \frac{f(b) \cdot a - f(a) \cdot b}{f(b) - f(a)}.$$
Beim Bisektions-Verfahren hatten wir hier die Formel
$$c := \frac{1}{2} \cdot (a + b)$$
verwendet.

2. Bei der Rückgabe des berechneten Wertes in Zeile 14 bzw. 16 ist es erforderlich, die Beträge der Funktionswerte an den Intervall-Grenzen `a` und `b` zu vergleichen, denn wir wissen nicht, ob die Folge $(a_n)_n$ oder die Folge $(b_n)_n$ gegen die Nullstelle von f konvergiert. Wir geben daher als Ergebnis die Intervall-Grenze zurück, für die der Betrag des Funktionswertes am kleinsten ist. Da wir wissen, dass der Funktionswert an der linken Intervall-Grenze immer kleiner als 0 ist, erhalten wir dort den Betrag der Funktion f, indem wir dem Funktionswert das Minuszeichen vorstellen.

```
1  regulaFalsi := procedure(f, a, b, n) {
2      assert(a < b, "Error: !(a < b)");
3      assert(f(a) < 0 && f(b) > 0, "Error: !(f(a) < 0 && f(b) > 0)");
4      fa := f(a); fb := f(b);
5      for (i in [1 .. n]) {
6          c  := (fb * a - fa * b) / (fb - fa); fc := f(c);
7          if (fc <= 0) {
8              a := c; fa := fc;
9          } else {
10             b := c; fb := fc;
11         }
12     }
13     if (-fa < fb) {
14         return a;
15     } else {
16         return b;
17     }
18 };
```

Abbildung 4.3: Implementierung der Regula Falsi in SetlX.

Tabelle 4.2 zeigt die ersten 12 Iterations-Schritte, wenn die Regula Falsi zur Berechnung der Nullstelle von $x - \cos(x)$ eingesetzt wird. Wir sehen, dass wir bereits im 9-ten Schritt dieselbe Genauigkeit erreicht haben, für die wir mit dem Bisektions-Verfahren 30 Schritte benötigt haben. Wir sehen auch, dass die rechte Intervall-Grenze immer konstant bleibt. Es sieht so aus, als ob wir mit der Regula Falsi ein Verfahren gefunden hätten, dass dem Bisektions-Verfahren überlegen wäre. Die nächste Aufgabe zeigt Ihnen jedoch, dass dem Verfahren eine ganz wichtige Eigenschaft fehlt, die das Bisektions-Verfahren besitzt: Das Verfahren ist nicht robust! Es gibt Funktionen,

bei denen die Regula Falsi zur Nullstellen-Bestimmung **wesentlich mehr** Iterationen benötigt als das Bisektions-Verfahren.

n	a_n	b_n	$f(a_n)$	$f(b_n)$
1:	0.000000000	1.000000000	-1.00000000e+00	4.59697694e-01
2:	0.685073357	1.000000000	-8.92992765e-02	4.59697694e-01
3:	0.736298997	1.000000000	-4.66003904e-03	4.59697694e-01
4:	0.738945356	1.000000000	-2.33925666e-04	4.59697694e-01
5:	0.739078130	1.000000000	-1.17191742e-05	4.59697694e-01
6:	0.739084782	1.000000000	-5.87046549e-07	4.59697694e-01
7:	0.739085115	1.000000000	-2.94066726e-08	4.59697694e-01
8:	0.739085132	1.000000000	-1.47305551e-09	4.59697694e-01
9:	0.739085133	1.000000000	-7.37890543e-11	4.59697694e-01
10:	0.739085133	1.000000000	-3.69623245e-12	4.59697694e-01
11:	0.739085133	1.000000000	-1.85199566e-13	4.59697694e-01
12:	0.739085133	1.000000000	-9.23913723e-15	4.59697694e-01

Tabelle 4.2: Die ersten 12 Schritte der Regula Falsi zur Lösung von $x - \cos(x) = 0$.

Aufgabe 22: Verwenden Sie die Regula Falsi zur Lösung der Gleichung

$$x^4 - 1 = 0.$$

Starten Sie mit dem Intervall $[0, 10]$. Zeigen Sie, dass für alle natürlichen Zahlen n mit $n \leq 1000$ die folgende Ungleichung für die linke Intervall-Grenze a_n gilt:

$$a_n \leq \frac{n}{1000}.$$

Die Lösung der Gleichung $x^4 - 1 = 0$ in dem Intervall ist $x = 1$. Aus der zu zeigenden Ungleichung kann beispielsweise gefolgert werden, dass $a_{100} \leq 0.1$ gilt. Der mit dem obigen Programm ermittelte Wert für a_{100} ist $a_{100} = 0.0985146583$. In diesem Fall hat die Regula Falsi also selbst nach 100 Iterationen nicht eine einzige korrekte Stelle im Ergebnis berechnen können! ◇

Lösung: Wir zeigen durch vollständige Induktion über n, dass für alle $n \leq 1000$ zum einen die Ungleichung $a_n \leq n \cdot 10^{-3}$ gilt und dass zum anderen b_n konstant ist, es gilt $b_n = 10$.

I.A.: $n = 1$. Es gilt

$$a_1 = 0 \leq 1 \cdot 10^{-3} \quad \text{und} \quad b_1 = 10.$$

I.S.: $n \mapsto n + 1$.

Die Funktion $f := (x \mapsto x^4 - 1)$ ist für nichtnegative Zahlen monoton steigend, dass heißt aus $0 \leq u \leq v$ folgt auch $f(u) \leq f(v)$. Es gilt

$$f(n \cdot 10^{-3}) = n^4 \cdot 10^{-12} - 1 \quad \text{und} \quad f(10) = 10^4 - 1.$$

Nach Induktions-Voraussetzung können wir a_n durch $n \cdot 10^{-3}$ abschätzen und aufgrund der Monotonie von f können wir dann $f(a_n)$ durch $f(n \cdot 10^{-3})$ abschätzen. Wenden wir daher für $a'_n = n \cdot 10^{-3}$ und $b_n = 10$ die Regula Falsi an um eine Näherung c'_n für die Nullstelle von f zu berechnen, so wird c'_n größer sein als der wahre Wert von c_n, der in dem Algorithmus tatsächlich auftritt. Es gilt:

$$\begin{aligned}
c'_n &= \frac{f(b_n)\cdot a'_n - f(a'_n)\cdot b_n}{f(b_n) - f(a'_n)} \\
&= \frac{f(10)\cdot n\cdot 10^{-3} - f\big(n\cdot 10^{-3}\big)\cdot 10}{f(10) - f\big(n\cdot 10^{-3}\big)} \\
&= \frac{\big(10^4-1\big)\cdot n\cdot 10^{-3} - \big(n^4\cdot 10^{-12}-1\big)\cdot 10}{10^4 - 1 - n^4\cdot 10^{-12} + 1} \\
&= 10^{-4}\cdot \frac{10\cdot n - n\cdot 10^{-3} - n^4\cdot 10^{-11} + 10}{1 - n^4\cdot 10^{-16}} \\
&= 10^{-3}\cdot \frac{n + 1 - n\cdot 10^{-4} - n^4\cdot 10^{-12}}{1 - n^4\cdot 10^{-16}}
\end{aligned}$$

Wir untersuchen nun, für welche natürlichen Zahlen n die Ungleichung $c'_n \leq 10^{-3}\cdot(n+1)$ gilt.

$$\begin{aligned}
& c'_n \leq 10^{-3}\cdot(n+1) \\
\Leftrightarrow\ & 10^{-3}\cdot\frac{n+1-n\cdot 10^{-4}-n^4\cdot 10^{-12}}{1-n^4\cdot 10^{-16}} \leq 10^{-3}\cdot(n+1) \\
\Leftrightarrow\ & n+1-n\cdot 10^{-4}-n^4\cdot 10^{-12} \leq (n+1)\cdot\big(1-n^4\cdot 10^{-16}\big) \\
\Leftrightarrow\ & n+1-n\cdot 10^{-4}-n^4\cdot 10^{-12} \leq (n+1)-(n+1)\cdot n^4\cdot 10^{-16} \\
\Leftrightarrow\ & -n\cdot 10^{-4}-n^4\cdot 10^{-12} \leq -(n+1)\cdot n^4\cdot 10^{-16} \\
\Leftrightarrow\ & n\cdot 10^{-4}+n^4\cdot 10^{-12} \geq (n+1)\cdot n^4\cdot 10^{-16} \\
\Leftarrow\ & n^4\cdot 10^{-12} \geq (n+1)\cdot n^4\cdot 10^{-16} \\
\Leftrightarrow\ & 1 \geq (n+1)\cdot 10^{-4} \\
\Leftrightarrow\ & 10^4 \geq n+1 \\
\Leftrightarrow\ & n \leq 9999
\end{aligned}$$

Solange $n < 1000$ ist, gilt also sicher $c'_n < 1$ und damit ist $f(c'_n)$ negativ. Daher gilt

$$a_{n+1} \leq a'_{n+1} = c'_n \leq 10^{-3}\cdot n \text{ und } b_{n+1} = b_n = 10. \qquad \square$$

4.2.2 Das Sekanten-Verfahren

Ein Problem bei der Regula Falsi scheint darin zu liegen, dass häufig eine Intervall-Grenze während der gesamten Iteration fest bleibt. Dies war schon bei der Bestimmung der Nullstelle der Funktion $x \mapsto x - \cos(x)$ der Fall. Eine Möglichkeit, dieses Problem zu umgehen besteht darin, dass wir anstatt eine Folge von Intervallen $\big([a_n, b_n]\big)_{n\in\mathbb{N}}$ zu bilden, einfach nur eine Folge von Punkten $\big(x_n\big)_{n\in\mathbb{N}}$ konstruieren. Den Punkt x_{n+1} bestimmen wir, indem wir durch die Punkte x_{n+1} und x_n eine Gerade legen und dann x_n als den Schnittpunkt dieser Geraden mit der x-Achse bestimmen. Das führt auf dieselbe Formel wie bei der Regula-Falsi, wir setzen nämlich

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

Dann brauchen wir nur noch zwei Startwerte x_1 und x_2 und die Rechnung kann los gehen. Abbildung 4.4 zeigt eine Implementierung des Sekanten-Verfahrens in SetlX. Testen wir dieses Programm mit der Funktion $x \mapsto x - \cos(x)$, so erhalten wir die in Tabelle 4.3 gezeigten Werte.

Wir sehen, dass jetzt bereits 7 Iterationen ausreichen, um die Lösung der Gleichung mit der geforderten Genauigkeit zu berechnen. Es sieht also so aus, als ob das Sekanten-Verfahren den anderen Verfahren überlegen ist. In der Tat kann gezeigt werden, dass das Sekanten-Verfahren, **wenn** es denn konvergiert, schneller konvergiert als die anderen Verfahren. Wir werden das später präzisieren. Das Problem ist, dass das Sekanten-Verfahren gar nicht immer konvergiert. Betrachten wir beispielsweise die Funktion

$$x \mapsto \frac{2}{x^2+1} - 1.$$

Diese Funktion hat bei $x = 1.0$ eine Nullstelle. Mit den Startwerten $a = 0$ und $b = 5.0$ produziert unser Programm die in Tabelle 4.4 gezeigten Werte.

```
1  secant := procedure(f, a, b, digits) {
2      fa := f(a);
3      fb := f(b);
4      while (abs(b - a) > (1/10)**(digits + 1)) {
5          c := (fb * a - fa * b) / (fb - fa);
6          a := b; b := c; fa := fb; fb := f(c);
7      }
8      return b;
9  };
```

Abbildung 4.4: Implementierung des Sekanten-Verfahrens in SetlX.

n	x_n	$f(x_n)$
1:	10.00000000000	+1.08390715e+01
2:	0.84466083134	+1.80675899e-01
3:	0.68946400911	-8.21230732e-02
4:	0.73796206792	-1.87910933e-03
5:	0.73909776898	+2.11474296e-05
6:	0.73908513008	-5.24715686e-09
7:	0.73908513322	-1.46275678e-14

Tabelle 4.3: Lösung der Gleichung $x - \cos(x) = 0$ mit dem Sekanten-Verfahren.

n	x_n	$f(x_n)$
1:	+5.000000e+00	-0.923076923
2:	+2.600000e+00	-0.742268041
3:	-7.252631e+00	-0.962687030
4:	+3.577905e+01	-0.998438891
5:	-1.165962e+03	-0.999998529
6:	+7.693592e+05	-1.000000000
7:	-5.237534e+11	-1.000000000
8:	+1.550094e+23	-1.000000000
9:	∞	-1.000000000

Tabelle 4.4: Divergenz des Sekanten-Verfahrens bei der Lösung von $\frac{2}{x^2+1} - 1 = 0$.

4.2.3 Das Illinois-Verfahren

Von den bisher vorgestellten Verfahren ist nur das Bisektions-Verfahren wirklich robust. Bei der Regula-Falsi ist das Problem, dass eine Intervall-Grenze stehen bleiben kann. Am Beispiel der Funktion $x \mapsto x^4 - 1$ haben wir gesehen, dass dies zu einer sehr langsamen Konvergenz führen kann. Beim Sekanten-Verfahren hatten wir dieses Problem behoben, aber dort kann es in ungünstigen Fällen passieren, dass das Verfahren überhaupt nicht mehr konvergiert. Das *Illinois-Verfahren* [8] versucht die Konvergenz der Regula Falsi auf andere Weise zu beschleunigen. Die Idee des Verfahrens ist eigentlich sehr naheliegend: Wenn bei der Regula Falsi eine der Intervall-Grenzen über zwei oder mehr Schritte konstant bleibt, dann wird der Funktionswert an der betreffenden Intervall-Grenze halbiert, so dass der Einfluss dieses Wertes bei der Berechnung der nächsten Näherung c_n nach der Formel

$$c_n := \frac{f(b_n) \cdot a_n - f(a_n) \cdot b_n}{f(b_n) - f(a_n)}$$

gemindert wird. Nehmen wir o.B.d.A. an, dass $f(a) < 0$ und $0 < f(b)$ ist, so führt das zur folgenden Definition der Folgen $\big(a_n\big)_{n\in\mathbb{N}}$ und $\big(b_n\big)_{n\in\mathbb{N}}$:

I.A.: $n = 1$. Wir setzen

$$a_1 := a, \quad b_1 := b, \quad \alpha_1 := 1, \quad \text{und} \quad \beta_1 := 1.$$

Die Werte α_n und β_n sind dabei Gewichtungs-Faktoren, die wir später benötigen.

I.S.: $n \mapsto n + 1$. Wir definieren ähnlich wie bei der Regula Falsi den Wert c_n als

$$c_n := \frac{\beta_n \cdot f(b_n) \cdot a_n - \alpha_n \cdot f(a_n) \cdot b_n}{\beta_n \cdot f(b_n) - \alpha_n \cdot f(a_n)}.$$

Der Unterschied zur Regula Falsi liegt in den Gewichtungs-Faktoren α_n und β_n. Die Werte für a_{n+1} und b_{n+1} werden durch dieselbe Fall-Unterscheidung wie bei der Regula Falsi festgelegt:

$$\langle a_{n+1}, b_{n+1}\rangle := \begin{cases} \langle a_n, c_n\rangle & \text{falls} \quad f(c_n) > 0 \\ \langle c_n, b_n\rangle & \text{falls} \quad f(c_n) \leq 0. \end{cases}$$

Falls wir nun feststellen, dass $b_{n+1} = b_{n-1}$ ist, so hat sich der Wert der rechten Intervall-Grenze während der letzten zwei Iterationen nicht geändert. Wir wollen diesen Wert daher beim nächsten Iterations-Schritt schwächer gewichten und setzen deshalb in diesem Fall

$$\beta_{n+1} = \frac{1}{2} \cdot \beta_n \quad \text{und} \quad \alpha_{n+1} := 1.$$

Ist umgekehrt $a_{n+1} = a_{n-1}$, so hat sich der Wert der linken Intervall-Grenze nicht geändert. Wir gewichten daher die linke Intervall-Grenze beim nächsten Iterations-Schritt schwächer und setzen

$$\beta_{n+1} = 1 \quad \text{und} \quad \alpha_{n+1} := \frac{1}{2} \cdot \alpha_n.$$

Die Umsetzung dieses Verfahrens sehen Sie in Abbildung 4.5. In den Variablen `oldA1` und `oldB1` speichern wir die Werte von a_{n-1} und b_{n-1}, in den Variablen `oldA2` und `oldB2` sind die Werte a_{n-2} und b_{n-2} gespeichert. Wir initialisieren diese Werte mit `om`, denn `om` bezeichnet in SetlX den undefinierten Wert. Falls wir in Zeile 19 feststellen, dass der Wert von $a_n = a_{n-2}$ ist, dann setzen wir den Wert α_{n+1} auf $\alpha_n/2$. Analog testen wir in Zeile 14, ob $b_n = b_{n-2}$ ist und setzen gegebenenfalls β_{n+1} auf $\beta_n/2$.

```
illinois := procedure(f, a, b, n) {
    assert(a < b, "a has to be less than b");
    assert(f(a) < 0 && 0 < f(b), "We need f(a) < 0 and 0 < f(b)!");
    [ fa, fb ] := [ f(a), f(b) ];
    oldA1 := om; oldB1 := om;
    oldA2 := om; oldB2 := om;
    alpha := 1; beta := 1;
    for (k in [1 .. n]) {
        c  := (beta * fb * a - alpha * fa * b) / (beta * fb - alpha * fa);
        fc := f(c);
        if (fc < 0) {
            a := c; fa := fc; alpha := 1;
            if (oldB2 == b) {
                beta /= 2;
            }
        } else if (fc > 0) {
            b := c; fb := fc; beta := 1;
            if (oldA2 == a) {
                alpha /= 2;
            }
        } else {
            return c;
        }
        oldA2 := oldA1; oldB2 := oldB1;
        oldA1 := a;     oldB1 := b;
    }
    return (a + b) / 2;
};
```

Abbildung 4.5: Implementierung des Illinois-Verfahrens zur Berechnung von Nullstellen.

Kapitel 5

Differenzierbare Funktionen

In diesem Kapitel kommen wir zum Kern der Analysis und führen den Begriff der *Ableitung* einer Funktion ein. Dies ist der wichtigste Begriff in der Analysis. Der Begriff wurde unabhängig von Isaac Newton und Gottfried Wilhelm Leibniz gefunden, die folgende formale Definition der Ableitung geht auf Augustin-Louis Cauchy zurück.

5.1 Der Begriff der Ableitung

Definition 49 (Ableitung) Es sei $D \subseteq \mathbb{R}$ ein Intervall. Eine Funktion $f : D \to \mathbb{R}$ ist im Punkt $\widehat{x} \in D$ *differenzierbar*, wenn der Grenzwert

$$\lim_{h \to 0} \frac{f(\widehat{x}+h) - f(\widehat{x})}{h}$$

existiert. In diesem Fall definieren wir

$$\frac{d\,f}{dx}(\widehat{x}) = \lim_{h \to 0} \frac{f(\widehat{x}+h) - f(\widehat{x})}{h}.$$

Wir bezeichnen den Wert $\frac{d\,f}{dx}(\widehat{x})$ als die *Ableitung* der Funktion f an der Stelle $\widehat{x}$. Gelegentlich werden wir für die Ableitung auch die Schreibweise $f'(\widehat{x})$ verwenden. ◇

Bemerkung: Beachten Sie, dass wir in der obigen Definition den Ausdruck

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

als Funktion von h auffassen. Dieser Ausdruck wird auch als *Differential-Quotient* bezeichnet. Er gibt die Steigung einer Sekante an, die die Funktion $x \mapsto f(x)$ in den Punkten $\widehat{x}$ und $\widehat{x}+h$ schneidet. Definieren wir

$$r(h) := f(\widehat{x}+h) - f(\widehat{x}) - h \cdot \frac{df}{dx}(\widehat{x}),$$

so gilt einerseits

$$\lim_{h \to 0} \frac{r(h)}{h} = \lim_{h \to 0} \frac{f(\widehat{x}+h) - f(\widehat{x})}{h} - \frac{df}{dx}(\widehat{x}) = \frac{df}{dx}(\widehat{x}) - \frac{df}{dx}(\widehat{x}) = 0,$$

und andererseits haben wir

$$f(\widehat{x}+h) = f(\widehat{x}) + h \cdot \frac{df}{dx}(\widehat{x}) + r(h).$$

Die Funktion $r(h)$ ist also der Fehler, der bei der linearen Approximation entsteht. $\diamond$

Bemerkung: Falls die Funktion f im Punkt $\widehat{x}$ differenzierbar ist, dann ist die Funktion dort auch stetig, denn es gilt

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

und $\lim\limits_{h\to 0} f(\widehat{x}+h) = f(\widehat{x})$ heißt gerade, dass f im Punkt $\widehat{x}$ stetig ist. $\square$

Beispiele:

1. Die konstante Funktion $f := (x \mapsto c)$ hat überall die Ableitung 0, denn es gilt

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

2. Die identische Funktion $id := (x \mapsto x)$ hat überall die Ableitung 1, denn es gilt:

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

3. Die Funktion $abs := (x \mapsto |x|)$, die den Absolutbetrag berechnet, ist im Punkte $\widehat{x} = 0$ nicht differenzierbar. Wir zeigen, dass der Grenzwert

$$\lim_{h\to 0} \frac{abs(h)-abs(0)}{h}$$

nicht existiert. Dazu betrachten wir zunächst die Folge $\left(\frac{1}{n}\right)_{n\in\mathbb{N}}$. Nehmen wir an, dass dieser Grenzwert existiert und den Wert a hat. Da

$$\lim_{n\to\infty} \tfrac{1}{n} = 0$$

ist, müsste nach Definition des Grenzwerts dann gelten:

$$a = \lim_{h\to 0} \frac{abs(h)-abs(0)}{h} = \lim_{n\to\infty} \frac{abs(\frac{1}{n})}{\frac{1}{n}} = \lim_{n\to\infty} \frac{\frac{1}{n}}{\frac{1}{n}} = 1.$$

Betrachten wir andererseits die Folge $\left(-\frac{1}{n}\right)_{n\in\mathbb{N}}$ und berücksichtigen, dass diese Folge ebenfalls gegen 0 konvergiert, so erhalten wir

$$a = \lim_{h\to 0} \frac{abs(h)-abs(0)}{h} = \lim_{n\to\infty} \frac{abs(-\frac{1}{n})}{-\frac{1}{n}} = \lim_{n\to\infty} \frac{\frac{1}{n}}{-\frac{1}{n}} = -1.$$

Da a nicht gleichzeitig die Werte $+1$ und -1 annehmen kann, müssen wir folgern, dass die Funktion abs an der Stelle $\widehat{x} = 0$ nicht differenzierbar ist. $\diamond$

Satz 50 (Ableitungs-Regeln) Es seien $f : D \to \mathbb{R}$ und $g : D \to \mathbb{R}$ Funktionen, die im Punkt $\widehat{x}$ differenzierbar sind. Dann gilt:

1. Die Funktion $f + g := \big(x \mapsto f(x) + g(x)\big)$ ist im Punkt $\widehat{x}$ differenzierbar und es gilt:
$$(f+g)'(\widehat{x}) = f'(\widehat{x}) + g'(\widehat{x}).$$
2. Die Funktion $f - g := \big(x \mapsto f(x) - g(x)\big)$ ist im Punkt $\widehat{x}$ differenzierbar und es gilt:
$$(f-g)'(\widehat{x}) = f'(\widehat{x}) - g'(\widehat{x}).$$
3. Die Funktion $f \cdot g := \big(x \mapsto f(x) \cdot g(x)\big)$ ist im Punkt $\widehat{x}$ differenzierbar und es gilt die Produkt-Regel:
$$(f \cdot g)'(\widehat{x}) = f'(\widehat{x}) \cdot g(\widehat{x}) + f(\widehat{x}) \cdot g'(\widehat{x}).$$
4. Ist $g(\widehat{x}) \neq 0$, dann ist die Funktion $\dfrac{f}{g} := \left(x \mapsto \dfrac{f(x)}{g(x)}\right)$ im Punkt $\widehat{x}$ differenzierbar und es gilt die Quotienten-Regel:
$$\left(\frac{f}{g}\right)'(\widehat{x}) = \frac{f'(\widehat{x}) \cdot g(\widehat{x}) - f(\widehat{x}) \cdot g'(\widehat{x})}{g(\widehat{x})^2}.$$

Beweis: Wir zeigen nur die Produkt-Regel. Es gilt:

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

Dabei haben wir im letzten Schritt ausgenutzt, dass eine differenzierbare Funktion auch stetig ist. Daher gilt

$$\lim_{h \to 0} g(\widehat{x}+h) = g(\widehat{x}).$$ □

Aufgabe 23: Zeigen Sie: Ist die Funktion g im Punkt $\widehat{x}$ differenzierbar und gilt $g(\widehat{x}) \neq 0$, so ist auch die Funktion $\dfrac{1}{g} := \left(x \mapsto \dfrac{1}{g(x)}\right)$ im Punkt $\widehat{x}$ differenzierbar und es gilt

$$\left(\frac{1}{g}\right)'(\widehat{x}) = -\frac{g'(\widehat{x})}{g(\widehat{x})^2}.$$

Folgern Sie aus diesem Ergebnis die Quotienten-Regel. ◇

Satz 51 (Ketten-Regel) Die Funktionen $f : \mathbb{R} \to \mathbb{R}$ sei differenzierbar im Punkt $\widehat{x} \in \mathbb{R}$ und die Funktion $g : \mathbb{R} \to \mathbb{R}$ sei differenzierbar im Punkt $\widehat{y} = f(\widehat{x})$. Dann ist auch die Funktion

$$g \circ f := \big(x \mapsto g(f(x))\big)$$

im Punkt $\widehat{x}$ differenzierbar und es gilt

$$(g \circ f)'(\widehat{x}) = g'(f(\widehat{x})) \cdot f'(\widehat{x}).$$

Beweis: Aus der Differenzierbarkeit von f und g folgt, dass es Funktionen $r_1(h)$ und $r_2(h)$ gibt, so dass gilt:

1. $f(\widehat{x} + h) = f(\widehat{x}) + h \cdot f'(\widehat{x}) + r_1(h) \quad$ mit $\lim\limits_{h \to 0} \dfrac{r_1(h)}{h} = 0,$
2. $g(\widehat{y} + h) = g(\widehat{y}) + h \cdot g'(\widehat{y}) + r_2(h) \quad$ mit $\lim\limits_{h \to 0} \dfrac{r_2(h)}{h} = 0.$

Damit finden wir für den Differential-Quotienten der Funktion $g \circ f$ im Punkt $\widehat{x}$:

$$\begin{aligned}
& \frac{(g \circ f)(\widehat{x} + h) - (g \circ f)(\widehat{x})}{h} \\
= \; & \frac{g\big(f(\widehat{x} + h)\big) - g\big(f(\widehat{x})\big)}{h} \\
= \; & \frac{g\big(f(\widehat{x}) + h \cdot f'(\widehat{x}) + r_1(h)\big) - g\big(f(\widehat{x})\big)}{h} \\
= \; & \frac{g\big(\widehat{y} + h \cdot f'(\widehat{x}) + r_1(h)\big) - g\big(\widehat{y}\big)}{h} \\
= \; & \frac{g(\widehat{y}) + \big(h \cdot f'(\widehat{x}) + r_1(h)\big) \cdot g'(\widehat{y}) + r_2\big(h \cdot f'(\widehat{x}) + r_1(h)\big) - g(\widehat{y})}{h} \\
= \; & f'(\widehat{x}) \cdot g'(\widehat{y}) + \frac{r_1(h)}{h} \cdot g'(\widehat{y}) + \frac{r_2\big(h \cdot f'(\widehat{x}) + r_1(h)\big)}{h}
\end{aligned}$$

Wenn wir jetzt den Grenzwert $h \to 0$ berechnen, dann müssen wir uns den letzten Term genauer ansehen. Es gilt

$$\begin{aligned}
\lim_{h \to 0} \frac{r_2\big(h \cdot f'(\widehat{x}) + r_1(h)\big)}{h} &= \lim_{h \to 0} \frac{r_2\big(h \cdot f'(\widehat{x}) + r_1(h)\big)}{h \cdot f'(\widehat{x}) + r_1(h)} \cdot \frac{h \cdot f'(\widehat{x}) + r_1(h)}{h} \\
&= \lim_{h \to 0} \frac{r_2\big(h \cdot f'(\widehat{x}) + r_1(h)\big)}{h \cdot f'(\widehat{x}) + r_1(h)} \cdot \lim_{h \to 0} \frac{h \cdot f'(\widehat{x}) + r_1(h)}{h} \\
&= \lim_{h \to 0} \frac{r_2\big(h\big)}{h} \cdot \left(\lim_{h \to 0} f'(\widehat{x}) + \frac{r_1(h)}{h} \right) \\
&= 0 \cdot \big(f'(\widehat{x}) + 0\big) \\
&= 0
\end{aligned}$$

Damit sehen wir:

$$\begin{aligned}
& \lim_{h \to 0} \frac{(g \circ f)(\widehat{x} + h) - (g \circ f)(\widehat{x})}{h} \\
= \; & \lim_{h \to 0} f'(\widehat{x}) \cdot g'(\widehat{y}) + \frac{r_1(h)}{h} \cdot g'(\widehat{y}) + \frac{r_2\big(h \cdot f'(\widehat{x}) + r_1(h)\big)}{h} \\
= \; & f'(\widehat{x}) \cdot g'(\widehat{y}) + \lim_{h \to 0} \frac{r_1(h)}{h} \cdot g'(\widehat{y}) + \lim_{h \to 0} \frac{r_2\big(h \cdot f'(\widehat{x}) + r_1(h)\big)}{h} \\
= \; & f'(\widehat{x}) \cdot g'(\widehat{y}) + 0 + 0 \\
= \; & f'(\widehat{x}) \cdot g'(\widehat{y}).
\end{aligned}$$

Der obige exakte Beweis ist recht umständlich. Wir geben daher zusätzlich eine Plausibilitätsbetrachtung. Nach Definition der Ableitung gilt

$$g'(\widehat{y}) = \lim_{h \to 0} \frac{g(\widehat{y}+h) - g(\widehat{y})}{h}$$

Für kleine Werte von h gilt daher ungefähr

$$g(\widehat{y}+h) \approx g(\widehat{y}) + g'(\widehat{y}) \cdot h.$$

Analog finden wir für die Funktion f

$$f(\widehat{x}+h) \approx f(\widehat{x}) + f'(\widehat{x}) \cdot h.$$

Damit finden wir für den Differential-Quotienten der Funktion $g \circ f$ im Punkt $\widehat{x}$:

$$\begin{aligned}
\frac{(g \circ f)(\widehat{x}+h) - (g \circ f)(\widehat{x})}{h} &= \frac{g\big(f(\widehat{x}+h)\big) - g\big(f(\widehat{x})\big)}{h} \\
&\approx \frac{g\big(f(\widehat{x}) + f'(\widehat{x}) \cdot h\big) - g\big(f(\widehat{x})\big)}{h} \\
&\approx \frac{g\big(f(\widehat{x})\big) + g'\big(f(\widehat{x})\big) \cdot f'(\widehat{x}) \cdot h - g\big(f(\widehat{x})\big)}{h} \\
&= \frac{g'\big(f(\widehat{x})\big) \cdot f'(\widehat{x}) \cdot h}{h} \\
&= g'\big(f(\widehat{x})\big) \cdot f'(\widehat{x})
\end{aligned}$$

Die linke Seite der Gleichung stellt den Differential-Quotienten der Funktion $g \circ f$ dar und muss daher für $h \to 0$ gegen die Ableitung $(g \circ f)(\widehat{x})$ konvergieren. □

Aufgabe 24: Zeigen Sie, dass für alle natürlichen Zahlen n gilt:

$$\frac{d\,x^n}{dx} = n \cdot x^{n-1}.$$

Satz 52 (Ableitung von Potenzreihen) Ist die Funktion f als Potenzreihe definiert,

$$\sum_{n=0}^{\infty} a_n \cdot x^n$$

und ist R der Konvergenz-Radius dieser Potenzreihe, so ist f für alle $x \in \mathbb{R}$ mit $|x| < R$ differenzierbar und es gilt

$$f'(x) = \sum_{n=1}^{\infty} n \cdot a_n \cdot x^{n-1}.$$

Der letzte Satz besagt, dass Potenzreihen innerhalb ihres Konvergenz-Radius gliedweise differenziert werden können. Ein Beweis dieses Satzes ist mit den uns zur Verfügung stehenden Hilfsmitteln nicht möglich.

Wir berechnen als nächstes die Ableitung einiger wichtiger Funktionen.

1. Die Exponential-Funktion $\exp(x)$ ist definiert als

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

   Nach dem letzten Satz gilt für die Ableitung der Exponential-Funktion

   $$\frac{d}{dx}\exp(x) = \sum_{n=1}^{\infty} \frac{n}{n!} \cdot x^{n-1} = \sum_{n=1}^{\infty} \frac{1}{(n-1)!} \cdot x^{n-1} = \sum_{n=0}^{\infty} \frac{1}{n!} \cdot x^n = \exp(x),$$

   die Ableitung der Exponential-Funktion ergibt also wieder die Exponential-Funktion!

2. Um den natürlichen Logarithmus ableiten zu können, betrachten wir die Gleichung

$$\ln\big(\exp(x)\big) = x.$$

Differenzieren wir beide Seiten dieser Gleichung nach x, so erhalten wir nach der Ketten-Regel

$$\ln'\big(\exp(x)\big) \cdot \exp(x) = 1,$$

denn die Ableitung der Exponential-Funktion ergibt ja wieder die Exponential-Funktion. Setzen wir hier $y := \exp(x)$, so haben wir

$$\ln'(y) \cdot y = 1, \quad \text{also} \quad \frac{d}{dy}\ln(y) = \frac{1}{y}.$$

3. Um die Ableitung der Funktion $x \mapsto \sin(x)$ berechnen zu können, betrachten wir die Definition von Sinus und Tangens am Einheitskreis: Aus der Definition von Sinus und Tangens

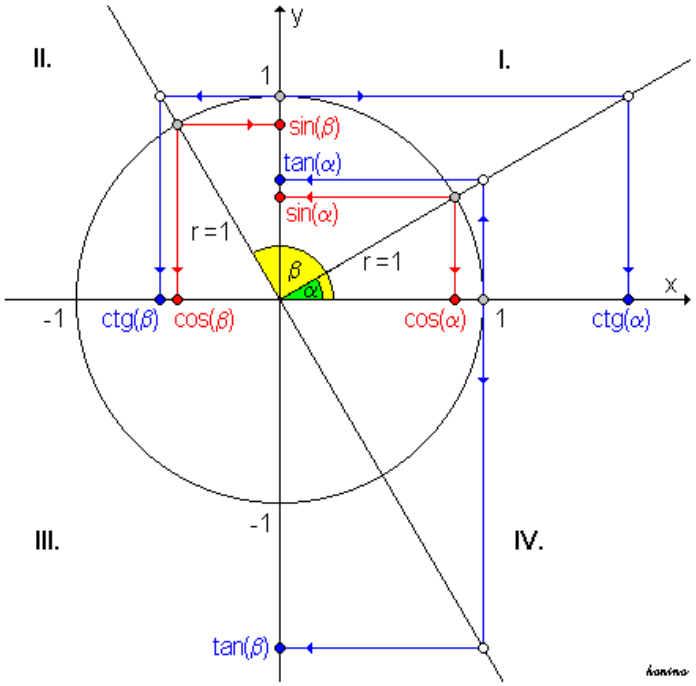

Abbildung 5.1: Die Winkel-Funktionen am Einheitskreis.

folgt die Ungleichung

$$\sin(\varphi) \le \varphi \le \tan(\varphi) = \frac{\sin(\varphi)}{\cos(\varphi)}$$

Division dieser Gleichung durch $\sin(\varphi)$ liefert

$$1 \le \frac{\varphi}{\sin(\varphi)} \le \frac{1}{\cos(\varphi)}$$

Wir bilden den Kehrwert und erhalten

$$1 \ge \frac{\sin(\varphi)}{\varphi} \ge \cos(\varphi)$$

Nun bilden wir den Grenzwert für $\varphi \to 0$:

$$1 \geq \lim_{\varphi \to 0} \frac{\sin(\varphi)}{\varphi} \geq \lim_{\varphi \to 0} \cos(\varphi)$$

Wegen $\lim_{\varphi \to 0} \cos(\varphi) = \cos(0) = 1$ folgt daraus

$$\lim_{\varphi \to 0} \frac{\sin(\varphi)}{\varphi} = 1.$$

Aus dem Geometrie-Unterricht ist das Additionstheorem für den Sinus bekannt:

$$\sin(x+y) = \sin(x) \cdot \cos(y) + \cos(x) \cdot \sin(y).$$

Daraus folgt einerseits

$$\begin{aligned} \sin(x) &= \sin\left(\tfrac{x+y}{2} + \tfrac{x-y}{2}\right) \\ &= \sin\left(\tfrac{x+y}{2}\right) \cdot \cos\left(\tfrac{x-y}{2}\right) + \cos\left(\tfrac{x+y}{2}\right) \cdot \sin\left(\tfrac{x-y}{2}\right) \end{aligned}$$

und andererseits gilt wegen $\sin(-x) = -\sin(x)$ und $\cos(-x) = \cos(x)$

$$\begin{aligned} \sin(y) &= \sin\left(\tfrac{x+y}{2} - \tfrac{x-y}{2}\right) \\ &= \sin\left(\tfrac{x+y}{2}\right) \cdot \cos\left(\tfrac{x-y}{2}\right) - \cos\left(\tfrac{x+y}{2}\right) \cdot \sin\left(\tfrac{x-y}{2}\right). \end{aligned}$$

Subtrahieren wir diese Gleichungen voneienander, so erhalten wir

$$\sin(x) - \sin(y) = 2 \cdot \cos\left(\tfrac{x+y}{2}\right) \cdot \sin\left(\tfrac{x-y}{2}\right).$$

Damit können wir die Ableitung des Sinus ausrechnen:

$$\begin{aligned} \lim_{h \to 0} \frac{\sin(x+h) - \sin(x)}{h} &= \lim_{h \to 0} \frac{2 \cdot \cos\left(\dfrac{x+h+x}{2}\right) \cdot \sin\left(\dfrac{x+h-x}{2}\right)}{h} \\ &= \lim_{h \to 0} \cos\left(x + \tfrac{h}{2}\right) \cdot \lim_{h \to 0} \frac{\sin\left(\dfrac{h}{2}\right)}{\dfrac{h}{2}} \\ &= \cos(x) \cdot \lim_{h \to 0} \frac{\sin(h)}{h} \\ &= \cos(x) \end{aligned}$$

Damit haben wir gezeigt, dass gilt:

$$\tfrac{d}{dx} \sin(x) = \cos(x).$$

4. Die Ableitung des Cosinus könnte in analoger Weise berechnet werden, es ist aber einfacher, wenn wir von den Gleichungen

$$\cos(x) = \sin\left(\tfrac{\pi}{2} - x\right) \quad \text{und} \quad \cos\left(\tfrac{\pi}{2} - x\right) = \sin(x)$$

ausgehen und die Ketten-Regel verwenden. Es ergibt sich

$$\begin{aligned}
\frac{d}{dx}\cos(x) &= \frac{d}{dx}\sin\Big(\frac{\pi}{2} - x\Big) \\
&= \cos\Big(\tfrac{\pi}{2} - x\Big) \cdot \frac{d}{dx}\Big(\frac{\pi}{2} - x\Big) \quad \text{nach der Ketten-Regel} \\
&= \sin(x) \cdot (-1) \\
&= -\sin(x).
\end{aligned}$$

5. Jetzt kann die Ableitung der Tangens-Funktion über die Quotienten-Regel berechnet werden:

$$\begin{aligned}
\frac{d}{dx}\tan(x) &= \frac{d}{dx}\left(\frac{\sin(x)}{\cos(x)}\right) \\
&= \frac{\Big(\frac{d}{dx}\sin(x)\Big) \cdot \cos(x) - \sin(x) \cdot \Big(\frac{d}{dx}\cos(x)\Big)}{\cos^2(x)} \\
&= \frac{\cos(x) \cdot \cos(x) - \sin(x) \cdot \big(-\sin(x)\big)}{\cos^2(x)} \\
&= \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} \\
&= \frac{1}{\cos^2(x)}
\end{aligned}$$

6. Die Ableitung der Arcus-Tangens-Funktion kann nun mit dem selben Trick berechnet werden, den wir schon bei der Berechnung der Ableitung des Logarithmus benutzt haben. Wir gehen diesmal von der Gleichungen

   $$\arctan\big(\tan(x)\big) = x$$

   aus und differenzieren beide Seiten dieser Gleichung. Nach der Ketten-Regel erhalten wir

   $$\arctan'\big(\tan(x)\big) \cdot \tfrac{d}{dx}\tan(x) = 1.$$

   Setzen wir hier die Ableitung für die Tangens-Funktion ein, so haben wir

   $$\frac{d}{dx}\arctan\big(\tan(x)\big) \cdot \frac{1}{\cos^2(x)} = 1.$$

   Multiplikation mit $\cos^2(x)$ ergibt

   $$\frac{d}{dx}\arctan\big(\tan(x)\big) = \cos^2(x).$$

   Den in dieser Gleichung auftretenden Term $\cos^2(x)$ müssen wir durch einen Term ausdrücken, in dem nur $\tan(x)$ auftritt. Dazu betrachten wir die Definition der Tangens-Funktion:

$$
\begin{aligned}
& & \tan^2(x) &= \frac{\sin^2(x)}{\cos^2(x)} & \\
\Leftrightarrow & & \tan^2(x) &= \frac{1-\cos^2(x)}{\cos^2(x)} & \text{wegen } \sin^2(x)+\cos^2(x)=1 \\
\Leftrightarrow & \cos^2(x)\cdot\tan^2(x) &= 1-\cos^2(x) & & \\
\Leftrightarrow & \cos^2(x)\cdot\tan^2(x)+\cos^2(x) &= 1 & & \\
\Leftrightarrow & \cos^2(x)\cdot\big(\tan^2(x)+1\big) &= 1 & & \\
\Leftrightarrow & \cos^2(x) &= \frac{1}{\tan^2(x)+1} & &
\end{aligned}
$$

Damit können wir also schreiben

$$\arctan'\big(\tan(x)\big) = \frac{1}{\tan^2(x)+1}.$$

Setzen wir jetzt $y = \tan(x)$, so erhalten wir

$$\frac{d}{dy}\arctan(y) = \frac{1}{y^2+1}.$$

Aufgabe 25: Zeigen Sie

$$\frac{d}{dx}\arcsin(x) = \frac{1}{\sqrt{1-x^2}}.$$ ◇

Aufgabe 26: Berechnen Sie die Ableitung der Funktion $x \mapsto \sqrt{x}$.

