

Steffen Goebbels
Stefan Ritter

Mathematik verstehen und anwenden

von den Grundlagen bis zu Fourier-Reihen
und Laplace-Transformation

3. Auflage



Springer Spektrum

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

Steffen Goebbels · Stefan Ritter

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

3., überarbeitete und erweiterte Auflage

Steffen Goebbels
Fachbereich Elektrotechnik und Informatik
Hochschule Niederrhein
Krefeld, Deutschland

Stefan Ritter
Fakultät für Elektro- und Informationstechnik
Hochschule Karlsruhe
Karlsruhe, Deutschland

ISBN 978-3-662-57393-8 ISBN 978-3-662-57394-5 (eBook)
<https://doi.org/10.1007/978-3-662-57394-5>

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Springer Spektrum

1. Aufl.: © Spektrum Akademischer Verlag Heidelberg 2011

2. und 3. Aufl.: © Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 2013, 2018

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Verantwortlich im Verlag: Andreas Rüdinger

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Spektrum ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature

Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Vorwort zur ersten Auflage

Sie halten ein weiteres Buch in den Händen, das in die Höhere Mathematik einführt. Falls Sie es nicht schon gekauft oder ausgeliehen haben, würden wir uns freuen, wenn Sie es täten. Keine Sorge – reich machen Sie uns damit nicht (insbesondere dann nicht, wenn Sie es nur ausleihen). Aber vielleicht hilft es Ihnen beim Einstieg ins Studium und später als Nachschlagewerk. Es gibt viele und manche sehr gute Bücher über Höhere Mathematik. Einige davon sind im Literaturverzeichnis aufgelistet. Wir maßen uns nicht an zu sagen, dass unseres besser ist. Wir freuen uns auch, wenn Sie es nur als Zweitbuch auswählen. Was das Buch von einigen anderen Werken unterscheidet, ist die Bandbreite. Da es aus dem Unterricht in den Bachelor-Studiengängen Maschinenbau, Elektrotechnik und Mechatronik an der Hochschule Karlsruhe und der Hochschule Niederrhein entstanden ist, berücksichtigt es die Einstiegsschwierigkeiten von Studierenden mit lückenhaften Vorkenntnissen und motiviert die Inhalte mit praktischen Beispielen aus den Ingenieurfächern. Gehören Sie zu dieser Gruppe, dann lassen Sie beim Lesen die ausführlichen Beweise zunächst aus. Wenn Sie tiefer in die Mathematik einsteigen wollen (oder müssen) und Sie die Verfahren wirklich verstehen wollen, finden Sie über die kommentierten Beweise hinaus ein reichhaltiges Angebot. Themen, die über ein Minimalprogramm (das das Überleben im Studium sicher stellt) hinausgehen, sind mit einem Stern (*) gekennzeichnet. Einige dieser Inhalte sind mathematischer Natur, andere stellen einen Bezug zu Anwendungen aus der Technik her. Studieren Sie eine Naturwissenschaft, so sehen Sie hier, wofür man die Mathematik praktisch benötigt. Darüber hinaus bieten die Kästen noch zusätzliche Hintergrundinformationen und weiteres Material zur Vertiefung des Stoffs.

Im ersten Kapitel werden Grundlagen wie Logik, Mengenlehre und Zahlen auf dem Niveau eines Mathematik-Vorkurses behandelt. Auch wenn Sie gute Vorkenntnisse haben, sollten Sie dieses Kapitel als Erstes durchblättern. Unserer Erfahrung nach werden hier die meisten Klausurfehler gemacht. Vielleicht sind auch einige Themen wie komplexe Zahlen oder Determinanten neu für Sie.

Danach können Sie entweder mit der Analysis in Kapitel 2 oder mit der Linearen Algebra in Kapitel 3 weitermachen. Die Analysis beschäftigt sich mit Grenzwerten, kümmert sich also um das unendlich Kleine und Große. Dazu gehört insbesondere die Differenzial- und Integralrechnung (Umgang mit momentanen Änderungen). Die Lineare Algebra benötigt man z. B. beim Lösen von linearen Gleichungssystemen, wie sie beispielsweise bei der Berechnung von Spannungen und Strömen in elektrischen Netzwerken auftreten.

Die weiteren Kapitel sind überwiegend unabhängig voneinander, setzen aber die Sätze der Analysis aus Kapitel 2 und einige Aussagen der Linearen Algebra aus Kapitel 3 voraus. Diese Abschnitte lesen sich natürlich am leichtesten der vorgegebenen Nummerierung folgend. In Kapitel 4 erweitern wir die Analysis aus Kapitel 2 auf Funk-

tionen mit mehreren Variablen, wie sie in unserer dreidimensionalen Welt auftreten. Viele Zusammenhänge in der Natur beschreiben Veränderungen und lassen sich als Differenzialgleichungen modellieren. Dazu sehen wir uns in Kapitel 5 einige ausgewählte Lösungsverfahren an. Die Fourier-Analyse nimmt aufgrund ihrer praktischen Bedeutung mit Kapitel 6 einen breiten Raum ein. Hier zerlegt man eine Schwingung in die einzelnen Frequenzen, aus denen sie zusammengesetzt ist. Das Buch schließt in Kapitel 7 mit einer kurzen Einführung in die Wahrscheinlichkeitsrechnung und Statistik, die man beispielsweise bei Simulationen, in der digitalen Signalverarbeitung und im Qualitätsmanagement benötigt.

Vorwort zur dritten Auflage

Nachdem wir in der zweiten Auflage insbesondere die Kapitel zur Fourier-Analyse und schließenden Statistik erweitert haben, wurde jetzt der Umfang des Buchs den Anregungen von Lesern folgend erheblich erweitert. Für einige neue Abbildungen haben wir Geobasisdaten der Kommunen und des Landes NRW verwendet, die im Rahmen einer Open-Data-Initiative bei GeoBasis NRW kostenlos bezogen werden können.

Zu jedem Kapitel finden Sie eine Aufgabensammlung. Die Lösungen sowie Korrekturen zu den bislang erschienenen Auflagen stehen auf der Internetseite zum Buch zur Verfügung: <http://www.springer.com/978-3-662-57393-8>

Dank

Wir möchten unseren Mitarbeitern und Kollegen in Karlsruhe und Krefeld danken, die uns bei der Erstellung des Buchs unterstützt haben. Ebenso bedanken wir uns bei vielen Studierenden und Tutoren für ihre Anregungen und konstruktive Kritik. Besonderer Dank gilt M.Eng. Michael Gref, Prof. Dr. Knut Schumacher, Prof. Dr. Roland Hoffmann, Prof. Dr. Johannes Blanke Böhne, Prof. Dr. Pohle-Fröhlich, Prof. Dr. Christoph Dalitz, Prof. Dr. Jochen Rethmann, Prof. Dr. Peer Ueberholz, Prof. Dr. Karlheinz Schöffler, Dipl.-Ing. Ralph Radmacher, Dipl.-Ing. Guido Janßen sowie Prof. Dr. Lorens Imhof und nicht zuletzt unseren Lehrern Prof. Dr. Rolf Joachim Nessel und Prof. Dr. Erich Martensen.

Wir haben eine Fülle von Beispielen verwendet, die sich im Laufe der Jahre angesammelt haben und deren Ursprung nicht immer nachvollziehbar war. Sollten wir hier Autoren unwissentlich zitieren, möchten wir uns dafür entschuldigen.

Zum Schluss möchten wir uns noch ganz besonders bei Herrn Dr. Rüdinger und Frau Lühker vom Springer-Verlag für die engagierte Unterstützung des Buchprojekts bedanken.

Inhaltsverzeichnis

Vorwort	v
1 Grundlagen	1
1.1 Mengenlehre	1
1.1.1 Mengenbegriff	2
1.1.2 Mengenoperationen	4
1.1.3 Abbildungen	7
1.2 Logik	12
1.2.1 Aussagenlogik	12
1.2.2 Prädikatenlogik	18
1.2.3 Beweise	23
1.3 Reelle Zahlen	25
1.3.1 Natürliche und ganze Zahlen	25
1.3.2 Rationale Zahlen	34
1.3.3 Reelle Zahlen	44
1.4 Rechnen mit reellen Zahlen	55
1.4.1 Potenzen und Wurzeln	55
1.4.2 Summen und Produkte, Binomischer Lehrsatz	57
1.4.3 Beträge und Ungleichungen	65
1.4.4 Über das Lösen von Gleichungen und Ungleichungen	71
1.5 Reelle Funktionen	77
1.5.1 Notation reeller Funktionen	77
1.5.2 Eigenschaften von reellen Funktionen	80
1.5.3 Umkehrfunktion	85
1.5.4 Verkettung von Funktionen	87
1.5.5 Signum- und Betragsfunktion	89
1.5.6 Polynome und gebrochen-rationale Funktionen	90
1.5.7 Potenz- und Wurfelfunktionen	101
1.5.8 Exponentialfunktionen und Logarithmen	102
1.5.9 Trigonometrische Funktionen	112
1.5.10 Hyperbel- und Areafunktionen	128
1.6 Komplexe Zahlen	131
1.6.1 Erweiterung der reellen Zahlen um eine imaginäre Einheit	132
1.6.2 Komplexe Arithmetik	133
1.6.3 Die Gauß'sche Zahlenebene	135
1.6.4 Euler'sche Gleichung und Polarform komplexer Zahlen	138
1.6.5 Komplexe Wechselstromrechnung*	144
1.6.6 Fundamentalsatz der Algebra	147
1.7 Lineare Gleichungssysteme und Matrizen	152

1.7.1	Lineare Gleichungssysteme	152
1.7.2	Matrizen, Zeilen- und Spaltenvektoren	154
1.7.3	Lösen linearer Gleichungssysteme	161
1.7.4	Inverse Matrix und transponierte Matrix	168
1.7.5	Symmetrische und orthogonale Matrizen	173
1.7.6	Dreiecksmatrizen, Bandmatrizen und LR-Zerlegung *	176
1.8	Determinanten	181
1.8.1	Definition und elementare Eigenschaften von Determinanten	182
1.8.2	Determinanten und lineare Gleichungssysteme	193
1.9	Aufgaben	199
2	Differenzial- und Integralrechnung	211
2.1	Folgen	211
2.1.1	Definition und Grundbegriffe von Folgen	212
2.1.2	Konvergenz und Divergenz von Folgen	216
2.1.3	Rechnen mit konvergenten Folgen	220
2.1.4	Konvergenzkriterien	223
2.1.5	Die Euler'sche Zahl e als Grenzwert von Folgen	226
2.1.6	Approximation reeller Potenzen	228
2.1.7	Bestimmte Divergenz	229
2.1.8	Häufungspunkte einer Folge *	232
2.1.9	Folgenkompaktheit und Cauchy-Folgen *	232
2.2	Zahlen-Reihen	236
2.2.1	Definition und Konvergenz einer Reihe	237
2.2.2	Rechnen mit konvergenten Reihen	240
2.2.3	Alternativen zur Definition der Reihenkonvergenz	241
2.2.4	Absolute Konvergenz	243
2.2.5	Konvergenzkriterien für Reihen	245
2.3	Grenzwerte von Funktionen und Stetigkeit	255
2.3.1	Umgebungen und Überdeckungen	255
2.3.2	Grenzwerte von Funktionen	257
2.3.3	Stetigkeit	270
2.3.4	Eigenschaften stetiger Funktionen	278
2.3.5	Unstetigkeitsstellen	285
2.4	Differenzierbarkeit und Ableitungen	288
2.4.1	Ableitung als Grenzwert des Differenzenquotienten	289
2.4.2	Ableitungsregeln	295
2.4.3	Newton-Verfahren	305
2.4.4	Das Differenzial	307
2.4.5	Höhere Ableitungen	310
2.5	Zentrale Sätze der Differenzialrechnung	314

2.5.1	Satz von Fermat: notwendige Bedingung für lokale Extrema	314
2.5.2	Mittelwertsätze der Differenzialrechnung	315
2.5.3	Regeln von L'Hospital	322
2.6	Integralrechnung	328
2.6.1	Definition des Integrals	329
2.6.2	Eigenschaften des Integrals	334
2.6.3	Hauptsatz der Differenzial- und Integralrechnung	339
2.6.4	Rechenregeln zur Integration	343
2.6.5	Numerische Integration	359
2.6.6	Uneigentliche Integrale	362
2.6.7	Volumen und Flächen	369
2.6.8	Lebesgue-Integral *	373
2.7	Satz von Taylor, Kurvendiskussion und Extremalprobleme	382
2.7.1	Taylor-Summen	382
2.7.2	Kurvendiskussion und Extremalprobleme	387
2.8	Potenzreihen	398
2.8.1	Unendliche Taylor-Summen und Potenzreihen	398
2.8.2	Einschub: Funktionenfolgen *	402
2.8.3	Konvergenz von Potenzreihen	411
2.8.4	Differenziation und Integration von Potenzreihen	415
2.8.5	Der Zusammenhang zwischen Potenzreihen und Taylor-Reihen	417
2.8.6	Die komplexe Exponentialfunktion	418
2.9	Aufgaben	420
3	Lineare Algebra	427
3.1	Vektoren in der Ebene und im Raum	427
3.1.1	Vektoren: Grundbegriffe und elementare Rechenregeln	427
3.1.2	Skalarprodukt und Orthogonalität	435
3.1.3	Vektorprodukt und Spatprodukt	442
3.1.4	Anwendungen des Skalar-, Vektor- und Spatprodukts	450
3.2	Analytische Geometrie	452
3.2.1	Geraden in der Ebene und im Raum	453
3.2.2	Ebenen im Raum	460
3.3	Vektorräume	466
3.3.1	Definition des Vektorraums	467
3.3.2	Lineare Unabhängigkeit, Basis und Dimension	474
3.3.3	Skalarprodukt und Norm	483
3.3.4	Orthogonalität, Orthogonal- und Orthonormalsysteme	488
3.4	Lineare Abbildungen	500
3.4.1	Lineare Abbildungen und Matrizen	500
3.4.2	Summe, skalares Vielfaches und Verkettung linearer Abbildungen . .	506

3.4.3	Kern und Bild einer linearen Abbildung, Dimensionssatz	508
3.4.4	Umkehrabbildung und inverse Matrix	515
3.4.5	Koordinaten- und Basistransformationen *	517
3.5	Lösungstheorie linearer Gleichungssysteme	521
3.5.1	Lösungsraum eines linearen Gleichungssystems	523
3.5.2	Berechnung von linearen elektrischen Netzwerken *	529
3.6	Eigenwerte und Eigenvektoren	537
3.6.1	Eigenwerte und Eigenvektoren	538
3.6.2	Diagonalisierung von Matrizen *	548
3.6.3	Hauptvektoren und Jordan-Normalform *	552
3.7	Normierte Vektorräume: Lineare Algebra trifft Analysis *	557
3.7.1	Norm	557
3.7.2	Banach- und Hilbert-Räume	560
3.7.3	L^p -Räume	562
3.7.4	Stetige Abbildungen zwischen normierten Vektorräumen	566
3.7.5	Einige zentrale Sätze der Funktionalanalysis	577
3.7.6	Sobolev-Räume	584
3.8	Aufgaben	585
4	Funktionen mit mehreren Variablen	589
4.1	Grenzwerte und Stetigkeit	592
4.2	Ableitungen von reellwertigen Funktionen mit mehreren Variablen	597
4.2.1	Ableitungsbegriffe	597
4.2.2	Implizite Differenziation und implizite Funktion	609
4.2.3	Höhere Ableitungen	610
4.2.4	Fehlerrechnung *	614
4.3	Extremwertrechnung	617
4.3.1	Lokale und globale Extrema	618
4.3.2	Extrema unter Nebenbedingungen *	631
4.3.3	Lineare Optimierung *	638
4.4	Integralrechnung mit mehreren Variablen	649
4.4.1	Integration über mehrdimensionale Intervalle	649
4.4.2	Integration über Normalbereiche	657
4.4.3	Substitutionsregel	661
4.4.4	Polar-, Zylinder- und Kugelkoordinaten	663
4.4.5	Lebesgue-Integral, L^p - und Sobolev-Räume *	668
4.5	Vektoranalysis	672
4.5.1	Vektorfelder	673
4.5.2	Kurven	674
4.5.3	Quellen, Senken und Wirbel in Vektorfeldern	678
4.5.4	Kurvenintegrale	680

4.5.5	Satz von Green *	688
4.5.6	Flächenintegrale *	690
4.5.7	Die Sätze von Gauß und Stokes *	694
4.6	Aufgaben	701
5	Gewöhnliche Differenzialgleichungen	705
5.1	Einführung	705
5.1.1	Beispiele für Differenzialgleichungen aus Physik und Technik	706
5.1.2	Grundbegriffe	710
5.1.3	Konstruktion einer Lösung, Existenz und Eindeutigkeit	715
5.1.4	Iterationsverfahren von Picard und Lindelöf	718
5.1.5	Runge-Kutta-Verfahren	719
5.2	Lösungsmethoden für Differenzialgleichungen erster Ordnung	721
5.2.1	Lineare Differenzialgleichungen erster Ordnung	722
5.2.2	Nicht-lineare Differenzialgleichungen erster Ordnung	735
5.3	Lineare Differenzialgleichungssysteme	748
5.3.1	Motivation: Eine Schaltung mit Induktivitäten	748
5.3.2	Grundbegriffe	749
5.3.3	Homogene Lösungen	753
5.3.4	Partikuläre Lösungen	757
5.3.5	Komplexe und mehrfache Eigenwerte *	762
5.4	Lineare Differenzialgleichungen höherer Ordnung	770
5.4.1	Lösung über ein lineares Differenzialgleichungssystem	770
5.4.2	Lösung mit einem Ansatz vom Typ der rechten Seite	777
5.4.3	Schwingungsgleichung *	782
5.5	Ausblick: Partielle Differenzialgleichungen und Finite-Elemente-Methode *	788
5.5.1	Eine schwingende Saite: Wellengleichung	788
5.5.2	Partielle Differenzialgleichungen zweiter Ordnung	790
5.5.3	Finite-Elemente-Methode	792
5.5.4	Beispiel für die Finite-Elemente-Methode in 2-D	799
5.6	Aufgaben	809
6	Fourier-Reihen und Integraltransformationen	813
6.1	Fourier-Reihen	814
6.1.1	Fourier-Koeffizienten und Definition der Fourier-Reihe	815
6.1.2	Sinus- und Kosinus-Form der Fourier-Reihe	821
6.1.3	Komplexwertige Funktionen und Fourier-Koeffizienten	823
6.1.4	Faltung	832
6.1.5	Konvergenz von Fourier-Reihen *	840
6.1.6	Gibbs-Phänomen	853
6.1.7	Entwicklung 2p-periodischer Funktionen	859
6.2	Fourier-Transformation	861

