Steffen Goebbels Stefan Ritter

Zusatzmaterial zum Buch „Mathematik verstehen und anwenden – von den Grundlagen bis zu Fourier-Reihen und Laplace-Transformation“

Spektrum Akademischer Verlag Heidelberg 2011

Stand 18.02.2013

Vorwort

Auf den folgenden Seiten finden Sie Zusatzmaterial zu einigen Kapiteln des Buchs. Teils handelt es sich um ergänzende Beispiele und Erklärungen, die hier auf auf Grund der Rückmeldungen einiger Leser aufgenommen wurden. Zum anderen Teil werden weitere Begriffe erläutert, die nicht in den Kontext des Lehrbuchs passten. Dazu gehören Beispiele partieller Differenzialgleichungen, die mit den Mitteln des Buchs zugänglich sind, sowie eine kurze Darstellung der χ^2-Verteilung zum Test, ob eine angenommene Wahrscheinlichkeitsverteilung tatsächlich vorliegt. Neben vielen weiteren Themen wird kurz auf die Z-Transformation eingegangen und die Rolle von Fensterfunktionen bei der Fouriertransformation erläutert.

Den Abschluss bilden einige Programme, mit denen Grafiken des Buchs erstellt und Berechnungen durchgeführt wurden.

Inhaltsverzeichnis

1 Zusatzmaterial zu Kapitel 1

Übersicht

1.1 Aussagenlogik

Die aussagenlogischen Formeln

$$\begin{aligned} S &= (\neg A \wedge \neg B \wedge C_{\text{in}}) \vee (\neg A \wedge B \wedge \neg C_{\text{in}}) \vee (A \wedge \neg B \wedge \neg C_{\text{in}}) \vee (A \wedge B \wedge C_{\text{in}}), \\ C_{\text{out}} &= (\neg A \wedge B \wedge C_{\text{in}}) \vee (A \wedge \neg B \wedge C_{\text{in}}) \vee (A \wedge B \wedge \neg C_{\text{in}}) \vee (A \wedge B \wedge C_{\text{in}}) \\ &= (A \wedge B) \vee (B \wedge C_{\text{in}}) \vee (A \wedge C_{\text{in}}) \end{aligned}$$

für die Summe und den Übertrag im Beispiel des Volladdierers sind in **disjunktiver Normalform**. Dabei werden die Klammerterme oder-verknüpft (also mit Disjunktionen verbunden). Innerhalb jedes Klammerterms gibt es nur aussagenlogische Variablen, die entweder negiert oder nicht-negiert vorkommen und und-verknüpft sind. Jede Formel lässt sich wie im Beispiel beschrieben durch Ablesen der Wertetabelle (siehe Tabelle 1.1) in eine disjunktive Normalform bringen. Ähnlich kann man jede Formel als **konjunktive Normalform** schreiben. Beim Addierwerk ergeben sich die konjunktiven Normalformen

$$\begin{aligned} S &= \neg(\neg A \wedge \neg B \wedge \neg C_{\text{in}}) \wedge \neg(\neg A \wedge B \wedge C_{\text{in}}) \wedge \neg(A \wedge \neg B \wedge C_{\text{in}}) \wedge \neg(A \wedge B \wedge \neg C_{\text{in}}) \\ &= (A \vee B \vee C_{\text{in}}) \wedge (A \vee \neg B \vee \neg C_{\text{in}}) \wedge (\neg A \vee B \vee \neg C_{\text{in}}) \wedge (\neg A \vee \neg B \vee C_{\text{in}}), \end{aligned}$$

Tab. 1.1: Wertetabelle eines Volladdierers

A	0	0	0	0	1	1	1	1
B	0	0	1	1	0	0	1	1
C_{in}	0	1	0	1	0	1	0	1
S	0	1	1	0	1	0	0	1
C_{out}	0	0	0	1	0	1	1	1

$$C_{\text{out}} = (A \vee B \vee C_{\text{in}}) \wedge (A \vee B \vee \neg C_{\text{in}}) \wedge (A \vee \neg B \vee C_{\text{in}}) \wedge (\neg A \vee B \vee C_{\text{in}}).$$

Hier haben wir in der Wertetabelle die Variablenwerte gesucht, die eine Null liefern sollen und haben wie bei der disjunktiven Normalform dazu Klammerterme aus und-verknüpften negierten und nicht-negierten Variablen erstellt. Negieren wir nun diese Klammerterme, so liefern die De Morgan'schen Regeln Terme mit Oder-Verknüpfungen. Jeder dieser Terme generiert die Null, zu der er erstellt wurde. Für alle anderen Variablenwerte liefert er eine Eins. Verbinden wir die so gewonnenen Terme mit Konjunktionen, so werden alle gewünschten Nullen (und keine weiteren) erzeugt.

Diese Normalformen eignen sich, um systematisch Formeln zu vereinfachen. Die konjunktive Normalform ist der Ausgangspunkt des Resolutionskalküls. Das ist ein Verfahren zur Prüfung auf Unerfüllbarkeit. Eine möglichst kurze disjunktive Normalform erhält man mittels **Karnaugh-Veitch-Diagrammen**. Bei einem solchen Diagramm wird die Wertetabelle geschickt aufgeschrieben, damit man Terme ablesen kann, die möglichst große Rechtecke von Einsen generieren. Für den Übertrag im Beispiel ergibt sich aus der Wertetabelle:

	$B = \neg C_{\text{in}} = 1$	$B = C_{\text{in}} = 1$	$\neg B = C_{\text{in}} = 1$	$\neg B = \neg C_{\text{in}} = 1$
$A = 1$	**1**	**<u>1</u>**	1	0
$\neg A = 1$	0	<u>1</u>	0	0

Die Variablen benachbarter Spalten (und Zeilen) unterscheiden sich durch genau eine Negation. Das gilt auch für die erste und letzte Spalte (Zeile). Für das Rechteck aus den fett gedruckten Werten und das aus den unterstrichenen Werten lesen wir die Formeln $A \wedge B$ sowie $B \wedge C_{\text{in}}$ ab. Die noch fehlende Eins ergibt sich über den Block $A \wedge C_{\text{in}}$, und wir erhalten insgesamt wie zuvor: $C_{\text{out}} = (A \wedge B) \vee (A \wedge C_{\text{in}}) \vee (B \wedge C_{\text{in}})$. Allgemein sucht man nach besonders großen Rechtecken, die eine oder mehrere Zeilen und Spalten umfassen und auch Ränder überschreiten dürfen.

Hintergrund: Resolutionskalkül

Eine aussagenlogische Formel ist **erfüllbar**, wenn es eine **Belegung** ihrer Variablen mit Wahrheitswerten gibt, so dass die Formel wahr wird. Sie ist **unerfüllbar**, wenn es keine einzige Belegung der Variablen gibt, für die die Formel wahr wird.

Eine wichtige Fragestellung ist, ob eine gegebene Formel unerfüllbar ist. In der Aussagenlogik ist der Test auf Unerfüllbarkeit aber sehr einfach, wenn auch ineffizient: Probiere alle Variablenbelegungen (wahr/falsch) aus. Bei n Variablen gibt es damit 2^n Versuche im schlechtesten Fall. Leider klappt das Ausprobieren bei Formeln der Prädikatenlogik nicht mehr. Hier gibt es Variablen, die irgendwelche Werte aus einem „Universum“ annehmen dürfen. Bei der Überprüfung der Unerfüllbarkeit müsste man auch hier alle (eventuell unendlich vielen) Variablenwerte durchprobieren, was aber nicht geht. Daher benutzt man den Resolutionskalkül, der auch Grundlage der Programmiersprache Prolog ist. Wir erklären nun die Funktionsweise des Resolutionskalküls anhand der Aussagenlogik, da er hier genau so funktioniert wie in der Prädikatenlogik aber viel einfacher zu verstehen ist.

Seien $A_1, A_2, A_3, \dots$ aussagenlogische Variablen. Eine endliche Menge $K \subset \{A_1, A_2, A_3, \dots\} \cup \{\neg A_1, \neg A_2, \neg A_3, \dots\}$ heißt eine **Klausel**.

Die Konjunktionsglieder einer konjunktiven Normalform können über Klauseln dargestellt werden.

Beispiel 1.1
$(A \vee B) \wedge (\neg A \vee C)$ führt zu den Klauseln $\{A, B\}$ und $\{\neg A, C\}$. ■

Die Menge $\mathbb{K}$ der Klauseln K zu den Konjunktionsgliedern einer aussagenlogischen Formel heißt die zugehörige **Klauselmenge**.

Beispiel 1.2
$(A \vee B) \wedge (\neg A \vee C)$ führt zur Klauselmenge $\{\{A, B\}, \{\neg A, C\}\}$. ■

Seien K_1, K_2 und R Klauseln. R heißt **Resolvente** von K_1 und K_2, falls es eine aussagenlogische Variable A gibt mit $A \in K_1$ und $\neg A \in K_2$ und $R = (K_1 \setminus \{A\}) \cup (K_2 \setminus \{\neg A\})$ (oder $A \in K_2$ und $\neg A \in K_1$ und $R = (K_1 \setminus \{\neg A\}) \cup (K_2 \setminus \{A\})$).

Sei $\mathbb{K}$ eine Klauselmenge. Wir bezeichnen die Menge aller Resolventen von Klauseln aus $\mathbb{K}$ mit $\mathrm{Res}(\mathbb{K}) := \mathbb{K} \cup \{R : R \text{ ist Resolvente zweier Klauseln von } \mathbb{K}\}$. Die Ergebnisse der Resolution können für weitere Resolutionen herangezogen werden. Damit erhalten wir die Mengen

$$\mathrm{Res}^0(\mathbb{K}) := \mathbb{K}, \quad \mathrm{Res}^{n+1}(\mathbb{K}) := \mathrm{Res}\left(\mathrm{Res}^n(\mathbb{K})\right),$$

$$\mathrm{Res}^*(\mathbb{K}) := \bigcup_{n=1}^{\infty} \mathrm{Res}^n(\mathbb{K}).$$

Beispiel 1.3
Resolution von $\{A, \neg B, C\}$ und $\{B, C, D\}$ ergibt $\{A, C, D\}$. ■

Der folgende Satz ist der Resolutionskalkül:
Eine Formel in konjunktiver Normalform mit Klauselmenge $\mathbb{K}$ ist genau dann unerfüllbar, falls $\emptyset \in \text{Res}^*(\mathbb{K})$.
Beweis

- Falls $\mathbb{K}$ unerfüllbar ist und weitere Klauseln hinzukommen, dann ist die neue Menge erst recht unerfüllbar. Denn die einzelnen Klauseln sind über $\wedge$ verknüpft. Damit ist also auch $\text{Res}(\mathbb{K})$ unerfüllbar.
- Falls $\mathbb{K}$ erfüllbar ist: Beide Eingangsklauseln eines Resolutionsschritts sind mit einer Variablenbelegung erfüllt. Damit ist auch die Resolvente erfüllt, z.B.: Ist $(A \vee B \vee \neg C) \wedge (\neg A \vee \neg C \vee D)$ erfüllt, so ist
  - bei $A =$ wahr aufgrund der zweiten Klausel $\neg C \vee D$ wahr.
  - Bei $A =$ falsch ist aufgrund der ersten Klausel $B \vee \neg C$ wahr.

  Also ist $B \vee \neg C \vee D$ bzw. $\{B, \neg C, D\}$ wahr.

Damit haben wir gezeigt: Bei Anwendung von Res ändert sich die Erfüllbarkeit nicht.

Falls $\emptyset \in \text{Res}^k(\mathbb{K})$: In $\text{Res}^{k-1}(\mathbb{K})$ gibt es zwei Klauseln vom Typ $\{A\}$ und $\{\neg A\}$. $A \wedge \neg A$ ist aber unerfüllbar, also ist $\text{Res}^{k-1}(\mathbb{K})$ unerfüllbar und somit auch $\mathbb{K}$, da gemäß der vorangehenden Überlegung beide Mengen die gleiche Erfüllbarkeit haben.

Man beachte, dass $\emptyset \in \text{Res}^*(\mathbb{K}) \Longleftrightarrow \emptyset \in \text{Res}^k(\mathbb{K})$ für ein $k \in \mathbb{N}$.

Es bleibt zu zeigen: Falls $\mathbb{K}$ unerfüllbar ist, dann kann man durch fortgesetzte Resolution die leere Menge erzeugen. Wir zeigen dies mit einer Vollständigen Induktion über die Anzahl der Variablen n einer Klauselmenge $\mathbb{K}$.

- Induktionsanfang für $n = 1$:
  $\mathbb{K} = \{\{A\}, \{\neg A\}\}$ ist die einzige unerfüllbare Klauselmenge (falls die Variable A genannt wird), und $\emptyset$ entsteht bei der Resolution.
- Induktionsannahme: Fortgesetzte Resolution liefert für jede unerfüllbare Klauselmenge mit n Variablen $A_1, A_2, \ldots, A_n$ die leere Menge.
- Induktionsschritt: Wir zeigen, dass auch für jede unerfüllbare Klauselmenge $\mathbb{K}$ mit $n+1$ Variablen $A_1, A_2, \ldots, A_{n+1}$ die leere Menge resolviert werden kann.
  - $\mathbb{K}_0$ entstehe aus $\mathbb{K}$ durch Weglassen von A_{n+1} aus allen Klauseln und Weglassen aller Klauseln mit $\neg A_{n+1}$.
  - $\mathbb{K}_1$ entstehe aus $\mathbb{K}$ durch Weglassen von $\neg A_{n+1}$ aus allen Klauseln und Weglassen aller Klauseln mit A_{n+1}.

  Da $\mathbb{K}$ unerfüllbar ist, ist $\mathbb{K}$ insbesondere für die Variablen $A_1, \ldots, A_n$ unerfüllbar, wenn wir A_{n+1} durch wahr oder falsch ersetzen:
  - Wählen wir $A_{n+1} =$ wahr, ist $\mathbb{K}_1$ unerfüllbar.
  - Wählen wir $A_{n+1} =$ falsch, ist $\mathbb{K}_0$ unerfüllbar.

  In beiden Fällen kann man mittels Resolution $\emptyset$ erzeugen:

$$\emptyset \in \text{Res}^*(\mathbb{K}_0), \quad \emptyset \in \text{Res}^*(\mathbb{K}_1).$$

Fügt man A_{n+1} den Klauseln aus $\mathbb{K}_0$ und $\neg A_{n+1}$ den Klauseln aus $\mathbb{K}_1$ wieder hinzu, aus denen sie zuvor entfernt wurden, so liefert die Resolution für das erweiterte $\tilde{\mathbb{K}}_0 \subset \mathbb{K}$ eine Menge, in der $\emptyset$ oder $\{A_{n+1}\}$ enthalten ist, für das erweiterte $\tilde{\mathbb{K}}_1 \subset \mathbb{K}$ ist $\emptyset$ oder $\{\neg A_{n+1}\}$ enthalten. Falls $\emptyset \in \tilde{\mathbb{K}}_0 \cup \tilde{\mathbb{K}}_1$, ist $\emptyset \in \mathrm{Res}^*(\mathbb{K})$, und wir sind fertig. Sonst sind ist

$$\{\{A_{n+1}\}, \{\neg A_{n+1}\}\} \subset \mathrm{Res}^*(\tilde{\mathbb{K}_0}) \cup \mathrm{Res}^*(\tilde{\mathbb{K}_1}) \subset \mathrm{Res}^*(\mathbb{K}).$$

Die Resolvente von beiden Mengen ist $\emptyset$: $\emptyset \in \mathrm{Res}^*(\mathbb{K})$.

■

Beispiel 1.4
Wir zeigen mittels Resolution die Unerfüllbarkeit der Klauselmenge

$$\{A, \neg B, C\}, \{B, C\}, \{\neg A, C\}, \{B, \neg C\}, \{\neg B\}\},$$

also der Formel

$$(A \vee \neg B \vee C) \wedge (B \vee C) \wedge (\neg A \vee C) \wedge (B \vee \neg C) \wedge \neg B.$$

Resolution von $\{B, C\}$ und $\{B, \neg C\}$ liefert $\{B\}$.
Resolution dieser Menge mit $\{\neg B\}$ ergibt $\emptyset$ und damit die Unerfüllbarkeit. ■

1.2 Prädikatenlogik

Wir konkretisieren die Beschreibung aus Beispiel 1.16 von Seite 21 (Logikprogrammierung) und definieren einige Fakten:

```
vater(otto, hans).
vater(otto, ute).
vater(hans, karl).
mutter(ute, anna).
```

Damit sollen die wahren Aussagen ausgedrückt werden, dass Otto Vater von Hans und Ute, Hans Vater von Karl und Ute Mutter von Anna ist. Wir stellen jetzt Fragen an ein Prolog-System:

```
?- vater(otto, ute).    ?- vater(otto,karl).    ?- vater(hans,karl).
wahr                    falsch                  wahr
```

Da Prolog annimmt, seine Datenbank würde die ganze Welt komplett beschreiben, ist wahr, was gefunden wird und falsch, was nicht vorhanden ist. Prolog erlaubt Anfragen mit Variablen (Aussageformen). Dabei wird nach Werten für die Variablen gesucht, die aus der Anfrage eine wahre Aussage machen. Die Frage `?- vater(otto,X)` heißt: Für welche $X \in \mathbb{M} := \{\texttt{otto}, \texttt{hans}, \texttt{ute}, \texttt{karl}, \texttt{anna}\}$ lässt sich $\texttt{vater}(\texttt{otto}, \texttt{X})$ aus den Fakten folgern. Die Antwort ist

```
X = hans    X = ute
```

Variablen dürfen selbstverständlich auch bei der Definition von Fakten benutzt werden:

```
gleich(A, A).
```

Dahinter verbirgt sich die wahre Aussage: Für alle $A \in \mathbb{M}$ gilt $A = A$. Mit diesem Fakt kann geprüft werden, ob z. B. zwei Konstanten gleich sind:

```
?- gleich(franz, franz)    ?- gleich(franz, joerg)
wahr                       falsch
```

Wir erweitern nun die Datenbasis um Regeln:

```
elternteil(A,B) :- vater(A,B).
elternteil(A,B) :- mutter(A,B).
geschwister(A,B) :- elternteil(C,A) elternteil(C,B).
```

Diese Notation bedeutet:

$$\text{Für alle } A, B \in \mathbb{M} : \text{vater(A,B)} \vee \text{mutter(A,B)} \Longrightarrow \text{elternteil(A,B)},$$

$$\text{Für alle } A, B, C \in \mathbb{M} : \text{elternteil(C,A)} \wedge \text{elternteil(C,B)} \Longrightarrow \text{geschwister(A,B)}.$$

Die Anfrage `?- geschwister(hans, ute)` wird wahr, da die Aussage aus den Fakten mittels der Regeln abgeleitet werden kann:

$$\begin{aligned} &\texttt{vater(otto, hans)} \wedge \texttt{vater(otto,ute)} \\ \Longrightarrow\ &\texttt{vater(otto, hans)} \wedge \texttt{elternteil(otto, ute)} \\ \Longrightarrow\ &\texttt{elternteil(otto, hans)} \wedge \texttt{elternteil(otto, ute)} \\ \Longrightarrow\ &\texttt{geschwister(hans, ute)} \end{aligned}$$

Wird nach `?- geschwister(X, Y)` gefragt, so gibt das System die folgenden Lösungen an:

```
X=hans  X=hans  X=ute   X=ute   X=karl  X=karl  X=joerg X=joerg
Y=hans  Y=ute   Y=hans  Y=ute   Y=karl  Y=joerg Y=karl  Y=joerg
```

Das Verfahren, mit dem Prolog alle Lösungen findet, heißt Backtracking oder Tiefensuche. Das Prinzip des Backtracking lässt sich sehr anschaulich an einem Labyrinth erklären. Ein Spieler, der einen Weg aus dem Irrgarten finden soll, kann so vorgehen: Wähle zunächst eine Richtung aus und gehe bis zur nächsten Verzweigung. Wähle dort wieder eine Richtung aus und fahre so fort, bis eine Sackgasse erreicht ist. Nun kehre zur letzten Verzweigung zurück und gehe in eine hier bisher noch nicht probierte Richtung. Wird wieder eine Sackgasse erreicht, wiederholt sich das Verfahren. Sind an einer Gabelung bereits alle weiterführenden Wege ausprobiert, so wird zur vorangehenden Verzweigung zurückgekehrt. Gelangt der Spieler schließlich an den Ausgangspunkt zurück und sind dort bereits alle Richtungen probiert, so gibt es kein Entkommen (kein korrektes Labyrinth). Andernfalls wird irgendwann der Ausgang gefunden.

Weggabelungen sind beim Prolog-System Regeln und Fakten gleichen Namens. Prolog probiert diese von oben nach unten. Entsprechend verfährt es aber auch bei der Verifikation der Aussagen auf der rechten Seite einer Regel von links nach rechts, so dass nach und nach ein ganz bestimmter Lösungsweg eingeschlagen wird, der aber bei Misslingen durch Rückkehr zur vorangehenden Gabelung korrigiert wird.

Genau genommen entspricht die Ausführung eines Prolog-Programms der Durchführung des oben beschriebenen Resolutionsalgorithmus für einen Unerfüllbarkeitstest von $\neg($ Programm $\Longrightarrow$ Anfrage $)$, d. h. von Programm $\wedge\ \neg$Anfrage. Die Reihenfolge der Resolutionsschritte ist dabei durch die Tiefensuche vorgegeben.

Die Tiefensuche in Prolog ersetzt die Notwendigkeit, selbst Lösungsalgorithmen zu schreiben, also genau festzulegen, **wie** ein Computer eine Aufgabe lösen soll. Prolog ist damit eine deklarative Sprache ähnlich wie die Datenbanksprache SQL, bei der ein Problem exakt beschrieben aber kein Lösungsalgorithmus vorgegeben wird (**was**). In den meisten anderen Programmiersprachen müssen wir leider Lösungsalgorithmen schreiben.

1.3 Beweise

Beispiel 1.5 (Halteproblem)
Wir beweisen, dass es es kein Computerprogramm A gibt, das in endlicher Zeit entscheiden kann, ob ein weiteres Computerprogramm seinerseits nach endlicher Zeit hält, d. h. zu einem Ergebnis kommt. Dieses **Halteproblem** spielt in der Informatik eine große Rolle, da es zeigt, dass nicht alles programmierbar ist. Der Trick des hier geführten indirekten Beweises besteht darin, dass wir annehmen, dass es das Programm A gibt und wir damit ein Programm B konstruieren, das als Eingabe ebenfalls ein Programm erwartet. Wir wenden dann B auf sich selbst an, um einen Widerspruch

zu erhalten. Die Anwendung auf sich selbst ist ein ganz typisches Vorgehen in der theoretischen Informatik.
Das Programm B sei (ohne formale Programmiersprache) wie folgt aufgebaut:

- Wende A auf das Eingabeprogramm an.
- Falls A feststellt, dass das Eingabeprogramm nicht hält, dann endet die Ausführung von B.
- Falls A feststellt, dass das Eingabeprogramm hält, dann geht B in eine Endlosschleife, d. h. B führt eine Anweisung unendlich oft aus und endet nie.

Jetzt starten wir das Programm B mit sich selbst als Eingabe. Falls B nach endlicher Zeit hält, so kann B nicht in die Endlosschleife gehen, und A muss feststellen, dass B nicht hält – wir haben einen Widerspruch. Also kann B nicht nach endlicher Zeit halten. Da A wegen der Annahme nach endlicher Zeit zu einem Ergebnis kommt, geht das nur, wenn A feststellt, dass B nach endlicher Zeit fertig wird und man damit in die Endlosschleife gelangt. Damit haben wir auch in dieser Situation einen Widerspruch zum Ergebnis von A. Die Widersprüche zeigen, dass es, entgegen der Annahme, das Programm A nicht geben kann. ■

1.4 Reelle Zahlen

Der Begriff des Körpers umfasst einfachere mathematische Strukturen, die in diesem Buch nicht weiter untersucht werden aber z. B. für Anwendungen in der Informatik, Kryptologie und mathematischen Physik durchaus wichtig sind.

Hat man eine Menge $\mathbb{K}$ mit nur einer Verknüpfung „$+$“ **oder** „$\cdot$“, die ein Assoziativgesetz erfüllt, für die es ein neutrales Element gibt und zu der sichergestellt ist, dass **jedes** Element von $\mathbb{K}$ ein inverses Element besitzt, so nennt man $\mathbb{K}$ zusammen mit der Verknüpfung eine **Gruppe**. Ist zudem ein Kommutativgesetz erfüllt, so spricht man von einer **kommutativen Gruppe** oder **abelschen Gruppe** nach dem Mathematiker Nils Henrik Abel (1802–1829). Hat man einen Körper $\mathbb{K}$, so bildet dieser bezüglich der Addition eine kommutative Gruppe. Ebenso ist $\mathbb{K} \setminus \{0\}$ zusammen mit der Multiplikation eine kommutative Gruppe. Das neutrale Element 0 der Addition muss ausgeschlossen werden, da es dazu kein inverses Element der Multiplikation gibt.

Beispiel 1.6 (Restklassen)
Definiert man auf der Menge $\mathbb{Z}_n := \{0, 1, 2, \ldots, n-1\}$ eine Addition wie in $\mathbb{N}$, wobei man bei einem Ergebnis $\geq n$ anschließend den Rest der Ganzzahldivision durch n nimmt (**modulo** n rechnet), so entsteht eine kommutative Gruppe. In $\mathbb{Z}_5$ (also $n = 5$) gilt beispielsweise $2 + 4 = 1$, $3 + 2 = 0$, $1 + 2 = 3$. Das inverse Element zu $0 < m < n$

ist also $n - m$, wir benötigen keine „negativen Zahlen“. Das Ziffernblatt einer Uhr kann als $\mathbb{Z}_{12}$ verstanden werden. Man nennt die Zahlen $0, \ldots, n-1$ Vertreter (Repräsentanten) von **Restklassen**. So steht beispielsweise die Zahl 2 für die Restklasse $\{2, 2+n, 2+2n, 2+3n, \ldots\}$. Alle Elemente einer Restklasse haben also den gleichen Rest bei Division durch n. Die n Restklassen haben offensichtlich keine gemeinsamen Elemente.

Statt mit den Zahlen $0, \ldots, n-1$ können wir auch direkt mit Restklassen rechnen: Zwei Restklassen werden addiert, indem man aus beiden Restklassen ein beliebiges Element auswählt und die beiden Elemente addiert. Die Ergebnisrestklasse ist die, in der die so berechnete Zahl liegt. ■

Man kann beweisen, dass für Primzahlen n die Gruppen $\mathbb{Z}_n$ aus dem vorangehenden Beispiel zu einem Körper werden, wenn man bei der Multiplikation ebenfalls den Rest der Ganzzahldivision nimmt. Das klappt aber nur bei Primzahlen n. Die allgemeine Struktur, die entsteht, wenn man nicht nur Primzahlen nimmt, nennt man einen **Restklassenring**. Verzichtet man in der Definition des Körpers in den Axiomen zur Multiplikation auf die Kommutativität, die Existenz eines neutralen Elements und inverser Elemente, so erhält man die Definition eines **Rings**. Damit ist insbesondere jeder Körper ein Ring, aber ein Ring muss sich bezüglich der Multiplikation nicht so schön verhalten wie ein Körper. Restklassenringe sind Beispiele für Ringe.