Hinweis: Verwenden Sie die Produkt-Regel. ◇

Aufgabe 27: Es sei $p \in \mathbb{Z}$ und $q \in \mathbb{N}$. Überlegen Sie, was die Ableitung der Funktion

$$x \mapsto x^{\frac{p}{q}}$$

ist und beweisen Sie Ihre Behauptung.

Hinweis: Betrachten Sie zunächst den Fall $p = 1$. ◇

5.2 Mittelwert-Sätze

Definition 53 (lokales Maximum) Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ hat im Punkt $\bar{x} \in \mathbb{R}$ ein *lokales Maximum*, wenn gilt:

$$\exists \varepsilon \in \mathbb{R}_+ : \forall x \in \mathbb{R} : |x - \bar{x}| < \varepsilon \to f(x) \leq f(\bar{x}).$$ ◇

Die in der obigen Definition auftretende Menge von Zahlen, deren Abstand von $\bar{x}$ kleiner ist als ε, bezeichnen wir auch als ε-Umgebung des Punktes $\bar{x}$, die ε-Umgebung des Punktes x ist also die Menge

$$U_\varepsilon(\bar{x}) := \big\{x \in \mathbb{R} \;\big|\; |x - \bar{x}| < \varepsilon\big\}.$$

Der Begriff des lokalen Maximums steht im Kontrast zu dem Begriff eines *globalen Maximums*. Eine Funktion $f : D \to \mathbb{R}$ hat in einem Punkt $\bar{x} \in D$ ein globales Maximum, wenn

$$\forall x \in D : f(x) \leq f(\bar{x})$$

gilt. Natürlich ist jedes globale Maximum auch ein lokales Maximum, aber die Umkehrung gilt im allgemeinen nicht. Der nächste Satz liefert ein notwendiges Kriterium für das Auftreten eines lokalen Maximums.

Satz 54 (Pierre de Fermat, 160?–1665) Hat die Funktion $f : D \to \mathbb{R}$ im Punkt $\bar{x}$ ein lokales Maximum, ist $U_\varepsilon(\bar{x}) \subseteq D$ und ist die Funktion f zusätzlich im Punkt $\bar{x}$ differenzierbar, so gilt

$$\frac{df}{dx}(\bar{x}) = 0.$$

Beweis: Wir betrachten zunächst die Folge $\left(\bar{x} + \frac{1}{n}\right)_{n\in\mathbb{N}}$. O.B.d.A. sei ε so klein gewählt, dass

$$\forall x \in \mathbb{R} : |x - \bar{x}| < \varepsilon \to f(x) \leq f(\bar{x})$$

gilt. Wenn $n > \frac{1}{\varepsilon}$ ist, liegt $\bar{x} + \frac{1}{n}$ in der ε-Umgebung von $\bar{x}$. Daher gilt für alle $n > \frac{1}{\varepsilon}$

$$f\left(\bar{x} + \tfrac{1}{n}\right) \leq f(\bar{x}).$$

Damit gilt für den Differential-Quotienten

$$\frac{f\left(\bar{x} + \dfrac{1}{n}\right) - f(\bar{x})}{\bar{x} + \dfrac{1}{n} - \bar{x}} = n \cdot \left(f\left(\bar{x} + \frac{1}{n}\right) - f(\bar{x})\right) \leq 0.$$

Da wir vorausgesetzt haben, dass die Funktion f im Punkt $\bar{x}$ differenzierbar ist, gilt

$$\frac{df}{dx}(\bar{x}) = \lim_{n\to\infty} \frac{f\left(\bar{x} + \dfrac{1}{n}\right) - f(\bar{x})}{\bar{x} + \dfrac{1}{n} - \bar{x}} \leq 0.$$

Wir betrachten nun die Folge $\left(\bar{x} - \frac{1}{n}\right)_{n\in\mathbb{N}}$. Wieder sei ε so gewählt, dass

$$\forall x \in \mathbb{R} : |x - \bar{x}| < \varepsilon \to f(x) \leq f(\bar{x})$$

gilt. Wenn $n > \frac{1}{\varepsilon}$ liegt daher $\bar{x} - \frac{1}{n}$ in der ε-Umgebung von $\bar{x}$. Daher gilt für alle $n > \frac{1}{\varepsilon}$

$$f\left(\bar{x} - \tfrac{1}{n}\right) \leq f(\bar{x}).$$

Damit gilt für den Differential-Quotienten

$$\frac{f\left(\bar{x} + \dfrac{1}{n}\right) - f(\bar{x})}{\bar{x} - \dfrac{1}{n} - \bar{x}} = -n \cdot \left(f\left(\bar{x} - \frac{1}{n}\right) - f(\bar{x})\right) \geq 0.$$

Da wir vorausgesetzt haben, dass die Funktion f im Punkt $\bar{x}$ differenzierbar ist, gilt

$$\frac{df}{dx}(\bar{x}) = \lim_{n\to\infty} \frac{f\left(\bar{x} - \dfrac{1}{n}\right) - f(\bar{x})}{\bar{x} - \dfrac{1}{n} - \bar{x}} \geq 0.$$

Wir haben jetzt also die beiden Ungleichungen

$$\frac{df}{dx}(\bar{x}) \leq 0 \quad \text{und} \quad \frac{df}{dx}(\bar{x}) \geq 0$$

gezeigt. Daraus folgt sofort $\frac{df}{dx}(\bar{x}) = 0$. □

Analog zur Definition eines lokalen Maximums kann auch der Begriff eines *lokalen Minimums* definiert werden. Auch in einem lokalen Minimum hat die Ableitung den Wert 0.

Satz 55 Ist die Funktion $f : [a, b] \to \mathbb{R}$ stetig, so nimmt f auf dem Intervall $[a, b]$ sowohl das Maximum als auch das Minimum an, es gibt also Punkte x_{min} und x_{max}, so dass gilt

$$\forall x \in [a, b] : f(x) \leq f(x_{max}) \quad \text{und} \quad \forall x \in [a, b] : f(x) \geq f(x_{min}). \qquad \diamond$$

Satz 56 (Michel Rolle, 1652 – 1719) Ist die Funktion $f : [a, b] \to \mathbb{R}$ differenzierbar und gilt außerdem $f(a) = f(b)$, dann gibt es ein $\bar{x} \in (a, b)$, so dass gilt

$$\frac{df}{dx}(\bar{x}) = 0.$$

Beweis: Es gibt zwei Fälle:

1. Die Funktion f ist konstant, für alle $x \in [a, b]$ gilt also $f(x) = f(a)$. Da die Ableitung einer konstanten Funktion den Wert 0 hat, gilt dann offenbar sogar für alle $x \in [a, b]$

   $$\frac{df}{dx}(x) = 0.$$

2. Da die Funktion f differenzierbar ist, ist sie auch stetig und nimmt daher sowohl ein Minimum als auch ein Maximum in dem Intervall $[a, b]$ an. Es gibt also x_{min} und x_{max} mit

   $$\forall x \in [a, b] : f(x) \leq f(x_{max}) \quad \text{und} \quad \forall x \in [a, b] : f(x) \geq f(x_{min}).$$

   Da wir jetzt voraussetzen können, dass die Funktion nicht konstant ist, und da weiterhin $f(a) = f(b)$ gilt, muss

   $$f\big(x_{min}\big) < f(a) \quad \text{oder} \quad f\big(x_{max}\big) > f(a)$$

   gelten. Daraus folgt

   $$x_{min} \notin \{a, b\} \quad \text{oder} \quad x_{max} \notin \{a, b\}.$$

   Damit hat die Funktion dann in x_{min} ein lokales Minimum oder in x_{max} ein lokales Maximum (oder beides) und nach dem Satz von Fermat folgt

   $$\frac{df}{dx}\big(x_{min}\big) = 0 \quad \text{oder} \quad \frac{df}{dx}\big(x_{max}\big) = 0.$$ □

Aus dem Satz von Rolle folgern wir später zwei wichtige Mittelwert-Sätze und den Satz von *L'Hôpital* (Guillaume François Antoine, Marquis de L'Hôpital, 1661–1704).

Satz 57 (Mittelwert-Satz der Differential-Rechnung, Augustin-Louis Cauchy, 1789–1857) Ist die Funktion $f : [a, b] \to \mathbb{R}$ für alle $x \in [a, b]$ differenzierbar, so gilt:

$$\exists c \in (a, b) : \frac{df}{dx}(c) = \frac{f(b) - f(a)}{b - a}.$$

Beweis: Wir definieren die Funktion $g : [a, b] \to \mathbb{R}$ durch

$$g(x) := f(x) - f(a) - \frac{f(b) - f(a)}{b - a} \cdot (x - a).$$

Da die Funktion f nach Voraussetzung differenzierbar ist, ist auch die Funktion g differenzierbar und es gilt

$$g(a) = f(a) - f(a) - \frac{f(b) - f(a)}{b - a} \cdot (a - a) = 0.$$

und

$$g(b) = f(b) - f(a) - \frac{f(b) - f(a)}{b - a} \cdot (b - a) = f(b) - f(a) - \big(f(b) - f(a)\big) = 0.$$

Damit gilt $g(a) = g(b)$ und folglich erfüllt die Funktion g die Voraussetzung des Satzes von Rolle. Also gibt es ein $c \in (a, b)$, so dass

$$\frac{dg}{dx}(c) = 0$$

gilt. Setzen wir hier die Definition von g ein, so haben wir

$$\begin{aligned} & \frac{dg}{dx}(c) = \frac{df}{dx}(c) - \frac{f(b) - f(a)}{b - a} = 0 \\ \Rightarrow \quad & \frac{df}{dx}(c) = \frac{f(b) - f(a)}{b - a} \end{aligned}$$

□

Abbildung 5.2 zeigt die geometrische Bedeutung des Mittelwert-Satzes: Es gibt eine Tangente an die Funktion, die dieselbe Steigung hat wie die Sekante, die durch die Punkte $\langle a, f(a\rangle)$ und $\langle b, f(b\rangle)$ geht.

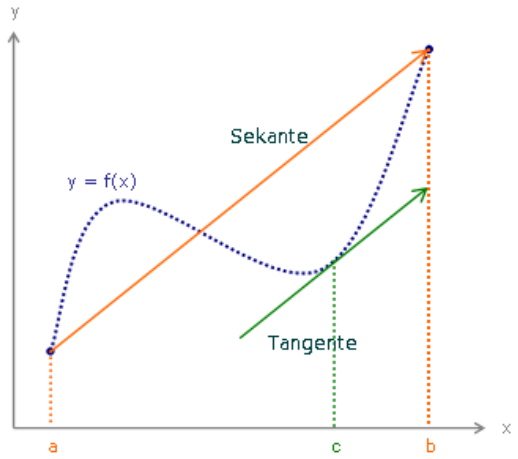

Abbildung 5.2: Geometrische Bedeutung des Mittelwert-Satzes.

Satz 58 (Erweiterter Mittelwert-Satz) Sind die Funktion $f, g : [a, b] \to \mathbb{R}$ für alle $x \in [a, b]$ differenzierbar und gilt $\frac{dg}{dx}(x) \neq 0$ für alle $x \in [a, b]$, so gilt:

$$\exists c \in (a, b) : \frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

◇

Bemerkung: Auf den ersten Blick mag es verwundern, dass nicht explizit $g(a) \neq g(b)$ gefordert wird. Dies folgt aber sofort aus der Bedingung $\forall x \in [a, b] : \frac{dg}{dx}(x) \neq 0$ und dem Satz von Rolle. ◇

Aufgabe 28: Beweisen Sie den erweiterten Mittelwert-Satz. Betrachten Sie dazu die Funktion

$$h(x) := \alpha \cdot f(x) - \beta \cdot g(x)$$

und bestimmen Sie α und β so, dass Sie auf die Funktion h den Satz von Rolle anwenden können.

◇

Der folgende Satz ist für die praktische Berechnung von Grenzwerten unentbehrlich.

Satz 59 (Guillaume François Antoine, Marquis de L'Hôpital, 1661 –1704)
Die Funktionen $f, g : (a, b) \to \mathbb{R}$ seien differenzierbar , es sei $c \in (a, b)$ und es gelte

1. $f(c) = g(c) = 0$ und
2. $\forall x \in (a, b) : g'(x) \neq 0$.

Dann gilt

$$\lim_{x \to c} \frac{f(x)}{g(x)} = \frac{f'(c)}{g'(c)}.$$

Beweis: Da die Funktion f und g im Punkt c differenzierbar sind, gibt es Funktionen $r_1(h)$ und $r_2(h)$, so dass gilt:

1. $f(c+h) = f(c) + h \cdot f'(c) + r_1(h)$ mit $\lim\limits_{h \to 0} \dfrac{r_1(h)}{h} = 0$.
2. $g(c+h) = g(c) + h \cdot g'(c) + r_2(h)$ mit $\lim\limits_{h \to 0} \dfrac{r_2(h)}{h} = 0$.

Wir haben die folgende Kette von Gleichungen:

$$\begin{aligned}
\lim_{h \to 0} \frac{f(c+h)}{g(c+h)} &= \lim_{h \to 0} \frac{f(c) + h \cdot f'(c) + r_1(h)}{g(c) + h \cdot g'(c) + r_2(h)} = \lim_{h \to 0} \frac{h \cdot f'(c) + r_1(h)}{h \cdot g'(c) + r_2(h)} \\
&= \lim_{h \to 0} \frac{f'(c) + \dfrac{r_1(h)}{h}}{g'(c) + \dfrac{r_2(h)}{h}} = \frac{f'(c) + \lim\limits_{h \to 0} \dfrac{r_1(h)}{h}}{g'(c) + \lim\limits_{h \to 0} \dfrac{r_2(h)}{h}} \\
&= \frac{f'(c)}{g'(c)}
\end{aligned}$$

□

Beispiel: Mit dem Satz von L'Hôpital können wir nun den Grenzwert $\lim\limits_{x \to 0} \dfrac{\sin(x)}{x}$ noch einmal berechnen:

$$\lim_{x \to 0} \frac{\sin(x)}{x} = \lim_{x \to 0} \frac{\cos(x)}{1} = \cos(0) = 1.$$

◇

Der Satz von L'Hôpital behält seine Gültigkeit, wenn x gegen Unendlich strebt. Sind $f, g : \mathbb{R} \to \mathbb{R}$ differenzierbare Funktionen, so dass der Grenzwert

$$\lim_{x \to \infty} \tfrac{f'(x)}{g'(x)}$$

existiert, und gilt entweder

$$\Big(\lim_{x \to \infty} f(x) = 0 \ \wedge \ \lim_{x \to \infty} g(x) = 0\Big) \quad \vee \quad \Big(\lim_{x \to \infty} f(x) = \infty \ \wedge \ \lim_{x \to \infty} g(x) = \infty\Big)$$

so folgt

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{x \to \infty} \frac{f'(x)}{g'(x)}.$$

Wir geben ein Beispiel. Es gilt

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

Der Satz von L'Hôpital lässt sich iteriert anwenden. Beispielsweise gilt

$$\lim_{x\to\infty} \frac{x^2}{\exp(x)} = \lim_{x\to\infty} \frac{2\cdot x}{\exp(x)} = \lim_{x\to\infty} \frac{2}{\exp(x)} = 0.$$

Definition 60 (Schnelleres Wachstum) Wir sagen, dass die Funktion $x \mapsto f(x)$ für $x \to \infty$ *schneller als* die Funktion $x \mapsto g(x)$ *wächst*, falls

$$\lim_{x\to\infty} \frac{g(x)}{f(x)} = 0$$

gilt. ◇

Aufgabe 29: Zeigen Sie, dass für alle natürlichen Zahlen n gilt:

$$\lim_{x\to\infty} \frac{x^n}{\exp(x)} = 0.$$

Damit sehen wir, dass die Exponential-Funktion schneller wächst als jede Potenz. ◇

Aufgabe 30:

1. Zeigen Sie, dass die Funktion $x \mapsto e^{\ln(x)\cdot\ln(x)}$ für alle $n \in \mathbb{N}$ schneller als die Funktion $x \mapsto x^n$ wächst.
2. Zeigen Sie, dass die Funktion $x \mapsto e^x$ schneller wächst als die Funktion $x \mapsto e^{\ln(x)\cdot\ln(x)}$. ◇

Aufgabe 31: Berechnen Sie den Grenzwert

$$\lim_{x\to 0} x \cdot \ln(x).$$ ◇

Aufgabe 32: Berechnen Sie den Grenzwert

$$\lim_{x\to\infty} \sqrt{x+\sqrt{x}} - \sqrt{x}\,.$$ ◇

5.3 Monotonie und Konvexität

Im Folgenden bezeichnet D entweder ein Intervall der Form

$$[a,b], \quad (a,b], \quad [a,b), \quad (a,b),$$

ein unbeschränktes Intervall der Form

$$[a,\infty), \quad (a,\infty), \quad (-\infty,b], \quad (-\infty,b)$$

oder die Menge $\mathbb{R}$ der reellen Zahlen.

Definition 61 (monoton) Eine Funktion $f : D \to \mathbb{R}$ ist *monoton steigend* g.d.w.

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

gilt. Die Funktion f ist *streng monoton steigend*, wenn die schärfere Bedingung

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

erfüllt ist. Weiter heißt f *monoton fallend*, wenn

$$\forall x, y \in D : x < y \rightarrow f(x) \geq f(y)$$

gilt. Analog ist f *streng monoton fallend*, falls die folgende Bedingung gilt:

$$\forall x, y \in D : x < y \rightarrow f(x) > f(y). \qquad \diamond$$

Satz 62 Eine differenzierbare Funktion $f : D \rightarrow \mathbb{R}$ ist genau dann monoton steigend, wenn gilt:

$$\forall x \in D : f'(x) \geq 0. \qquad \diamond$$

Beweis: Da es sich bei diesem Beweis um eine "genau-dann-wenn"-Aussage handelt, spalten wir den Beweis in zwei Teile auf.

"$\Rightarrow$": Wir nehmen zunächst an, dass f monoton steigend ist und zeigen, dass dann $f'(x) \geq 0$ gilt. Die Ableitung ist definiert als der Grenzwert

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

Wir zeigen, dass der Differential-Quotient

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

für alle $h \neq 0$ größer oder gleich 0 ist. Zm Nachweis dieser Behauptung führen wir eine Fallunterscheidung bezüglich des Vorzeichens von h durch.

(a) Fall: $h > 0$.
Aus $h > 0$ folgt $x + h > x$. Aus der Monotonie von f folgt dann, dass $f(x+h) \geq f(x)$ ist. Also gilt $f(x+h) - f(x) \geq 0$ und daraus folgt

$$\frac{f(x+h) - f(x)}{h} \geq 0.$$

(b) Fall: $h < 0$.
Aus $h < 0$ folgt nun $x - h < x$. Aus der Monotonie von f folgt jetzt die Ungleichung $f(x+h) \leq f(x)$. Also haben wir $f(x+h) - f(x) \leq 0$. Wegen $h < 0$ gilt dann insgesamt

$$\frac{f(x+h) - f(x)}{h} \geq 0.$$

Da der Differential-Quotient in jedem Fall größer-gleich 0 ist und die Ableitung $f'(x)$ als Grenzwert des Differential-Quotienten für h gegen 0 definiert ist, muss $f'(x) \geq 0$ gelten.

"$\Leftarrow$": Wir nehmen nun an, dass für alle $x \in D$ die Ungleichung $f'(x) \geq 0$ gilt und zeigen, dass f dann monoton steigend ist. Diesen Beweis führen wir indirekt. Wir nehmen an, es gäbe $x, y \in D$ mit

$$x < y \quad \text{aber} \quad f(x) > f(y).$$

Nach dem Mittelwert-Satz der Differential-Rechnung gibt es dann ein $z \in [x, y]$, so dass

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

gilt. Aus $x < y$ folgt $y - x \geq 0$ und aus $f(x) > f(y)$ folgt $f(y) - f(x) < 0$. Damit hätten wir dann aber $f'(z) < 0$ im Widerspruch zur Voraussetzung. $\square$

In Analogie zum letzten Satz kann gezeigt werden, dass eine differenzierbare Funktion $f : D \rightarrow \mathbb{R}$ genau dann monoton fallend ist, wenn für alle $x \in D$ die Ungleichung $f'(x) \leq 0$ gilt.

Aufgabe 33: Die Funktion $f : D \rightarrow \mathbb{R}$ sei differenzierbar und es gelte

$$\forall x \in D : f'(x) > 0.$$

Zeigen Sie, dass die Funktion f dann streng monoton steigend ist.

Bemerkung: Die Funktion $x \mapsto x^3$ ist streng monoton steigend, aber an der Stelle $x = 0$ verschwindet die Ableitung dieser Funktion. Dies zeigt, dass sich die Aussage des letzten Satzes nicht umkehren lässt.

Definition 63 (strenges lokales Minimum)
Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ hat im Punkt $\bar{x} \in \mathbb{R}$ ein *strenges lokales Minimum*, wenn gilt:

$$\exists \varepsilon \in \mathbb{R}_+ : \forall x \in \mathbb{R} : |x - \bar{x}| < \varepsilon \wedge \ x \neq \bar{x} \rightarrow f(x) > f(\bar{x}). \qquad \diamond$$

Bemerkung: Der Begriff des *strengen lokalen Maximum* lässt sich analog definieren.

Satz 64 Die Funktion $f : \mathbb{R} \to \mathbb{R}$ sei zweimal differenzierbar, die zweite Ableitung $f''(x)$ sei stetig und für ein $x_0 \in \mathbb{R}$ gelte

$$f'(x_0) = 0 \ \wedge \ f''(x_0) > 0.$$

Dann hat die Funktion f in x_0 ein strenges lokales Minimum.

Beweis: Da die zweite Ableitung $f''(x)$ stetig ist, können wir $\varepsilon := f''(x_0) > 0$ setzen und finden dann ein $\delta > 0$, so dass

$$\forall x \in \mathbb{R} : |x - x_0| < \delta \rightarrow |f''(x) - f''(x_0)| < \varepsilon = f''(x_0).$$

gilt. Subtrahieren wir $|f''(x) - f''(x_0)|$ auf beiden Seiten dieser Gleichung, so folgt, dass für alle $x \in \mathbb{R}$ mit $|x - x_0| < \delta$ die Ungleichung

$$f''(x_0) - |f''(x) - f''(x_0)| > 0$$

gilt. Wir behaupten, dass dann

$$f''(x) > 0 \quad \text{für alle } x \in \mathbb{R} \text{ mit } |x - x_0| < \delta \tag{5.1}$$

gilt. Zum Nachweis dieser Behauptung führen wir eine Fallunterscheidung bezüglich der relativen Größe von $f''(x)$ und $f''(x_0)$ durch.

1. Fall: $f''(x) < f''(x_0)$. Dann gilt

   $$|f''(x) - f''(x_0)| = f''(x_0) - f''(x).$$

   Also folgt aus der Ungleichung $f''(x_0) - |f''(x) - f''(x_0)| > 0$ die Ungleichung

   $$f''(x_0) - \big(f''(x_0) - f''(x)\big) > 0$$

   und wegen $f''(x_0) - \big(f''(x_0) - f''(x)\big) = f''(x)$ haben wir damit die Behauptung $f''(x) > 0$ gezeigt.

2. Fall: $f''(x) \geq f''(x_0)$.

   In diesem Fall folgt die Behauptung sofort aus der Voraussetzung $f''(x_0) > 0$ und der Transitivität der Relation $>$.

Die Ungleichung (5.1) zeigt uns, dass die Funktion $x \mapsto f'(x)$ in der δ-Umgebung von x_0 streng monoton steigend ist. Da außerdem $f'(x_0) = 0$ gilt, folgt insgesamt

$$f'(x) < 0 \quad \text{für alle } x \in U_\delta(x_0) \text{ mit } x < x_0 \quad \text{und}$$

$$f'(x) > 0 \quad \text{für alle } x \in U_\delta(x_0) \text{ mit } x > x_0.$$

Damit ist die Funktion f innerhalb der δ-Umgebung $U_\delta(x_0)$ für $x < x_0$ streng monoton fallend

und für $x > x_0$ streng monoton wachsend. Dann muss f aber ein lokales Minimum in x_0 haben. □

Bemerkung: Falls für die Funktion f die Bedingung

$$f'(x_0) = 0 \wedge f''(x_0) < 0.$$

erfüllt ist, dann hat die Funktion an der Stelle x_0 ein strenges lokales Maximum. ◇

Definition 65 (konvex, konkav) Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt *konvex* genau dann, wenn

$$\forall x_1, x_2 \in \mathbb{R} : \forall t \in [0,1] : f\big(t \cdot x_1 + (1-t) \cdot x_2\big) \leq t \cdot f(x_1) + (1-t) \cdot f(x_2)$$

gilt. Geometrisch bedeutet dies, dass die Funktionswerte der Funktion f unterhalb der Sekante durch die Punkte $\langle x_1, f(x_1)\rangle$ und $\langle x_2, f(x_2)\rangle$ liegen. Abbildung 5.3 auf Seite 80 zeigt dies anschaulich: In dem Intervall (x_1, x_2) liegen die Werte der Funktion f unterhalb der Gerade g, die durch die beiden Punkte $\langle x_1, f(x_1)\rangle$ und $\langle x_2, f(x_2)\rangle$ geht. Die Gleichung dieser Geraden ist

$$g(t) = \frac{t - x_1}{x_2 - x_1} \cdot f(x_2) + \frac{t - x_2}{x_1 - x_2} \cdot f(x_1).$$

Sie können dies sofort verifizieren, denn offenbar ist $g(t)$ in der Variablen t linear und andererseits gilt

$$g(x_1) = \frac{x_1 - x_1}{x_2 - x_1} \cdot f(x_2) + \frac{x_1 - x_2}{x_1 - x_2} \cdot f(x_1) = f(x_1)$$

und analog sehen wir, dass auch $g(x_2) = f(x_2)$ ist.

Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt *konkav* genau dann, wenn

$$\forall x_1, x_2 \in \mathbb{R} : \forall t \in [0,1] : f\big(t \cdot x_1 + (1-t) \cdot x_2\big) \geq t \cdot f(x_1) + (1-t) \cdot f(x_2)$$

gilt. Hier liegen die Funktionswerte der Funktion f also oberhalb der Sekante durch die Punkte $\langle x_1, f(x_1)\rangle$ und $\langle x_2, f(x_2)\rangle$.

Abbildung 5.4 auf Seite 80 zeigt eine konkave Funktion f zusammen mit einer Sekante g. Es ist deutlich zu sehen, dass hier die Funktionswerte oberhalb der Sekante liegen. ◇

Lemma 66 (Invarianz der Konvexität unter linearen Transformationen)
Die Funktion $f : \mathbb{R} \to \mathbb{R}$ sei konvex und es sei $\alpha \in \mathbb{R}$. Definieren wir die Funktion $g : \mathbb{R} \to \mathbb{R}$ als

$$g(x) := f(x) + \alpha \cdot x,$$

so ist auch die Funktion g konvex. Eine entsprechende Aussage gilt auch für konkave Funktionen.

Aufgabe 34: Beweisen Sie das vorangehende Lemma.

Satz 67 Die Funktion $f : \mathbb{R} \to \mathbb{R}$ sei zweimal differenzierbar und die Funktion $x \mapsto f''(x)$ sei stetig. Dann gilt

$$f \text{ konvex} \quad \Leftrightarrow \quad \forall x \in \mathbb{R} : f''(x) \geq 0.$$

Beweis: Wir spalten den Beweis in zwei Teile auf.

"⇒": Wir führen den Nachweis indirekt und nehmen an, dass es ein $x_0 \in \mathbb{R}$ gibt, so dass $f''(x_0) < 0$ ist. Ähnlich wie bei Beweis von Satz 64 folgt daraus, dass es eine δ_1-Umgebung $U_{\delta_1}(x_0)$ gibt, so dass

$$f''(x) < 0 \quad \text{für alle } x \in U_{\delta_1}(x_0)$$

gilt. Wir definieren eine Funktion $g : \mathbb{R} \to \mathbb{R}$ durch

$$g(x) := f(x) - x \cdot f'(x_0).$$

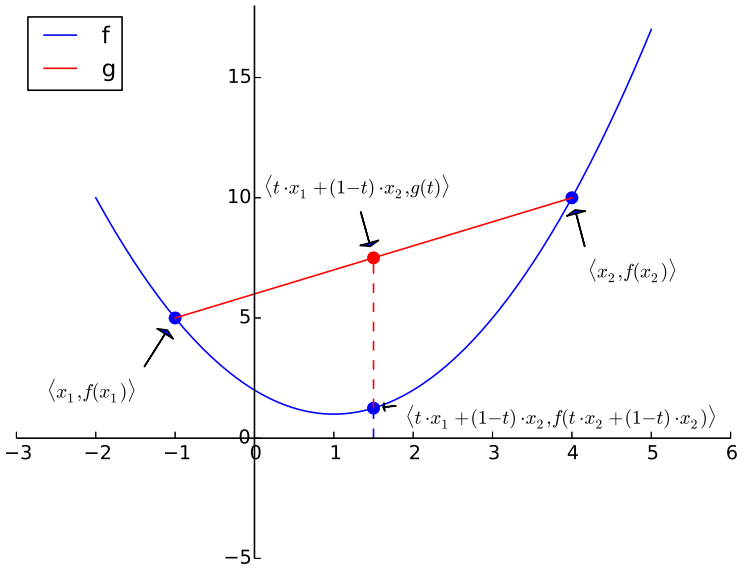

Abbildung 5.3: Eine konvexe Funktion f zusammen mit einer Sekante g.

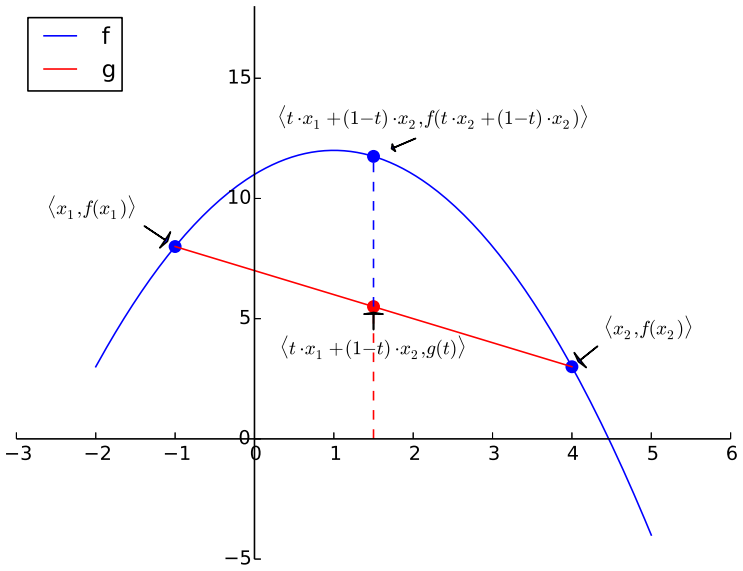

Abbildung 5.4: Eine konkave Funktion f zusammen mit einer Sekante g.

Dann gilt

$$g'(x) = f'(x) - f'(x_0) \quad \text{und} \quad g''(x) = f''(x).$$

Daraus folgt durch Einsetzen

$$g'(x_0) = 0 \quad \text{und} \quad g''(x_0) < 0.$$

Damit hat die Funktion g im Punkt x_0 ein lokales Maximum. Also gibt es eine δ_2-Umgebung von x_0, so dass

$$g(x) < g(x_0) \quad \text{für alle } x \in U_{\delta_2}(x_0)$$

gilt. O.B.d.A. können wir voraussetzen, dass $\delta_2 \leq \delta_1$ gilt. Nach dem Lemma 66 wissen wir, dass die Funktion g ebenfalls konvex ist. Definieren wir

$$x_1 := x_0 - \frac{\delta_2}{2}, \quad x_2 := x_0 + \frac{\delta_2}{2} \quad \text{und} \quad t := \frac{1}{2},$$

so folgt also

$$t \cdot g(x_1) + (1-t) \cdot g(x_2) \geq g\big(t \cdot x_1 + (1-t) \cdot x_2\big) \tag{5.2}$$

Nun gilt

$$t \cdot x_1 + (1-t) \cdot x_2 = \frac{1}{2} \cdot x_0 - \frac{1}{2} \cdot \frac{\delta_2}{2} + \frac{1}{2} \cdot x_0 + \frac{1}{2} \cdot \frac{\delta_2}{2} = x_0.$$

Damit folgt aus der Ungleichung (5.2) die Ungleichung

$$\frac{1}{2} \cdot g(x_1) + \frac{1}{2} \cdot g(x_2) \geq g(x_0). \tag{5.3}$$

Andererseits folgt aus der Tatsache, dass sowohl x_1 als auch x_2 in der δ_1-Umgebung von x_0 liegen, dass

$$g(x_1) < g(x_0) \quad \text{und} \quad g(x_2) < g(x_0)$$

gilt. Multiplizieren wir diese beiden Gleichungen mit $\frac{1}{2}$ und addieren sie, so ergibt sich

$$\tfrac{1}{2} \cdot g(x_1) + \tfrac{1}{2} \cdot g(x_2) < g(x_0)\ .$$

Diese Ungleichung steht aber im Widerspruch zur Ungleichung (5.3).

"⇐": Es seien x_1, x_2 und $t \in [0,1]$ gegeben. O.B.d.A. sei weiter $x_1 < x_2$. Wir definieren zunächst

$$x_0 := t \cdot x_1 + (1-t) \cdot x_2$$

Es lässt sich sofort nachrechnen, dass dann $x_1 < x_0 < x_2$ gilt. Nach dem Mittelwert-Satz der Differential-Rechnung gibt es jeweils ein $c_1 \in [x_1, x_0]$ und ein $c_2 \in [x_0, x_2]$, so dass

$$f'(c_1) = \frac{f(x_0) - f(x_1)}{x_0 - x_1} \quad \text{und} \quad f'(c_2) = \frac{f(x_2) - f(x_0)}{x_2 - x_0}$$

gilt. Da $f''(x) \geq 0$ ist, wissen wir außerdem, dass die Funktion $f'(x)$ monoton steigend ist. Da offenbar $c_1 \leq c_2$ ist, folgt daraus die Ungleichung $f'(c_1) \leq f'(c_2)$ und damit gilt

$$\frac{f(x_0) - f(x_1)}{x_0 - x_1} \leq \frac{f(x_2) - f(x_0)}{x_2 - x_0}. \tag{5.4}$$

Es gilt

$$x_0 - x_1 = t \cdot x_1 + (1-t) \cdot x_2 - x_1 = (1-t) \cdot (x_2 - x_1)$$

und genauso sehen wir

$$x_2 - x_0 = x_2 - \big(t \cdot x_1 + (1-t) \cdot x_2\big) = t \cdot (x_2 - x_1).$$

Multiplizieren wir daher die Ungleichung (5.4) mit $t \cdot (1-t) \cdot (x_2 - x_1)$, so erhalten wir die Ungleichung

$$t \cdot \big(f(x_0) - f(x_1)\big) \leq (1-t) \cdot \big(f(x_2) - f(x_0)\big).$$

Addieren wir auf beiden Seiten der Gleichung $(1-t)\cdot f(x_0)$ und $t\cdot f(x_1)$ und setzen dann noch für x_0 den Wert $t\cdot x_1 + (1-t)\cdot x_2$ ein, so erhalten wir die Ungleichung

$$f\big(t\cdot x_1 + (1-t)\cdot x_2\big) \leq t\cdot f(x_1) + (1-t)\cdot f(x_2).$$

Das ist aber gerade die Konvexität der Funktion f. □

5.4 Die Exponential-Funktion

Wir wollen in diesem Abschnitt zeigen, dass für die früher definierte Exponential-Funktion, die wir als

$$\exp(x) := \sum_{n=0}^{\infty} \frac{1}{n!}\cdot x^n$$

definiert haben, die Gleichung

$$\exp(x) = e^x \quad \text{mit } e := \sum_{n=0}^{\infty} \frac{1}{n!}$$

gilt. Die oben definierte Zahl e hat den Wert

$$e = 2.718\,281\,828\,459\,045\,235\,360\,287\,471\,352\,662\,497\,757\,247\,093\,699\,959\,574\,966\,967\,627\,724\cdots$$

und wird als Eulersche Zahl (Leonhard Euler, 1707–1783) bezeichnet. Zum Nachweis der oben behaupteten Gleichung benötigen wir das folgende Lemma.

Lemma 68 Ist die Funktion $f:\mathbb{R}\to\mathbb{R}$ für alle $x\in\mathbb{R}$ differenzierbar und gilt

$$f'(x) = 0 \quad \text{für alle } x\in\mathbb{R}$$

so ist die Funktion f konstant: Es gibt dann ein $c\in\mathbb{R}$ so dass

$$f(x) = c \quad \text{für alle } x\in\mathbb{R} \text{ ist.}$$

Beweis: Wir führen den Beweis indirekt und nehmen an, dass die Funktion f nicht konstant ist. Es gibt dann also zwei Zahlen $x_1, x_2\in\mathbb{R}$, so dass

$$x_1 \neq x_2 \quad \text{und} \quad f(x_1)\neq f(x_2)$$

gilt. O.B.d.A. sei $x_1 < x_2$. Nach dem Mittelwert-Satz gibt es nun ein $c\in[x_1,x_2]$, so dass

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

gilt. Nach Voraussetzung wissen wir, dass $f'(c)=0$ ist. Also haben wir

$$0 = \frac{f(x_2)-f(x_1)}{x_2-x_1}.$$

Multiplikation dieser Gleichung mit x_2-x_1 liefert die Gleichung

$$0 = f(x_2)-f(x_1)$$

und daraus folgt sofort $f(x_1)=f(x_2)$. Damit ist die Annahme $f(x_1)\neq f(x_2)$ widerlegt. □

Aufgabe 35: Zeigen Sie: Ist die Funktion $f:\mathbb{R}\to\mathbb{R}$ zweimal differenzierbar und gilt $f''(x)=0$ für alle $x\in\mathbb{R}$, so gibt es Zahlen $c,d\in\mathbb{R}$, so dass

$$\forall x\in\mathbb{R}: f(x) = c\cdot x + d$$

gilt. Überlegen Sie, wie Sie diese Aussage so verallgemeinern können, dass die verallgemeinerte Aussage für beliebige n-mal differenzierbare Funktionen $f:\mathbb{R}\to\mathbb{R}$ gilt, für deren n-te Ableitung

$f^{(n)}(x) = 0$ für alle $x \in \mathbb{R}$ ist. ◇

Aufgabe 36: Zeigen Sie, dass für alle $x \in \mathbb{R}$

$$\exp(x) \cdot \exp(-x) = 1$$

gilt. Bei Ihrem Beweis sollen Sie die Gleichung $\exp(x+y) = \exp(x) \cdot \exp(y)$ nicht benutzen! Folgern Sie aus der von Ihnen gezeigten Gleichung, dass die Exponential-Funktion keine Nullstelle hat. ◇

Aus dem letzten Lemma folgt eine wichtige Charakterisierung der Exponential-Funktion.

Lemma 69 Die Funktion $f : \mathbb{R} \to \mathbb{R}$ sei für alle $x \in \mathbb{R}$ differenzierbar und es gelte

$$f'(x) = \lambda \cdot f(x) \quad \text{für ein } \lambda \in \mathbb{R}.$$

Dann gibt es ein $c \in \mathbb{R}$, so dass

$$f(x) = c \cdot \exp(\lambda \cdot x) \quad \text{für alle } x \in \mathbb{R} \text{ ist.}$$

Beweis: Wir definieren die Funktion $g : \mathbb{R} \to \mathbb{R}$ als

$$g(x) := f(x) \cdot \exp(-\lambda \cdot x).$$

Dann ist die Funktion g differenzierbar und es gilt

$$\begin{aligned} g'(x) &= f'(x) \cdot \exp(-\lambda \cdot x) + f(x) \cdot (-\lambda) \cdot \exp(-\lambda \cdot x) \\ &= \lambda \cdot f(x) \cdot \exp(-\lambda \cdot x) - \lambda \cdot f(x) \cdot \exp(-\lambda \cdot x) \\ &= 0 \end{aligned}$$

Nach dem letzten Lemma (Lemma 68) muss die Funktion g konstant sein. Damit gilt

$$g(x) = g(0) = f(0) \cdot \exp(0) = f(0) \cdot 1 = f(0).$$

Wir definieren $c := f(0)$. Setzen wir in der letzten Gleichung die Definition der Funktion g ein, so haben wir

$$f(x) \cdot \exp(-\lambda \cdot x) = c.$$

Mutliplizieren wir diese Gleichung mit $\exp(\lambda \cdot x)$ und berücksichtigen, dass wir in der letzten Aufgabe gezeigt haben, dass $\exp(\lambda \cdot x) \cdot \exp(-\lambda \cdot x) = 1$ ist, dann erhalten wir die Gleichung

$$f(x) = c \cdot \exp(\lambda \cdot x).$$ □

Aus dem letzten Satz können wir nun die Funktional-Gleichung der Exponential-Funktion folgern.