6.2.1	Fourier-Integral	861
6.2.2	Fourier-Umkehrtransformation	865
6.2.3	Fourier-Koeffizienten und Fourier-Transformation	867
6.2.4	Eigenschaften der Fourier-Transformation	869
6.2.5	Faltung	874
6.3	Laplace-Transformation	878
6.3.1	Von der Fourier- zur Laplace-Transformation	878
6.3.2	Rechnen mit der Laplace-Transformation	882
6.3.3	Laplace-Transformation in der Systemtheorie *	894
6.4	Diskrete Fourier-Transformation	903
6.4.1	Ausgangspunkt: Koeffizienten einer Fourier-Reihe	905
6.4.2	Diskrete Fourier-Transformation	908
6.4.3	Diskrete Faltung *	919
6.4.4	FFT-Algorithmus	923
6.4.5	Numerische Berechnung von Fourier-Koeffizienten	928
6.4.6	Abtastatz für trigonometrische Polynome	930
6.4.7	Abtastung 2p-periodischer Funktionen und Leck-Effekt (Leakage) *	937
6.4.8	Numerische Berechnung der Fourier-Transformation	939
6.4.9	Abtastatz der Fourier-Transformation	941
6.4.10	Leck-Effekt und Fensterfunktionen *	950
6.4.11	Zusammenfassung	954
6.5	Wavelets und schnelle Wavelet-Transformation *	954
6.5.1	Idee der Wavelet-Transformation	955
6.5.2	Eindimensionale Wavelet-Transformation mit orthogonalen Wavelets	959
6.5.3	Zweidimensionale diskrete Wavelet-Transformation	963
6.6	Aufgaben	965
7	Wahrscheinlichkeitsrechnung und Statistik	969
7.1	Beschreibende Statistik	970
7.1.1	Grundbegriffe	970
7.1.2	Empirische Verteilungsfunktionen	975
7.1.3	Lageparameter	977
7.1.4	Streuungsparameter	982
7.1.5	Zweidimensionale Häufigkeitsverteilungen und Korrelation	984
7.1.6	Kovarianzmatrix	988
7.1.7	Lineare Regressionsrechnung	992
7.2	Wahrscheinlichkeitsrechnung	997
7.2.1	Zufallsexperimente und Ereignisse	997
7.2.2	Wahrscheinlichkeit und Satz von Laplace	999
7.2.3	Kombinatorik	1003
7.2.4	Unabhängige Ereignisse und bedingte Wahrscheinlichkeiten	1008

7.2.5	Zufallsvariablen	1018
7.2.6	Lage- und Streuungsparameter von Zufallsvariablen	1032
7.2.7	Gesetz der großen Zahlen	1042
7.2.8	Zentraler Grenzwertsatz	1047
7.2.9	Integrale über Zufallsvariablen *	1054
7.3	Schließende Statistik	1056
7.3.1	Punktschätzungen	1057
7.3.2	Begriffe der Fehlerrechnung *	1061
7.3.3	Intervallschätzungen	1063
7.3.4	Hypothesentests	1071
7.4	Aufgaben	1076
	Literaturverzeichnis	1083
	Index	1087

Einige Aspekte der **numerischen Mathematik** finden sich über das Buch verteilt:

- Kondition, Stabilität und Konsistenz 573
- Direktes Lösen linearer Gleichungssysteme: Gauß-Algorithmus 162; Cramer'sche Regel 197; LR-Zerlegung 176; Cholesky-Zerlegung 627
- Iteratives Lösen linearer Gleichungssysteme: Fixpunktiterationen, Jacobi-, Gauß-Seidel-, Richardson-Verfahren 179, 575; Gradientenverfahren, Verfahren der konjugierten Gradienten 623
- Nullstellensuche: Intervallschachtelung (Bisektion) 279; Newton-Verfahren und Levenberg-Marquard-Verfahren 305, 637
- Interpolation: algebraische Interpolation 92; Splines 312; trigonometrische Interpolation 930
- Ausgleichsrechnung, Regressionsrechnung: Gauß'sche Normalgleichungen 996; Hauptachsentransformation 991
- Quadraturverfahren: summierte Mittelpunkts-, Trapez- und Simpson-Regel 359
- Numerisches Lösen von Differenzialgleichungen: Cauchy-Euler-Polygonzugverfahren 715; Runge-Kutta-Verfahren, Verfahren von Heun 719; Differenzenverfahren und Finite-Elemente-Methode 792
- Optimierung: Gradientenabstieg 603, 623; Lagrange-Multiplikatoren 631; Lineare Optimierung 638
- Integraltransformationen: FFT-Algorithmus 923; schnelle Wavelet-Transformation 954



1 Grundlagen

Übersicht

1.1	Mengenlehre	1
1.2	Logik	12
1.3	Reelle Zahlen	25
1.4	Rechnen mit reellen Zahlen	55
1.5	Reelle Funktionen	77
1.6	Komplexe Zahlen	131
1.7	Lineare Gleichungssysteme und Matrizen	152
1.8	Determinanten	181
1.9	Aufgaben	199

In diesem Kapitel wiederholen wir den Schulstoff bis zum Beginn der Oberstufe. Das wäre allerdings langweilig, wenn wir nicht schon vor dem Hintergrund der späteren Anwendungen darüber hinausgehende Inhalte einflechten würden (z. B. das Rechnen mit komplexen Zahlen). Außerdem wird in diesem Kapitel eine korrekte mathematische Schreibweise eingeführt. Vielfach herrscht Verwirrung, wann man ein Gleichheitszeichen, wann ein Folgerungszeichen und wann ein Äquivalenzzeichen benutzt. Deshalb beginnen wir mit Grundbegriffen aus Mengenlehre und Logik. Dann kommen wir zu Zahlen und zum Rechnen.

1.1 Mengenlehre

Beim Beschreiben der Objekte, mit denen wir uns beschäftigen werden, helfen Mengennotationen. Sie bilden die Basis der in der Mathematik verwendeten Sprache.

1.1.1 Mengenbegriff

Definition 1.1 (Mengenbegriff von Cantor, 1845–1918, hier 1895)

Eine **Menge** M ist eine gedankliche Zusammenfassung von unterscheidbaren Objekten. Diese Objekte nennt man **Elemente** von M .

Eine Menge M kann beschrieben werden durch Auflistung der Elemente. Diese Auflistung wird in geschweifte Klammern gesetzt. Die Menge, die aus den Zahlen 1, 2 und 3 gebildet wird, ist $\{1, 2, 3\}$. Die Menge B aller Buchstaben lautet $\{a, b, c, \dots, z\}$.

Definition 1.2 (Mengenschreibweisen)

- Wir schreiben $x \in M$, um auszudrücken, dass x ein Element der Menge M ist und $x \notin M$, um zu sagen, dass x nicht in M liegt, d. h. kein Element der Menge M ist.
- Eine Menge, die kein Element besitzt, heißt **leere Menge** und wird mit $\emptyset$ oder $\{\}$ notiert.
- Zwei Mengen M und N heißen **gleich** ($M = N$) genau dann, wenn sie die gleichen Elemente besitzen. Das Symbol „ $=$ “ wird generell in der Mathematik für Gleichheit, „ $\neq$ “ für Ungleichheit verwendet.
- M heißt **Teilmenge** von N , $M \subset N$, genau dann, wenn jedes Element von M auch Element von N ist. Ist M nicht Teilmenge von N , so schreibt man $M \not\subset N$.
- Für $M \subset N$ ist das **Komplement** von M bezüglich N definiert als Menge aller Elemente aus N , die nicht in M enthalten sind. Diese Menge wird mit $\overline{M}$ oder alternativ mit $\mathcal{C}_N M$ bezeichnet. Wenn aus dem Zusammenhang die Menge N bekannt ist, ist die Kurzschreibweise $\overline{M}$ verständlich. Anderenfalls ist es besser, die Menge N explizit mit $\mathcal{C}_N M$ anzugeben.
- M und N heißen **disjunkt** (**elementfremd**) genau dann, wenn sie keine gemeinsamen Elemente besitzen.
- Die **Potenzmenge** $\mathcal{P}(M)$ einer Menge M ist die Menge, die alle Teilmengen von M enthält. Sie ist also insbesondere eine Menge, deren Elemente selbst wieder Mengen sind.

Gemäß dieser Definition ist auch $N \subset N$. Viele Autoren verwenden die Schreibweise $M \subset N$ nur, falls M und N zusätzlich verschieden sind. $M \subseteq N$ erlaubt dann auch Gleichheit. Hier verwenden wir ausschließlich das Symbol $\subset$ für beide Situationen.

Beispiel 1.1

Wir betrachten die Menge aller Vokale $V = \{a, e, i, o, u\}$ und die Menge aller Buchstaben des Alphabets $B = \{a, b, c, \dots, z\}$:

- Das Element $a \in V$ ist ein Vokal, aber $b \notin V$ ist ein Konsonant.
- Es ist $V \subset B$, denn jeder Vokal ist gleichzeitig auch ein Buchstabe.

- Dagegen ist $B \not\subset V$, denn es gibt Buchstaben, die keine Vokale sind.
- Das Komplement von V bezüglich B ist die Menge aller Buchstaben, die nicht Vokale sind. $\overline{V} = C_B V$ besteht aus den Konsonanten. ■

Aufgrund der Definition der Gleichheit spielt die Reihenfolge, in der Elemente angegeben werden, keine Rolle. Es reicht, jedes Element genau einmal anzugeben.

Satz 1.1 (Potenzmenge)
Eine Menge M mit n Elementen besitzt 2^n verschiedene Teilmengen, d. h., die Potenzmenge $\mathcal{P}(M)$ besitzt 2^n Elemente.

Beweis: Für jedes Element von M kann man entscheiden, ob das Element in eine Teilmenge aufgenommen werden soll. Diese n Entscheidungen zwischen zwei Alternativen führen zu den 2^n verschiedenen Teilmengen, siehe Abbildung 1.1. □

Beispiel 1.2
 $\mathcal{P}(\{a, b, c\}) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, c\}, \{a, b, c\}\}$ hat $2^3 = 8$ Elemente. Jedes Element ist selbst eine Menge, nämlich eine Teilmenge von $\{a, b, c\}$. ■

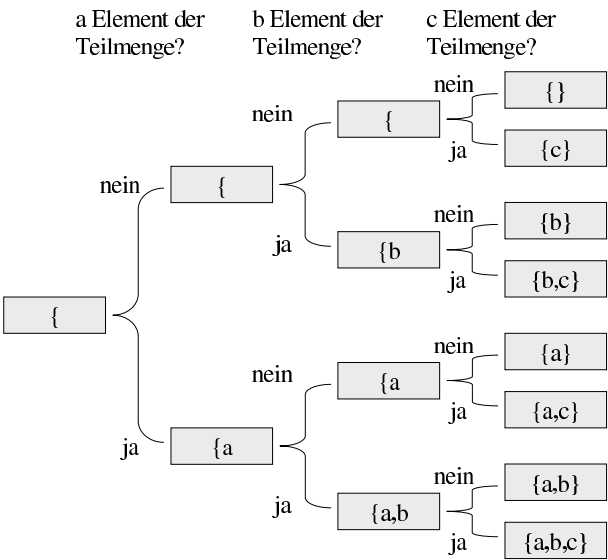


Abb. 1.1: Entscheidungsbaum zum Auffinden aller 2^3 Teilmengen von $\{a, b, c\}$

Oft kann man nicht jedes Element einer Menge M explizit hinschreiben. Man verwendet dann die Schreibweise

$$M = \{x \in G : x \text{ erfüllt eine Bedingung}\}. \tag{1.1}$$

Der Doppelpunkt wird als „wofür gilt“ gesprochen. Dabei ist G eine bereits definierte Grundmenge. Zum Beispiel ist $\{x \in \{1, 2, 3, 4, 5\} : x^2 \in \{4, 9, 25, 36\}\} = \{2, 3, 5\}$. Durch diese Schreibweise vermeidet man mögliche Widersprüche. Dazu gibt es später Hintergrundinformationen auf Seite 5.

1.1.2 Mengenoperationen

Definition 1.3 (Mengenoperationen)

Es seien M, N Mengen.

- Der **Durchschnitt** von M und N (M geschnitten mit N) ist die Menge aller Elemente, die sowohl in M als auch in N enthalten sind:

$$M \cap N = \{x : x \in M \text{ und } x \in N\}.$$

- In der **Vereinigung** von M und N (M vereinigt mit N) sind genau alle Elemente beider Mengen enthalten:

$$M \cup N = \{x : x \in M \text{ oder } x \in N\}.$$

- Die **Differenz** von M und N (M ohne N) entsteht, indem man aus M alle Elemente entfernt, die in N enthalten sind.

$$M \setminus N = \{x : x \in M \text{ und } x \notin N\}.$$

- Das **Kreuzprodukt** von M und N (M Kreuz N) ist die Menge

$$M \times N = \{(x, y) : x \in M \text{ und } y \in N\}.$$

Die Elemente (x, y) von $M \times N$ sind **Paare** von Elementen $x \in M$ und $y \in N$.

- Das **n -fache Kreuzprodukt** von M (M hoch n) ist

$$M^n = \{(x_1, x_2, \dots, x_n) : x_1, x_2, \dots, x_n \in M\}.$$

Die Elemente von M^n sind **n -Tupel** von Elementen aus M .

Diese Operationen kann man durch **Venn-Diagramme** veranschaulichen. Die Mengen werden dabei mit Hilfe von Kreisen oder Ellipsen dargestellt, siehe Abbildung 1.2. Die Punkte innerhalb eines Kreises sind die Elemente der durch ihn repräsentierten Menge. Eine Schnittmenge $A \cap B$ besteht z. B. aus den Punkten der Überlappung der Kreisflächen der Mengen A und B .

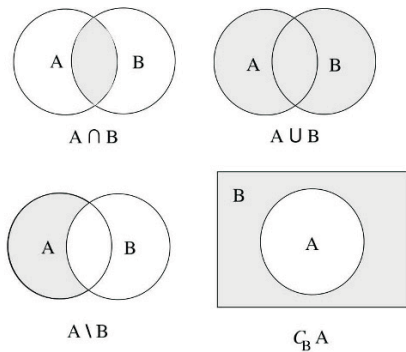


Abb. 1.2: Mengenoperationen, dargestellt als Venn-Diagramme

Beispiel 1.3

Für $M = \{1, 2, 3\}$ und $N = \{2, 3, 4\}$ erhalten wir:

- a) $M \cap N = \{2, 3\}$ und $M \cup N = \{1, 2, 3, 4\}$,
- b) $M \setminus N = \{1\}$ und $N \setminus M = \{4\}$,
- c) $M \times N = \{(x, y) : x \in M \text{ und } y \in N\}$ bzw.

$$M \times N = \{(1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), (3, 2), (3, 3), (3, 4)\},$$

- d) $\mathcal{P}(M) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, M\}$.

■

Hintergrund: Die Allmenge

Die Angabe einer existierenden Grundmenge bei der Definition einer neuen Menge wie in (1.1) verhindert ein schwerwiegendes Problem der Mengenlehre: Ohne Grundmenge könnte man auf die Idee kommen, die Menge aller Mengen, die sogenannte **Allmenge** zu definieren. Die Existenz der Allmenge führt aber zu unlösbaren Widersprüchen, so dass diese Menge nicht existieren kann. Wir nehmen an, die Allmenge A würde existieren. Die Elemente der Allmenge sind wie bei der Potenzmenge Mengen. Deren Elemente können auch wieder Mengen sein usw. Wir können damit A in zwei Teilmengen A_1 und A_2 zerlegen, wobei A_1 die Menge aller Mengen ist, die sich selbst nicht enthalten. A_2 ist die Menge aller Mengen, die sich selbst enthalten:

$$A = A_1 \cup A_2 \text{ mit } A_1 = \{x \in A : x \notin x\}, A_2 = \{x \in A : x \in x\}.$$

A_1 und A_2 sind disjunkt, da sich eine Menge nicht gleichzeitig selbst enthalten und nicht selbst enthalten kann. In welcher der beiden Mengen ist nun A_1 ? Falls $A_1 \in A_1$, dann ist laut Definition dieser Menge $A_1 \notin A_1$. Also muss $A_1 \notin A_1$ und $A_1 \in A_2$ sein. Nach Definition von A_2 ist aber dann $A_1 \in A_1$ im Widerspruch zu $A_1 \notin A_1$. Die Annahme, dass A existiert, hat zu diesem Widerspruch geführt. A kann daher nicht existieren.