Hintergrund: RSA-Public-Key-Verschlüsselung

Beim **RSA-Verschlüsselungsverfahren** wird eine als Zahl aus $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$ vorliegende Information bijektiv auf ein weiteres Element des $\mathbb{Z}_n$ abgebildet und damit verschlüsselt. Die Abbildung geschieht mittels eines öffentlichen Schlüssels. Für die Umkehrabbildung benötigt man dagegen einen privaten Schlüssel, der nicht allgemein zugänglich ist. So kann jeder eine Nachricht verschlüsseln, aber nur Besitzer des privaten Schlüssels können sie wieder entschlüsseln.

Damit die Entschlüsselung ohne Kenntnis des privaten Schlüssels (durch Raten) sehr schwierig wird, muss die natürliche Zahl n sehr groß sein. Man nimmt aber nicht irgend ein n, sondern beginnt mit zwei großen verschiedenen Primzahlen p und q, über die dann $n := p \cdot q$ definiert ist. Weiterhin seien $a, b \in \mathbb{N}$ so gewählt, dass $a \cdot b = 1 + k \cdot (p-1)(q-1)$ für einen Faktor $k \in \mathbb{N}_0$ gilt, dass also b das Inverse zu a bei der Multiplikation in $\mathbb{Z}_{(p-1)(q-1)}$ ist (dazu kann man eine Erweiterung des Euklid'schen Algorithmus einsetzen, vgl. z. B. (Hachenberger, 2005, S.186 unten)). Der öffentliche Schlüssel ist das Zahlenpaar (n, b). Der private Schlüssel ist der Vektor (n, a). Eine Zahl x wird verschlüsselt über den Rest y der Ganzzahldivision von x^b durch n. Umgekehrt erhält man aus y die Information x zurück über den Rest der Ganzzahldivision von y^a durch n. Einen Beweis finden Sie beispielsweise in (Hachenberger, 2005, S. 191).

Hintergrund: Alternativen zum Vollständigkeitsaxiom

Die antiken Griechen haben sich auf rationale Zahlen beschränkt, die aber für ihre Geometrie nicht ausreichten. Euxodos (ca. 408–355 v. Chr.) hat die Lücke mit einer Proportionenlehre geschlossen. Bis zur Arbeit von Richard Dedekind (1831–1916) gab es aber außer diesem geometrischen Verständnis keine saubere Definition einer entsprechenden Zahlenmenge $\mathbb{R}$, was insbesondere beim Umgang mit Grenzwerten problematisch war.

Ein Dedekind'scher Schnitt besteht aus zwei nicht-leeren Teilmengen $A, B \subset \mathbb{Q}$ mit $A \cup B = \mathbb{Q}$, so dass jedes Element von A kleiner als alle Elemente von B ist. Jedes Paar (A, B), das diese Bedingung erfüllt, entspricht einer reellen Zahl (die zwischen den Mengen A und B liegt und gleich $\sup A = \inf B$ ist). Man kann die reelle Zahl auch als (A, B) schreiben. Das war zuvor damit gemeint, dass die reellen Zahlen nur bis auf Umbenennung der Zahlen eindeutig festgelegt sind. Im Gegensatz zum Zugang über das Vollständigkeitsaxiom hat man hier eine konkrete, wenn auch ungewohnte, Darstellung der Zahlen. Sie ist gleichwertig mit der üblichen Dezimalbruchdarstellung, auf die wir noch genauer eingehen.

Statt des Vollständigkeitsaxioms kann man auch direkt die Existenz von Grenzwerten fordern. Äquivalent zum Vollständigkeitsaxiom ist, dass jede Cauchy-Folge (siehe Definition 2.8 auf Seite 218) von reellen Zahlen einen Grenzwert in $\mathbb{R}$ besitzt. Da wir hier die Zugehörigkeit des Supremums zu $\mathbb{R}$ verlangen, ergibt sich diese Alternative als Satz 2.8, ebenfalls Seite 218.

1.5 Reelle Funktionen

Im Buch werden die gängigen reellen Funktionen behandelt. Bei der praktischen Arbeit sind die in Abbildung 1.1 skizzierten Funktionsgraphen besonders wichtig, die Sie sich einprägen sollten.

1.6 Faktorzerlegung und Polynomdivision

Auch wenn es oft schwierig ist, die Nullstellen von Polynomen zu finden, so ist in Klausuren ist die Welt meist einfacher. Im Gegensatz zu realen Anwendungen haben hier die Polynome ganzzahlige Koeffizienten und meist auch ganzzahlige Nullstellen. Diese kann man gezielt durch Ausprobieren ermitteln: Sei $x_0 \in \mathbb{Z}$ eine ganzzahlige Nullstelle des Polynoms $p(x)$, also

$$0 = p(x_0) = a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n \Longrightarrow a_0 = x_0(-a_1 - a_2 x_0 - \cdots - a_n x_0^{n-1}).$$

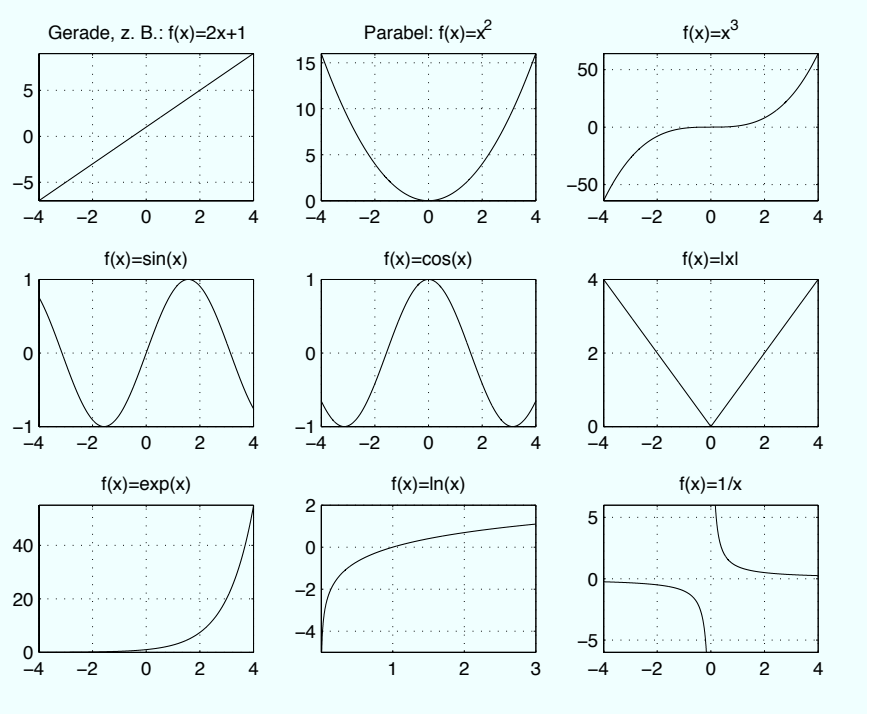

Abb. 1.1: Wichtige Funktionsgraphen

Damit ist bei ganzzahligen Koeffizienten $a_0, a_1, \ldots, a_{n-1} \in \mathbb{Z}$ eine ganzzahlige Nullstelle x_0 ein Teiler des Koeffizienten a_0.

Beispiel 1.7
Falls das Polynom $p(x) = x^3 + 2x^2 - 5x - 6$ eine ganzzahlige Nullstelle hat, dann muss sie als Teiler von -6 aus der Menge $\{-6, -3, -2, -1, 1, 2, 3, 6\}$ stammen. Tatsächlich sind die Nullstellen -3, -1 und 2. ■

1.7 Matrizen, Zeilen- und Spaltenvektoren

Beispiel 1.8 (Umrechnung von Farbwerten)
Die Farbe eines Punktes kann über seinen Rotanteil R, Grünanteil G und Blauanteil B angegeben werden. Beim analogen PAL-Fernsehsignal werden dagegen die Helligkeit Y (Luminanz, Schwarzweißbild) und die Farbdaten (Chrominanz) U und V verwendet. Der Vorteil der YUV-Darstellung besteht darin, dass das menschliche Auge Helligkeitsunterschiede viel deutlicher als Farbunterschiede wahrnimmt. Damit kann man mehr Speicherplatz für die Helligkeitsdaten verwenden und die Farbinformationen komprimiert ablegen. Die Helligkeit wäre eigentlich die Summe von R, G und B. Allerdings nimmt das Auge die Farben unterschiedlich intensiv wahr, so dass die Anteile gewichtet werden (vgl. (Schenk und Rigoll, 2010, S. 204)): $Y := 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B$. Die Farbinformation U ist die mit dem Faktor 0,492 gewichtete Differenz $B - Y$ und V die mit 0,877 gewichtete Differenz $R - Y$. Daraus ergibt sich die folgende Umrechnung in Matrixform:

$$\begin{pmatrix} Y \\ U \\ V \end{pmatrix} = \begin{bmatrix} 0{,}299 & 0{,}587 & 0{,}114 \\ -0{,}147 & -0{,}289 & 0{,}436 \\ 0{,}615 & -0{,}515 & -0{,}1 \end{bmatrix} \cdot \begin{pmatrix} R \\ G \\ B \end{pmatrix}.$$

Möchte man aus der YUV-Information wieder eine RGB-Information machen, so muss man das entsprechende Gleichungssystem lösen, also z. B. die inverse Matrix (hier gerundet) mit dem Vektor $(Y, U, V)^\top$ multiplizieren:

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} \approx \begin{bmatrix} 1 & 0 & 1{,}14 \\ 1 & -0{,}395 & -0{,}581 \\ 1 & 2{,}033 & 0 \end{bmatrix} \cdot \begin{pmatrix} Y \\ U \\ V \end{pmatrix}.$$

■

2 Zusatzmaterial zu Kapitel 2

Übersicht

2.1 Konvergenz und Divergenz von Folgen

Beispiel 2.1
Die Folge $\left(\frac{1}{a^n}\right)_{n=1}^{\infty}$ konvergiert für jeden Wert $a > 1$ gegen 0. Zu einem $\varepsilon > 0$ müssen wir zum Beweis eine Stelle n_0 finden, so dass für $n > n_0$ gilt:

$$\left|\frac{1}{a^n} - 0\right| = \frac{1}{a^n} = a^{-n} < \varepsilon.$$

Da alle Werte positiv sind, können wir auf beide Seiten der Ungleichung den Logarithmus anwenden, so dass dazu $-n\ln(a) < \ln(\varepsilon)$ äquivalent ist. Das ist wegen $-\ln(a) < 0$ für $n > -\frac{\ln(\varepsilon)}{\ln(a)}$ erfüllt, so dass wir $n_0 \geq -\frac{\ln(\varepsilon)}{\ln(a)}$ wählen können. ■

2.2 Zahlen-Reihen

Beispiel 2.2 (Z-Transformation)
Gegeben sei eine Folge $(a_k)_{k=0}^{\infty}$, die einer Wachstumsbedingung $|a_k| \leq C\alpha^k$ für von k unabhängige Konstanten C und $\alpha > 0$ genügt. Aus der Folge kann man mittels der

Reihe $\sum_{k=0}^{\infty} a_k \frac{1}{z^k}$ eine Funktion mit der Variablen z generieren. Für jeden Zahlenwert z erhält man eine andere Reihe. Die Funktion ist für die Werte von z definiert, für die die Reihe konvergiert. Die Abbildung der Folge auf die angegebene Funktion heißt **Z-Transformation**, die Funktion heißt die **Z-Transformierte** der Folge. Die Z-Transformation weist für Folgen von Zahlen ähnliche Eigenschaften auf wie die Laplace-Transformation für Funktionen.

Wegen der Wachstumsbedingung ist $\left|a_k \frac{1}{z^k}\right| \leq C \left|\frac{\alpha}{z}\right|^k$. Verlangen wir $|z| > \alpha$, so ist $\left|\frac{\alpha}{z}\right| < 1$. Damit liegt eine konvergente geometrische Reihe als Majorante vor. Die Z-Transformierte der Folge ist also in jedem Fall für Werte $|z| > \alpha$ erklärt.

Wir berechnen als Beispiel die Z-Transformation der monoton wachsenden Fibonacci-Folge $a_0 = 0$, $a_1 = 1$, $a_{k+2} = a_k + a_{k+1}$ für $k \in \mathbb{N}_0$. Da $|a_{k+2}| = a_k + a_{k+1} \leq 2a_{k+1}$ ist, erhält man durch Iteration $|a_{k+2}| \leq 2^{k+1} a_1 = \frac{1}{2} 2^{k+2}$. Außerdem ist $|a_0| = 0 < \frac{1}{2} 2^0$ und $|a_1| = 1 = \frac{1}{2} 2^1$. Die Folge erfüllt damit die Wachstumsbedingung mit $C = \frac{1}{2}$ und $\alpha = 2$. Die Z-Transformierte $A(z)$ ist daher für $|z| > 2$ über eine konvergente Reihe erklärt. Wir berechnen nun eine explizite Darstellung von $A(z)$.

Für die Transformierte der (verschobenen) Folge $(a_{k+1})_{k=0}^{\infty}$, die eine Wachstumsbedingung mit $C = 1$ und $\alpha = 2$ erfüllt, ist

$$\sum_{k=0}^{\infty} a_{k+1} \frac{1}{z^k} = z \sum_{k=0}^{\infty} a_{k+1} \frac{1}{z^{k+1}} = z \left[-a_0 + \sum_{k=0}^{\infty} a_k \frac{1}{z^k} \right] = -a_0 z + z A(z) = z A(z).$$

Diese Transformierte existiert ebenfalls für $|z| > 2$.

Entsprechend gilt für die um zwei Glieder verschobene Folge $(a_{k+2})_{k=0}^{\infty}$, die eine Wachstumsbedingung mit $C = 2$ und $\alpha = 2$ erfüllt:

$$\sum_{k=0}^{\infty} a_{k+2} \frac{1}{z^k} = z^2 \sum_{k=0}^{\infty} a_{k+2} \frac{1}{z^{k+2}} = z^2 \left[-a_0 - \frac{a_1}{z} + \sum_{k=0}^{\infty} a_k \frac{1}{z^k} \right] = -z + z^2 A(z),$$

wobei auch diese Darstellung wieder für $|z| > 2$ erklärt ist.

Damit erhalten wir durch Transformation beider Seiten der Definitionsgleichung $a_{k+2} = a_k + a_{k+1}$ für $|z| > 2$:

$$-z + z^2 A(z) = A(z) + z A(z) \Longleftrightarrow (z^2 - z - 1) A(z) = z \Longleftrightarrow A(z) = \frac{z}{z^2 - z - 1}.$$

Hier haben wir ausgenutzt, dass konvergente Reihen gliedweise addiert werden dürfen.

■

Beispiel 2.3 (Anwendung des Vergleichskriteriums)
Sei $p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$ ein Polynom vom Grad n und $q_m(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_0$ ein Polynom vom Grad m, das keine Nullstellen größer als n_0 besitzt, $a_n \neq 0$, $a_m \neq 0$. Wir überlegen uns, dass die Reihe

$$\sum_{k=n_0}^{\infty} \frac{p_n(k)}{q_m(k)}$$

genau dann konvergiert, wenn $m > n+1$ ist. Für große Werte von k verhält sich der Zähler im Wesentlichen wie $a_n k^n$ (d. h. $c|a_n|k^n \leq |q_n(k)| \leq C|a_n|k^n$ für Konstanten $c, C > 0$ und alle genügend großen k) und der Nenner wie $b_m k^m$, der Bruch also wie $\frac{a_n}{b_m}\frac{1}{k^{m-n}}$. Für $m-n \geq 2$ ist damit eine konvergente Majorante $\sum_{k=n_0}^{\infty}\frac{1}{k^{m-n}}$ gefunden, für $m-n \leq 1$ eine divergente Minorante. Damit folgt die Aussage direkt mit dem Vergleichskriterium für Reihen. ■

2.3 Newton-Verfahren

Sei $(x_n)_{n=1}^{\infty}$ die mit dem Newton-Verfahren auf Seite 283 gebildete Folge. Diese konvergiert, wenn die Bedingung (2.24) erfüllt ist. Die gute Nachricht ist aber, dass man die Bedingungen in den Anwendungen nicht unbedingt prüfen muss: Wenn die Folge $(x_n)_{n=1}^{\infty}$ gegen eine Zahl x_0 konvergiert, dann ist der Grenzwert x_0 in jedem Fall eine Nullstelle. Es kann also nichts Falsches herauskommen. Genauer gilt für eine stetig differenzierbare Funktion f, deren Ableitung stets von null verschieden ist, aufgrund der Stetigkeit von f und f':

$$x_0 = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty}\left[x_n - \frac{f(x_n)}{f'(x_n)}\right] = \left[\lim_{n\to\infty} x_n\right] - \frac{f(\lim_{n\to\infty} x_n)}{f'(\lim_{n\to\infty} x_n)} = x_0 + \frac{f(x_0)}{f'(x_0)}.$$

Vergleichen wir die linke und rechte Seite, so muss $f(x_0) = 0$ sein.

In einer Anwendung können wir also in paar Iterationen ausrechnen und abschätzen, ob sich die Werte einer Zahl nähern, die dann Nullstelle ist. Die schlechte Nachricht ist, dass die Folge ohne Einschränkung an den Startpunkt nicht immer konvergiert.

2.4 Integralrechnung

Beispiel 2.4 (Anwendung der Substitutionsregel in der Physik)
Die kinetische Arbeit ist als Kraft F mal Weg s definiert. Ändert sich die Kraft selten, so kann man die Arbeit als Summe über die Wegstücke mit konstanter Kraft bilden, ändert sich die Kraft als Funktion $F(s)$ des zurückgelegten Wegs kontinuierlich, so führt das zum Integral $W = \int_a^b F(s)\,ds$. Hier ist der Weg das Intervall $[a, b]$ (für Wege im dreidimensionalen Raum werden Kurvenintegrale verwendet). Kennt man die (bijektive) Weg-Zeit-Funktion $s(t)$, so können wir mittels Substitution $s = s(t)$ (links steht eine Variable s, rechts eine Funktion s) zu einer Integration über die Zeit t übergehen ($ds = s'(t)\,dt$):

$$W = \int_a^b F(s)\,ds = \int_{s^{-1}(a)}^{s^{-1}(b)} F(s(t))s'(t)\,dt.$$

Hier ist $s^{-1}(a)$ der Startzeitpunkt t_a bei a und $s^{-1}(b)$ der Ankunftszeitpunkt t_b bei b. Die Ableitung $s'(t)$ ist die Geschwindigkeit $v(t)$ zum Zeitpunkt t. Damit erhalten wir

$$W = \int_{t_a}^{t_b} F(s(t))v(t)\,dt.$$

Die Kraft $F(s(t))$ kann als Masse m mal Beschleunigung $a(t) = v'(t) = s''(t)$ geschrieben werden. Mit der Substitution $v = v(t)$, $dv = v'(t)\,dt$ ergibt sich mit $v_a = v(t_a)$ und $v_b = v(t_b)$

$$W = \int_{t_a}^{t_b} F(s(t))v(t)\,dt = \int_{t_a}^{t_b} mv'(t)v(t)\,dt = \int_{v(t_a)}^{v(t_b)} mv\,dv = \Big[\frac{m}{2}v^2\Big]_{v_a}^{v_b}.$$

Die Substitutionsregel wird häufig kalkülhaft angewendet. Das klappt, wenn man wie in der Physik die Variablen der Funktionen weglassen kann, da sie aus dem Zusammenhang bekannt sind. Ist die Kraft F aufgrund des Zusammenhangs zum Zeitpunkt t gemeint, so handelt es sich bei F nicht um $F(t)$ sondern um $F(s(t))$, da wir $F(s)$ als Funktion der Strecke eingeführt haben. So wird beispielsweise aus

$$\int_{t_a}^{t_b} F(s(t))v(t)\,dt = \int_{t_a}^{t_b} mv(t)a(t)\,dt = \int_{t_a}^{t_b} mv(t)v'(t)\,dt = \int_{v_a}^{v_b} mv\,dv$$

kurz

$$\int_{t_a}^{t_b} Fv\,dt = \int_{t_a}^{t_b} mva\,dt = \int_{t_a}^{t_b} mv\frac{dv}{dt}\,dt = \int_{v_a}^{v_b} mv\,dv.$$

Durch die Substitution entsteht dv aus $\frac{dv}{dt}\,dt$ quasi durch „Kürzen“. ■

2.5 Uneigentliche Integrale

Beispiel 2.5 (Gammafunktion)
Zu jedem $n \in \mathbb{N}$ zeigen wir mittels Vollständiger Induktion, dass

$$\Gamma(n) := \int_0^\infty x^{n-1}e^{-x}\,dx = (n-1)!$$

gilt. Für $n = 1$ ist $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1 = 0!$, wie wir in einem Beispiel 2.104 haben. Unter der Annahme, dass $\Gamma(n) = (n-1)!$ ist, müssen wir nun nur noch zeigen, dass $\Gamma(n+1) = n!$ ist. Das klappt mittels partieller Integration:

$$\begin{aligned}\Gamma(n+1) &= \int_0^\infty x^n e^{-x}\,dx = \lim_{t\to\infty}\int_0^t x^n e^{-x}\,dx \\ &= \lim_{t\to\infty}\big[-x^n e^{-x}\big]_0^t + \lim_{t\to\infty}\int_0^t nx^{n-1}e^{-x}\,dx\end{aligned}$$

$$= \quad 0 - 0 + n \int_0^\infty x^{n-1} e^{-x}\, dx = n \cdot \Gamma(n) = n \cdot (n-1)! = n!.$$

Das Integral existiert nicht nur für natürliche Zahlen n sondern auch, wenn man für n positive reelle Zahlen einsetzt. Über das Integral ist die **Gammafunktion** definiert. Sie erklärt die Fakultät über $\mathbb{N}$ hinaus.

Wir rechnen mit dem Majoranten-Kriteruim nach, dass man für n tatsächlich reelle Zahlen einsetzen darf, dass also $\Gamma(x) := \int_0^\infty t^{x-1} e^{-t}\, dt$ für jeden festen Parameterwert $x \in]0, \infty[$ existiert. Die Gammafunktion ist auch für negative, nicht-ganzzahlige reelle Zahlen erklärt - allerdings nicht über die hier eingesetzte Integraldarstellung.

- Für $x = 1$ haben wir den Wert bereits zu $0! = 1$ berechnet.
- Für $x \in]0, 1[$ liegt sowohl ein unbeschränkter Integrationsbereich als auch ein unbeschränkter Integrand vor: Die Funktion $t^{x-1} e^{-t}$ ist wegen $x - 1 < 0$ für $t \to 0+$ bestimmt divergent, und damit ist der Integrand nicht beschränkt.
  Wir behandeln beide Probleme separat, indem wir das Integral aufteilen:
  $$\int_0^\infty t^{x-1} e^{-t}\, dt = \int_0^1 t^{x-1} e^{-t}\, dt + \int_1^\infty t^{x-1} e^{-t}\, dt.$$
  Wegen $|t^{x-1} e^{-t}| \le t^{x-1} e^0 = t^{x-1}$ und
  $$\lim_{u \to 0+} \int_u^1 t^{x-1}\, dt = \frac{1}{x} \lim_{u \to 0+} [t^x]_{t=u}^{t=1} = \frac{1}{x}$$
  gibt es eine integrierbare Majorante, und das erste Integral $\int_0^1 t^{x-1} e^{-t}\, dt$ existiert. Auf dem unbeschränkten Integrationsbereich $[1, \infty[$ ist $|t^{x-1} e^{-t}| \le 1^{x-1} e^{-t} = e^{-t}$, wobei damit auch hier eine integrierbare Majorante gefunden ist.
- Für $x > 1$ müssen wir uns wegen $\lim_{t \to \infty} t^{x-1} = \infty$ etwas anstrengen, um eine integrierbare Majorante zu finden:
  $$|t^{x-1} e^{-t}| = \underbrace{\frac{t^{x-1}}{\exp\left(\frac{t}{2}\right)}}_{=:g(t)} \exp\left(-\frac{t}{2}\right) \le C \exp\left(-\frac{t}{2}\right).$$
  Dazu müssen wir noch zeigen, dass der Faktor $g(t)$ ist mit einer Konstante C beschränkt ist. Zunächst gibt es zu $x \in [1, \infty[$ eine Zahl $n \in \mathbb{N}$ mit $n \le x < n + 1$. Mittels n-maliger Anwendung des Satzes von L'Hospital für den Fall $\left[\frac{\infty}{\infty}\right]$ erhalten wir
  $$\lim_{t \to \infty} g(t) = \lim_{t \to \infty} \frac{t^{x-1}}{\exp\left(\frac{t}{2}\right)} = \lim_{t \to \infty} \frac{(x-1)(x-2)(x-3)\cdots(x-n) t^{x-1-n}}{\frac{1}{2^n} \exp\left(\frac{t}{2}\right)} = 0.$$
  Der Grenzwert ist null, da für $x = n$ bereits der Zähler null ist. Anderenfalls führt der Exponent $x - 1 - n < 0$ zum Grenzwert. Aufgrund des Grenzwertes gibt es eine Stelle t_0, so dass für $t > t_0$ gilt: $0 \le g(t) \le 1$. Da die Funktion g stetig ist, nimmt sie auf $[0, t_0]$ ihr Maximum und Minimum an und ist insbesondere dort beschränkt. Sie ist damit auf dem ganzen Intervall $[0, \infty[$ mit einer Konstante C beschränkt.

$\exp\left(-\frac{t}{2}\right)$ ist in diesem Fall eine integrierbare Majorante mit $\int_1^\infty \exp\left(-\frac{t}{2}\right)\,dt = 2\exp\left(-\frac{1}{2}\right) = \frac{2}{\sqrt{e}}$.