Satz 70 (Funktional-Gleichung der Exponential-Funktion) Für alle $x, y \in \mathbb{R}$ gilt

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

Beweis: Für ein gegebenes, festes $y \in \mathbb{R}$ definieren wir die Funktion $f : \mathbb{R} \to \mathbb{R}$ durch

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

Dann gilt

$$f_y'(x) = 1 \cdot \exp(x + y) = f_y(x).$$

Nach dem letzten Lemma gilt also

$$f_y(x) = c \cdot \exp(x). \tag{5.5}$$

Da diese Gleichung auch für $x = 0$ gilt und da $\exp(0) = 1$ ist, haben wir

$$f_y(0) = c.$$

Setzen wir hier die Definition von $f_y(x)$ ein, so folgt

$$\exp(0 + y) = c, \quad \text{also } c = \exp(y).$$

Setzen wir dies zusammen mit der Definition von f_y in Gleichung (5.5) ein, so erhalten wir

$$\exp(x + y) = \exp(y) \cdot \exp(x). \qquad \square$$

Bemerkung: Mit Hilfe der Funktional-Gleichung der Exponential-Funktion können wir nun für beliebige $\lambda \in \mathbb{R}_0$ und $x \in \mathbb{R}$ den Ausdruck λ^x definieren. Wir betrachten zunächst den Spezialfall $\lambda = e$: Ist $n \in \mathbb{N}$, so können wir mit Hilfe der Funktional-Gleichung durch eine leichte Induktion nach n zeigen, dass

$$\exp(n) = e^n$$

ist. Aufgrund der Gleichung

$$\exp(x) \cdot \exp(-x) = 1$$

folgt daraus, dass auch für negative ganze Zahlen $m \in \mathbb{Z}$

$$\exp(m) = e^m$$

gilt, denn wenn $m = -n$ mit $n \in \mathbb{N}$ ist, haben wir

$$e^m = e^{-n} = \frac{1}{e^n} = \frac{1}{\exp(n)} = \exp(-n) = \exp(m).$$

Ist nun $\frac{p}{q} \in \mathbb{Q}$, wobei $p \in \mathbb{Z}$ und $q \in \mathbb{N}$ gilt, so haben wir nach dem bisher gezeigten

$$e^p = \exp(p).$$

Ziehen wir hier die q-te Wurzel, so haben wir

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

gezeigt, denn es gilt

$$\left(\exp\Big(\frac{p}{q}\Big)\right)^q = \exp\Big(q \cdot \frac{p}{q}\Big) = \exp(p).$$

Damit haben wir also nun für alle rationalen Zahlen $r \in \mathbb{Q}$ die Gleichung

$$e^r = \exp(r)$$

gezeigt. Es stellt sich die Frage, wie wir am sinnvollsten den Wert von Ausdrücken wie

$$e^{\sqrt{2}}$$

definieren können. Es ist naheliegend, für beliebige reelle Zahlen $x \in \mathbb{R}$ den Wert e^x als

$$e^x := \exp(x)$$

zu definieren. Für beliebige $\lambda \in \mathbb{R}_+$ setzen wir dann

$$\lambda^x := \texttt{exp}\big(x \cdot \ln(\lambda)\big).$$

Mit Hilfe der Funktional-Gleichung der Exponential-Funktion können Sie nun leicht nachweisen, dass für die so definierte Potenz die aus der Schule bekannten Potenz-Gesetze gelten. $\diamond$

Kapitel 6

Anwendungen der Theorie

In diesem Kapitel stellen wir verschiedene Anwendungen der bisher entwickelten Theorie vor. Zunächst zeigen wir, wie sich bestimmte transzendente Funktionen wie der natürliche Logarithmus und die trigonometrischen Funktionen effektiv mit Hilfe von Reihen berechnen lassen. Anschließend diskutieren wir, wann eine Funktion sich durch ein Polynom interpolieren lässt. Danach besprechen wir das Newton'sche Verfahren zur Bestimmung von Nullstellen, dass dann anwendbar ist, wenn die Funktion, deren Nullstelle bestimmt werden soll, differenzierbar ist. Außerdem untersuchen wir die Konvergenz von Fixpunkt-Verfahren und zeigen, wie sich lineare Gleichungs-Systeme mit Hilfe von Fixpunkt-Verfahren approximativ lösen lassen.

6.1 Taylor-Reihen

Es sei $f : \mathbb{R} \to \mathbb{R}$ eine Funktion, die beliebig oft differenzierbar ist. Wir stellen uns die Frage, ob es möglich ist, eine solche Funktion als Potenzreihe darzustellen, wir fragen also, ob es eine Folge $\bigl(a_n\bigr)_{n\in\mathbb{N}}$ gibt, so dass

$$f(x) = \sum_{n=0}^{\infty} a_n \cdot x^n \tag{6.1}$$

gilt. Falls eine solche Folge $\bigl(a_n\bigr)_{n\in\mathbb{N}}$ existiert, dann möchten wir diese Folge berechnen können. Wenn die Gleichung (6.1) gültig ist, dann können wir den Koeffizienten a_0 dadurch berechnen, dass wir in dieser Gleichung $x = 0$ setzen. Wir erhalten dann

$$f(0) = a_0 + \sum_{n=1}^{\infty} a_n \cdot 0^n = a_0. \tag{6.2}$$

Um den Koeffizienten a_1 zu berechnen, differenzieren wir Gleichung (6.1):

$$\frac{df}{dx}(x) = a_1 \cdot 1 \cdot x^0 + \sum_{n=2}^{\infty} a_n \cdot n \cdot x^{n-1}. \tag{6.3}$$

Setzen wir in dieser Gleichung $x = 0$, so finden wir

$$\frac{df}{dx}(0) = a_1 + \sum_{n=2}^{\infty} a_n \cdot n \cdot 0^{n-1} = a_1. \tag{6.4}$$

Allgemein können wir den Koeffizienten a_k dadurch bestimmen, dass wir Gleichung (6.1) k-mal nach x differenzieren und anschließend $x = 0$ setzen. Wir beweisen zunächst durch Induktion über

k, dass für alle $k \in \mathbb{N}_0$

$$\begin{aligned} f^{(k)}(x) &= \sum_{n=k}^{\infty} a_n \cdot n \cdot (n-1) \cdot \cdots \cdot \big(n-(k-1)\big) \cdot x^{n-k} \\ &= \sum_{n=k}^{\infty} a_n \cdot \left(\prod_{i=0}^{k-1}(n-i)\right) \cdot x^{n-k} \\ &= \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} \cdot a_n \cdot x^{n-k} \end{aligned} \tag{6.5}$$

gilt. Hierbei bezeichnet $f^{(k)}(x)$ die k-te Ableitung der Funktion f an der Stelle x.

I.A.: $k = 0$. Es gilt

$$\begin{aligned} f^{(0)}(x) &= f(x) \\ &= \sum_{n=0}^{\infty} a_n \cdot x^n \\ &= \sum_{n=k}^{\infty} \frac{n!}{(n-0)!} \cdot a_n \cdot x^{n-k}. \end{aligned}$$

I.S.: $k \mapsto k+1$. Es gilt

$$\begin{aligned} f^{(k+1)}(x) &= \frac{df}{dx}^{(k)}(x) \\ &\overset{IV}{=} \frac{d}{dx} \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} \cdot a_n \cdot x^{n-k} \\ &= \sum_{n=k+1}^{\infty} \frac{n!}{(n-k)!} \cdot (n-k) \cdot a_n \cdot x^{n-k-1} \\ &= \sum_{n=k+1}^{\infty} \frac{n!}{(n-k-1)!} \cdot a_n \cdot x^{n-(k+1)} \\ &= \sum_{n=k+1}^{\infty} \frac{n!}{\big(n-(k+1)\big)!} \cdot a_n \cdot x^{n-(k+1)} \end{aligned}$$

Damit ist der Beweis von Gleichung (6.5) abgeschlossen. Setzen wir in dieser Gleichung für x den Wert 0 ein, so erhalten wir

$$\begin{aligned} f^{(k)}(0) &= \frac{k!}{(k-k)!} \cdot a_k + \sum_{n=k+1}^{\infty} \frac{n!}{(n-k)!} \cdot a_n \cdot 0^{n-k} \\ &= k! \cdot a_k \end{aligned}$$

Dividieren wir diese Gleichung durch $k!$, so haben wir für die Koeffizienten der Taylor-Reihe die Formel

$$a_k = \frac{f^{(k)}(0)}{k!}$$

gefunden. Also definieren wir für eine Funktion $f : \mathbb{R} \to \mathbb{R}$, die im Punkt $x = 0$ beliebig oft differenzierbar ist, die der Funktion f zugeordnete *Taylor-Reihe* als

$$\boxed{\mathit{taylor}(f,x) := \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} \cdot x^n.} \tag{6.6}$$

Bemerkung: Im Allgemeinen wissen wir nicht, ob die Reihe $taylor(f, x)$ konvergiert. Selbst wenn die Reihe konvergiert folgt daraus noch nicht, dass $f(x) = taylor(f, x)$ ist. Als Beispiel dazu betrachten wir die Funktion $f : \mathbb{R} \to \mathbb{R}$, die durch

$$f(x) := \begin{cases} \exp\left(-\frac{1}{x^2}\right) & \text{falls } x \neq 0 \\ 0 & \text{falls } x = 0 \end{cases}$$

definiert ist. Im Buch von Otto Forster [1] wird gezeigt, dass für diese Funktion die Werte sämtlicher Ableitungen an der Stelle $x = 0$ verschwinden. Damit gilt dann $taylor(f, x) = 0$. ◇

6.1.1 Der Abbruch-Fehler bei der Taylor-Reihe

Um zu untersuchen, wann die Taylor-Reihe $taylor(f, x)$ gegen $f(x)$ konvergiert, definieren wir zu einer gegebenen Funktion f und einer natürlichen Zahl $n \in \mathbb{N}_0$ den *Abbruch-Fehler vom Grad n* als

$$error_n(x) := f(x) - \sum_{i=0}^{n} \frac{f^{(i)}(0)}{i!} \cdot x^i.$$

Der Abbruch-Fehler gibt also an, wie groß der Fehler ist, wenn wir die Berechnung der Summe einer Taylor-Reihe nach dem n-ten Glied abbrechen. Wir berechnen eine Abschätzung für den Abbruch-Fehler $error_n(x)$. Dazu benutzen wir den erweiterten Mittelwert-Satz. Zunächst bemerken wir, dass für alle $k = 0, \cdots, n$ die k-te Ableitung des Abbruch-Fehlers vom Grad n den Wert 0 hat:

$$error_n^{(k)}(0) = 0$$

Dies folgt aus der Definition des Abbruch-Fehlers, denn wir hatten die Taylor-Reihe ja gerade so definiert, dass Sie mit der Funktion f an der Stelle 0 in allen Ableitungen übereinstimmt. Jetzt wenden wir auf die Funktionen $error_n(x)$ und $g_0(x) := x^{n+1}$ in dem Intervall $[0, x]$ den erweiterten Mittelwert-Satz an. Dann gibt es ein $\chi_1 \in [0, x]$, so dass

$$\frac{\frac{d}{dx} error_n(\chi_1)}{\frac{d}{dx} g_0(\chi_1)} = \frac{error_n(x) - error_n(0)}{g_0(x) - g_0(0)} \tag{6.7}$$

gilt. Für die Ableitung der Funktion $g_0(x) = x^{n+1}$ finden wir $\frac{d}{dx} g_0(x) = (n+1) \cdot x^n$. Wegen $error_n(0) = 0$ und $g_0(0) = 0$ vereinfacht sich Gleichung (6.7) zu

$$\frac{error_n^{(1)}(\chi_1)}{(n+1) \cdot \chi_1^n} = \frac{error_n(x)}{x^{n+1}}. \tag{6.8}$$

Nun wenden wir in dem Interval $[0, \chi_1]$ den erweiterten Mittelwert-Satz auf die beiden Funktionen $error_n^{(1)}(x)$ und $g_1(x) := (n+1) \cdot x^n$ an. Dann gibt es ein $\chi_2 \in [0, \chi_1]$, so dass

$$\frac{\frac{d}{dx} error_n^{(1)}(\chi_2)}{\frac{d}{dx} g_1(\chi_2)} = \frac{error_n^{(1)}(\chi_1) - error_n^{(1)}(0)}{g_1(\chi_1) - g_1(0)} \tag{6.9}$$

gilt. Für die Ableitung der Funktion $g_1(x) = (n+1) \cdot x^n$ finden wir $\frac{d}{dx} g_1(x) = (n+1) \cdot n \cdot x^{n-1}$. Wegen $error_n^{(1)}(0) = 0$ und $g_1(0) = 0$ vereinfacht sich Gleichung (6.9) unter Berücksichtigung von Gleichung (6.8) zu

$$\frac{error_n^{(2)}(\chi_2)}{(n+1)\cdot n\cdot \chi_2^{n-1}} = \frac{error_n^{(1)}(\chi_1)}{(n+1)\cdot \chi_1^{n}} = \frac{error_n(x)}{x^{n+1}}.$$

Dieses Spiel können wir fortsetzen. Wenn wir k-mal den erweiterten Mittelwert-Satz anwenden und $k \leq n$ ist, erhalten wir ein $\chi_k \in [0, \chi_{k-1}]$, so dass gilt:

$$\frac{error_n^{(k)}(\chi_k)}{\frac{(n+1)!}{(n+1-k)!}\cdot \chi_k^{n+1-k}} = \frac{error_n(x)}{x^{n+1}} \tag{6.10}$$

Um diese Behauptung per Induktion nach k zu beweisen, bemerken wir, dass der Induktions-Anfang $k = 1$ bereits bewiesen wurde. Im Induktions-Schritt wenden wir in dem Interval $[0, \chi_k]$ auf die beiden Funktionen $error_n^{(k)}(\chi_k)$ und $g_k(x) := \frac{(n+1)!}{(n+1-k)!}\cdot x^{n+1-k}$ den erweiterten Mittelwert-Satz an. Wir finden dann ein $\chi_{k+1} \in [0, \chi_k]$, so dass

$$\frac{\frac{d}{dx} error_n^{(k)}(\chi_{k+1})}{\frac{d}{dx} g_k(\chi_{k+1})} = \frac{error_n^{(k)}(\chi_k) - error_n^{(k)}(0)}{g_k(\chi_k) - g_k(0)} \tag{6.11}$$

gilt. Für die Ableitung der Funktion $g_k(x)$ finden wir

$$\begin{aligned}
\frac{d}{dx} g_k(x) &= \frac{(n+1)!}{(n+1-k)!}\cdot (n+1-k)\cdot x^{n+1-k-1} \\
&= \frac{(n+1)!}{(n+1-(k+1))!}\cdot x^{n+1-(k+1)} \\
&= g_{k+1}(x)
\end{aligned}$$

Wegen $error_n^{(k)}(0) = 0$ und $g_k(0) = 0$ vereinfacht sich Gleichung (6.11) zu

$$\frac{error_n^{(k+1)}(\chi_{k+1})}{g_{k+1}(\chi_{k+1})} = \frac{error_n^{(k)}(\chi_k)}{g_{k+1}(\chi_k)}$$

Berücksichtigen wir hier noch die Induktions-Voraussetzung (6.10), so haben wir

$$\frac{error_n^{(k+1)}(\chi_{k+1})}{g_{k+1}(\chi_{k+1})} = \frac{error_n(x)}{x^{n+1}} \tag{6.12}$$

gefunden und dadurch die Formel (6.10) per Induktion nachgewiesen. Setzen wir in der Gleichung (6.10) für k den Wert n ein, so haben wir

$$\frac{error_n^{(n)}(\chi_n)}{(n+1)!\cdot \chi_n} = \frac{error_n(x)}{x^{n+1}} \tag{6.13}$$

gezeigt. Wir wenden nun den erweiterten Mittelwert-Satz auf die Funktionen $error_n^{(n)}(x)$ und $x \mapsto (n+1)!\cdot x$ an. Dann erhalten wir ein $\chi \in [0, \chi_n] \subseteq [0, x]$, so dass

$$\frac{\frac{d}{dx} error_n^{(n)}(\chi)}{\frac{d}{dx}(n+1)!\cdot x(\chi)} = \frac{error_n^{(n)}(\chi_n) - error_n^{(n)}(0)}{(n+1)!\cdot \chi_n - (n+1)!\cdot 0} \tag{6.14}$$

gilt. Wegen $error_n^{(n)}(0) = 0$ haben wir also

$$\frac{error_n^{(n+1)}(\chi)}{(n+1)!} = \frac{error_n^{(n)}(\chi_n)}{(n+1)!\cdot \chi_n}. \tag{6.15}$$

Um diese Gleichung zu vereinfachen, errinnern wir daran, dass $error_n(x)$ als

$$error_n(x) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} \cdot x^k$$

definiert ist. Wenn wir die $(n+1)$-te Ableitung der Funktion $error_n(x)$ bilden, dann bleibt von der Summe nichts über, es gilt also

$$error_n^{(n+1)}(x) = f^{(n+1)}(x).$$

Setzen wir dieses Ergebnis in Gleichung (6.15) ein und berücksichtigen Gleichung (6.13), so finden wir

$$\frac{f^{(n+1)}(\chi)}{(n+1)!} = \frac{error_n(x)}{x^{n+1}}. \tag{6.16}$$

Setzen wir hier die Definition von $error_n(x)$ ein und multiplizieren die Gleichung mit x^{n+1}, so haben wir gezeigt, dass es ein $\chi \in [0, x]$ gibt, so dass

$$\boxed{f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} \cdot x^k + f^{(n+1)}(\chi) \cdot \frac{x^{n+1}}{(n+1)!}} \tag{6.17}$$

gilt. Diese Formel bezeichnen wir als die *Taylor-Entwicklung* (Brook Taylor, 1685 – 1731) der Funktion f mit *Lagrange'schem Restglied* (Joseph Louis Lagrange, 1736 – 1813).

6.2 Beispiele von Taylor-Entwicklungen

Wir zeigen nun, wie wir transzendente Funktionen mit Hilfe der Taylor-Entwicklungen approximieren können. Dadurch werden diese Funktionen einer numerischen Behandlung zugänglich.

6.2.1 Berechnung des natürlichen Logarithmus

Wir beginnen mit dem natürlichen Logarithmus $x \mapsto \ln(x)$. Dieser ist als die Umkehrfunktion der Exponential-Funktion definiert, es gilt also

$$\ln\bigl(\exp(x)\bigr) = x.$$

Da die Exponential-Funktion immer positiv ist, ist der natürliche Logarithmus für $x \leq 0$ nicht definiert ist. Wir betrachten daher die Funktion $f(x) := \ln(1+x)$. Zunächst berechnen wir die Ableitungen dieser Funktion. Wir beweisen durch Induktion, dass für alle natürlichen Zahlen $n \geq 1$ die n-te Ableitung der Funktion f die folgende Form hat:

$$f^{(n)}(x) = (-1)^{n+1} \cdot \frac{(n-1)!}{(1+x)^n}$$

I.A.: $n = 1$. Es gilt

$$\frac{d}{dx} f(x) = \frac{d}{dx} \ln(1+x) = \frac{1}{1+x} = (-1)^{1+1} \cdot \frac{(1-1)!}{(1+x)^1}$$

I.S.: $n \mapsto n+1$. Wir haben

$$\begin{aligned}
f^{(n+1)}(x) &= \frac{d}{dx} f^{(n)}(x) \\
&\overset{IV}{=} \frac{d}{dx} \left((-1)^{n+1} \cdot \frac{(n-1)!}{(1+x)^n} \right) \\
&= (-1)^{n+1} \cdot (n-1)! \cdot \frac{(-n)}{(1+x)^{n+1}} \\
&= (-1)^{(n+1)+1} \cdot \frac{n!}{(1+x)^{n+1}}
\end{aligned}$$

Daraus folgt sofort

$$f^{(n)}(0) = (-1)^{n+1} \cdot \frac{(n-1)!}{(1+0)^n} = (-1)^{n+1} \cdot (n-1)!$$

Damit erhalten wir für die Taylor-Entwicklungen der Funktion $\ln(1+x)$ das Ergebnis

$$\mathit{taylor}\big(x \mapsto \ln(1+x), x\big) = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \frac{(k-1)! \cdot x^k}{k!} = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \frac{x^k}{k}. \tag{6.18}$$

Wir wollen nun zeigen, dass diese Taylor-Reihe tatsächlich gegen $\ln(1+x)$ konvergiert, wir wollen also zeigen, dass

$$\mathit{taylor}\big(x \mapsto \ln(1+x), x\big) = \ln(1+x)$$

gilt. Dazu betrachten wir die Taylor-Entwicklung mit dem Lagrange'schen Restglied:

$$\ln(1+x) = \sum_{k=1}^{n} (-1)^{k+1} \cdot \frac{x^k}{k} + (-1)^n \cdot \frac{1}{(1+\chi)^{n+1}} \cdot \frac{x^{n+1}}{n+1} \tag{6.19}$$

Für den Abbruch-Fehler haben wir also

$$\mathit{error}_n(x) = (-1)^n \cdot \frac{1}{(1+\chi)^{n+1}} \cdot \frac{x^{n+1}}{n+1}$$

mit $\chi \in [0, x]$ und für $x \in (-1, 1]$ geht dieser Wert für $n \to \infty$ gegen 0. Damit haben wir insgesamt $\mathit{taylor}\big(x \mapsto \ln(1+x), x\big) = \ln(1+x)$ für $x \in (-1, 1]$ gezeigt und folglich können wir

$$\boxed{\ln(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \frac{x^k}{k}}$$

schreiben.[1] Setzen wir hier für x den Wert 1 ein, so haben wir die Formel

$$\ln(2) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} \pm \cdots$$

gefunden. Um den Abbruch-Fehler abzuschätzen, setzen wir in $\mathit{error}_n(x)$ für x den Wert 1 ein und finden

$$|\mathit{error}_n(1)| \leq \frac{1}{n+1}.$$

Um $\ln(2)$ also nach der obigen Formel auf eine Genauigkeit von 10^{-9} berechnen zu können, müssten wir 1 000 000 000 Terme aufsummieren! Es geht auch effizienter. Dazu ersetzen wir in Gleichung (6.18) x durch $-x$ und erhalten

$$\ln(1-x) = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \frac{(-x)^k}{k} = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \frac{(-1)^k \cdot x^k}{k} = -\sum_{k=1}^{n} \frac{x^k}{k} \tag{6.20}$$

[1] Die Formel gilt auch für für negative x, deren Betrag kleiner als 1 ist, aber das können wir mit unseren Mitteln nicht beweisen.

Subtrahieren wir diese Gleichung von der Gleichung (6.19), so erhalten wir

$$\begin{aligned}
\ln\Big(\frac{1+x}{1-x}\Big) &= \ln(1+x) - \ln(1-x) \\
&= \sum_{k=1}^{\infty}(-1)^{k+1}\cdot\frac{x^k}{k} + \sum_{k=1}^{n}\frac{x^k}{k} \\
&= \sum_{k=1}^{\infty}\big((-1)^{k+1}+1\big)\cdot\frac{x^k}{k} \\
&= 2\cdot\sum_{n=0}^{\infty}\frac{x^{2\cdot n+1}}{2\cdot n+1} \qquad (6.21)
\end{aligned}$$

Setzen wir hier für x den Wert $\frac{1}{3}$ ein, so erhalten wir

$$\ln\left(\frac{1+\frac{1}{3}}{1-\frac{1}{3}}\right) = \ln\left(\frac{\frac{4}{3}}{\frac{2}{3}}\right) = \ln(2) = 2\cdot\sum_{n=0}^{\infty}\frac{1}{2\cdot n+1}\cdot\left(\frac{1}{3}\right)^{2\cdot n+1}$$

Um den Fehler e abzuschätzen, den wir erhalten, wenn wir diese Reihe nach dem Glied $2\cdot n+1$ abbrechen, schätzen wir den Abbruch-Fehler wie folgt ab:

$$\begin{aligned}
&\left|\ln\left(\frac{1+\frac{1}{3}}{1-\frac{1}{3}}\right) - 2\cdot\sum_{k=0}^{n}\frac{1}{2\cdot k+1}\cdot\left(\frac{1}{3}\right)^{2\cdot k+1}\right| \\
= \;& 2\cdot\left|\sum_{k=n+1}^{\infty}\frac{1}{2\cdot k+1}\cdot\left(\frac{1}{3}\right)^{2\cdot k+1}\right| \\
\leq \;& 2\cdot\sum_{k=n+1}^{\infty}\left(\frac{1}{3}\right)^{2\cdot k+1} &&= \; 2\cdot\sum_{k=0}^{\infty}\left(\frac{1}{3}\right)^{2\cdot n+2\cdot k+3} \\
= \;& 2\cdot\left(\frac{1}{3}\right)^{2\cdot n+3}\sum_{k=0}^{\infty}\left(\frac{1}{3}\right)^{2\cdot k} &&= \; 2\cdot\left(\frac{1}{3}\right)^{2\cdot n+3}\sum_{k=0}^{\infty}\left(\frac{1}{9}\right)^{k} \\
= \;& 2\cdot\left(\frac{1}{3}\right)^{2\cdot n+3}\frac{1}{1-\frac{1}{9}} &&= \; 2\cdot\left(\frac{1}{3}\right)^{2\cdot n+3}\frac{9}{8} \\
= \;& \frac{1}{4}\cdot\left(\frac{1}{3}\right)^{2\cdot n+1}
\end{aligned}$$

Wir wollen $\ln(2)$ auf eine Genauigkeit von 10^{-9} berechnen. Also wählen wir n so, dass gilt:

$$\begin{aligned}
& \frac{1}{4}\cdot\left(\frac{1}{3}\right)^{2\cdot n+1} && \leq 10^{-9} \\
\Leftrightarrow \quad & \left(\frac{1}{3}\right)^{2\cdot n+1} && \leq 4\cdot 10^{-9} \\
\Leftrightarrow \quad & -\ln(3)\cdot(2\cdot n+1) && \leq \ln(4) - 9\cdot\ln(10) \\
\Leftrightarrow \quad & (2\cdot n+1) && \geq \frac{9\cdot\ln(10)-\ln(4)}{\ln(3)} \\
\Leftrightarrow \quad & n && \geq 0.5\cdot\left(\frac{9\cdot\ln(10)-\ln(4)}{\ln(3)} - 1\right) \approx 8.3 \\
\Leftarrow \quad & n && \geq 9
\end{aligned}$$

Um $\ln(2)$ auf eine Genauigkeit von 10^{-9} zu berechnen reicht es also aus, wenn wir in der Formel (6.21) die ersten 9 Glieder der Summe berücksichtigen. Wir erhalten

$$\ln(2) \approx 2 \cdot \sum_{n=0}^{9} \frac{1}{2 \cdot n + 1} \left(\frac{1}{3}\right)^{2 \cdot n + 1} \approx 0.69314718054981171974$$

Der wirkliche Fehler ist sogar noch kleiner, er beträgt etwa 10^{-11}. Das liegt daran, dass wir bei der Abschätzung der Summe durch die geometrische Reihe den Faktor $\dfrac{1}{2 \cdot k + 1}$ vernachlässigt haben.

Das Verfahren, das wir oben benutzt haben um $\ln(2)$ zu berechnen, lässt sich verallgemeinern. Ist die Aufgabe gegeben, für eine gegebene reelle Zahl r den natürlichen Logarithmus $\ln(r)$ zu berechnen, so setzen wir

$$\begin{aligned}
& r = \frac{1+x}{1-x} \\
\Leftrightarrow \quad & (1-x) \cdot r = 1 + x \\
\Leftrightarrow \quad & r - x \cdot r = 1 + x \\
\Leftrightarrow \quad & r - 1 = x + x \cdot r \\
\Leftrightarrow \quad & r - 1 = x \cdot (1 + r) \\
\Leftrightarrow \quad & \frac{r-1}{r+1} = x
\end{aligned}$$

Bei gegebenem r bestimmen wir also x nach der Formel $x = \dfrac{r-1}{r+1}$. Für das so bestimmte x gilt dann

$$\ln(r) = \ln\Big(\frac{1+x}{1-x}\Big) = 2 \cdot \sum_{n=0}^{\infty} \frac{x^{2 \cdot n + 1}}{2 \cdot n + 1} \quad \text{mit } x = \frac{r-1}{r+1} \tag{6.22}$$

Falls $r \leq 2$ ist, gilt $x \leq \frac{1}{3}$ und dann konvergiert die obige Reihe sehr gut. In modernen Rechnern werden reelle Zahlen y in der Form

$$y = s \cdot r \cdot 2^n \quad \text{mit } s \in \{-1, +1\}, \quad r \in [1, 2) \quad \text{und } n \in \mathbb{Z}$$

dargestellt. Ist y positiv, so lässt sich der natürliche Logarithmus nach der Formel

$$\ln(y) = \ln(r) + n \cdot \ln(2)$$

berechnen, wobei $\ln(r)$ mit Hilfe der Formel (6.22) gefunden wird.

Aufgabe 37: Berechnen Sie die Taylor-Reihen für die Funktionen $x \mapsto \sin(x)$ und $x \mapsto \cos(x)$ geben Sie eine Abschätzung für den Abbruch-Fehler an. Folgern Sie außerdem die auf Leonard Euler (1707 — 1783) zurück gehende Eulersche Formel

$$\boxed{e^{i \cdot x} = \cos(x) + i \cdot \sin(x)},$$

bei der i die imaginäre Einheit bezeichnet, es gilt also $i \cdot i = -1$. ◇

6.2.2 Berechnung des Arcus-Tangens

Die direkte Berechnung der Taylor-Reihe einer Funktion mit Hilfe der Formel (6.6) ist unter Umständen sehr mühsam. Wollen wir beispielsweise die Funktion $x \mapsto \arctan(x)$ in einer Taylor-Reihe entwickeln, so berechnen wir die ersten fünf Ableitungen wie folgt:

1. $\arctan^{(1)}(x) = \dfrac{1}{1+x^2}$.

2. $\arctan^{(2)}(x) = -2 \cdot \dfrac{x}{(1+x^2)^2}$.

3. $\arctan^{(3)}(x) = 2 \cdot \dfrac{3 \cdot x^2 - 1}{(1+x^2)^3}$.

4. $\arctan^{(4)}(x) = -24 \cdot \dfrac{x \cdot (x^2 - 1)}{(1+x^2)^4}$.

5. $\arctan^{(5)}(x) = 24 \cdot \dfrac{1 + 5 \cdot x^4 - 10 \cdot x^2}{(1+x^2)^5}$.

Aufgabe 38: Versuchen Sie, eine allgemeine Formel für die n-te Ableitung der Funktion $x \mapsto \arctan(x)$ zu finden. Beweisen Sie die Richtigkeit Ihrer Formel.

Hinweis: Es gilt

$$\frac{1}{1+x^2} = \frac{1}{2} \cdot \left(\frac{1}{1 + i \cdot x} + \frac{1}{1 - i \cdot x} \right). \qquad \diamond$$

Wir gehen in der Vorlesung einen anderen Weg um die Taylor-Reihe der Arkustangens-Funktion zu berechnen. Dazu stellen wir die Ableitung $\frac{d}{dx} \arctan(x)$ durch eine geometrische Reihe dar:

$$\frac{d}{dx} \arctan(x) = \frac{1}{1+x^2} = \sum_{n=0}^{\infty} \left(-x^2\right)^n = \sum_{n=0}^{\infty} (-1)^n \cdot x^{2 \cdot n}.$$

Die Ableitung der Taylor-Reihe muss diese Reihe ergeben und außerdem muss die Reihe an der Stelle 0 den Wert 0 haben, denn es gilt $\arctan(0) = 0$. Damit finden wir

$$\arctan(x) = \sum_{n=0}^{\infty} (-1)^n \cdot \frac{x^{2 \cdot n + 1}}{2 \cdot n + 1} \tag{6.23}$$

Da $\tan\left(\frac{\pi}{4}\right) = 1$, also $\arctan(1) = \frac{\pi}{4}$ ist, haben wir die Formel

$$\frac{\pi}{4} = \sum_{n=0}^{\infty} (-1)^n \cdot \frac{1}{2 \cdot n + 1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} \pm \cdots \tag{6.24}$$

gefunden. Für einen vollständigen Beweis dieser Formel müssten wir den Abbruch-Fehler nach der Lagrange'schen Formel berechnen. Das würde uns jetzt allerdings zuviel Zeit kosten.

6.2.3 Berechnung von π^*

Zur effizienten Berechnung von π ist die Formel (6.24) nicht geeignet. Aus dem Beweis des Kriteriums von Leibniz für die Konvergenz alternierender Summen folgt, dass der Abbruch-Fehler durch das erste weggelassene Glied abgeschätzt werden kann. Für die obige Formel heißt das, dass der Abbruch-Fehler wie folgt abgeschätzt werden kann:

$$\left| \arctan(x) - \sum_{k=0}^{n} (-1)^k \cdot \frac{1}{2 \cdot k + 1} \right| \leq \frac{1}{2 \cdot (n+1) + 1}$$

Überlegen wir, wieviele Glieder der Summe benötigt werden, um $\frac{\pi}{4}$ auf eine Genauigkeit von 10^{-9} zu berechnen. Dann muss n die folgende Ungleichung erfüllen:

$$\begin{aligned}
& \frac{1}{2 \cdot n + 3} && \leq && 10^{-9} \\
\Leftrightarrow \quad & 2 \cdot n + 3 && \geq && 10^9 \\
\Leftrightarrow \quad & n && \geq && 0.5 \cdot (10^9 - 3) \\
\Leftrightarrow \quad & n && \geq && 499\,999\,998.5 \\
\Leftarrow \quad & n && \geq && 499\,999\,999
\end{aligned}$$

Wir müssten wir also etwa 500 Millionen Terme aufsummieren um die geforderte Genauigkeit zu erreichen. Um eine Formel zu erhalten, die schneller konvergiert, gehen wir von den Additions-Theoremen von Sinus und Cosinus aus. Diese lauten:

$$\sin(\alpha + \beta) = \sin(\alpha) \cdot \cos(\beta) + \cos(\alpha) \cdot \sin(\beta) \quad \text{und}$$
$$\cos(\alpha + \beta) = \cos(\alpha) \cdot \cos(\beta) - \sin(\alpha) \cdot \sin(\beta).$$

Teilen wir die erste Gleichung durch die zweite Gleichung, so folgt

$$\frac{\sin(\alpha + \beta)}{\cos(\alpha + \beta)} = \frac{\sin(\alpha) \cdot \cos(\beta) + \cos(\alpha) \cdot \sin(\beta)}{\cos(\alpha) \cdot \cos(\beta) - \sin(\alpha) \cdot \sin(\beta)}.$$

Da $\tan(x) = \dfrac{\sin(x)}{\cos(x)}$ ist, können wir die linke Seite dieser Gleichung durch $\tan(\alpha + \beta)$ ersetzen. Auf der rechten Seite der Gleichung kürzen wir durch $\cos(\alpha) \cdot \cos(\beta)$. Dann erhalten wir

$$\tan(\alpha + \beta) = \frac{\frac{\sin(\alpha)}{\cos(\alpha)} + \frac{\sin(\beta)}{\cos(\beta)}}{1 - \frac{\sin(\alpha)}{\cos(\alpha)} \cdot \frac{\sin(\beta)}{\cos(\beta)}}.$$

Ersetzen wir hier noch die Brüche der Form $\dfrac{\sin(x)}{\cos(x)}$ durch $\tan(x)$, so haben wir das Additions-Theorem für den Tangens gefunden:

$$\tan(\alpha + \beta) = \frac{\tan(\alpha) + \tan(\beta)}{1 - \tan(\alpha) \cdot \tan(\beta)} \tag{6.25}$$

In dieser Formel setzen wir $\alpha = \arctan(x)$ und $\beta = \arctan(y)$ ein und erhalten

$$\tan(\arctan(x) + \arctan(y)) = \frac{\tan(\arctan(x)) + \tan(\arctan(y))}{1 - \tan(\arctan(x)) \cdot \tan(\arctan(y))}$$

Nehmen wir nun von beiden Seiten dieser Gleichung den Arkustangens und berücksichtigen, dass $\tan(\arctan(x)) = x$ und $\tan(\arctan(y)) = y$ gilt, so erhalten wir das Additions-Theorem für den Arkustangens:

$$\arctan(x) + \arctan(y) = \arctan\left(\frac{x + y}{1 - x \cdot y}\right) \tag{6.26}$$

Hier setzen wir nun $y := x$. Das liefert

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

Wir wollen $\arctan(1)$ berechnen. Daher wählen wir x so, dass Folgendes gilt:

$$\begin{aligned}
& \frac{2 \cdot x}{1 - x^2} && = && 1 \\
\Leftrightarrow \quad & 2 \cdot x && = && 1 - x^2 \\
\Leftrightarrow \quad & x^2 + 2 \cdot x + 1 && = && 2 \\
\Leftarrow \quad & x && = && \sqrt{2} - 1
\end{aligned}$$

Wir können also $x = \sqrt{2} - 1$ wählen und dann π nach der Formel

$$\pi = 4 \cdot \arctan(1) = 8 \cdot \arctan\bigl(\sqrt{2} - 1\bigr) = 8 \cdot \sum_{n=0}^{\infty} (-1)^n \cdot \frac{(\sqrt{2} - 1)^{2 \cdot n + 1}}{2 \cdot n + 1} \tag{6.27}$$

berechnen. Wegen $\sqrt{2} - 1 \approx 0.4142$ konvergiert diese Reihe recht gut. Um auszurechnen, wieviele Glieder benötigt werden um π auf eine Genauigkeit von 10^{-9} zu berechnen, schätzen wir den Abbruch-Fehler mit dem Leibniz-Kriterium ab, wobei wir zur Vereinfachung den Nenner $2 \cdot n + 1$ durch 1 abschätzen:

$$\begin{array}{rrcl}
 & 8 \cdot \bigl(\sqrt{2} - 1\bigr)^{2 \cdot (n+1) + 1} & \leq & 10^{-9} \\
\Leftrightarrow & \ln(8) + (2 \cdot n + 3) \cdot \ln\bigl(\sqrt{2} - 1\bigr) & \leq & -9 \cdot \ln(10) \\
\Leftrightarrow & (2 \cdot n + 3) \cdot \ln\bigl(\sqrt{2} - 1\bigr) & \leq & -9 \cdot \ln(10) - \ln(8) \\
\Leftrightarrow & 2 \cdot n + 3 & \geq & -\dfrac{9 \cdot \ln(10) + \ln(8)}{\ln\bigl(\sqrt{2} - 1\bigr)} \\
\Leftrightarrow & n & \geq & -0.5 \cdot \left(\dfrac{9 \cdot \ln(10) + \ln(8)}{\ln\bigl(\sqrt{2} - 1\bigr)} + 3 \right) \approx 11.4
\end{array}$$

Also reicht es sicher aus, die ersten 12 Glieder der Summe zu berücksichtigen um π auf eine Genauigkeit von 10^{-9} zu berechnen. Führen wir die Rechnung durch, so finden wir

$$\pi \approx \sum_{k=0}^{12} (-1)^k \cdot \frac{(\sqrt{2} - 1)^{2 \cdot k + 1}}{2 \cdot k + 1} \approx 3.141592653601609.$$

Der tatsächliche Fehler ist hier kleiner als $2 \cdot 10^{-11}$.