Wir haben den Mengenbegriff ohne ein Axiomensystem kennengelernt. Durch die fehlende mathematische Präzisierung können solche Probleme dann auftreten. Die Axiome von Zermelo-Fraenkel beheben diesen Missstand, aber auch hier stößt die Mathematik an ihre Grenzen, da man die Widerspruchsfreiheit der Axiome nicht beweisen kann.

Satz 1.2 (Eigenschaften von Mengen)

Es seien M, N und K Mengen. Dann gilt:

- Aus $M \subset N$ und $N \subset K$ folgt $M \subset K$.
- Aus $M \subset N$ und $N \subset M$ folgt $M = N$.
- **Kommutativgesetz** (Vertauschung der Reihenfolge):

$$M \cup N = N \cup M, \quad M \cap N = N \cap M.$$

- **Assoziativgesetz** (andere Klammerung):

$$M \cup (N \cup K) = (M \cup N) \cup K, \quad M \cap (N \cap K) = (M \cap N) \cap K.$$

- **Distributivgesetz** (Ausmultiplikation):

$$M \cap (N \cup K) = (M \cap N) \cup (M \cap K), \quad M \cup (N \cap K) = (M \cup N) \cap (M \cup K).$$

- Für $M, N \subset K$ gelten die **De Morgan'schen Regeln**:

$$\mathcal{C}_K(M \cup N) = \mathcal{C}_K M \cap \mathcal{C}_K N, \quad \mathcal{C}_K(M \cap N) = \mathcal{C}_K M \cup \mathcal{C}_K N.$$

In diesem Sinne wird unter der Bildung des Komplements aus der Vereinigung ein Durchschnitt und umgekehrt.

Auf den rechten Seiten der De Morgan'schen Regeln haben wir auf Klammern verzichtet. Dabei verwenden wir die Konvention, dass das Komplement am stärksten bindet, außerdem bindet der Schnitt enger als die Vereinigung: $\mathcal{C}_K A \cup B \cap C = (\mathcal{C}_K A) \cup (B \cap C)$.

Beweis: Wir zeigen exemplarisch das erste Distributivgesetz. Ist x ein Element von $M \cap (N \cup K)$, dann liegt x sowohl in M als auch in mindestens einer der beiden Mengen N oder K . Falls x Element von N ist, dann ist es auch Element von $M \cap N$. Anderenfalls muss es in $M \cap K$ liegen. In jedem Fall ist x also in $(M \cap N) \cup (M \cap K)$. Wir haben damit gezeigt, dass

$$M \cap (N \cup K) \subset (M \cap N) \cup (M \cap K).$$

Ist umgekehrt x ein Element der rechten Menge, so liegt es in $M \cap N$ oder in $M \cap K$ (oder in beiden Mengen). In jedem Fall liegt x in M und in mindestens einer der beiden Mengen N oder K , also in $M \cap (N \cup K)$, so dass auch

$$(M \cap N) \cup (M \cap K) \subset M \cap (N \cup K)$$

und damit die Gleichheit gezeigt ist. $\square$

Mit den Distributivgesetzen kann man unter Zuhilfenahme der anderen Regeln Mengenoperationen durch „Ausmultiplizieren“ umformen, z. B. so:

$$\begin{aligned} (A \cup B) \cap (C \cup D) &= [(A \cup B) \cap C] \cup [(A \cup B) \cap D] \\ &= (A \cap C) \cup (B \cap C) \cup (A \cap D) \cup (B \cap D), \\ (A \cap B) \cup (C \cap D) &= (A \cup C) \cap (B \cup C) \cap (A \cup D) \cap (B \cup D). \end{aligned}$$

Dies entspricht dem Ausmultiplizieren von reellen Zahlen

$$(a + b) \cdot (c + d) = a \cdot c + b \cdot c + a \cdot d + b \cdot d,$$

wobei man $+$ durch $\cup$ bzw. $\cap$ und $\cdot$ durch $\cap$ bzw. $\cup$ ersetzt.

Beispiel 1.4

Wir vereinfachen den Mengenausdruck $(X \cap ((\mathcal{C}_M X) \cup Y)) \cup (Z \cap (Y \cup Z))$ für Mengen $X, Y, Z \subset M$ mit den Distributivgesetzen:

$$\begin{aligned} &(X \cap ((\mathcal{C}_M X) \cup Y)) \cup (Z \cap (Y \cup Z)) \\ &= ((X \cap (\mathcal{C}_M X)) \cup (X \cap Y)) \cup ((Z \cap Y) \cup (Z \cap Z)) \\ &= \emptyset \cup (X \cap Y) \cup (Y \cap Z) \cup Z = (X \cap Y) \cup Z. \end{aligned} \quad \blacksquare$$

1.1.3 Abbildungen

Vom Zeitpunkt $t = 0$ bis $t = t_1$ fährt ein Zug mit zunächst konstanter Geschwindigkeit (gleichförmige Bewegung), bremst dann aber und kommt genau zum Zeitpunkt t_1 im Bahnhof zum Stehen. Die Abfahrt ist zum späteren Zeitpunkt $t = t_2$. Ab diesem Moment beschleunigt der Zug wieder.

Jedem Zeitpunkt t kann man nun eine zurückgelegte Wegstrecke $s(t)$ zuordnen. So kann die Position des Zugs zu jedem Zeitpunkt bestimmt werden. Die Zeitpunkte werden auf Wegstrecken abgebildet. In Abbildung 1.3 ist diese **Abbildung** der Zeitpunkte auf Wegstrecken $s(t)$ als **Funktionsgraph** dargestellt. Tatsächlich werden bei der Fahrplanerstellung und der Zugüberwachung überlagerte Weg-Zeit-Diagramme aller Züge in einem Streckenbereich verwendet. Daran kann man Zugabstände und Kreuzungspunkte erkennen. Durch Anpassung der Diagramme werden bei Verspätungen neue Planungen vorgenommen.

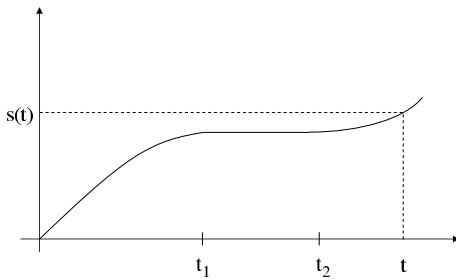


Abb. 1.3: Weg-Zeit-Diagramm

Definition 1.4 (Abbildungen)

Seien E und F nicht-leere Mengen. Eine **Abbildung** (oder **Funktion**) f von E in F (Schreibweise: $f : E \rightarrow F$) ist eine Vorschrift, die jedem Element $x \in E$ eindeutig ein Element $y \in F$ zuordnet (Schreibweisen: $y = f(x)$, $f : x \mapsto y$). Dabei heißt y das **Bild** von x unter f , man bezeichnet y auch als den **Funktionswert** von f im Punkt (oder an der Stelle) x . Man nennt dabei x das **Argument** der Funktion, das ist der Wert, den man in die Funktion „einsetzt“. E heißt der **Definitionsbereich** von f . Man schreibt dafür häufig $E = D(f)$. Ist $E_0 \subset E$, so heißt die Menge

$$f(E_0) = \{f(x) : x \in E_0\} \subset F$$

das **Bild** von E_0 unter der Abbildung f . Das Bild der Menge E heißt die **Wertemenge** oder der **Wertebereich** von f . Man schreibt dafür häufig $W(f)$.

Mit anderen Worten: f ordnet jedem Element des Definitionsbereichs genau ein Element des Wertebereichs zu. Wendet man f auf eine Teilmenge E_0 des Definitionsbereichs an, so erhält man die Teilmenge des Wertebereichs, die die Funktionswerte zu allen Elementen von E_0 enthält.

Beim Weg-Zeit-Diagramm ist E eine Menge von Zeitpunkten und F eine Menge von Strecken. Die Abbildung $s : E \rightarrow F$ ordnet jedem Zeitpunkt eine Strecke zu. Umgekehrt kann man fragen, welche Zeitpunkte zu vorgegebenen Strecken gehören:

Definition 1.5 (Urbild einer Abbildung)

Seien E und F nicht-leere Mengen und $f : E \rightarrow F$ eine Abbildung. Ist $F_0 \subset F$, so heißt die Menge

$$f^{-1}(F_0) := \{x \in E : f(x) \in F_0\} \subset E$$

das **Urbild** von F_0 . Insbesondere ist $f^{-1}(F) = E$.

$f^{-1}(F_0)$ ist also die Menge aller Elemente von E , die von f auf ein Element von F_0 abgebildet werden.

Beispiel 1.5

Sei $f : \{1, 2, 3, 4\} \rightarrow \{3, 4, 5, 6, 7\}$ mit $f(1) = 5$, $f(2) = 5$, $f(3) = 4$, $f(4) = 7$ (siehe Abbildung 1.4). Dann ist $f(\{1, 2, 3, 4\}) = \{4, 5, 7\}$ der Wertebereich von f und $f^{-1}(\{3, 5, 7\}) = f^{-1}(\{5, 7\}) = \{1, 2, 4\}$, $f^{-1}(\{3, 6\}) = \emptyset$, $f(f^{-1}(\{3, 5, 7\})) = f(\{1, 2, 4\}) = \{5, 7\}$. ■

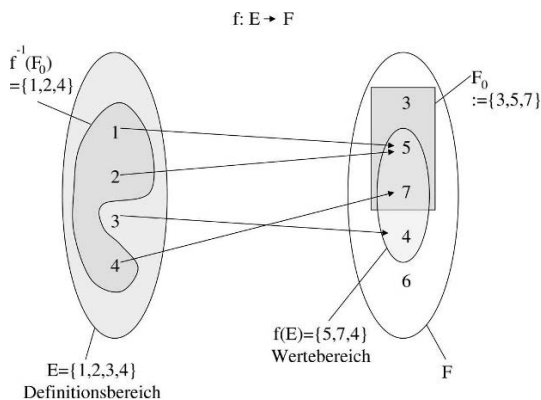


Abb. 1.4: Beispiel zur Definition der Abbildung

Definition 1.6 (Gleichheit)

Zwei Abbildungen $f, g : E \rightarrow F$ heißen gleich ($f = g$) genau dann, wenn sie für jedes Element von E das gleiche Bild in F liefern: $f(x) = g(x)$ für alle $x \in E$.

Gleiche Abbildungen haben insbesondere den gleichen Definitions- und damit den gleichen Wertebereich. Zusätzlich müssen gleiche Abbildungen in die gleiche Zielmenge F abbilden.

Beispiel 1.6

$f : \{1, 2, 3\} \rightarrow \{1, 4, 9\}$ mit $f(1) = 1$, $f(2) = 4$ und $f(3) = 9$ sowie $g : \{1, 2, 3\} \rightarrow \{1, 4, 9\}$ mit $g(x) := x^2$ sind gleiche Abbildungen. Dagegen ist $h : \{1, 2, 3, 4\} \rightarrow \{1, 4, 9, 16\}$ mit $h(x) = x^2$ nicht gleich f oder g . Auch $\tilde{h} : \{1, 2, 3\} \rightarrow \{1, 4, 9, 16\}$ mit $\tilde{h}(x) = x^2$ wäre eine andere Abbildung. Wir werden später allerdings meist die Zielmenge als Wertebereich wählen, so dass diese Pingeligkeit keine Rolle spielt. ■

Gleiche Abbildungen haben die gleichen Eigenschaften:

Definition 1.7 (Eigenschaften von Abbildungen)

Sei $f : E \rightarrow F$.

- f heißt **injektiv** (oder **eindeutig**) genau dann, wenn für je zwei Elemente $x_1, x_2 \in E$ mit $x_1 \neq x_2$ gilt: $f(x_1) \neq f(x_2)$.

- f heißt **surjektiv** oder *Abbildung von E auf F genau dann, wenn zu jedem $y \in F$ mindestens ein $x \in E$ existiert mit $f(x) = y$, d. h. $f(E) = F$.*
- f heißt **bijektiv** genau dann, wenn f injektiv und surjektiv ist.

Injektivität der Abbildung $f : E \rightarrow F$ bedeutet, dass jedes Element von F höchstens einmal als Bild auftritt. Surjektivität bedeutet, dass jedes Element von F mindestens einmal als Bild auftritt. Bei der Bijektivität erscheint jedes Element von F genau einmal als Bild.

Beispiel 1.7

- a) f aus Beispiel 1.5 ist weder injektiv noch surjektiv.
 b) $f : \{1, 2\} \rightarrow \{3, 4, 5\}$ mit $f : 1 \mapsto 3, 2 \mapsto 4$, ist injektiv, aber nicht surjektiv.
 c) $f : \{1, 2\} \rightarrow \{3\}$ mit $f : 1 \mapsto 3, 2 \mapsto 3$, ist surjektiv, aber nicht injektiv. ■

Nach Definition einer Abbildung $f : E \rightarrow F$ gibt es zu jedem $x \in E$ genau ein $y \in F$ mit $y = f(x)$, zu jedem Urbild $x \in E$ gibt es genau ein Bild. Injektivität ist quasi die umgekehrte Eigenschaft: Zu jedem Bild existiert genau ein Urbild. Eine nicht-injektive Abbildung kann man zu einer injektiven machen, indem man den Definitionsbereich einschränkt. Ob dieser Schritt sinnvoll ist, hängt oft vom Zusammenhang ab.

Man kann leicht eine Abbildung in eine surjektive Abbildung überführen, indem man F auf die Wertemenge $f(E)$ reduziert.

Satz 1.3 (Existenz der Umkehrabbildung)

*Ist $f : E \rightarrow F$ bijektiv, so existiert eine eindeutige Abbildung $f^{-1} : F \rightarrow E$, die jedem $y \in F$ ein $x \in E$ zuordnet mit $f(x) = y$. Diese heißt die **Umkehrabbildung** (oder **Umkehrfunktion**) $f^{-1} : F \rightarrow E$ von f .*

Beweis: Zu jedem $y \in F$ gibt es mindestens ein $x \in E$ mit $f(x) = y$, da f surjektiv ist (F ist also der Wertebereich von f). Da f zudem injektiv ist, kann es nicht mehr als ein $x \in E$ mit $f(x) = y$ geben. Zu jedem $y \in F$ gibt es also genau ein $x \in E$ mit $f(x) = y$. Darüber ist eine eindeutige Abbildung (nämlich die Umkehrabbildung) erklärt. □

Beispiel 1.8

- a) $f : \{1, 2\} \rightarrow \{3, 4\}$ mit $f : 1 \mapsto 3, 2 \mapsto 4$, ist injektiv und surjektiv, also bijektiv. Damit existiert die Umkehrabbildung $f^{-1} : \{3, 4\} \rightarrow \{1, 2\}$ mit $f^{-1} : 3 \mapsto 1, 4 \mapsto 2$.
 b) Die Abbildung s aus dem Weg-Zeit-Diagramm Abbildung 1.3 ist nicht injektiv (und damit nicht bijektiv), da das Fahrzeug hält und damit vielen Zeitpunkten die gleiche Strecke zugeordnet wird. Kennt man also eine zurückgelegte Strecke, so weiß man nicht in jedem Fall, zu welchem Zeitpunkt sie gehört. Es gibt keine Umkehrabbildung. ■

Achtung: Die Schreibweise f^{-1} kann irreführend sein. $f^{-1}(x)$ ist das Element, das die Umkehrfunktion dem Element x zuordnet. Ist $f(x) \neq 0$ eine Zahl, so schreibt man für deren Kehrwert $\frac{1}{f(x)} = f(x)^{-1}$. Der Kehrwert ist aber etwas völlig anderes als der Wert der Umkehrabbildung. Leider kennzeichnet man beide Werte mit dem Exponenten -1 .

Beispiel 1.9 (Caesar-Code)

Vor 2 000 Jahren verschickte bereits Julius Caesar verschlüsselte Nachrichten. Dabei verwendete er einen sehr einfachen Code: Zu einer festzulegenden Zahl $n \in \{1, 2, \dots, 25\}$ (bei 26 Buchstaben) wurde jeder Buchstabe eines Textes (nur Buchstaben, keine Leerzeichen oder sonstige Sonderzeichen) durch einen Buchstaben ersetzt, der zyklisch n Stellen im Alphabet später steht. Man muss also den Schlüssel n kennen, um einen Text zu entschlüsseln (oder maximal 25 Möglichkeiten durchprobieren).

Sei M die Menge aller Texte und $f_n : M \rightarrow M$ die Abbildung, die alle Buchstaben eines Textes um n nach rechts verschiebt. Dann ist f bijektiv. Die Umkehrabbildung verschiebt die Buchstaben um n zyklisch im Alphabet nach links. Für $n = 3$ ist $f(\text{„DIE-SISTEINTEXT“}) = \text{„GLHVLVWHLQWHAW“}$. Möchte man einen Text mit Leerzeichen verschlüsseln, so werden diese im ersten Schritt entfernt. Da man damit eine Information verliert, wird die Verschlüsselungsabbildung nicht injektiv: „DIE SEE“ und „DIESE E“ werden identisch verschlüsselt.