■

Man kann die Konvergenz unendlicher Reihen über die Existenz von uneigentlichen Integralen zeigen, die eine Majorante der Reihe sind, z. B. so:

$$\sum_{k=2}^{n} \frac{1}{k^2} \le \sum_{k=2}^{n} \int_{k-1}^{k} \frac{1}{x^2}\,dx = \int_1^n \frac{1}{x^2}\,dx = \left[-\frac{1}{x}\right]_1^n = -\frac{1}{n} + 1.$$

Die monoton wachsende Reihe ist daher beschränkt und somit konvergent mit

$$\sum_{k=2}^{\infty} \frac{1}{k^2} \le \int_1^\infty \frac{1}{x^2}\,dx = 1.$$

Über einen Vergleich mit Integralen lässt sich auch die Divergenz von Reihen nachweisen, z. B. ist die harmonische Reihe nicht beschränkt wegen

$$\sum_{k=1}^{n} \frac{1}{k} \ge \sum_{k=1}^{n} \int_{k}^{k+1} \frac{1}{x}\,dx = \int_1^{n+1} \frac{1}{x}\,dx = [\ln|x|]_1^{n+1} = \ln(n+1).$$

Die hier vorgenommene Abschätzungstechnik von Summen gegen Integrale eignet sich, um eine Näherungsformel für die Fakultät zu motivieren. Ist ein Kürzen von Fakultäten nicht möglich, dann kann man nicht gut mit ihnen rechnen. Wünschenswert wäre eine Darstellung über Potenzen. Dies leistet näherungsweise die **Stirling'sche Formel**

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Je größer n ist, um so genauer wird $n!$ durch die rechte Seite berechnet. Präziser formuliert gibt es für jedes n eine Zahl $0 \le \xi_n \le 1$ mit

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{\xi_n}{12n}\right),$$

wobei der zusätzliche Korrekturfaktor für $n \to \infty$ gegen 1 strebt. Dass diese Formel in etwa stimmen kann, sieht man so: Zunächst verwenden wir eine Rechenregel für den Logarithmus:

$$\ln(n!) = \ln(2 \cdot 3 \cdots n) = \ln(2) + \ln(3) + \cdots + \ln(n) = \sum_{k=2}^{n} \ln(k).$$

Dann können wir wie zuvor sowohl nach unten als auch nach oben gegen ein Integral abschätzen. Dabei nutzen wir aus, dass der Logarithmus (streng) monoton wachsend ist:

$$\sum_{k=2}^{n} \ln(k) \ge \int_1^n \ln(x)\,dx = [x \cdot \ln(x) - x]_1^n = n(\ln(n) - 1) + 1,$$

$$\sum_{k=2}^{n} \ln(k) \le \int_2^{n+1} \ln(x)\,dx = [x\cdot\ln(x)-x]_2^{n+1} = (n+1)(\ln(n+1)-1) - 2\ln(2) + 2.$$

Einsetzen in die Exponentialfunktion liefert

$$n! = \exp(\ln(n!)) \ge \exp(n(\ln(n)-1)+1) = e\frac{\mathrm{e}^{n\ln(n)}}{e^n} = e\left(\frac{n}{e}\right)^n$$

und andererseits

$$n! \le \exp((n+1)(\ln(n+1)-1) - 2\ln(2) + 2) = \frac{(n+1)^{n+1}}{e^{n+1}}\frac{e^2}{2^2},$$

also $n! = n\cdot(n-1)! \le \frac{e^2}{4} n \left(\frac{n}{e}\right)^n$.

Wir erkennen in beiden Abschätzungen die dominierende Größe $\left(\frac{n}{e}\right)^n$. Eine Präzisierung (beachte, dass für $n \ge 2$ gilt: $e \le \sqrt{2\pi n} \le \frac{e^2}{4} n$) führt zur Stirling'schen Formel.

2.6 Kurvendiskussion und Extremalprobleme

Mittels Differenzialrechnung findet man nur lokale Extrema. Um darüber globale Extrema zu ermitteln, muss man Zusatzüberlegungen anstellen:

Hat man zu einer auf einem Intervall $[a,b]$ differenzierbaren Funktion über die Bedingung $f'(x)=0$ alle möglichen lokalen Extremstellen in $]a,b[$ ermittelt, dann können globale Extrema (die ja insbesondere lokale Extrema sind) nur an diesen Stellen und an den Randpunkten a und b liegen. Da eine stetige Funktion auf $[a,b]$ sowohl das globale Maximum als auch das globale Minimum annimmt, befinden sich beide tatsächlich an jeweils mindestens einer dieser Stellen, und man muss nur die Funktionswerte miteinander vergleichen.

Ist der Definitionsbereich unbeschränkt, dann kann man Grenzwerte in die Überlegung einbeziehen:

Sei f auf $\mathbb{R}$ differenzierbar mit existierenden Grenzwerten $g_1 = \lim_{x\to-\infty} f(x)$ und $g_2 = \lim_{x\to\infty} f(x)$ (wobei $\pm\infty$ zugelassen sind), und sei M der größte Funktionswert an einer Nullstelle der Ableitung. Falls M größer oder gleich g_1 und g_2 ist, dann liegt dort ein globales Maximum. Sei m der kleinste Funktionswert an einer Nullstelle der Ableitung. Falls m kleiner oder gleich g_1 und g_2 ist, dann liegt dort ein globales Minimum.

Wir beweisen die Aussage für das globale Maximum, indem wir annehmen, dass es einen größeren Funktionswert M' gibt und dies zum Widerspruch führen. M werde an der Stelle x_0 und M' an der Stelle x_1 angenommen.

Wir betrachten nun den Fall $x_1 < x_0$. Wegen $\lim_{x\to-\infty} f(x) \le M < M'$ gibt es eine Stelle $x_2 < x_1$ mit $f(x_2) < M'$. Die stetige Funktion f nimmt auf dem Intervall $[x_2, x_0]$ ein Maximum an, das größer oder gleich M' ist. Aufgrund der Funktionswerte am Rand muss dies an einer Stelle x_3 im Inneren $]x_2, x_0[$ liegen. Da f differenzierbar ist,

ist $f'(x_3) = 0$. Das steht aber im Widerspruch dazu, dass M der größte Funktionswert an den Nullstellen von f' ist.

Im Fall $x_0 < x_1$ erhält man den gleichen Widerspruch mittels des Grenzwertes $\lim_{x\to\infty} f(x) \leq M < M'$. Der Beweis für das globale Minimum ist völlig analog.

Beispiel 2.6
Wir suchen die globalen Extrema der Funktion $f(x) := \frac{x}{x^2+1}$. Die Nullstellen der Ableitung $f'(x) = \frac{1-x^2}{(x^2+1)^2}$ sind -1 und 1 mit $f(-1) = -\frac{1}{2}$ und $f(1) = \frac{1}{2}$. Da $\lim_{x\to\pm\infty} f(x) = 0$ ist, liegt das globale Maximum an der Stelle 1 und das globale Minimum an der Stelle -1. Insbesondere muss man hier zur Klassifikation der Extrema die zweite Ableitung gar nicht mehr ausrechnen. ■

2.7 Konvergenz von Potenzreihen

Zur Folge der Fibonacci-Zahlen $a_0 = 0$, $a_1 = 1$, $a_k = a_{k-2} + a_{k-1}$ für $k \geq 2$ bilden wir eine Potenzreihe $\sum_{k=0}^{\infty} a_k x^k$. Da der Grenzwert $\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = \Phi = \frac{1+\sqrt{5}}{2}$ der goldene Schnitt ist, ist der Konvergenzradius der Kehrwert des goldenen Schnitts: $\rho = \frac{1}{\Phi}$. Mittels dieser Potenzreihe kann man die Binet-Formel herleiten, die eine explizite Darstellung der a_k ohne Rekursion liefert. Im Buch erhalten wir die Formel über eine Eigenwertbetrachtung am Ende des Kapitels zur Linearen Algebra. Wir gehen hier auf den Potenzreihenansatz kurz ein (vgl. (Heuser, 2009, S. 378)):

Über die Berechnung der Ableitungen entwickeln wir die Funktion $g(x) := \frac{1}{x-x_0}$ für ein festes $x_0 \in \mathbb{R}$ in eine Potenzreihe.

$$g'(x) = -\frac{1}{(x-x_0)^2},\ g''(x) = 2\frac{1}{(x-x_0)^3}, \ldots, g^{(n)}(x) = n!(-1)^n \frac{1}{(x-x_0)^{n+1}}.$$

Damit ist $g(0) = -\frac{1}{x_0}$, $g^{(n)}(0) = -n!\frac{1}{x_0^{n+1}}$, und wir erhalten die Reihendarstellung

$$g(x) = \frac{1}{x-x_0} = -\sum_{k=0}^{\infty} \frac{1}{x_0^{k+1}} x^k, \quad x \in]-|x_0|, |x_0|[,$$

Zum Konvergenzradius:

$$\lim_{k\to\infty} \left| \frac{1}{x_0^{k+1}} \right|^{1/k} = \exp\left(-\ln|x_0| \lim_{k\to\infty} \frac{k+1}{k} \right) = \frac{1}{|x_0|}.$$

Damit haben wir den Konvergenzradius $|x_0|$. Mehr war auch nicht zu erwarten, da die Funktion g an der Stelle x_0 nicht definiert ist.

Jetzt können wir die Binet-Formel für die Glieder der Fibonacci-Folge $a_0 = 0$, $a_1 = 1$, $a_k = a_{k-1} + a_{k-1}$ $(k \geq 2)$ herleiten. Dazu sei f die Grenzfunktion der über die Fibonacci-Folge gebildeten Potenzreihe

$$f(x) = \sum_{k=0}^{\infty} a_k x^k,$$

deren Konvergenzradius der Kehrwert des goldenen Schnitts ist. Den brauchen wir hier aber noch gar nicht. Wichtig ist nur, dass er größer als null ist: Die Folge der Fibonacci-Zahlen ist monoton wachsend. Damit ist für $k \geq 2$ $a_{k+1} = a_k + a_{k-1} \leq 2a_k$, also $a_{k+1} \leq 2^k a_1 = 2^k$. Damit hat die Potenzreihe für $|x| < \frac{1}{2}$ die als geometrische Reihe konvergente Majorante $\sum_{k=0}^{\infty}(2x)^k$ und damit einen Konvergenzradius $\rho \geq \frac{1}{2}$.

Wir können also über die Funktion f im Intervall $\left]-\frac{1}{2}, \frac{1}{2}\right[$ verfügen. Hier erfüllt sie die Gleichung

$$f(x) - xf(x) - x^2 f(x) = x,$$

die sich direkt durch Einsetzen der Potenzreihe ergibt:

$$\begin{aligned} f(x) - xf(x) - x^2 f(x) &= \sum_{k=0}^{\infty} a_k x^k - \sum_{k=0}^{\infty} a_k x^{k+1} - \sum_{k=0}^{\infty} a_k x^{k+2} \\ &= \sum_{k=0}^{\infty} a_k x^k - \sum_{k=1}^{\infty} a_{k-1} x^k - \sum_{k=2}^{\infty} a_{k-2} x^k \\ &= a_0 x^0 + a_1 x^1 - a_0 x^1 + \sum_{k=2}^{\infty} \underbrace{(a_k - a_{k-1} - a_{k-2})}_{=0} x^k = x. \end{aligned}$$

Damit erhalten wir für f die Darstellung $f(x) = -\frac{x}{x^2+x-1}$. Der Nenner hat die Nullstellen $x_1 = \frac{-1+\sqrt{5}}{2} = \frac{1}{\Phi}$ und $x_2 = \frac{-1-\sqrt{5}}{2} = -\Phi$, mit denen wir die Partialbruchzerlegung $f(x) = \frac{1}{x_1 - x_2}\left[\frac{-x_1}{x-x_1} + \frac{x_2}{x-x_2}\right]$ erhalten. Jetzt kommt die eingangs berechnete Reihendarstellung von g sowohl für $x_0 = x_1$ als auch für $x_0 = x_2$ zum Zuge:

$$f(x) = \frac{1}{x_1 - x_2}\left[x_1 \sum_{k=0}^{\infty} \frac{1}{x_1^{k+1}} x^k - x_2 \sum_{k=0}^{\infty} \frac{1}{x_2^{k+1}} x^k\right] = \frac{1}{\sqrt{5}} \sum_{k=0}^{\infty}\left[\frac{1}{x_1^k} - \frac{1}{x_2^k}\right] x^k.$$

Jetzt haben wir zwei Potenzreihendarstellungen für f. Die Koeffizienten beider Reihen (die sich aus den Ableitungen von f am Entwicklungsmittelpunkt 0 berechnen), müssen gleich sein:

$$a_k = \frac{1}{\sqrt{5}}\left[\frac{1}{x_1^k} - \frac{1}{x_2^k}\right] = \frac{1}{\sqrt{5}}\left[\Phi^k - \left(-\frac{1}{\Phi}\right)^k\right].$$

Das ist die Binet-Formel. Mit ihr erhält man wegen $\Phi > 1$ und $\lim_{k\to\infty}\left[-\frac{1}{\Phi}\right]^{k+1} = \lim_{k\to\infty}\left[-\frac{1}{\Phi}\right]^k = 0$ den Grenzwert

$$\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = \lim_{n\to\infty} \frac{\Phi^{k+1} - \left[-\frac{1}{\Phi}\right]^{k+1}}{\Phi^k - \left[-\frac{1}{\Phi}\right]^k} = \Phi.$$

2.8 Differenziation und Integration von Potenzreihen

Beispiel 2.7
Um zur Potenzreihendarstellung der Exponentialfunktion zu gelangen, haben wir bereits über ihre Ableitungen verfügt. Hätten wir exp direkt über die Potenzreihe definiert, so hätten wir durch gliedweises Ableiten nun auch die Ableitungen:

$$\frac{d}{dx}e^x = \frac{d}{dx}\sum_{k=0}^{\infty}\frac{x^k}{k!} = \sum_{k=0}^{\infty}\frac{d}{dx}\frac{x^k}{k!} = \sum_{k=1}^{\infty}\frac{kx^{k-1}}{k!} = \sum_{k=1}^{\infty}\frac{x^{k-1}}{(k-1)!} = \sum_{k=0}^{\infty}\frac{x^k}{k!} = e^x.$$

■

Beispiel 2.8
Der Sinus Cardinalis $\operatorname{sinc}(x)$ ist die an der Stelle $x = 0$ stetig mit dem Wert 1 ergänzte Funktion $\frac{\sin(x)}{x}$. Mit der Potenzreihe des Sinus erhalten wir für $x \neq 0$ die Darstellung

$$\operatorname{sinc}(x) = \frac{1}{x}\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k+1}}{(2k+1)!} = \sum_{k=0}^{\infty}(-1)^k\frac{x^{2k}}{(2k+1)!}.$$

Da die Potenzreihe des Sinus auf $\mathbb{R}$ konvergiert, konvergiert auch diese Reihe für alle $x \neq 0$. Die Konvergenz am Entwicklungsmittelpunkt $x = 0$ gegen 1 ist offensichtlich. Damit haben wir wieder eine Potenzreihe mit $\rho = \infty$. Die Grenzfunktion ist nicht nur stetig auf $\mathbb{R}$, sondern sogar beliebig oft differenzierbar.

Da die Reihe insbesondere für den Entwicklungsmittelpunkt 0 den Wert 1 hat, erhalten wir über die Potenzreihe den Grenzwert $\lim_{x\to 0}\frac{\sin(x)}{x} = 1$ ohne die im Buch gemachten geometrischen Überlegungen.

Jetzt wissen wir zudem, dass der Sinus Cardinalis unendlich oft differenzierbar auf $\mathbb{R}$ ist. ■

3 Zusatzmaterial zu Kapitel 3

Übersicht

3.1 Skalarprodukt und Norm

Beispiel 3.1
Suchmaschinen im Internet nutzen das Standardskalarprodukt im $\mathbb{R}^n$, um die Ähnlichkeit einer Anfrage mit einer Internetseite zu bewerten. Dazu zählt man alle relevanten Wörter der Anfrage und schreibt die Anzahlen in jeweils einen Vektor. Jede Stelle des Vektors entspricht dabei genau einem aussagekräftigen Wort der jeweiligen Sprache. Die Raumdimension n ist also sehr groß, da mehrere zehntausend Wörter einer Sprache zu berücksichtigen sind. Ebenso wird das mit der Anfrage zu vergleichende Dokument in einen Vektor überführt. Anfrage und Dokument sind dann ähnlich, wenn das Skalarprodukt der beiden Vektoren dividiert durch die Beträge bzw. Normen der Vektoren nahe bei 1 ist. Im $\mathbb{R}^2$ oder $\mathbb{R}^3$ entspricht dies dem Kosinus des Winkels zwischen den Vektoren, der damit nahe bei Null liegt. Anfrage und Dokument sind also ähnlich, wenn die zugehörigen Vektoren fast in die gleiche Richtung zeigen. ■

Hintergrund: Skalarprodukt für einen komplexen Vektorraum

Sei $(V, +; \mathbb{C}, \cdot)$ ein komplexer Vektorraum. Eine Abbildung

$$„ \cdot “ : \quad V \times V \mapsto \mathbb{C}$$

heißt **Skalarprodukt**, falls sie für beliebige Vektoren $\vec{a}, \vec{b}, \vec{c} \in V$ und Skalare $\lambda \in \mathbb{C}$ die folgenden Regeln (Axiome) erfüllt:

1. **Positive Definitheit**: Das Skalarprodukt eines Vektors mit sich selbst ist eine nichtnegative reelle Zahl: $\vec{a} \cdot \vec{a} \geq 0$. Aus $\vec{a} \cdot \vec{a} = 0$ folgt $\vec{a} = \vec{0}$.

2. **Die Abbildung ist hermitesch**: $\vec{a} \cdot \vec{b} = \overline{\vec{b} \cdot \vec{a}}$. Diese Eigenschaft ersetzt die bei reellen Vektorräumen geforderte Symmetrie. Wegen $\vec{a} \cdot \vec{a} = \overline{\vec{a} \cdot \vec{a}}$ ist insbesondere das Skalarprodukt eines Vektors mit sich selbst eine reelle Zahl. Das müssten wir bei der positiven Definitheit damit gar nicht fordern.
3. **Multiplikation mit Skalar**: $(\lambda\vec{a}) \cdot \vec{b} = \overline{\lambda}(\vec{a} \cdot \vec{b})$ aber $\vec{a} \cdot (\lambda\vec{b}) = \lambda(\vec{a} \cdot \vec{b})$. Homogenität liegt also nur beim zweiten Argument vor. Zieht man aus dem ersten Argument einen Skalar vor das Skalarprodukt, so muss man ihn komplex konjugieren.
4. **Additivität in beiden Argumenten**: $(\vec{a}+\vec{b}) \cdot \vec{c} = \vec{a} \cdot \vec{c} + \vec{b} \cdot \vec{c}$, $\vec{a} \cdot (\vec{b}+\vec{c}) = \vec{a} \cdot \vec{b} + \vec{a} \cdot \vec{c}$.

Das komplexe Standardskalarprodukt zweier komplexer Vektoren $\vec{a} = (a_1, a_2, \ldots, a_n)$ und $\vec{b} = (b_1, b_2, \ldots, b_n) \in \mathbb{C}^n$ ist definiert als

$$\vec{a} \cdot \vec{b} := \sum_{k=1}^{n} \overline{a_k} \cdot b_k. \tag{3.1}$$

3.2 Eigenwerte und Eigenvektoren

Der Satz von Cayley-Hamilton

Das charakteristische Polynom $p(s)$ einer $(n \times n)$-Matrix $\mathbf{A}$ hat eine interessante Eigenschaft, auf die im Buch nicht eingegangen wird. Setzt man die Matrix in das Polynom ein, d. h. berechnet man $p(\mathbf{A})$, so entsteht die Nullmatrix. Das ist der **Satz von Cayley-Hamilton**.

Da $\mathbf{A}$ eine quadratische Matrix ist, lassen sich die Potenzen von $\mathbf{A}$ durch entsprechend häufige Multiplikation mit $\mathbf{A}$ erzeugen, und $p(\mathbf{A})$ ist tatsächlich eine wohldefinierte $(n \times n)$-Matrix.

Wir beweisen den Satz für den Spezialfall, dass es eine Basis des $\mathbb{C}^n$ aus Eigenvektoren von $\mathbf{A}$ gibt. Jeder Vektor lässt sich dann als Linearkombination von Eigenvektoren schreiben. Die Nullmatrix ist genau die Matrix, die bei Multiplikation mit jedem Vektor den Nullvektor liefert. Wir müssen also nur zeigen, dass $p(\mathbf{A})$ multipliziert mit jedem Eigenvektor den Nullvektor ergibt. Sei dazu $\vec{d}$ ein Eigenvektor zum Eigenwert s. Wegen

$$\mathbf{A}^k\vec{d} = \mathbf{A}^{k-1}s\vec{d} = s\mathbf{A}^{k-1}\vec{d} = s^2\mathbf{A}^{k-2}\vec{d} = \cdots = s^k\vec{d}$$

ist $p(\mathbf{A})\vec{d} = p(s)\vec{d}$. Da s ein Eigenwert ist, gilt aber $p(s) = 0$ und damit $p(\mathbf{A})\vec{d} = \vec{0}$.

Adjungierte Matrix

Bei Matrizen mit komplexen Einträgen verbindet man die Transposition häufig mit der komplexen Konjugation. Das führt zum folgenden Begriff.

Die **Adjungierte** $\mathbf{A}^*$ einer $(m \times n)$-Matrix $\mathbf{A}$ ist die $(n \times m)$-Matrix, die aus $\mathbf{A}$ durch Transposition und anschließender komplexer Konjugation aller Einträge entsteht: $\mathbf{A}^* := \overline{\mathbf{A}^\top}$.
Beispielsweise ist

$$\begin{bmatrix} 1-2j & 2+j & 3 \\ 4 & 5 & 6j \end{bmatrix}^* = \begin{bmatrix} 1+2j & 4 \\ 2-j & 5 \\ 3 & -6j \end{bmatrix}.$$

Die adjungierte Matrix einer reellen Matrix ist die transponierte Matrix. Eine Matrix $\mathbf{A}$ heißt **selbstadjungiert** genau dann, wenn $\mathbf{A} = \mathbf{A}^*$ ist.

Hintergrund: Eigenwerte selbstadjungierter Matrizen

Analog zu symmetrischen reellen Matrizen verhalten sich selbstadjungierte komplexe Matrizen.

Es sei $\mathbf{A} \in \mathbb{C}^{n\times n}$ eine selbstadjungierte Matrix.

1. Alle Eigenwerte von $\mathbf{A}$ sind reell.
2. Alle Eigenvektoren zu **verschiedenen** (reellen) Eigenwerten sind orthogonal bezüglich des komplexen Standardskalarprodukts (3.1) auf Seite 24, d. h., ihr komplexes Standardskalarprodukt ist null.

Wir beweisen beide Aussagen:

1. Sei $s \in \mathbb{C}$ ein Eigenwert mit Eigenvektor $\vec{d} \in \mathbb{C}^n$, $\vec{d} \neq \vec{0}$. Insbesondere ist das komplexe Standardskalarprodukt $\overline{\vec{d}^\top} \cdot \vec{d} = |\vec{d}|^2$ ungleich null. Wegen

$$s\overline{\vec{d}^\top} \cdot \vec{d} = \overline{\vec{d}^\top} \cdot \mathbf{A}\vec{d} = \overline{(\mathbf{A}^* \cdot \vec{d})^\top} \cdot \vec{d} = \overline{(\mathbf{A} \cdot \vec{d})^\top} \cdot \vec{d} = \overline{s\vec{d}^\top} \cdot \vec{d} = \overline{s}\overline{\vec{d}^\top} \cdot \vec{d}$$

   muss daher $s = \overline{s}$ und damit s reell sein.
2. Sind $\vec{d}_1$ und $\vec{d}_2 \in \mathbb{C}^n$ Eigenvektoren zu verschiedenen (laut a) reellen) Eigenwerten $s_1 \neq s_2 \in \mathbb{R}$, so sind diese bezüglich des komplexen Standardskalarprodukts orthogonal ($s_1 = \overline{s_1}$):

$$\begin{aligned} s_1\left(\overline{\vec{d}_1^\top} \cdot \vec{d}_2\right) &= \left(\overline{s_1}\overline{\vec{d}_1^\top}\right) \cdot \vec{d}_2 = \overline{(\mathbf{A} \cdot \vec{d}_1)^\top} \cdot \vec{d}_2 = \overline{\vec{d}_1^\top}\mathbf{A}^* \cdot \vec{d}_2 = \overline{\vec{d}_1^\top}\mathbf{A} \cdot \vec{d}_2 \\ &= \overline{\vec{d}_1^\top} s_2 \vec{d}_2 = s_2\left(\overline{\vec{d}_1^\top}\vec{d}_2\right), \end{aligned}$$

   denn $\overline{\vec{d}_1^\top}\vec{d}_2$ muss wegen $s_1 \neq s_2$ gleich null sein.

Diagonalisierung von Matrizen

Entsprechend dem Satz 3.39 (vgl. Kasten oben) sind für jede komplexe Matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ äquivalent:

1. $\mathbf{A}$ ist selbstadjungiert.
2. Es existiert eine **unitäre** Matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$, d. h. $\mathbf{X}^{-1} = \mathbf{X}^*$, und eine reelle Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ mit $\mathbf{D} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{X}^*\mathbf{A}\mathbf{X}$.

Symmetrische und selbstadjungierte Matrizen lassen sich zu einer reellen Diagonalmatrix diagonalisieren. Generell kann man eine Matrix zu einer komplexen Diagonalmatrix diagonalisieren, wenn sie mit ihrer Adjungierten kommutiert, d. h., wenn sie die Bedingung $\mathbf{A}^* \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A}^*$ erfüllt. Solche Matrizen heißen **normal**.

Insbesondere sind offensichtlich reelle symmetrische und komplexe selbstadjungierte Matrizen normal, da $\mathbf{A}^* \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A}^*$.

Selbst wenn man eine Matrix verwendet, die nicht diagonalisierbar ist, so kann man sie unter Zuhilfenahme von sogenannten Hauptvektoren (eine Verallgemeinerung von Eigenvektoren) doch noch nahezu in Diagonalform bringen. Eine entsprechende Darstellung ist die Jordan-Normalform.

4 Zusatzmaterial zu Kapitel 4

Übersicht

4.1 Funktionen mit mehreren Variablen

Hintergrund: Currying

Wir werden mit einer Funktion $f(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$ häufig arbeiten, indem wir für zwei Variablen feste Werte einsetzen und dann eine Funktion einer Variablen erhalten. Traut man sich zu, mit Abbildungen zu arbeiten, deren Werte wieder Abbildungen sind, so kann man dieses Prinzip auch als Hintereinanderausführung mehrerer Abbildungen schreiben. Dabei handelt sich um ein Konzept der funktionalen Programmierung aus der Informatik, das **Currying** heißt. Es wird z. B. durch die Programmiersprache Scala unterstützt, die bei der Implementierung von Twitter verwendet wurde.

Wir schreiben $f(x, y, z) = [[f_1(x)](y)](z)$, wobei f_1 eine Abbildung mit Definitionsbereich $\mathbb{R}$ und Werten ist, die selbst Abbildungen sind. Die Abbildung $f_2 := [f_1(x)]$ ist ebenfalls eine Abbildung von $\mathbb{R}$ in eine Menge von Abbildungen. Die Abbildung $f_3 := [[f_1(x)](y)] : \mathbb{R} \to \mathbb{R}$ berechnet schließlich in Abhängigkeit von z die gesuchte Zahl. Zu einem Wert $x \in \mathbb{R}$ liefert f_1 also die von x abhängende Abbildung $f_2 := [f_1(x)] : y \to f_3$. Dabei ist $f_3 : \mathbb{R} \to \mathbb{R}$ eine von x und y abhängende Abbildung mit $f_3(z) := f(x, y, z)$. Für verschiedene Werte von x entstehen (gegebenenfalls)

verschiedene Abbildungen f_2, in denen x konstant ist. Verschiedene y führen dann zu (gegebenenfalls) verschiedenen Abbildungen f_3. In f_3 sind x und y konstante Parameter, die Variable ist z.