Die Machin'sche Formel*

Es gibt noch eine elegantere Möglichkeit, die Kreiszahl π mit Hilfe des Arkustangens zu berechnen. Es gilt nämlich die Machin'sche Formel (John Machin, 1686 – 1751):

$$\boxed{\frac{\pi}{4} = 4 \cdot \arctan\Bigl(\frac{1}{5}\Bigr) - \arctan\Bigl(\frac{1}{239}\Bigr)}$$

Beweis: Wegen $\dfrac{\pi}{4} = \arctan(1)$ ist die Machin'sche Formel äquivalent zu

$$\arctan(1) = 4 \cdot \arctan\Bigl(\frac{1}{5}\Bigr) - \arctan\Bigl(\frac{1}{239}\Bigr).$$

Wir addieren auf beiden Seiten dieser Gleichung den Wert $\arctan\bigl(\frac{1}{239}\bigr)$ und sehen dann, dass die Machin'sche Gleichung zu der Gleichung

$$\arctan(1) + \arctan\Bigl(\frac{1}{239}\Bigr) = 4 \cdot \arctan\Bigl(\frac{1}{5}\Bigr).$$

äquivalent ist. Wir wenden nun auf beiden Seiten dieser Gleichung das Additions-Theorem des Arkustangens an und finden

$$\arctan\left(\frac{1 + \frac{1}{239}}{1 - \frac{1}{239}}\right) = 2 \cdot \arctan\left(\frac{\frac{2}{5}}{1 - \frac{1}{25}}\right)$$

Dies vereinfachen wir zu

$$\arctan\Bigl(\frac{240}{238}\Bigr) = 2 \cdot \arctan\Bigl(\frac{10}{24}\Bigr)$$

Kürzen liefert

$$\arctan\Big(\frac{120}{119}\Big) = 2 \cdot \arctan\Big(\frac{5}{12}\Big)$$

Hier können wir auf der rechten Seite das Additions-Theorem des Arkustangens ein zweites Mal anwenden und finden

$$\arctan\Big(\frac{120}{119}\Big) = \arctan\left(\frac{\frac{10}{12}}{1 - \frac{25}{12} \cdot \frac{25}{12}}\right)$$

Elementare Bruchrechnung zeigt die Gültigkeit der Gleichung

$$\frac{\frac{10}{12}}{1 - \frac{25}{12} \cdot \frac{25}{12}} = \frac{120}{119}.$$

Damit haben wir die Machin'sche Formel bewiesen. □

Aufgabe 39*: Leiten Sie die folgende Formel aus dem Additions-Theorem des Arcus-Tangens her und berechnen Sie damit π auf eine Genauigkeit von 10^{-9}:

$$\frac{\pi}{4} = 2 \cdot \arctan\Big(\frac{1}{2}\Big) - \arctan\Big(\frac{1}{7}\Big).$$

6.3 Polynom-Interpolation

Nach dem wir uns im letzten Abschnitt damit beschäftigt haben für eine gegebene Funktion f eine Folge von Polynomen zu konstruieren, deren Ableitungen im Punkt 0 mit der Funktion f übereinstimmen, zeigen wir jetzt, wie sich Polynome konstruieren lassen, die mit eine Funktion f an vorgegebenen Punkten übereinstimmen. Sind $n+1$ Paare der Form

$$\langle x_0, y_0\rangle, \langle x_1, y_1\rangle, \cdots \langle x_n, y_n\rangle$$

gegeben, so besteht die Aufgabe der *Polynom-Interpolation* darin, ein Polynom $p(x)$ vom n zu finden, so dass

$$\forall i \in \{0, 1, \cdots, n\} : p(x_i) = y_i$$

gilt. Wir zeigen sofort, dass diese Aufgabe lösbar ist. Zu einer gegebenen Liste von $n+1$ verschiedenen *Stützstellen*

$$[x_0, x_1, \cdots, x_n]$$

definieren wir für alle $k \in \{0, 1, \cdots, n\}$ das k-te *Lagrange'sche Polynom* (Joseph Louis Lagrange, 1736 – 1813) vom Grad n wie folgt:

$$\boxed{L_k\big([x_0, \cdots, x_n]; x\big) := \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i}} \tag{6.28}$$

Für die Folge der Stützstellen $[-1, 0, 1]$ lauten die Lagrange'schen Polynome beispielsweise

1. $L_0\big([-1,0,1];x\big) := \dfrac{(x-0)\cdot(x-1)}{(-1-0)\cdot(-1-1)} = \dfrac{1}{2}\cdot x^2 - \dfrac{1}{2}\cdot x$
2. $L_1\big([-1,0,1];x\big) := \dfrac{\big(x-(-1)\big)\cdot(x-1)}{(0-(-1))\cdot(0-1)} = -x^2 + 1$
3. $L_2\big([-1,0,1];x\big) := \dfrac{\big(x-(-1)\big)\cdot(x-0)}{(1-(-1))\cdot(1-0)} = \dfrac{1}{2}\cdot x^2 + \dfrac{1}{2}\cdot x$

Ist eine Liste $[x_0, x_1, \cdots, x_n]$ von $n+1$ verschiedenene Stützstellen gegeben, so haben die Lagrange'schen Polynome eine sehr nützliche Eigenschaft. Um diese Eigenschaft einfacher schreiben zu

können, definieren wir für natürliche Zahlen j und k das sogenannte *Kronecker-Delta* wie folgt:

$$\delta_{k,j} = \begin{cases} 1 & \text{falls} \quad j = k; \\ 0 & \text{falls} \quad j \neq k. \end{cases} \tag{6.29}$$

Damit gilt nun

$$L_k\big([x_0, x_1, \cdots, x_n]; x_j\big) = \delta_{k,j}. \tag{6.30}$$

Diese Eigenschaft lässt sich durch einfaches Nachrechnen bestätigen. Wir betrachten die Fälle $j = k$ und $j \neq k$ getrennt:

1. Fall: $j = k$. Dann haben wir

$$L_k\big([x_0, \cdots, x_n]; x_k\big) = \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{x_k - x_i}{x_k - x_i} = 1 = \delta_{k,k}$$

2. Fall: $j \neq k$. Dann haben wir

$$\begin{aligned} L_k\big([x_0, \cdots, x_n]; x_j\big) &= \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{x_j - x_i}{x_k - x_i} \\ &= \frac{(x_j - x_0) \cdot \ldots \cdot (x_j - x_j) \cdot \ldots \cdot (x_j - x_n)}{(x_k - x_0) \cdot \ldots \cdot (x_k - x_j) \cdot \ldots \cdot (x_k - x_n)} \\ &= 0 \\ &= \delta_{j,k} \end{aligned}$$

Die Eigenschaft (6.30) macht es jetzt einfach, Polynome zu konstruieren, die an den Stützstellen $x_0, x_1, \cdots, x_n$ die vorgegebenen Werte y_0, y_1, $\cdots$, y_n annehmen. Wir definieren

$$p(x) := \sum_{k=0}^{n} y_k \cdot L_k(x).$$

Dann gilt $p(x_j) = y_j$ für alle $j = 0, 1, \cdots, n$, denn wir haben

$$p(x_j) = \sum_{k=0}^{n} y_k \cdot L_k(x_j) = \sum_{k=0}^{n} y_j \cdot \delta_{j,k} = y_j \cdot \delta_{j,j} = y_j.$$

Also löst das oben definierte Polynom $p(x)$ das Interpolations-Problem

$$\langle x_0, y_0\rangle,\ \langle x_1, y_1\rangle,\ \cdots,\ \langle x_n, y_n\rangle.$$

Beispiel: Wollen wir ein Polynom $p(x)$ konstruieren, welches das Interpolations-Problem

$$\langle -1, 1\rangle,\ \langle 0, 0\rangle \text{ und } \langle 1, 1\rangle$$

löst, so können wir mit den oben gefundenen Lagrange'schen Polynomen das Polynom $p(x)$ wie folgt definieren:

$$\begin{aligned} p(x) &= 1 \cdot L_0\big([-1, 0, 1]; x\big) + 0 \cdot L_1\big([-1, 0, 1]; x\big) + 1 \cdot L_2\big([-1, 0, 1]; x\big) \\ &= \big(\tfrac{1}{2} \cdot x^2 - \tfrac{1}{2} \cdot x\big) + \big(\tfrac{1}{2} \cdot x^2 + \tfrac{1}{2} \cdot x\big) \\ &= x^2 \end{aligned}$$

Aufgabe 40: Bei einer Klausur können insgesamt n Punkte erreicht werden. Bestimmen Sie ein Polynom $p(x)$ vom Grade 1, so dass

$$p(n) = 1.0 \quad \text{und} \quad p\Big(\frac{n}{2}\Big) = 4.0$$

gilt. Hat ein Teilnehmer einer Klausur k von n Punkten erreicht, so ist $p(k)$ die Note, mit der die Leistung bewertet wird. ◇

6.3.1 Interpolation nach Newton*

Bei der Rechnung mit den oben definierten Lagrange'schen Polynomen tritt in der Praxis ein Problem auf. Hat man für eine gegebene Zahl von Stützstellen das Interpolations-Problem gelöst und erhält man nun eine zusätzliche Stützstelle, so ist es erforderlich, alle Lagrange'schen Polynome noch einmal zu berechnen, denn die Lagrange'schen Polynome vom Grad $n+1$ haben mit den Lagrange'schen Polynomen vom Grad n nur wenig zu tun. Hier ist der Ansatz von Newton besser geeignet. Bei dem Newton'schen Ansatz schreibt sich ein Interpolations-Polynom vom Grad n in der Form

$$\begin{aligned} p_n(x) &= \sum_{k=0}^{n} c_k \cdot \prod_{i=0}^{k-1}(x - x_i) \\ &= c_0 + c_1 \cdot (x - x_0) + c_2 \cdot (x - x_0) \cdot (x - x_1) + \cdots + c_n \cdot \prod_{i=0}^{n-1}(x - x_i) \end{aligned} \tag{6.31}$$

Der Vorteil dieses Ansatzes besteht darin, dass das Newton'sche Interpolations-Polynom vom Grad $n+1$ unmittelbar aus dem Newton'schen Interpolations-Polynom vom Grad n wie folgt hervorgeht:

$$p_{n+1}(x) = p_n(x) + c_{n+1} \cdot \prod_{i=0}^{n}(x - x_i).$$

Ist eine Interpolations-Aufgabe

$$\langle x_0, y_0\rangle,\ \langle x_1, y_1\rangle,\ \cdots,\ \langle x_n, y_n\rangle,$$

gegeben, so können die Koeffizienten c_k für $k = 0, 1, \cdots, n$ der Reihe nach wie folgt berechnet werden.

1. Um c_0 zu bestimmen, setzen wir in Gleichung (6.31) für x den Wert x_0 ein. Dann fallen alle Terme bis auf den ersten Term weg und wir erhalten

   $$y_0 = c_0.$$

2. Um c_1 zu bestimmen, setzen wir in Gleichung (6.31) für x den Wert x_1 ein. Dann fallen alle Terme bis auf die ersten beiden Term weg und wir erhalten

   $$y_1 = c_0 + c_1 \cdot (x_1 - x_0).$$

   Setzen wir hier für c_0 den im letzten Schritt gefundenen Wert y_0 ein, so finden wir

   $$c_1 = \frac{y_1 - y_0}{x_1 - x_0}.$$

3. Um c_2 zu bestimmen, setzen wir in Gleichung (6.31) für x den Wert x_2 ein. Dann fallen alle Terme bis auf die ersten drei Terme weg und wir erhalten

   $$y_2 = c_0 + c_1 \cdot (x_2 - x_0) + c_2 \cdot (x_2 - x_0) \cdot (x_2 - x_1).$$

   Hier setzen wir für c_0 und c_1 die in den letzten Schritten gefundenen Werte ein und haben dann

$$
\begin{aligned}
y_2 &= y_0 + \frac{y_1 - y_0}{x_2 - x_0} \cdot (x_2 - x_0) + c_2 \cdot (x_2 - x_0) \cdot (x_2 - x_1) \\
y_2 - y_0 &= \frac{y_1 - y_0}{x_2 - x_0} \cdot (x_2 - x_0) + c_2 \cdot (x_2 - x_0) \cdot (x_2 - x_1) \\
\frac{y_2 - y_0}{x_2 - x_0} &= \frac{y_1 - y_0}{x_2 - x_0} + c_2 \cdot (x_2 - x_1) \\
\frac{y_2 - y_0}{x_2 - x_0} - \frac{y_1 - y_0}{x_2 - x_0} &= c_2 \cdot (x_2 - x_1) \\
\frac{\dfrac{y_2 - y_0}{x_2 - x_0} - \dfrac{y_1 - y_0}{x_2 - x_0}}{x_2 - x_1} &= c_2
\end{aligned}
$$

Die obige Rechnung gibt Anlass zur Definition der sogenannten *dividierten Differenzen* vom Rang k, die wir jetzt für alle $k = 1, \cdots, n$ durch Induktion über k definieren.

I.A.: $k = 1$. Für alle $i = 0, \cdots, n$ setzen wir

$$[x_k]_{\text{dd}} := y_k.$$

I.S.: $k \mapsto k + 1$. Für $i = 0, \cdots, n - k$ setzen wir

$$[x_i, x_{i+1}, \cdots, x_{i+k}]_{\text{dd}} := \frac{[x_{i+1}, \cdots, x_{i+k}]_{\text{dd}} - [x_i, \cdots, x_{i+k-1}]_{\text{dd}}}{x_{i+k} - x_i}.$$

Die dividierten Differenzen der Ordnung $k+1$ berechnen sich also aus den dividierten Differenzen der Ordnung k durch Bildung einer Differenz und einer anschließenden Division. Dieser Umstand erklärt ihren Namen. Für die Koeffizienten c_k in dem Newton'schen Ansatz (6.31) gilt nun

$$c_k = [x_0, x_1, \cdots, x_k]_{\text{dd}}, \tag{6.32}$$

das Interpolations-Problem ein Polynom $p(x)$ zu finden, für das

$$p(x_0) = y_0, \quad p(x_1) = y_1, \quad \cdots \quad \text{und} \quad p(x_n) = y_n$$

gilt, wird also durch das Polynom

$$\boxed{p(x) = \sum_{k=0}^{n} [x_0, x_1, \cdots, x_k]_{\text{dd}} \cdot \prod_{i=0}^{k-1} (x - x_i)} \tag{6.33}$$

gelöst.

6.3.2 Der Interpolations-Fehler

Wir untersuchen als nächstes, wie groß der Fehler bei der Polynom-Interpolation werden kann. Dazu beweisen wir zunächst den folgenden Hilfs-Satz. Dieser Satz ist eine Verallgemeinerung des Satzes von Rolle (Satz 56).

Satz 71 Ist $f : \mathbb{R} \to \mathbb{R}$ eine Funktion, die n-mal differenzierbar ist und die $n + 1$ verschiedene Null-Stellen

$$x_0 < x_1 < \cdots < x_n$$

hat, so gibt es ein $\xi \in [x_0, x_n]$ mit $f^{(n)}(\xi) = 0$.

Beweis: Wir zeigen, dass für alle $k = 0, 1, \cdots, n$ die k-te Ableitung $f^{(k)}(x)$ in dem Interval $[x_0, x_n]$ mindestens $n + 1 - k$ verschiedene Nullstellen hat. Diesen Nachweis führen wir durch Induktion über k.

I.A.: $k = 0$. Es gilt $f^{(0)}(x) = f(x)$ und da die Funktion f nach Voraussetzung $n+1$ Nullstelle hat, folgt die Behauptung.

I.S.: $k \mapsto k+1$. Nach Induktions-Voraussetzung hat die Funktion $f^{(k)}(x)$ mindestens $n+1-k$ verschiedene Nullstellen. Nehmen wir an, diese Nullstellen seien der Größe nach geordnet als

$$y_1 < y_2 < \cdots y_{n+1-k}.$$

Wegen $k+1 \leq n$ ist die Funktion $f^{(k)}(x)$ nach Voraussetzung differenzierbar und nach dem Satz von Rolle hat die Ableitung dieser Funktion jeweils zwischen zwei Nullstellen y_i und y_{i+1} eine Nullstelle, für $i = 1, \cdots, n-k-1$ gibt es also $z_i \in (y_i, y_{i+1}) \subseteq [x_0, x_n]$ mit

$$\frac{d}{dx} f^{(k)}(z_i) = f^{(k+1)}(z_i) = 0.$$

Setzen wir in der gerade bewiesenen Behauptung für k den Wert n ein, so sehen wir, dass die Funktion $f^{(n)}(x)$ in dem Interval $[x_0, x_n]$ mindestens $n+1-n = 1$ Nullstelle hat, also gibt es das gesuchte $\xi \in [x_0, x_n]$. □

Aufgabe 41: Es sei $p(x)$ ein Polynom vom Grad $n \geq 1$. Zeigen Sie, dass $p(x)$ höchstens n verschiedene Nullstellen hat. Folgern Sie daraus, dass es zu n gegebenen Paaren

$$\langle x_0, y_0 \rangle, \langle x_1, y_1 \rangle, \cdots, \langle x_n, y_n \rangle,$$

genau ein Polynom $p(x)$ gibt, so dass $p(x_i) = y_i$ für alle $i = 0, 1, \cdots, n$ gilt. ◇

Satz 72 Ist die Funktion $f : \mathbb{R} \to \mathbb{R}$ mindestens $(n+1)$-mal differenzierbar, ist $p(x)$ eine Polynom vom Grad kleiner gleich n, sind $x_0 < x_1 < \cdots < x_n$ Punkte mit

$$f(x_i) = p(x_i) \text{ für alle } i = 0, 1, \cdots, n,$$

und ist $\bar{x} \in [x_0, x_n]$, so gibt es ein $\zeta \in [x_0, x_n]$, so dass für den Interpolations-Fehler $f(\bar{x}) - p(\bar{x})$ gilt:

$$\boxed{f(\bar{x}) - p(\bar{x}) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} \cdot \prod_{i=0}^{n} (\bar{x} - x_i).}$$

Beweis: Falls $\bar{x} \in \{x_0, x_1, \cdots, x_n\}$ ist, dann folgt sofort $f(\bar{x}) - p(\bar{x}) = 0$ und da dann auch

$$\prod_{i=0}^{n} (\bar{x} - x_i) = 0$$

gilt, ist die Behauptung in diesem Fall offensichtlich. Andernfalls definieren wir die Funktion

$$g(x) := f(x) - p(x) - \left(\prod_{i=0}^{n} \frac{x - x_i}{\bar{x} - x_i} \right) \cdot \big(f(\bar{x}) - p(\bar{x}) \big).$$

Mit f ist auch die Funktion g mindestens $(n+1)$-mal differenzierbar. Außerdem gilt wegen $p(x_k) = f(x_k)$ für alle $k = 0, 1, \cdots, n$

$$g(x_k) = f(x_k) - p(x_k) - \left(\prod_{i=0}^{n} \frac{x_k - x_i}{\bar{x} - x_i} \right) \cdot \big(f(\bar{x}) - p(\bar{x}) \big) = 0,$$

denn der Faktor $x_k - x_i$ verschwindet im Falle $i = k$. Außerdem haben wir

$$\begin{aligned} g(\bar{x}) &= f(\bar{x}) - p(\bar{x}) - \left(\prod_{i=0}^{n} \frac{\bar{x} - x_i}{\bar{x} - x_i} \right) \cdot \big(f(\bar{x}) - p(\bar{x}) \big) \\ &= f(\bar{x}) - p(\bar{x}) - 1 \cdot \big(f(\bar{x}) - p(\bar{x}) \big) \\ &= 0. \end{aligned}$$

Damit hat die Funktion insgesamt $n+2$ verschiedene Nullstellen. Wenden wir jetzt auf die Funktion

g den eben gezeigten Hilfs-Satz an, so finden wir ein $\zeta \in [x_0, x_n]$ mit

$$g^{(n+1)}(\zeta) = 0.$$

Wir bilden die $(n+1)$-te Ableitung von g und finden

$$g^{(n+1)}(x) = f^{(n+1)}(x) - 0 - (n+1)! \cdot \left(\prod_{i=0}^{n} \frac{1}{\bar{x} - x_i} \right) \cdot \big(f(\bar{x}) - p(\bar{x}) \big), \tag{6.34}$$

denn die $(n+1)$-te Ableitung eines Polynoms vom Grad n ist 0 und die $(n+1)$-te Ableitung des Polynoms

$$\prod_{i=0}^{n} \frac{x - x_i}{\bar{x} - x_i}$$

ist $(n+1)!$ mal der Koeffizient der Potenz x^{n+1}. Setzen wir in Gleichung (6.34) die Nullstelle ζ ein, so finden wir

$$\begin{aligned}
& 0 = f^{(n+1)}(\zeta) - (n+1)! \cdot \left(\prod_{i=0}^{n} \frac{1}{\bar{x} - x_i} \right) \cdot \big(f(\bar{x}) - p(\bar{x}) \big) \\
\Leftrightarrow \quad & (n+1)! \cdot \left(\prod_{i=0}^{n} \frac{1}{\bar{x} - x_i} \right) \cdot \big(f(\bar{x}) - p(\bar{x}) \big) = f^{(n+1)}(\zeta) \\
\Leftrightarrow \quad & f(\bar{x}) - p(\bar{x}) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} \cdot \prod_{i=0}^{n} (\bar{x} - x_i).
\end{aligned}$$

□

Aufgabe 42: Für die Funktion $x \mapsto \sin(x)$ soll im Intervall $[0, \frac{\pi}{2}]$ ein Tabelle erstellt werden, so dass der bei linearer Interpolation entstehende Interpolations-Fehler kleiner als 10^{-5} ist. Das Intervall $[0, \frac{\pi}{2}]$ soll zu diesem Zweck in gleich große Intervalle aufgeteilt werden. Berechnen Sie die Anzahl der Einträge, die für die Erstellung der Tabelle notwendig ist. ◇

6.4 Der Banach'sche Fixpunkt-Satz

Der Banach'sche Fixpunkt-Satz, der 1922 von Stefan Banach (1892 – 1945) bewiesen wurde, ist ein wichtiges Hilfsmittel zur Lösung von Gleichungen. Wir werden den Banach'schen Fixpunkt-Satz nur für den Spezialfall der reellen Zahlen formulieren und beweisen. In der Mathematik wird dieser Satz in einem abstrakteren Rahmen verwendet, die Menge der reellen Zahlen wird dann durch einen *vollständigen metrischen Raum* ersetzt.

Bevor wir den Banach'schen Fixpunkt-Satz formulieren und beweisen, wollen wir das damit verbundene Fixpunkt-Verfahren motivieren. Wir haben bereits einmal ein solches angewendet: In dem Kapitel 3 über Folgen und Reihen hatten wir die Gleichung

$$x = \cos(x)$$

mit Hilfe eines Fixpunkt-Verfahrens gelöst. Wir hatten damals induktiv eine Folge $\big(x_n\big)_{n \in \mathbb{N}}$ definiert, indem wir $x_1 := 0$ und $x_{n+1} := \cos(x_n)$ definiert hatten. Das in Abbildung 3.1 auf Seite 19 gezeigte Programm berechnet die Folge $\big(x_n\big)_{n \in \mathbb{N}}$ und wir hatten gesehen, dass diese Folge gegen einen Grenzwert

$$\bar{x} := \lim_{n \to \infty} x_n$$

konvergiert, der wegen

$$\cos(\bar{x}) = \cos\left(\lim_{n \to \infty} x_n \right) = \lim_{n \to \infty} \cos(x_n) = \lim_{n \to \infty} x_{n+1} = \bar{x}$$

auch eine Lösung der Fixpunkt-Gleichung $x = \cos(x)$ ist. Wir wollen dieses Verfahren nun verallgemeinern. Wir nehmen dazu an, dass eine stetige Funktion f gegeben ist und wir eine Lösung

der Fixpunkt-Gleichung

$$x = f(x)$$

iterativ bestimmen wollen, indem wir induktiv eine Folge $\big(x_n\big)_{n\in\mathbb{N}}$ definieren, wobei $x_{n+1} = f(x_n)$ gelten soll. Den Startwert x_1 werden wir dabei weitgehend willkürlich wählen. Zunächst ist Folgendes klar: Sollte die Folge $\big(x_n\big)_{n\in\mathbb{N}}$ gegen einen Grenzwert $\bar{x}$ konvergieren, so ist dieser Grenzwert eine Lösung der Fixpunkt-Gleichung $x = f(x)$, denn es gilt

$$f(\bar{x}) = f\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = \bar{x}.$$

Dabei haben wir bei der Gleichung

$$f\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} f(x_n)$$

die Stetigkeit von f ausgenutzt. Weniger klar ist die Antwort auf die Frage, wann die durch $x_{n+1} = f(x_n)$ definierte Folge konvergiert. Die entscheidende Antwort auf diese Frage hat Stefan Banach gegeben: Wenn die Funktionswerte von f näher beieinander liegen als die Argumente, dann konvergiert die Folge. Bevor wir das im Detail formal untersuchen, wollen wir die dahinter liegende Anschauung verstehen. Nehmen wir an, dass die Fixpunkt-Gleichung $f(x) = x$ eine Lösung $\bar{x}$ hat und nehmen wir weiter an, dass eine (nicht besonders gute) Näherung x_1 für $\bar{x}$ gegeben ist. Wenn nun die Funktionswerte von f näher beieinander liegen als die Argumente, dann können wir die Näherung x_1 zu einer Näherung x_2 verbessern, indem wir

$$x_2 := f(x)$$

definieren. Warum ist x_2 besser als x_1? Um diese Frage zu beantworten, betrachten wir den Abstand von x_2 und $\bar{x}$. Es gilt

$$\begin{array}{lll} & |x_2 - \bar{x}| & \\ = & |f(x_1) - f(\bar{x})| & \text{denn } x_2 = f(x_1) \text{ und } \bar{x} = f(\bar{x}) \\ < & |x_1 - \bar{x}|, & \end{array}$$

wobei wir bei der letzten Ungleichung benutzt haben, dass die Funktionswerte von f näher beieinander liegen als die Argumente. Um diese Argumentation wasserdicht zu machen, definieren wir zunächst formal unter welchen Umständen die Funktionswerte einer Funktion näher beieinander liegen als die Argumente. Die nun folgende Definition ist etwas schärfer, als Sie es vielleicht im ersten Moment vermuten würden. Das werden wir später noch diskutieren.

Definition 73 (kontrahierend) Eine Funktion $f : [a, b] \to [a, b]$ ist eine *kontrahierende Abbildung* wenn es eine Zahl $q < 1$ gibt, so dass

$$\forall x, y \in [a, b] : |f(x) - f(y)| \leq q \cdot |x - y|.$$

gilt. Wir bezeichnen die Zahl q als den *Kontraktions-Koeffizienten*. ◇

Beispiel: Die Funktion

$$\cos : [0, 1] \to [0, 1]$$

ist eine kontrahierende Abbildung. Zunächst müssen wir uns davon überzeugen, dass diese Funktion wohldefiniert ist. Dazu muss aus $x \in [0, 1]$ folgen, dass auch $\cos(x) \in [0, 1]$ gilt. Dies folgt aus $\cos(0) = 1$, $\cos(1) \approx 0.54$ und der Tatsache, dass die Kosinus-Funktion in dem Intervall $[0, 1]$ monoton fallend ist, denn $\dfrac{d}{dx}\cos(x) = -\sin(x)$ und für alle $x \in \big[0, \pi\big]$ gilt $\sin(x) \geq 0$. Seien nun Zahlen $x, y \in [0, 1]$ gegeben. Nach dem Mittelwert-Satz der Differenzial-Rechnung gibt es dann ein $\zeta \in [x, y]$ mit

$$\frac{\cos(x) - \cos(y)}{x - y} = \cos'(\zeta) = -\sin(\zeta).$$

Die Sinus-Funktion nimmt in dem Intervall $[0,1]$ ihr Maximum in dem Punkt 1 an, es gilt

$$\forall t \in [0,1] : \sin(t) \leq \sin(1) \approx 0.8414709848\cdots \leq 0.85.$$

Also haben wir folgende Abschätzung

$$|\cos(x) - \cos(y)| = \sin(\zeta) \cdot |x - y| \leq 0.85 \cdot |x - y| \quad \text{für alle } x, y \in [0,1]. \qquad \diamond$$

Das letzte Beispiel verallgemeinern wir zu einem Satz.

Satz 74 Ist die Funktion $f : [a,b] \to [a,b]$ in dem Intervall $[a,b]$ differenzierbar und gibt es eine Zahl $q < 1$, so dass

$$\forall t \in [a,b] : |f'(t)| \leq q$$

gilt, dann ist die Abbildung f kontrahierend mit dem Kontraktions-Koeffizienten q.

Beweis: Es seien $x, y \in [a,b]$. Nach dem Mittelwert-Satz der Differenzial-Rechnung gibt es dann ein $\zeta \in [x,y]$ mit

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

Nehmen wir auf beiden Seiten dieser Ungleichung den Betrag, so erhalten wir

$$\frac{|f(x) - f(y)|}{|x - y|} = |f'(\zeta)| \leq q,$$

denn wir hatten ja vorausgesetzt, dass die Ungleichung $|f'(t)| < q$ für alle $t \in [a,b]$ gilt. Multiplizieren wir diese Ungleichung mit $|x - y|$, so erhalten wir

$$|f(x) - f(y)| \leq q \cdot |x - y|$$

und nach Definition einer kontrahierenden Abbildung ist das die Behauptung. $\square$

Satz 75 Ist $f : [a,b] \to [a,b]$ eine kontrahierende Abbildung, so ist f auch stetig.

Aufgabe 43: Beweisen Sie den letzten Satz. $\diamond$

Satz 76 (Banach'scher Fixpunkt-Satz) Es sei $f : [a,b] \to [a,b]$ eine kontrahierende Abbildung mit dem Kontraktions-Koeffizienten q und x_0 sei eine Zahl aus dem Intervall. Definieren wir die Folge $\big(x_n\big)_{n\in\mathbb{N}}$ induktiv durch

$$x_{n+1} = f(x_n),$$

so konvergiert diese Folge. Setzen wir

$$\bar{x} := \lim_{n\to\infty} x_n,$$

so gilt $f(\bar{x}) = \bar{x}$ und darüber hinaus gilt die Abschätzung

$$|x_n - \bar{x}| \leq \frac{q^n}{1-q} \cdot |x_1 - x_0|.$$

Beweis: Wir starten den Beweis mit einer im ersten Moment skurril anmutenden Formel:

$$x_n - x_0 = (x_n - x_{n-1}) + (x_{n-1} - x_{n-2}) + \cdots + (x_2 - x_1) + (x_1 - x_0) = \sum_{i=1}^{n} (x_i - x_{i-1}).$$

Diese Summe wird als *Teleskop-Summe* bezeichnet. Daraus folgt sofort

$$x_n = x_0 + \sum_{i=1}^{n} (x_i - x_{i-1}).$$

Damit gilt dann aber

$$\lim_{n\to\infty} x_n = x_0 + \sum_{i=1}^{\infty}(x_i - x_{i-1}),$$

wenn wir noch zeigen können, dass die auf der rechten Seite dieser Formel auftretende Reihe konvergiert. Dazu zeigen wir, dass die geometrische Reihe eine Majorante dieser Reihe ist. Konkret zeigen wir durch Induktion, dass für alle $i \in \mathbb{N}$

$$|x_{i+1} - x_i| \leq q^i \cdot |x_1 - x_0|$$

gilt. Der Induktions-Anfang ist trivial. Im Induktions-Schritt haben wir

$$|x_{i+2} - x_{i+1}| = |f(x_{i+1}) - f(x_i)| \leq q \cdot |x_{i+1} - x_i| \overset{\text{IV}}{\leq} q \cdot q^i \cdot |x_1 - x_0| = q^{i+1} \cdot |x_1 - x_0|.$$

Nachdem wir jetzt wissen, dass die Folge $\big(x_n\big)_{n\in\mathbb{N}}$ konvergiert, zeigen wir, dass der Grenzwert $\bar{x}$ ein Fixpunkt der Funktion f ist. Da f als kontrahierende Abbildung auch stetig ist, können wir die Anwendung der Funktion f mit der Grenzwert-Bildung vertauschen. Also haben wir

$$f(\bar{x}) = f\Big(\lim_{n\to\infty} x_n\Big) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} x_n = \bar{x}$$

und dies zeigt, dass $\bar{x}$ ein Fixpunkt der Funktion f ist.

Den Abstand zwischen $\bar{x}$ und x_n können wir abschätzen, wenn wir $\bar{x}$ als unendliche Reihe schreiben und gleichzeitig x_n durch eine Teleskop-Summe ausdrücken.

$$\begin{array}{rcll}
|\bar{x} - x_n| & = & \Big| x_0 + \sum_{i=1}^{\infty}(x_i - x_{i-1}) - \Big(x_0 + \sum_{i=1}^{n}(x_i - x_{i-1})\Big) \Big| & \\
& = & \Big| \sum_{i=n+1}^{\infty}(x_i - x_{i-1}) \Big| & \leq \quad \sum_{i=n+1}^{\infty} \big|x_i - x_{i-1}\big| \\
& = & \sum_{i=n}^{\infty} \big|x_{i+1} - x_i\big| & \leq \quad \sum_{i=n}^{\infty} q^i \cdot \big|x_1 - x_0\big| \\
& = & \big|x_1 - x_0\big| \cdot \sum_{i=n}^{\infty} q^i & = \quad \big|x_1 - x_0\big| \cdot \sum_{i=0}^{\infty} q^{n+i} \\
& = & \big|x_1 - x_0\big| \cdot q^n \cdot \sum_{i=0}^{\infty} q^i & = \quad \big|x_1 - x_0\big| \cdot q^n \cdot \dfrac{1}{1-q}
\end{array}$$

Die obige Ungleichungskette zeigt insgesamt die Gültigkeit der Abschätzung

$$|\bar{x} - x_n| \leq \frac{q^n}{1-q} \cdot |x_1 - x_0|$$

und damit ist der Beweis ist abgeschlossen. □

Setzen wir in der eben gezeigten Abschätzung für n den Wert 1 ein, so erhalten wir

$$|\bar{x} - x_1| \leq \frac{q}{1-q} \cdot |x_1 - x_0|.$$

Wenn wir in dieser Ungleichung x_1 durch x_m und x_0 durch x_{m-1} ersetzen, behält die Ungleichung ihre Gültigkeit, denn wir können x_{m-1} ja als den Startwert einer neuen Folge $\big(y_n\big)_{n\in\mathbb{N}}$ ansehen, für die wir $y_0 = x_{m-1}$ und $y_1 = f(y_0) = f(x_{m_1}) = x_m$ setzen. Wir haben dann also

$$|\bar{x} - x_m| \leq \frac{q}{1-q} \cdot |x_m - x_{m-1}|.$$

Wenn wir q kennen, können wir damit die Güte der bisher erreichten Approximation abschätzen.

Beispiel: Wir haben oben gesehen, dass die Abbildung $\cos : [0,1] \to [0,1]$ kontrahierend ist mit einem Kontraktions-Koeffizienten $q \leq 0.85$. Wollen wir den Fixpunkt dieser Abbildung auf eine Genauigkeit von ε berechnen, so müssen wir folglich solange iterieren, bis

$$\frac{q}{1-q} \cdot |x_m - x_{m-1}| \leq \varepsilon,$$

und das ist äquivalent zu

$$|x_m - x_{m-1}| \leq \frac{1-q}{q} \cdot \varepsilon,$$

Setzen wir hier $q = 0.85$ und $\varepsilon = 10^{-6}$, so ist die Abbruchbedingung also

$$|x_m - x_{m-1}| \leq \frac{1-0.85}{0.85} \cdot 10^{-6} \approx 1.77 \cdot 10^{-7}.$$

Wir können auch die Zahl der Iterationen abschätzen, die notwendig sind um den Fixpunkt mit einer Genauigkeit von ε zu berechnen. Wir gehen dazu von der Ungleichung

$$|x_n - \bar{x}| \leq \frac{q^n}{1-q} \cdot |x_1 - x_0|.$$

aus. Starten wir die Fixpunkt-Iteration mit $x_0 = 0$, so erhalten wir $x_1 = \cos(0) = 1$ und damit können wir die Anzahl der Iterationen n abschätzen:

$$\begin{aligned}
& |x_n - \bar{x}| \leq \varepsilon \\
\Leftarrow \quad & \frac{q^n}{1-q} \cdot |x_1 - x_0| \leq \varepsilon \\
\Leftrightarrow \quad & \frac{q^n}{1-q} \cdot |1 - 0| \leq \varepsilon \\
\Leftrightarrow \quad & \frac{q^n}{1-q} \leq \varepsilon \\
\Leftrightarrow \quad & n \cdot \ln(q) - \ln(1-q) \leq \ln(\varepsilon) \\
\Leftrightarrow \quad & n \cdot \ln(q) \leq \ln(\varepsilon) + \ln(1-q) \\
\Leftrightarrow \quad & n \geq \frac{\ln(\varepsilon) + \ln(1-q)}{\ln(q)} \\
\Leftrightarrow \quad & n \geq \frac{\ln(10^{-6}) + \ln(1-0.85)}{\ln(0.85)} \\
\Leftrightarrow \quad & n \geq \frac{-6 \cdot \ln(10) + \ln(1-0.85)}{\ln(0.85)} \\
\Leftrightarrow \quad & n \geq 96.7
\end{aligned}$$

Damit benötigen wir also höchstens 97 Iterationen um die gewünschte Genauigkeit zu erzielen. Wir haben die einzelnen Werte der Folge, die sich bei der iterativen Lösung der Gleichung $x = \cos(x)$ ergibt, in Tabelle 3.1 auf Seite 19 angegeben. Diese Tabelle zeigt, dass bereits nach 36 Iterationen eine Genauigkeit von 10^{-6} erreicht ist. Der Grund dafür, dass es deutlich schneller ging, als wir mit der obigen Abschätzung berechnet haben, liegt darin, dass der Kontraktions-Koeffizient q den Wert von $\sin(x)$ in dem Intervall $[0,1]$ abschätzen muss. Da die Sinus-Funktion in diesem Intervall monoton wächst, haben wir den Kontraktions-Koeffizient als $\sin(1) \approx 0.85$ berechnet. Für die Lösung $\bar{x}$ der Fixpunkt-Gleichung gilt aber $\sin(\bar{x}) \approx 0.67$, so dass der Kontraktions-Koeffizient kleiner wird, je mehr wir uns der Lösung annähern.

n	x_n
1	0.8414709848078965
10	0.48132935526234627
100	0.1696653247073242
1000	0.05462012602579727
10000	0.017314486231827124
100000	0.005476997236720512
1000000	0.0017320423900648602
10000000	5.477222534834344E-4
100000000	1.7320506994644328E-4

Tabelle 6.1: Werte der durch $x_{n+1} = \sin(x_n)$ definierten Folge $\big(x_n\big)_{n\in\mathbb{N}}$.

Aufgabe 44: Lösen Sie für $y = 10^6$ und $y = 10^{-6}$ die Gleichung $x \cdot \exp(x) = y$ durch eine einfache Fixpunkt-Iteration. Berechnen Sie die Lösung x jeweils auf eine Genauigkeit von 10^{-7}.

Hinweis: Diese Aufgabe ist in erster Linie als Programmier-Aufgabe gedacht. ◇

Bei der Definition einer kontrahierenden Abbildung haben Sie sich vielleicht gewundert, warum es nicht reicht zu fordern, dass

$$|f(x) - f(y)| < |x - y|$$

gilt. Eine Funktion, die nur diese schwächere Bedingung erfüllt, wollen wir als eine *schwach kontrahierende Abbildung* bezeichnen. In der Tat kann gezeigt werden, dass auch für eine schwach kontrahierende Abbildung f die Folge $x_{n+1} := f(x_n)$ für beliebige Startwerte gegen eine Lösung der Fixpunkt-Gleichung $x = f(x)$ konvergiert. Allerdings ist die Konvergenz unter Umständen nur sehr langsam. Als Beispiel betrachten wir die Funktion $x \mapsto \sin(x)$. Offenbar hat die Gleichung $x = \sin(x)$ bei $\bar{x} = 0$ einen Fixpunkt. Setzen wir $x_1 := 1$ und berechnen wir die Folge x_n numerisch, so erhalten wir die in Tabelle 6.1 gezeigten Ergebnisse. Selbst nach 10^8 Iterationen haben wir die Lösung der Gleichung $x = \sin(x)$ erst auf drei Stellen genau berechnet. Der Grund dafür ist, dass die Ableitung der Sinus-Funktion der Cosinus ist, und dieser hat an der Stelle $x = 0$ den Betrag 1. Es ist zwar so, dass die Funktion $x \mapsto \sin(x)$ schwach kontrahierend ist, aber es gibt kein $q < 1$, so dass für alle $x, y \in \mathbb{R}$ die Ungleichung

$$|\sin(x) - \sin(y)| \leq q \cdot |x - y|$$

erfüllt ist. Das ist der Grund für die äußerst langsame Konvergenz der Folge $\big(x_n\big)_{n\in\mathbb{N}}$.

6.4.1 Beschleunigung der Fixpunkt-Iteration

Auch mit 36 Iterationen ist die Fixpunkt-Iteration bei der Lösung der Gleichung $x = \cos(x)$ dem Bisektions-Verfahren unterlegen. Wir stellen uns daher die Frage, ob wir die Konvergenz der Fixpunkt-Iteration beschleunigen können. Falls die kontrahierende Abbildung f differenzierbar ist, so ist der Kontraktions-Koeffizient durch den Betrag der Ableitung von f gegeben. Wir überlegen uns daher, wie wir die Abbildung so verändern können, dass sich einerseits der Fixpunkt nicht ändert, aber andererseits der Betrag der Ableitung kleiner wird. Dazu formen wir die Fixpunkt-Gleichung $x = f(x)$ wie folgt um:

$$\begin{array}{rrcll} & x & = & f(x) & \mid + \alpha \cdot x \\ \Leftrightarrow & (1+\alpha) \cdot x & = & f(x) + \alpha \cdot x & \mid \cdot \dfrac{1}{1+\alpha} \\ \Leftrightarrow & x & = & \dfrac{f(x) + \alpha \cdot x}{1+\alpha} & \end{array}$$

Damit haben wir also die Funktion

$$g(x) = \frac{f(x) + \alpha \cdot x}{1 + \alpha}$$

gefunden, welche dieselben Fixpunkte hat wie die ursprüngliche Funktion f. Für die Ableitung gilt

$$g'(x) = \frac{f'(x) + \alpha}{1 + \alpha}$$

Der Betrag diese Ableitung wird für den Fixpunkt $\bar{x}$ dann am kleinsten, wenn wir

$$\alpha = -f'(\bar{x})$$

wählen. Wählen wir beispielsweise $\alpha = 0.7$, so finden wir die Lösung der Fixpunkt-Gleichung $x = \cos(x)$ mit dem Programm in Abbildung 6.1 die in Tabelle 6.2 gezeigten Werte. Wir sehen, dass bereits nach fünf Iterations-Schritten eine Genauigkeit von mehr als 10^{-6} erreicht ist.

```
1  x     := 0;
2  alpha := read("input alpha");
3  for (i in [1 .. 12]) {
4      x := 1 / (1 + alpha) * (cos(x) + alpha * x);
5      print("$i$: $x$");
6  }
```

Abbildung 6.1: Berechnung der durch $x_{n+1} = \dfrac{\cos(x) + \alpha \cdot x_n}{1 + \alpha}$ definierten Folge.

n	x_n	n	x_n
1	0.588235294117647	7	0.739085133207671
2	0.731579943641669	8	0.739085133215044
3	0.738956362842702	9	0.739085133215159
4	0.739083130793540	10	0.739085133215161
5	0.739085102132028	11	0.739085133215161
6	0.739085132732678	12	0.739085133215161

Tabelle 6.2: Die ersten 12 Glieder der durch $x_{n+1} = \frac{\cos(x)+\alpha\cdot x_n}{1+\alpha}$ definierten Folge für $\alpha = 0.7$.