Der Caesar-Code ist sehr einfach und nicht sicher. Heute sind Verfahren wie die RSA-Verschlüsselung etabliert, siehe Beispiel 1.32 auf Seite 46. ■

Direkt aus der Definition der Umkehrfunktion erhält man:

Lemma 1.1 (Umkehrfunktion)

Es sei $f : E \rightarrow F$ bijektiv mit Umkehrfunktion $f^{-1} : F \rightarrow E$. Die Funktionen erfüllen die Beziehungen

$$\begin{aligned} f^{-1}(f(x)) &= x \quad \text{für alle } x \in E, \\ f(f^{-1}(y)) &= y \quad \text{für alle } y \in F. \end{aligned}$$

Die Umkehrung der Umkehrfunktion ist wieder die Ausgangsfunktion. Umkehrabbildungen werden wir später z. B. beim Lösen von Gleichungen verwenden. Ist der Wert von x gesucht mit $f(x) = y$, so ist $x = f^{-1}(y)$, z. B. für $f(x) = x^3$ und $f^{-1}(y) = y^{\frac{1}{3}}$: Aus $x^3 = y$ folgt $x = y^{\frac{1}{3}}$. In Kapitel 1.5.3 sehen wir uns (nach der Einführung der reellen Zahlen) Umkehrfunktionen zu reellwertigen Funktionen etwas genauer an.

1.2 Logik

1.2.1 Aussagenlogik

Überall im täglichen Leben, insbesondere in der Mathematik, wird man mit Aussagen konfrontiert, die entweder wahr oder falsch sein können. Eine Aussage kann nicht zugleich wahr und falsch sein.

Definition 1.8 (Aussage)

*Unter einer **Aussage** A versteht man ein sprachliches Gebilde, welches einen der beiden **Wahrheitswerte** wahr (w) oder falsch (f) hat.*

Alternativ verwendet man auch die Zahl 1 statt „wahr“ oder w sowie 0 für „falsch“ oder f .

Wahre Aussagen sind:

- Krefeld liegt am Rhein.
- Es gibt unendlich viele Primzahlen.
- $3 + 4 = 7$.

Falsche Aussagen sind:

- Die Erde ist eine Scheibe.
- $3 + 4 = 8$.

Es gibt aber auch Aussagen, von denen wir (zum Zeitpunkt, an dem wir dies schreiben) nicht sicher wissen, ob sie wahr oder falsch sind:

- Es gibt außerirdisches Leben.
- $P \neq NP$ (eines der im Internet zu findenden Millenium-Probleme, deren Lösung mit einem sehr hohen Preisgeld belohnt wird).

Das kann sich aber ändern, so war bis vor wenigen Jahren nicht bekannt, ob die Fermat'sche Vermutung wahr ist. Die Aussage lautet: „Die Gleichung $a^n + b^n = c^n$ hat für natürliche Zahlen $n > 2$ keine ganzzahligen Lösungen a, b, c .“ 1995 wurde von Andrew Wiles der Nachweis veröffentlicht, dass die Aussage wahr ist. Laut einer Pressemeldung vom 09.08.2010 auf www.heise.de soll auch $P \neq NP$ nachgewiesen worden sein. Allerdings bleibt abzuwarten, ob der Beweisansatz tatsächlich funktioniert.

Folgende Formulierungen sind keine Aussagen im mathematischen Sinn, da sie nicht eindeutig wahr oder falsch sind:

- Krefeld ist schön.
- Mathe ist schwierig.

Wir bezeichnen Aussagen mit Variablen (Platzhaltern) wie A bzw. B , die die Werte „wahr“ oder „falsch“ annehmen können, und verknüpfen sie mit sogenannten logischen Operatoren, die wir im Folgenden definieren. Einzelne Variablen, aber auch durch Verknüpfung mit logischen Operatoren gebildete Ausdrücke heißen **aussagenlogische Formeln**, die wir wiederum mit Variablennamen abkürzen und mit logischen Operatoren verknüpfen können. Abhängig von den Wahrheitswerten der Variablen nehmen aussagenlogische Formeln dann ebenfalls entweder den Wert „wahr“ oder den Wert „falsch“ an.

Wir unterscheiden später in diesem Text im Sprachgebrauch nicht mehr zwischen Aussagen und aussagenlogischen Formeln.

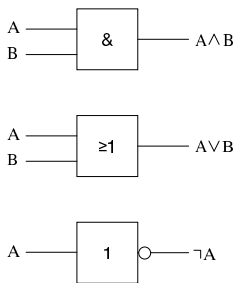


Abb. 1.5: Darstellung logischer Verknüpfungen als Gatter gemäß IEC 60617-12

Definition 1.9 (Verknüpfungen/Operatoren)

- Die Formel $A \vee B$ (sprich: A oder B) ist (für eine konkrete Belegung der Variablen) wahr genau dann, wenn (bei dieser Belegung) die Formeln A oder B (oder beide) wahr sind, also für wahre Aussagen stehen. Ist weder A noch B wahr, so ist die Formel falsch. $A \vee B$ ist eine **Disjunktion**.
- Die Formel $A \wedge B$ (sprich: A und B) ist wahr genau dann, wenn A und B beide wahr sind. Ist mindestens eine der Formeln A oder B falsch, so ist $A \wedge B$ falsch. $A \wedge B$ ist eine **Konjunktion**.
- Die Formel $\neg A$ (sprich: nicht A) ist wahr genau dann, wenn A falsch ist, sonst ist sie falsch. Statt $\neg A$ ist auch die Schreibweise $\bar{A}$ gebräuchlich, die auch für das Komplement von Mengen verwendet wird. $\neg A$ ist eine **Negation**.

Diese Verknüpfungen sind als integrierte Schaltkreise preiswert erhältlich. In Abbildung 1.5 sind die dabei verwendeten Symbole angegeben. Verknüpfungen von aussagenlogischen Formeln kann man über Wahrheitswertetabellen darstellen. Hier verwenden wir 0 für falsch und 1 für wahr. Negation, Konjunktion und Disjunktion sind so in Tabelle 1.1 angegeben.

Zwei aussagenlogische Formeln sind **gleich** „=“ genau dann, wenn sie bei jeder Belegung der Variablen mit Wahrheitswerten den gleichen Wahrheitswert annehmen. Verwenden wir ab jetzt statt „=“ das Symbol „:=“, so handelt es sich um eine definierende

Tab. 1.1: Wertetabelle der aussagenlogischen Verknüpfungen

A	B	$\neg A$	$A \wedge B$	$A \vee B$
0	0	1	0	0
0	1	1	0	1
1	0	0	0	1
1	1	0	1	1

Gleichheit. Hier weist man einem Ausdruck links vom Zeichen den Wert der rechten Seite zu. Häufig sieht man auch ein mit einem Ausrufungszeichen gekennzeichnetes Gleichheitszeichen $\stackrel{!}{=}$ oder ein anderes so markiertes Symbol. Das Ausrufungszeichen bedeutet „soll sein“. Man verlangt also die Gleichheit und berechnet dann, was nötig ist, um die Gleichheit zu erhalten.

Eine aussagenlogische Formel heißt **erfüllbar** genau dann, wenn es eine Belegung der Variablen gibt, die die Formel wahr werden lässt. Sie heißt **unerfüllbar** genau dann, wenn die Formel bei jeder Belegung der Variablen falsch ist, z.B. ist $A \wedge \neg A$ unerfüllbar. Umgekehrt heißt eine Formel, die bei jeder Variablenbelegung wahr ist, eine **Tautologie**. Beispielsweise ist $A \vee \neg A$ eine Tautologie.

Die Logik-Verknüpfungen weisen große Parallelen zu den Mengenoperationen auf. Der Negation $\neg$ entspricht bei Mengen das Komplement, der Oder-Verknüpfung $\vee$ die Vereinigung $\cup$ und der Und-Verknüpfung $\wedge$ der Schnitt $\cap$. Man kann die Aussagenlogik nachbilden, indem man die Wahrheitswerte „falsch“ durch die leere Menge $\emptyset$ und „wahr“ durch eine nicht-leere Menge, z.B. $\{1\}$, ausdrückt. Statt der Logik-Verknüpfungen kann man nun die Mengen-Verknüpfungen verwenden. Das Komplement (als Negation) ist dann bezüglich $\{1\}$ zu berechnen.

Es verwundert daher nicht, dass die Rechenregeln der Logik, die man über Wahrheitswertetabellen nachweist, aussehen wie die der Mengenlehre:

Satz 1.4 (Rechenregeln für Logik-Verknüpfungen)

Seien A, B und C aussagenlogische Formeln. Dann gilt:

■ **Kommutativgesetze:**

$$A \wedge B = B \wedge A, \quad A \vee B = B \vee A,$$

■ **Assoziativgesetze:**

$$(A \wedge B) \wedge C = A \wedge (B \wedge C), \quad (A \vee B) \vee C = A \vee (B \vee C),$$

■ **Distributivgesetze:**

$$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C), \quad A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C).$$

Die Klammern geben die Reihenfolge der Operationen vor. Da sie umständlich sind, legt man fest, dass $\neg$ enger bindet als $\wedge$ und $\wedge$ enger bindet als $\vee$. (Punkt- vor Strichrechnung, $\wedge$ kann mit der Multiplikation und $\vee$ mit der Addition verglichen werden.) Diese Prioritäten entsprechen genau denen für Mengenoperationen. Wegen des Assoziativgesetzes spielt die Reihenfolge bei der Auswertung des gleichen Operators keine Rolle. Damit können wir in vielen Fällen auf Klammern verzichten. Zum Beispiel ist $\neg A \vee B \wedge C = (\neg A) \vee (B \wedge C)$.

In der Digitaltechnik wird häufig ein exklusives Oder „xor“ bzw. $\oplus$ verwendet:

$$A \oplus B := A \wedge \neg B \vee \neg A \wedge B.$$

Diese Formel ist nur dann wahr, wenn entweder A oder B , aber nicht beide wahr sind. Mittels xor lässt sich ein binär als Liste von Nullen und Einsen gespeichertes Dokument verschlüsseln. Als Schlüssel dient ein weiteres Dokument, das stellenweise mit dem ersten xor-verknüpft wird. Dieser Vorgang ist bijektiv. Die Umkehrabbildung besteht in der erneuten Verknüpfung.

Viele Fehler geschehen durch falsche Negation. Hier sind die bereits von der Komplementbildung bei Mengen bekannten **De Morgan'schen Regeln** hilfreich, die man ebenfalls durch Aufstellen der Wahrheitswertetabelle nachrechnet:

Satz 1.5 (De Morgan'sche Regeln)

Seien A und B aussagenlogische Formeln. Dann gilt:

$$\begin{aligned}\neg(A \wedge B) &= \neg A \vee \neg B \\ \neg(A \vee B) &= \neg A \wedge \neg B.\end{aligned}$$

Die Negation der Aussage „Sie ist jung und schön.“ ist daher „Sie ist alt **oder** hässlich.“.

Beispiel 1.10 (Addierwerk)

In einem Computer werden Zahlen im **Dualsystem** (Zweiersystem) dargestellt (vgl. Seite 27). Dabei gibt es nur die Ziffern 0 und 1 (falsch und wahr), statt Zehnerpotenzen werden Potenzen von 2 verwendet. Die Zahl 10110101 im Zweiersystem entspricht der Dezimalzahl $1 \cdot 1 + 0 \cdot 2 + 1 \cdot 4 + 0 \cdot 8 + 1 \cdot 16 + 1 \cdot 32 + 0 \cdot 64 + 1 \cdot 128 = 181$. Zwei Dualzahlen werden addiert wie Dezimalzahlen, allerdings findet ein Übertrag zur nächsten Stelle schon dann statt, wenn die Summe größer als 1 ist:

$$\begin{array}{rcccc} & 1 & 0 & 1 & 0 \\ + & & & 1 & 1 & 1 \\ \hline & 1 & 1 & & & \\ \hline 1 & 0 & 0 & 0 & 1 \end{array}$$

Wir betrachten die Summe zweier Ziffern A und B und eines Übertrags C_{in} . Das Ergebnis ist eine Ziffer S und der nächste Übertrag C_{out} . Die folgenden Formeln können in der Wertetabelle (siehe Tabelle 1.2) abgelesen werden, indem man Terme für die Spalten erstellt, in denen S bzw. C_{out} den Wert 1 annimmt. Diese Terme werden dann mit Oder verknüpft.

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

Eine Schaltung, die diese Logik realisiert, heißt **Volladdierer**. Zwei Zahlen werden

Tab. 1.2: 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

addiert, indem man die Ziffernaddition für jede Stelle von rechts nach links durchführt und den Übertrag C_{out} einer Stelle als Übertrag C_{in} der nächsten Stelle verwendet (siehe Abbildung 1.6). ■

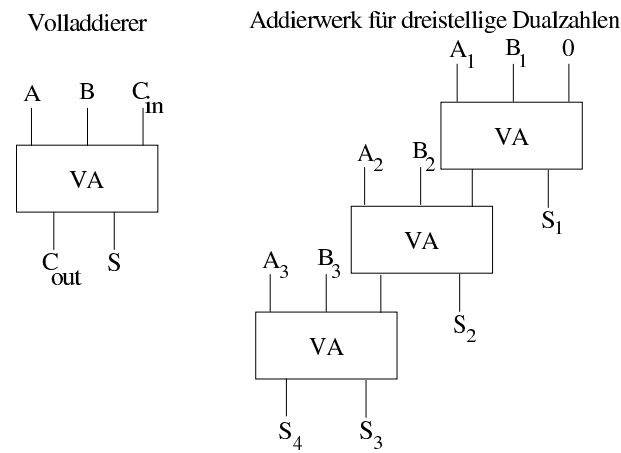


Abb. 1.6: Additionswerk mittels Volladdierer

Hintergrund: Normalformen

Die aussagenlogischen Formeln für die Summe und den Übertrag im Beispiel 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 in eine disjunktive Normalform bringen. Ähnlich kann man jede Formel als **konjunktive Normalform** schreiben. Beim Addierwerk ergeben sich die konjunktiven Normalformen

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

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 den Variablenwerten, zu denen er erstellt wurde. Für alle anderen Werte 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, vgl. (Goebbels und Rethmann, 2014, Kap. 1.2.6). Eine möglichst kurze disjunktive Normalform erhält man mittels eines **Karnaugh-Veitch-Diagramms**. 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.

1.2.2 Prädikatenlogik

Zur Vereinfachung haben wir in der Aussagenlogik Aussagen durch Variablen ersetzt, die für den Wahrheitswert der Aussagen stehen. In der Prädikatenlogik kommt nun ein anderer Typ von Variablen hinzu: Die Aussagen dürfen selbst noch von Parametern abhängen.

Man nennt eine Aussage, die von den Werten einer oder mehrerer Variablen abhängt, die Werte aus einer gewissen Grundmenge annehmen dürfen, eine **Aussageform**.

Eine Aussageform hat im Allgemeinen keinen bestimmten Wahrheitswert. Erst wenn die Variablen (z. B. $x_1, x_2, \dots, x_n$) durch feste Werte ersetzt werden, entsteht eine Aussage, von der feststeht, ob sie wahr oder falsch ist. Ersetzt man nur einen Teil der Variablen, hat man eine Aussageform mit den restlichen Variablen.

Beispielsweise ist $x_1 = x_2$ eine Aussageform, in der wir für x_1 und x_2 Zahlen einsetzen können. Die Aussageform wird zu einer wahren Aussage, wenn wir für x_1 und x_2 die gleiche Zahl einsetzen. Wenn wir nur für x_1 die Zahl 4711 einsetzen, dann erhalten wir die neue Aussageform $4711 = x_2$.

Wie bei Aussagen, die wir durch aussagenlogische Variablen ersetzt haben, ersetzen wir auch Aussageformen ihrerseits durch Variablen A, B usw., die nun aber in Abhängigkeit der Variablen, die innerhalb der Aussageform vorkommen, mit Werten wahr und falsch belegt werden. Wir schreiben dann beispielsweise $A(x_1, x_2, \dots, x_n)$ und sprechen vom **Prädikat** A .

Wir können z. B. die Aussageform $x_1 = x_2$ mit $A(x_1, x_2)$ bezeichnen, wobei A genau dann den Wert „wahr“ annimmt, wenn man für x_1 und x_2 die gleiche Zahl einsetzt.

Unterscheidet sich also mindestens einer der Werte $x_1, x_2, \dots, x_n$ von den Werten $y_1, y_2, \dots, y_n$, so kann auch $A(x_1, x_2, \dots, x_n)$ einen anderen Wahrheitswert als $A(y_1, y_2, \dots, y_n)$ haben.

Aus der Aussagenlogik wird so die **Prädikatenlogik**. Prädikat und Aussageform sind für uns Synonyme.

Beispiel 1.11

- a) $A(x) := x^2 > 30$ ist eine Aussageform. $\neg A(x)$ lautet $x^2 \leq 30$. $A(x)$ wird zur wahren Aussage, wenn man $x = 6$ einsetzt. Für $x = 5$ ist $A(x)$ falsch.
- b) Für $x \in \{\text{„rot“}, \text{„gelb“}, \text{„grün“}\}$ wird die Aussageform $A(x) := \text{„}x \text{ ist eine Ampelfarbe.“}$ zu einer wahren Aussage. Für $x = \text{„blau“}$ wird sie zu einer falschen Aussage.



Wie aussagenlogische Variablen kann man Prädikate mittels der Logik-Vernüpfungen zu prädikatenlogischen Formeln verknüpfen. Insbesondere sind aus wahren Folgerun-

gen (Folgerungspfeil: „ $\implies$ “) und Äquivalenzen (Äquivalenzzeichen: „ $\iff$ “) alle Berechnungen und Beweise zusammengesetzt, siehe Tabelle 1.3.

Tab. 1.3: Wertetabelle der Folgerung und Äquivalenz

A	B	$A \implies B$	$A \iff B$
0	0	1	1
0	1	1	0
1	0	0	0
1	1	1	1

Definition 1.10 (Implikation)

Seien A und B aussagen- oder prädikatenlogische Formeln. Die **Folgerung** bzw. **Implikation** $A \implies B$ ist definiert als $\neg A \vee B$ und wird als „Aus A folgt B .“ gesprochen.