4.2 Ableitungen von reellwertigen Funktionen mit mehreren Variablen

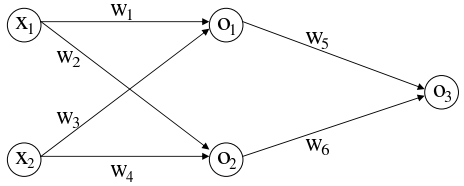

Abb. 4.1: Neuronales Netz

Beispiel 4.1 (Mit dem Gradientenverfahren lernendes Neuronales Netz)
Neuronale Netze bilden das Verhalten von vernetzten Nervenzellen nach. Sie berechnen dazu aus Eingabewerte einen oder mehrere Ausgabewerte. In Abbildung 4.1 ist ein einfaches dreischichtiges Netz dargestellt. Die Neuronen der ersten Schicht nehmen Eingangswerte x_1 und x_2 an und leiten diese an die beiden Neuronen der zweiten, mittleren Schicht weiter. Dabei werden die Werte jedoch gewichtet. Als Eingabe erhält das obere Neuron der mittleren Schicht den Wert $w_1x_1 + w_3x_2 = (w_1, w_3) \cdot (x_1, x_2)$ und das untere den Wert $w_2x_1 + w_4x_2 = (w_2, w_4) \cdot (x_1, x_2)$. Die Gewichte w_k werden in einem Lernvorgang so gewählt, dass das Netz am Ende ein gewünschtes Verhalten zeigt. Darauf gehen wir gleich ein. Die Neuronen der mittleren Schicht berechnen nun eine Ausgabe, die an die dritte Schicht weitergereicht wird. Einfache Schwellenwertneuronen vergleichen die Eingabe mit einem Schwellenwert. Ist sie kleiner, wird 0 weitergereicht, sonst 1. Das erweist sich jedoch für die Bestimmung der Gewichte beim Lernen als schwierig, da der Ausgabewert über eine Sprungfunktion und nicht über eine differenzierbare Funktion berechnet wird. Hier verwenden wir stattdessen eine **Aktivierungsfunktion**, die einen Ausgabewert zwischen 0 und 1 berechnet. Üblich ist die Funktion $g(x) := \frac{1}{1+e^{-x}}$, da sich ihre Ableitung an einer Stelle x direkt aus dem Funktionswert $g(x)$ ergibt und so die Formeln einfacher werden:

$$g'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \frac{1+e^{-x}-1}{1+e^{-x}} = \frac{1}{1+e^{-x}} \left[1 - \frac{1}{1+e^{-x}}\right] = g(x)[1-g(x)].$$

Man nennt g eine **sigmoide Funktion**, da sie monoton wachsend mit $\lim_{x\to-\infty} g(x) = 0$ und $\lim_{x\to\infty} g(x) = 1$ ist. Der „mittlere Wert“ $\frac{1}{2}$ wird für $x = 0$ angenommen.

Damit entspricht 0 einem Schwellenwert. Möchte man einen anderen Schwellenwert Φ realisieren, so kann man $g(x-\Phi)$ verwenden. Das erreicht man, indem man z. B. statt der Eingabe $w_1x_1+w_3x_2$ die Eingabe $w_1x_1+w_3x_2+(-\Phi)\cdot 1$ verwendet. So wird der Schwellenwert zu einem weiteren Gewicht, und wir müssen beim Lernen nur Gewichte optimieren. Um das Beispiel kurz zu halten, verzichten wir hier auf die Anpassung der Schwellenwerte über Gewichte und lassen den Term $(-\Phi)\cdot 1$ weg.

Wir bezeichnen mit $o_1 = o_1(w_1, w_3)$, $o_2 = o_2(w_2, w_4)$ und $o_3 = o_3(w_1, \ldots, w_6)$ die Werte, die die Neuronen der mittleren und letzten Schicht berechnen. Im Folgenden lassen wir nur wegen der Übersichtlichkeit die Argumente teilweise weg.

Das Netz berechnet also aus vorgegebenen Werten x_1 und x_2 in Abhängigkeit von den Gewichten den Zahlenwert

$$\begin{aligned} o_3(w_1,\ldots,w_6) &= g(w_5o_1(w_1,w_3)+w_6o_2(w_2,w_4)) \\ &= g(w_5g(w_1x_1+w_3x_2)+w_6g(w_2x_1+w_4x_2)). \end{aligned}$$

Wir berechnen die partiellen Ableitungen

$$\begin{aligned} \frac{\partial o_3}{\partial w_5}(w_1,\ldots,w_6) &= g'(w_5o_1+w_6o_2)o_1 = g(w_5o_1+w_6o_2)[1-g(w_5o_1+w_6o_2)]o_1 \\ &= o_3[1-o_3]o_1, \\ \frac{\partial o_3}{\partial w_6}(w_1,\ldots,w_6) &= o_3[1-o_3]o_2, \\ \frac{\partial o_3}{\partial w_1}(w_1,\ldots,w_6) &= o_3[1-o_3]w_5o_1[1-o_1]x_1, \\ \frac{\partial o_3}{\partial w_2}(w_1,\ldots,w_6) &= o_3[1-o_3]w_6o_2[1-o_2]x_1, \\ \frac{\partial o_3}{\partial w_3}(w_1,\ldots,w_6) &= o_3[1-o_3]w_5o_1[1-o_1]x_2, \\ \frac{\partial o_3}{\partial w_4}(w_1,\ldots,w_6) &= o_3[1-o_3]w_6o_2[1-o_2]x_2 \end{aligned}$$

und kennen damit $\operatorname{grad} o_3(w_1,\ldots,w_6)$. Möchte man für eine spezielle Eingabe x_1, x_2 die Gewichte so bestimmen, dass die Ausgabe o_3 möglichst nahe an einer vorgegebenen Ausgabe o ist, so kann man den Fehler $f(w_1,\ldots,w_6) := (o-o_3(w_1,\ldots,w_6))^2$ minimieren. Dazu kann man von einer Startbelegung der Gewichte ausgehend sukzessive Schritte in Richtung des negativen Gradienten $-\operatorname{grad} f(w_1,\ldots,w_6) = -2(o-o_3(w_1,\ldots,w_6))[-\operatorname{grad} o_3(w_1,\ldots,w_6)]$ machen. Bei jedem Schritt addiert man zum aktuellen Vektor der Gewichte den Wert $\lambda 2(o-o_3)\operatorname{grad} o_3$, wobei λ die Schrittweite bezeichnet, die auch Lernrate genannt wird. Wir minimieren den Fehler also mit dem Gradientenverfahren. Ein Neuronales Netz soll auch bei anderen Eingabewerten geeignete Ausgaben liefern. Daher benutzt man die Methode des steilsten Abstiegs abwechselnd für verschiedene Eingaben und ihre vorgegebenen Ausgaben. So wird das Netz trainiert, und man hat die Hoffnung, dass es sich für „ähnliche“ Eingaben auch ähnlich verhält und so Muster erkennen kann. ■

Beispiel 4.2 (Eine andere Sicht auf das Newton-Verfahren)
Wir können das eindimensionale Newton-Verfahren, mit dem wir näherungsweise Nullstellen einer differenzierbaren Funktion $f : \mathbb{R} \to \mathbb{R}$ bestimmt haben, als modifiziertes Gradientenverfahren auffassen. Die Nullstellenbestimmung ist äquivalent mit der Suche nach einem (globalen) Minimum der Funktion $[f(x)]^2$. Beginnend an einer Stelle x_0 geht man hier ein Stück in Richtung des negativen Gradienten (der negativen Ableitung) $-2f(x)f'(x)$ bis zu der Stelle, an der die Tangente an f im Punkt $(x_0, f(x_0))$ die x-Achse schneidet. Diese Stelle ist x_1, und der Vorgang wiederholt sich. Der negative Gradient bestimmt hier also, ob nach rechts oder nach links weitergesucht werden soll. Wie weit in diese Richtung gegangen wird, wird allerdings einfacher als beim Gradientenverfahren berechnet, indem die Funktion durch die Tangente angenähert wird. ■

Beispiel 4.3 (Partielle Differenzialgleichung)
Häufig sucht man Funktionen, über deren Ableitungsverhalten man aufgrund von Naturgesetzen etwas weiß. Wie sehen beispielsweise die (total differenzierbaren) Funktionen $f : \mathbb{R}^2 \to \mathbb{R}$ aus, die die Gleichung

$$a\frac{\partial f}{\partial x}(x,y) + b\frac{\partial f}{\partial y}(x,y) = 0$$

für alle $(x, y) \in \mathbb{R}^2$ erfüllen? Dabei ist $\vec{0} \neq (a, b) \in \mathbb{R}^2$ ein konstanter Vektor. Diese Aufgabenstellung ist ein Beispiel für eine **partielle Differenzialgleichung**, also eine Gleichung, in der partielle Ableitungen einer gesuchten Funktion auftreten. Im Folgenden werden wir vereinzelt beispielhaft auf partielle Differenzialgleichungen eingehen. Dem Spezialfall, dass die gesuchte Funktion nur von einer Variable abhängt, haben wir dagegen ein eigenes Kapitel gewidmet (Kapitel 5).

Wir können dieses Beispiel mit der Richtungsableitung umschreiben zu

$$\sqrt{a^2+b^2}\left(\frac{\partial f}{\partial x}(x,y), \frac{\partial f}{\partial y}(x,y)\right) \cdot \begin{pmatrix} \frac{a}{\sqrt{a^2+b^2}} \\ \frac{b}{\sqrt{a^2+b^2}} \end{pmatrix} = \sqrt{a^2+b^2}\frac{\partial f}{\partial \frac{1}{\sqrt{a^2+b^2}}(a,b)}(x,y) = 0.$$

Die Steigung einer Lösung ist an jeder Stelle in Richtung von (a, b) gleich null, die Lösung ist konstant auf allen Geraden mit Richtungsvektor (a, b). Diese Geraden heißen die **Charakteristiken** der Differenzialgleichung. Die Koordinatenform der Charakteristiken lautet $bx - ay = c$, wobei man für jede Konstante $c \in \mathbb{R}$ eine andere Gerade erhält. Da f auf diesen Geraden konstant ist, gibt es nur für jedes c einen anderen Funktionswert. Jede Lösung f hat damit die Gestalt $f(x, y) = g(bx - ay)$ für eine Funktion $g : \mathbb{R} \to \mathbb{R}$, die nur eine Variable hat. Genauer lässt sich g ohne weitere Anforderungen an die Lösung nicht bestimmen. So erfüllt beispielsweise $f(x, y) = \sin(bx - ay)$ aber auch $f(x, y) = \exp(bx - ay)$ die Differenzialgleichung. ■

4.3 Lokale und globale Extrema

Wir gehen hier kurz auf den Begriff **Spektralnorm** ein, der für numerische Rechenverfahren benötigt wird.

Wir benutzen die notwendige Bedingung für ein lokales Extremum, um den größten Eigenwert einer Matrix zu finden. Dieser kann z. B. beim Lösen eines Differenzialgleichungssystems (siehe Kapitel 5.3) verwendet werden. Ein Problem der Linearen Algebra wird mit Mitteln der Analysis gelöst.

Sei $\mathbf{A}$ eine invertierbare, reelle $(n \times n)$-Matrix. Mit $\vec{x} \in \mathbb{R}^n$ bezeichnen wir hier Spaltenvektoren. Wir suchen

$$\|\mathbf{A}\|_2 := \max\left\{\frac{|\mathbf{A}\vec{x}|}{|\vec{x}|} : 0 \neq \vec{x} \in \mathbb{R}^n\right\}.$$

Dabei ist wie zuvor

$$|\vec{x}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\vec{x} \cdot \vec{x}}$$

und entsprechend

$$|\mathbf{A}\vec{x}| = \sqrt{(\mathbf{A}\vec{x}) \cdot (\mathbf{A}\vec{x})} = \sqrt{\vec{x}^\top \mathbf{A}^\top \mathbf{A}\vec{x}}.$$

Man kann nachrechnen, dass die Menge nach oben beschränkt ist, so dass das Supremum existiert. Mit einem Stetigkeitsargument zeigt man, dass das Supremum tatsächlich auch als Maximum und damit als lokales Extremum der Funktion

$$f(\vec{x}) := \frac{|\mathbf{A}\vec{x}|}{|\vec{x}|} = \sqrt{\frac{\vec{x}^\top \mathbf{A}^\top \mathbf{A}\vec{x}}{\vec{x} \cdot \vec{x}}}$$

angenommen wird. Darauf gehen wir hier nicht weiter ein. Stattdessen benutzen wir die notwendige Bedingung $\operatorname{grad} f(\vec{x}) = \vec{0}$ für ein Extremum an einer Stelle $\vec{x}$. Dazu berechnen wir die partiellen Ableitungen. Mittels Kettenregel ist

$$\begin{aligned}\frac{\partial}{\partial x_k}\left[\vec{x}^\top \mathbf{A}^\top \mathbf{A}\vec{x}\right] &= \frac{\partial}{\partial x_k}|\mathbf{A}\vec{x}|^2 = \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\left[\sum_{l=1}^{n} a_{i,l}x_l\right]^2 = \sum_{i=1}^{n} 2\left[\sum_{l=1}^{n} a_{i,l}x_l\right] a_{i,k} \\ &= 2\sum_{i=1}^{n} a_{i,k}(A\vec{x})_i = 2(\mathbf{A}^\top \mathbf{A}\vec{x})_k.\end{aligned}$$

Damit und mit Ketten- sowie Quotientenregel erhalten wir:

$$\begin{aligned}\frac{\partial}{\partial x_k} f(\vec{x}) &= \frac{1}{2\sqrt{\frac{\vec{x}^\top \mathbf{A}^\top \mathbf{A}\vec{x}}{\vec{x}\cdot\vec{x}}}} \frac{\partial}{\partial x_k}\left[\frac{\vec{x}^\top \mathbf{A}^\top \mathbf{A}\vec{x}}{\vec{x} \cdot \vec{x}}\right] \\ &= \frac{1}{2\frac{|\mathbf{A}\vec{x}|}{|\vec{x}|}} \frac{2(\mathbf{A}^\top \mathbf{A}\vec{x})_k(\vec{x}\cdot\vec{x}) - \vec{x}^\top \mathbf{A}^\top \mathbf{A}\vec{x}\, 2x_k}{(\vec{x}\cdot\vec{x})^2} \\ &= \frac{|\vec{x}|}{|\mathbf{A}\vec{x}|} \frac{|\vec{x}|^2(\mathbf{A}^\top \mathbf{A}\vec{x})_k - |\mathbf{A}\vec{x}|^2 x_k}{|\vec{x}|^4} = \frac{1}{|\mathbf{A}\vec{x}||\vec{x}|}(\mathbf{A}^\top \mathbf{A}\vec{x})_k - \frac{|\mathbf{A}\vec{x}|}{|\vec{x}|^3}x_k.\end{aligned}$$

Aus der notwendigen Bedingung grad $f(\vec{x}) = \vec{0}$ für $\vec{x} \neq 0$ folgt

$$\mathbf{A}^\top \mathbf{A}\vec{x} - \left(\frac{|\mathbf{A}\vec{x}|}{|\vec{x}|}\right)^2 \vec{x} = \vec{0}.$$

Wird das Maximum an einer Stelle $\vec{x}_0$ angenommen, so ist die Stelle $\vec{x}_0$ ein Eigenvektor von $\mathbf{A}^\top \mathbf{A}$ zum Eigenwert $\left(\frac{|\mathbf{A}\vec{x}_0|}{|\vec{x}_0|}\right)^2 = \|\mathbf{A}\|_2^2$.

Hat man umgekehrt einen beliebigen Eigenvektor $\vec{y}$ von $\mathbf{A}^\top \mathbf{A}$ zum Eigenwert λ, so ist $\lambda \geq 0$ und insbesondere reell, denn

$$|\mathbf{A}\vec{y}|^2 = \vec{y}^\top \mathbf{A}^\top \mathbf{A}\vec{y} = \vec{y}^\top \lambda \vec{y} = \lambda |\vec{y}|^2,$$

also $\lambda = \frac{|\mathbf{A}\vec{y}|^2}{|\vec{y}|^2} \geq 0$.

Außerdem ist $\|\mathbf{A}\|_2 \geq \frac{|\mathbf{A}\vec{y}|}{|\vec{y}|} = \sqrt{\lambda}$. Damit ist also $\|\mathbf{A}\|_2^2$ größer oder gleich jedem Eigenwert. Da $\|\mathbf{A}\|_2^2$ mit dem Eigenwert des Eigenvektors $\vec{x}_0$ übereinstimmt, ist dieser Eigenwert $\|\mathbf{A}\|_2^2 = \left(\frac{|A\vec{x}_0|}{|\vec{x}_0|}\right)^2$ insbesondere der größte Eigenwert von $\mathbf{A}^\top \mathbf{A}$.

Die sogenannte **Spektralnorm** $\|\mathbf{A}\|_2$ ist gleich der Wurzel aus dem größten Eigenwert von $\mathbf{A}^\top \mathbf{A}$. Diese Matrix-Norm passt zum Betrag (Euklid'sche Norm) der Vektoren. Mit ihr kann man die Größe eines Vektors abschätzen, der durch Multiplikation mit einer Matrix entsteht. Für $\vec{x} \neq \vec{0}$ ist

$$|\mathbf{A}\vec{x}| = \frac{|\mathbf{A}\vec{x}|}{|\vec{x}|}|\vec{x}| \leq \|\mathbf{A}\|_2 |\vec{x}|.$$

Auch für $\vec{x} = \vec{0}$ gilt die Abschätzung, da auf beiden Seiten eine Null steht.

Im Falle einer reellen, symmetrischen Matrix $\mathbf{A}$, d. h. $\mathbf{A}^\top = \mathbf{A}$, sind die Eigenwerte von $\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}$ die quadrierten Eigenwerte von $\mathbf{A}$. In diesem Fall ist $\|\mathbf{A}\|_2$ gleich dem Betrag des betragsmäßig größten Eigenwertes von $\mathbf{A}$. Handelt es sich dabei um einen einfachen Eigenwert, sind also alle anderen Eigenwerte betragsmäßig echt kleiner, so kann man diesen leicht mit dem Computer ausrechnen. Dazu startet man mit einem Vektor $\vec{v}_0 \neq 0$ und berechnet sukzessive $\vec{v}_k := \mathbf{A}\vec{v}_{k-1}$ durch Multiplikation mit der Matrix $\mathbf{A}$. Bei geschickter Wahl von $\vec{v}_0$ konvergiert die Folge $(\vec{v}_k)_{k=0}^\infty$ gegen einen Eigenvektor zum betragsmäßig größten Eigenwert (siehe z. B. (Arens et al., 2012, S.614)).

4.4 Extrema unter Nebenbedingungen

Beispiel 4.4 (Lineare Optimierung)
Sucht man das Maximum einer stetig differenzierbaren Funktion $f(x, y)$ auf einer beschränkten Menge, deren Rand stückweise über eine geeignete Nebenbedingung $g(x, y) = 0$ beschrieben werden kann, so erhält man im Inneren der Menge Kandidaten über die Bedingung grad $f(x, y) = \vec{0}$. Auf jedem Stück des Randes ergeben sich

Kandidaten mit dem Satz über die Lagrange-Multiplikatoren. Maxima können nur an den Kandidatenstellen des Inneren und der Randstücke sowie an den Berührungspunkten der Randstücke liegen.

Betrachtet man für $c_1, c_2 \in \mathbb{R}$ die Funktion $f(x, y) = c_1 x + c_2 y$, so beschreibt f die Höhen einer Ebene im $\mathbb{R}^3$. Anschaulich ist klar, dass ein Maximum von f daher nur auf dem Rand angenommen werden kann. Ist der Rand aus Strecken zusammengesetzt (ein Polygonzug), so kann ein Maximum nur an den Ecken des Randes liegen. Das ist genau die Situation, die man bei der **Linearen Optimierung** zur Lösung vieler betriebswirtschaftlicher Probleme betrachtet. Hier benötigt man keine Lagrange-Multiplikatoren und muss lediglich die Ecken des Randes berücksichtigen. Allerdings weisen die Probleme häufig sehr viele Variablen auf, so dass dies nicht so einfach ist. ■

Beispiel 4.5 (Mehrdimensionales Newton-Verfahren)
Bei der Nullstellensuche mit dem eindimensionalen Newton-Verfahren (vgl. Beispiel 4.2 auf Seite 30) gelangt man von einer Stelle x_k zu einer Stelle x_{k+1}, die näher an einer Nullstelle liegen soll, indem man x_{k+1} als Nullstelle der Gerade $f(x_k) + f'(x_k)(x - x_k)$ sucht. Diese ist eine Taylor-Entwicklung von f an der Stelle x_k. Also ist x_{k+1} ein Minimum von $[f(x_k) + f'(x_k)(x - x_k)]^2$. Entsprechend kann man mit einer mehrdimensionalen Taylor-Entwicklung für $f : \mathbb{R}^n \to \mathbb{R}$ die Nullstellensuche gestalten: Minimiere $[f(\vec{x}_k) + \text{grad}\, f(\vec{x}_k) \cdot (\vec{x} - \vec{x}_k)]^2$. Diese Taylor-Entwicklung nähert die Funktion $f^2(\vec{x})$ nur sehr gut in der Nähe von $\vec{x}_k$ an. Statt ein Minimum auf $\mathbb{R}^n$ zu suchen, kann es daher besser sein, $\vec{x}_{k+1}$ als ein lokales Minimum auf einer Umgebung $\{\vec{x} \in \mathbb{R}^n : |\vec{x} - \vec{x}_k| \leq \delta(\vec{x}_k)\}$ für ein geeignetes $\delta(\vec{x}_k) > 0$ zu bestimmen und dann von dort aus weiterzusuchen. Ein solches Minimum findet man unter Verwendung eines Lagrange-Multiplikators wie zu Beginn des vorangehenden Beispiels beschrieben: Im Inneren der Umgebung ist die notwendige Bedingung $\text{grad}([f(\vec{x}_k) + \text{grad}\, f(\vec{x}_k)(\vec{x} - \vec{x}_k)]^2) = \vec{0}$, auf dem Rand lässt sich eine notwendige Bedingung mit einem Lagrange-Multiplikator über die Nebenbedingung $0 = g(\vec{x}) = |\vec{x} - \vec{x}_k|^2 - [\delta(\vec{x}_k)]^2$ formulieren. Damit existiert ein $\lambda \in \mathbb{R}$, so dass als notwendige Bedingung für ein Minimum in $\vec{x}$

$$2[f(\vec{x}_k) + \text{grad}\, f(\vec{x}_k)(\vec{x} - \vec{x}_k)] \,\text{grad}\, f(\vec{x}_k) + \lambda \cdot 2(\vec{x} - \vec{x}_k) = \vec{0}$$

gilt. Für $\lambda = 0$ ist der Fall eines Minimums im Inneren der Umgebung eingeschlossen. Der resultierende Algorithmus heißt **Levenberg-Marquardt-Verfahren**. ■

4.5 Kurvenintegrale

Beispiel 4.6
Für die Kurve K mit der Parametrisierung $(\vec{x}(t) := (\cos t, \sin t), [0, \pi])$, die einen mathematisch positiv durchlaufenen Halbkreis mit Radius 1 beschreibt, berechnen wir das

Kurvenintegral des Vektorfelds $\vec{V}(x_1, x_2) = (V_1(x_1, x_2), V_2(x_1, x_2))$ mit $V_1(x_1, x_2) := -x_2$, $V_2(x_1, x_2) := x_1$:

$$\begin{aligned}\int_K \vec{V}\, d\vec{x} &= \int_0^\pi V_1(\cos t, \sin t)(-\sin t) + V_2(\cos t, \sin t)\cos t\, dt \\ &= \int_0^\pi (-\sin t)(-\sin t) + \cos(t)\cos(t)\, dt = \int_0^\pi 1\, dt = \pi.\end{aligned}$$

■

Unter vernünftigen Voraussetzungen sind sind auf einfach zusammenhängenden Gebieten Wegunabhängigkeit und Wirbelfreiheit gleichbedeutend. Insbesondere hat ein wirbelfreies Feld $\vec{V}$ ein Potenzial φ, also $\vec{V} = \operatorname{grad}\varphi$. Ist das Feld zusätzlich quellenfrei, so ist $0 = \operatorname{div}\vec{V} = \operatorname{div}\operatorname{grad}\varphi = \Delta\varphi$ mit dem Laplace-Operator $\Delta := \sum_{k=1}^{3} \frac{\partial^2}{\partial x_k^2}$. Ein wirbel- und quellenfreies Feld hat damit ein Potenzial φ, das die **Laplace-Differenzialgleichung**

$$\Delta\varphi(x_1, x_2, x_3) = \frac{\partial^2\varphi}{\partial x_1^2}(x_1, x_2, x_3) + \frac{\partial^2\varphi}{\partial x_2^2}(x_1, x_2, x_3) + \frac{\partial^2\varphi}{\partial x_3^2}(x_1, x_2, x_3) = 0$$

erfüllt. Das Potenzial φ ist also eine Lösung dieser Gleichung, in der eine Bedingung an zweite partielle Ableitungen von φ gestellt werden. Hier handelt es sich also wieder um eine partielle Differenzialgleichung (s.o.). Ersetzt man die Null auf der rechten Seite durch eine andere Funktion mit den Variablen x_1, x_2, x_3, so bekommt diese Differenzialgleichung den Namen **Poisson-Gleichung** oder **Potenzialgleichung**.

Beispiel 4.7
Jetzt betrachten wir die Jordan-Kurve K mit der Parametrisierung $(\vec{x}(t) := (\cos t, \sin t, 0), [0, 2\pi])$ und das Vektorfeld $\vec{V}(x_1, x_2, x_3) = (-x_2, x_1, 0)$:

$$\int_K \vec{V}\, d\vec{x} = \int_0^{2\pi} \sin^2 t + \cos^2 t\, dt = 2\pi.$$

Hätten wir Wegunabhängigkeit, so müsste sich null ergeben. Tatsächlich ist auch die Rotation nicht $\vec{0}$: $\operatorname{rot}\vec{V}(x_1, x_2, x_3) = (0, 0, 2)$. ■

4.6 Satz von Green

Der Satz von Green wird bewiesen, indem man sich aus den Bedingungen an das Gebiet E eine geeignete Parametrisierung der Randkurve konstruiert und dann das Kurvenintegral mittels des Hauptsatzes der Differenzial- und Integralrechnung in das Integral über E überführt. In einem ganz einfachen Spezialfall wollen wir den Satz beweisen. Dazu betrachten wir ein Rechteck $E :=]a_1, b_1[\times]a_2, b_2[$ mit den gegenüberliegenden Eckpunkten (a_1, a_2) und (b_1, b_2). Die positiv durchlaufene Randkurve von E setzt sich zusammen aus den vier Kurven $((t, a_2), [a_1, b_1])$, $((b_1, t), [a_2, b_2])$,

$((a_1+b_1-t,b_2),[a_1,b_1])$ und $((a_1,a_2+b_2-t),[a_2,b_2])$. Dann gilt für ein stetig differenzierbares Vektorfeld $\vec{V}=(V_1,V_2)$ mit dem Satz von Fubini:

$$\begin{aligned}
&\iint_E \left[\frac{\partial V_2}{\partial x}(x,y)-\frac{\partial V_1}{\partial y}(x,y)\right] d(x,y)\\
&= \int_{a_2}^{b_2}\int_{a_1}^{b_1}\frac{\partial V_2}{\partial x}(x,y)\,dx\,dy-\int_{a_1}^{b_1}\int_{a_2}^{b_2}\frac{\partial V_1}{\partial y}(x,y)\,dy\,dx\\
&= \int_{a_2}^{b_2}[V_2(b_1,y)-V_2(a_1,y)]\,dy-\int_{a_1}^{b_1}[V_1(x,b_2)-V_1(x,a_2)]\,dx\\
&= \int_{a_1}^{b_1}V_1(x,a_2)\,dx+\int_{a_2}^{b_2}V_2(b_1,y)\,dy+\int_{a_1}^{b_1}V_1(x,b_2)\cdot(-1)\,dx+\int_{a_2}^{b_2}V_2(a_1,y)\cdot(-1)\,dy\\
&= \int_{\partial E}\vec{V}\,d\vec{x},
\end{aligned}$$

da wir im vorletzten Schritt genau die Kurvenintegrale entlag den oben angegebenen vier Randkurven erhalten haben.