Das oben skizzierte Verfahren der Konvergenz-Beschleunigung hat einen Schönheitsfehler: Wir haben α so gewählt, dass $f'(\bar{x}) + \alpha$ möglichst klein wird. Das Problem dabei ist, dass wir $\bar{x}$ gar nicht kennen und daher auch $f'(\bar{x})$ unbekannt ist. Eine mögliche Lösung besteht darin, dass wir für α in jedem Schritt $-f'(x_n)$ einsetzen. Das führt auf folgende Definition für die Folge $\big(x_n\big)_{n\in\mathbb{N}}$:

$$x_{n+1} = \frac{f(x_n) - f'(x_n) \cdot x_n}{1 - f'(x_n)} \tag{6.35}$$

Abbildung 6.2 zeigt ein Programm zur Umsetzung dieser Idee. Hier haben wir die ersten 7 Werte mit Gleichung (6.35) berechnet, die für den Fall $f(x) = \cos(x)$ die Form

$$x_{n+1} := \frac{cos(x) + \sin(x) \cdot x}{1 + \sin(x)}$$

annimmt. Da diese Gleichung aber komplexer als die ursprüngliche Gleichung $x_{n+1} = \cos(x_n)$ ist, müssen wir damit rechnen, dass die Rundungfehler höher sind als bei der Iteration. Zur Eliminierung dieser Rundungfehler führen wir daher in den Zeilen 8 – 11 noch eine Nach-Iteration mit

der Gleichung $x_{n+1} = \cos(x_n)$ durch. Tabelle 6.3 zeigt die von diesem Programm berechneten Werte. Diesmal ist die Genauigkeit von 10^{-6} bereits nach 4 Schritten erreicht, de facto ist der im vierten Schritt berechnete Wert sogar auf 9 Stellen hinter dem Komma genau. Weiter sehen Sie, dass durch die Nach-Iteration die letzte angezeigte Stelle verändert wird.

```
1  x := 0;
2  n := 7;
3  for (i in [1 .. 7]) {
4      alpha := -sin(x);
5      x := (cos(x) - alpha * x) / (1 - alpha);
6      print("$i$: $x$");
7  }
8  for (i in [n+1 .. n+4]) {
9      x := cos(x);
10     print("$i$: $x$");
11 }
```

Abbildung 6.2: Berechnung der durch $x_{n+1} = \frac{\cos(x_n)+\sin(x_n)\cdot x}{1+\sin(x_n)}$ definierten Folge.

n	x_n	n	x_n
1	1.0	7	0.7390851332151608
2	0.7503638678402440	8	0.7390851332151606
3	0.7391128909113617	9	0.7390851332151607
4	0.7390851333852842	10	0.7390851332151607
5	0.7390851332151604	11	0.7390851332151607
6	0.7390851332151608		

Tabelle 6.3: Ausgabe des in Abbildung 6.2 gezeigten Programms.

6.4.2 Das Newton'sche Verfahren zur Berechnung von Nullstellen

Oft besteht die Aufgabe darin, eine Nullstelle einer Funktion $g(x)$ zu finden. Dieses Problem ist dazu äquivalent, eine Fixpunkt-Gleichung zu lösen, denn es gilt

$$g(x) = 0 \;\Leftrightarrow\; x + g(x) = x.$$

Eine Nullstelle der Funktion $g(x)$ ist also ein Fixpunkt der Funktion $f(x) = x + g(x)$. Setzen wir in Gleichung (6.35) für $f(x)$ die Funktion $x + g(x)$ ein, so erhalten wir wegen

$$\frac{d}{dx}\big(x + g(x)\big) = 1 + g'(x)$$

die Gleichung

$$\begin{aligned} x_{n+1} &= \frac{x_n + g(x_n) - \big(1 + g'(x_n)\big) \cdot x_n}{1 - \big(1 + g'(x_n)\big)} \\ &= \frac{x_n + g(x_n) - x_n - g'(x_n) \cdot x_n}{-g'(x_n)} \\ &= \frac{-g(x_n) + g'(x_n) \cdot x_n}{g'(x_n)} \\ &= x_n - \frac{g(x_n)}{g'(x_n)} \end{aligned}$$

Wir haben also die Iterations-Vorschrift

$$\boxed{x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}} \tag{6.36}$$

gefunden. Dieses Verfahren wird als *Newton'sches Verfahren* bezeichnet. Die Gleichung (6.36) lässt sich geometrisch interpretieren: Legen wir im Punkt $\langle x_n, g(x_n)\rangle$ eine Tangente an die Funktion g, so schneidet diese Tangente die x-Achse im Punkt

$$x_n - \frac{g(x_n)}{g'(x_n)}.$$

Der neue Wert x_{n+1} ist also die Näherung, die wir erhalten, wenn wir die Funktion g durch die Tangente im Punkt $\langle x_n, g(x_n)\rangle$ ersetzen and dann die Nullstelle berechnen, die diese Tangente hat.

Aufgabe 45: Beweisen Sie diese Behauptung. ◇

Beispiel: Als Anwendung des Newton'schen Verfahrens betrachten wir die Berechnung der k-ten Wurzel ($k \in \mathbb{N}$ mit $k \geq 2$) einer gegebenen Zahl $a > 0$. Wegen

$$x = \sqrt[k]{a} \quad \Leftrightarrow \quad x^k - a = 0$$

setzen wir $g(x) := x^k - a$ und bestimmen die Nullstellen der Funktion $g(x)$ mit dem Newton'schen Verfahren. Für die Ableitung der Funktion $g(x)$ finden wir

$$g'(x) = k \cdot x^{k-1}.$$

Damit lautet die Iterations-Vorschrift

$$x_{n+1} = x_n - \frac{x_n^k - a}{k \cdot x_n^{k-1}} = \frac{1}{k}\left((k-1) \cdot x_n + \frac{a}{x_n^{k-1}}\right).$$

Berechnen wir mit diesem Verfahren die dritte Wurzel aus 2, so lautet die Iterations-Vorschrift

$$x_{n+1} = \frac{1}{3} \cdot \left(2 \cdot x_n + \frac{2}{x_n^2}\right)$$

Lassen wir die Folge mit 1 starten so finden wir die Werte

1.0, 1.333333333, 1.263888889, 1.259933494, 1.259921050

und der letzte Wert stimmt im Rahmen der Rechengenauigkeit mit $\sqrt[3]{2}$ überein.

Das Newton'sche Verfahren ist nicht robust, denn im Allgemeinen konvergiert das Verfahren nicht. Als Beispiel betrachten wir die Funktion

$$g(x) = x^3 - 2 \cdot x + 2.$$

Es gilt $g(-2) = -8 - 2 \cdot (-2) + 2 = -2 < 0$ und $g(2) = 8 - 2 \cdot 2 + 2 = 6 > 0$. Nach dem Zwischenwert-Satz über stetige Funktionen muss die Funktion g daher in dem Intervall $[-2, 2]$ eine Nullstelle haben. Das Newton'sche Verfahren ergibt die Formel

$$x_{n+1} = x_n - \frac{x_n^3 - 2 \cdot x_n + 2}{3 \cdot x_n^2 - 2}.$$

Wählen wir als Start-Wert $x_0 := 0$, so erhalten wir

$$x_1 = 0 - \frac{2}{-2} = 1.$$

Für x_2 finden wir

$$x_2 = 1 - \frac{1 - 2 + 2}{3 - 2} = 1 - 1 = 0 = x_0.$$

Wir sind also wieder bei unserem Start-Wert angekommen! Damit gilt allgemein

$$x_n = \begin{cases} 0 & \text{falls } n \,\%\, 2 = 0, \\ 1 & \text{falls } n \,\%\, 2 = 1. \end{cases}$$

Folglich konvergiert das Verfahren in diesem Fall nicht. Wählen wir den Startwert $x_0 = -0.5$, so konvergiert das Verfahren gegen die Lösung $-2.23070764576\cdots$, die außerhalb des Intervalls $[-2, 2]$ liegt. Wählen wir als Startwert $x_0 = -1.5$, so konvergiert das Verfahren gegen die Lösung $-1.769292354238631\cdots$. Diese Beispiele zeigen, dass das Newton'sche Verfahren ohne weitere Einschränkungen nicht zuverlässig ist.

6.4.3 Analyse des Newton'schen Verfahrens

Wir wollen in diesem Abschnitt herausfinden, unter welchen Umständen das Newton'sche Verfahren konvergiert und wollen außerdem die Geschwindigkeit der Konvergenz des Verfahrens untersuchen. Als erstes benötigen wir einen Hilfssatz über konvexe Funktionen.

Lemma 77 Die Funktion $f : [a, b] \to \mathbb{R}$ sei zweimal differenzierbar und konvex. Ist $x_0 \in [a, b]$ und definieren wir die lineare Funktion $g : [a, b] \to \mathbb{R}$ als

$$g(x) := f(x_0) + f'(x_0) \cdot (x - x_0),$$

so ist die Funktion g die Tangente an f im Punkt $\langle x_0, f(x_0)\rangle$ und es gilt

$$g(x) \leq f(x) \quad \text{für alle } x \in [a, b].$$

Anschaulich bedeutet dies, dass die Tangente immer unterhalb einer konvexen Funktion liegt.

Beweis: Zunächst gilt offenbar

$$g(x_0) = f(x_0) + f'(x_0) \cdot (x_0 - x_0) = f(x_0),$$

so dass die Werte der Funktionen f und g im Punkt x_0 übereinstimmen. Für die Ableitung von $g(x)$ finden wir

$$g'(x) = f'(x_0),$$

so dass die Gerade g dieselbe Steigung hat wie die Funktion f an der Stelle x_0. Damit ist g aber die Tangente an die Funktion f an der Stelle x_0. Zum Beweis der Ungleichung $g(x) \leq f(x)$ definieren wir die Funktion $h : [a - x_0, b - x_0] \to \mathbb{R}$ als

$$h(x) := f(x + x_0).$$

Nach Gleichung (6.17) gilt für die Funktion h die Gleichung

$$h(x) = h(0) + h'(0) \cdot x + \tfrac{1}{2} \cdot h^{(2)}(\chi) \cdot x^2,$$

wobei χ ein Element des Intervalls $[a - x_0, b - x_0]$ ist, über das wir sonst nichts wissen. Aufgrund der Gleichungen

$$f(x) = h(x - x_0), \quad h(0) = f(x_0), \quad h'(0) = f'(x_0) \quad \text{und} \quad h^{(2)}(\chi) = f^{(2)}(\chi + x_0),$$

folgt daraus für die Funktion f die Gleichung

$$f(x) = f(x_0) + f'(x_0) \cdot (x - x_0) + \tfrac{1}{2} \cdot f^{(2)}(\chi + x_0) \cdot (x - x_0)^2.$$

Definieren wir $\varphi := \chi + x_0$, so gilt $\varphi \in [a, b]$ und wir haben

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0) \cdot (x - x_0) + \frac{1}{2} \cdot f^{(2)}(\varphi) \cdot (x - x_0)^2 \\ &= g(x) + \frac{1}{2} \cdot f^{(2)}(\varphi) \cdot (x - x_0)^2 \\ &\geq g(x), \end{aligned}$$

denn da wir angenommen hatten, dass die Funktion f zweimal differenzierbar und konvex ist, gilt

$$f^{(2)}(\varphi) \geq 0$$

und das Quadrat $(x - x_0)^2$ ist sicher immer größer oder gleich 0. □

Satz 78 Die Funktion $f : [a, b] \to \mathbb{R}$ sei zweimal differenzierbar und konvex, es gelte $f(a) < 0$, $f(b) > 0$ und

$$\forall x \in [a, b] : f'(x) > 0.$$

Falls der Startwert $x_0 \in [a, b]$ so gewählt wird, dass $f(x_0) > 0$ ist, dann ist die durch das Newton'sche Verfahren definierte Folge $\big(x_n\big)_{n\in\mathbb{N}}$ monoton fallend und beschränkt und damit konvergent. Für den Grenzwert $\bar{x} := \lim\limits_{n\to\infty} x_n$ gilt $f(\bar{x}) = 0$.

Beweis: Nach dem Zwischenwert-Satz hat die Funktion eine Nullstelle ξ in dem Intervall $[a, b]$. Da $f'(x)$ für alle $x \in [a, b]$ echt größer als Null ist, kann f keine zweite Nullstelle haben, denn sonst hätten wir einen Widerspruch zum Satz von Rolle. Also gilt

$$\forall x \in [a, \xi) : f(x) < 0 \quad \text{und} \quad \forall x \in (\xi, b] : f(x) > 0.$$

Wir zeigen zunächst durch Induktion über n, dass

$$f(x_n) \geq 0 \quad \text{für alle } n \in \mathbb{N}_0 \text{ gilt.}$$

I.A. $n = 0$.

Die Ungleichung $f(x_0) > 0$ gilt nach Voraussetzung. √

I.S. $n \mapsto n + 1$.

Nach der Definition ist x_{n+1} die Nullstelle der Tangente g an die Funktion f im Punkt x_n, es gilt also $g(x_{n+1}) = 0$. Die Tangente g hat nach dem letzten Lemma die Form

$$g(x) = f(x_n) + f'(x_n) \cdot (x - x_n)$$

und ebenfalls nach dem letzten Lemma gilt $g(x) \leq f(x)$. Da nun $g(x_{n+1}) = 0$ ist, folgt aus der Ungleichung $g(x) \leq f(x)$ durch Einsetzen von x_{n+1} die Ungleichung

$$0 \leq f(x_{n+1}).$$

Damit ist die Induktion abgeschlossen. √

Nach Definition von x_{n+1} gilt

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

Einerseits haben wir gerade gezeigt, dass $f(x_n) \geq 0$ ist, andererseits ist die Voraussetzung, dass für alle $x \in [a, b]$ die Ungleichung $f'(x) > 0$ gilt. Daraus folgt aber

$$\frac{f(x_n)}{f'(x_n)} \geq 0$$

und damit gilt

$$x_{n+1} \leq x_n.$$

Dies zeigt, dass die Folge $\big(x_n\big)_{n\in\mathbb{N}}$ monoton fallend ist. Da wir andererseits wissen, dass $f(x_n) \geq 0$ ist, muss $x_n \geq \xi$ gelten. Also ist die Folge x_n durch ξ nach unten beschränkt. Als monoton fallend und beschränkte Folge hat $\big(x_n\big)_{n\in\mathbb{N}}$ damit einen Grenzwert $\bar{x}$. Für diesen Grenzwert gilt

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

Aus der Gleichung

$$\bar{x} = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})} \quad \text{folgt sofort} \quad 0 = -\frac{f(\bar{x})}{f'(\bar{x})}$$

und daraus folgt durch Multiplikation mit $-f'(\bar{x})$ die gesuchte Gleichung $f(\bar{x}) = 0$. □

Analyse der Konvergenz-Geschwindigkeit

Benutzen wir das Newton'sche Verfahren zur Berechnung von $\sqrt{2}$, so erhalten wir die folgenden Ergebnisse:

```
x0 = 2
x1 = 1.500000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
x2 = 1.416666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666
x3 = 1.414215686274509803921568627450980392156862745098039215686274509803921568627450980392156862745098039215686
x4 = 1.414213562374689910626295578890134910116559622115744044584905019200054371835389268358990040
x5 = 1.414213562373095048801689623502530243614981925776197428498289498623195824228923621784941830
x6 = 1.414213562373095048801688724209698078569671875377234001561013133113265255630339978531787160
x7 = 1.414213562373095048801688724209698078569671875376948073176679737990732478462107038850387530
x8 = 1.414213562373095048801688724209698078569671875376948073176679737990732478462107038850387530
x9 = 1.414213562373095048801688724209698078569671875376948073176679737990732478462107038850387530
```

Wir sehen, dass x_2 auf 2 Stellen hinter dem Komma mit $\sqrt{2}$ übereinstimmt, bei x_3 sind bereits 5 Stellen richtig, x_4 hat eine Genauigkeit von 11 Stellen, x_5 stimmt auf 24 Stellen mit $\sqrt{2}$ überein, bei x_6 sind es 48 Stellen und x_7 hat bereits eine Genauigkeit von 100 Stellen. Wir beobachten, dass sich die Zahl der korrekten Stellen mit jeder Operation etwa verdoppelt. Diese Phänomen wollen wir nun genauer untersuchen. Dazu entwicken wir die Funktion $f(x)$ an der Stelle x_n in einer Taylor-Reihe, die wir nach dem linearen Glied abbrechen. Wir erhalten

$$f(x) = f(x_n) + f'(x_n) \cdot (x - x_n) + \frac{1}{2} \cdot f^{(2)}(\varphi) \cdot (x - x_n)^2$$

Hier setzen wir für x den Wert $\bar{x}$, also die Nullstelle von f ein und erhalten

$$0 = f(x_n) + f'(x_n) \cdot (\bar{x} - x_n) + \frac{1}{2} \cdot f^{(2)}(\varphi) \cdot (\bar{x} - x_n)^2$$

Wir subtrahieren $f(x_n)$ und teilen anschließend durch $f'(x_n)$. Das liefert die Gleichung

$$-\frac{f(x_n)}{f'(x_n)} = \bar{x} - x_n + \frac{1}{2} \cdot \frac{f^{(2)}(\varphi)}{f'(x_n)} \cdot (\bar{x} - x_n)^2$$

Jetzt addieren wir x_n und subtrahieren $\bar{x}$. Das liefert

$$x_n - \frac{f(x_n)}{f'(x_n)} - \bar{x} = \frac{1}{2} \cdot \frac{f^{(2)}(\varphi)}{f'(x_n)} \cdot (\bar{x} - x_n)^2.$$

An dieser Stelle bemerken wir, dass $x_n - \dfrac{f(x_n)}{f'(x_n)} = x_{n+1}$ ist und haben damit

$$x_{n+1} - \bar{x} = \frac{1}{2} \cdot \frac{f^{(2)}(\varphi)}{f'(x_n)} \cdot (\bar{x} - x_n)^2.$$

Wir definieren nun $\varepsilon_n := |x_n - \bar{x}|$. Der Wert ε_n gibt also den Betrag des Approximations-Fehler an, den wir nach der n-ten Iteration des Newton'schen Verfahrens haben. Damit schreibt sich die letzte Gleichung als

$$\varepsilon_{n+1} = \frac{1}{2} \cdot \left| \frac{f^{(2)}(\varphi)}{f'(x_n)} \right| \cdot \varepsilon_n^2.$$

In den Fällen, in denen wir den Ausdruck $\dfrac{1}{2} \cdot \left| \dfrac{f^{(2)}(\varphi)}{f'(x_n)} \right|$ durch eine Konstante K abschätzen können, haben wir dann

$$\varepsilon_{n+1} \leq K \cdot \varepsilon_n^2.$$

Die Zahl der korrekten Stellen nach der n-ten Iteration ist in etwa durch

$$\lambda_n \approx -\log_{10}(\varepsilon_n)$$

gegeben. Logarithmieren wir die obige Gleichung zur Basis 10 und multiplizieren mit -1, so erhalten wir

$$\lambda_{n+1} \geq 2 \cdot \lambda_n - \log_{10}(K).$$

Zur Konkretisierung unserer Überlegungen betrachten wir wieder die Funktion $f(x) = x^2 - 2$. Hier gilt

$$f'(x) = 2 \cdot x \geq 2 \quad \text{falls } x \geq 1 \text{ ist}$$

und weiter gilt $f^{(2)}(x) = 2$. Damit gilt

$$\frac{1}{2} \cdot \left| \frac{f^{(2)}(\varphi)}{f'(x_n)} \right| \leq \frac{1}{2}$$

und wegen $\log_{10}\left(\frac{1}{2}\right) \approx -0.3$ haben wir die Abschätzung

$$\lambda_{n+1} = 2 \cdot \lambda_n + 0.3$$

gefunden, die in der Tat zeigt, dass sich die Anzahl der korrekten Stellen bei jedem Schritt mehr als verdoppelt.

Aufgabe 46: Analysieren Sie das Newton'sche Verfahren zur Berechnung der dritten Wurzel aus 2 und berechnen Sie, wieviele Iterationen höchstens notwendig sind um den Wert von $\sqrt[3]{2}$ auf eine Genauigkeit von 10^{-101} zu berechnen. ◇

6.5 Iterative Lösung linearer Gleichungs-Systeme*

Es sei eine $n \times n$ Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ und ein Vektor $\vec{b} \in \mathbb{R}^n$ gegeben. Eine Möglichkeit, das Gleichungs-System

$$\mathbf{A}\,\vec{x} = \vec{b}$$

zu lösen, haben Sie im ersten Semester kennegelernt: Es ist das Gauß'sche Eliminations-Verfahren (Carl Friedrich Gauß, 1777 – 1855). Es gibt allerdings Situationen, in denen dieses Verfahren zu aufwendig ist. Dies ist beispielsweise dann der Fall, wenn einerseits n groß ist und wenn andererseits die meisten Komponenten der Matrix A den Wert 0 haben. Solche Matrizen heißen *dünn besetzte* Matrizen. Diese Art von Matrizen tritt bei der numerischen Lösung von partiellen Differential-Gleichungen auf. Das Gauß'sche Eliminations-Verfahren besitzt eine Komplexität von

$O(n^3)$ und ist damit für große dünn besetzte Matrizen ungeeignet. Hier ist der Rechenaufwand bei iterative Verfahren geringer. Ein weiteres Problem ist, dass die Rundungsfehler beim Gauß'schen Eliminations-Verfahren sehr groß werden können. Demgegenüber sind iterative Verfahren selbstkorrigierend.

Ist $\mathbf{A}$ gegeben, so lautet die Gleichung $\mathbf{A}\,\vec{x} = \vec{b}$ in Komponenten-Schreibweise

$$\sum_{j=1}^{n} a_{ij} \cdot x_j = b_i \quad \text{für alle } i = 1, \cdots, n.$$

Die Idee besteht darin, diese Gleichung in eine Fixpunkt-Gleichung zu transformieren. Dazu formen wir die Gleichung wie folgt um:

$$\begin{aligned}
& \sum_{j=1}^{n} a_{ij} \cdot x_j &=& \; b_i \\
\Leftrightarrow \quad & a_{ii} \cdot x_i + \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot x_j &=& \; b_i \\
\Leftrightarrow \quad & a_{ii} \cdot x_i &=& \; b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot x_j \\
\Leftrightarrow \quad & x_i &=& \; \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot x_j \right).
\end{aligned}$$

Diese Umformung liefert uns die Iterations-Vorschrift

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot x_j^{(n)} \right)$$

mit der wir versuchen können, eine Lösung der Fixpunkt-Gleichung zu finden. Das Verfahren, das wir auf diese Weise erhalten, wird als *Gesamtschritt-Verfahren* oder auch als *Jacobi-Verfahren* (Carl Gustav Jacob Jacobi, 1804 – 1851) bezeichnet. Die Frage lautet jetzt, wann die durch dieses Verfahren definierte Iteration konvergiert. Um eine positive Antwort auf diese Frage geben zu können, benötigen wir die folgenden Definitionen.

Definition 79 (Starkes Zeilen-Summen-Kriterium) Eine $n \times n$ Matrix $A \in \mathbb{R}^{n\times n}$ erfüllt das *starke Zeilen-Summen-Kriterium* falls es eine Zahl $q < 1$ gibt, so dass gilt

$$\sum_{\substack{j=1\\ j\neq i}}^{n} \big|a_{ij}\big| \leq q \cdot \big|a_{ii}\big| \quad \text{für alle } i = 1, \cdots, n.$$

Falls $a_{ii} \neq 0$ ist, ist die in der obigen Definition gegebene Ungleichung äquivalent zu der Ungleichung

$$\frac{1}{\big|a_{ii}\big|} \cdot \sum_{\substack{j=1\\ j\neq i}}^{n} \big|a_{ij}\big| \leq q.$$

Als nächstes führen wir ein Konzept ein, das den Begriff des Betrags auf Vektoren aus dem $\mathbb{R}^n$ verallgemeinert.

Definition 80 (Maximums-Norm) Ist $\vec{x} \in \mathbb{R}^n$, so definieren wir die Maximums-Norm von $\vec{x}$ als

$$\big\|\vec{x}\big\|_\infty := \max\big\{\, |x_i| \mid i \in \{1, \cdots, n\} \,\big\}.$$

Mit der so definierten Norm können wir so rechnen, wie wir das von Beträgen bei reellen Zahlen gewöhnt sind, insbesondere gilt auch die *Dreiecks-Ungleichung*, wir haben also für beliebige Vektoren $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$\left\|\vec{x}+\vec{y}\right\|_\infty \leq \left\|\vec{x}\right\|_\infty + \left\|\vec{y}\right\|_\infty.$$

Außerdem haben wir für reelle Zahlen $\alpha \in \mathbb{R}$ und Vektoren $\vec{x} \in \mathbb{R}^n$ die Gleichung

$$\left\|\alpha\,\vec{x}\right\|_\infty = |\alpha| \cdot \left\|\vec{x}\right\|_\infty$$

Hierbei bezeichnet $\alpha\,\vec{x}$ die komponentenweise Multiplikation des Vektors $\vec{x}$ mit der Zahl α. Schließlich gilt für alle $\vec{x} \in \mathbb{R}^n$ die Ungleichung $0 \leq \left\|\vec{x}\right\|_\infty$ wobei Gleichheit nur im Fall $\vec{x} = \vec{0}$ auftritt:

$$\left\|\vec{x}\right\|_\infty = 0 \;\Rightarrow\; \vec{x} = \vec{0} \quad \text{für alle } \vec{x} \in \mathbb{R}.$$

Der folgende Satz beantwortet nun die oben gestellte Frage nach der Konvergenz des Gesamtschritt-Verfahrens in einem für Anwendungen wichtigen Spezial-Fall.

Satz 81 Wenn die Matrix $A \in \mathbb{R}^{n\times n}$ das starke Zeilen-Summen-Kriterium erfüllt, dann konvergiert das Gesamtschritt-Verfahren für jeden Start-Vektor. Bezeichnen wir die Vektoren der Iteration mit $\vec{x}^{(n)}$ und definieren wir

$$\vec{x}^{(\infty)} := \lim_{n\to\infty} \vec{x}^{(n)}$$

und ist weiter q die Zahl aus dem starken Zeilen-Summen-Kriterium, dann gilt außerdem die Abschätzung

$$\left\|\vec{x}^{(\infty)} - \vec{x}^{(n)}\right\|_\infty \leq \tfrac{q^n}{1-q} \cdot \left\|\vec{x}^{(0)} - \vec{x}^{(1)}\right\|_\infty.$$

Beweis: Wir definieren eine Funktion $f : \mathbb{R}^n \to \mathbb{R}^n$ komponentenweise:

$$f_i(\vec{x}) = \tfrac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot x_j \right).$$

Die i-te Komponente der Funktion f berechnet also gerade die i-te Komponente des Vektors $\vec{x}^{(n+1)}$. Wir zeigen nun, dass für beliebige Vektoren $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$\left\|f(\vec{x}) - f(\vec{y})\right\|_\infty \leq q \cdot \left\|\vec{x} - \vec{y}\right\|_\infty$$

gilt, denn dann ist f eine kontrahierende Abbildung, und die Behauptung des Satzes folgt aus dem Banach'schen Fixpunkt-Satz, den wir zwar nur für Funktionen $f : \mathbb{R} \to \mathbb{R}$ bewiesen haben, der aber auch für vektorwertige Funktionen im $\mathbb{R}^n$ gilt, wenn wir den Betrag durch die Maximums-Norm ersetzen. Wir schätzen die Differenz $\left|f_i(\vec{x}) - f_i(\vec{y})\right|$ wie folgt ab:

$$\begin{aligned}
\left|f_i(\vec{x}) - f_i(\vec{y})\right| &= \left| \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot x_j \right) - \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot y_j \right) \right| \\
&= \left| \frac{1}{a_{ii}} \cdot \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} \cdot (y_j - x_j) \right| \quad \leq \quad \frac{1}{|a_{ii}|} \cdot \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \cdot |x_j - y_j| \\
&\leq \frac{1}{|a_{ii}|} \cdot \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \cdot \left\|\vec{x} - \vec{y}\right\|_\infty \quad \leq \quad \left(\frac{1}{|a_{ii}|} \cdot \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \right) \cdot \left\|\vec{x} - \vec{y}\right\|_\infty \\
&\leq q \cdot \left\|\vec{x} - \vec{y}\right\|_\infty
\end{aligned}$$

Da diese Ungleichung für alle $i = 1, \cdots, n$ gilt, haben wir insgesamt die Abschätzung

$$\left\|f(\vec{x}) - f(\vec{y})\right\|_\infty \leq q \cdot \left\|\vec{x} - \vec{y}\right\|_\infty$$

gezeigt und die Behauptung folgt nun aus dem Banach'schen Fixpunkt-Satz. □

Abbildung 6.3 auf Seite 116 zeigt eine einfache Implementierung des Jacobi-Verfahrens. Die Funktion `jakobi`() bekommt drei Argumente:

1. a ist eine $n \times n$ Matrix, die durch eine Liste dargestellt wird, deren Elemente selbst wie-

der Listen der Länge n sind. Auf das Matrix-Element a_{ij} wird in SetlX dann durch den Ausdruck $a[i][j]$ zugegriffen.

2. b ist eine n-dimensionaler Vektor, der durch eine Liste der Länge n dargestellt wird.

3. k ist die Anzahl der durchzuführenden Iterationen.

```
1  jacobi := procedure(a, b, k) {
2      n := #b;
3      assert(#a    == n, "wrong number of equations");
4      assert(#a[1] == n, "wrong number of variables");
5      x := xNew := [ 0 : i in [ 1 .. n ] ];
6      for (l in [1 .. k]) {
7          for (i in [1 .. n]) {
8              xNew[i] := b[i];
9              for (j in [ 1 .. n ]) {
10                 if (i != j) {
11                     xNew[i] -= a[i][j] * x[j];
12                 }
13             }
14             xNew[i] /= a[i][i];
15         }
16         x := xNew;
17         print("$l$: $x$");
18     }
19     return x;
20 };
21
22 demo := procedure() {
23     a := [ [ 4.0, 1.0, 0.0 ],
24            [ 1.0, 4.0, 1.0 ],
25            [ 0.0, 1.0, 4.0 ] ];
26     b := [ 5.0, 6.0, 5.0 ];
27     k := 35;
28     x := jacobi(a, b, k);
29     print("x = $x$");
30 };
```

Abbildung 6.3: Implementierung des Jacobi-Verfahrens.

Die Funktion `jacobi`(a, b, x) versucht, mit Hilfe des Jacobi-Verfahrens das lineare Gleichungs-System

$$\mathbf{a}\,\vec{x} = \vec{b}$$

zu lösen und arbeitet im Detail wie folgt:

1. In Zeile 5 wid der Start-Vektor $\vec{x}^{(0)}$ als Null-Vektor definiert. Die Variable `xNew` speichert den Wert $\vec{x}^{(n+1)}$.

2. Die Schleife, die in Zeile 6 beginnt, führt insgesamt k Iterationen des Jacobi-Verfahrens aus.

3. Die Schleife, die in Zeile 7 beginnt, berechnet den nächsten Wert $\vec{x}^{(n+1)}$ gemäß der Formel

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j} \cdot x_j \right).$$

Der Index i läuft dabei über die Komponenten des Vektors $\vec{x}^{(n+1)}$.

Die Funktion `demo()` ruft die Funktion so auf, dass anschließend das Gleichungs-System

$$\begin{pmatrix} 4.0 & 1.0 & 0.0 \\ 1.0 & 4.0 & 1.0 \\ 0.0 & 1.0 & 4.0 \end{pmatrix} \vec{x} = \begin{pmatrix} 5.0 \\ 6.0 \\ 5.0 \end{pmatrix}$$

gelöst werden kann. Die Lösung dieses Systems ist

$$\vec{x} = \begin{pmatrix} 1.0 \\ 1.0 \\ 1.0 \end{pmatrix}.$$

Diese Lösung wird nach 35 Iterationen gefunden. Tabelle 6.4 auf Seite 118 zeigt den Verlauf der Rechnung. Die Matrix A erfüllt das starke Zeilen-Summen-Kriterium mit $q = \frac{1}{2}$. Die exakte Lösung wird nach 35 Schritten gefunden.

Das Gauß-Seidel-Verfahren

Betrachten wir die Implementierung des Gesamtschritt-Verfahrens, so liegt die folgende Optimierung auf der Hand. Wenn wir mit der Formel

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} \cdot x_j^{(n)} \right)$$

die i-te Komponente von $\vec{x}^{(n+1)}$ berechnen, dann kennen wir bereits die neuen Komponenten $x_1^{(n+1)}, \cdots, x_{i-1}^{(n+1)}$. Da diese Komponenten (hoffentlich) näher an der Lösung liegen als die Komponenten $x_1^{(n)}, \cdots, x_{i-1}^{(n)}$ des alten Vektors $\vec{x}^{(n)}$, liegt es nahe, für $j < i$ die neuen Komponenten $x_j^{(n+1)}$ an Stelle der alten Komponenten $x_j^{(n)}$ zu benutzen. Damit kommen wir zu der zunächst kompliziert aussehenden Iterations-Formel

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{j=1}^{i-1} a_{ij} \cdot x_j^{(n+1)} - \sum_{j=i+1}^{n} a_{ij} \cdot x_j^{(n)} \right).$$

Das Verfahren, das auf dieser Formel basiert, wird als *Einzelschritt-Verfahren* oder auch *Gauß-Seidel-Verfahren* (Philipp Ludwig von Seidel, 1821 – 1896) bezeichnet.

Abbildung 6.4 zeigt, wie wir die Implementierung ändern müssen, wenn wir das Gauß-Seidel-Verfahren anwenden wollen. Die Implementierung des Gauß-Seidel-Verfahrens ist einfacher als die Implementierung des Jacobi-Verfahrens, denn wir benötigen keine Hilfsvariable `xNew` mehr. Tabelle 6.5 zeigt, dass das Gauß-Seidel-Verfahren für unser Beispiel etwa doppelt so schnell konvergiert wie das Jacobi-Verfahren. Es lässt sich zeigen, dass das Gauß-Seidel-Verfahren zur Lösung immer dann konvergiert, wenn die Matrix **A** des linearen Gleichungs-Systems

$$\mathbf{A}\vec{x} = \vec{b}$$

das starke Zeilen-Summen-Kriterium erfüllt. Allerdings ist die theoretische Analyse des Gauß-Seidel-Verfahren schwieriger als die Untersuchung des Jacobi-Verfahrens, so dass wir von einer detailierteren Diskussion aus Zeitgründen absehen müssen.

n	$x_1^{(n)}$	$x_2^{(n)}$	$x_3^{(n)}$
1	1.25	1.5	1.25
2	0.875	0.875	0.875
3	1.03125	1.0625	1.03125
4	0.984375	0.984375	0.984375
5	1.00390625	1.0078125	1.00390625
6	0.998046875	0.998046875	0.998046875
7	1.00048828125	1.0009765625	1.00048828125
8	0.999755859375	0.999755859375	0.999755859375
9	1.00006103515625	1.0001220703125	1.00006103515625
10	0.999969482421875	0.999969482421875	0.999969482421875
11	1.0000076293945312	1.0000152587890625	1.0000076293945312
12	0.9999961853027344	0.9999961853027344	0.9999961853027344
13	1.0000009536743164	1.0000019073486328	1.0000009536743164
14	0.9999995231628418	0.9999995231628418	0.9999995231628418
15	1.0000001192092896	1.000000238418579	1.0000001192092896
16	0.9999999403953552	0.9999999403953552	0.9999999403953552
17	1.0000000149011612	1.0000000298023224	1.0000000149011612
18	0.9999999925494194	0.9999999925494194	0.9999999925494194
19	1.0000000018626451	1.0000000037252903	1.0000000018626451
20	0.9999999990686774	0.9999999990686774	0.9999999990686774
21	1.0000000002328306	1.0000000004656613	1.0000000002328306
22	0.9999999998835847	0.9999999998835847	0.9999999998835847
23	1.0000000000291038	1.0000000000582077	1.0000000000291038
24	0.9999999999854481	0.9999999999854481	0.9999999999854481
25	1.000000000003638	1.000000000007276	1.000000000003638
26	0.999999999998181	0.999999999998181	0.999999999998181
27	1.0000000000004547	1.0000000000009095	1.0000000000004547
28	0.9999999999997726	0.9999999999997726	0.9999999999997726
29	1.0000000000000568	1.0000000000001137	1.0000000000000568
30	0.9999999999999716	0.9999999999999716	0.9999999999999716
31	1.000000000000007	1.0000000000000142	1.000000000000007
32	0.9999999999999964	0.9999999999999964	0.9999999999999964
33	1.0000000000000009	1.0000000000000018	1.0000000000000009
34	0.9999999999999996	0.9999999999999996	0.9999999999999996
35	1.0	1.0	1.0

Tabelle 6.4: Konvergenz des Jacobi-Verfahrens.

```
gaussSeidel := procedure(a, b, k) {
    n := #b;
    assert(#a    == n, "wrong number of equations");
    assert(#a[1] == n, "wrong number of variables");
    x := [ 0 : i in [ 1 .. n ] ];
    for (l in [1 .. k]) {
        for (i in [1 .. n]) {
            x[i] := b[i];
            for (j in [ 1 .. n ]) {
                if (i != j) {
                    x[i] -= a[i][j] * x[j];
                }
            }
            x[i] /= a[i][i];
        }
        print("$l$: $x$");
    }
    return x;
};
```

Abbildung 6.4: Implementierung des Gauß-Seidel-Verfahrens.

n	$x_1^{(n)}$	$x_2^{(n)}$	$x_3^{(n)}$
1	1.25	1.1875	0.953125
2	0.953125	1.0234375	0.994140625
3	0.994140625	1.0029296875	0.999267578125
4	0.999267578125	1.0003662109375	0.999908447265625
5	0.999908447265625	1.0000457763671875	0.9999885559082031
6	0.9999885559082031	1.0000057220458984	0.9999985694885254
7	0.9999985694885254	1.0000007152557373	0.9999998211860657
8	0.9999998211860657	1.0000000894069672	0.9999999776482582
9	0.9999999776482582	1.000000011175871	0.9999999972060323
10	0.9999999972060323	1.0000000013969839	0.999999999650754
11	0.999999999650754	1.000000000174623	0.9999999999563443
12	0.9999999999563443	1.0000000000218279	0.999999999994543
13	0.999999999994543	1.0000000000027285	0.9999999999993179
14	0.9999999999993179	1.000000000000341	0.9999999999999147
15	0.9999999999999147	1.0000000000000426	0.9999999999999893
16	0.9999999999999893	1.0000000000000053	0.9999999999999987
17	0.9999999999999987	1.0000000000000009	0.9999999999999998
18	0.9999999999999998	1.0	1.0
19	1.0	1.0	1.0

Tabelle 6.5: Konvergenz des Gauß-Seidel-Verfahrens.

Kapitel 7

Integral-Rechnung

7.1 Einführung des Integral-Begriffs

In diesem Kapitel beschäftigen wir uns mit der Frage, wie die Fläche zwischen einer Kurve, die durch eine Funktion definiert ist, und der x-Achse berechnet werden kann. Um die Dinge konkret zu machen betrachten wir die Funktion $x \mapsto x^2$ und fragen, welche Fläche von dieser Kurve und der x-Achse im Intervall $[0, 1]$ eingeschlossen wird. Diese Fläche ist in Abbildung 7.1 grau dargestellt.

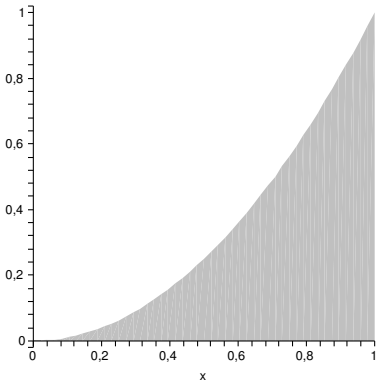

Abbildung 7.1: Die Fläche unter der Parabel $x \mapsto x^2$ im Intervall $[0, 1]$.