Die Formel $A \implies B$ sei für jeden möglichen Wert der Variablen der Aussageformen wahr. Dann gilt:

- Wenn A wahr ist, dann muss auch B wahr sein. Man nennt A eine **hinreichende Bedingung** für B . Kann man zeigen, dass A wahr ist, dann hat man auch B gezeigt.

Achtung: Bestimmt man alle Variablenwerte, für die eine hinreichende Bedingung A wahr ist, so erhält man in der Regel nur einige und nicht alle Werte, für die die gefolgerte Aussageform B wahr wird.

- Umgekehrt muss B wahr sein, damit A überhaupt wahr werden kann. Daher bezeichnet man B als **notwendige Bedingung** für A . Ist eine notwendige Bedingung B für gewisse Variablenwerte erfüllt, so weiß man noch nicht, ob die zu untersuchende Aussage A für entsprechende Werte auch wahr ist. Nur wenn eine notwendige Bedingung B nicht erfüllt ist, weiß man, dass die zu untersuchende Aussage A falsch ist.

Sucht man alle Werte $x \in M$, für die eine Aussageform $A(x)$ wahr wird, so kann man mit der notwendigen Bedingung $B(x)$ die Kandidaten für x einschränken.

- Aus einer falschen Aussage kann man mittels wahrer Folgerung alles schließen.

Beispiel 1.12

Mit dem Satz von Fermat (Satz 2.34) und Folgerung 2.7 auf Seite 314 werden wir prominente Bedingungen für die Existenz von Extremwerten kennenlernen, die Sie vermutlich bereits aus der Schulzeit kennen:

- Eine notwendige Bedingung für die Existenz eines lokalen Extremums einer differenzierbaren Funktion an einer Stelle x_0 ist $f'(x_0) = 0$. Nur falls $f'(x_0) = 0$ ist, kann

in x_0 ein Extremum vorliegen. Die Bedingung kann aber auch erfüllt sein, wenn x_0 keine Extremstelle ist. Mit dem Folgerungspfeil geschrieben lautet der Satz von Fermat:

$$\text{„Differenzierbares } f \text{ hat ein lokales Extremum in } x_0.“ \implies f'(x_0) = 0.$$

- Eine hinreichende Bedingung für die Existenz eines lokalen Extremums einer differenzierbaren Funktion an einer Stelle x_0 ist $f'(x_0) = 0$ und $f''(x_0) \neq 0$. Ist diese Bedingung erfüllt, weiß man, dass in x_0 ein Extremum vorliegt. Die Bedingung muss aber nicht für alle Extremstellen erfüllt sein.

$$f'(x_0) = 0 \wedge f''(x_0) \neq 0 \implies \text{„}f \text{ hat ein lokales Extremum in } x_0.“ \quad \blacksquare$$

Beispiel 1.13

Sei x eine beliebige (reelle) Zahl. Dann wird die Aussageform $x = 2 \implies x^2 = 4$ zu einer wahren Aussage.

- Ist nämlich $x \neq 2$, so ist die Aussage $x = 2$ falsch und die Folgerung wahr.
- Ist $x = 2$, so ist auch die Aussage $x^2 = 4$ wahr, und die Folgerung ist ebenfalls wahr.

Man beachte, dass für $x := -2$ dagegen die Aussageform $x^2 = 4 \implies x = 2$ zu einer falschen Aussage wird. $x^2 = 4$ ist nämlich wahr, aber $x = 2$ ist falsch. Damit ist die Folgerung falsch. ■

Die Implikation ist **transitiv**, d. h., es gilt (es ist stets wahr):

$$[A \implies B \implies C] \implies [A \implies C],$$

ist also $A \implies B \implies C$ wahr, so ist auch $A \implies C$ wahr.

Definition 1.11 (Äquivalenz)

Seien A und B aussagen- oder prädikatenlogische Formeln. Die **Äquivalenz** $A \iff B$ ist erklärt als

$$(A \implies B) \wedge (B \implies A).$$

Die Äquivalenz ist also nur wahr, wenn A und B entweder beide wahr oder beide falsch sind.

Gleichheit von Formeln A und B liegt genau dann vor, wenn $A \iff B$ für jede Belegung der Variablen wahr ist. Gleichheit von Aussagen und Aussageformen entspricht also einer stets wahren Äquivalenz der Formeln.

Auch die Äquivalenz ist **transitiv**, d. h., es gilt $[A \iff B \iff C] \implies [A \iff C]$. Sucht man beispielsweise nach Lösungen einer Gleichung $A(x)$ (z. B. $A(x) := [x - 2 = 1]$), d. h. nach Werten x , für die die Aussageform $A(x)$ wahr wird, so macht man häufig Äquivalenzumformungen

$$A(x) \iff B(x) \iff C(x) \iff \dots \iff D(x),$$

die für jeden Wert x wahr sind. So stellt man sicher, dass man bei Betrachtung von $D(x)$ statt $A(x)$ tatsächlich richtige Lösungen findet (ist $D(x)$ wahr, so auch $A(x)$, $D(x) \implies A(x)$) und keine Lösungen übersieht (ist $A(x)$ wahr, so auch $D(x)$, $A(x) \implies D(x)$).

Beispiel 1.14

Für jeden Zahlenwert von x sind die folgenden Aussagen wahr:

$$x - 2 = 1 \iff x = 3, \quad x - 2 = 1 \implies x = 3 \vee x = 1,$$

$$x^2 = 4 \iff x = 2 \vee x = -2, \quad x = 2 \implies x^2 = 4.$$

Im Gegensatz zur Äquivalenz ist eine Folgerung auch dann noch wahr, wenn man zusätzliche Lösungen dazu bekommt, z. B. ist $x = 2 \implies x^2 = 4$ wahr, aber $x = 2 \iff x^2 = 4$ ist **nicht** für alle Werte von x wahr (s. o.). Für $x = -2$ kommt man nicht mehr von rechts nach links. ■

An dieser Stelle ist eine Bemerkung zum Aufschreiben länglicher Rechnungen nötig. Ein Leser muss verstehen, wie Rechenschritte zusammenhängen. Den Zusammenhang drückt man über die Symbole „ $\iff$ “, „ $\implies$ “ sowie „ $=$ “ aus:

$$(x + 1)(x - 1) = x^2 - x + x - 1 = x^2 - 1 = x^2 - 1^2.$$

Das ist eine Aussageform. Da für jeden Zahlenwert von x alle vier Terme den gleichen Wert haben, wird die Aussageform für jede Zahl x zu einer wahren Aussage. Dagegen macht die Schreibweise

$$(x + 1)(x - 1) \iff x^2 - x + x - 1 \iff x^2 - 1 \iff x^2 - 1^2$$

keinen Sinn, da die Terme $(x + 1)(x - 1)$, $x^2 - x + x - 1$, $x^2 - 1$ und $x^2 - 1^2$ keine Aussagen mit einem Wahrheitswert sind. Sinnvoll ist dagegen die Schreibweise

$$(x + 1)(x - 1) = 0 \iff x^2 - x + x - 1 = 0 \iff x^2 = 1 \iff x = 1 \vee x = -1,$$

wobei man statt $\iff$ die Implikation $\implies$ benutzt, falls bei einer Umformung weitere Lösungen hinzukommen (s. o.): $x + 1 = 0 \implies x^2 = 1$.

Schreibt man bei einer Rechnung „ $\implies$ “ oder „ $\iff$ “, so drückt man damit aus, dass diese logischen Verknüpfungen für alle relevanten Werte der Aussageformen wahr werden. Um die Formulierung „für alle relevanten Werte“ eleganter und explizit auszudrücken, bietet die Sprache der Mathematik mit **Quantoren** eine Formulierung:

- Der **Allquantor** $\forall$ steht für den Text „für alle“. $\forall x \in E : A(x)$ ist die Aussage, die in Textform lautet: „Für alle Elemente x von E gilt: Die Aussageform $A(x)$ wird eine wahre Aussage.“ Oder anders formuliert: $A(x)$ wird für jedes $x \in E$ wahr. Wahre Aussagen sind beispielsweise:

- $\forall x \in \{\text{„rot“}, \text{„gelb“}, \text{„grün“}\} : \text{„}x \text{ ist eine Ampelfarbe.“}$,
- $\forall x \in \{-3, -2, -1, 0, 1, 2, 3\} : (x^2 = 4 \iff x = 2 \vee x = -2)$.

- Der **Existenzquantor** $\exists$ steht für den Text „es existiert“. $\exists x \in E : A(x)$ ist die Aussage, die in Textform lautet: „Es existiert (mindestens) ein Element von E , so dass, ersetzt man x durch dieses Element, $A(x)$ wahr wird.“ Anders formuliert: $A(x)$ wird für ein $x \in E$ wahr. Wahre Aussagen sind:

- $\exists x \in \{\text{„blau“}, \text{„gelb“}, \text{„grün“}\} : \text{„}x \text{ ist eine Ampelfarbe.“}$,
- $\exists x \in \{1, 2, 3\} : x^2 = 4$.

Quantoren darf man hintereinander schalten: $\forall x \in E \exists y \in F : A(x, y)$ ist die Aussage: „Zu jedem $x \in E$ existiert ein $y \in F$ (das für jedes x ein anderes Element sein kann), so dass $A(x, y)$ wahr ist“.

Im Umgang mit Quantoren sind einige Regeln zu beachten:

- Die Reihenfolge verschiedener Quantoren darf nicht vertauscht werden. Es ist ein Unterschied, ob man sagt „Zu jedem $x \in E$ existiert ein $y \in F$, das von x abhängig sein darf, so dass ...“ oder „Es existiert ein $y \in F$ (das nicht von x abhängt), so dass für alle $x \in E$ gilt: ...“.
- Bei Negation muss man die Quantoren austauschen. Wenn etwas nicht für alle $x \in E$ gilt, dann gibt es ein $x \in E$, für das es nicht gilt. Wenn ein $x \in E$ nicht existiert, so dass eine Aussageform wahr wird, dann wird sie für alle $x \in E$ nicht wahr:

$$\begin{aligned} \neg[\forall x \in E \exists y \in F : A(x, y)] &= \exists x \in E : \neg[\exists y \in F : A(x, y)] \\ &= \exists x \in E \forall y \in F : \neg A(x, y). \end{aligned}$$

Damit die Aussagen besser lesbar sind, werden wir in diesem Buch statt der Quantoren Text verwenden.

Beispiel 1.15

- a) Wie lautet die Negation der Aussage „Alle Wege führen nach Rom.“?

Antwort: „Es gibt einen Weg, der nicht nach Rom führt.“

- b) Wie lautet die Negation der Aussage „Es gibt eine Straße mit Schlaglöchern.“?

Antwort: „Alle Straßen sind frei von Schlaglöchern.“ ■

Beispiel 1.16

Mit Prädikatenlogik kann man programmieren, vgl. (Goebbels und Rethmann, 2014, Kap. 1.2.6). Aus „Programmieren in Logik“ ist der Name der Programmiersprache Prolog abgekürzt. Zunächst definiert man in Prolog eine Datenbasis, die aus wahren Aussagen besteht. Zusätzlich werden Regeln (wahre Folgerungen) aufgestellt. Kann eine Aussage aus den Fakten mittels der Regeln abgeleitet werden, ist sie auch wahr, anderenfalls gilt sie als falsch. Damit hat man ein Expertensystem (wissensbasiertes System), das auf Fragen antworten kann. ■

1.2.3 Beweise

In der Mathematik werden wahre Aussagen als **Sätze** und **Hilfssätze** formuliert und sind zu beweisen. Ein Hilfssatz wird auch mit **Lemma** bezeichnet. Ein **Beweis** ist eine wahre logische Folgerung (Implikation) der zu zeigenden Aussage aus bereits bewiesenen Aussagen (bekannten Sätzen) unter Verwendung von Begriffsbildungen (Definitionen).

Da man mit den Folgerungen irgendwo beginnen muss, ergibt sich die Notwendigkeit, gewisse grundlegende Aussagen als **Axiome** einer Theorie zu akzeptieren (als wahr anzusehen), ohne sie zu beweisen. Im vorangehenden Beispiel übernehmen die Fakten des Prolog-Programms die Rolle von Axiomen.

Wir gehen im Folgenden aus von einem zu beweisenden Satz B (Behauptung) und bezeichnen die Bedingungen, unter denen er gilt, mit A (Annahme). Zu A gehören natürlich alle bisherigen Folgerungen aus den Axiomen. Darüber hinaus können aber auch weitere Aussagen bei der Formulierung eines Satzes gefordert werden.

Zu zeigen ist also die Implikation:

$$A \implies B.$$

Dazu gibt es zwei Ansätze:

- Man zeigt mit der Information, dass A wahr ist, mittels Zwischenaussagen, dass auch B wahr ist:

$$A \implies C_1 \implies C_2 \implies \dots \implies B.$$

Dies ist ein **direkter Beweis**. Dahinter steckt der **Modus Ponens**: Ist A eine wahre Aussage, und ist die Folgerung $A \implies B$ ebenfalls wahr, dann ist auch B eine wahre Aussage.

- Man nimmt an, dass B falsch ist und zeigt (mittels wahrer Folgerungen), dass dann auch A falsch ist. Unter der Voraussetzung, dass A wahr ist, ist das ein Widerspruch, und die Annahme muss falsch sein: B ist wahr.

$$\neg B \implies C_1 \implies C_2 \implies \dots \implies \neg A.$$

Dies ist ein **indirekter Beweis** oder ein **Beweis durch Widerspruch**, da man das Gegenteil der zu zeigenden Aussage zum Widerspruch führt.

Beispiel 1.17

Wir beweisen direkt die Aussage: „Für alle $a, b \geq 0$ gilt: $\frac{a+b}{2} \geq \sqrt{ab}$ “. Auf die hier verwendeten Rechenregeln gehen wir im Detail in Kapitel 1.3 ein.

Wir betrachten Aussageformen für alle Zahlen $a, b \geq 0$ und beginnen (unter Auslassung der Quantoren) mit einer Aussageform, die für alle diese Zahlen wahr ist:

$$(a - b)^2 \geq 0 \implies a^2 - 2ab + b^2 \geq 0 \xrightarrow{+4ab} a^2 + 2ab + b^2 \geq 4ab$$

$$\implies (a+b)^2 \geq 4ab \xRightarrow{\sqrt{\dots}} a+b \geq 2\sqrt{ab} \xRightarrow{/2} \frac{a+b}{2} \geq \sqrt{ab},$$

wobei im vorletzten Schritt die Einschränkung $a, b \geq 0$ verwendet wurde. Wir haben damit einen Spezialfall von Satz 1.16 auf Seite 69 gezeigt. ■

Bei einem indirekten Beweis erhält man durch die Negation der zu zeigenden Aussage eine zusätzliche Information, die man im Beweis benutzen kann.

Beispiel 1.18

Wir beweisen die Aussage „Ist n^2 gerade (d. h. durch 2 teilbar), dann ist n gerade“ durch indirekte Schlussweise. Zunächst ist $A := „n^2$ ist gerade“ und $B := „n$ ist gerade“. Um $A \implies B$ indirekt zu beweisen, zeigen wir $\neg B \implies \neg A$, d. h. die Implikation „Ist n ungerade (d. h. nicht durch 2 teilbar), dann ist n^2 ungerade“. Sei nun n ungerade, d. h. $n = 2k + 1$ mit einer nicht-negativen ganzen Zahl k . Dann folgt

$$n = 2k + 1 \implies n^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1,$$

wobei $2(2k^2 + 2k)$ offensichtlich eine gerade Zahl ist, die durch die Addition von eins ungerade wird. n^2 ist also ungerade, d. h., es gilt $\neg B \implies \neg A$, und hiermit ist $A \implies B$ gezeigt. ■

Beispiel 1.19 (Halteproblem *)

Wir beweisen, dass es kein Computerprogramm A gibt, das in endlicher Zeit entscheiden kann, ob ein beliebiges weiteres Computerprogramm seinerseits nach endlicher Zeit hält, d. h. zu einem Ergebnis kommt. Dieses unentscheidbare **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 anzunehmen, dass es das Programm A gibt. Damit können wir 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. Einen ähnlichen Trick haben wir auch schon bei der Betrachtung der Allmenge im Kasten auf Seite 5 benutzt.

Das Programm B sei (ohne formale Programmiersprache) wie folgt aufgebaut:

- B wendet 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 das Programm A entgegen der Annahme nicht geben kann. ■

Häufig erkennt man, dass verschiedene Aussagen völlig gleich bewiesen werden können und beschränkt sich auf eine dieser Aussagen. Es kann auch vorkommen, dass eine Einschränkung für den Beweis keine Rolle spielt, aber viel Schreibarbeit erspart. In diesen Situationen findet man häufig die Abkürzung o. B. d. A., die für „ohne Beschränkung der Allgemeinheit“ steht. Am Ende von Beweisen findet man bisweilen auch die Abkürzung „q. e. d.“, die „quod erat demonstrandum“ bedeutet: „was zu beweisen war“. Wir benutzen hier dagegen ein leeres Kästchen.

1.3 Reelle Zahlen

In der Umgangssprache unterscheidet man häufig nicht zwischen Zahlen mit und ohne Nachkommateil, Zahlen, die man als Brüche schreiben kann und Zahlen, die unendlich viele Nachkommastellen ohne regelmäßige Wiederholung besitzen. In der Mathematik beginnt man dagegen systematisch mit einer einfachen Zahlenmenge und erweitert diese sukzessive so, dass die gängigen Rechenarten möglich werden. Dabei findet man die unterschiedlichen Zahlentypen. So benötigt man Brüche, wenn man dividieren möchte, reelle Zahlen, wenn die Quadratwurzel aus nicht-negativen Zahlen erklärt sein soll, und komplexe Zahlen, wenn die Quadratwurzel auch aus negativen Zahlen benötigt wird. Wir beginnen mit den natürlichen Zahlen und erweitern diese Zahlenmenge sukzessive zu den reellen Zahlen. Später folgt dann der Übergang zu komplexen Zahlen.