Beispiel 4.8
Mit dem Satz von Green können wir den Flächeninhalt eines zweidimensionalen Normalbereichs E über ein Kurvenintegral berechnen. Dazu betrachten wir das Vektorfeld $\vec{V}(x,y)=(V_1(x,y),V_2(x,y))=(0,x)$ mit $\frac{\partial V_2}{\partial x}(x,y)-\frac{\partial V_1}{\partial y}(x,y)=1$ und erhalten

$$\iint_E 1\,d(x,y)=\int_{\partial E}\vec{V}\,d\vec{x}.$$

Beispielsweise ergibt sich so die Fläche für $E=[0,2]\times[3,6]$ mit einer Randkurve, die aus vier Strecken im Gegenuhrzeigersinn zusammengesetzt ist, zu

$$\iint_{[0,2]\times[3,6]}1\,d(x,y)=\int_0^2 0\cdot1+t\cdot0\,dt+\int_3^6 0\cdot0+2\cdot1\,dt+\int_2^0 0\cdot1+t\cdot0\,dt+\int_6^3 0\cdot0+0\cdot1\,dt=6.$$

■

4.7 Flächenintegrale

Um die Definition der Oberflächenintegrals genauer zu verstehen, betrachten wir ein Parameterintervall $S=]a_1,b_1[\times]a_2,b_2[$ mit einer Parameterfunktion $\vec{F}(u,v)=(u,v,z(u,v))$. Die Fläche ist der Funktionsgraph der Funktion $z(u,v)$ und damit besonders anschaulich. Zu jedem $x=u$ und $y=v$ wird also eine Höhe $z(u,v)$ berechnet. Über jedem Teilintervall

$$\Big[\underbrace{a_1+(i-1)\frac{b_1-a_1}{n}}_{=:u_{i-1}},\underbrace{a_1+i\frac{b_1-a_1}{n}}_{=u_i}\Big]\times\Big[\underbrace{a_2+(k-1)\frac{b_2-a_2}{m}}_{=:v_{k-1}},\underbrace{a_2+k\frac{b_2-a_2}{m}}_{=v_k}\Big],$$

$1 \le i \le n$, $1 \le k \le m$, hat die Fläche näherungsweise den Flächeninhalt des von den beiden Vektoren

$$\left(\frac{b_1 - a_1}{n}, 0, \frac{b_1 - a_1}{n}\frac{\partial z}{\partial u}(u_{i-1}, v_{k-1})\right) = \frac{b_1 - a_1}{n}\frac{\partial \vec{F}}{\partial u}(u_{i-1}, v_{k-1})$$

und

$$\left(0, \frac{b_2 - a_2}{m}, \frac{b_2 - a_2}{m}\frac{\partial z}{\partial v}(u_{i-1}, v_{k-1})\right) = \frac{b_2 - a_2}{m}\frac{\partial \vec{F}}{\partial v}(u_{i-1}, v_{k-1})$$

aufgespannten Parallelogramms. Multiplizieren wir auf jedem Teilintervall einen Funktionswert von g (an der linken unteren Ecke (u_{i-1}, v_{k-1})) mit dem über das Vektorprodukt berechneten Flächeninhalt und summieren über alle Teilintervalle, so erhalten wir die Zahl

$$\frac{(b_1 - a_1)\cdot(b_2 - a_2)}{n \cdot m}\sum_{i=1}^{n}\sum_{k=1}^{m} g\left(\vec{F}(u_{i-1}, v_{k-1})\right)\cdot\left|\frac{\partial \vec{F}}{\partial u}(u_{i-1}, v_{k-1}) \times \frac{\partial \vec{F}}{\partial v}(u_{i-1}, v_{k-1})\right|.$$

Lassen wir n und m gegen unendlich streben und damit die Zerlegung des Intervalls immer feiner werden, so erhalten wir das Oberflächenintegral aus der Definition.

4.8 Satz von Gauß, der Divergenzsatz

Mit dem Satz von Gauß erhalten wir auch eine neue Sicht auf die Laplace- oder Poisson-Differenzialgleichung (s.o.), die in der Theorie partieller Differenzialgleichungen als **Green'sche Formel** (und nicht, wie zu erwarten wäre, als Gauß'sche Formel) bekannt ist. Damit erhalten wir gleichzeitig neue Einblicke über Potenziale als Lösung einer Laplace-Gleichung, falls das Feld quellenfrei ist, oder sonst als Lösung einer Poisson-Gleichung. Wählen wir im Satz von Gauß die Funktion $\vec{V}(x, y, z) = \operatorname{grad} f(x, y, z) = \nabla f(x, y, z)$ für eine geeignete reellwertige Funktion f (z. B. wäre mit dem elektrischen Potenzial φ die Wahl $f := -\varphi$ möglich, und die elektrische Feldstärke wäre $\vec{E} = \vec{V} = -\operatorname{grad}\varphi$), so ergibt sich die Green'sche Formel

$$\begin{aligned}
\iiint_E \Delta f(x, y, z)\, d(x, y, z) &= \iiint_E \operatorname{div}\operatorname{grad} f(x, y, z)\, d(x, y, z) \\
&= \iiint_E \operatorname{div}\vec{V}(x, y, z)\, d(x, y, z) \overset{\text{Gauß}}{=} \int_{\partial E} \vec{V}\cdot\vec{N}\, d\sigma \\
&= \int_{\partial E} \operatorname{grad} f \cdot \vec{N}\, d\sigma = \int_{\partial E} \frac{\partial f}{\partial \vec{N}}\, d\sigma.
\end{aligned}$$

Bei der Laplace-Gleichung ist eine Funktion f gesucht, bei der $\Delta f(x, y, z) = 0$ ist. Man nennt eine solche Funktion eine **harmonische Funktion**. Für diese verschwindet insbesondere das Integral auf der linken Seite. Zum Beispiel ist das Potenzial eines wirbel- und quellenfreien Feldes eine harmonische Funktion. Hier muss notwendiger

Weise auf dem Rand von E das Integral $\int_{\partial E} \frac{\partial f}{\partial \vec{N}}\, d\sigma$ der Richtungsableitung von f (des Potenzials) in Richtung der äußeren Normalen gleich null sein. Hat man allgemeiner die Poisson-Gleichung $\Delta f(x,y,z) = g(x,y,z)$, so muss jede Lösung f auch $\int_{\partial E} \frac{\partial f}{\partial \vec{N}}\, d\sigma = \iiint_E g(x,y,z)\, d(x,y,z)$ erfüllen. Üblicher Weise sucht man hier eine Lösung f der Differenzialgleichung, die auf dem Rand von E einer vorgegebenen Bedingung (Randbedingung) genügt, z. B. die **Neumann-Randbedingung** $\frac{\partial f}{\partial \vec{N}} = h(x,y,z)$ auf ∂E. Damit es dann überhaupt eine Lösung geben kann, muss die vorgegebene Funktion h die Gleichung $\int_{\partial E} h\, d\sigma = \iiint_E g(x,y,z)\, d(x,y,z)$ erfüllen. Die vorgegebenen Daten (Funktionen) g und h müssen also zueinander passen.

Mit der Green'schen Formel kann man zeigen, dass Eigenschaften einer Lösungsfunktion im Inneren des Gebiets E durch Eigenschaften auf dem Rand festgelegt sind, z. B. wird ein Maximum einer Lösung f der Laplace-Gleichung auf dem Rand angenommen (**Maximumprinzip**). Damit lässt sich dann auch die Eindeutigkeit einer Lösung bei vorgegebener Randbedingung nachweisen.

Mit einer Erweiterung der Green'schen Formel findet man auch eine in der Praxis wichtige Darstellung der Lösungen der Poisson-Gleichung. Ist eine Funktion f gesucht, die auf einem geeigneten Gebiet E die Gleichung $\Delta f(x,y,z) = g(x,y,z)$ erfüllt und für die gleichzeitig die **Dirichlet-Randbedingung** $f(x,y,z) = h(x,y,z)$ auf dem Rand ∂E für eine dort vorgegebene Funktion h gilt, dann hat die Lösung f an jeder Stelle $\vec{x}_0 \in E$ die Darstellung

$$f(\vec{x}_0) = \int_{\partial E} h(\vec{x}) \frac{\partial G(\vec{x}, \vec{x_0})}{\partial \vec{N}(\vec{x})}\, d\sigma + \iiint_E g(\vec{x}) G(\vec{x}, \vec{x}_0)\, d\vec{x}.$$

Hier ist $\vec{x}_0$ eine feste Stelle, an der der Wert der Lösung berechnet wird, und $\vec{x}$ ist jeweils die Integrationsvariable. $G(\vec{x}, \vec{x_0})$ ist für jede Stelle $\vec{x}_0$ eine Funktion in der Variable $\vec{x}$. Eine solche Funktion, die aus den vorgegebenen Daten die Lösung generiert, existiert und heißt **Green'sche Funktion**. Neben der Stelle $\vec{x}_0$ hängt sie nur vom Gebiet E ab. Für gängige Gebiete ist die Green'sche Funktion bekannt, so dass man die Lösung der Differenzialgleichung über die beiden Integrale berechnen kann. Ist beispielsweise E eine Kugel mit Radius r um den Nullpunkt, so ist für $\vec{x}_0 \neq \vec{0}$ (siehe (Strauss, 1995, S. 198)):

$$G(\vec{x}, \vec{x}_0) = \frac{1}{4\pi} \left[-\frac{1}{|\vec{x} - \vec{x}_0|} + \frac{r}{|\vec{x}_0| \left| \vec{x} - \frac{r^2 \vec{x}_0}{|\vec{x}_0|^2} \right|} \right].$$

Hier hat man allerdings die Schwierigkeit, dass die Funktion in jeder Umgebung von $\vec{x}_0$ unbeschränkt ist, so dass $\iiint_E f(\vec{x}) G(\vec{x}, \vec{x}_0)\, d\vec{x}$ nur sinnvoll ist, wenn man wie bei uneigentlichen Integralen eine Kugel um die Stelle $\vec{x}_0$ ausspart und deren Radius gegen null gehen lässt. Da $\vec{x}_0 \in E$ ist, gilt $|\vec{x}_0| < r$. Daher ist $\left| \frac{r^2 \vec{x}_0}{|\vec{x}_0|^2} \right| = \frac{r^2}{|\vec{x}_0|} > r$, so dass der Punkt $\frac{r^2 \vec{x}_0}{|\vec{x}_0|^2}$ nicht im Gebiet liegt und keine Schwierigkeiten bei der Integration macht.

Die Green'sche Funktion kann interpretiert werden als ein elektrisches Potenzial, wobei nur an der Stelle $\vec{x}_0$ eine Punktladung vorliegt.

5 Zusatzmaterial zu Kapitel 5

Übersicht

5.1 Numerische Lösung von Differenzialgleichungen

Hintergrund: Differenzenverfahren und Finite Elemente

Mit der Euler-Cauchy-Polygonzugmethode und dem Iterationsverfahren von Picard und Lindelöf haben wir bereits Ansätze gesehen, mit denen wir Differenzialgleichungen näherungsweise (numerisch) lösen können. Wir betrachten hier ganz kurz zwei gängige andere Verfahren. Dazu wählen wir beispielhaft eine sehr einfache Randwertaufgabe: Gesucht ist eine Funktion $y : [0, 1] \to \mathbb{R}$ mit $y(0) = y(1) = 0$ (Randbedingung) und

$$y''(x) = f(x), \quad x \in [0, 1].$$

Für eine stetige Inhomogenität erhält man die Lösung exakt durch doppelte Integration:

$$y(x) = \int_0^x \left[\int_0^t f(u)\,du \right] dt - x \int_0^1 \left[\int_0^t f(u)\,du \right] dt.$$

Hier sorgt der rechte Term für die Einhaltung der Randbedingung.

Der vielleicht offensichtlichste Ansatz zur Bestimmung der Funktionswerte von y an endlich vielen Stellen des Intervalls besteht darin, dass wir die Ableitungen durch Differenzenquotienten annähern. Wir möchten dabei y an den $n + 1$ Stellen $t_k = \frac{k}{n}$, $k \in \{0, 1, \ldots, n\}$ berechnen. Für $k \in \{1, \ldots, n - 1\}$ ist

$$\begin{aligned} y''(t_k) &\approx \frac{y'(t_{k+1}) - y(t_k)'}{t_{k+1} - t_k} = n[y'(t_{k+1}) - y(t_k)'] \\ &\approx n\left[\frac{y(t_{k+1}) - y(t_k)}{t_{k+1} - t_k} - \frac{y(t_k) - y(t_{k-1})}{t_k - t_{k-1}} \right] = n^2[y(t_{k+1}) - 2y(t_k) + y(t_{k-1})]. \end{aligned}$$

Damit hat man ein eindeutig lösbares Gleichungssystem mit zwei Gleichungen $y(t_0) = 0$ und $y(t_n) = 0$ aus den Randbedingungen sowie $n - 1$ Gleichungen

$$n^2[y(t_{k+1}) - 2y(t_k) + y(t_{k-1})] = f(t_k).$$

Rechenverfahren, bei denen man die Ableitungen so durch Differenzenquotienten ersetzt, heißen **Differenzenverfahren**.

Betrachtet man partielle Differenzialgleichungen auf komplizierteren Gebieten als Intervalle, so gibt es mit den Differenzen an den Rändern Probleme, da man vielfach den Rand nicht genau mit äquidistanten Zerlegungspunkten treffen kann. Für Randwertprobleme hat sich hier ein anderes Verfahren durchgesetzt, das wir uns für die gleiche gewöhnliche Differenzialgleichung anschauen:

Bei der **Finite-Elemente-Methode** überführt man das Randwertproblem mittels partieller Integration in ein „schwaches Problem“, das dann über einem endlich dimensionalen Vektorraum von Funktionen („Ansatzfunktionen“) gelöst wird. Dazu muss nur ein lineares Gleichungssystem gelöst werden.

Zunächst multiplizieren wir beide Seiten der Differenzialgleichung mit einer stetig differenzierbaren Ansatzfunktion v, die die Randbedingung $v(0) = v(1) = 0$ erfüllt, und integrieren dann partiell:

$$\begin{aligned} y''(x) = f(x) \quad &\Longrightarrow \quad y''(x)v(x) = f(x)v(x) \Longrightarrow \int_0^1 y''(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx \\ &\Longrightarrow \quad [y'(x)v(x)]_0^1 - \int_0^1 y'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx \\ &\Longrightarrow \quad \int_0^1 y'(x)v'(x)\,dx = -\int_0^1 f(x)v(x)\,dx. \end{aligned} \tag{5.1}$$

Erfüllt y die Differenzialgleichung, so erfüllt y auch das „schwache Problem“ (5.1) für jede Ansatzfunktion v. Wir zerlegen nun das Intervall $[0, 1]$ wieder in n Teilintervalle $\left[\frac{k-1}{n}, \frac{k}{n}\right]$, $k \in \{1, 2, \ldots, n\}$ und betrachten stetige Ansatzfunktionen, die die Randbedingung erfüllen und auf jedem Teilintervall einer Geraden $a_k x + b_k$ entsprechen. Diese Funktionen dürfen Knicke an den Stellen $\frac{k}{n}$ haben, die einzelnen Geradenstücke müssen aber stetig zusammenpassen.

Jetzt suchen wir eine Näherungslösung y_n aus der Menge dieser Ansatzfunktionen, die für alle Ansatzfunktionen v das schwache Problem löst. Statt die exakte Lösung zu bestimmen, begnügen wir uns also mit einer Näherungslösung, die aus Geradenstücken zusammengesetzt ist. Je mehr Geradenstücke verwendet werden, je feiner also die Zerlegung wird, desto genauer entspricht die Näherungslösung der exakten Lösung.

Dass y_n an den endlich vielen Zerlegungsstellen nicht differenzierbar ist, spielt bei der Berechnung des Integrals keine Rolle. Auf jedem Teilintervall sind die Ableitungen der Ansatzfunktionen konstant, so dass für jede Wahl von $v(x)$ eine Gleichung für diese Konstanten entsteht.

Ist $y_n(x) = a_k x + b_k$ und $v(x) = c_k x + d_k$ auf $\left[\frac{k-1}{n}, \frac{k}{n}\right]$, so gilt:

$$\int_{\frac{k-1}{n}}^{\frac{k}{n}} y'(x)v'(x)\,dx = \int_{\frac{k-1}{n}}^{\frac{k}{n}} a_k c_k\,dx = \frac{c_k}{n} \cdot a_k.$$

Für eine Funktion $v(x)$ erhalten wir also die Gleichung

$$\sum_{k=1}^{n} \frac{c_k}{n} \cdot a_k = -\int_0^1 f(x)v(x)\,dx.$$

Unter Verwendung von n (linear unabhängigen) Ansatzfunktionen $v(x)$ entsteht so ein Gleichungssystem, dessen Lösung die Koeffizienten a_k der gesuchten Näherungslösung liefert. Die Konstanten b_k ergeben sich aus Randbedingung und Stetigkeit.

Bei partiellen Differenzialgleichungen mit zwei Variablen muss man ein Gebiet im $\mathbb{R}^2$ in einfache Teilmengen zerlegen. Dabei verwendet man häufig Dreiecke (Triangulierung). Die Zerlegung in endlich viele einfache Teilmengen führt zum Namen **Finite Elemente**.

5.2 Partielle Differenzialgleichungen

Hintergrund: Charakterisierung partieller Differenzialgleichungen

Im Rahmen von Beispielen werden im Buch die Schwingungsgleichung und oben die Laplace- und Poisson-Differenzialgleichung erwähnt. Hier handelt es sich um partielle (und nicht um geöhnliche) Differenzialgleichungen. Besonders häufig stößt man auf Gleichungen mit höchstens zweiten partiellen Ableitungen:

Eine homogene lineare partielle Differenzialgleichung zweiter Ordnung mit konstanten Koeffizienten $a, b, c, d, e, f \in \mathbb{R}$ für Funktionen mit zwei Variablen hat die allgemeine Gestalt

$$a\frac{\partial^2 u(x,y)}{\partial x^2} + b\frac{\partial^2 u(x,y)}{\partial x \partial y} + c\frac{\partial^2 u(x,y)}{\partial y^2} + d\frac{\partial u(x,y)}{\partial x} + e\frac{\partial u(x,y)}{\partial y} + fu(x,y) = 0.$$

Dabei ist eine Funktion $u(x,y)$ gesucht, so dass für alle Punkte (x,y) eines Gebiets $G \subset \mathbb{R}^2$ die Gleichung erfüllt ist.

Bei stetigen partiellen Ableitungen bis zur Ordnung zwei sagt der Satz von Schwarz, dass die Reihenfolge der Ableitungen vertauscht werden darf. Damit können wir den zweiten Summanden auch so schreiben:

$$b\frac{\partial^2 u(x,y)}{\partial x \partial y} = \frac{b}{2}\frac{\partial^2 u(x,y)}{\partial x \partial y} + \frac{b}{2}\frac{\partial^2 u(x,y)}{\partial y \partial x}.$$

Wir schreiben jetzt die Koeffizienten der zweiten Ableitungen in eine Matrix:

$$\mathbf{A} := \begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{bmatrix}, \quad \det \mathbf{A} = ac - \frac{b^2}{4}.$$

Die Matrix ist symmetrisch und wird für eine Typ-Einteilung verwendet. Dabei werden geometrische Begriffe verwendet. Falls

- $\det \mathbf{A} > 0$ ist ($ac > \frac{b^2}{4} > 0$), so heißt die Gleichung **elliptisch**. Durch Übergang zu neuen Variablen r und s (Substitution) kann man eine Differenzialgleichung ohne gemischte partielle Ableitungen gewinnen:

$$\frac{\partial^2 v(r,s)}{\partial^2 r} + \frac{\partial^2 v(r,s)}{\partial^2 s} + \cdots = 0.$$

  Der Term mit den zweiten Ableitungen erinnert an die linke Seite der Gleichung $x^2 + y^2 = 1$, die einen Kreis, also eine spezielle Ellipse, beschreibt.
- $\det \mathbf{A} < 0$ ist ($ac < \frac{b^2}{4}$), so heißt die Gleichung **hyperbolisch**. Eine solche Gleichung kann man in die folgende Form ohne gemischte Ableitungen bringen:

$$\frac{\partial^2 v(r,s)}{\partial^2 r} - \frac{\partial^2 v(r,s)}{\partial^2 s} + \cdots = 0.$$

  Der Term mit den zweiten Ableitungen erinnert jetzt die linke Seite der Gleichung $x^2 - y^2 = 1$, die eine Hyperbel als Lösungsmenge hat.
- $\det \mathbf{A} = 0$ ist ($ac = \frac{b^2}{4}$), so heißt die Gleichung **parabolisch**, und man erhält eine Darstellung mit nur einer partiellen Ableitung zweiter Ordnung:

$$\frac{\partial^2 v(r,s)}{\partial^2 r} + \cdots = 0.$$

Eine entsprechende Klassifikation ist auch bei mehr Variablen über Bedingungen an die Eigenwerte der entsprechenden Matrix möglich.

Bei der (zweidimensionalen) Poisson-Gleichung ist $a = c = 1$ und $b = 0$. Damit ist $\det \mathbf{A} > 0$, es handelt sich um eine elliptische Gleichung. Elliptische Differenzialgleichungen werden mit einer Randbedingung versehen. In den Anwendungen beschreiben sie zeitunabhängige Probleme wie z. B. das Höhenprofil einer Membran, auf die in jedem Raumpunkt eine feste Kraft wirkt. Über sie werden oft Zustände minimaler Energie beschrieben.

Bei der Schwingungsgleichung für eine Saite ist a eine Konstante größer null, $c = -1$ und ebenfalls $b = 0$, so dass $\det \mathbf{A} < 0$ ist. Es handelt sich also um eine hyperbolische Differenzialgleichung. Diese beschreiben in der Regel zeitabhängige Phänomene und besitzen üblicher Weise neben Randbedingungen eine Anfangsbedingung.

Parabolische Gleichungen findet man ebenfalls mit einer Mischung aus Anfangs- und Randbedingungen. Hier wird der Situation bei elliptischen Gleichungen noch eine zeitliche Dimension hinzugefügt. Das ist typisch für Wärmeleitung und Diffusionsprozesse.

6 Zusatzmaterial zu Kapitel 6

Übersicht

6.1 Komplexwertige Funktionen und Fourier-Koeffizienten

Nachdem wir im Buch nachgerechnet haben, dass die Fourier-Koeffizienten tatsächlich dafür sorgen, dass die Fourier-Partialsummen eine 2π-periodische Funktion f möglichst gut annähern, sehen wir uns noch kurz einen anderen Zugang zur Erklärung der Formeln für die Fourier-Koeffizienten an. Dazu nehmen wir an, dass die Fourier-Partialsummen so gegen die Ausgangsfunktion f streben, dass wir ein Integral über f schreiben dürfen als Grenzwert über Integrale der Partialsummen. Das ist z. B. erlaubt, wenn die Partialsummen gleichmäßig gegen f konvergieren. Zwar ist das in vielen Situationen nicht der Fall, aber wir nehmen es jetzt trotzdem einmal an. Dann gilt für $l \in \mathbb{N}_0$:

$$
\begin{aligned}
\int_{-\pi}^{\pi} f(t)\cdot\cos(lt)\,dt &= \int_{-\pi}^{\pi}\left[a_0+\sum_{k=1}^{\infty}(a_k\cos(kt)+b_k\sin(kt))\right]\cdot\cos(lt)\,dt\\
&= \int_{-\pi}^{\pi}\lim_{n\to\infty}\left[a_0+\sum_{k=1}^{n}(a_k\cos(kt)+b_k\sin(kt))\right]\cdot\cos(lt)\,dt\\
&= \lim_{n\to\infty}\int_{-\pi}^{\pi}\left[a_0+\sum_{k=1}^{n}(a_k\cos(kt)+b_k\sin(kt))\right]\cdot\cos(lt)\,dt\\
&= \int_{-\pi}^{\pi}a_0\cos(lt)\,dt+\lim_{n\to\infty}\sum_{k=1}^{n}\left[a_k\int_{-\pi}^{\pi}\cos(kt)\cos(lt)\,dt+b_k\int_{-\pi}^{\pi}\sin(kt)\cos(lt)\,dt\right]
\end{aligned}
$$

$$= \int_{-\pi}^{\pi} a_0 \cos(lt)\,dt + \lim_{n\to\infty} \sum_{k=1}^{n} a_k \int_{-\pi}^{\pi} \cos(kt)\cos(lt)\,dt.$$

Beim vorletzten Gleichheitszeichen wird der Grenzwert mit dem Integral vertauscht, was wie oben beschrieben nur unter zusätzlichen Voraussetzungen erlaubt ist.

Falls $l = 0$ ist, ergibt sich $\int_{-\pi}^{\pi} f(t)\,dt = \int_{-\pi}^{\pi} a_0\,dt = 2\pi a_0$, so dass wir wieder die Definition des Fourier-Koeffizienten a_0 erkennen. Für $l > 0$ werden nach der Orthogonalität der trigonometrischen Funktionen alle bis auf einen Summanden null und $\int_{-\pi}^{\pi} f(t) \cdot \cos(lt)\,dt = a_l\pi$. Analog erhält man die Darstellung der Sinus-Koeffizienten b_l durch Multiplikation der Reihe mit $\sin(lt)$ und anschließender Integration. Wenn die Fourier-Reihe die Ausgangsfunktion ergeben soll, dann müssen die Fourier-Koeffizienten genau so aussehen wie von uns definiert.

6.2 Fourier-Transformation

Als Beispiel berechnen wir die Fourier-Transformierte der Funktion $f(t) = \cos^{2n} t$ für $t \in [-\pi/2, \pi/2]$ und $f(t) = 0$ für $|t| > \pi/2$. Funktionen dieses Typs eignen sich als Fensterfunktionen, so erhält man bis auf Faktoren für $n = 1$ das Hann-Fenster (s. u.). Da f gerade ist, ist die Fourier-Transformierte reell.