Wir können versuchen, diese Fläche durch sogenannte *Treppen-Funktionen* zu approximieren. Eine Treppen-Funktion ist eine Funktion, die stückweise konstant ist. Abbildung 7.2 auf Seite 121 zeigt eine Approximation der Funktion $x \mapsto x^2$ durch eine Treppen-Funktion. In dieser Abbildung haben wir das Intervall $[0, 1]$ in 10 gleich große Teilintervalle aufgeteilt. Wollen wir im allgemeinen Fall die Fläche unter eine Funktion f in einem vorgegebenen Intervall $[a, b]$ berechnen, so unterteilen wir das Intervall $[a, b]$ in n Teilintervalle der Form

$$\big[a + (i-1) \cdot h, a + i \cdot h\big] \quad \text{mit } h := \frac{b-a}{n} \text{ und } i = 1, \cdots, n.$$

Um dann die Fläche berechnen zu können, approximieren wir diese Fläche zum einen durch eine Treppen-Funktion, die oberhalb der zu integrierenden Funktion liegt und zum anderen durch eine Treppen-Funktion, die unterhalb der zu integrierenden Funktion liegt. Wir nehmen zur Vereinfachung zunächst an, dass die Funktion f monoton steigend ist. In diesem Fall definieren wir zu einer vorgegebenen Zahl n von Intervallen die obere Treppen-Funktion $f_n^{\downarrow}(x)$ wie folgt:

$$f_n^{\downarrow}(x) := f(a + i \cdot h) \quad \text{falls } x \in \big(a + (i-1) \cdot h,\, a + i \cdot h\big) \text{ und } i = 1, \cdots, n. \tag{7.1}$$

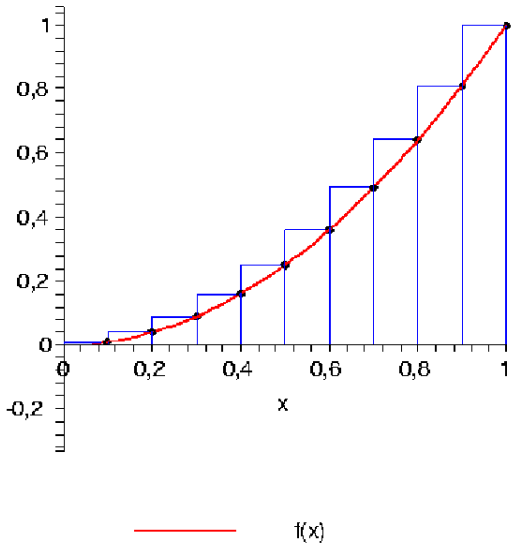

Abbildung 7.2: Die Fläche unter der Parabel $x \mapsto x^2$ im Intervall $[0, 1]$.

Da wir f als monoton steigend vorausgesetzt haben, gilt

$$\forall x \in \big(a + (i-1) \cdot h, a + i \cdot h\big) : f(x) \leq f(a + i \cdot h) = f^{\downarrow}(x),$$

die Funktion f liegt also unterhalb der Treppen-Funktion $f^{\downarrow}$. Wie die Treppen-Funktion an den Randpunkten der Intervalle $[a + (i-1) \cdot h, a + i \cdot h]$ definiert wird, ist unwichtig. Abbildung 7.2 zeigt die Funktion $f(x) = x^2$ und die zugehörige Treppen-Funktion $f^{\downarrow}_{10}(x)$. Die Fläche unter einer solchen Treppen-Funktion kann leicht berechnet werden: Schreiben wir

$$\int_a^b f(x)\,dx$$

für die Fläche unter einer Funktion f in dem Intervall $[a, b]$, so gilt für die Treppen-Funktion $f^{\downarrow}_n$ offenbar

$$\int_a^b f^{\downarrow}_n(x)\,dx = \sum_{i=1}^{n} f(a + i \cdot h) \cdot h \tag{7.2}$$

denn die Fläche eines Rechtecks berechnet sich als das Produkt von Breite und Höhe. Die Höhe des Rechtecks im i-ten Teilintervall hat den Wert $f(a + i \cdot h)$ und die Breite des Teilintervalls ist h.

Genau wie wir die Fläche unter der Funktion f von oben approximieren können, können wir diese Fläche auch von unten approximieren. Die dazu notwendige Treppen-Funktion $f^{\uparrow}_n(x)$ definieren wir analog zu Gleichung (7.1)

$$f^{\uparrow}_n(x) = f(a + (i-1) \cdot h) \quad \text{falls } x \in (a + (i-1) \cdot h, a + i \cdot h), \tag{7.3}$$

wir werten also diesmal in jedem Intervall die Funktion f am linken Eckpunkt aus, denn dort ist die Funktion f am kleinsten, denn wir haben ja vorausgesetzt, dass die Funktion f monoton

steigend ist. Für diese Treppen-Funktion berechnet sich die Fläche also nach der Formel

$$\int_a^b f_n^{\uparrow}(x)\,dx = \sum_{i=1}^{n} f\big(a + (i-1)\cdot h\big)\cdot h = \sum_{i=0}^{n-1} f(a + i\cdot h)\cdot h. \tag{7.4}$$

Für die Fläche, die zwischen der Funktion f und der x-Achse liegt, haben wir insgesamt die Abschätzung

$$\sum_{i=0}^{n-1} f(a + i\cdot h)\cdot h \leq \int_a^b f(x)\,dx \leq \sum_{i=1}^{n} f(a + i\cdot h)\cdot h$$

gefunden. Die linke Summe bezeichnen wir als *Unter-Summe*, die Summe auf der rechten Seite der Ungleichungskette nennen wir *Ober-Summe*. Wir können hoffen, dass für wachsende Werte von n die Werte von Ober-Summe und Unter-Summe gegen den selben Grenzwert konvergieren. Dazu berechnen wir zunächst die Differenz dieser beiden Summen:

$$\begin{array}{ll}
 & \sum\limits_{i=1}^{n} f(a + i\cdot h)\cdot h - \sum\limits_{i=0}^{n-1} f(a + i\cdot h)\cdot h \\
= & f(a + n\cdot h)\cdot h - f(a)\cdot h \\
= & \Big(f\Big(a + n\cdot \frac{b-a}{n}\Big) - f(a)\Big)\cdot \frac{b-a}{n} \\
= & \Big(f(a + b - a) - f(a)\Big)\cdot \frac{b-a}{n} \\
= & \Big(f(b) - f(a)\Big)\cdot \frac{b-a}{n} \\
\xrightarrow[n\to\infty]{} & 0
\end{array}$$

Damit ist klar, dass bei einer monoton steigenden Funktion die Ober-Summe für eine wachsende Zahl n von Intervallen gegen den selben Wert konvergiert wie die Unter-Summe. Daher definieren wir

$$\int_a^b f(x)\,dx := \lim_{n\to\infty} \sum_{i=1}^{n} f\Big(a + i\cdot \frac{b-a}{n}\Big)\cdot \frac{b-a}{n} \tag{7.5}$$

Diesen Grenzwert nennen wir auch das *Integral* von f in dem Intervall $[a, b]$. Die so gegebene Definition ist zunächst nur für monoton wachsende Funktionen schlüssig, aber es ist offensichtlich, dass das Integral für monoton fallende Funktionen auf die selbe Weise berechnet werden kann. Ist nun eine Funktion f in dem Intervall f weder monoton fallend noch monoton steigend, so können wir versuchen, dass Intervall so in Teilintervalle aufzuspalten, dass f in jedem Teilintervall monoton fallend oder monoton steigend ist. Da wir auf jedem dieser Teilintervalle das Integral nach der Formel (7.5) berechnen können, können wir dann auch insgesamt das Integral nach dieser Formel berechnen.

Aufgabe 47: Berechnen Sie das Integral

$$\int_0^b x^2\,dx$$

nach der in (7.5) angegebenen Formel.

Hinweis: Es gilt

$$\sum_{i=1}^{n} i^2 = \frac{n}{6}\cdot(n+1)\cdot(2\cdot n+1). \qquad \diamond$$

Definition 82 (Integral) Ist $f : [a, b] \to \mathbb{R}$ eine Funktion, so dass der Grenzwert

$$\lim_{n\to\infty} \sum_{i=1}^{n} f\Big(a + i \cdot \frac{b-a}{n}\Big) \cdot \frac{b-a}{n}$$

existiert, so nennen wir f *integrierbar* und definieren das *Integral* von f in dem Intervall $[a, b]$ als

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f\Big(a + i \cdot \frac{b-a}{n}\Big) \cdot \frac{b-a}{n}. \qquad \diamond$$

Der so eingeführte Integral-Begriff hat folgende Eigenschaften:

1. Linearität: Sind f und g zwei Funktionen, so dass das Integral über f bzw. g in dem Intervall $[a, b]$ definiert ist und sind $\alpha, \beta \in \mathbb{R}$, so ist auch das Integral der Funktion

   $$x \mapsto \alpha \cdot f(x) + \beta \cdot g(x)$$

   definiert und es gilt

   $$\int_a^b \alpha \cdot f(x) + \beta \cdot g(x)\,dx = \alpha \cdot \int_a^b f(x)\,dx + \beta \cdot \int_a^b g(x)\,dx.$$

   Diese Eigenschaft folgt aus der Tatsache, dass sowohl die Summe als auch das Produkt zweier konvergenter Folgen wieder konvergent sind.

2. Monotonie: Sind f und g zwei Funktionen, so dass das Integral über f und g in dem Intervall $[a, b]$ definiert ist, so gilt:

   $$\Big(\forall x \in [a, b] : f(x) \le g(x)\Big) \;\Rightarrow\; \int_a^b f(x)\,dx \;\le\; \int_a^b g(x)\,dx.$$

   Auch diese Eigenschaft folgt aus der entsprechenden Eigenschaft konvergenter Folgen.

Satz 83 (Mittelwert-Satz der Integral-Rechnung)
Es sei $f : [a, b] \to \mathbb{R}$ eine stetige Funktion. Dann existiert ein $\xi \in [a, b]$, so dass gilt:

$$\int_a^b f(x)\,dx = f(\xi) \cdot (b - a).$$

Beweis: Da die Funktion f stetig ist, nimmt f auf dem Intervall $[a, b]$ ein Minimum und ein Maximum in den Punkten x_{min} und x_{max} an. Dann gilt

$$\forall x \in [a, b] : f\big(x_{min}\big) \le f(x) \le f\big(x_{max}\big).$$

Aufgrund der Monotonie des Integral-Operators folgt daraus sofort

$$\begin{aligned}
& \int_a^b f\big(x_{min}\big)\,dx \le \int_a^b f(x)\,dx \le \int_a^b f\big(x_{max}\big)\,dx \\
\Leftrightarrow\quad & f\big(x_{min}\big) \cdot (b-a) \le \int_a^b f(x)\,dx \le f\big(x_{max}\big) \cdot (b-a) \\
\Leftrightarrow\quad & f\big(x_{min}\big) \le \frac{1}{b-a} \cdot \int_a^b f(x)\,dx \le f\big(x_{max}\big)
\end{aligned}$$

Die obige Ungleichungskette zeigt, dass

$$\frac{1}{b-a} \cdot \int_a^b f(x)\,dx \;\in \big[f(x_{min}), f(x_{max})\big]$$

gilt. Aufgrund des Zwischenwert-Satzes für stetige Funktionen (Satz 48 auf Seite 50) nimmt die stetige Funktion f jeden Wert in dem Intervall $\big[f\big(x_{min}\big), f\big(x_{max}\big)\big]$ an. Also gibt es ein $\xi \in [a, b]$, so dass

$$f(\xi) = \frac{1}{b-a} \cdot \int_a^b f(x)\,dx$$

gilt und dass ist äquivalent zu

$$f(\xi) \cdot (b-a) = \int_a^b f(x)\,dx. \qquad \square$$

Der Mittelwert-Satz versetzt uns in die Lage, einen Zusammenhang zwischen Differential-Rechnung und Integral-Rechnung herzustellen.

Satz 84 (Ableitung von Integralen)
Die Funktion $f : [a,b] \to \mathbb{R}$ sei stetig. Definieren wir für $x \in [a,b]$ die Funktion $F : [a,b] \to \mathbb{R}$ durch

$$F(x) := \int_a^x f(t)\,dt,$$

so ist die Funktion F im Intervall $[a,b]$ differenzierbar und es gilt

$$\frac{dF}{dx}(x) = f(x).$$

Beweis: Es gilt

$$\begin{aligned}
& \frac{dF}{dx}(x) \\
= \; & \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} \\
= \; & \lim_{h\to 0} \frac{\int_a^{x+h} f(t)\,dt - \int_a^x f(t)\,dt}{h} && \text{nach Definition von } F \\
= \; & \lim_{h\to 0} \frac{1}{h} \int_x^{x+h} f(t)\,dt \\
= \; & \lim_{h\to 0} \frac{1}{h} \cdot (x+h-x) \cdot f(\xi_h) && \text{nach dem Mittelwertsatz für ein } \xi_h \in [x, x+h] \\
= \; & \lim_{h\to 0} f(\xi_h) \\
= \; & f(x) && \text{wegen } \xi_h \in [x, x+h].
\end{aligned}$$

Damit ist die Behauptung bewiesen. $\square$

Bemerkung: Der letzte Satz zeigt uns, dass der Differentiations-Operator

$$\frac{d\,\cdot}{dx} := \left(f \mapsto \frac{df}{dx} \right)$$

zu dem Integral-Operator

$$\int_a^x \cdot\, dt := \left(f \mapsto \int_a^x f(t)\,dt \right)$$

invers ist: Wenden wir auf eine Funktion zunächst den Integral-Operator an und wenden wir dann auf die resultierende Funktion den Differentiations-Operator an, so erhalten wir wieder die ursprüngliche Funktion:

$$\boxed{\frac{d}{dx} \int_a^x f(t)\,dt = f(x).}$$

Diese Aussage lässt sich im Wesentlichen auch umkehren. Diese Umkehrung ist der Hauptsatz der Differential- und Integral-Rechnung. Bevor wir diesen Satz in Angriff nehmen können, benötigen

wir noch eine Definition.

Definition 85 (Stamm-Funktion)
Ist $f : [a, b] \to \mathbb{R}$ eine Funktion, so ist die Funktion $F : [a, b] \to \mathbb{R}$ eine *Stamm-Funktion* von f, falls F differenzierbar ist und

$$\frac{dF}{dx}(x) = f(x)$$

gilt. In diesem Fall schreiben wir auch $F(x) = \int f(x)\,dx$. Der Ausdruck $\int f(x)\,dx$ wird als *unbestimmtes Integral* bezeichnet. ◇

Die Schreibweise $F(x) = \int f(x)\,dx$ ist problematisch, denn die Stamm-Funktion einer gegebenen Funktion ist nicht eindeutig. Ist $F(x)$ eine Stamm-Funktion von $f(x)$ und definieren wir die Funktion G durch $G(x) := F(x) + c$ für eine beliebige Konstante c, so ist natürlich auch $G(x)$ eine Stamm-Funktion von f, denn es gilt

$$\frac{dG}{dx}(x) = \frac{dF}{dx}(x) + \frac{dc}{dx} = f(x) + 0 = f(x).$$

Umgekehrt gilt, dass zwei verschiedene Stamm-Funktionen zu einer Funktion f sich nur um eine Konstante unterscheiden. Dies sehen wir wie folgt: Angenommen, F_1 und F_2 seien zwei Stamm-Funktionen einer Funktion f, es gelte also

$$\frac{dF_1}{dx}(x) = f(x) \quad \text{und} \quad \frac{dF_2}{dx}(x) = f(x).$$

Dann definieren wir die Funktion H als $H(x) := F_1(x) - F_2(x)$. Damit gilt

$$\frac{dH}{dx}(x) = \frac{dF_1}{dx}(x) - \frac{dF_2}{dx}(x) = f(x) - f(x) = 0.$$

Nach Lemma 68 gibt es nun eine Konstante c, so dass $H(x) = c$ gilt. Wir kommen jetzt zu einem zentralen Ergebnis dieser Vorlesung.

Satz 86 (Hauptsatz der Differential- und Integral-Rechnung)
Die Funktion $f : [a, b] \to \mathbb{R}$ sei stetig, $F : [a, b] \to \mathbb{R}$ sei eine Stamm-Funktion von f und es gelte $u, v \in [a, b]$. Dann gilt

$$\boxed{\int_u^v f(x)\,dx = F(v) - F(u).}$$

Beweis: Definieren wir die Funktion $G(x)$ durch

$$G(x) := \int_a^x f(t)\,dt \quad \text{für alle } x \in [a, b],$$

so besagt Satz 84, dass die Funktion G eine Stamm-Funktion der Funktion f ist. Ist nun F eine beliebige weitere Stamm-Funktion von f, so haben wir gerade gesehen, dass die Stamm-Funktionen G und F sich nur um eine Konstante c unterscheiden können, es gilt also

$$F(x) = G(x) + c.$$

Damit haben wir

$$\begin{aligned} F(v) - F(u) &= G(v) + c - \big(G(u) + c\big) \\ &= G(v) - G(u) \\ &= \int_a^v f(t)\,dt - \int_a^u f(t)\,dt \\ &= \int_u^v f(t)\,dt. \end{aligned}$$

□

Der letzte Satz gibt Anlass zu einer Schreibweise. Für eine Funktion F definieren wir

$$F(x)\Big|_u^v := F(v) - F(u).$$

Den letzten Satz können wir eine wichtige Schlussfolgerung ziehen: Ist $f : [a, b] \to \mathbb{R}$ eine differenzierbare Funktion, so ist f offenbar eine Stamm-Funktion der Funktion $f'(x)$. Also gilt:

$$\boxed{\int_u^v f'(x)\,dx = f(v) - f(u) = f(x)\Big|_u^v} \tag{7.6}$$

Dies zeigt uns, dass der Integral-Operator in gewisser Weise zum Differential-Operator invers ist.

7.2 Regeln zur Berechnung von Integralen

Der Hauptsatz der Differential- und Integral-Rechnung ermöglicht es, Regeln zur Berechnung von Integralen aufzustellen. Tabelle 7.1 zeigt die Stamm-Funktionen der wichtigsten Funktionen. Um diese Tabelle zu verifizieren reicht es aus, die in der rechten Spalte der Tabelle angegebene Funktion nach x zu differenzieren. Wir führen dies exemplarisch für den Eintrag $\ln(x)$ vor. Es gilt

$$\begin{aligned} \frac{d}{dx}\big(x \cdot \ln(x) - x\big) &= \frac{d}{dx}\big(x \cdot \ln(x)\big) - \frac{d}{dx}x \\ &= 1 \cdot \ln(x) + x \cdot \frac{1}{x} - 1 \quad \text{(Produkt-Regel)} \\ &= \ln(x) + 1 - 1 \\ &= \ln(x) \end{aligned}$$

Damit ist gezeigt, dass $\int \ln(x)\;dx = x \cdot \ln(x) - x$ gilt. Die übrigen Einträge der Tabelle können auf ähnliche Weise verifiziert werden.

Bemerkenswert ist vielleicht noch die Stamm-Funktion der Funktion $x \mapsto \frac{1}{x}$: Für $x > 0$ gilt sicher

$$\frac{d}{dx}\ln(x) = \frac{1}{x}$$

Außerdem gilt nach der Ketten-Regel für $x < 0$

$$\frac{d}{dx}\ln(-x) = -1 \cdot \frac{1}{-x} = \frac{1}{x}.$$

Diese beiden Gleichungen können wir zu

$$\frac{d}{dx}\ln\big(|x|\big) = \frac{1}{x}$$

zusammen fassen und folglich haben wir

$$\int \frac{1}{x}\,dx = \ln\big(|x|\big).$$

Wir können die einzelnen Einträge der Tabelle zwar durch Differenzieren leicht verifizieren, aber

Funktion $f(x)$	Stamm-Funktion $\int f(x)\,dx$
x^α mit $\alpha \neq -1$	$\dfrac{1}{\alpha+1} \cdot x^{\alpha+1}$
$\dfrac{1}{x}$	$\ln\big(\lvert x\rvert\big)$
$\exp(x)$	$\exp(x)$
$\sin(x)$	$-\cos(x)$
$\cos(x)$	$\sin(x)$
$\tan(x)$	$-\ln\big(\lvert\cos(x)\rvert\big)$
$\ln(x)$	$x \cdot \ln(x) - x$
$\dfrac{1}{1+x^2}$	$\arctan(x)$
$\dfrac{1}{\sqrt{1-x^2}}$	$\arcsin(x)$
$\dfrac{1}{\sqrt{1+x^2}}$	$\ln\big(x+\sqrt{1+x^2}\big)$
$\arctan(x)$	$x \cdot \arctan(x) - \dfrac{1}{2} \cdot \ln\big(1+x^2\big)$

Tabelle 7.1: Tabelle einiger Integrale

dabei bleibt die Frage offen, wie die Einträge dieser Tabelle gefunden wurden. Wir stellen jetzt einige Sätze auf, mit deren Hilfe wir Stamm-Funktionen gegebener Funktionen berechnen können. Wir erhalten diese Sätze indem wir die Regeln, die wir zur Differenzierung aufgestellt haben, umdrehen.

7.2.1 Die Substitutions-Regel

Wir beginnen damit, dass wir die Ketten-Regel umstellen.

$$\begin{array}{lrcll}
 & \dfrac{d}{dx}h(f(x)) & = & f'(x)\cdot h'(f(x)) & \text{Ketten-Regel} \\
\Rightarrow & \displaystyle\int \frac{d}{dx}h(f(x))\,dx & = & \displaystyle\int f'(x)\cdot h'(f(x))\,dx \;+\; c & \text{Stamm-Funktion bilden} \\
\Rightarrow & h(f(x)) & = & \displaystyle\int f'(x)\cdot h'(f(x))\,dx \;+\; c & \text{Definition der Stamm-Funktion.}
\end{array}$$

Nach dem Hauptsatz können wir hier bei den Integralen Grenzen einsetzen. Dann erhalten wir

$$\int_a^b f'(x)\cdot h'(f(x))\,dx = h\big(f(b)\big) - h\big(f(a)\big) \tag{7.7}$$

Wir definieren nun $g(x) := h'(x)$. Dann ist h eine Stamm-Funktion der Funktion g, es gilt also $h(x) = \int g(x)\,dx \;+\; d$, was wir nach dem Hauptsatz auch als

$$h\big(f(b)\big) - h\big(f(a)\big) = \int_{f(a)}^{f(b)} g(x)\,dx \tag{7.8}$$

schreiben können. Ersetzen wir in Gleichung (7.7) h' durch g so erhalten wir zusammen mit Gleichung (7.8) die *Substitutions-Regel*:

$$\boxed{\int_a^b f'(x) \cdot g\big(f(x)\big)\,dx = \int_{f(a)}^{f(b)} g(x)\,dx}. \tag{7.9}$$

Sie können sich die Substitutions-Regel mit Hilfe der folgenden suggestiven Pseudo-Ableitung leicht merken. Wir gehen aus von dem unbestimmten Integral

$$\int g(y)\,dy.$$

Hier führen wir die Variablen-Transformation $y = f(x)$ durch. Dann gilt

$$\frac{dy}{dx} = f'(x)$$

Wir rechnen nun mit dem Ausdruck $\dfrac{dy}{dx}$ so, als ob es ein gewöhnlicher Bruch wäre und stellen die letzte Gleichung nach dy um. Dann haben wir

$$dy = f'(x) \cdot dx$$

Ersetzen wir in dem ursprünglichen Integral dy durch diesen Ausdruck, so haben wir

$$\int g(y)\,dy = \int g\big(f(x)\big) \cdot f'(x)\,dx$$

Das ist aber genau die Substitutions-Regel für unbestimmte Integrale.

Aufgabe 48: Berechnen Sie das Integral $\displaystyle\int_0^x \tan(t)\,dt$ mit Hilfe der Substitutions-Regel. ◇

Lösung: Es gilt:

$$\begin{aligned}
\int_0^x \tan(t)\,dt &= \int_0^x \frac{\sin(t)}{\cos(t)}\,dt \\
&= -\int_0^x \big(-\sin(t)\big) \cdot \frac{1}{\cos(t)}\,dt \\
&= -\int_{\cos(0)}^{\cos(x)} \frac{1}{y}\,dy \\
&= -\Big(\ln\big(|\cos(x)|\big) - \ln\big(|\cos(0)|\big)\Big) \\
&= \ln\big(|\cos(0)|\big) - \ln\big(|\cos(x)|\big) \\
&= \ln(1) - \ln\big(|\cos(x)|\big) \\
&= -\ln\big(|\cos(x)|\big)
\end{aligned}$$

Damit ist gezeigt, dass die Funktion $x \mapsto -\ln\big(|\cos(x)|\big)$ eine Stamm-Funktion der Funktion $x \mapsto \tan(x)$ ist:

$$\int \tan(x)\,dx = -\ln\big(|\cos(x)|\big) + c.$$

Wir zeigen, wie sich die Stamm-Funktion von $\tan(x)$ suggestiver berechnen lässt. Bei dem Integral

$$\int \frac{\sin(t)}{\cos(t)}\,dt$$

führen wir die Variablen-Transformation $x = \cos(t)$ durch. Dann gilt

$$\frac{dx}{dt} = -\sin(t) \;\Leftrightarrow\; dt = \frac{dx}{-\sin(t)}.$$

Ersetzen wir in dem Integral

$$\int \frac{\sin(t)}{\cos(t)}\,dt \quad \text{den Ausdruck } dt \text{ durch} \quad \frac{dx}{-\sin(t)}$$

und $\cos(t)$ durch x, so erhalten wir

$$\int \frac{\sin(t)}{\cos(t)}\,dt \;=\; \int \frac{\sin(t)}{x} \cdot \frac{dx}{-\sin(t)} \;=\; -\int \frac{1}{x}\,dx = -\ln(|x|) = -\ln\bigl(|\cos(t)|\bigr).$$

Das Rechnen mit den *infinitesimalen* Größen dx und dt ist offensichtlich intuitiver als die formale Anwendung der Substitutions-Regel. □

Aufgabe 49: Berechnen Sie das Integral

$$\int \frac{x}{1+x^2}\,dx$$

mit Hilfe der Substitutions-Regel. ◇

7.2.2 Partielle Integration

Als nächstes überlegen wir, wie wir aus der Produkt-Regel der Differential-Rechnung eine Regel zur Berechnung von Integralen gewinnen können. Es gilt

$$\begin{array}{lll}
\frac{d}{dx}\Bigl(f(x)\cdot g(x)\Bigr) &=& f'(x)\cdot g(x) + f(x)\cdot g'(x) \qquad \text{(Umstellen)} \\
\Leftrightarrow\; f'(x)\cdot g(x) &=& \frac{d}{dx}\Bigl(f(x)\cdot g(x)\Bigr) - f(x)\cdot g'(x) \qquad \text{(Stamm-Funktion)} \\
\Leftrightarrow\; \int f'(x)\cdot g(x)\,dx &=& \int \frac{d}{dx}\Bigl(f(x)\cdot g(x)\Bigr)\,dx - \int f(x)\cdot g'(x)\,dx
\end{array}$$

Die Stamm-Funktion von $\frac{d}{dx}\Bigl(f(x)\cdot g(x)\Bigr)$ ist natürlich $f(x)\cdot g(x)$, also haben wir

$$\boxed{\int f'(x)\cdot g(x)\,dx = f(x)\cdot g(x) - \int f(x)\cdot g'(x)\,dx} \tag{7.10}$$

gefunden. Diese Gleichung setzt die Stamm-Funktionen von $f'(x)\cdot g(x)$ und $f(x)\cdot g'(x)$ in Beziehung: Die Stamm-Funktion der ersten Funktion kann auf die Stamm-Funktion der zweiten Funktion zurück geführt werden. Die Regel zur partiellen Integration lässt sich nicht nur für die Berechnung der Stamm-Funktion eines Produkts einsetzen, sondern sie wird auch benutzt um die Integrale von Umkehr-Funktionen zu bestimmen. Als Beispiel zeigen wir, wie sich das Integral der Funktion $\ln(x)$ mittels partieller Integration bestimmen lässt.

$$\begin{aligned}
\int \ln(x)\,dx &= \int 1 \cdot \ln(x)\,dx \\
&= x \cdot \ln(x) - \int x \cdot \frac{d}{dx}\ln(x)\,dx \\
&= x \cdot \ln(x) - \int x \cdot \frac{1}{x}\,dx \\
&= x \cdot \ln(x) - \int 1\,dx \\
&= x \cdot \ln(x) - x
\end{aligned}$$

Aufgabe 50: Berechnen Sie die Stamm-Funktion $\int \arctan(x)\,dx$. ◇

7.2.3 Das Integral von Umkehr-Funktionen*

Wir zeigen noch einen anderen Weg, mit dem das Integral der Umkehr-Funktion einer Funktion berechnet werden kann. Es sei also eine Funktion $f : [a, b] \to \mathbb{R}$ geben, die eine Umkehr-Funktion hat. Wir setzen zur Vereinfachung jetzt voraus, dass die Funktion f monoton steigend ist, wenn f monoton fallend ist, liegen die Dinge analog. Die Umkehr-Funktion von f sei die Funktion

$$g : \big[f(a), f(b)\big] \to \mathbb{R}, \quad \text{für alle } x \in [a, b] \text{ gilt also } g\big(f(x)\big) = x.$$

Analog gilt dann für alle $y \in \big[f(a), f(b)\big]$ die Gleichung $f\big(g(y)\big) = y$. Abbildung 7.3 zeigt die Funktion. Durch Spiegelung an der Winkelhalbierung geht die Funktion f in die Umkehr-Funktion über, wenn wir also die y-Achse als x-Achse ansehen, zeigt die Abbildung die Umkehr-Funktion. Wir betrachten jetzt die Fläche des Rechtecks, dessen linke unter Ecke im Ursprung des Koordinaten-Systems liegt und dessen rechte obere Ecke die Koordinaten $\big\langle b, f(b)\big\rangle$ hat. Dieses Rechtecks hat den Flächeninhalt $b \cdot f(b)$. Diese Fläche setzt sich aus drei Teilen zusammen, die in der Figur unterschiedlich markiert sind.

1. Links unten findet sich ein kleines Rechteck, das mit diagonalen Streifen markiert ist. Die linke untere Ecke dieses Rechtecks liegt ebenfalls im Ursprung des Koordinaten-Systems, die rechte obere Ecke hat die Koordinaten $\big\langle a, f(a)\big\rangle$. Damit hat dieses Rechteck die Fläche $a \cdot f(a)$.

2. Die vertikal schraffierte Fläche ist die Fläche unter der Kurve $f(x)$ und ist folglich gegeben durch das Integral

$$\int_a^b f(x)\,dx.$$

3. Die horizontal schraffierte Fläche ist die Fläche unter der Kurve $g(y)$ der Umkehr-Funktion und ist daher gegeben durch

$$\int_{f(a)}^{f(b)} g(y)\,dy.$$

Natürlich ist die gesamte Fläche des Rechtecks die Summe aller drei Teile, es gilt also

$$b \cdot f(b) = a \cdot f(a) + \int_a^b f(t)\,dt + \int_{f(a)}^{f(b)} g(t)\,dt \tag{7.11}$$

Ersetzen wir in dieser Gleichung b durch x und stellen die Gleichung nach dem ersten Integral um, so erhalten wir

$$\int_a^x f(t)\,dt = x \cdot f(x) - a \cdot f(a) - \int_{f(a)}^{f(x)} g(t)\,dt \tag{7.12}$$

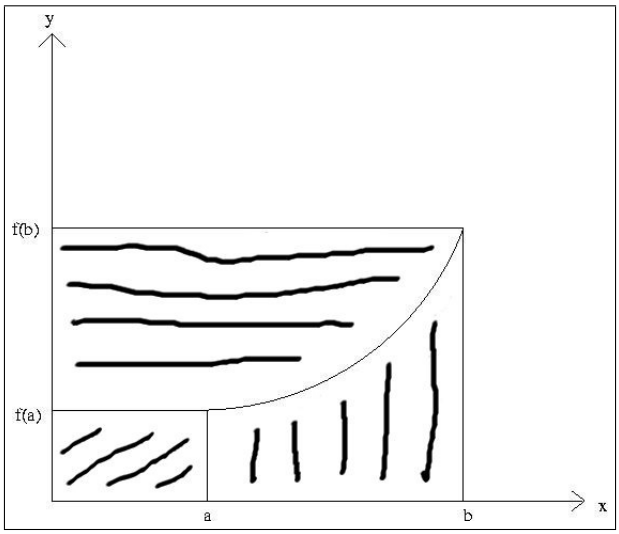

Abbildung 7.3: Die Funktion $f : [a, b] \to \mathbb{R}$

Sind wir nur an Stamm-Funktionen interessiert, so können wir den konstanten Term $a \cdot f(a)$ weglassen. Dann schreibt sich die letzte Gleichung als

$$\int f(x)\,dx = x \cdot f(x) - G\big(f(x)\big) \quad \text{mit} \quad G(x) := \int g(x)\,dx. \tag{7.13}$$

Beispiel: Wir betrachten die Funktion $f(x) = \ln(x)$ mit der Umkehr-Funktion $g(x) = \exp(x)$. Wegen $\int \exp(x)\,dx = \exp(x)$ gilt

$$\int \ln(x) = x \cdot \ln(x) - \exp\big(\ln(x)\big) = x \cdot \ln(x) - x. \qquad \diamond$$

7.2.4 Berechnung der Fläche eines Kreises

Wir zeigen, wie sich der Flächen-Inhalt eines Kreises berechnen lässt. Ein Kreis vom Radius 1 kann nach dem Satz des Pythagoras in der Koordinaten-Ebene durch die Relation $x^2 + y^2 = 1$ dargestellt werden. Lösen wir diese Gleichung nach x auf, so wird der Kreis durch die beiden Funktionen $x \mapsto \sqrt{1 - x^2}$ und $x \mapsto -\sqrt{1 - x^2}$ für $x \in [-1, 1]$ beschrieben. Wir beschränken uns nun auf den Viertelkreis im ersten Quadranten der Koordinaten-Ebene. Die Fläche dieses Viertelkreises ist durch das Integral

$$I := \int_0^1 \sqrt{1 - x^2}\,dx$$

gegeben, dass wir nun berechnen. Zunächst führen wir die Koordinaten-Transformation $x = \sin(t)$ durch. Dann erhalten wir

$$\frac{dx}{dt} = \cos(t), \quad \text{also } dx = \cos(t) \cdot dt.$$

Wegen $0 = \sin(0)$ und $1 = \sin(\pi/2)$ haben wir

$$I = \int_0^{\frac{\pi}{2}} \sqrt{1 - \sin^2(t)} \cdot \cos(t)\,dt = \int_0^{\frac{\pi}{2}} \cos(t) \cdot \cos(t)\,dt.$$

Dieses Integral versuchen wir nun mit partieller Integration zu vereinfachen. Wir setzen in Gleichung 7.10 $f'(t) := \cos(t)$ und $g(t) = \cos(t)$. Wegen $f(t) = \sin(t)$ und $g'(t) = -\sin(t)$ erhalten wir

dann

$$I = \sin(t) \cdot \cos(t)\Big|_0^{\frac{\pi}{2}} + \int_0^{\frac{\pi}{2}} \sin(t) \cdot \sin(t)\, dt = \int_0^{\frac{\pi}{2}} \sin^2(t)\, dt$$

Damit haben wir jetzt

$$I = \int_0^{\frac{\pi}{2}} \cos^2(t)\, dt = \int_0^{\frac{\pi}{2}} \sin^2(t)\, dt$$

Daraus folgt

$$\begin{aligned}
\Leftrightarrow \quad 2 \cdot I &= \int_0^{\frac{\pi}{2}} \cos^2(t)\, dt + \int_0^{\frac{\pi}{2}} \sin^2(t)\, dt \\
\Leftrightarrow \quad 2 \cdot I &= \int_0^{\frac{\pi}{2}} \left(\cos^2(t)\, dt + \sin^2(t)\right) dt \\
\Leftrightarrow \quad 2 \cdot I &= \int_0^{\frac{\pi}{2}} 1\, dt \\
\Leftrightarrow \quad 2 \cdot I &= t\,\Big|_0^{\frac{\pi}{2}} \\
\Leftrightarrow \quad 2 \cdot I &= \frac{\pi}{2} \\
\Leftrightarrow \quad I &= \frac{\pi}{4}
\end{aligned}$$

Damit haben wir für die Fläche eines Viertelkreises den Wert $\frac{\pi}{4}$ gefunden, der ganze Kreis hat also die Fläche π.

7.3 Berechnung der Bogenlänge

Gelegentlich tritt das Problem auf, die Länge der Strecke zu berechnen, die durch einen Funktions-Graphen beschrieben wird. Geht es darum, die Bogenlänge des Graphen der Funktion $f : [a, b] \to \mathbb{R}$ in dem Intervall $[a, b]$ zu berechnen, so zerlegen wir das Intervall in $[a, b]$ in n Teilintervalle der Länge

$$h = \frac{b-a}{n}.$$

Das Teilstück des Graphen, das von dem Punkt $\langle x, f(x)\rangle$ zu dem Punkt $\langle x+h, f(x+h)\rangle$ kann für kleine Werte von h näherungsweise durch die Sekante von dem Punkt $\langle x, f(x)\rangle$ zu dem Punkt $\langle x+h, f(x+h)\rangle$ approximiert werden. Nach dem Satz des Pythagoras hat diese Sekante die Länge

$$ds = \sqrt{h^2 + \big(f(x+h) - f(x)\big)^2}.$$

Nach Definition der Ableitung von f gilt für kleine Werte von h

$$\frac{f(x+h) - f(x)}{h} \approx f'(x).$$

Also haben wir

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

Setzen wir diesen Wert in der obigen Formel für ein, so erhalten wir

$$ds \approx \sqrt{h^2 + \big(f'(x) \cdot h\big)^2} = \sqrt{1 + \Big(\frac{df}{dx}\Big)^2} \cdot h.$$

Die Länge l des gesamten Funktions-Graphen erhalten wir, wenn wir diesen Ausdruck von a bis b

aufintegrieren:

$$\boxed{l = \int_a^b \sqrt{1 + \left(\frac{df}{dx}\right)^2}\, dx}.$$

Beispiel: Wir berechnen die Länge l des Kreisbogens, der durch die Funktion $f : [0,1] \to \mathbb{R}$ mit

$$f(x) = \sqrt{1 - x^2}$$

definiert ist. Anschaulich sollte diese Länge ein Viertel des Umfangs eines Kreises mit Radius 1 betragen. Es gilt

$$\frac{df}{dx} = \frac{1}{2} \cdot \frac{-2 \cdot x}{\sqrt{1-x^2}} = \frac{-x}{\sqrt{1-x^2}}.$$

Also haben wir für die Länge

$$\begin{aligned} l &= \int_0^1 \sqrt{1 + \frac{x^2}{1-x^2}}\, dx \\ &= \int_0^1 \frac{1}{\sqrt{1-x^2}}\, dx \end{aligned}$$

An dieser Stelle wenden wir die Substitutions-Regel

$$\int_a^b f'(x) \cdot g\big(f(x)\big)\, dx = \int_{f(a)}^{f(b)} g(x)\, dx$$

an, wobei wir

$$f(x) := \sin(x), \quad g(x) := \frac{1}{\sqrt{1-x^2}}, \quad a := 0 \quad \text{und} \quad b := \frac{\pi}{2}$$

setzen. Wegen

$$f'(x) = \cos(x), \quad \sin(0) = 0 \quad \text{und} \quad \sin(\pi/2) = 1$$

folgt dann

$$\begin{aligned} l &= \int_0^{\pi/2} \cos(x) \cdot \frac{1}{\sqrt{1-\sin^2(x)}}\, dx \\ &= \int_0^{\pi/2} \cos(x) \cdot \frac{1}{\cos(x)}\, dx \\ &= \int_0^{\pi/2} 1\, dx \\ &= x\Big|_0^{\pi/2} \\ &= \tfrac{\pi}{2} \end{aligned}$$

und das ist genau der Umfang eines Viertel-Kreises. ◇

Aufgabe 51: Berechnen Sie die Bogenlänge der Parabel $x \mapsto x^2$ im Intervall $[0,1]$. ◇

7.4 Uneigentliche Integrale

Ist bei dem Ausdruck $\int_a^b f(x)\,dx$ die Funktion f an einer der Intervall-Grenzen nicht definiert, so bezeichnen wir den Ausdruck als uneigentliches Integral. Wir betrachten nur den Fall, dass die Funktion f nur in dem halboffenen Intervall $(a, b]$ definiert ist, während f in dem Punkt a nicht definiert sein soll. Beispielsweise ist die Funktion $x \mapsto \frac{1}{\sqrt{x}}$ im Punkt $x = 0$ nicht definiert. In diesem Fall definieren wir

$$\int_a^b f(x)\,dx := \lim_{\substack{h\to 0\\ h>0}} \int_{a+h}^b f(x)\,dx.$$

Wir betrachten als Beispiel das Integral der Funktion $x \mapsto \dfrac{1}{\sqrt{x}}$ in dem Intervall $[0, 1]$. Es gilt

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{\substack{h\to 0\\ h>0}} \int_h^1 \frac{1}{\sqrt{x}}\,dx = \lim_{\substack{h\to 0\\ h>0}} 2\cdot\sqrt{x}\,\Big|_h^1 = \lim_{\substack{h\to 0\\ h>0}} 2\cdot\sqrt{1} - 2\cdot\sqrt{h} = 2.$$

Ein anderer Fall liegt vor, wenn eine der Integrations-Grenzen den Wert Unendlich hat. Wir betrachten nur den Fall $b = \infty$. In diesem Fall definieren wir

$$\int_a^\infty f(t)\,dt := \lim_{x\to\infty} \int_a^x f(t)\,dt.$$

Wir betrachten als Beispiel die Funktion $x \mapsto \dfrac{1}{x^2}$ in dem Intervall $[1, \infty)$. Es gilt

$$\int_1^\infty \frac{1}{t^2}\,dt = \lim_{x\to\infty} \int_1^x \frac{1}{t^2}\,dt = \lim_{x\to\infty} -\frac{1}{t}\,\Big|_1^x = \lim_{x\to\infty} -\frac{1}{x} + \frac{1}{1} = 1.$$

Indem wir Reihen und uneigentliche Integrale in Beziehung setzen, können wir unter Umständen leicht sehen, dass eine Reihe konvergiert, denn es gilt der folgende Satz.