1.3.1 Natürliche und ganze Zahlen

Man erhält die **natürlichen Zahlen**, indem man eine erste natürliche Zahl als Zeichen „1“ definiert und dann festlegt, dass mit jeder natürlichen Zahl n auch die Zeichenkette, die entsteht, wenn man an n die Zeichen „+1“ anhängt, eine natürliche Zahl repräsentiert. Diese verstehen wir als Nachfolger von n . Damit ist $\{„1“, „1 + 1“, „1 + 1 + 1“, „1 + 1 + 1 + 1“, \dots\}$ die Menge der natürlichen Zahlen. Als Abkürzung ersetzen wir die Zeichenketten durch die bekannten Zahlen im Zehnersystem:

$$\mathbb{N} := \{1, 2, 3, \dots, 9, 10, 11, \dots\}.$$

$\mathbb{N}_0 := \{0, 1, 2, 3, \dots\}$ ist die Menge der natürlichen Zahlen mit 0. In $\mathbb{N}_0$ kennen wir die übliche Addition und Multiplikation. Subtraktion und Division können Ergebnisse haben, die nicht mehr zu $\mathbb{N}_0$ gehören. Man kann die natürlichen Zahlen mathema-

tisch sauber mittels der Peano-Axiome einführen, die die vorangehende Konstruktion formalisieren.

Man erweitert nun $\mathbb{N}_0$ so, dass die Subtraktion nicht aus $\mathbb{N}_0$ hinausführt und erhält

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

die Menge der **ganzen Zahlen**. Die Einführung negativer Zahlen war ein Meilenstein in der Mathematik. Erst seit dem 16. Jahrhundert werden sie systematisch verwendet. Für ganze Zahlen verwendet man hauptsächlich die Symbole i und j (sofern diese im jeweiligen Zusammenhang nicht durch die imaginäre Einheit der komplexen Zahlen belegt sind) sowie k, l, m, n , aber auch weitere wie p und q , wenn aus dem Zusammenhang die Bedeutung klar ist.

Häufig werden Teilmengen definiert, indem Elemente ausgewählt werden, die bestimmten Bedingungen genügen. Die Menge der geraden Zahlen ist die Menge aller ganzen Zahlen, die das Doppelte einer anderen ganzen Zahl sind. Dies schreibt man knapp

$$\mathbb{Z}_g = \{m \in \mathbb{Z} : m = 2 \cdot n, \quad n \in \mathbb{Z}\}.$$

Analog sind die ungeraden Zahlen definiert durch

$$\mathbb{Z}_u = \{m \in \mathbb{Z} : m = 2 \cdot n + 1, \quad n \in \mathbb{Z}\}.$$

Achtung: Die folgenden Fehler im Umgang mit negativen Zahlen fallen in Klausuren immer wieder auf:

- Multipliziert man mit einer negativen Zahl, so sollte man das Produkt **nicht** angeben als $n \cdot -m$, da man das Multiplikationszeichen leicht übersieht und dann eine Subtraktion vornimmt. Der Malpunkt wird schnell überlesen, und es ist auch üblich, ihn ganz wegzulassen. Schreiben Sie $n \cdot (-m)$ oder kurz $n(-m)$, also z. B. $3 \cdot (-4)$ statt $3 \cdot -4$.
- Punktrechnung geht vor Strichrechnung: $21 = (1 + 2) \cdot (3 + 4) \neq 1 + 2 \cdot (3 + 4) = 15$. Die Klammer darf nicht weggelassen werden.
- Minus mal Minus ist Plus: $(-n) \cdot (-m) = (-1) \cdot (-1) \cdot n \cdot m = n \cdot m$.
- Subtraktion negativer Zahlen: $-4 - (-3) = -4 + 3 = -1 \neq -7$.

Dass solche Fehler nicht nur in Klausuren geschehen, sieht man am Bilanzierungsfehler der Bad Bank der Hypo Real Estate, der im Oktober 2011 bekannt wurde. Die Rheinische Post vom 31.10.2011 berichtete auf Seite A7, dass gut 55 Milliarden Euro (55 000 000 000 Euro) auf der falschen Seite der Bilanz eingetragen wurden (Vorzeichenfehler). Zieht man sie dort ab und addiert sie zur anderen Seite, dann entsteht allerdings eine Differenz von 110 Milliarden Euro. Hier stimmt vermutlich auch mit dem Zeitungsbericht etwas nicht, da diese Summe nie genannt wurde.

1.3.1.1 Ordnung

Die natürlichen Zahlen sind **total geordnet**, d. h., für zwei natürliche Zahlen m und n ist mindestens eine der beiden folgenden Aussagen wahr:

- m ist kleiner oder gleich n , d. h. $m \leq n$ (d. h., n ist direkter oder indirekter Nachfolger von m) oder
- n ist kleiner oder gleich m , d. h. $n \leq m$ (d. h., n ist direkter oder indirekter Vorgänger von m).

Definition 1.12 (Ordnungsrelation)

Eine **Ordnungsrelation** „ $\leq$ “ auf einer beliebigen Menge E muss genau die folgenden Axiome für alle $m, n, r \in E$ erfüllen:

- **Reflexivität:** $n \leq n$,
- **Transitivität:** $(n \leq m) \wedge (m \leq r) \implies n \leq r$,
- **Antisymmetrie:** $(n \leq m) \wedge (m \leq n) \implies n = m$.

E heißt **total geordnet** genau dann, wenn für jedes Paar von Elementen $n, m \in E$ gilt:

$$(m \leq n) \vee (n \leq m).$$

Diese Bedingungen sind offensichtlich für den bekannten Vergleich „ $\leq$ “ auf $\mathbb{N}$ erfüllt.

Zusätzlich zu $\leq$ werden wir die Zeichen $\geq$ für größer oder gleich, $<$ für echt kleiner und $>$ für echt größer verwenden, also

$$n \geq m := m \leq n, \quad n < m := (n \leq m) \wedge (n \neq m), \quad n > m := (n \geq m) \wedge (n \neq m).$$

Beispiel 1.20 (Lexikographische Ordnung)

Die Einträge im Index des Buchs oder in einem Lexikon sind **lexikographisch** geordnet. Ausgehend von der Reihenfolge $A \leq B \leq C \leq \dots \leq Z$ der Buchstaben des Alphabets werden dabei die Wörter zunächst nach dem ersten Buchstaben sortiert. Innerhalb der Gruppen mit gleichem ersten Buchstaben wird dann nach dem zweiten sortiert usw. So ist „Adam $\leq$ Eva“ wegen $A \leq E$ und „Fahrrad $\leq$ Fahrzeug“ wegen $R \leq Z$. Die Reflexivität, Transitivität und die Antisymmetrie prüft man hier leicht nach. ■

1.3.1.2 Zahlendarstellung

Wir sind es gewohnt, dass natürliche Zahlen im Dezimalsystem mit der Ziffernmenge $\{0, 1, \dots, 9\}$ angegeben werden. Der Wert, für den eine Ziffer steht, hängt von der Position innerhalb der Ziffernfolge ab. Im Dezimalsystem gilt beispielsweise

$$123 = 1 \cdot 10^2 + 2 \cdot 10^1 + 3 \cdot 10^0.$$

Dabei geben die einzelnen Stellen Faktoren zu den Potenzen $10^0 = 1, 10^1 = 10, 10^2 = 100, \dots, 10^n = \underbrace{10 \cdot 10 \cdots 10}_{n\text{-mal}}$. Allgemeiner kann man statt der (wegen unserer zehn Finger willkürlich gewählten) Basis 10 auch eine andere natürliche Zahl $b > 1$ als Basis des Zahlensystems benutzen:

$$\begin{aligned} a &= (a_n a_{n-1} \dots a_0)_b \\ &= a_n \cdot b^n + a_{n-1} \cdot b^{n-1} + \dots + a_0 \cdot b^0, \quad a_k \in \{0, 1, \dots, b-1\}, \end{aligned}$$

wobei $b^0 := 1$ und $b^k := b \cdot b^{k-1}, k \in \mathbb{N}$ (siehe Seite 36). Wie im Dezimalsystem steht links die höchstwertige Stelle, während man ganz rechts die Einer schreibt. In der Digitaltechnik werden Zahlendarstellungen zu Basen verwendet, die eine Zweierpotenz sind (siehe Tabelle 1.4).

Tab. 1.4: Zahlendarstellungen mit Zweierpotenzen

Basis b	Name	Ziffern
2	Dual- oder Binärsystem	$\{0, 1\}$
8	Oktalsystem	$\{0, 1, \dots, 7\}$
16	Hexadezimalsystem	$\{0, 1, \dots, 9, a, b, c, d, e, f\}$

Beispiel 1.21

Wir stellen die Hexadezimalzahl $(4e20b)_{16}$ im Dezimalsystem dar:

$$11 \cdot 16^0 + 0 \cdot 16^1 + 2 \cdot 16^2 + 14 \cdot 16^3 + 4 \cdot 16^4 = 11 + 512 + 57\,344 + 262\,144 = 320\,011.$$

■

Die Umwandlung einer Zahl aus der Darstellung zur Basis b ins Dezimalsystem kann durch geschickte Klammerung effizient mit dem **Horner-Schema** erfolgen, das wir in Kapitel 1.5.6.4 behandeln.

Beispiel 1.22 (Subtraktion auf dem Computer*)

Die Subtraktion ganzer Zahlen wird im Computer auf die Addition zurückgeführt, indem man negative Zahlen geschickt darstellt. Das funktioniert aber nur bei einer festen Stellenzahl, wobei die höchste Stelle das Vorzeichen angibt (1 bei einer negativen Zahl). Bei einer negativen Zahl verändert man aber auch die anderen Stellen, damit die Addition möglichst einfach wird.

- **Einerkomplement:** Bei einer Dualzahl werden alle Nullen durch Einsen und Einsen durch Nullen ersetzt. Durch die Subtraktion $1111_2 - y$ erhält man das Einerkomplement der vierstelligen Dualzahl y . Für $y = 1011_2$ ist $1111_2 - y = 0100_2$ das

Einerkomplement. Stellt man negative Zahlen als Einerkomplement der entsprechenden positiven Zahl dar, so kann man die Subtraktion wie folgt auf die Addition zurückführen. Dazu sehen wir uns die Differenz zweier vierstelliger Dualzahlen x und y an:

$$x - y = x + (1111_2 - y) - 1111_2 = x + (1111_2 - y) - 10000_2 + 0001_2.$$

- Ist die fünfte (höchste) Stelle von $x + (1111_2 - y)$ durch einen Übertrag gesetzt (gleich eins), so ist $x - y$ positiv. Um in diesem Fall aus $x + (1111_2 - y)$ wieder $x - y$ zu erhalten, müssen wir 1111_2 abziehen. Das machen wir aber in zwei Schritten: Mit -10000_2 lassen wir die führende Stelle weg und müssen schließlich noch 1 addieren.
- Ist die fünfte (höchste) Stelle von $x + (1111_2 - y)$ nicht gesetzt (gleich null), so ist $x - y$ negativ oder null. Das Ergebnis im Einerkomplement erhalten wir über

$$1111_2 - (-1) \cdot [x + (1111_2 - y) - 1111_2] = x + (1111_2 - y).$$

Nach dem Addieren von Einerkomplementen muss man also einen evtl. auftretenden Übertrag an der höchsten Stelle als 1 zum bisherigen Ergebnis addieren. Gibt es keinen Übertrag, so entfällt diese Addition, und der Wert von $x - y$ liegt bereits als Einerkomplement vor.

Ein Nachteil des Einerkomplements ist, dass die Null zwei Darstellungen besitzt, z. B. bei vier Stellen 0000_2 und 1111_2 . Das lässt sich vermeiden:

- Das **Zweierkomplement** einer n -stelligen Dualzahl erhält man, indem zum Einerkomplement eins addiert wird. Sollte es einen Übertrag an der höchsten Stelle geben, so lässt man diesen weg. Für $y \neq 0$ ist

$$x - y = x - y + 1111_2 + 0001_2 - 10000_2 = x + \underbrace{[(1111_2 - y) + 0001_2]}_{\text{Zweierkomplement}} - 10000_2.$$

Addiert man zu x das Zweierkomplement von y , so kann es an der höchsten Stelle einen Übertrag geben. Gibt es ihn, so kann man 10000_2 abziehen und erhält ein nicht-negatives Ergebnis, z. B. $1110_2 - 1011_2 = 1110_2 + 0101_2 - 10000_2 = 0011_2$. Hat man jedoch keinen Übertrag, so ist $x - y$ negativ, und die Darstellung als Zweierkomplement ist bereits berechnet:

$$1111_2 - (-1) \cdot [x + (1111_2 - y) + 0001_2 - 10000_2] + 0001_2 = x + (1111_2 - y) + 0001_2.$$

Während man beim Einerkomplement einen Übertrag an der höchsten Stelle berücksichtigen muss, kann man ihn beim Rechnen mit dem Zweierkomplement einfach weglassen.

Stellt man allgemeiner n -stellige Zahlen in einem Zahlensystem zur Basis b dar, so funktioniert der gleiche Trick:

$$x - y = x + (b^n - 1 - y) + 1 - b^n. \quad \blacksquare$$

1.3.1.3 Primzahlen

In der Kryptographie spielen Verschlüsselungen mit **Primzahlen** eine wichtige Rolle (RSA-Verfahren, siehe Seite 46).

Definition 1.13 (Primzahl)

Eine natürliche Zahl mit Ausnahme der Eins, die nur durch sich selbst und durch die Eins ohne Rest teilbar ist, heißt **Primzahl**.

Die ersten Primzahlen lauten: 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, ...

Satz 1.6 (Primfaktorzerlegung)

Jede natürliche Zahl $a > 1$ lässt sich als endliches Produkt von Primzahlen (Primfaktoren) p_k schreiben:

$$a = p_1 \cdot p_2 \cdots p_n.$$

Dabei sind die verwendeten Primzahlen und ihre Anzahl eindeutig bestimmt.

Beweis: (Skizze) Es ist sowohl die Existenz einer Primfaktorzerlegung als auch deren Eindeutigkeit zu zeigen. Beides lässt sich mit einem (indirekten) Beweis durch Widerspruch bewerkstelligen. Wir beschränken uns auf den Nachweis der Existenz, der etwas einfacher als der Beweis der Eindeutigkeit ist:

Wir nehmen an, dass es natürliche Zahlen ohne eine Zerlegung gibt. Dann gibt es eine kleinste unter diesen Zahlen, die wir mit n bezeichnen. n ist keine Primzahl (sonst hätten wir die Primfaktorzerlegung), lässt sich also durch eine Zahl m mit $1 < m < n$ teilen. Nun besitzen aber die kleineren Zahlen $\frac{n}{m}$ und m eine Primfaktorzerlegung, also auch $n = \frac{n}{m} \cdot m$. Damit haben wir einen Widerspruch zur Annahme, die falsch ist. $\square$

Wir erhalten beispielsweise die Primfaktorzerlegungen

$$10 = 2 \cdot 5, \quad 6 = 2 \cdot 3, \quad 18 = 2 \cdot 3 \cdot 3, \quad 13 = 13 \quad \text{und} \quad 252 = 2 \cdot 2 \cdot 3 \cdot 3 \cdot 7.$$

Satz 1.7 (Größte Primzahl)

Es gibt unendlich viele Primzahlen. Damit kann es keine größte Primzahl geben.

Beweis: Wir führen wieder einen (indirekten) Beweis mittels Widerspruch. Dazu nehmen wir an, dass es nur endlich viele Primzahlen gibt. Diese Annahme gibt uns eine Zusatzinformation, die wir im Beweis verwenden können. Denn nun existiert eine größte Primzahl p . Da jede natürliche Zahl als Produkt der endlich vielen Primzahlen

$p_1, p_2, \dots, p_n = p$ geschrieben werden kann, ist auch die Zahl $q := p_1 \cdot p_2 \cdot p_3 \cdot \dots \cdot p_n + 1$ als ein solches Produkt darstellbar. Dies ist aber falsch, da wegen der $+1$ keine der Primzahlen $p_1, \dots, p_n$ die Zahl q teilt. Damit kann die Annahme „es gibt nur endlich viele Primzahlen“ nicht stimmen. $\square$

Definition 1.14 (Größter gemeinsamer Teiler)

*Der **größte gemeinsame Teiler** $\text{ggT}(n, m)$ zweier natürlicher Zahlen n und m ist die größte natürliche Zahl, die sowohl n als auch m teilt.*

Zum Beispiel ist $\text{ggT}(15, 10) = 5$. Man erhält den größten gemeinsamen Teiler, indem man alle gemeinsamen Primfaktoren beider Zahlen (gemäß ihrer gemeinsamen Vielfachheit) miteinander multipliziert.

Statt vorhandene Primfaktorzerlegungen zu nutzen, kann man auch den **Euclid'schen Divisionsalgorithmus** verwenden. Zur Berechnung des größten gemeinsamen Teilers dividiert man dabei die größere Zahl (diese sei n) durch die kleinere (diese sei m). Entsteht dabei ein Rest $0 < r < m$, so wiederholt man die Division mit den neuen Zahlen $n' := m$ und $m' := r$. Entsteht dabei wieder ein Rest r' , so setzt sich das Verfahren fort mit den Zahlen $n'' := m'$ und $m'' := r'$. Man dividiert sukzessive so lange, bis schließlich ein Rest 0 auftritt. Der zuletzt verwendete Divisor (Nenner) ist der größte gemeinsame Teiler.