Für $n = 1$ ist mittels partieller Integration und $\sin^2(u) + \cos^2(u) = 1$ für $v \neq 0$ und $v \neq 2$:

$$\begin{aligned}
f^\wedge(v) &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2(u)\cos(vu)\,du \\
&= \left[\frac{1}{v}\cos^2(u)\sin(vu)\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} + \frac{1}{v}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2\cos(u)\sin(u)\sin(vu)\,du \\
&= 0 - \left[\frac{1}{v^2}2\cos(u)\sin(u)\cos(vu)\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} + \frac{1}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} [-2\sin^2(u) + 2\cos^2(u)]\cos(vu)\,du \\
&= \frac{4}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2(u)\cos(vu)\,du - \frac{1}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2\cos(vu)\,du \\
&= \frac{4}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2(u)\cos(vu)\,du - \frac{4}{v^3}\sin\left(\frac{\pi}{2}v\right).
\end{aligned}$$

Damit ist

$$f^\wedge(v) = -\frac{1}{1-\frac{4}{v^2}}\frac{4}{v^3}\sin\left(\frac{\pi}{2}v\right) = \frac{2}{4-v^2}\frac{2}{v}\sin\left(\frac{\pi}{2}v\right).$$

Für $n > 1$ erhalten wir analog mit partieller Integration für $v \neq 2k$, $0 \leq k \leq n$:

$$f^\wedge(v) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2n}(u)\cos(vu)\,du$$

$$
\begin{aligned}
&= 0 + \frac{1}{v}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2n\cos^{2n-1}(u)\sin(u)\sin(vu)\,du \\
&= 0 + \frac{1}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos(vu)\left[2n(2n-1)\cos^{2n-2}(u)[-\sin^2(u)] + 2n\cos^{2n-1+1}(u)\right]du \\
&= \frac{2n}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos^{2n}(u)\cos(vu)\,du - \frac{2n(2n-1)}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos(vu)\cos^{2n-2}(u)[1-\cos^2(u)]\,du \\
&= \left[\frac{2n}{v^2} + \frac{2n(2n-1)}{v^2}\right]\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos^{2n}(u)\cos(vu)\,du \\
&\quad -\frac{2n(2n-1)}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos^{2n-2}(u)\cos(vu)\,du.
\end{aligned}
$$

Wir erhalten die Rekursionsformel:

$$f^\wedge(v) = -\frac{v^2}{v^2-4n^2}\frac{2n(2n-1)}{v^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos^{2n-2}(u)\cos(vu)\,du$$

und damit

$$f^\wedge(v) = \frac{(2n)!}{\prod_{k=1}^{n}((2k)^2-v^2)}\frac{2}{v}\sin\left(\frac{\pi}{2}v\right).$$

Da die Fourier-Transformierte stetig ist, ergeben sich ihre Werte an den Stellen $v = 2m$, $0 \le m \le n$, durch stetige Ergänzung mit dem Satz von L'Hospital:

$$f^\wedge(0) = \frac{(2n)!}{\prod_{k=1}^{n}(2k)^2}\lim_{v\to 0}\frac{2}{v}\sin\left(\frac{\pi}{2}v\right) = \frac{(2n)!}{4^n(n!)^2}\pi,$$

und für $m > 0$ ist

$$
\begin{aligned}
f^\wedge(2m) &= \frac{(2n)!}{4^{n-1}\prod_{k\in\{1,2,\ldots,n\}\setminus\{m\}}(k^2-m^2)}\frac{\pi}{2m}\frac{\cos(\pi m)}{-4m} \\
&= \frac{(-1)^{m+1}(2n)!}{4^n\prod_{k\in\{1,2,\ldots,n\}\setminus\{m\}}(k^2-m^2)}\frac{\pi}{2m^2}.
\end{aligned}
$$

Hintergrund: Mehrdimensionale Fourier-Reihen und Fourier-Transformation

Die Theorie der Fourier-Reihen lässt sich direkt auf Funktionen mit mehreren Variablen übertragen. Ist $f : \mathbb{R}^n \to \mathbb{C}$ eine Funktion, die in jeder der n Variablen 2π-periodisch ist, so lässt sich diese unter geeigneten Voraussetzungen wie in einer Dimension als **Fourier-Reihe** schreiben:

$$f(\vec{t}) = \sum_{\vec{k}\in\mathbb{Z}^n} f^\wedge(\vec{k})e^{j\vec{k}\cdot\vec{t}}.$$

Dabei ist $\vec{k}\cdot\vec{t}$ das Skalarprodukt $k_1t_1+\cdots+k_nt_n$. Die **Fourier-Koeffizienten** sind hier

$$f^\wedge(\vec{k}\,) := \frac{1}{(2\pi)^n}\int_{[0,2\pi]\times\cdots\times[0,2\pi]} f(\vec{t})e^{-j\vec{k}\cdot\vec{t}}\,d\vec{t}.$$

In der Fourier-Reihe wird über $\mathbb{Z}^n$ summiert. Falls die Reihe absolut konvergiert, spielt es keine Rolle, in welcher Reihenfolge die abzählbare Menge $\mathbb{Z}^n$ durchlaufen wird. Anderenfalls wirkt sich die Reihenfolge aber auf das Konvergenzverhalten aus. Üblich sind beispielsweise radiale Partialsummen

$$\sum_{\vec{k}\in\mathbb{Z}^n} f^\wedge(\vec{k})e^{j\vec{k}\cdot\vec{t}} := \lim_{\rho\to\infty}\sum_{\vec{k}\in\mathbb{Z}^n \text{ mit } |\vec{k}|<\rho} f^\wedge(\vec{k})e^{j\vec{k}\cdot\vec{t}}$$

oder quadratische Partialsummen

$$\sum_{\vec{k}\in\mathbb{Z}^n} f^\wedge(\vec{k})e^{j\vec{k}\cdot\vec{t}} := \lim_{\rho\to\infty}\sum_{\vec{k}\in\mathbb{Z}^n \text{ mit } |k_1|<\rho,\ldots,|k_n|<\rho} f^\wedge(\vec{k})e^{j\vec{k}\cdot\vec{t}}.$$

Beide Definitionen können sich unterschiedlich verhalten. Diese Schwierigkeiten gibt es in einer Dimension nicht.

Auch die Fourier-Transformation kann auf Funktionen mit mehreren Variablen ausgedehnt werden. Sei $f:\mathbb{R}^n\to\mathbb{C}$ eine Funktion, für die

$$\int_{\mathbb{R}^n}|f(\vec{t})|\,dt := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}|f(t_1,t_2,\ldots,t_n)|\,dt_1\ldots dt_{n-1}dt_n$$

als reelle Zahl existiert. Hier setzen wir also bereits eine absolute Konvergenz voraus, so dass wir nicht wie bei den Fourier-Reihen verschiedene Grenzwerte betrachten müssen. Damit ist die **Fourier-Transformation**

$$f^\wedge(\vec{\omega}\,) := \int_{\mathbb{R}^n} f(\vec{t})e^{j\vec{t}\cdot\vec{\omega}}\,d\vec{t}$$

wohldefiniert und hat als Verallgemeinerung der eindimensionalen Fourier-Transformation vergleichbare Eigenschaften. Ist insbesondere f zerfallend, d. h. $f(t_1,t_2,\ldots,t_n) = f_1(t_1)\cdot f_2(t_2)\cdots f_n(t_n)$, so wird aus der mehrdimensionalen Fourier-Transformation das Produkt eindimensionaler:

$$f^\wedge(\vec{\omega}\,) = f_1^\wedge(\omega_1)\cdot f_2^\wedge(\omega_2)\cdots f_n^\wedge(\omega_n).$$

Das gilt aber nur in diesem Spezialfall. Weiterführende Literatur: z. B. (Zygmund, 2002, Teil 2, Kap. 17) und (Stein und Shakarchi, 2003, Kap. 6).

6.3 Diskrete Fourier-Transformation

Hintergrund: Zweidimensionale diskrete Fourier-Transformation

Möchten wir Bilddaten transformieren, so liegt kein Vektor sondern eine Matrix mit Helligkeitsinformationen der Bildpunkte vor. Eine Matrix von Funktionswerten haben wir auch, wenn wir zweidimensionale Fourier-Koeffizienten oder eine zweidimensionale Fourier-Transformation näherungsweise berechnen wollen (vgl. Kasten auf Seite 45). Seien $\mathbf{A} \in \mathbb{C}^{m \times n}$, w eine m-te und v eine n-te primitive Einheitswurzel. Die **diskrete Fourier-Transformation** zu w und v bildet dann $\mathbf{A}$ auf eine Matrix $\mathbf{B} \in \mathbb{C}^{m \times n}$ ab, wobei für die Komponenten (deren Index wir wie zuvor bei 0 beginnen) gilt:

$$b_{i,k} := \sum_{r=0}^{m-1} \sum_{s=0}^{n-1} a_{r,s} w^{ir} v^{ks} = \sum_{r=0}^{m-1} w^{ir} \sum_{s=0}^{n-1} a_{r,s} v^{ks}.$$

Für jeden festen Wert von r berechnet die innere Summe die k-te Komponente der eindimensionalen diskreten Fourier-Transformation $\mathrm{DFT}_v((a_{r,0}, a_{r,1}, \ldots, a_{r,n-1}))$. Mit diesen Ergebnissen wird dann eine weitere eindimensionale Fourier-Transformation zur Wurzel w durchgeführt:

$$b_{i,k} = \mathrm{DFT}_w \left(\left(\mathrm{DFT}_v((a_{r,0}, a_{r,1}, \ldots, a_{r,n-1}))_k \right)_{r=0}^{m-1} \right)_i.$$

Mit $m + n$ eindimensionalen diskreten Fourier-Transformationen ist die zweidimensionale diskrete Fourier-Transformation berechnet. Denn man kann zunächst für jedes $r \in \{0, 1, \ldots, m-1\}$ die innere Transformation ausrechnen und sich die Ergebnisse merken. Für jedes $k \in \{0, 1, \ldots, n-1\}$ wird nun aus den zuvor berechneten Transformierten ein neuer Vektor gebildet, indem man deren k-te Komponenten zu einem neuen Vektor zusammenfasst. Dieser Vektor wird dann transformiert, so dass zu den inneren m Transformationen noch n äußere kommen.

Eine Umkehrtransformation erhält man nun über

$$a_{i,k} := \frac{1}{m \cdot n} \sum_{r=0}^{m-1} \sum_{s=0}^{n-1} b_{r,s} w^{-ir} v^{-ks}.$$

In höheren Raumdimensionen wird dieser Ansatz analog verwendet.

6.4 Abtastsatz der Fourier-Transformation

Bandbegrenztheit ist eine sehr einschränkende Bedingung. Ist eine bandbegrenzte Funktion auf einem kleinen Stück der x-Achse (in einem offenen Intervall) gleich null,

so sind zwangsläufig alle Funktionswerte auf ganz $\mathbb{R}$ gleich null. Das folgt aus dem hier nicht behandelten Satz von Paley-Wiener und Eigenschaften von Funktionen einer komplexen Variablen. Damit haben wir ein echtes Problem bei der Anwendung des Satzes, da wir im realen Leben nicht unendlich viele Funktionswerte berücksichtigen können. Wir müssen innerhalb eines beschränkten Intervalls zwischen einem Anfangs- und einem Endzeitpunkt abtasten. Dazu setzen wir alle Werte der Funktion außerhalb dieses Intervalls zu null und verletzen so die Bandbegrenztheit. Glücklicher Weise gilt der Abtastsatz dann aber auch noch näherungsweise. Ist beispielsweise für eine auf $\mathbb{R}$ stetige Funktion f mit $\int_{-\infty}^{\infty} |f(t)|\,dt < \infty$ statt der Bandbegrenztheit „nur“ $\int_{-\infty}^{\infty} |f^\wedge(\omega)|\,d\omega$ eine endliche Zahl, so gilt für jedes $t \in \mathbb{R}$ (siehe Brown (1967))

$$f(t) = \lim_{\Delta t \to 0+} \sum_{k=-\infty}^{\infty} f(k\Delta t)\,\mathrm{sinc}\left(\frac{\pi}{\Delta t}(t - k\Delta t)\right). \tag{6.1}$$

Verlangt man zusätzlich noch, dass $\int_{-\infty}^{\infty} |f(t)|^2\,dt$ endlich ist, so ist diese Konvergenz sogar gleichmäßig (siehe Butzer, Splettstößer und Stens (1988)). Ohne Bandbegrenztheit gilt auch die Berechnung von Werten der Transformierten über Werte der Ausgangsfunktion mittels

$$f^\wedge(\omega) = \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t)\exp\left(-j\omega k\Delta t\right). \tag{6.2}$$

nicht mehr. Den Aliasing-Fehler, der dann entsteht, diskutieren wir jetzt. Dazu wenden wir die Poisson-Summationsformel (siehe Kasten auf Seite 717) zunächst für eine Funktion g an, die ihrerseits die Fourier-Transformierte einer Funktion h ist: $g(\omega) := h^\wedge(\omega)$. Sei also $h : \mathbb{R} \to \mathbb{C}$ mit $\int_{-\infty}^{\infty} |h(t)|\,dt < \infty$. Als Fourier-Transformierte ist $g = h^\wedge$ stetig. Um die weiteren Voraussetzungen der Summationsformel für $g = h^\wedge$ zu erfüllen, muss zusätzlich $\int_{-\infty}^{\infty} |h^\wedge(\omega)|\,d\omega < \infty$ sein (so dass sich die Fourier-Umkehrtransformation aus der Fourier-Transformation berechnen lässt) und

$$\sum_{k=-\infty}^{\infty} |(h^\wedge)^\wedge(k)| = 2\pi \sum_{k=-\infty}^{\infty} \left|\frac{1}{2\pi}(h^\wedge)^\wedge(-k)\right| = 2\pi \sum_{k=-\infty}^{\infty} |h(k)| < \infty$$

erfüllt sein. Weiterhin muss $\sum_{k=-\infty}^{\infty} h^\wedge(\omega + k2\pi)$ gleichmäßig gegen eine Grenzfunktion konvergieren. Dann sind die Voraussetzungen für $g = h^\wedge$ erfüllt, und wir erhalten im ersten Schritt mit der Summationsformel und im zweiten Schritt über die Fourier-Umkehrtransformation:

$$\begin{aligned}\sum_{k=-\infty}^{\infty} h^\wedge(\omega + 2k\pi) &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} (h^\wedge)^\wedge(k) e^{jk\omega} \\ &= \sum_{k=-\infty}^{\infty} \frac{1}{2\pi}(h^\wedge)^\wedge(-k) e^{-jk\omega} = \sum_{k=-\infty}^{\infty} h(k) e^{-jk\omega}. \end{aligned} \tag{6.3}$$

Im Beweis des Abtastsatzes wird $f^\wedge$ zu einer 2π-periodischen Funktion fortgesetzt. Dabei werden alle von null verschiedenen Funktionswerte in das Periodenintervall abgebildet. Hat man keine Bandbegrenzung, dann ist das etwas schwieriger. Zur Funktion f und $\Delta t > 0$ definieren wir uns die Hilfsfunktion $h(t) := \Delta t f(t\Delta t)$, so dass mit den Rechenregeln der Fourier-Transformation $h^\wedge(\omega) = f^\wedge\left(\frac{\omega}{\Delta t}\right)$ ist. Die Funktion f muss jetzt nicht bandbegrenzt sein, allerdings soll h die schwächeren Voraussetzungen der Gleichung (6.3). Dann wird (6.3) für jeden Wert $\omega_0 \in \mathbb{R}$ zu

$$\sum_{k=-\infty}^{\infty} f^\wedge\left(\frac{1}{\Delta t}\omega_0 + k\frac{2\pi}{\Delta t}\right) = \sum_{k=-\infty}^{\infty} \Delta t f(k\Delta t) e^{-jk\omega_0}.$$

Setzen wir $\omega_0 := \omega\Delta t$, so erhalten wir für alle $\omega \in \mathbb{R}$:

$$\sum_{k=-\infty}^{\infty} f^\wedge\left(\omega + k\frac{2\pi}{\Delta t}\right) = \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) e^{-j\omega k\Delta t}.$$

Damit ist

$$f^\wedge(\omega) - \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) e^{-j\omega k\Delta t} = -\underbrace{\left[\sum_{k=1}^{\infty} f^\wedge\left(\omega - k\frac{2\pi}{\Delta t}\right) + \sum_{k=1}^{\infty} f^\wedge\left(\omega + k\frac{2\pi}{\Delta t}\right)\right]}_{\text{Aliasing-Fehler}}. \tag{6.4}$$

Diese Darstellung entspricht genau dem Aliasing-Fehler, den wir auch schon bei der Berechnung von Fourier-Koeffizienten ermittelt haben. Der Aliasing-Fehler verschwindet, wenn f bandbegrenzt ist und die Shannon-Nyquist-Bedingung erfüllt ist. Dann erhält man die Formel aus dem Abtastsatz. Sonst führen die zusätzlichen Summen zu dem in Abbildung 6.27 dargestellten Effekt.

Für die Diskretisierung der Fourier-Transformation müssen wir uns auf Funktionswerte aus einem (Zeit-)Intervall $[-R, R]$ beschränken. Auch wenn wir R so groß wählen, dass die zu transformierende Funktion f außerhalb von $[-R, R]$ sehr klein ist, transformieren wir dann nicht die Funktion $f(t)$, sondern eine Funktion $f(t) \cdot 1_{[-R,R]}(t)$, wobei der als **Fensterfunktion** gewählte Rechteckimpuls $1_{[-R,R]}(t)$ auf dem Intervall $[-R, R]$ gleich 1 und sonst gleich 0 ist. Die daraus resultierende Verfälschung ist der Leck-Effekt.

Für den Rechteckimpuls kennen wir bereits die Fourier-Transformierte:

$$[\mathcal{F}1_{[-R,R]}(t)](\omega) = \left[\mathcal{F}1_{[-1,1]}\left(\frac{t}{R}\right)\right](\omega) = R\left[\mathcal{F}1_{[-1,1]}(t)\right](R\omega) = 2R\,\mathrm{sinc}(R\omega).$$

Der Faltungssatz im Frequenzbereich (siehe insbesondere die Anmerkung zu den Voraussetzungen) besagt nun, dass

$$\mathcal{F}[f \cdot 1_{[-R,R]}](\omega) = \frac{R}{\pi} f^\wedge(\omega) * \mathrm{sinc}(R\omega).$$

Man berechnet diese Funktion statt $f^\wedge(\omega)$. Das ist analog zur Berechnung von Fourier-Koeffizienten mit einem zu kleinen Abtastintervall (kleiner als eine volle Periodenlänge), die ebenfalls zur Verknüpfung mit der sinc-Funktion führt. Statt mit einer Rechteckfunktion kann man f auch mit anderen Fensterfunktionen multiplizieren, die außerhalb eines Intervalls $[-R, R]$ gleich null sind.

Ist f nicht die Nullfunktion, so ist leider das Produkt von f mit einer Fensterfunktion g nicht bandbegrenzt, und wir müssen bei der Berechnung von $(f \cdot g)^\wedge(\omega)$ mittels diskreter Fourier-Transformation einen Aliasing-Fehler in Kauf nehmen, der durch die Wahl der Fensterfunktion beeinflusst wird. Sei $[(f \cdot g)^\wedge]^*(\omega)$ der mit der diskreten Fourier-Transformation ermittelte zugehörige Wert. Dann lässt sich der Gesamtfehler mit der Dreiecksungleichung so abschätzen:

$$\begin{aligned} |f^\wedge(\omega) - [(f \cdot g)^\wedge]^*(\omega)| &= |f^\wedge(\omega) - (f \cdot g)^\wedge(\omega) + (f \cdot g)^\wedge(\omega) - [(f \cdot g)^\wedge]^*(\omega)| \\ &\leq \underbrace{|f^\wedge(\omega) - (f \cdot g)^\wedge(\omega)|}_{\text{Leck-Fehler}} + \underbrace{|(f \cdot g)^\wedge(\omega) - [(f \cdot g)^\wedge]^*(\omega)|}_{\text{Aliasing-Fehler}}. \end{aligned}$$

Um den hier auftretenden Aliasing-Fehler zu berechnen, muss man für f in die Fehlerdarstellung (6.4) die Funktion $f \cdot g$ einsetzen. Ist die Ausgangsfunktion f bandbegrenzt und ist die Fensterfunktion g so gewählt, dass $(f \cdot g)^\wedge = f^\wedge * g^\wedge$ möglichst gut $f^\wedge$ annähert, so wird nicht nur der Leck-Fehler sondern auch der Aliasing-Fehler klein. Denn dann sind die Werte $|(f \cdot g)^\wedge(\omega)|$ zu großen $|\omega|$ wegen $f^\wedge(\omega) = 0$ sehr klein, und damit werden die beiden Summen des Fehlers in (6.4) ebenfalls klein.

Die $2R\,\mathrm{sinc}(R\omega)$-Funktion als Transformierte des Rechteck-Fensters $1_{[-R,R]}(t)$ hat ein absolutes Maximum bei $\omega = 0$ aber unendlich viele weitere positive Maxima und negative Minima. Durch die Faltung mit $f^\wedge$ wird damit statt der exakten Transformierten $f^\wedge(\omega)$ eine Mittelung der Werte von $f^\wedge(\omega)$ berechnet. Dabei gehen durch das Maximum bei 0 vor allem Werte zu nahe benachbarten ω ein. Allerdings gehen durch die vielen weiteren Minima und Maxima auch weiter entfernte Werte ein, die zum Leck-Fehler beitragen und auch außerhalb des Bandbereichs von f Werte ungleich null produzieren. Wählt man R größer, so zieht sich $\mathrm{sinc}(R\omega)$ um den Nullpunkt zusammen, und weiter entfernte Werte werden weniger berücksichtigt, Leck- und Aliasing-Fehler werden kleiner.

Günstiger als Rechteckfenster sind Fensterfunktionen g, für die $|g^\wedge|$ kleinere Nebenmaxima hat und mit denen man über $f^\wedge * g^\wedge$ die gesuchte Transformierte $f^\wedge$ besser annähern kann. Beispielsweise hat die Dreiecksfunktion (**Bartlett-Fenster**) $g(t) := 1 - \left|\frac{t}{R}\right|$ für $-R \leq t \leq R$ und $g(t) := 0$ für $|t| > R$ die Fourier-Transformierte

$$g^\wedge(\omega) = R\frac{\sin^2\left(\frac{\omega R}{2}\right)}{\left(\frac{\omega R}{2}\right)^2} = R\,\mathrm{sinc}^2\left(\frac{\omega R}{2}\right).$$

Diese Funktion $g^\wedge$ entspricht dem Fejér-Kern bei 2π-periodischen Funktionen, und die Faltung von $f^\wedge$ mit $g^\wedge$ strebt für $R \to \infty$ entsprechend gut gegen $f^\wedge$. Das ist

ein Ergebnis der **Approximationstheorie**. In der Nachrichtentechnik werden andere gängige Fensterfunktionen wie das Hann- und das Hamming-Fenster eingesetzt.

Für $0 \leq a \leq 1$ betrachten wir

$$g(t) := \begin{cases} a + (1-a)\cos\left(\frac{\pi}{R}t\right) & : \quad -R \leq t \leq R \\ 0 & : \quad |t| > R. \end{cases}$$

Für $a = 1$ ist $g(t)$ die Fensterfunktion eines Rechteckfensters, für $a = \frac{1}{2}$ spricht man vom **Hann-Fenster**. In diesem Fall ist

$$g(t) := \begin{cases} \frac{1}{2} + \frac{1}{2}\cos\left(\frac{\pi}{R}t\right) = \cos^2\left(\frac{\pi}{2R}t\right) & : \quad -R \leq t \leq R \\ 0 & : \quad |t| > R, \end{cases}$$

so dass man auch von einem $\cos^2$-Fenster spricht. Bei der digitalen Signalverarbeitung hat man festgestellt, dass ein kleiner Sprung an den Rändern des Fensters einen positiven Effekt haben kann. Das ist für $a = 0{,}54$ der Fall. Hier heißt $g(t)$ die Fensterfunktion des **Hamming-Fensters**, wobei die Fensterfunktion des Hann-Fensters auf einen kleinen Sockel der Höhe 0,08 gesetzt wird (siehe Abbildung 6.1):

$$g(t) := \begin{cases} 0{,}54 + 0{,}46\cos\left(\frac{\pi}{R}t\right) = 0{,}08 + 0{,}92\cos^2\left(\frac{\pi}{2R}t\right) & : \quad -R \leq t \leq R \\ 0 & : \quad |t| > R. \end{cases}$$

Wir berechnen die Fourier-Transformierten dieser Fensterfunktionen: Da $g(t)$ eine gerade Funktion ist, ist $g^\wedge(\omega)$ eine reelle Funktion, wir erhalten

$$\begin{aligned} g^\wedge(t) &= \int_{-R}^{R} \left[a + (1-a)\cos\left(\frac{\pi}{R}t\right)\right] \exp(j\omega t)\, dt \\ &= \int_{-R}^{R} \left[a + (1-a)\cos\left(\frac{\pi}{R}t\right)\right] \cos(\omega t)\, dt \\ &= a\left[\frac{\sin(\omega t)}{\omega}\right]_{t=-R}^{t=R} + (1-a)\int_{-R}^{R} \cos\left(\frac{\pi}{R}t\right)\cos(\omega t)\, dt \\ &= a \cdot 2R\,\mathrm{sinc}(\omega R) + (1-a)\int_{-R}^{R} \cos\left(\frac{\pi}{R}t\right)\cos(\omega t)\, dt, \end{aligned}$$

wobei wir das verbleibende Integral mittels doppelter partieller Integration lösen:

$$\begin{aligned} &\int_{-R}^{R} \cos\left(\frac{\pi}{R}t\right)\cos(\omega t)\, dt \\ &= \left[\frac{R}{\pi}\sin\left(\frac{\pi}{R}t\right)\cos(\omega t)\right]_{t=-R}^{t=R} - \int_{-R}^{R} \frac{R}{\pi}\sin\left(\frac{\pi}{R}t\right)\left[-\omega\sin(\omega t)\right] dt \\ &= \frac{R}{\pi}\left[\left[-\frac{R}{\pi}\cos\left(\frac{\pi}{R}t\right)\omega\sin(\omega t)\right]_{t=-R}^{t=R} + \int_{-R}^{R} \frac{R}{\pi}\cos\left(\frac{\pi}{R}t\right)\omega^2\cos(\omega t)\, dt\right] \end{aligned}$$

$$= \quad 2\frac{R^2}{\pi^2}\omega\sin(\omega R) + \frac{R^2}{\pi^2}\omega^2 \int_{-R}^{R} \cos\left(\frac{\pi}{R}t\right)\cos(\omega t)\,dt.$$

Damit ist

$$\int_{-R}^{R} \cos\left(\frac{\pi}{R}t\right)\cos(\omega t)\,dt = \frac{2\frac{R^2}{\pi^2}}{1-\frac{R^2}{\pi^2}\omega^2}\omega\sin(\omega R) = \frac{2\omega}{\frac{\pi^2}{R^2}-\omega^2}\sin(\omega R),$$

und wir erhalten

$$\begin{aligned} g^\wedge(\omega) &= a\cdot 2R\,\mathrm{sinc}(\omega R) + (1-a)\cdot\frac{2\omega}{\frac{\pi^2}{R^2}-\omega^2}\sin(\omega R) \\ &= 2R\,\mathrm{sinc}(\omega R)\left[a+(1-a)\frac{\omega^2}{\frac{\pi^2}{R^2}-\omega^2}\right] = 2R\,\mathrm{sinc}(\omega R)\left[a+(1-a)\frac{1}{\left(\frac{\pi}{R\omega}\right)^2-1}\right]. \end{aligned}$$

Für $|\omega| \to \infty$ strebt der Term in der eckigen Klammer gegen $a-(1-a) = 2a-1$. Im Fall des Hann-Fensters, also für $a = \frac{1}{2}$, werden dadurch die Schwingungen der sinc-Funktion zusätzlich gedämpft. Beim Hamming-Fenster strebt der Term in der eckigen Klammer zwar nicht gegen null, dafür ist er aber nahe beim Hauptmaximum kleiner als beim Hann-Fenster, siehe Abbildung 6.1.