Satz 87 (Integral-Vergleichskriterium) Es sei $f : [1, \infty) \to \mathbb{R}_+$ eine monoton fallende Funktion. Dann gilt

$$\lim_{n\to\infty} \sum_{i=1}^n f(i) \text{ existiert} \quad \text{g.d.w.} \quad \lim_{x\to\infty} \int_1^x f(t)\,dt \text{ existiert.}$$

Beweis: Für alle $i \in \mathbb{N}$ mit $i > 1$ und alle $x \in [i-1, i]$ gelten die Ungleichungen

$$i - 1 \le x \quad \text{und} \quad x \le i.$$

Da die Funktion f monoton fallend ist, folgt daraus

$$f(i) \le f(x) \quad \text{und} \quad f(x) \le f(i-1).$$

Aus der Monotonie des Integral-Operators folgt nun

$$\int_{i-1}^i f(i)\,dx \le \int_{i-1}^i f(x)\,dx \quad \text{und} \quad \int_{i-1}^i f(x)\,dx \le \int_{i-1}^i f(i-1)\,dx$$

Das Integral über die konstanten Funktionen $x \mapsto f(i)$ bzw. $x \mapsto f(i-1)$ liefert nur den konstanten Funktions-Wert multipliziert mit der Länge des Intervalls, die natürlich 1 ist. Also haben wir die Ungleichungen

$$f(i) \le \int_{i-1}^i f(x)\,dx \quad \text{und} \quad \int_{i-1}^i f(x)\,dx \le f(i-1)$$

gezeigt. Summieren wir diese Ungleichungen für alle i aus der Menge $\{2, \cdots, n\}$, so erhalten wir

$$\sum_{i=2}^{n} f(i) \le \int_1^n f(x)\,dx \quad \text{und} \quad \int_1^n f(x)\,dx \le \sum_{i=2}^{n} f(i-1)$$

Bei der ersten Ungleichung addieren wir noch $f(1)$, bei der zweiten Ungleichung ersetzen wir die Summations-Variable i durch $i+1$ und erhalten:

$$\sum_{i=1}^{n} f(i) \le f(1) + \int_1^n f(x)\,dx \quad \text{und} \quad \int_1^n f(x)\,dx \le \sum_{i=1}^{n-1} f(i)$$

Falls nun das Integral $\int_1^\infty f(x)\,dx$ existiert, so gilt

$$\sum_{i=1}^{n} f(i) \le f(1) + \int_1^n f(x)\,dx \le f(1) + \int_1^\infty f(x)\,dx$$

Damit ist dann die Folge

$$\left(\sum_{i=1}^{n} f(i)\right)_{n\in\mathbb{N}}$$

monoton und beschränkt, folglich konvergent. Existiert umgekehrt der Grenzwert $\sum_{i=1}^{\infty} f(i)$, so ist aufgrund der zweiten Ungleichung auch die Folge

$$\left(\int_1^n f(x)\,dx\right)_{n\in\mathbb{N}}$$

beschränkt. Da diese Folge außerdem monoton ist, folgt die Konvergenz. □

Beispiel: Wir betrachten das Integral

$$\int_1^\infty \frac{1}{x^\alpha}\,dx \quad \text{für } \alpha > 0.$$

Falls $\alpha \neq 1$ ist, gilt

$$\begin{aligned} I(\alpha) &:= \int_1^\infty \frac{1}{t^\alpha}\,dt \\ = \lim_{x\to\infty} \int_1^x \frac{1}{t^\alpha}\,dt &= \lim_{x\to\infty} \frac{t^{1-\alpha}}{1-\alpha}\Big|_1^x \\ = \lim_{x\to\infty} \frac{x^{1-\alpha}-1}{1-\alpha}. & \end{aligned}$$

Im Falle $\alpha = 1$ haben wir

$$I(\alpha) = \ln(x).$$

Nun hängt alles davon ab, ob $\alpha < 1$, $\alpha = 1$ oder $\alpha > 1$ ist.

1. Falls $\alpha < 1$ ist, haben wir $1-\alpha > 0$ und damit gilt

   $$I(\alpha) = \lim_{x\to\infty} \frac{x^{1-\alpha}-1}{1-\alpha} = \infty$$

   Also konvergiert die Summe

   $$\sum_{i=1}^{\infty} \frac{1}{i^\alpha}$$

   in diesem Fall nicht.

2. Falls $\alpha = 1$ ist, gilt

$$I(1) = \lim_{x\to\infty} \int_1^x \frac{1}{t}\,dt = \lim_{x\to\infty} \ln(x) = \infty$$

Im letzten Schritt haben wir hier ausgenutzt, dass

$$\lim_{x\to\infty} \ln(x) = \infty$$

gilt. Das folgt daraus, dass die Funktion $x \mapsto \ln(x)$ einerseits monoton steigend ist und dass andererseits

$$\ln\big(\exp(n)\big) = n \quad \text{für alle } n \in \mathbb{N}$$

gilt, denn die letzte Gleichung zeigt, dass die Funktion $x \mapsto \ln(x)$ beliebig groß wird.

Insgesamt können wir nun aus dem Integral-Vergleichskriterium folgern, dass die harmonische Reihe divergiert:

$$\sum_{i=1}^{\infty} \frac{1}{i} = \infty.$$

3. Falls $\alpha > 1$ ist, haben wir $1 - \alpha < 0$ und damit gilt

$$I(\alpha) = \lim_{x\to\infty} \frac{x^{1-\alpha} - 1}{1-\alpha} = \frac{1}{\alpha - 1}$$

In diesem Fall konvergiert also die Summe

$$\sum_{i=1}^{\infty} \frac{1}{i^{\alpha}}.$$

Aufgabe 52: Untersuchen Sie mit Hilfe des Integral-Vergleichskriteriums, ob die Reihe

$$\sum_{n=1}^{\infty} \frac{1}{n \cdot (n+1)} \qquad \text{konvergiert.}$$ ◇

7.5 Numerische Integration*

Die Gleichung $x = \cos(x)$ haben wir zwar numerisch lösen können, aber wir waren nicht in der Lage, einen algebraischen Ausdruck für die Lösung anzugeben. Auch bei der Integration von Funktionen ist es nicht immer möglich, einen algebraischen Ausdruck für die Stamm-Funktion einer Funktion anzugeben. Beispielsweise konnte Liouville (Joseph Liouville, 1809 – 1882) beweisen, dass das unbestimmte Integral

$$\int \exp(-x^2)\,dx$$

nicht auf die uns bisher bekannten Funktionen zurück geführt werden kann [9]. Für die praktische Berechnung von Integralen benötigen wir daher numerische Methoden.

7.5.1 Die Trapez-Regel

Die wohl naheliegenste Methode besteht darin, die zu integrierende Funktion stückweise linear zu interpolieren und dann das zur berechnende Integral durch das Integral der stückweise linearen Funktion zu approximieren. Ist das Integral $\int_a^b f(t)\,dt$ zu berechnen, so wird zunächst das Intervall $[a, b]$ in n gleich große Teilintervalle zerlegt. Das i-te Teilintervall ist $[a + (i-1)\cdot h, a + i \cdot h]$ mit $h := \frac{b-a}{n}$ und $i = 1, \cdots, n$. Wir setzen $x_i := a + i \cdot h$. Innerhalb des i-ten Teilintervalls wird die Funktion $f(x)$ dann durch die Gerade

$$g(x) = f\big(x_{i-1}\big) \cdot \frac{x - x_i}{x_{i-1} - x_i} + f\big(x_i\big) \cdot \frac{x - x_{i-1}}{x_i - x_{i-1}}$$

approximiert. Das Integral über $g(x)$ im Intervall $[x_{i-1}, x_i]$ ist dann

$$\begin{aligned}
&\int_{x_{i-1}}^{x_i} g(x)\,dx \\
= &\int_{x_{i-1}}^{x_i} f\big(x_{i-1}\big)\cdot\frac{x-x_i}{x_{i-1}-x_i} + f\big(x_i\big)\cdot\frac{x-x_{i-1}}{x_i-x_{i-1}}\,dx \\
= &\frac{f\big(x_{i-1}\big)}{x_{i-1}-x_i}\cdot\int_{x_{i-1}}^{x_i}(x-x_i)\,dx + \frac{f\big(x_i\big)}{x_i-x_{i-1}}\cdot\int_{x_{i-1}}^{x_i}(x-x_{i-1})\,dx \\
= &\frac{f\big(x_{i-1}\big)}{x_{i-1}-x_i}\cdot\frac{1}{2}\cdot(x-x_i)^2\Big|_{x_{i-1}}^{x_i} + \frac{f\big(x_i\big)}{x_i-x_{i-1}}\cdot\frac{1}{2}\cdot(x-x_{i-1})^2\Big|_{x_{i-1}}^{x_i} \\
= &-\frac{f\big(x_{i-1}\big)}{x_{i-1}-x_i}\cdot\frac{1}{2}\cdot(x_{i-1}-x_i)^2 + \frac{f\big(x_i\big)}{x_i-x_{i-1}}\cdot\frac{1}{2}\cdot(x_i-x_{i-1})^2 \\
= &\frac{1}{2}\cdot f\big(x_{i-1}\big)\cdot(x_i-x_{i-1}) + \frac{1}{2}\cdot f\big(x_i\big)\cdot(x_i-x_{i-1}) \\
= &\frac{1}{2}\cdot\Big(f\big(x_i\big)+f\big(x_{i-1}\big)\Big)\cdot(x_i-x_{i-1})
\end{aligned}$$

Die letzte Formel lässt sich geometrisch deuten, denn der Ausdruck

$$\tfrac{1}{2}\cdot\big(f(x_i)+f(x_{i-1})\big)\cdot(x_i-x_{i-1})$$

beschreibt gerade die Fläche eines Trapezes mit der Breite $h = x_i - x_{i-1}$ und Seiten der Länge $f(x_i)$ und $f(x_{i-1})$. Daher wird dieses Verfahren zur Berechnung eines Integrals auch als *Trapez-Regel* bezeichnet. Um das Integral über das gesamte Intervall $[a, b]$ zu berechnen, müssen wir die Integrale über die einzelnen Intervalle lediglich aufsummieren. Wir erhalten dann

$$\begin{aligned}
\int_a^b f(x)\,dx &\approx \sum_{i=1}^{n}\frac{1}{2}\cdot\big(f(x_i)+f(x_{i-1})\big)\cdot(x_i-x_{i-1}) \\
&= \sum_{i=1}^{n}\frac{1}{2}\cdot\big(f(a+i\cdot h)+f(a+(i-1)\cdot h)\big)\cdot h \\
&= \left(\frac{1}{2}\cdot f(a) \;+\; \sum_{i=1}^{n-1} f\Big(a+i\cdot\frac{b-a}{n}\Big) \;+\; \frac{1}{2}\cdot f(b)\right)\cdot\frac{b-a}{n}
\end{aligned}$$

Abbildung 7.4 auf Seite 138 zeigt die numerische Berechnung des Integrals $\int_0^1 \exp(x^2)\,dx$ mit Hilfe der Trapez-Regel. In diesem Fall ist $n = 3$. Führen wir die Berechnung tatsächlich aus, so erhalten wir die Näherung

$$\int_0^1 \exp(-x^2)\,dx \approx 0.7399864752\cdots.$$

Der exakte Wert ist 0.7468241328, unter Berücksichtigung der Tatsache, dass wir $n = 3$ gewählt haben ist die Näherung also gar nicht so schlecht. Erhöhen wir n auf den Wert 300, so erhalten wir mit der Trapez-Regel die Näherung 0.7468234519, der Fehler ist dann also kleiner als 10^{-6}. Im nächsten Satz geben wir eine theoretische Analyse des Fehlers.

Satz 88 Ist die Funktion $f : [a, b] \to \mathbb{R}$ zweimal differenzierbar, gilt

$$\big|f^{(2)}(x)\big| \le K \quad \text{für alle } x \in [a, b]$$

und definieren wir gemäß der Trapez-Regel

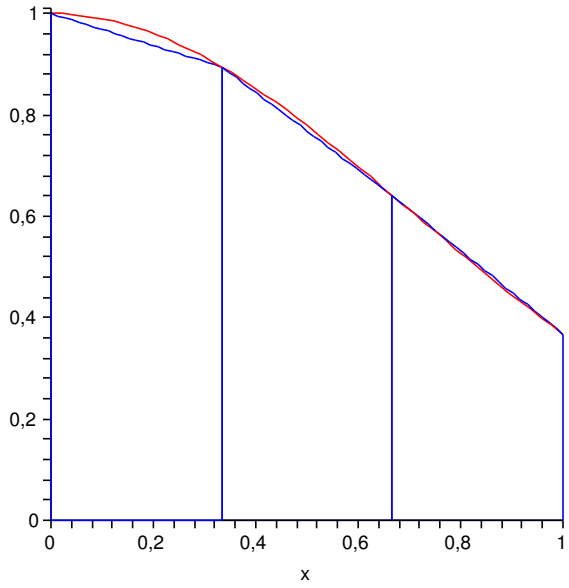

Abbildung 7.4: Berechnung des Integrals $\int_0^1 \exp(-x^2)\,dx$ mit Hilfe der Trapez-Regel.

$$I_{\text{Trapez}} := \left(\frac{1}{2}\cdot f(a) \;+\; \sum_{i=1}^{n-1} f\Big(a + i\cdot\frac{b-a}{n}\Big) \;+\; \frac{1}{2}\cdot f(b)\right)\cdot\frac{b-a}{n}$$

so kann der Unterschied zwischen dem exakten Integral und der Näherung I_{Trapez} wie folgt abgeschätzt werden:

$$\left|\int_a^b f(x)\,dx - I_{\text{Trapez}}\right| \;\leq\; \frac{K}{12}\cdot\frac{(b-a)^3}{n^2}$$

Beweis: Bei der Ableitung der Trapez-Regel haben wir $f(x)$ durch ein lineares Polynom interpoliert. Nach Satz 72 gilt für den Unterschied zwischen $f(x)$ und dem interpolierenden Polynom $p(x)$ vom Grad n

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\cdot\prod_{i=0}^{n}(x - x_i).$$

Dabei ist ξ ein nicht näher bekannter Wert aus dem Intervall $[a, b]$. Bei linearer Interpolation ist $n = 1$ und also haben wir

$$f(x) - p(x) = \frac{f^{(2)}(\xi)}{2}\cdot(x - x_{i-1})\cdot(x - x_i).$$

Aufgrund der Voraussetzung über die zweite Ableitung von $f(x)$ haben wir also

$$|f(x) - p(x)| \leq \frac{K}{2}\cdot|(x - x_{i-1})\cdot(x - x_i)| \tag{7.14}$$

In dem Intervall $[x_{i-1}, x_i]$ gilt $x - x_{i-1} \geq 0$ und $x - x_i \leq 0$, also haben wir

$$|(x - x_{i-1})\cdot(x - x_i)| = -(x - x_{i-1})\cdot(x - x_i).$$

Integrieren wir die Ungleichung 7.14 in dem Intervall $[x_{i-1}, x_i]$, so erhalten wir

$$
\begin{aligned}
& \left| \int_{x_{i-1}}^{x_i} f(x)\,dx - \int_{x_{i-1}}^{x_i} p(x)\,dx \right| \\
\leq & \int_{x_{i-1}}^{x_i} \left| f(x)\,dx - p(x) \right| dx \\
\leq & \frac{K}{2} \cdot \int_{x_{i-1}}^{x_i} \left| (x - x_{i-1}) \cdot (x - x_i) \right| dx \\
\leq & -\frac{K}{2} \cdot \int_{x_{i-1}}^{x_i} (x - x_{i-1}) \cdot (x - x_i)\,dx \\
= & -\frac{K}{2} \cdot \frac{-1}{6} \cdot (x_i - x_{i-1})^3 \\
= & \frac{K}{12} \cdot (x_i - x_{i-1})^3
\end{aligned}
$$

Diese Ungleichung gilt nun für jedes $i = 1, \cdots, n$. Summieren wir diese Ungleichungen für alle Intervalle auf, so erhalten wir

$$
\begin{aligned}
\left| \int_a^b f(x)\,dx - I_{\text{Trapez}} \right| & \leq \sum_{i=1}^{n} \frac{K}{12} \cdot (x_i - x_{i-1})^3 \\
& \leq \sum_{i=1}^{n} \frac{K}{12} \cdot \Big(\frac{b-a}{n} \Big)^3 \\
& \leq n \cdot \frac{K}{12} \cdot \Big(\frac{b-a}{n} \Big)^3 \\
& \leq \frac{K}{12} \cdot \frac{(b-a)^3}{n^2}
\end{aligned}
$$

Falls wir also die Zahl der Intervalle verzehnfachen, sinkt der Fehler auf ein Hundertstel. □

Aufgabe 53: Berechnen Sie, in wieviele Teilintervalle das Intervall $[0, 1]$ aufgeteilt werden muss, wenn das Integral

$$\int_0^1 e^{-x^2}\,dx$$

mit Hilfe der Trapez-Regel mit einer Genauigkeit von 10^{-6} berechnet werden soll. ◇

7.5.2 Die Simpson'sche Regel

Anstatt die zu integrierende Funktion f durch ein lineares Polynom zu interpolieren, können wir f auch durch ein Polynom zweiten Grades interpolieren. Wir brauchen dann natürlich drei Stützstellen. Daher integrieren wir jetzt nicht mehr über ein Intervall $[x_{i-1}, x_i]$, sondern nehmen statt dessen das Intervall $[x_{i-1}, x_{i+1}]$ und benutzen x_{i-1}, x_i und x_{i+1} als Stützstellen. Die Funktion $f(x)$ approximieren wir in dem Intervall mit der Methode von Lagrange nach der Formel

$$
\begin{aligned}
p(x) = & f(x_{i-1}) \cdot \frac{(x - x_i) \cdot (x - x_{i+1})}{(x_{i-1} - x_i) \cdot (x_{i-1} - x_{i+1})} \\
+ & f(x_i) \cdot \frac{(x - x_{i-1}) \cdot (x - x_{i+1})}{(x_i - x_{i-1}) \cdot (x_i - x_{i+1})} \\
+ & f(x_{i+1}) \cdot \frac{(x - x_{i-1}) \cdot (x - x_i)}{(x_{i+1} - x_{i-1}) \cdot (x_{i+1} - x_i)}
\end{aligned}
$$

Setzen wir hier $x_{i-1} = x_i - h$ und $x_{i+1} = x_i + h$ und integrieren dann $p(x)$ in dem Intervall $[x_i - h, x_i + h]$, so erhalten wir mit Hilfe von *SymPy* das Ergebnis

$$\int_{x_{i-1}}^{x_{i+1}} p(x)\,dx = \frac{1}{6} \cdot \big(f(x_{i-1}) + 4 \cdot f(x_i) + f(x_{i+1})\big) \cdot (x_{i+1} - x_{i-1})$$

Unterteilen wir das Intervall $[a, b]$ lediglich in zwei Teilintervalle $[a, \frac{a+b}{2}]$ und $[\frac{a+b}{2}, b]$, so können wir die obige Formel direkt verwenden. Wir erhalten dann die *Kepler'sche Fassregel* (Johannes Kepler, 1571 – 1630).

$$\int_a^b f(x)\,dx \approx \frac{1}{6} \cdot \Big(f(a) + 4 \cdot f\Big(\frac{a+b}{2}\Big) + f(b)\Big) \cdot (b - a) \tag{7.15}$$

Berechnen wir das Integral $\int_0^1 \exp(-x^2)\,dx$ mit dieser Regel, so erhalten wir 0.7471804290 und der Fehler ist bereits kleiner als $4 \cdot 10^{-4}$.

Unterteilen wir das Intervall $[a, b]$ in n Intervalle und ist darüber hinaus n eine gerade Zahl, so können wir die Regel 7.15 jeweils auf die beiden benachbarten Intervalle $[x_{2\cdot i}, x_{2\cdot i+1}]$ und $[x_{2\cdot i+1}, x_{2\cdot i+2}]$ anwenden, wobei der Index i über alle Elemente der Menge $\{0, \cdots, (n-1)/2\}$ läuft. Setzen wir $h := \frac{b-a}{n}$ und $x_i := a + i \cdot h$, so erhalten wir die Formel

$$\begin{aligned} \int_a^b f(x)\,dx &\approx \frac{h}{3} \cdot \sum_{i=0}^{(n-1)/2} \Big(f(x_{2\cdot i}) + 4 \cdot f(x_{2\cdot i+1}) + f(x_{2\cdot i+2})\Big) \\ &= \frac{h}{3} \cdot \Big(f(x_0) \;+\; 4 \cdot \sum_{i=0}^{(n-1)/2} f(x_{2\cdot i+1}) \;+\; 2 \cdot \sum_{i=1}^{(n-1)/2} f(x_{2\cdot i}) \;+\; f(x_n)\Big) \end{aligned} \tag{7.16}$$

Die obige Formel trägt den Namen Simpson'sche Regel (Thomas Simpson, 1710 – 1761). Falls die Funktion f insgesamt viermal differenzierbar ist und falls darüber hinaus die vierte Ableitung von f der Ungleichung

$$|f^{(4)}(x)| \leq K$$

genügt, so lässt sich der Fehler, der bei der Verwendung der Simpson'schen Regel entsteht, durch

$$\frac{K}{180} \cdot \frac{(b-a)^5}{n^4}$$

abschätzen. Verdoppeln wir die Zahl n der Intervalle, so verkleinert sich der Fehler also um das 16-fache! Wählen wir beispielswiese $n = 20$ und berechen das Integral $\int_0^1 \exp(-x^2)\,dx$, so erhalten wir den Wert 0.746824183 und der Fehler ist kleiner als 10^{-7}.

Aufgabe 54:

(a) Berechnen Sie mit Hilfe der Kepler'schen Fass-Regel eine Approximation für das Integral

$$\int_0^{\frac{1}{2}} \sin(x)\,dx.$$

(b) Geben Sie eine möglichst genaue Abschätzung für den Approximations-Fehler.

(c) Vergleichen Sie ihr Ergebnis mit dem exakten Wert. ◇

Aufgabe 55: Gegenstand dieser Aufgabe ist die numerische Berechnung der Summe

$$\sum_{k=1}^{\infty} \frac{1}{k^2}.$$

Gehen Sie zur Berechnung dieser Summe in folgenden Schritten vor.

(a) Approximieren Sie die Rest-Summe $\sum_{k=n}^{\infty} \frac{1}{k^2}$ durch ein geeignetes Integral.

Hinweis: Es gilt $f(k) = \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} f(k)\,dt \approx \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} f(t)\,dt.$

(b) Berechnen Sie eine obere Abschätzung für den Approximations-Fehler, den Sie bei der Integration in Teil (a) erhalten.

Hinweis: Schätzen Sie die auftretenden Summen durch Integrale nach oben ab.

(c) Berechnen Sie, wir groß Sie n wählen müssen, damit der Approximations-Fehler kleiner als 10^{-6} bleibt.

(d) Geben Sie nun einen Näherungs-Wert für die Summe $\sum_{k=1}^{\infty} \frac{1}{k^2}$ an, der sich von dem exakten Ergebnis um weniger als 10^{-6} unterscheidet.

Kapitel 8

Die Kreiszahl π und die Euler'sche Zahl e sind irrational*

In diesem Kapitel zeigen wir, dass sowohl die Kreiszahl π, die als Fläche eines Kreises mit dem Radius 1 definiert ist, als auch die Euler'sche Zahl e, die als Grenzwert der Reihe

$$\exp(1) := \sum_{k=0}^{\infty} \frac{1}{k!}$$

festgelegt ist, irrational sind. Da der Nachweis der Irrationalität von e einfacher ist, beginnen wir damit.

8.1 Die Euler'sche Zahl e ist irrational

Nach Definition von e gilt

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

Für alle $n \in \mathbb{N}$ definieren wir die n-te Partial-Summe s_n als

$$s_n := \sum_{k=0}^{n} \frac{1}{k!}.$$

Als nächstes definieren wir für alle natürlichen Zahlen $n \in \mathbb{N}$ den n-ten Rest

$$r_n := e - s_n = \sum_{k=0}^{\infty} \frac{1}{k!} - \sum_{k=0}^{n} \frac{1}{k!} = \sum_{k=n+1}^{\infty} \frac{1}{k!}$$

Offenbar gilt

$$0 < r_n, \tag{8.1}$$

denn der n-te Rest r_n enthält auf jeden Fall den Term $\frac{1}{(n+1)!}$ und der ist positiv. Wir wollen nun den n-ten Rest r_n nach oben hin abschätzen. Dazu benötigen wir zunächst die folgende Abschätzung, die für alle $k > n+1$ gültig ist:

$$\begin{aligned}
& \frac{(n+1)!}{k!} && < && \frac{1}{(n+1)^{k-(n+1)}} \\
\Leftrightarrow \quad & \frac{k!}{(n+1)!} && > && (n+1)^{k-(n+1)} \\
\Leftrightarrow \quad & \underbrace{(n+2)\cdot(n+3)\cdot\ldots\cdot(k-1)\cdot k}_{k-(n+1) \text{ Faktoren}} && > && \underbrace{(n+1)\cdot\ldots\cdot(n+1)}_{k-(n+1) \text{ Faktoren}}
\end{aligned}$$

Die letzte Ungleichung ist richtig, denn die Faktoren auf der linken Seite haben die Form $(n+1)+i$ mit $i \in \{1, \cdots, k-(n+1)\}$ und offenbar gilt

$$(n+1)+i > n+1 \quad \text{für } i \in \{1, \cdots, k-(n+1)\},$$

so dass zu jedem Faktor in dem Produkt auf der linken Seite ein kleinerer Faktor auf der rechten Seite korrespondiert. Falls $k = (n+1)$ ist, gilt

$$\frac{(n+1)!}{k!} = \frac{1}{(n+1)^{k-(n+1)}} = 1,$$

was man unmittelbar durch Einsetzen bestätigen kann. Wir haben also insgesamt Folgendes gezeigt:

$$\frac{(n+1)!}{k!} < \frac{1}{(n+1)^{k-(n+1)}} \quad \text{für alle } k > n+1 \tag{8.2}$$

und für $k = n+1$ haben wir die Gleichheit beider Seiten. Nun gilt für alle $n \in \mathbb{N}$ mit $n \geq 1$ die folgende Ungleichungs-Kette:

$$\begin{aligned}
r_n &= \sum_{k=n+1}^{\infty} \frac{1}{k!} \\
&= \frac{1}{(n+1)!} \cdot \sum_{k=n+1}^{\infty} \frac{(n+1)!}{k!} \\
&< \frac{1}{(n+1)!} \cdot \sum_{k=n+1}^{\infty} \frac{1}{(n+1)^{k-(n+1)}} && \text{(nach Gleichung (8.2))} \\
&= \frac{1}{(n+1)!} \cdot \sum_{k=0}^{\infty} \frac{1}{(n+1)^{k}} && \text{(Index-Verschiebung)} \\
&= \frac{1}{(n+1)!} \cdot \frac{1}{1-\frac{1}{n+1}} && \text{(geometrische Reihe)} \\
&= \frac{1}{(n+1)!} \cdot \frac{1}{\frac{n+1-1}{n+1}} && \text{(Hauptnenner)} \\
&= \frac{1}{(n+1)!} \cdot \frac{n+1}{n} \\
&= \frac{1}{n! \cdot n}
\end{aligned}$$

Multiplizieren wir die resultierende Ungleichung mit $n!$, so sehen wir, dass

$$n! \cdot r_n < \frac{1}{n}$$

gilt. Fassen wir diese Gleichung zusammen mit der Gleichung (8.1), so haben wir insgesamt

$$0 < n! \cdot r_n < \frac{1}{n} \quad \text{falls } n \geq 1 \text{ ist.} \tag{8.3}$$

Damit ist aber klar, dass der Ausdruck $n! \cdot r_n$ für $n \geq 1$ keine natürliche Zahl sein kann.

Theorem 89 Die Eulersche Zahl e ist irrational.

Beweis: Wir nehmen an, dass e rational ist. Dann gibt es natürliche Zahlen $p, q \in \mathbb{N}$ mit $q \geq 1$ und

$$e = \frac{p}{q}.$$

Wir betrachten den Ausdruck $q! \cdot r_q = q! \cdot (e - s_q)$ und setzten dort für e den Wert $\frac{p}{q}$ ein:

$$q! \cdot r_q = q! \cdot \Big(\frac{p}{q} - \sum_{k=0}^{q} \frac{1}{k!} \Big) = (q-1)! \cdot p - \sum_{k=0}^{q} \frac{q!}{k!} \in \mathbb{Z},$$

denn $(q-1)! \cdot p$ ist auf jeden Fall eine natürliche Zahl und für $k \leq q$ hat der Ausdruck $\frac{q!}{k!}$ die Form

$$\frac{q!}{k!} = \frac{1 \cdot 2 \cdot \ldots \cdot k \cdot (k+1) \cdot (k+2) \cdot \ldots \cdot (q-1) \cdot q}{1 \cdot 2 \cdot \ldots \cdot k} = (k+1) \cdot (k+2) \cdot \ldots \cdot (q-1) \cdot q$$

und das ist ebenfalls eine natürliche Zahl. Damit haben wir aber einen Widerspruch, denn die Aussagen

$$0 < q! \cdot r_q < \tfrac{1}{q} \quad \text{und} \quad q! \cdot r_q \in \mathbb{Z}$$

sind unvereinbar. □

8.2 Die Kreiszahl π ist irrational

Zur Vorbereitung des Beweises benötigen wir das folgende Lemma.

Lemma 90 Für alle natürlichen Zahlen n gilt:

$$\int_0^\pi f(x) \cdot \sin(x)\, dx = \sum_{k=0}^{n} (-1)^k \cdot \big(f^{(2k)}(\pi) + f^{(2k)}(0) \big) + (-1)^{n+1} \cdot \int_0^\pi f^{(2n+2)}(x) \cdot \sin(x)\, dx.$$

Beweis: Zur Abkürzung definieren wir

$$I := \int_0^\pi f(x) \cdot \sin(x)\, dx$$

und

$$S_n := \sum_{k=0}^{n} (-1)^k \cdot \big(f^{(2k)}(\pi) + f^{(2k)}(0) \big) + (-1)^{n+1} \cdot \int_0^\pi f^{(2n+2)}(x) \cdot \sin(x)\, dx.$$

Die Behauptung

$$I = S_n$$

wird nun durch Induktion nach n bewiesen. Dabei werden wir sowohl im Induktions-Anfang als auch im Induktions-Schritt zwei partielle Integrationen durchführen.

I.A.: $n = 0$.

Nach Definition von I gilt

$$I = \int_0^\pi f(x) \cdot \sin(x)\, dx$$

Wir integrieren partiell und setzen $u(x) = f(x)$ und $v'(x) = \sin(x)$. Dann gilt $u'(x) = f'(x)$ und $v(x) = -\cos(x)$. Also haben wir

$$I = -f(x) \cdot \cos(x)\Big|_0^\pi + \int_0^\pi f'(x) \cdot \cos(x).$$

Für den ersten Summanden auf der rechten Seite dieser Gleichung finden wir

$$-f(x) \cdot \cos(x)\Big|_0^\pi = -f(\pi) \cdot \cos(\pi) + f(0) \cdot \cos(0) = f(\pi) + f(0).$$

Um das verbleibende Integral zu berechnen führen wir eine erneute partielle Integration durch, bei der wir diesmal $u(x) = f'(x)$ und $v'(x) = \cos(x)$ setzen. Dann gilt $u'(x) = f^{(2)}(x)$

und $v(x) = \sin(x)$. Damit finden wir für das Integral I den Wert

$$I = f(\pi) + f(0) - \int_0^\pi f^{(2)}(x) \cdot \sin(x)\, dx.$$

Auf der anderen Seite haben wir

$$\begin{aligned} S_0 &= \sum_{k=0}^{0} (-1)^k \cdot \left(f^{(2k)}(\pi) + f^{(2k)}(0)\right) + (-1)^{0+1} \cdot \int_0^\pi f^{(2\cdot 0+2)}(x) \cdot \sin(x)\, dx \\ &= (-1)^0 \cdot \left(f^{(0)}(\pi) + f^{(0)}(0)\right) - \int_0^\pi f^{(2)}(x) \cdot \sin(x)\, dx \\ &= f(\pi) + f(0) - \int_0^\pi f^{(2)}(x) \cdot \sin(x)\, dx \\ &= I. \end{aligned}$$

I.S.: $n \mapsto n+1$

Zur Abkürzung definieren wir

$$J_n = \int_0^\pi f^{(2n+2)}(x) \cdot \sin(x)\, dx$$

Wir berechnen J_n durch partielle Integration und setzen $u(x) := f^{(2n+2)}(x)$ und $v'(x) := \sin(x)$. Dann haben wir $u'(x) = f^{(2n+3)}(x)$ und $v(x) = -\cos(x)$. Das liefert

$$J_n = -f^{(2n+2)}(x) \cdot \cos(x)\Big|_0^\pi + \int_0^\pi f^{(2n+3)}(x) \cdot \cos(x)\, dx.$$

Wegen $\cos(\pi) = -1$ und $\cos(0) = 1$ vereinfacht sich der erste Summand auf der rechten Seite wie folgt:

$$-f^{(2n+2)}(x) \cdot \cos(x)\Big|_0^\pi = f^{(2n+2)}(\pi) + f^{(2n+2)}(0).$$

Das auf der rechten Seite der obigen Gleichung verbleibende Integral berechnen wir durch eine weitere partielle Integration. Diesmal setzen wir $u(x) = f^{(2n+3)}(x)$ und $v'(x) = \cos(x)$. Dann haben wir $u'(x) = f^{(2n+3)}(x)$ und $v(x) = \sin(x)$ und für das Integral finden wir

$$\begin{aligned} \int_0^\pi f^{(2n+3)}(x) \cdot \cos(x)\, dx &= f^{(2n+3)}(x) \cdot \sin(x)\Big|_0^\pi - \int_0^\pi f^{(2n+4)}(x) \cdot \sin(x)\, dx \\ &= -\int_0^\pi f^{(2n+4)}(x) \cdot \sin(x)\, dx \end{aligned}$$

Insgesamt haben wir damit für J_n den Ausdruck

$$\begin{aligned} J_n &= f^{(2n+2)}(\pi) + f^{(2n+2)}(0) - \int_0^\pi f^{(2n+4)}(x) \cdot \sin(x)\, dx \\ &= f^{(2n+2)}(\pi) + f^{(2n+2)}(0) - J_{n+1} \end{aligned}$$

gefunden. Jetzt rechnen wir wie folgt:

$$\begin{aligned} I &\overset{IV}{=} \sum_{k=0}^{n} (-1)^k \cdot \left(f^{(2k)}(\pi) + f^{(2k)}(0)\right) + (-1)^{n+1} \cdot J_n \\ &= \sum_{k=0}^{n} (-1)^k \cdot \left(f^{(2k)}(\pi) + f^{(2k)}(0)\right) + (-1)^{n+1} \cdot \left(f^{(2n+2)}(\pi) + f^{(2n+2)}(0) - J_{n+1}\right) \\ &= \sum_{k=0}^{n+1} (-1)^k \cdot \left(f^{(2k)}(\pi) + f^{(2k)}(0)\right) + (-1)^{n+2} \cdot J_{n+1} \;=\; S_{n+1} \end{aligned}$$

Damit haben wir gezeigt, dass $I = S_{n+1}$ gilt und die Induktion ist abgeschlossen. □

Theorem 91 Die Kreiszahl π ist irrational.

Beweis: Wir führen den Beweis indirekt und nehmen an, dass $\pi \in \mathbb{Q}$ ist. Dann gibt es Zahlen $p, q \in \mathbb{N}$ mit

$$\pi = \frac{p}{q}.$$

Für beliebige $n \in \mathbb{N}$ definieren wir das Polynom $g_n(x)$ wie folgt:

$$g_n(x) := \frac{1}{n!} \cdot x^n \cdot (p - q \cdot x)^n.$$

Es gilt

$$\begin{aligned} g_n(\pi - x) &= \frac{1}{n!} \cdot \left(\frac{p}{q} - x\right)^n \cdot \left(p - q \cdot \left(\frac{p}{q} - x\right)\right)^n \\ &= \frac{1}{n!} \cdot \left(\frac{p}{q} - x\right)^n \cdot (q \cdot x)^n \\ &= \frac{1}{n!} \cdot (p - q \cdot x)^n \cdot x^n \\ &= g_n(x) \end{aligned}$$

Diese Gleichung überträgt sich natürlich auf die Ableitungen und daher haben wir

$$g_n^{(k)}(\pi - x) = (-1)^k \cdot g_n^{(k)}(x).$$

Offenbar ist g_n ein Polynom vom Grad $2 \cdot n$ und damit ist klar, dass $g_n^{(2n+2)}(x) = 0$ ist. Setzen wir in der Behauptung des letzten Lemmas für f die Funktion $g_n(x)$ ein, so folgt daher

$$\int_0^\pi g_n(x) \cdot \sin(x)\, dx = \sum_{k=0}^{n} (-1)^k \cdot \Big(g_n^{(2k)}(\pi) + g_n^{(2k)}(0)\Big).$$

Wir zeigen, dass alle Summanden in der Summe auf der rechten Seite dieser Gleichung ganze Zahlen sind. Dabei reicht es aus, dies für die Summanden der Form $g_n^{(2k)}(0)$ zu zeigen, denn es gilt

$$g_n^{(2k)}(\pi) = (-1)^{2 \cdot k} \cdot g_n^{(2k)}(\pi - \pi) = g_n^{(2k)}(0).$$

Wir zeigen mit Hilfe einer Fallunterscheidung, dass $g_n^{(k)}(0) \in \mathbb{N}$ für alle $k \in \mathbb{N}$ gilt.

(1) Fall: $k < n$.

Da das Polynom $g_n(x)$ die Form

$$g_n(x) = \frac{1}{n!} \cdot \sum_{i=n}^{2 \cdot n} c_i \cdot x^i$$

mit Koeffizienten $c_i \in \mathbb{Z}$ hat, folgt, dass für $k < n$

$$g_n^{(k)}(x) = \frac{1}{n!} \cdot \sum_{i=n}^{2 \cdot n} \frac{i!}{(i-k)!} \cdot c_i \cdot x^{i-k}$$

gilt. Jeder Term dieser Summe enthält mindestens den Faktor x. Setzen wir hier für x den Wert 0 ein, so wird daher jeder Term in der Summe 0. Damit gilt

$$g_n^{(k)}(0) = 0 \in \mathbb{N}.$$

(2) Fall: $k \geq n$.

Diesmal verschwinden beim Ableiten alle Summanden mit Index $i < k$. Wir haben also

$$\begin{aligned} g_n^{(k)}(x) &= \frac{1}{n!} \cdot \sum_{i=k}^{2\cdot n} \frac{i!}{(i-k)!} \cdot c_i \cdot x^{i-k} \\ &= \sum_{i=k}^{2\cdot n} \frac{k!}{n!} \cdot \frac{i!}{k! \cdot (i-k)!} \cdot c_i \cdot x^{i-k} \\ &= \sum_{i=k}^{2\cdot n} \frac{k!}{n!} \cdot \binom{i}{k} \cdot c_i \cdot x^{i-k} \end{aligned}$$

Für $x = 0$ folgt dann

$$g_n^{(k)}(0) = \frac{k!}{n!} \cdot \binom{k}{k} \cdot c_k = \frac{k!}{n!} \cdot c_k \in \mathbb{Z},$$

denn wenn $k \geq n$ ist, ist $\frac{k!}{n!}$ eine natürliche Zahl.