Durch den Algorithmus werden mit jeder Division die beiden Zahlen kleiner, bis schließlich der Rest der Division 0 ist (spätestens, wenn durch 1 geteilt wird). Um den Algorithmus zu beweisen, kann man zeigen, dass

$$\text{ggT}(n, m) = \text{ggT}(n', m') = \text{ggT}(n'', m'') = \dots$$

ist. Man nennt diese Eigenschaft eine **Invariante**. Das ist eine Eigenschaft, die bei jedem Durchlauf eines Algorithmus erhalten bleibt. In der Informatik beweist man über Invarianten die Korrektheit von Programmen (vgl. Beispiel 1.30 auf Seite 43).

Beispiel 1.23

Zu $n = 18$ und $m = 12$ liefert der erste Schritt des Algorithmus $18 : 12 = 1$ Rest 6. Mit dem Nenner und dem Rest geht es in den zweiten Schritt: $12 : 6 = 2$ Rest 0. Damit ist $\text{ggT}(18, 12) = 6$. $\blacksquare$

Beispiel 1.24

Der rechteckige Boden eines Raums soll mit möglichst großen quadratischen Teppichplatten (keine Fugen) ausgelegt werden. Wie groß ist die maximale Kantenlänge der Platten bei einer Grundfläche von $2,10 \text{ m} \cdot 1,80 \text{ m}$, wenn keine Platten geschnitten werden sollen? Die maximale Kantenlänge in cm ist $\text{ggT}(210, 180) = 30$. Es werden dann je 7 Platten in 6 Reihen verlegt. $\blacksquare$

Definition 1.15 (Kleinstes gemeinsames Vielfaches)

Das **kleinste gemeinsame Vielfache** $\text{kgV}(n, m)$ von zwei natürlichen Zahlen n und m ist die kleinste natürliche Zahl, die sowohl von n als auch von m geteilt wird.

Es gilt:

$$\text{kgV}(n, m) = \frac{n \cdot m}{\text{ggT}(n, m)}.$$

Das kleinste gemeinsame Vielfache ist genau das Produkt aller Primfaktoren der beiden Zahlen, wobei jeweils die maximale Vielfachheit eines Faktors aus beiden Primfaktorzerlegungen gewählt ist.

Beispiel 1.25

$$\text{kgV}(18, 12) = \frac{18 \cdot 12}{6} = 36. \quad \blacksquare$$

1.3.1.4 Fakultät und Binomialkoeffizient

Die natürlichen Zahlen werden zum Abzählen von Mengen benötigt. Wichtige Begriffe in diesem Zusammenhang sind Fakultäten und Binomialkoeffizienten.

Definition 1.16 (Fakultät)

Die **Fakultät** einer natürlichen Zahl n ist erklärt durch das Produkt

$$n! := n \cdot (n - 1) \cdot (n - 2) \cdots 2 \cdot 1.$$

Zusätzlich setzt man $0! := 1$.

Man erhält beispielsweise $3! = 3 \cdot 2 \cdot 1 = 6$ und $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

$n!$ wächst sehr schnell mit n : $5! = 120$, $10! = 3\,628\,800$, $20! \approx 2,432902008 \cdot 10^{18}$. Dabei verwenden wir $\approx$ für „ungefähr (näherungsweise) gleich“.

Häufig benötigt man die **rekursive** Darstellung für $n > 0$:

$$n! = n \cdot (n - 1)!, \quad 0! = 1.$$

Hier wird die Fakultät quasi über sich selbst (d. h. rekursiv) definiert. Das ist so, als würde man sich am eigenen Schopf aus dem Sumpf ziehen. Im Gegensatz zu diesem Bild macht aber diese Darstellung der Fakultät Sinn, da man sich damit sukzessive bis zum bekannten Startwert $0! = 1$ vorarbeiten kann.

Die Zahl $n!$ gibt an, auf wie viele verschiedene Weisen man n Objekte in einer Liste anordnen kann: Für den ersten Listenplatz gibt es n Möglichkeiten, für den zweiten bleiben dann noch $n - 1$, bis schließlich nur eine Möglichkeit für den letzten Listenplatz übrig ist. Insgesamt erhält man also $n \cdot (n - 1) \cdot (n - 2) \cdots 2 \cdot 1 = n!$ Möglichkeiten. Jede Anordnung heißt eine **Permutation** der n Objekte.

Mit Potenzen kann man oft leichter Rechnen als mit Fakultäten. Vergrößern wir in $n!$ jeden Faktor zu n , so erhalten wir die ganz grobe Abschätzung $n! \leq n^n$. Tatsächlich verhält sich die Fakultät ähnlich wie n^n . Dieser Zusammenhang heißt Stirling'sche Formel und wird später auf Seite 369 behandelt.

In der Kombinatorik – aber auch im Umgang mit Polynomen – sind Ausdrücke wichtig, die sich aus mehreren Fakultäten zusammensetzen.

Definition 1.17 (Binomialkoeffizient)

Es seien $n \in \mathbb{N}_0$ und $m \in \mathbb{Z}$. Dann heißt die natürliche Zahl

$$\binom{n}{m} := \begin{cases} \frac{n!}{(n-m)! \cdot m!} & \text{falls } n \geq m \geq 0, \\ 0 & \text{falls } m < 0 \text{ oder } m > n, \end{cases}$$

der **Binomialkoeffizient** von n und m , gelesen: „ n über m “.

Die Zahl $\binom{n}{m}$ drückt aus, wie viele verschiedene genau m -elementige Teilmengen man aus einer Menge mit n Elementen bilden kann. Statt von m -elementigen Teilmengen spricht man auch von **Kombinationen** von m verschiedenen Elementen aus einer Menge von n Elementen. Da eine Kombination eine Menge ist, spielt die Reihenfolge ihrer Elemente keine Rolle. Beispielsweise betrachten wir Kombinationen von drei Elementen aus der Menge $\{1, 2, 3, 4, 5\}$. Eine Kombination ist z. B. $\{1, 3, 4\}$. $\{3, 1, 4\}$ ist keine weitere Kombination, sie stimmt mit $\{1, 3, 4\}$ überein. Die Anzahl der Kombinationen erhält man nun so: Zunächst zählen wir auch unterschiedliche Reihenfolgen als unterschiedliche Kombinationen. Dann ergeben sich für das erste Element der Kombination n Möglichkeiten, für das zweite $n-1$, bis schließlich für das m -te noch $n-m+1$ Elemente zur Verfügung stehen. Die Anzahl ist also $n!/(n-m)!$. Beim Lotto bildet man Teilmengen von $m=6$ Elementen aus einer Menge mit $n=49$ Elementen. Würde die Reihenfolge der Zahlen eine Rolle spielen, so hätte man

$$49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44 = \frac{49 \cdot 48 \cdots 1}{43 \cdot 42 \cdots 1} = \frac{49!}{(49-6)!}$$

verschiedene mögliche Ziehungsergebnisse. Nun müssen wir noch ermitteln, wie oft wir die gleiche Kombination gezählt haben. m Zahlen kann man aber gerade auf $m!$ Weisen anordnen: Für die erste Position gibt es m Möglichkeiten, für die zweite $m-1$ usw. Dividieren wir durch die Anzahl der Mehrfachzählungen, ergibt sich genau der Binomialkoeffizient, der aufgrund dieser Überlegung insbesondere eine natürliche Zahl ist. Beim Lotto gibt es also $\frac{49!}{(49-6)! \cdot 6!} = \binom{49}{6} = 13\,983\,816$ verschiedene Zahlenkombinationen aus sechs Zahlen.

Satz 1.8 (Rechenregeln für Binomialkoeffizienten)

Für $n, m \in \mathbb{N}_0$ gilt:

- a) $\binom{n}{m} = \binom{n}{n-m},$
- b) $\binom{n}{0} = \binom{n}{n} = 1$ und $\binom{n}{n-1} = \binom{n}{1} = n.$
- c) Eine wichtige Beziehung, mit der man die Binomialkoeffizienten sukzessive berechnen kann, ist für $1 \leq m \leq n$:

$$\binom{n}{m-1} + \binom{n}{m} = \binom{n+1}{m}. \quad (1.2)$$

Beweis: Die Regeln folgen direkt aus der Definition des Binomialkoeffizienten.

a) Für $n < m$ sind beide Seiten null, sonst gilt:

$$\binom{n}{m} = \frac{n!}{(n-m)! \cdot m!} = \frac{n!}{m! \cdot (n-m)!} = \frac{n!}{(n-(n-m))! \cdot (n-m)!} = \binom{n}{n-m}.$$

b) $\binom{n}{0} = \frac{n!}{(n-0)! \cdot 0!} = 1$ und $\binom{n}{n} = \frac{n!}{(n-n)! \cdot n!} = 1$ sowie für $n \geq 1$:

$$\binom{n}{n-1} = \frac{n!}{1! \cdot (n-1)!} = n \text{ und } \binom{n}{1} = \frac{n!}{(n-1)! \cdot 1!} = n.$$

Für $n = 0$ sind nach Definition auch die beiden Binomialkoeffizienten gleich null.

$$\begin{aligned} \text{c) } \binom{n}{m-1} + \binom{n}{m} &= \frac{n!}{(n-m+1)! (m-1)!} + \frac{n!}{(n-m)! m!} = \frac{n! m + n! (n-m+1)}{(n-m+1)! m!} \\ &= \frac{n! (n+1)}{(n-m+1)! m!} = \binom{n+1}{m}. \end{aligned} \quad \square$$

Die Binomialkoeffizienten können mittels (1.2) über ein Schema (**Pascal'sches Dreieck**, siehe Abbildung 1.7) berechnet werden, wobei am Rand stets der Wert 1 steht. Im Inneren des Dreiecks erhält man den Wert eines Binomialkoeffizienten durch Addition der beiden Vorgänger.

1.3.2 Rationale Zahlen

Als Nächstes soll auch jede Division mit einer Zahl aus $\mathbb{Z} \setminus \{0\}$ erklärt sein. Dazu erweitert man $\mathbb{Z}$ zu

$$\mathbb{Q} := \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{Z} \setminus \{0\} \right\},$$

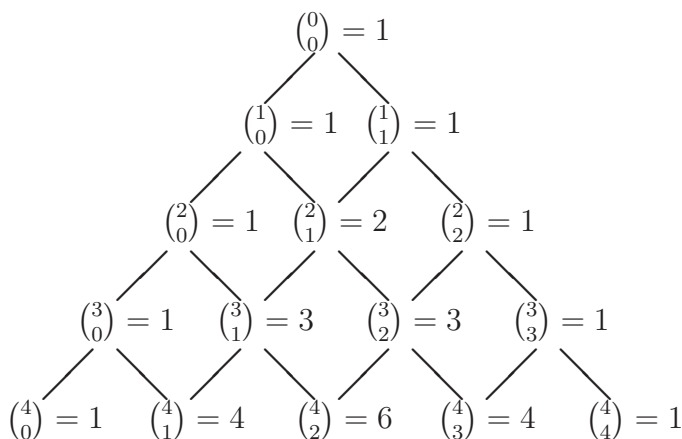


Abb. 1.7: Pascal'sches Dreieck zur Berechnung der Binomialkoeffizienten

die Menge der **rationalen Zahlen (Brüche)**. Die rationalen Zahlen $\mathbb{Q}$ sind also Brüche ganzer Zahlen. Ein und dieselbe rationale Zahl $x = \frac{p}{q}$ kann in verschiedener Weise als Bruch dargestellt werden, z. B. $\frac{2}{3} = \frac{4}{6} = \frac{20}{30}$. Die Darstellung wird eindeutig, wenn verlangt wird, dass p und q teilerfremd sind, d. h. eine weitere Kürzung des Bruchs nicht mehr möglich ist, und dass der Nenner positiv ist.

In der Schule benutzt man eine Darstellung von Brüchen als gemischte Zahlen: $\frac{10}{3} = 3 + \frac{1}{3}$ wird geschrieben als $3\frac{1}{3}$. Hier besteht die Gefahr der Verwechslung mit $3 \cdot \frac{1}{3}$. Da es üblich ist, den Multiplikationspunkt ganz wegzulassen, weiß man nicht, ob $3\frac{1}{3} = 3 + \frac{1}{3}$ oder $3\frac{1}{3} = 3 \cdot \frac{1}{3}$ gemeint ist. Daher raten wir von der Verwendung gemischter Zahlen ab.

Die ganzen und die rationalen Zahlen sind total geordnet. Wir benutzen die üblichen Symbole $\geq$, $>$, $\leq$ und $<$.

1.3.2.1 Rechnen mit rationalen Zahlen (Bruchrechnung)

In diesem Abschnitt stellen wir zur Erinnerung die wichtigsten Rechenregeln und elementare Begriffe für den Umgang mit rationalen Zahlen übersichtlich zusammen.

- **Kürzen** (Entfernen von gleichen Faktoren in Zähler und Nenner):

$$\frac{m \cdot p}{m \cdot q} = \frac{p}{q}, \quad m \cdot q \neq 0.$$

Achtung: In der Regel ist $\frac{m+p}{m+q} \neq \frac{p}{q}$.

- **Erweitern** ist umgekehrtes Kürzen, man multipliziert Zähler und Nenner mit einem gleichen Faktor.

- **Addition:** Hier muss man die Summanden gleichnamig machen, d. h. den Nenner durch Erweitern angleichen:

$$\frac{m}{n} + \frac{p}{q} := \frac{m \cdot q + p \cdot n}{n \cdot q}.$$

Das Ergebnis kann nun gekürzt werden. Eleganter ist es, den gemeinsamen Nenner als das kleinste gemeinsame Vielfache von n und q zu wählen, z. B.:

Beispiel 1.26

$$\frac{3}{6} + \frac{4}{15} = \frac{3}{2 \cdot 3} + \frac{4}{3 \cdot 5} = \frac{3 \cdot 5 + 4 \cdot 2}{2 \cdot 3 \cdot 5} = \frac{23}{30}. \quad \blacksquare$$

- **Multiplikation**

$$\frac{m}{n} \cdot \frac{p}{q} := \frac{m \cdot p}{n \cdot q} = \frac{mp}{nq}.$$

- **Division** ist Multiplikation mit dem Kehrwert

$$\frac{\frac{m}{n}}{\frac{p}{q}} = \frac{m}{n} \cdot \frac{q}{p} = \frac{mq}{np}.$$

Werden beispielsweise Widerstände R_1 und R_2 parallel geschaltet, so ergibt sich als Gesamtwiderstand R_g :

$$R_g = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2}} = \frac{1}{\frac{R_2}{R_1 R_2} + \frac{R_1}{R_1 R_2}} = \frac{1}{\frac{R_2 + R_1}{R_1 R_2}} = \frac{R_1 R_2}{R_2 + R_1}.$$

- **Potenzen und Wurzeln:** Für $q \in \mathbb{Q}$ und $n \in \mathbb{N}$ ist

$$q^n = \underbrace{q \cdot q \cdot \dots \cdot q}_{n\text{-mal}}, \quad q^{-n} := \frac{1}{q^n}, \quad q \neq 0.$$

q heißt **Basis**, n heißt **Exponent**, q^n die **n -te Potenz** von q . Für $q \neq 0$ ist $q^0 := 1$, 0^0 ist nicht definiert, da man sonst Widersprüche bekommt (wir sehen uns mit Hilfe der Differenzialrechnung in Beispiel 2.82 auf Seite 327 an, dass 0^0 jede Zahl sein kann). Aus Bequemlichkeit wird diese Regel häufig lockerer gehandhabt, wenn x eine Variable ist und man x^0 betrachtet. Für alle $x \neq 0$ ist $x^0 = 1$, so dass man das Symbol fortsetzt (stetige Fortsetzung der Funktion, siehe Seite 285) und auch für $x = 0$ mit dem Wert 1 rechnet. Keinesfalls darf man aber ohne zusätzliche Überlegungen $0^0 = 1$ rechnen.

Mit dem negativen Exponenten -1 haben wir für einen Kehrwert die Darstellung $\frac{1}{q} = q^{-1}$. Der Exponent -1 wird als Symbol auch für Umkehrabbildungen verwendet, hat dabei aber eine völlig andere Bedeutung.

Es gelten die **Potenzregeln** für $p, q \in \mathbb{Q}$, $q \neq 0$ und $n, m \in \mathbb{Z}$:

$$\begin{aligned} q^n \cdot q^m &= q^{n+m}, & [q^n]^m &= q^{n \cdot m}, \\ p^n \cdot q^n &= (p \cdot q)^n, & \frac{p^n}{q^n} &= \left(\frac{p}{q}\right)^n, & q &\neq 0. \end{aligned}$$

Als **n -te Wurzel** einer rationalen Zahl $q \geq 0$ bezeichnet man die (nicht-negative) Lösung x der Gleichung

$$x^n = q \wedge x \geq 0 \iff x = \sqrt[n]{q}.$$

Für $n = 2$ spricht man von der Quadratwurzel, für $n = 3$ von der Kubikwurzel. So gilt beispielsweise $2^2 = 4 \implies 2 = \sqrt{4}$, und entsprechend gilt $\sqrt{9} = 3$ und $\sqrt[3]{27} = 3$. Allerdings gibt es nicht immer eine n -te Wurzel innerhalb der Menge $\mathbb{Q}$, dafür benötigen wir die reellen Zahlen.

■ Prozentrechnung

$x \%$ (x **Prozent**, x vom Hundert) ist nur eine andere Schreibweise für $\frac{x}{100}$.