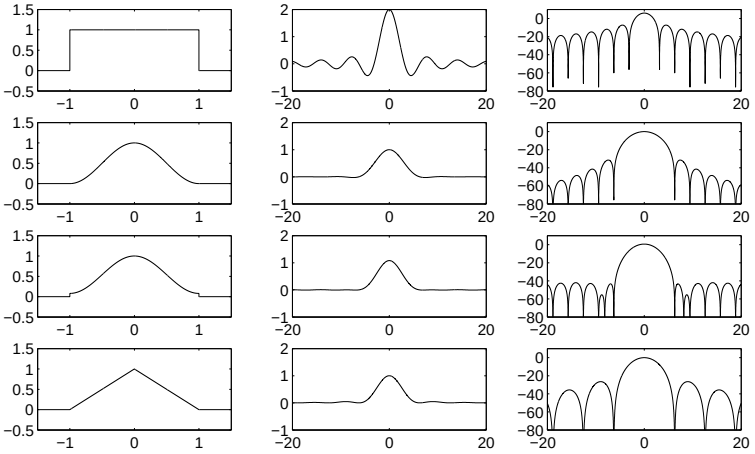

Abb. 6.1: Rechteck-, Hann-, Hamming- und Bartlett-Fenster: Links sind die Fensterfunktionen $g(t)$ zu $R = 1$ und in der Mitte die zugehörigen Transformierten $g^\wedge(\omega)$ dargestellt. Die Darstellung trügt: Keine der Fensterfunktionen ist bandbegrenzt. Das sieht man besser, wenn man eine logarithmische Darstellung der Graphen benutzt. Dazu ist rechts $10\cdot\lg(|g^\wedge(\omega)|^2)$ eingezeichnet (Intensitätsverhältnisse in Dezibel).

Hintergrund: Wavelets und schnelle Wavelet-Transformation

In die Berechnung von Fourier-Koeffizienten und von Werten der Fourier-Transformierten gehen alle Werte der Ausgangsfunktion ein. Schöner wäre es, wenn sich nur Funktionswerte aus kleinen Intervallen auf gewisse zugehörige Werte der Fourier-Transformierten auswirken würden. So lassen sich in der Praxis nur endlich viele Abtastwerte aus einem endlichen Intervall verwenden, die Werte außerhalb des Intervalls spielen aber leider für die exakte Berechnung aller Werte der Transformierten eine Rolle. Bei zu analysierenden Signalen liegen sie zum Teil in der Zukunft. Hier kann man sich mit Fensterfunktionen behelfen, die die nicht verfügbaren Daten ausblenden, aber leider auch zu einem Fehler führen. Die Verwendung von Fenstern ist eine Lokalisierung der Fourier-Transformation.

Die Ursache des globalen Verhaltens der Fourier-Transformation liegt darin, dass die verwendeten Sinus- und Kosinus-Funktionen nicht nur lokal Werte ungleich null annehmen. Bei der **Wavelet-Transformation** kann man dagegen die Sinus- und Kosinus-Funktionen durch Funktionen ersetzen, die auf immer kleineren Intervallen von null verschieden sind. Bei einer lokalen Änderung ändern sich dann nur die Koeffizienten bzw. Werte der Transformierten, die einen lokalen Bezug haben, die anderen bleiben unverändert.

Die Idee der Wavelet-Transformation besteht darin, eine Ausgangsfunktion zunächst auf eine feine Nährung abzubilden. Diese Näherung kann dann durch eine gröbere Näherung plus der Differenz zur feineren Näherung ausgedrückt werden. So verfährt man weiter und drückt die gröbere Näherung wieder durch eine noch gröbere plus eine Differenz aus. Um die feine Näherung zu erhalten, benötigt man schließlich die gröbste Näherung zuzüglich aller Differenzen. Das ist letztlich auch der Ansatz der Fourier-Reihe. Mit höheren Frequenzen kommen immer feinere Differenzen dazu. Diese werden bei Fourier-Reihen aber global berechnet und bei der Wavelet-Transformation nach Möglichkeit lokal.

Mit einer Wavelet-Transformation kann man verlustbehaftet Daten packen, indem man die Werte der Differenzen quantisiert (in Klassen einteilt), also mit geringerer Genauigkeit speichert.

Wir betrachten die Wavelet-Transformation am Beispiel des (einfachen, aber leider unstetigen) Haar-Wavelets und beginnen mit einer **Skalierungsfunktion** Φ, über die die feine Nährung der Ausgangsfunktion beschrieben wird. Beim Haar-Wavelet ist

$$\Phi(x) := \begin{cases} 1, & 0 \le x < 1, \\ 0, & \text{sonst.} \end{cases}$$

Durch Verschiebung erhalten wir die Funktionen

$$\Phi_{0,k}(x) := \Phi(x - k).$$

Wir können jede auf $\mathbb{R}$ definierte Funktion, die auf jedem Intervall $[k, k+1[$ für $k \in \mathbb{Z}$ konstant ist, als (unendliche) Linearkombination der Funktionen $\Phi_{0,k}(x)$ schreiben. Damit wir aber auch Funktionen, die nicht stückweise konstant sind, näherungsweise als

eine solche Linearkombination ausdrücken können, benötigen wir beliebig kleine konstante Stücke, die sich aneinandersetzen lassen. Diese erhalten wir durch Skalierung von $\Phi_{0,k}$:

$$\Phi_{i,k}(x) := 2^{i/2}\Phi(2^i x - k).$$

In der Tat kann man sich mit diesen Funktionen jeder integrierbaren Funktion beliebig genau nähern, wenn man als Fehlermaß das Integral über Fehlerquadrate nimmt, das bereits zur Definition der Fourierkoeffizienten geführt hat.

Die Faktoren 2^i entsprechen den Frequenzen der Sinus- und Kosinus-Terme bei Fourier-Reihen. Neu ist die durch $-k$ verursachte Translation, die den lokalen Bezug der Funktionen herstellt.

Den von den Funktionen $\Phi_{i,k}(x)$ für festes i und alle $k \in \mathbb{Z}$ erzeugte Vektorraum von Funktionen nennen wir V_i. Dann gilt:

$$V_i \subset V_{i+1}.$$

Wir haben also eine Skala von Verfeinerungen, man spricht von einer **Multiskalenanalyse**. Bei der Wavelet-Transformation stellt man zunächst die gegebene Funktion möglichst gut über eine Funktion in V_n für ein vorgegebenes n dar. Dann schreibt man diese als Funktion aus V_{n-1} (der nächst gröberen Skala) plus einer Differenz (zwischen den beiden Skalen). Der Ansatz wird fortgesetzt, bis man eine Funktion aus V_0 plus n Differenzen erhält.

Die beste Approximation (im quadratischen Mittel) an eine gegebene Funktion aus V_n erhält man (wie Fourier-Partialsummen) über eine Orthogonalprojektion. Der Vorfaktor $2^{i/2}$ der $\Phi_{i,k}(x)$ sorgt dafür, dass die Funktionen zu festem i bezüglich des (reellen) Skalarprodukts $f \bullet g := \int_{-\infty}^{\infty} f(x)g(x)\,dx$ nicht nur paarweise orthogonal, sondern auch zu 1 normiert sind. (Verwendet man eine andere Skalierungsfunktion als die des Haar-Wavelets, so bilden die $\Phi_{i,k}(x)$ nicht unbedingt eine Orthonormalbasis, müssen aber eine Riesz-Basis beschreiben. Die Rechnungen werden dann etwas schwieriger als bei orthogonalen Funktionen. Darauf gehen wir hier nicht ein.) Man erhält als beste Approximation aus V_n an f die Funktion

$$\sum_{k\in\mathbb{Z}} (\Phi_{n,k} \bullet f)\Phi_{n,k}(x) = \sum_{k\in\mathbb{Z}} \int_{-\infty}^{\infty} \Phi_{n,k}(t)f(t)\,dt \cdot \Phi_{n,k}(x).$$

Die für die Orthogonalprojektion zu berechnenden Integrale können näherungsweise mittels Quadraturverfahren aus Abtastwerten berechnet werden.

Bislang haben wir noch nicht erklärt, was ein Wavelet ist. Der Begriff kommt ins Spiel, wenn wir uns überlegen, wie man eine Differenz darstellen kann. Wegen $V_0 \subset V_1$ lässt sich Φ als Linearkombination der $\Phi_{1,k}$ schreiben:

$$\Phi(x) = \sum_{k\in\mathbb{Z}} \frac{c_k}{\sqrt{2}}\Phi_{1,k}(x) = \sum_{k\in\mathbb{Z}} c_k \Phi(2^1 x - k). \tag{6.5}$$

Die Gleichung nennt man **Verfeinerungsgleichung** oder **Skalierungsgleichung** (two-scale relation). Die Koeffizienten c_k sind für die Wavelet-Transformation sehr wichtig.

Speziell für die Skalierungsfunktion des Haar-Wavelets gilt $c_0 = c_1 = 1$ und $c_k = 0$ für $k \in \mathbb{Z} \setminus \{0, 1\}$:

$$\Phi(x) = \Phi(2x) + \Phi(2x-1) = \frac{1}{\sqrt{2}}\Phi_{1,0}(x) + \frac{1}{\sqrt{2}}\Phi_{1,1}(x).$$

Diese Beziehung gilt aufgrund der Konstruktion nicht nur für die ersten beiden Skalen, sondern ganz allgemein:

$$\Phi_{i,k}(x) = \sum_{l\in\mathbb{Z}} \frac{c_l}{\sqrt{2}}\Phi_{i+1,2k+l}(x). \tag{6.6}$$

Diese Gleichung im Rahmen des Haar-Wavelets lautet

$$\Phi_{i,k}(x) = \frac{1}{\sqrt{2}}\left[\Phi_{i+1,2k}(x) + \Phi_{i+1,2k+1}(x)\right].$$

Damit können wir nun Funktionen angeben, mit denen die Differenz der näherungsweisen Darstellung einer Funktion in zwei benachbarten Skalen beschrieben werden kann:

Das **Haar-Wavelet** ist die Funktion

$$\Psi(x) := 2\Phi(2x) - \Phi(x) = \begin{cases} 1, & 0 \le x < \frac{1}{2}, \\ -1, & \frac{1}{2} \le x < 1, \\ 0, & \text{sonst.} \end{cases}$$

Damit ist

$$\Phi(x) + \Psi(x) = 2\Phi(2x) \tag{6.7}$$

und

$$\Phi(x) - \Psi(x) = 2\Phi(2x-1). \tag{6.8}$$

Wie zuvor für Φ verschieben und skalieren wir Ψ und schreiben

$$\Psi_{i,k}(x) := 2^{i/2}\Psi(2^i x - k).$$

Damit erhalten wir aus (6.7):

$$\begin{aligned} \Phi_{1,2k}(x) &= \sqrt{2}\Phi(2x-2k) = \sqrt{2}\Phi(2[x-k]) = \frac{\sqrt{2}}{2}\left[\Phi(x-k) + \Psi(x-k)\right] \\ &= \frac{1}{\sqrt{2}}\left[\Phi_{0,k}(x) + \Psi_{0,k}(x)\right]. \end{aligned}$$

Analog ergibt sich aus (6.8):

$$\Phi_{1,2k+1}(x) = \frac{1}{\sqrt{2}}\left[\Phi_{0,k}(x) - \Psi_{0,k}(x)\right].$$

Entsprechendes gilt für höhere Skalen ($i \in \mathbb{N}_0$):

$$\Phi_{i+1,2k}(x) = \frac{1}{\sqrt{2}}\left[\Phi_{i,k}(x) + \Psi_{i,k}(x)\right], \quad \Phi_{i+1,2k+1}(x) = \frac{1}{\sqrt{2}}\left[\Phi_{i,k}(x) - \Psi_{i,k}(x)\right]. \tag{6.9}$$

Damit können wir die Funktionen $\Phi_{i+1,k}$ einer höheren Skala als Linearkombination von Funktionen $\Phi_{i,k}$ und Differenzen $\Psi_{i,k}$ $(k \in \mathbb{Z})$ schreiben.

Wir haben ein Wavelet zur Haar-Skalierungsfunktion betrachtet. Zu jeder Skalierungsfunktion, bei der $\Phi_{i,k}(x)$ für festes i ein Orthonormalsystem ist, erhält man ein zugehöriges Wavelet über den Ansatz (siehe (Louis, Maaß und Rieder, 1998, S. 122))

$$\Psi(x) = \sum_{k\in\mathbb{Z}} (-1)^k c_{1-k} \Phi(2x-k), \tag{6.10}$$

wobei $(c_k)_{k\in\mathbb{Z}}$ die Folge der Koeffizienten der Verfeinerungsgleichung (6.5) ist. Man beachte, dass das Haar-Wavelet der Spezialfall für $c_0 = c_1 = 1$ (und sonst null) ist. In diesem Fall erhalten wir mit (6.7) und (6.8):

$$\begin{aligned}\sum_{k\in\mathbb{Z}} (-1)^k c_{1-k} \Phi(2x-k) &= \Phi(2x-0) - \Phi(2x-1) = \frac{1}{2}[2\Phi(2x) - 2\Phi(2x-1)] \\ &= \frac{1}{2}[\Phi(x) + \Psi(x) - \Phi(x) + \Psi(x)] = \Psi(x).\end{aligned}$$

Abweichend von dieser Darstellung (6.10) findet man in der Literatur auch die Definition

$$\Psi(x) := \sum_{k\in\mathbb{Z}} (-1)^{k+1} c_{1-k} \Phi(2x-k),$$

bei der das Wavelet genau ein entgegengesetzes Vorzeichen hat.

Aus (6.10) erhalten wir für höhere Skalen

$$\begin{aligned}\Psi_{i,k}(x) &= 2^{i/2} \sum_{l\in\mathbb{Z}} (-1)^l c_{1-l} \Phi(2[2^i x - k] - l) \\ &= 2^{i/2} \sum_{l\in\mathbb{Z}} (-1)^l c_{1-l} \Phi(2^{i+1} x - [2k+l]) \\ &= 2^{i/2} \sum_{l\in\mathbb{Z}} (-1)^l c_{1-l} \frac{1}{2^{\frac{i+1}{2}}} \Phi_{i+1,2k+l}(x) \\ &= \sum_{l\in\mathbb{Z}} (-1)^l \frac{c_{1-l}}{\sqrt{2}} \Phi_{i+1,2k+l}(x).\end{aligned} \tag{6.11}$$

Die **schnelle Wavelet-Transformation** basiert auf den Gleichungen (6.6) und (6.11). Wir beginnen mit der besten Approximation $g_n \in V_n$ an f:

$$g_n(x) = \sum_{k\in\mathbb{Z}} d_{n,k} \Phi_{n,k}(x),$$

wobei die Koeffizienten $d_{n,k} = \Phi_{n,k} \bullet f$ wie oben beschrieben mittels Quadratur berechnet werden können. Verwenden wir beispielsweise das Haar-Wavelet, so können wir eine Funktion f auf jedem der Teilintervalle $[k2^{-n}, (k+1)2^{-n}[$ durch eine konstante Funktion mit Wert $f(k2^{-n})$ ersetzen. Damit erhalten wir

$$d_{n,k} \approx f(k2^{-n}) \int_{-\infty}^{\infty} \Phi_{n,k}(x)\,dx = f(k2^{-n}) \int_{-\infty}^{\infty} 2^{n/2} \Phi(2^n x - k)\,dx$$

$$= \quad f(k2^{-n})2^{-n/2}\int_{-\infty}^{\infty}\Phi(u)\,du = 2^{-n/2}f(k2^{-n}).$$

Hier entsprechen (bis auf Normierung) die Koeffizienten $d_{n,k}$, die die Eingangsdaten für die schnelle Wavelet-Transformation sind, genau äquidistant gebildeten Abtastwerten der darzustellenden Funktion f.

Jetzt verlangen wir, dass die Funktionen $\{\Phi_{i,k}, \Psi_{i,k} : k \in \mathbb{Z}\}$ für jedes feste i ein Orthonormalsystem bilden. Beim Haar-Wavelet ist offensichtlich $\Phi \bullet \Phi = \int_0^1 1\,dx = 1$, $\Psi \bullet \Psi = \int_0^1 1\,dx = 1$, $\Phi \bullet \Psi = \int_0^{\frac{1}{2}} 1\,dx + \int_{\frac{1}{2}}^1 -1\,dx = 0$. Durch Skalierung und Translation ergibt sich daraus sofort, dass tatsächlich für jedes feste i ein Orthonormalsystem vorliegt.

Eine Funktion

$$g_{i+1}(x) = \sum_{k\in\mathbb{Z}} d_{i+1,k}\Phi_{i+1,k}(x) \in V_{i+1}$$

(die ja für $i+1 = n$ als beste Approximation an f bereits berechnet ist) kann wegen (6.9) in der i-ten Skala geschrieben werden als Orthogonalprojektion

$$g_{i+1}(x) = \sum_{k\in\mathbb{Z}} d_{i,k}\Phi_{i,k}(x) + \sum_{k\in\mathbb{Z}} h_{i,k}\Psi_{i,k}(x),$$

wobei sich die Koeffizienten über Skalarprodukte berechnen. Wegen der Orthonormalität für festes i erhalten wir mit (6.6):

$$\begin{aligned} d_{i,k} \quad &= \quad g_{i+1} \bullet \Phi_{i,k} = \sum_{m\in\mathbb{Z}} d_{i+1,m}\Phi_{i+1,m} \bullet \Phi_{i,k} \\ &\overset{(6.6)}{=} \quad \sum_{m\in\mathbb{Z}} d_{i+1,m}\Phi_{i+1,m} \bullet \left(\sum_{l\in\mathbb{Z}} \frac{c_l}{\sqrt{2}}\Phi_{i+1,2k+l}\right) \\ &= \quad \sum_{m\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} d_{i+1,m}\frac{c_l}{\sqrt{2}}\Phi_{i+1,m} \bullet \Phi_{i+1,2k+l} = \sum_{l\in\mathbb{Z}} d_{i+1,2k+l}\frac{c_l}{\sqrt{2}}. \end{aligned}$$

Hier haben wir eine Summe mit dem Skalarprodukt (also einem Integral) vertauscht. Solange die Summe wie beim Haar-Wavelet endlich ist, ist das kein Problem. Ebenso ergibt sich mit (6.11):

$$\begin{aligned} h_{i,k} \quad &= \quad g_{i+1} \bullet \Psi_{i,k} = \sum_{m\in\mathbb{Z}} d_{i+1,m}\Phi_{i+1,m} \bullet \Psi_{i,k} \\ &\overset{(6.11)}{=} \quad \sum_{m\in\mathbb{Z}} d_{i+1,m}\Phi_{i+1,m} \bullet \left(\sum_{l\in\mathbb{Z}} (-1)^l \frac{c_{1-l}}{\sqrt{2}}\Phi_{i+1,2k+l}\right) \\ &= \quad \sum_{l\in\mathbb{Z}} d_{i+1,2k+l}(-1)^l\frac{c_{1-l}}{\sqrt{2}}. \end{aligned}$$

Damit können wir mit einer Darstellung von $g \in V_n$ beginnen und daraus mit den beiden vorangehenden Gleichungen eine Darstellung in V_{n-1} mit Koeffizienten $d_{n-1,k}$ plus einer Differenz mit Koeffizienten $h_{n-1,k}$ berechnen. Das wird fortgesetzt, bis man bei V_0 ankommt. Gespeichert werden müssen die Koeffizienten $d_{0,k}$ der Darstellung in V_0 und alle Koeffizienten $h_{i,k}$ der Differenzen, $0 \le i < n$, $k \in \mathbb{Z}$. Obwohl als Indexmenge

$\mathbb{Z}$ verwendet wird, sind beim Haar-Wavelet beispielsweise nur zwei Werte c_l von null verschieden. Es müssen also nur Summen mit zwei Summanden berechnet werden.

Aus den Werten $d_{i+1,k}$ einer Stufe $i+1$ werden ungefähr halb so viele Werte $d_{i,k}$ der Stufe i berechnen. Aus m Werten der Stufe $i = n$ werden damit ungefähr $m\left(1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{n-1}}\right) \leq 2m$ Werte. Ebenso werden ca. $2m$ Koeffizienten $h_{i,k}$ bestimmt. Jeder einzelne Koeffizient wird über die endliche Summe in „konstanter“ Zeit berechnet. Damit hängt die Rechenzeit der schnellen Wavelet-Transformation (bei Summation über eine endliche Indexmenge) linear von der Anzahl der gegebenen Abtastwerte (oder alternativ: linear von der Anzahl der zu berechnenden Koeffizienten) ab. Sie ist damit effizienter als die schnelle Fourier-Transformation. Auch sieht man, wie sich lokale Änderungen an der Ausgangsfunktion auswirken. Beim Haar-Wavelet beeinflussen sie nur wenige Koeffizienten in der V_n-Darstellung. Durch die kurzen Summen bei der Berechnung der $d_{i,k}$ und $h_{i,k}$ ändern sich auch nur wenige dieser Koeffizienten. Die Auswirkungen bleiben lokal.

Die Rekonstruktion der Näherung g der Ausgangsfunktion f geschieht nun durch Summation über die berechneten Differenzen:

$$g(x) = \sum_{k\in\mathbb{Z}} d_{0,k}\Phi_{0,k}(x) + \sum_{i=0}^{n-1}\sum_{k\in\mathbb{Z}} h_{i,k}\Psi_{i,k}(x).$$

Im Vergleich zu Fourier-Koeffizienten, die von der Frequenz abhängen, hängen die hier verwendeten Koeffizienten $h_{i,k}$ von zwei Parametern ab: i beschreibt die Skala und entspricht damit der Frequenz bei Fourier-Koeffizienten. Der Ort auf der x-Achse wird durch k adressiert. Durch diesen Parameter wirken sich lokale Änderungen an einer Ausgangsfunktion auch nur auf die entsprechenden lokalen Koeffizienten aus.

7 Zusatzmaterial zu Kapitel 7

Übersicht

7.1 Modellbildung und Häufigkeit

Beispiel 7.1 (Bilderkennung)
Viele Verfahren zur Bilderkennung stützen sich auf die beschreibende Statistik. Möchte man beispielsweise berechnen, ob auf einem Foto eine Null „0“ oder ein Pluszeichen „+“ abgebildet ist, so kann man das über ein Merkmal tun. Die statistischen Elemente sind die Schwarzweißfotos, auf denen die schwarzen Zeichen bildfüllend zentriert abgebildet seien. Jedes Bild besteht aus einer Matrix mit schwarzen und weißen Punkten. Wir betrachten ein Merkmal, dass eine Spalte mit den meisten schwarzen Punkten ermittelt. Gibt es insgesamt n Spalten und liegt der Wert des Merkmals im Bereich der Spalten $\left[\frac{n}{4}, \frac{3n}{4}\right]$, also in der Bildmitte, so ist ein Pluszeichen wahrscheinlich. Liegt das Maximum außerhalb dieses Intervalls, also an den Rändern, so ist eine Null wahrscheinlich. ■

7.2 Lineare Regressionsrechnung

Auf Seite 819 werden die Parameter der Regressionsgeraden über die für ein Minimierungsproblem notwendige Bedingung, dass der Gradient der Nullvektor sein muss, berechnet. Es fehlt eine strenge mathematische Begründung dafür, dass es sich bei den

berechneten Parametern wirklich um eine Minimalstelle handelt. Dazu verwenden wir die hinreichende Bedingung der positiven Definitheit der Hesse-Matrix. Diese ist hier die konstante Matrix $\begin{bmatrix} 2n & 2n\overline{x} \\ 2n\overline{x} & 2\sum_{i=1}^n x_i^2 \end{bmatrix}$.
Die erste Hauptabschnittsdeterminante ist $2n > 0$, die zweite ist

$$\det \begin{bmatrix} 2n & 2n\overline{x} \\ 2n\overline{x} & 2\sum_{i=1}^n x_i^2 \end{bmatrix} = 4n\sum_{i=1}^n x_i^2 - 4n^2\overline{x}^2 = 4\left[n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2\right]. \quad (7.1)$$

Als Anwendung des Skalarprodukts ist

$$\begin{aligned} \sum_{i=1}^n x_i &= (x_1, x_2, \ldots, x_n) \cdot (1, 1, \ldots, 1) \\ &= |(x_1, x_2, \ldots, x_n)| \cdot |(1, 1, \ldots, 1)| \cdot \cos(\varphi) = \sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{n} \cdot \cos(\varphi), \end{aligned}$$

wobei φ der Vektor zwischen $(x_1, x_2, \ldots, x_n)$ und $(1, 1, \ldots, 1)$ ist. Damit erhalten wir $\left(\sum_{i=1}^n x_i\right)^2 = \cos^2(\varphi) \cdot n\sum_{i=1}^n x_i^2$ und insbesondere $\left(\sum_{i=1}^n x_i\right)^2 \leq n\sum_{i=1}^n x_i^2$, wobei Gleichheit nur gilt, wenn $\cos^2(\varphi) = 1$ ist. Das ist nur der Fall, wenn alle x_i den gleichen Wert haben, was wir aber mit der Forderung, dass ein Wert von $\overline{x}$ verschieden sein muss, ausgeschlossen haben. Damit ist aber die Determinante in (7.1) positiv, und die Hesse-Matrix ist positiv definit. Es liegt also ein (lokales) Minimum vor.

7.3 Wahrscheinlichkeitsrechnung

Bei der Modellierung von Zufallsexperimenten muss man einen geeigneten Wahrscheinlichkeitsraum finden. Dass dies gar nicht so einfach ist, zeigt das folgende Beispiel: Eine Familie hat zwei Kinder. Bei einem Besuch kommt zufällig ein Kind in den Raum. (Die beiden Kinder werfen eine Münze, wer in den Raum geht.) Gesucht ist die Wahrscheinlicheit, dass wenn dieses Kind ein Junge ist, auch das zweite Kind ein Junge ist. Wenn man davon ausgeht, dass Jungen und Mädchen mit der gleichen Wahrscheinlichkeit $\frac{1}{2}$ geboren werden und die Geburten stochastisch unabhängig voneinander sind, dann sollte die Wahrscheinlichkeit dafür, dass das zweite Kind ein Junge ist, ebenfalls $\frac{1}{2}$ sein. Also muss in der folgenden Überlegung ein Fehler stecken:

$\Omega := \{JJ, MM, JM, MJ\}$, wobei der erste Buchstabe der Elementarereignisse das Geschlecht des Erstgeborenen angibt. Jedes Elementarereignis hat die Wahrscheinlichkeit $\frac{1}{4}$. Das Ereignis, dass mindestens ein Kind ein Junge ist (was ja beobachtet wird),

ist damit $E := \{JJ, JM, MJ\}$ mit $P(E) = \frac{3}{4}$. Die Wahrscheinlichkeit, dass unter dieser Nebenbedingung beide Kinder Jungen sind, ist

$$P(\{JJ\}|E) = \frac{P(\{JJ\} \cap E)}{P(E)} = \frac{P(\{JJ\})}{P(E)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} \neq \frac{1}{2}.$$

Was ist hier falsch?

Tatsächlich ist die Definition von E nicht äquivalent zur Nebenbedingung, dass zufällig ein Junge gesehen wird. Denn die Wahrscheinlichkeit dafür sollte $\frac{1}{2}$ und nicht $\frac{3}{4}$ sein. E modelliert das Ereignis, dass mindestens ein Kind ein Junge ist. Das ist aber davon verschieden, dass auch ein Junge gesehen wird, denn wenn z. B. ein Mädchen und ein Junge zur Familie gehören, kann die Münze auf das Mädchen fallen.