Insgesamt wissen wir jetzt, dass das Integral

$$I_n := \int_0^\pi g_n(x) \cdot \sin(x)\, dx$$

für alle $n \in \mathbb{N}$ eine ganze Zahl ist. Für alle $x \in [0, \pi]$ gilt nun

$$0 \leq \sin(x) \quad \text{und} \quad 0 \leq g_n(x).$$

Also muss auch

$$0 < I_n$$

gelten. Die Ungleichung ist echt, denn die beiden Funktionen $\sin(x)$ und $g_n(x)$ haben nur bei $x = 0$ und $x = \pi$ eine Nullstelle. Außerdem gilt für alle $x \in [0, \pi]$

$$\sin(x) \leq 1 \quad \text{und} \quad g_n(x) \leq \frac{1}{n!} \cdot \pi^n \cdot p^n.$$

Die letzte dieser beiden Ungleichungen folgt aus der Tatsache, dass einerseits $x \leq \pi$ und andererseits $p - q \cdot x \leq p$ ist. Aus den beiden oberen Ungleichungen folgt durch Intergration

$$0 \leq I_n \leq \pi \cdot \frac{\pi^n \cdot p^n}{n!}$$

Nun gilt

$$\lim_{n\to\infty} \pi \cdot \frac{\pi^n \cdot p^n}{n!} = 0$$

Daher gibt es ein $n \in \mathbb{N}$, so dass $\pi \cdot \frac{\pi^n \cdot p^n}{n!} < 1$ ist und für dieses n haben wir

$$0 < I_n < 1.$$

Das ist aber ein Widerspruch dazu, dass wir oben nachgewiesen haben, dass I_n eine ganze Zahl ist. □

8.3 Tranzendente Zahlen

Definition 92 (Algebraische Zahlen)
Eine Zahl $r \in \mathbb{R}$ heißt *algebraisch* genau dann, wenn es ein Polynom

$$p(x) = \sum_{i=0}^{n} a_i \cdot x^i \quad \text{mit } a_i \in \mathbb{Z} \text{ für alle } i = 0, 1, \cdots, n$$

gibt, so dass r Nullstelle des Polynoms p ist, es muss also gelten

$$p(r) = \sum_{i=0}^{n} a_i \cdot r^i = 0. \qquad \diamond$$

Bei der obigen Definition ist die Forderung, dass die Koeffizienten a_i ganze Zahlen sind, entscheidend, denn sonst könnten wir zu beliebigem $r \in \mathbb{R}$ einfach das Polyom

$$p_r(x) := x - r$$

definieren und offenbar gilt $p_r(r) = 0$. Ein solches Polyom ist zur Definition einer algebraischen Zahl aber nur zugelassen, wenn r eine ganze Zahl ist. $\diamond$

Beispiel: Jede rationale Zahl r ist eine algebraische Zahl, denn wenn $r \in \mathbb{Q}$ ist, dann gibt es ganze Zahlen a und b mit $b \neq 0$, so dass

$$r = \frac{a}{b}$$

gilt. Damit können wir ein Polynom p als

$$p(x) := a - b \cdot x$$

definieren. Für dieses Polynom gilt dann

$$p(r) = p\left(\frac{a}{b}\right) = a - b \cdot \frac{a}{b} = a - a = 0$$

und damit ist gezeigt, dass jede rationale Zahl r algebraisch ist. Der Begriff der algebraischen Zahlen ist also eine Verallgemeinerung des Begriffs der rationalen Zahlen. $\diamond$

Beispiel: Die Zahl $\sqrt{2}$ ist eine algebraische Zahl, denn wenn wir das Polynom p als

$$p(x) := x^2 - 2$$

definieren, gilt offenbar

$$p\left(\sqrt{2}\right) = \left(\sqrt{2}\right)^2 - 2 = 2 - 2 = 0.$$

Dieses Beispiel zeigt, dass es sich bei dem Begriff der algebraischen Zahlen um eine echte Verallgemeinerung des Begriffs der rationalen Zahlen handelt, denn wir haben ja bereits im ersten Semester gesehen, dass die Zahl $\sqrt{2}$ keine rationale Zahl ist. $\diamond$

Aufgabe 56*: Zeigen Sie, dass die Zahl $\sqrt{2} + \sqrt{3}$ eine algebraische Zahl ist. Zeigen Sie außerdem, dass diese Zahl keine rationale Zahl ist. $\diamond$

Definition 93 (Transzendente Zahl) Eine Zahl $x \in \mathbb{R}$ ist genau dann *transzendent*, wenn x nicht algebraisch ist.

Es lässt sich zeigen, dass die Menge aller Polynome mit ganzzahligen Koeffizienten abzählbar ist. Damit ist natürlich auch die Menge der algebraischen Zahlen abzählbar. Da die Menge der reellen Zahlen überabzählbar ist, muss es also sehr viele reelle Zahlen geben, die transzendent sind. Allerdings ist der Nachweis der Transzendenz einer Zahl in der Regel recht aufwändig.

Theorem 94 (Charles Hermite, 1873)
Die Eulersche Zahl e ist transzendent.

Theorem 95 (Ferdinand von Lindemann, 1882)
Die Kreiszahl π ist transzendent.

Leider bleibt in dieser Vorlesung keine Zeit mehr zum Nachweis dieser beiden Theoreme. Unter

http://www.mathematik.uni-muenchen.de/~fritsch/euler.pdf

finden Sie eine Ausarbeitung des Nachweises der Transzendenz von e, einen Nachweis der Transzendenz von π finden Sie in dem folgenden Artikel von Herrn Prof. Fritsch:

http://www.mathematik.uni-muenchen.de/~fritsch/pi.pdf.

Aufgabe 57*: Zeigen Sie, dass für alle natürlichen Zahlen $n \in \mathbb{N}$

$$\int_0^\infty t^n \cdot e^{-t} dt = n!$$

gilt. ◇

Bemerkung: Die letzte Gleichung motiviert die folgende Definition der *Gamma-Funktion*. Wir setzen

$$\Gamma(x) := \int_0^\infty t^{x-1} \cdot e^{-t} dt.$$

Mit dieser Definition gilt

$$\Gamma(n+1) = n!$$

und daher können wir die Gamma-Funktion als eine Erweiterung der Fakultäts-Funktion auf die natürlichen Zahlen auffassen.

Kapitel 9

Fourier-Analyse*

Bei der Fourier-Analyse (Jean Baptiste Joseph Fourier; 1768 - 1830) zerlegen wir eine periodische Funktion in Sinus-Schwingungen verschiedener Frequenzen. Dieses Verfahren wird in der Praxis zur Ton- und Bild-Verarbeitung eingesetzt. Außerdem ist die Fourier-Analyse ein wichtiges Hilfsmittel zur Lösung von Differenzial-Gleichungen. Im Rahmen dieser Vorlesung werden wir die Fourier-Analyse allerdings nur zur Berechnung unendlichen Reihen einsetzen, denn für die anderen Anwendungen reicht die Zeit nicht aus.

Bei der Fourier-Analyse gehen wir davon aus, dass eine Funktion $f : \mathbb{R} \to \mathbb{R}$ gegeben ist, die *periodisch* mit der Periode $2 \cdot \pi$ ist, d. h. es gilt

$$\forall x \in \mathbb{R} : f(x + 2 \cdot \pi) = f(x).$$

Ein triviales Beispiel für eine periodische Funktion ist die konstante Funktion $x \mapsto c$. Das typische Beispiel einer periodischen Funktion ist die Funktion $x \mapsto \sin(x)$, denn es gilt $\sin(x+2\cdot\pi) = \sin(x)$. Genauso ist auch die Funktion $x \mapsto \cos(x)$ periodisch mit der Periode $2 \cdot \pi$. Weitere Beispiele für periodische Funktionen sind die Funktionen

$$x \mapsto \sin(n \cdot x) \quad \text{und} \quad x \mapsto \cos(n \cdot x) \quad \text{für } n \in \mathbb{N}.$$

Aus diesen Funktionen lassen sich weitere periodische Funktionen durch Linear-Kombination erhalten, denn wenn f und g zwei periodische Funktionen mit der Periode $2 \cdot \pi$ sind, so ist natürlich auch die Funktion

$$x \mapsto \alpha \cdot f(x) + \beta \cdot g(x) \quad \text{für } \alpha, \beta \in \mathbb{R}$$

eine periodische Funktion der Periode $2 \cdot \pi$. Die grundlegende Idee bei der Fourier-Analyse besteht nun darin, dass sich jede halbwegs normale[1] periodische Funktion als unendliche Linear-Kombination der oben vorgestellten Funktionen darstellen lässt. Genauer definieren wir folgendes:

Definition 96 (Fourier-Reihe) Es seien $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}\setminus\{0\}}$ Folgen reeller Zahlen. Dann bezeichnen wir den Ausdruck

$$\frac{1}{2} \cdot a_0 \; + \; \sum_{k=1}^{\infty} a_k \cdot \cos(k \cdot x) \; + \; \sum_{k=1}^{\infty} b_k \cdot \sin(k \cdot x) \tag{9.1}$$

als die mit den Folgen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ gebildete Fourier-Reihe.

9.1 Berechnung der Fourier-Koeffizienten

Die zentrale Frage bei der *Fourier-Analyse* ist es, für eine gegebene periodische Funktion f die *Fourier-Koeffizienten* a_k und b_k zu berechnen. Ist f eine periodische Funktion mit der Periode

[1] Es gibt periodische Funktionen, die sich nicht in einer Fourier-Reihe entwickeln lassen. Diese Funktionen sind aber relativ exotisch, so dass wir uns damit nicht weiter befassen.

$2 \cdot \pi$ und gilt

$$f(x) = \frac{1}{2} \cdot a_0 \; + \; \sum_{k=1}^{\infty} a_k \cdot \cos(k \cdot x) \; + \; \sum_{k=1}^{\infty} b_k \cdot \sin(k \cdot x), \tag{9.2}$$

so können wir den Koeffizienten a_0 dadurch gewinnen, dass wir die Funktion $f(x)$ in dem Intervall $[0, 2 \cdot \pi]$ integrieren. Wir erhalten dann

$$\int_0^{2\cdot\pi} f(x)\,dx \;=\; \frac{1}{2} \cdot a_0 \cdot \int_0^{2\cdot\pi} 1\,dx \;+\; \int_0^{2\cdot\pi} \sum_{k=1}^{\infty} a_k \cdot \cos(k \cdot x)\,dx \;+\; \int_0^{2\cdot\pi} \sum_{k=1}^{\infty} b_k \cdot \sin(k \cdot x)\,dx$$

Das erste Integral auf der rechten Seite können wir ausführen, die anderen Integrale vertauschen wir mit den unendlichen Reihen[2]. Das liefert

$$\int_0^{2\cdot\pi} f(x)\,dx \;=\; \frac{1}{2} \cdot a_0 \cdot 2 \cdot \pi \;+\; \sum_{k=1}^{\infty} a_k \cdot \int_0^{2\cdot\pi} \cos(k \cdot x)\,dx \;+\; \sum_{k=1}^{\infty} b_k \cdot \int_0^{2\cdot\pi} \sin(k \cdot x)\,dx \tag{9.3}$$

Nun gilt für alle $k \in \mathbb{N}$ mit $k \geq 1$

$$\begin{aligned} \int_0^{2\cdot\pi} \sin(k \cdot x)\,dx &= \frac{-1}{k} \cdot \cos(k \cdot x)\Big|_0^{2\cdot\pi} \\ &= \frac{-1}{k} \cdot \big(\cos(k \cdot 2 \cdot \pi) - \cos(0)\big) = \frac{-1}{k} \cdot (1 - 1) = 0, \end{aligned} \tag{9.4}$$

denn $\cos(k \cdot 2 \cdot \pi) = \cos(0) = 1$. Genauso sehen wir für alle $k \geq 1$

$$\int_0^{2\cdot\pi} \cos(k \cdot x)\,dx = \frac{1}{k} \cdot \sin(k \cdot x)\Big|_0^{2\cdot\pi} = \frac{1}{k} \cdot \big(\sin(k \cdot 2 \cdot \pi) - \sin(0)\big) = 0, \tag{9.5}$$

denn $\sin(k \cdot 2 \cdot \pi) = \sin(0) = 0$. Setzen wir die Gleichungen (9.5) und (9.4) in Gleichung (9.2) ein, so erhalten wir die Gleichung

$$\int_0^{2\cdot\pi} f(x)\,dx \;=\; \pi \cdot a_0 \quad \text{bzw.} \quad a_0 = \frac{1}{\pi} \int_0^{2\cdot\pi} f(x)\,dx \tag{9.6}$$

Damit haben wir den Koeffizienten a_0 bestimmt. Um die Koeffizienten a_k für $k \geq 0$ zu bestimmen, multplizieren wir die Gleichung (9.2) mit $\cos(n \cdot x)$, wobei $n \in \mathbb{N}$ mit $n \geq 1$ ist. Anschließend integrieren wir über das Intervall $[0, 2 \cdot \pi]$. Dann haben wir

$$\begin{aligned} \int_0^{2\cdot\pi} f(x) \cdot \cos(n \cdot x)\,dx &= \frac{1}{2} \cdot a_0 \int_0^{2\cdot\pi} \cos(n \cdot x)\,dx \\ &+ \int_0^{2\cdot\pi} \cos(n \cdot x) \cdot \sum_{k=1}^{\infty} a_k \cdot \cos(k \cdot x)\,dx \\ &+ \int_0^{2\cdot\pi} \cos(n \cdot x) \cdot \sum_{k=1}^{\infty} b_k \cdot \sin(k \cdot x)\,dx \end{aligned}$$

Vertauschen wir Integration und Summation, so erhalten wir

$$\begin{aligned} \int_0^{2\cdot\pi} f(x) \cdot \cos(n \cdot x)\,dx &= \frac{1}{2} \cdot a_0 \int_0^{2\cdot\pi} \cos(n \cdot x)\,dx \\ &+ \sum_{k=1}^{\infty} a_k \cdot \int_0^{2\cdot\pi} \cos(n \cdot x) \cdot \cos(k \cdot x)\,dx \\ &+ \sum_{k=1}^{\infty} b_k \cdot \int_0^{2\cdot\pi} \cos(n \cdot x) \cdot \sin(k \cdot x)\,dx \end{aligned} \tag{9.7}$$

[2] Eine genaue Analyse, wann diese Vertauschung zulässig ist, geht über den Rahmen dieser Vorlesung hinaus.

Wir berechnen als nächstes die Integrale, die in dieser Formel auftreten. Das erste Integral hat nach Gleichung (9.5) den Wert 0. Zur Berechnung der anderen Integrale definieren wir

$$I_{n,k} := \int_0^{2\cdot\pi} \cos(n\cdot x)\cdot\cos(k\cdot x)\,dx \quad \text{und} \quad J_{n,k} := \int_0^{2\cdot\pi} \sin(n\cdot x)\cdot\sin(k\cdot x)\,dx$$

Wir berechnen $I_{n,k}$ durch partielle Integration. Für $n \neq 0$ gilt

$$\begin{aligned} I_{n,k} &= \int_0^{2\cdot\pi} \cos(n\cdot x)\cdot\cos(k\cdot x)\,dx \\ &= \frac{1}{n}\cdot\sin(n\cdot x)\cdot\cos(k\cdot x)\Big|_0^{2\cdot\pi} + \frac{k}{n}\cdot\int_0^{2\cdot\pi}\sin(n\cdot x)\cdot\sin(k\cdot x)\,dx \\ &= \frac{k}{n}\cdot J_{n,k}. \end{aligned}$$

Damit haben wir $I_{n,k}$ auf $J_{n,k}$ zurück geführt. Jetzt berechnen wir $J_{n,k}$ durch partielle Integration. Für $n \neq 0$ gilt

$$\begin{aligned} J_{n,k} &= \int_0^{2\cdot\pi} \sin(n\cdot x)\cdot\sin(k\cdot x)\,dx \\ &= \frac{-1}{n}\cdot\cos(n\cdot x)\cdot\sin(k\cdot x)\Big|_0^{2\cdot\pi} + \frac{k}{n}\cdot\int_0^{2\cdot\pi}\cos(n\cdot x)\cdot\cos(k\cdot x)\,dx \\ &= \frac{k}{n}\cdot\ I_{n,k} \end{aligned}$$

Damit haben wir die Berechnung von $J_{n,k}$ auf $I_{n,k}$ zurück geführt. Insgesamt haben wir die Gleichungen

$$I_{n,k} = \frac{k}{n}\cdot J_{n,k} \quad \text{und} \quad J_{n,k} = \frac{k}{n}\cdot I_{n,k}$$

gefunden. Setzen wir die zweite Gleichung in die erste Gleichung ein, so folgt

$$I_{n,k} = \frac{k^2}{n^2}\cdot I_{n,k} \quad \text{also} \quad \left(1-\frac{k^2}{n^2}\right)\cdot I_{n,k} = 0.$$

Falls $k \neq n$ ist, folgt daraus sofort

$$I_{n,k} = 0 \quad \text{und} \quad J_{n,k} = 0 \quad \text{für } n \neq k.$$

In dem Fall $k = n$ hat die bisherige Rechnung uns nicht viel weiter gebracht. In diesem Fall wissen wir lediglich, dass $I_{n,n} = J_{n,n}$ gilt. Hier hilft uns eine Addition weiter:

$$\begin{aligned} 2\cdot I_{n,n} &= I_{n,n} + J_{n,n} \\ &= \int_0^{2\cdot\pi}\cos(n\cdot x)\cdot\cos(n\cdot x)\,dx \ + \ \int_0^{2\cdot\pi}\sin(n\cdot x)\cdot\sin(n\cdot x)\,dx \\ &= \int_0^{2\cdot\pi}\cos^2(n\cdot x)+\sin^2(n\cdot x)\,dx \\ &= \int_0^{2\cdot\pi} 1\,dx \\ &= 2\cdot\pi \end{aligned}$$

Teilen wir beide Seiten der Gleichung durch 2, so erhalten wir als Ergebnis

$$I_{n,n} = \int_0^{2\cdot\pi} \cos(n\cdot x)\cdot\cos(n\cdot x)\,dx \;=\; \pi.$$

Wegen $J_{n,n} = I_{n,n}$ gilt auch

$$J_{n,n} = \int_0^{2\cdot\pi} \sin(n\cdot x)\cdot\sin(n\cdot x)\,dx \;=\; \pi.$$

Insgesamt haben wir also

$$I_{n,k} = J_{n,k} = \pi\cdot\delta_{n,k},$$

wobei $\delta_{n,k}$ das früher definierte Kronecker-Delta bezeichnet. Um die Gleichung 9.7 nach den Koeffizienten a_k auflösen zu können, müssen wir noch die Integrale

$$H_{n,k} := \int_0^{2\cdot\pi} \cos(n\cdot x)\cdot\sin(k\cdot x)\,dx$$

berechnen. Wir könnten dieses Integral auf dieselbe Art berechnen, mit der wir oben die Integrale $\int_0^{2\cdot\pi}\cos(n\cdot x)\cdot\cos(k\cdot x)\,dx$ und $\int_0^{2\cdot\pi}\sin(n\cdot x)\cdot\sin(k\cdot x)\,dx$ berechnet haben. Es gibt aber noch einen anderen Weg, den wir jetzt aufzeigen. Aus dem Additions-Theorem für die Sinus-Funktion

$$\sin(\alpha+\beta) = \sin(\alpha)\cdot\cos(\beta) + \cos(\alpha)\cdot\sin(\beta)$$

folgt sofort

$$\sin(\alpha)\cdot\cos(\beta) = \frac{1}{2}\cdot\sin(\alpha+\beta) + \frac{1}{2}\cdot\sin(\alpha-\beta).$$

Damit gilt

$$\begin{aligned}
H_{n,k} &= \int_0^{2\cdot\pi} \cos(n\cdot x)\cdot\sin(k\cdot x)\,dx \\
&= \frac{1}{2}\cdot\int_0^{2\cdot\pi} \sin\big((k+n)\cdot x\big)\,dx \;+\; \frac{1}{2}\cdot\int_0^{2\cdot\pi} \sin\big((k-n)\cdot x\big)\,dx \\
&= 0
\end{aligned}$$

nach Gleichung (9.4). Für $n > 0$ schreibt sich damit die Formel 9.7 wie folgt

$$\begin{aligned}
\int_0^{2\cdot\pi} f(x)\cdot\cos(n\cdot x)\,dx &= \frac{1}{2}\cdot a_0\int_0^{2\cdot\pi}\cos(n\cdot x)\,dx \\
&+ \sum_{k=1}^{\infty} a_k\cdot\int_0^{2\cdot\pi}\cos(n\cdot x)\cdot\cos(k\cdot x)\,dx \\
&+ \sum_{k=1}^{\infty} b_k\cdot\int_0^{2\cdot\pi}\cos(n\cdot x)\cdot\sin(k\cdot x)\,dx \\
&= 0 + \sum_{k=1}^{\infty} a_k\cdot I_{n,k} + \sum_{k=1}^{\infty} b_k\cdot H_{n,k} \\
&= \sum_{k=1}^{\infty} a_k\cdot\pi\cdot\delta_{n,k} + \sum_{k=1}^{\infty} b_k\cdot 0 \\
&= a_n\cdot\pi
\end{aligned}$$

Damit haben wir für den Fourier-Koeffizienten a_n die Formel

$$a_n = \frac{1}{\pi}\cdot\int_0^{2\cdot\pi} f(x)\cdot\cos(n\cdot x)\,dx \tag{9.8}$$

gefunden. Vergleichen wir diese Formel mit der Formel 9.6, so sehen wir, dass diese Gleichung auch für $n = 0$ richtig ist. Um die Koeffizienten b_n zu berechnen, multplizieren wir die Gleichung 9.2 mit $\sin(n{\cdot}x)$ und integrieren über das Intervall $[0, 2{\cdot}\pi]$. Dann erhalten wir nach einer Rechnung, die ganz analog zur Berechnung der Koeffizienten a_k verläuft, das Ergebnis

$$b_n = \frac{1}{\pi} \int_0^{2\cdot\pi} f(x) \cdot \sin(n{\cdot}x)\, dx. \tag{9.9}$$

9.2 Konvergenz

Wir müssen noch die Frage beantworten, für welche Funktionen f die mit Hilfe der Gleichungen (9.8) (9.9) und (9.2) aufgestellte Fourier-Reihe gegen f konvergiert. Wir wollen uns mit einem Satz begnügen, der im wesentlichen auf Dirichlet (Johann Peter Gustav Lejeune Dirichlet; 1805 - 1859) zurück geht. Zuvor benötigen wir noch zwei Definitionen.

Definition 97 (Einschränkung einer Funktion) Ist $f : \mathbb{R} \to \mathbb{R}$ eine Funktion und ist $[a, b]$ ein nicht-leeres Intervall, so definieren wir die *Einschränkung* von f auf $[a, b]$ als die Funktion

$$f\restriction_{[a,b]}\colon [a, b] \to \mathbb{R} \quad \text{mit } f\restriction_{[a,b]}(x) = f(x) \text{ für alle } x \in [a, b].$$

Definition 98 (stetig differenzierbar) Eine Funktion $f : [a, b] \to \mathbb{R}$ ist *stetig differenzierbar* falls f differenzierbar ist und außerdem die Ableitung $f' : [a, b] \to \mathbb{R}$ stetig ist.

Definition 99 (stückweise stetig differenzierbar) Eine Funktion $f : [0, 2\pi] \to \mathbb{R}$ ist *stückweise stetig differenzierbar* falls es Zahlen x_0, x_1, $\cdots$, x_n gibt mit

$$0 = x_0 \le x_1 \le x_2 \cdots \le x_{n-1} \le x_n = 2{\cdot}\pi$$

gibt, so dass für alle $i = 1, \cdots, n$ gilt:

$f\restriction_{[x_{i-1},x_i]}$ ist stetig differenzierbar.

Ein Beispiel für eine stückweise stetige Funktion sehen Sie in Abbildung 9.1. Die Ableitung dieser Funktion weist in den Punkten 0, π und $-\pi$ Sprünge auf.

Satz 100 Es gelte

(1) $f : \mathbb{R} \to \mathbb{R}$ ist stetig mit der Periode $2 \cdot \pi$,

(2) $f\restriction_{[0,2\pi]}$ ist stückweise stetig differenzierbar,

(3) $a_k = \frac{1}{\pi} \cdot \int_0^{2\cdot\pi} f(x) \cdot \cos(k{\cdot}x)\, dx$ und $b_k = \frac{1}{\pi} \cdot \int_0^{2\cdot\pi} f(x) \cdot \sin(k{\cdot}x)\, dx$,

dann gilt

$$f(x) = \frac{1}{2} \cdot a_0 + \sum_{k=1}^{\infty} a_k \cdot \cos(k{\cdot}x) + \sum_{k=1}^{\infty} b_k \cdot \sin(k{\cdot}x).$$

Der Beweis dieses Satzes benötigt Hilfsmittel, die im Rahmen der Vorlesung nicht eingeführt werden können.

9.3 Beispiele

Um es gleich bei der Berechnung der Fourier-Koeffizienten einfacher zu haben, definieren wir für eine Funktion $f : \mathbb{R} \to \mathbb{R}$ die Begriffe *gerade* und *ungerade*:

(1) f ist *gerade* g.d.w. $\forall x \in \mathbb{R} : f(-x) = f(x)$.

(2) f ist *ungerade* g.d.w. $\forall x \in \mathbb{R} : f(-x) = -f(x)$.

Ist die Funktion $f : \mathbb{R} \to \mathbb{R}$ ungerade und ist f integrierbar, so gilt für beliebige Zahlen a

$$\int_{-a}^{a} f(x)\,dx = 0. \tag{9.10}$$

Beweis: Es gilt

$$\int_{-a}^{a} f(x)\,dx = \int_{-a}^{0} f(x)\,dx + \int_{0}^{a} f(x)\,dx.$$

In dem Integral über das Intervall $[-a, 0]$ führen wir die Variablen-Transformation $y = -x$ durch. Dann gilt $dy = -dx$, und $y(-a) = a$, $y(0) = 0$. Damit gilt

$$\int_{-a}^{0} f(x)\,dx = -\int_{a}^{0} f(-y)\,dy = \int_{0}^{a} f(-y)\,dy = -\int_{0}^{a} f(y)\,dy = -\int_{0}^{a} f(x)\,dx$$

Also haben wir insgesamt

$$\int_{-a}^{a} f(x)\,dx = \int_{-a}^{0} f(x)\,dx + \int_{0}^{a} f(x)\,dx = -\int_{0}^{a} f(x)\,dx + \int_{0}^{a} f(x)\,dx = 0. \qquad \square$$

Ist die Funktion $f : \mathbb{R} \to \mathbb{R}$ gerade, so kann ein Integral über ein zum Punkt $x = 0$ symmetrisches Intervall wie folgt vereinfacht werden:

$$\int_{-a}^{a} f(x)\,dx = 2 \cdot \int_{0}^{a} f(x)\,dx. \tag{9.11}$$

Aufgabe 58: Beweisen Sie die Gleichung 9.11.

Satz 101 Ist die Funktion $f : \mathbb{R} \to \mathbb{R}$ periodisch mit der Periode $2 \cdot \pi$, so gilt

$$\int_{0}^{2\cdot\pi} f(x)\,dx = \int_{-\pi}^{\pi} f(x)\,dx \tag{9.12}$$

Aufgabe 59: Beweisen Sie den letzten Satz.

Die letzten beiden Gleichungen können wir zusammenfassen.

Korollar 102 Ist die Funktion $f : \mathbb{R} \to \mathbb{R}$ einerseits periodisch mit der Periode $2 \cdot \pi$ und andererseits ungerade, so gilt

$$\int_{0}^{2\cdot\pi} f(x)\,dx = 0 \tag{9.13}$$

Beweis: Es gilt

$$\int_{0}^{2\cdot\pi} f(x)\,dx = \int_{-\pi}^{\pi} f(x)\,dx = 0.$$

Korollar 103 Ist die Funktion $f : \mathbb{R} \to \mathbb{R}$ einerseits periodisch mit der Periode $2 \cdot \pi$ und andererseits gerade, so gilt

$$\int_{0}^{2\cdot\pi} f(x)\,dx = 2 \cdot \int_{0}^{\pi} f(x)\,dx \tag{9.14}$$

Beweis: Es gilt

$$\int_0^{2\cdot\pi} f(x)\,dx = \int_{-\pi}^{\pi} f(x)\,dx = 2\cdot\int_0^{\pi} f(x)\,dx.$$

9.3.1 Fourier-Analyse der Sägezahn-Funktion

Wir berechnen als erstes die Fourier-Reihe für die *Sägezahn*-Funktion $s : \mathbb{R} \to \mathbb{R}$, die im Intervall $[0, 2\cdot\pi]$ wie folgt definiert ist:

$$s(x) = \begin{cases} x & \text{falls } x \leq \pi, \\ 2\cdot\pi - x & \text{falls } x \geq \pi. \end{cases}$$

Diese Funktion wird periodisch auf ganz $\mathbb{R}$ fortgesetzt. Abbildung 9.1 zeigt diese Funktion. Wir berechnen nun die Fourier-Koeffizienten dieser Funktion.

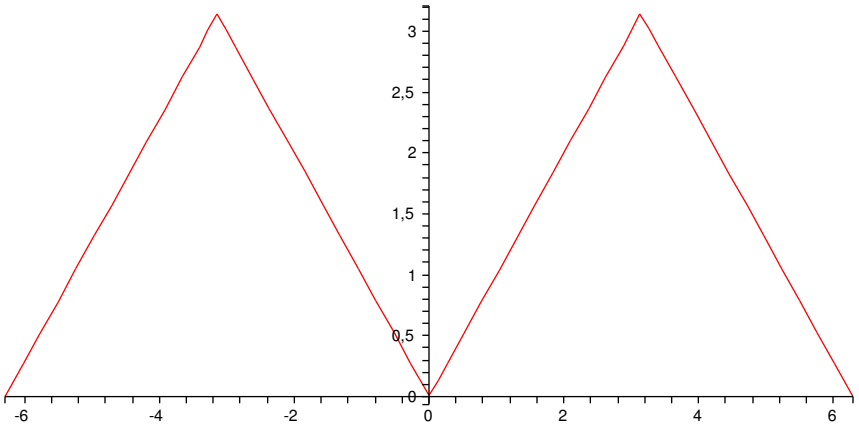

Abbildung 9.1: Die Sägezahn-Funktion.

(1) Die Koeffizienten a_n ergeben sich als

$$a_n = \frac{1}{\pi}\cdot\int_0^{2\cdot\pi} s(x)\cdot\cos(n{\cdot}x)\,dx = \frac{2}{\pi}\cdot\int_0^{\pi} s(x)\cdot\cos(n{\cdot}x)\,dx = \frac{2}{\pi}\cdot\int_0^{\pi} x\cdot\cos(n{\cdot}x)\,dx,$$

denn die Funktion $x \mapsto s(x)\cdot\cos(n{\cdot}x)\,dx$ ist einerseits periodisch mit der Periode $2\cdot\pi$ und andererseits gerade. Im Falle $n = 0$ haben wir

$$a_0 = \frac{2}{\pi}\cdot\int_0^{\pi} x\cdot\cos(0{\cdot}x)\,dx = \frac{2}{\pi}\cdot\int_0^{\pi} x\,dx = \frac{2}{\pi}\cdot\frac{x^2}{2}\Big|_0^{\pi} = \frac{2}{\pi}\cdot\frac{\pi^2}{2} = \pi.$$

Andernfalls berechnen wir das Integral mit Hilfe partieller Integration. Wir setzen $v'(x) = \cos(n{\cdot}x)$ und $u(x) = x$. Dann gilt $v(x) = \frac{1}{n}\cdot\sin(n{\cdot}x)$ und $u'(x) = 1$, also haben wir für $n \geq 1$

$$\begin{aligned} a_n &= \frac{2}{\pi}\cdot\left(x\cdot\frac{1}{n}\sin(n{\cdot}x)\Big|_0^{\pi} - \frac{1}{n}\cdot\int_0^{\pi}\sin(n{\cdot}x)\,dx\right) \\ &= \frac{2}{\pi}\cdot\left(0 - \frac{1}{n^2}\cdot\cos(n{\cdot}x)\Big|_0^{\pi}\right) \\ &= \frac{2}{\pi}\cdot\frac{1}{n^2}\cdot\Big((-1)^n - 1\Big), \end{aligned}$$

denn $\cos(n \cdot \pi) = (-1)^n$. Damit haben wir insgesamt

$$a_0 = \pi, \quad a_{2\cdot n+1} = \frac{-4}{\pi \cdot n^2} \quad \text{und} \quad a_{2\cdot(n+1)} = 0 \quad \text{für } n \in \mathbb{N}.$$

(2) Die Koeffizienten b_n ergeben sich als

$$b_n = \frac{1}{\pi} \cdot \int_0^{2\cdot\pi} s(x) \cdot \sin(n \cdot x)\, dx = 0$$

denn die Funktion $s(x) \cdot \sin(n x)$ ist sowohl periodisch mit der Periode $2 \cdot \pi$ als auch ungerade.

Insgesamt haben wir jetzt die Formel

$$s(x) = \frac{\pi}{2} - \frac{4}{\pi} \cdot \sum_{k=0}^{\infty} \frac{1}{(2\cdot k+1)^2} \cdot \cos\big((2\cdot k+1) \cdot x\big)$$

gefunden. Setzen wir hier für x den Wert π ein, so ergibt sich wegen $\cos\big((2\cdot k+1) \cdot \pi\big) = -1$

$$\begin{aligned}
\pi &= \frac{\pi}{2} - \frac{4}{\pi} \cdot \sum_{k=0}^{\infty} \frac{1}{(2\cdot k+1)^2} \cdot \cos\big((2\cdot k+1) \cdot \pi\big) \\
\Leftrightarrow \quad \frac{\pi}{2} &= \frac{4}{\pi} \cdot \sum_{k=0}^{\infty} \frac{1}{(2\cdot k+1)^2} \\
\Leftrightarrow \quad \frac{\pi^2}{8} &= \sum_{k=0}^{\infty} \frac{1}{(2\cdot k+1)^2}.
\end{aligned}$$

Damit sind wir jetzt in der Lage, die Reihe

$$\sigma := \sum_{n=1}^{\infty} \frac{1}{n^2}$$

zu berechnen. Zunächst zerlegen wir die Reihe in einen Teil, der nur über die geraden Indices läuft und einen Teil, der über die ungeraden Indices läuft:

$$\begin{aligned}
\sigma &= \sum_{n=1}^{\infty} \frac{1}{(2\cdot n)^2} + \sum_{n=0}^{\infty} \frac{1}{(2\cdot n+1)^2} \\
\Leftrightarrow \quad \sigma &= \frac{1}{4} \cdot \sum_{n=1}^{\infty} \frac{1}{n^2} + \frac{\pi^2}{8} \\
\Leftrightarrow \quad \sigma &= \frac{1}{4} \cdot \sigma + \frac{\pi^2}{8} \\
\Leftrightarrow \quad \frac{3}{4} \cdot \sigma &= \frac{\pi^2}{8} \\
\Leftrightarrow \quad \sigma &= \frac{\pi^2}{6}
\end{aligned}$$

Damit haben wir also die Formel

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

gezeigt. Die Frage nach dem Wert dieser Reihe war 1644 als *Basel'sches Problem* von Pietro Mengoli gestellt worden. Führende Mathematiker des 17-ten Jahrhunderts hatten sich erfolglos mit dieser Frage beschäftigt. Im Jahre 1735 gelang es Leonard Euler (1707 - 1783), dieses Problem zu lösen.

Aufgabe 60: Die Funktion p sei auf dem Intervall $[-\pi, \pi]$ definiert durch

$$p(x) = x^2.$$

Die Funktion werde so auf $\mathbb{R}$ fortgesetzt, dass die resultierende Funktion die Periode $2 \cdot \pi$ hat.

(1) Berechnen Sie die Fourier-Reihe von p.

(2) Berechnen Sie mit Hilfe der Fourier-Reihe von p einen Wert für die Reihe

$$\sum_{n=1}^{\infty} \frac{1}{n^2}.$$

Kapitel 10

Rundungsfehler*

Die meisten komplexen Probleme lassen sich nur mit numerischen Verfahren lösen. Wir haben bereits verschiedene numerische Verfahren kennengelernt, beispielsweise Verfahren zur Berechnung von Nullstellen sowie Verfahren zur numerischen Integration. Allen diesen Verfahren ist gemeinsam, dass zwei verschiedene Arten von Fehlern auftreten:

(1) Ein *Approximations-Fehler* tritt auf, wenn wir einen Wert λ berechnen wollen, zu dessen Berechnung wir nur eine Näherungsformel existiert. Oft ist der Approximations-Fehler ein *Abbruch-Fehler*, der seine Ursache darin hat, dass wir nur endlich viele Glieder einer unendlichen Reihe berechnen können. Wollen wir beispielsweise die Euler'sche Zahl e nach der Formel

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}$$

berechnen, so können wir auf dem Rechner diese Summe nicht gegen unendlich laufen lassen, sondern müssen die Summe nach endlich vielen Gliedern abbrechen, wir berechnen also als Approximation für e eine Reihe der Form

$$e_n := \sum_{k=0}^{n} \frac{1}{k!}$$

und müssen dann n so groß wählen, dass der Abbruch-Fehler

$$e - e_n = \sum_{k=n+1}^{\infty} \frac{1}{k!}$$

unterhalb einer vorgegeben Schranke bleibt.

(2) Zusätzlich zum Approximations-Fehler gibt es noch die Rundungsfehler, die im Laufe der Rechnung entstehen. Diese Rundungsfehler haben Ihre Ursache darin, dass Fließkomma-Zahlen auf dem Rechner mit einer vorgegebenen Genauigkeit dargestellt werden. Rechnen wir in der Sprache *Java* mit einer Fließkomma-Zahl vom Typ `float`, so stehen zur Darstellung der Stellen hinter dem Komma lediglich 23 Bits zur Verfügung. Werden nun zwei solche Zahlen multipliziert, so könnten bis zu 47 Bits notwendig sein, um alle Stellen hinter dem Komma korrekt wiedergeben zu können. Da aber zum Abspeichern des Ergebnisses lediglich 23 Bits zur Verfügung stehen, um die Ziffern hinter dem Komma abzuspeichern, bleibt nichts anderes übrig, als das Ergebnis auf 23 Bits zu runden. Der dadurch entstehende Fehler wird als Rundungsfehler bezeichnet.

Die Auswirkungen von Rundungsfehlern werden oft unterschätzt. Unter

`http://www.devtopics.com/20-famous-software-disasters-part-2/`

findet sich eine Liste der 20 spektakulärsten Software-Fehler, die Katastrophen ausgelöst haben. In mehreren Fällen waren Rundungsfehler ein Teil des Problems. Um einen ersten Eindruck von

der Wirkung von Rundungsfehlern zu bekommen, betrachten wir das in Abbildung 10.1 gezeigte Programm zur Berechnung der Reihe

$$\sum_{n=1}^{\infty} \frac{1}{n}.$$

```
1  harmonic := procedure() {
2      oldSum := 0.0;
3      sum := 1.0;
4      n := 1;
5      while (oldSum < sum) {
6          oldSum := sum;
7          n += 1;
8          sum += 1/n;
9      }
10     print("sum = $sum$, n = $n$");
11 };
12 harmonic();
```

Abbildung 10.1: Berechnung von $\sum_{n=1}^{\infty} \frac{1}{n}$.

Sie erwarten jetzt vielleicht, dass diese Programm nie terminiert, aber wenn wir dieses Programm in einer Datei mit dem Namen "`harmonic.stlx`" speichern und es dann mit dem Befehl

```
setlX --real32 harmonic.stlx
```

starten, dann erhalten wir nach wenigen Sekunden die Meldung:

```
sum = 13.05426, n = 200001.
```

Da wir früher bewiesen haben, dass die Partialsummen $\sum_{k=1}^{n} \frac{1}{k}$ für wachsende Werte von n beliebig groß werden, fragen wir uns, was bei der Rechnung schief gelaufen ist. Die Antwort ist, dass für $n = 200001$ der Wert $\frac{1}{n}$ so klein ist, dass die Summe

$$13.05426 + \tfrac{1}{n}$$

so nahe bei 13.05426 liegt, dass sie auf den Wert 13.05426 abgerundet wird. Um diesen Effekt näher zu beschreiben, definiert man für einen vorgegebenen Rechner die sogenannte *Maschinen-Konstante* *eps* als die kleinste positive Zahl, die, wenn sie auf diesem Rechner zu 1 addiert wird, eine Ergebnis größer als 1 ergibt. Die formale Definition lautet

$$\mathit{eps} := \min\bigl(\{x \in \mathbb{R} \mid x > 0 \wedge 1 \oplus x > 1\}\bigr).$$

Hier bezeichnet $\oplus$ die auf dem Rechner implementierte Addition. Abbildung 10.2 zeigt ein einfaches Programm zur Berechnung der Maschinen-Konstante. Bei der im IEEE-Standard 754 definierten 32-Bit-Architektur erhalten wir für *eps* den Wert

$$\mathit{eps}_{32} = 9.536745 \cdot 10^{-7},$$

bei einer 64-Bit-Architektur lautet das Ergebnis

$$\mathit{eps}_{64} = 8.88178419700125 \cdot 10^{-16},$$

und wenn wir mit 128-Bit rechnen, haben wir

$$\mathit{eps}_{128} = 7.703719777548943412223911770339695 \cdot 10^{-34}.$$

Bei modernen Rechnern, die den IEEE-Standard 754 implementieren, können wir davon ausgehen, dass der relative Rundungsfehler bei der Ausführung einer Grundrechenoperation durch die

```
1  maschinenKonstante := procedure() {
2      eps := 1.0;
3      old := eps;
4      while (1.0 + eps > 1.0) {
5          old := eps;
6          eps /= 2;
7      }
8      return old;
9  };
```

Abbildung 10.2: Berechnung der Maschinen-Konstante *eps*.

Maschinen-Konstante *eps* beschränkt ist.

Leider muss die Vorlesung aus Zeitgründen an dieser Stelle enden. Der interessierten Leser sei daher auf die Literatur, insbesondere den Artikel von Goldberg [10] verwiesen.

Literaturverzeichnis

[1] Forster, Otto: *Analysis I, Differential- und Integralrechnung einer Veränderlichen.* Vieweg & Teubner, 11te Auflage, 2011.

[2] Grauert, Hans und Ingo Lieb: *Differential- und Integralrechnung I.* Springer, 1967.

[3] Courant, Richard: *Differential and Integral Calculus*, volume 1. Blackie & Son Limited, 2nd edition, 1937.

[4] Wrede, Robert and Murray Spiegel: *Advanced Calculus.* The McGraw-Hill Companies, 3rd edition, 2010.

[5] Landau, Edmund: *Grundlagen der Analysis.* Akademische Verlagsgesellschaft, 1930. http://www.scribd.com/doc/2452802/Landau-Edmund-Grundlagen-der-Analysis.

[6] Rudin, Walter: *Principles of Mathematical Analysis.* McGraw-Hill International, 3rd edition, 1976.

[7] Dedekind, Richard: *Stetigkeit und irrationale Zahlen.* Friedrich Vieweg und Sohn, 1872.

[8] Dowell, M. and P. Jarrat: *A modified regula falsi method for computing the root of an equation.* BIT Numerical Mathematics, 11(2):168–174, 1971.

[9] Rosenlicht, Maxwell: *Integration in finite terms.* The American Mathematical Monthly, 79(9):963–972, November 1972.

[10] Goldberg, David: *What every computer scientist should know about floating-point arithmetic.* ACM Computing Surveys, 23(1):5–48, 1991. Available from `http://www.validlab.com/goldberg/paper.ps`.