Mit unserem Modell können wir also die Situation nicht richtig abbilden. Daher wählen wir ein anderes Modell:

$\Omega := \{JJ1, MM1, JM1, MJ1, JJ2, MM2, JM2, MJ2\}$, wobei der erste Buchstabe der Elementarereignisse das Geschlecht des Erstgeborenen angibt und die dritte Stelle bestimmt, ob der Erstgeborene (1) oder der Zweitgeborene (2) in den Raum kommt. Dann ist das Ereignis, dass ein Junge in den Raum kommt $E := \{JJ1, JJ2, JM1, MJ2\}$ und

$$P(\{JJ1, JJ2\}|E) = \frac{P(\{JJ1, JJ2\} \cap E)}{P(E)} = \frac{P(\{JJ1, JJ2\})}{P(E)} = \frac{\frac{2}{8}}{\frac{4}{8}} = \frac{1}{2}.$$

7.4 Punktschätzungen

In der Statistik werden häufig **Maximum-Likelihood-Schätzer** verwendet. Hier nimmt man aus plausiblen Gründen an, dass eine bestimmte Wahrscheinlichkeitsverteilung vorliegt und bestimmt einen Parameter der Verteilung über ein Experiment. Durch die Annahme ist das Verfahren bereits mit einer Unsicherheit verbunden, denn der zu bestimmende Parameter ist nicht notwendiger Weise wie in der Definition der Punktschätzung ein Parameter des Wahrscheinlichkeitsraums. Vielmehr wird dies nur angenommen. Die Annahme einer Wahrscheinlichkeitsverteilung lässt sich allerdings mit statistischen Mitteln überprüfen.

Um den Parameter der angenommenen Verteilung zu finden, definiert man eine Funktion (**Likelihood-Funktion**), die die Wahrscheinlichkeit des vorliegenden Experimentausgangs in Abhängigkeit des Parameters beschreibt. Ein Parameter, für den die Funktion maximal wird, ist dann die Maximum-Likelihood-Schätzung (siehe z. B. (Arens et al., 2012, S. 1370)). Man bestimmt den Parameter also so, dass der beobachtete Experimentausgang für diesen Parameter am wahrscheinlichsten ist.

Nimmt man beispielsweise an, dass eine Binomialverteilung $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ mit $n = 3$ vorliegt und beobachtet den Wert $X = 1$, so schätzt man p so, dass

für dieses p die Wahrscheinlichkeit von $X = 1$ im Vergleich zu anderen Werten von p maximal wird. Wir suchen also ein p, für das die Funktion $f(p) = \binom{3}{1}p^1(1-p)^{3-1} = 3(p - 2p^2 + p^3)$ ein Maximum annimmt. Für ein solches p muss die Ableitung null werden: $0 = f'(p) = 3-12p+9p^2$. Die Nullstellen sind 1 und $\frac{1}{3}$. Wegen $f''\left(\frac{1}{3}\right) = -6 < 0$ wird hier ein (lokales) Maximum angenommen. An den Intervallrändern $p = 1$ und $p = 0$ ist $f(p) = 0$. Da die stetige Funktion f auf $[0, 1]$ ihr Maximum annimmt, liegt dieses bei $p = \frac{1}{3}$ mit Wert $\frac{4}{9}$. Damit ist $\frac{1}{3}$ die Maximum-Likelihood-Schätzung für den Parameter p.

7.5 Konfidenzintervall für eine Wahrscheinlichkeit

Wie groß muss man eine Stichprobe wählen?

Häufig ist die Breite des Konfidenzintervalls bereits vorgegeben, d. h., eine gewisse Genauigkeit der Schätzung wird erwartet. Um diese Genauigkeit zu treffen, kann man die Anzahl der Experimente n variieren. Wie groß muss man n bei einem vorgegebenen $\varepsilon > 0$ wählen, damit die Wahrscheinlichkeit, dass die zu schätzende unbekannte Wahrscheinlichkeit p im Intervall $[\hat{p} - \varepsilon, \hat{p} + \varepsilon]$ liegt, mindestens $1 - \alpha$ ist?

Dazu wählen wir n so groß, dass $x \cdot \sqrt{\frac{p(1-p)}{n}} < \varepsilon$ bzw. $x \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \varepsilon$ ist. Da p nicht bekannt ist und $\hat{p}$ von der Anzahl der Experimente abhängt, nutzen wir aus, dass $p(1-p) \leq \frac{1}{4}$ und $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$ sind ($t(1-t)$ ist auf $[0, 1]$ nicht-negativ und hat das Maximum bei $t = \frac{1}{2}$). Damit verlangen wir $x \cdot \frac{1}{2\sqrt{n}} < \varepsilon$ bzw. $n > \left(\frac{x}{2\varepsilon}\right)^2$. Soll also die zu schätzende Wahrscheinlichkeit selbst mit Wahrscheinlichkeit 0,99 in einem Konfidenzintervall mit Radius $\varepsilon = 0{,}01$ (maximale Abweichung: ein Prozentpunkt) um den Schätzwert liegen, so wird das mit $n > \left(\frac{2{,}5758}{2 \cdot 0{,}01}\right)^2 \approx 16\,587$ Stichprobenwerten erreicht. Verlangt man nur eine Sicherheit von 0,9 für das Konfidenzintervall, so ergeben sich die Parameter $\alpha = 0{,}1$, $\Phi(x) = 1 - \frac{\alpha}{2} = 0{,}95$ und $x = 1{,}6449$. Erlaubt man zudem eine Abweichung von fünf Prozentpunkten ($\varepsilon = 0{,}05$), so ist $n > \left(\frac{1{,}6449}{2 \cdot 0{,}05}\right)^2 \approx 271$ zu wählen.

7.6 Vergleich zweier geschätzter Wahrscheinlichkeiten

Neben dem Vergleich zweier Wahrscheinlichkeiten kann man auch eine komplette vorliegende (aber unbekannte) Wahrscheinlichkeitsverteilung gegen eine angenommene testen. Die Nullhypothese beim so genannten **Chi-Quadrat-Test** (χ^2-Test) ist dabei, dass die angenommene Verteilung vorliegt. Beim Test wird eine nicht-negative Zufallsvariable χ^2 eingesetzt, die bei wahrer Nullhypothese einer **Chi-Quadrat-Verteilung**

(χ^2-Verteilung) genügt (siehe Abbildung 7.1). Sei $x_{1-\alpha}$ die Stelle, an der die χ^2-

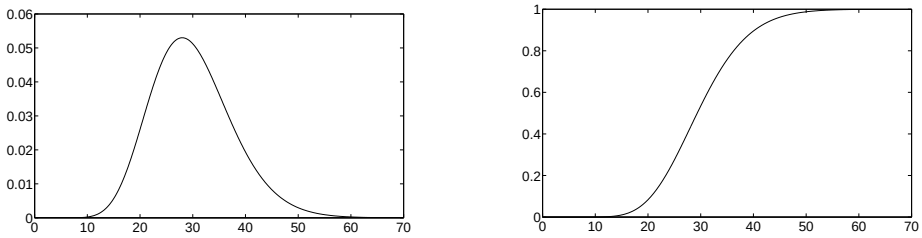

Abb. 7.1: Dichtefunktion und Verteilungsfunktion der χ^2-Verteilung zum Freiheitsgrad 30

Verteilung den Wert $1-\alpha$ annimmt. $x_{1-\alpha}$ ist das $1-\alpha$-Quantil der Verteilung. Dann ist Wahrscheinlichkeit, einen Wert aus $[0, x_{1-\alpha}]$ zu beobachten gleich $1-\alpha$ und einen Wert aus $]x_{1-\alpha}, \infty[$ zu beobachten, gleich α. Die Nullhypothese zum Niveau $1-\alpha$ ist abzulehnen, wenn ein Wert größer als $x_{1-\alpha}$ beobachtet wird, da dieses Ergebnis bei einem kleinen α sehr unwahrscheinlich ist. Die Annahme, dass eine χ^2-Verteilung vorliegt, ist damit unwahrscheinlich, so dass die Nullhypothese vermutlich nicht wahr ist. Genau so haben wir auch mit der Normalverteilung bei der Statistischen Prozesslenkung argumentiert.

Eine χ^2-Verteilung zum Freiheitsgrad (d. h. Parameter) n entsteht als Verteilung einer Zufallsvariablen $Z := \sum_{k=1}^{n} X_k^2$, die als Summe der Quadrate von n stochastisch unabhängigen, standardnormalverteilten Zufallsvariablen $X_1, \ldots, X_n$ definiert ist. Für jeden Wert $n \in \mathbb{N}$ erhält man eine andere Verteilung. Die Werte der Verteilungen findet man in Tabellenwerken (siehe Tabelle 7.1).

Tab. 7.1: Quantile der χ^2-Verteilungsfunktionen zu $n = 5$, 10, 20 und 30 Freiheitsgraden, erste Zeile: Wert der Verteilungsfunktion, zweite Zeile: Argument, bei dem der Wert angenommen wird (zugehöriges Quantil)

$n = 5$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	11,0705	11,6443	12,3746	13,3882	15,0863	16,7496	20,5150	22,1053	25,7448
$n = 10$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	18,3070	19,0207	19,9219	21,1608	23,2093	25,1882	29,5883	31,4198	35,5640
$n = 20$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	31,4104	32,3206	33,4624	35,0196	37,5662	39,9968	45,3147	47,4985	52,3860
$n = 30$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	43,7730	44,8336	46,1599	47,9618	50,8922	53,6720	59,7031	62,1619	67,6326

Die Zufallsvariable χ^2 wird für den Test so berechnet: Es möge m mögliche Zufallswerte geben. Falls es unendlich viele gibt, so muss man diese in m Klassen gruppieren,

also wie in der beschreibenden Statistik eine Klasseneinteilung vornehmen. Wenn man nun n Experimente, n sehr viel größer als m, durchführt, so erwartet man für den k-ten Zufallswert oder die k-te Klasse „m mal angenommene Eintrittswahrscheinlichkeit des Zufallswerts“ Werte, die bei Experimenten beobachtet werden sollten. Diese positive Anzahl sei a_k. Nun werden in einem Experiment tatsächlich n Stichprobenwerte $X_1(\omega), \ldots, X_n(\omega)$ beobachtet, und es wird gezählt, wie oft jeder der m Zufallswerte angenommen wird. Diese Anzahlen seien $n_1, \ldots, n_m$. Damit ist

$$\chi^2(\omega) := \sum_{i=1}^{m} \frac{(n_i - a_i)^2}{a_i}.$$

Diese Zufallsvariable ist (falls die Anzahlen n_i hinreichend groß sind) annähernd χ^2-verteilt zum Freiheitsgrad $m - 1$. Je kleiner der Wert der Variablen ist, desto besser passen die beobachteten Ergebnisse mit der zu testenden Wahrscheinlichkeitsverteilung überein. Ist der Wert jedoch so groß, dass er außerhalb des $1 - \alpha$-Quantils der χ^2-Verteilung zum Freiheitsgrad $m - 1$ liegt, so liegt vermutlich nicht die angenommene Wahrscheinlichkeitsverteilung vor.

Beispiel 7.2
Wir testen, ob ein Würfel gezinkt ist. Unsere Nullhypothese ist, dass er in Ordnung ist und alle Zahlen mit der Wahrscheinlichkeit $\frac{1}{6}$ auftreten. Jetzt würfeln wir 6 000 mal. Wir erwarten, dass jede Augenzahl 1 000 mal auftritt. Tatsächlich erhalten wir die folgenden Anzahlen:

1	2	3	4	5	6
800	900	1 100	1 200	990	1 010

Ist die Nullhypothese zum Niveau $1 - \alpha = 0{,}95$ abzulehnen? Die zu berechnende Zufallsvariable hat den Wert

$$\frac{1}{1000}\left[(800 - 1000)^2 + 100^2 + 100^2 + 200^2 + 10^2 + 10^2\right] = 100{,}2.$$

Dieser Wert ist (erheblich) größer als der Wert 11,0705 des 0,95-Quantils (siehe Tabelle 7.1 für $n = 6 - 1 = 5$) und sogar größer als das 0,9999-Quantil. Damit muss von einem gezinkten Würfel ausgegangen werden. Hätten wir

1	2	3	4	5	6
1 020	1 010	1 004	990	980	996

beobachtet, so würde daraus der Wert $\frac{1}{1000}[400 + 100 + 16 + 100 + 400 + 16] = 1{,}032$ berechnet, der weit unterhalb des 0,95-Quantils liegt. Damit gibt es in diesem Fall kein Anhaltszeichen für einen gezinkten Würfel. ■

8 MATLAB-Programme

Übersicht

MATLAB® ist eine mathematische Programmierumgebung der Firma The MathWorks, vgl. z. B. Schott (2004).

8.1 Berechnung eines Apfelmännchens

Wir berechnen die Figur aus Abbildung 2.4 von Seite 198: Zu jeder komplexen Startzahl berechnet die Funktion *mandelbrot* in Abbildung 8.1 die Höhe des Graphen als Anzahl der Iterationen bis zum Erreichen des Wertes 16. Für $(n+1)^2$ Punkte aus dem Intervall $[-2,1] \times [-1,1]$, die als komplexe Zahl geschrieben werden, wird diese Funktion aufgerufen und schließlich als Oberfläche mit dem *surf*-Befehl geplottet.

8.2 Numerische Berechnung von Fourier-Koeffizienten

Das MATLAB®-Programm in Abbildung 8.2 berechnet eine Fourier-Partialsumme des Dirichlet-Kerns D_3 mit den $2n+1$ Koeffizienten $f\hat{}(-n)$ bis $f\hat{}(n)$. Es stellt anschließend die mit diesen Koeffizienten gebildete Fourier-Reihe dar.

```
function erg = mandelbrot(c)
z=c;
k=0;
while (k<30)&&(abs(z)<16)
  z=z*z+c;
  k=k+1;
end
erg=k;


n = 100;
x = -2 :3/n : 1;
y = -1 :2/n : 1;
Z=zeros(n+1,n+1);
for l=1:n+1,
   for k=1:n+1,
      Z(l,k)=mandelbrot(x(k)+j*y(l));
   end
end
surf(x,y,Z);
```

Abb. 8.1: 3D-Plot eines Apfelmännchens

8.3 Numerische Berechnung der Fourier-Transformation

In MATLAB® sieht die numerische Berechnung von Funktionswerten einer Fourier-Transformierten z. B. wie in Abbildung 8.3 aus, wobei die Funktion f, die zum übergebenen Argumentvektor einen Vektor mit Funktionswerten der Ausgangsfunktion liefert, vorab in einem *m-File* zu definieren ist.

```
% insgesamt 2n+1 Abtastpunkte der Ausgangsfunktion:
n=3;
% Anzahl der Funktionswerte zur Darstellung der Reihe
% pro pi-Intervall:
m=30;
% Abstand der Samples:
dt=2*pi/(2*n+1);
%
% Vektor mit Abtaststellen:
tvec=0:dt:2*n*dt;
% Vektor mit zugehoerigen Funktionswerten, hier zum
% Dirichlet-Kern D3(t)=1+2cos(t)+2cos(2t)+2cos(3t):
fvec=ones(1,2*n+1)+2*cos(tvec)+2*cos(2*tvec)+2*cos(3*tvec);
%
% Schnelle Fourier-Transformation mit
% w=exp(-j*2*pi/(2n+1)), Normierung und Vertauschung
% der beiden Vektorhaelften. Erhalte Vektor der
% naeherungsweisen Fourier-Koeffizienten:
Fvec=fftshift(fft(fvec)/(2*n+1));
%
% berechne 4m+1 Funktionswerte der Ergebnis-Fourier-Summe:
rvec=zeros(1,4*m+1);
for k=-n:1:n
   rvec=rvec+Fvec(n+1+k)*exp(j*k*yvec);
end
%
% Graph der Fourier-Reihe:
plot(yvec, real(rvec(1:4*m+1)),'-k');
```

Abb. 8.2: Berechnung einer Fourier-Partialsumme mittels FFT

```
% f ist sehr klein ausserhalb von [-R,R]:
R=30*pi;
% n ist die Haelfte der Abtastpunkte:
n=30;
%
% Abstand der Samples:
dt=R/n;
% Abstand zweier berechneter Werte der Transformierten
% (unabhaengig von n)
dy=pi/R;
%
% Abtaststellen und Werte:
tvec=-n*dt:dt:(n-1)*dt;
fvec=f(tvec);
%
% Schnelle Fourier-Transformation mit Einheitswurzel
% w=exp(-j*2*pi/(2*n)) und Vertauschung der Haelften.
Fvec=fftshift(dt*fft(fvec));
%
% Multipliziere den k-ten Eintrag mit (-1)^(n-(k-1)):
%
Fvec=Fvec.*(((-1)*ones(1,2*n)).^(n*ones(1,2*n)-[0:1:2*n-1]));
%
% Diskretisierung des Frequenzbereichs passend zur Rechnung:
yvec=-n*dy:dy:(n-1)*dy;
%
% Darstellung der Transformierten:
plot(yvec, real(Fvec(1:2*n)),'.-k');
```

Abb. 8.3: Fourier-Transformation mittels FFT

Literaturverzeichnis

Lehrbücher für Höhere Mathematik

Burg K., Haf H. und Wille F. (2008, 2009) Höhere Mathematik für Ingenieure Band 1: Analysis, Band 2: Lineare Algebra, Band 3: Gewöhnliche Differenzialgleichungen, Distributionen, Integraltransformationen. Vieweg+Teubner, Wiesbaden. (Ein Grundlagenwerk zur Ingenieurmathematik)

Dobner G. und Dobner H.-J. (2004) Gewöhnliche Differenzialgleichungen. Fachbuchverlag Leipzig/Hanser, München. (Schöne und kompakte Einführung in gewöhnliche Differenzialgleichungen)

Dobner H.-J. und Engelmann B. (2002 und 2003) Analysis Band 1 und 2. Fachbuchverlag Leipzig/Hanser, München. (Schöne und kompakte Einführung in die Differenzial- und Integralrechnung und mehrdimensionale Analysis)

Dürrschnabel K. (2004) Mathematik für Ingenieure. Teubner, Wiesbaden. (Elementar gehaltene Einführung in die Ingenieurmathematik)

Gellrich R. und Gellrich C. (2003) Mathematik – Ein Lehr- und Übungsbuch Band 1–4. Harri Deutsch, Frankfurt a. M. (Schöne Beispiele, einfach geschrieben)

Hachenberger D. (2005) Mathematik für Informatiker. Pearson, München. (Schwerpunkt liegt auf Grundstrukturen)

Kreyszig E. (1998) Advanced Engineering Mathematics. Wiley, New York. (Ein Klassiker aus USA, der viele Themen der höheren Ingenieurmathematik behandelt)

Martensen E. (1998) Analysis I-IV. BI-Hochschultaschenbücher und Spektrum Hochschultaschenbücher, Bibliographisches Institut und Spektum-Verlag, Mannheim.

Meyberg K. und Vachenauer P. (1997) Höhere Mathematik Band 1 und 2. Springer, Berlin Heidelberg. (Insbesondere in Band 2 stehen viele Hintergrundinformationen zu Differenzialgleichungen und Integraltransformationen)

Papula L. (2008) Mathematik für Ingenieure und Naturwissenschaftler Band 1–3. Vieweg, Braunschweig. (Sehr beliebt bei Studierenden der Ingenieurfächer, auch bei wenig Vorkenntnissen gut verständlich)

Precht M., Voit K. und Kraft R. (1991) Mathematik 2 für Nichtmathematiker. Oldenbourg, München.

Rießinger T. (2007) Mathematik für Ingenieure. Springer, Berlin Heidelberg. (Kurzweilige und leicht verständliche Einführung in die Ingenieurmathematik)

Schott D. (2004) Ingenieurmathematik mit MATLAB®. Fachbuchverlag Leipzig/Hanser, München. (Kurze Anhänge, die zeigen, wie die behandelte Mathematik mit dem Mathematiksystem MATLAB® umgesetzt werden kann)

Stingl P. (2004) Mathematik für Fachhochschulen. Hanser, München. (Deckt bis auf einige Details ebenfalls den gesamten Stoff des Bachelor-Studiums in Ingenieurfächern ab)

Westermann T. (2008) Mathematik für Ingenieure. Springer, Berlin Heidelberg. (Deckt den gesamten Stoff des Bachelor-Studiums in Ingenieurfächern ab)

Lehrbücher zur Wahrscheinlichkeitsrechnung und Statistik

Beck-Bornholdt H.-P. und Dubben H.-H. (2006) Der Hund, der Eier legt. Rowohlt, Reinbek. (Eine kurzweilige Anleitung zum Fälschen von Statistiken)

Beucher O. (2007) Wahrscheinlichkeitsrechnung und Statistik mit MATLAB. Springer, Berlin Heidelberg.

Henze N. (2003) Stochastik für Einsteiger. Vieweg, Wiesbaden. (Gut lesbare Einführung in die Stochastik)

Menges G. (1982) Die Statistik: 12 Stationen des statistischen Arbeitens. Gabler, Wiesbaden. (Hier finden sich einige Anmerkungen zur Geschichte der Statistik).

Quatember A. (2008) Statistik ohne Angst vor Formeln. Pearson, München. (Auch mit Formeln, aber sehr anschaulich)

von Randow G. (2005) Das Ziegenproblem. Rowohlt, Hamburg. (Kurzweiliges Büchlein zu Phänomenen der Wahrscheinlichkeitsrechnung und Statistik)

Sachs M. (2003) Wahrscheinlichkeitsrechnung und Statistik für Ingenieurstudenten an Fachhochschulen. Fachbuchverlag Leipzig/Hanser, München. (Empfehlung insbesondere für die beschreibende Statistik)

Theden P. und Colsman H. (2002) Qualitätstechniken – Werkzeuge zur Problemlösung und ständigen Verbesserung. Hanser, München. (Heftchen zu QM-Techniken)

Weitere Referenzen

Arens T. et al. (2012) Mathematik. Spektrum, Heidelberg. (Umfassende, mathematisch genaue Darstellung der Mathematik, schöne Illustrationen)

Bauer H. (1968) Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie. de Gruyter, Berlin. (Das Standardwerk zur Wahrscheinlichkeitsrechnung für Mathematiker)

Brown J. L. Jr. (1997) On the error in reconstructing a non-bandlimited function by means of the bandpass sampling theorem. J. Math. Anal. Appl. 18 , S. 75–84. Erratum Ibid. 21 (1968), S. 699. (Erweiterung des Abtastsatzes für nicht-bandbeschränkte Funktionen)

Burg K., Haf H. und Wille F. (2004) Funktionentheorie. Vieweg, Wiesbaden. (Komplexe Analysis für Ingenieure und Naturwissenschaftler)

Burg K., Haf H. und Wille F. (2009) Partielle Differentialgleichungen und funktionalanalytische Grundlagen. Vieweg+Teubner, Wiesbaden. (In diesem Buch nicht behandelte Themen der Ingenieurmathematik)

Butz T. (2007) Fouriertransformation für Fußgänger. Teubner, Wiesbaden. (Viele Aspekte der Fourier-Analysis)

Butzer P. L., Splettstößer W. und Stens R. L. (1988) The sampling theorem and linear prediction in signal analysis. Jahresber. Deutsch. Mathe.-Verein. 90, S. 1–70. (Erweiterung des Abtastsatzes für nicht-bandbeschränkte Funktionen)

Ebbinghaus H.-D. et al. (1992) Zahlen. Springer, Berlin Heidelberg. (Ein ganzes Buch über Zahlen mit vielen historischen Anmerkungen)

Endl K. und Luh W. (1989) Analysis I. Aula, Wiesbaden. (Mathematisch genaue Darstellung der Differenzialrechnung)

Erwe F. (1962) Differential- und Integralrechnung, Band 1: Elemente der Infinitesimalrechnung und Differentialrechnung. Bibliographisches Institut, Mannheim. (Mathematisch genaue Einführung in die Analysis)

Müller-Fonfara R. (2008) Mathematik verständlich. Bassermann Verlag, München. (Leicht verständliche Einführung in Analysis, Lineare Algebra und Wahrscheinlichkeitsrechnung)

Freud R. (Hrsg.) (1990) Große Augenblicke aus der Geschichte der Mathematik. BI-Wissenschaftsverlag, Mannheim. (Unterhaltsam und gleichzeitig informativ)

Gramlich G. (2009) Lineare Algebra: Eine Einführung. Fachbuchverlag Leipzig/Hanser, München.

Hanke-Bourgeois M. (2008) Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens. Vieweg+Teubner, Wiesbaden.

Heuser H. (2009) Lehrbuch der Analysis Teil 1. Teubner, Wiesbaden. (Ein Klassiker zur Differenzial- und Integralrechnung)

Heuser H. (2004) Lehrbuch der Analysis Teil 2. Teubner, Wiesbaden. (Differenzial- und Integralrechnung mit mehreren Variablen, Fourier-Analysis, Geschichte der Analysis)

Hohloch E. et al. (2009) Brücken zur Mathematik Band 1–7. Cornelsen, Berlin.

Knorrenschild M. (2010) Numerische Mathematik: Eine beispielorientierte Einführung. Fachbuchverlag Leipzig/Hanser, München. (Schöne und kompakte Einführung in die Numerik für Ingenieure)

Kolmogorov A. (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin Heidelberg. (Die Geburtsstunde der modernen Wahrscheinlichkeitsrechnung)

Krämer W. (1992) Wie lügt man mit Statistik. In: Stochastik in der Schule 11, S. 3–24. (Schöne Beispiele)

Logothetis N. (1992) Managing for Total Quality: from Deming to Taguchi and SPC. Prentice Hall, Englewood Cliffs, NJ. (QM-Techniken, insbesondere statistische Prozesslenkung)

Louis A. K., Maaß P. und Rieder A. (1998) Wavelets. Teubner, Stuttgart. (Mathematisch fundierte Einführung in die Theorie der Wavelets)

Natanson I. P. (1955) Konstruktive Funktionentheorie. Akademie-Verlag, Berlin. (Klassische Speziallektüre zur Approximationstheorie und Fourieranalysis)

Schenk J. und Rigoll G. (2010) Mensch-Maschine-Kommunikation. Springer, Berlin Heidelberg.

Schüffler K. (1991) Mathematik in der Wirtschaftswissenschaft. Hanser, München.

Stein E. M. und Shakarchi R. (2003) Fourier Analysis – An Introduction, Princeton University Press, Princeton, NJ.

Strauss. W. A. (1995) Partielle Differentialgleichungen. Vieweg, Wiesbaden. (Beispielorientierte Einführung in das Gebiet der partiellen Differenzialgleichungen)

Zygmund A. (2002) Trigonometric Series. Cambridge University Press, Cambridge. (Der Klassiker zu Fourier-Reihen von 1935 in der dritten Auflage)

Index

http://www.springer.com/978-3-8274-3007-6

Mathematik verstehen und anwenden – von den Grundlagen bis zu Fourier-Reihen und Laplace-Transformation

Goebbels, S.; Ritter, S.

2013, XII, 948 S. 215 Abb., Softcover
ISBN: 978-3-8274-3007-6

http://www.springer.com/978-3-662-57393-8

Mathematik verstehen und anwenden – von den Grundlagen bis zu Fourier-Reihen und Laplace-Transformation

Goebbels, S.; Ritter, S.

2018, XIII, 1099 S. 235 Abb., 28 Abb. in Farbe.,
Softcover
ISBN: 978-3-662-57393-8