Damit ist $x \%$ von y das Produkt $\frac{x}{100} \cdot y$. Entsprechend erhält man $x \%$ mehr als y als Produkt $(1 + \frac{x}{100}) \cdot y$. Zur Umrechnung einer Zahl in eine Prozentangabe kann man mit $100 \% = 1$ multiplizieren: $0,8 = 0,8 \cdot 100 \% = 80 \%$.

Wie viel Mehrwertsteuer ist bei einem Steuersatz von 19% in einem Preis von x Euro enthalten? Die Antwort ist **nicht** $0,19 \cdot x$ Euro, da die Steuer vom noch nicht versteuerten Betrag y berechnet wird, also $y \cdot 1,19 = x$. Dieser ist $y = x/1,19$. Die Steuer beträgt also $x - x/1,19 = \frac{0,19}{1,19} \cdot x$.

Im Ausverkauf wird häufig mit „reduzierte Artikel nochmals 20% günstiger“ geworben. Um welchen Prozentsatz ist ein bereits mit 30% reduzierter Artikel dann herabgesetzt? Die Antwort ist **nicht** 50% , denn der Preis muss mit $0,7 \cdot 0,8 = 0,56$ multipliziert werden. Damit ist er nur 44% günstiger.

Achtung: Möchte man zwei Prozentsätze x und y zu $x + y$ addieren, so sagt man x plus y **Prozentpunkte**. 20% plus 30 Prozentpunkte sind also 50% .

1.3.2.2 Dezimalbruchdarstellung rationaler Zahlen

Rationale Zahlen besitzen eine Dezimalbruchdarstellung, die aus einer ganzen Zahl vor dem Komma und dann aus endlich oder unendlich vielen Nachkommastellen besteht. $32,3523 = 32 + 3 \cdot 10^{-1} + 5 \cdot 10^{-2} + 2 \cdot 10^{-3} + 3 \cdot 10^{-4}$. Bei unendlich vielen Nachkommastellen muss man unendlich viele Summanden addieren. Damit beschäftigen wir uns später im Kapitel über Reihen in Beispiel 2.19 auf Seite 237.

Lemma 1.2 (Dezimalbruchdarstellung rationaler Zahlen)

- a) Jede rationale Zahl $q \in \mathbb{Q}$ besitzt eine Dezimalbruchdarstellung mit endlich vielen Nachkommastellen oder eine periodische Dezimalbruchdarstellung, d. h., ab einer Nachkommastelle wiederholt sich ein Ziffernblock endlicher Länge unendlich oft. Dieser Ziffernblock heißt **Periode**.
- b) Umgekehrt ist jeder endliche oder periodische Dezimalbruch eine rationale Zahl.

Eine Periode wird mit einem Balken gekennzeichnet: $\frac{1}{3} = 0,3333\dots = 0,\overline{3}$ und $\frac{116}{990} = 0,1171717\dots = 0,1\overline{17}$.

- Beweis:** a) Hat man einen Bruch $\frac{p}{q}$, so kann man die Dezimalbruchdarstellung durch fortgesetzte schriftliche Division gewinnen. Bei jeder Division durch q erhält man einen Rest $0 \leq r < q$. Erreicht man $r = 0$, so bricht die Dezimalbruchdarstellung ab. Sonst erhält man spätestens nach q Divisionen einen Rest, der schon einmal aufgetreten ist. Spätestens ab hier wiederholen sich Ziffern, die Periode entsteht.
- b) Dass umgekehrt jeder endliche Dezimalbruch eine rationale Zahl ist, ist offensichtlich. Dass auch alle periodischen Dezimalbrüche rationale Zahlen sind, sieht man so:

$$\begin{aligned} & b_n b_{n-1} \dots b_0, c_1 c_2 \dots c_m \overline{a_1 a_2 \dots a_k} \\ &= b_n b_{n-1} \dots b_0, c_1 c_2 \dots c_m + 10^{-m} \cdot 0, \overline{a_1 a_2 \dots a_k}. \end{aligned}$$

Es bleibt also zu zeigen, dass der periodische Dezimalbruch $r = 0, \overline{a_1 a_2 \dots a_k}$ als Bruch schreibbar ist.

$$10^k \cdot r = a_1 a_2 \dots a_k, \overline{a_1 a_2 \dots a_k}.$$

Subtrahieren wir r , erhalten wir $(10^k - 1) \cdot r = a_1 a_2 \dots a_k$. Weiter folgt

$$r = \frac{a_1 a_2 \dots a_k}{10^k - 1} = \frac{a_1 a_2 \dots a_k}{\underbrace{99 \dots 9}_{k \text{ Stellen}}}. \quad (1.3)$$

Damit haben wir aber die Darstellung als Bruch. □

Im Beweis haben wir insbesondere gesehen, wie man mit (1.3) einen periodischen Dezimalbruch als Bruch umschreibt. Dabei gibt es eine Besonderheit für Neunerperioden:

$$0,\overline{9} = \frac{9}{9} = 1.$$

Eine Zahl kann also zwei Dezimalbruchdarstellungen haben.

Beispiel 1.27

$$0,12\overline{34} = 0,12 + 0,01 \cdot 0,\overline{34} = \frac{12}{100} + \frac{1}{100} \cdot \frac{34}{99} = \frac{3}{25} + \frac{17}{4950} = \frac{611}{4950}.$$

$31 : 8 = 3,875$	$5 : 12 = 0,41\overline{6}$
$\underline{24}$	50
70	$\underline{48}$
$\underline{64}$	20
60	$\underline{12}$
$\underline{56}$	80
40	$\underline{72}$
$\underline{40}$	80
0	$\underline{72}$

■

1.3.2.3 Abzählbarkeit *

Dieser Abschnitt richtet sich an mathematisch Interessierte und ist für praktische Anwendungen eher unwichtig. Dennoch ist das Thema spannend. Alle bislang eingeführten Zahlenmengen haben unendlich viele Elemente. Aber man hat den naiven Eindruck, dass es z. B. mehr Brüche als natürliche Zahlen gibt. Dieser Eindruck ist falsch! Die Mengen

$$\mathbb{N} \subset \mathbb{N}_0 \subset \mathbb{Z} \subset \mathbb{Q}$$

sind **abzählbar**, das heißt, man kann jedem Element der Menge eindeutig ein Element aus $\mathbb{N}$ zuordnen und umgekehrt. Es gibt also eine bijektive Abbildung auf $\mathbb{N}$. Diese gibt die Position einer Zahl in einer unendlich langen Aufzählung an, die alle Zahlen der jeweiligen Menge umfasst. In diesem Sinne ist die unendlich große Menge $\mathbb{Q}$ nicht mächtiger als die unendlich große Menge $\mathbb{N}$.

Wir zeigen die Abzählbarkeit von $\mathbb{Z}$ und $\mathbb{Q}$:

- $\mathbb{Z}$ kann in eine Abzählreihenfolge gebracht werden: $0, 1, -1, 2, -2, 3, -3, 4, -4, \dots$
Man erhält so alle Elemente von $\mathbb{Z}$. Eine bijektive Abbildung $f : \mathbb{N} \rightarrow \mathbb{Z}$ ist

$$f(k) := \begin{cases} -\frac{k-1}{2} & \text{falls } k \text{ ungerade ist,} \\ \frac{k}{2} & \text{falls } k \text{ gerade ist.} \end{cases}$$

- Bei $\mathbb{Q}$ ist es schwieriger, eine Abzählreihenfolge herzustellen. Hier hilft die **Cantor-Diagonalisierung**. Wir schreiben alle positiven Brüche in das Schema aus Abbildung 1.8 mit unendlich vielen Zeilen und Spalten. Dabei treten Mehrfachzählungen durch Kürzen auf (z. B. 1 auf der Hauptdiagonalen). Wir können nun die Brüche in eine Reihenfolge bringen, indem wir die unendliche Tabelle diagonalenweise (von links unten nach rechts oben) lesen und Mehrfachnennungen weglassen. Das funktioniert, da jede Diagonale eine endliche Länge hat, und wir erhalten die Reihenfolge: $1 = \frac{1}{1}, \frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, 4, \dots$ Negative Brüche kann man genauso abzählen, dann beide Zählungen wie bei $\mathbb{Z}$ zusammenführen und die Null als erste Zahl ergänzen.

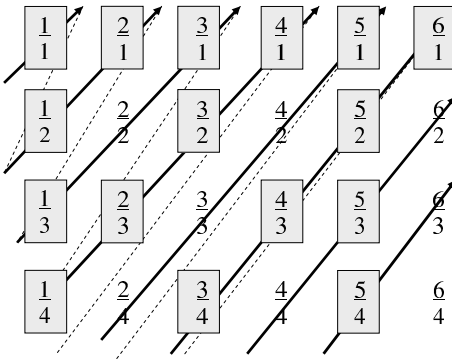


Abb. 1.8: Cantors erstes Diagonalargument

Beispiel 1.28 (Abzählbarkeit und Kreuzprodukt von Mengen)

Wir überlegen uns nun, dass auch das Kreuzprodukt zweier abzählbarer Mengen $A = \{a_1, a_2, \dots\}$ und $B = \{b_1, b_2, \dots\}$ wieder abzählbar ist. Das funktioniert genau wie beim Cantor-Diagonalargument. Wir müssen lediglich die Werte von Zähler und Nenner als erstes und zweites Element eines Zahlenpaars auffassen. Wir erhalten so ein Schema mit unendlich vielen Zeilen und Spalten:

$$\begin{array}{cccc}
 (a_1, b_1) & (a_1, b_2) & (a_1, b_3) & \dots \\
 (a_2, b_1) & (a_2, b_2) & (a_2, b_3) & \dots \\
 (a_3, b_1) & (a_3, b_2) & (a_3, b_3) & \dots \\
 \vdots & & &
 \end{array}$$

Wenn wir dieses Schema entlang der endlich langen Diagonalen durchlaufen, dann haben wir eine Abzählreihenfolge:

$$(a_1, b_1), (a_2, b_1), (a_1, b_2), (a_3, b_1), (a_2, b_2), (a_1, b_3), \dots$$

Im Gegensatz zu den Brüchen haben wir hier nicht das Problem, dass durch Kürzen gleiche Elemente entstehen können.

Wir wissen jetzt also, dass z.B. die Mengen $\mathbb{N} \times \mathbb{N}$ oder $\mathbb{Z} \times \mathbb{Q}$ abzählbar sind. Wir können das Kreuzprodukt auch mehrfach hintereinander anwenden und erhalten immer noch abzählbare Mengen, z.B. $\mathbb{N} \times \mathbb{Z} \times \mathbb{Q}$ oder $\mathbb{N}^4 := \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \mathbb{N}$. ■

Der Umgang mit dem Unendlichen ist ungewohnt. Ein überraschendes Beispiel dazu ist das **Hilbert'sche Hotel**. Dieses hat abzählbar unendlich viele Zimmer und ist mit unendlich vielen Gästen komplett ausgebucht. Trotzdem kann es noch einen weiteren Gast aufnehmen: Wenn alle Gäste in das jeweils nächste Zimmer wechseln, dann wird das erste Zimmer für den neuen Gast frei. Hätte das Hotel nur endlich viele Zimmer, dann würde der Gast des letzten Zimmers keinen neuen Raum bekommen. So muss aber kein Gast das Hotel verlassen. Sind bereits alle Zimmer mit den natürlichen Zahlen besetzt, dann gibt es sogar noch immer genug Platz für alle weiteren rationalen Zahlen.

1.3.2.4 Vollständige Induktion

Oft treten Gleichungen oder Ungleichungen auf, die eine natürliche Zahl als Parameter enthalten. Ein Beispiel hierfür ist die Ungleichung

$$2^n \geq n$$

für $n \in \mathbb{N}$. Soll man zeigen, dass eine solche Aussage für alle Werte des Parameters gilt, so könnte man eine Vielzahl von Werten einsetzen. Das ist aber mühsam und unsicher. Außerdem kann man die Aussage nicht für alle unendlich vielen Werte nachrechnen. Die Lösung besteht in einem Beweisverfahren, das **Vollständige Induktion** heißt. Man verwendet es, um zu zeigen, dass eine Aussageform $A(n)$ für jeden Wert $n \in \mathbb{N}$ (oder $n \in \mathbb{N}_0$ oder allgemeiner: für alle Elemente n einer abzählbaren Menge) wahr wird. Die zu zeigende Aussage ist also:

$$\text{Für alle } n \in \mathbb{N} \text{ gilt: } A(n),$$

z. B.: Für alle $n \in \mathbb{N}$ gilt: $2^n \geq n$.

Dazu zeigt man zunächst, dass die Aussage $A(1)$ wahr ist. Dann zeigt man, dass wenn für ein $n \in \mathbb{N}$ die Aussage $A(n)$ gilt, auch die Aussage $A(n+1)$ wahr ist. Damit ist man bereits fertig und hat die Aussage für jeden Parameterwert als wahr identifiziert. Das ist wie beim Umfallen von Dominosteinen: Weiß man, dass, wenn ein Stein fällt ($A(n)$ ist wahr) auch sein Nachfolger fällt (dann ist $A(n+1)$ wahr), so reicht das Umwerfen des ersten Steins ($A(1)$ ist wahr), um in einer Kettenreaktion auch alle anderen umzuwerfen: Da $A(1)$ wahr ist, ist auch $A(2)$ wahr. Da $A(2)$ wahr ist, ist auch $A(3)$ wahr usw.

Der Beweis durch Vollständige Induktion wird üblicherweise in drei Schritten aufgeschrieben:

- Induktionsanfang: Zeige, dass $A(1)$ richtig (wahr) ist, der erste Stein fällt also. Im Beispiel ist die Aussage $2^1 = 2 \geq 1$ wahr.

Betrachtet man statt $\mathbb{N}$ eine andere abzählbare Menge, so muss man die Aussage für deren erstes Element zeigen, also z. B. $A(0)$ für $\mathbb{N}_0$.

- **Induktionsannahme:** Wir müssen im folgenden Induktionsschluss zeigen, dass der $n + 1$ -te Dominostein fällt (dass also $A(n + 1)$ wahr ist), wenn der n -te fällt (also $A(n)$ wahr ist). Also müssen wir nur etwas in der Situation tun, in der $A(n)$ wahr ist. Deshalb nimmt man $A(n)$ als wahr an (Induktionsannahme).

Im Beispiel nehmen wir an, dass $2^n \geq n$ für eine Zahl n gilt.

- **Induktionsschluss:** Nun müssen wir unter der Annahme, dass $A(n)$ wahr ist, tatsächlich nachrechnen, dass auch $A(n + 1)$ wahr ist, die Aussage also für die nächste Zahl gilt. Im Beispiel:

$$2^{n+1} = 2 \cdot 2^n \stackrel{\text{Induktionsannahme}}{\geq} 2n = n + n \geq n + 1.$$

Damit haben wir die Ungleichung für alle $n \in \mathbb{N}$ bewiesen. Denn für $n = 1$ wird die Induktionsannahme durch den Induktionsanfang von einer angenommen wahren Aussage zu einer tatsächlich wahren Aussage. Aus dieser folgt mit dem Induktionsschluss, dass die Aussageform auch für $n = 2$ gilt. Die Annahme wird damit auch für $n = 2$ wahr, so dass man mit dem Induktionsschluss die Aussage für $n = 3$ bewiesen hat, usw.

Wir erinnern uns an die Definition der Folgerung $A(n) \implies A(n + 1)$: Die Folgerung ist genau dann wahr, wenn entweder $A(n)$ falsch ist oder wenn gleichzeitig $A(n)$ und $A(n + 1)$ wahr sind. Damit beweisen wir im Induktionsschluss, dass

$$\text{für alle } n \in \mathbb{N} \text{ gilt: } A(n) \implies A(n + 1). \quad (1.4)$$

Mit (1.4) kann man von der mit dem Induktionsanfang gezeigten Gültigkeit für $A(1)$ streng logisch (mit dem Modus Ponens) auf die Gültigkeit für alle Elemente schließen, die man sukzessive wie bei einer Kette von Dominosteinen als Nachfolger erreicht:

$$A(1) \stackrel{(1.4)}{\implies} A(2) \stackrel{(1.4)}{\implies} A(3) \stackrel{(1.4)}{\implies} A(4) \stackrel{(1.4)}{\implies} \dots$$

Beispiel 1.29

Wir beweisen, dass $2^n \leq n!$ für alle natürlichen Zahlen $n \geq 4$ gilt.

- **Induktionsanfang:** Für $n = 4$ ist $2^4 = 16 \leq 4! = 24$ wahr. Man beachte, dass wir die Aussage erst ab $n = 4$ beweisen müssen, daher beginnen wir nicht mit $n = 1$. Tatsächlich gilt sie auch nicht für $n = 3$.
- **Induktionsannahme:** $2^n \leq n!$ gelte für ein $n \geq 4$. Wenn man sich zum ersten Mal mit Vollständiger Induktion beschäftigt, dann erscheint die Induktionsannahme sehr merkwürdig: Man nimmt die Aussage an, die man eigentlich zeigen soll. Tatsächlich muss man aber nur beweisen, dass, wenn die Aussage für eine Zahl n gilt, sie auch für den nächsten Wert $n + 1$ gilt. Man muss also nur für den Fall etwas tun, in dem sie tatsächlich für einen Wert n gilt. Deshalb macht man den Induktionsschluss nur unter dieser Annahme.

- Induktionsschluss: Zu zeigen ist die Aussage für n ersetzt durch $n + 1$, also $2^{n+1} \leq (n + 1)!$. Dazu benutzen wir die Induktionsannahme:

$$2^{n+1} = 2 \cdot 2^n \leq 2 \cdot n! \leq (n + 1) \cdot n!, \text{ also } 2^{n+1} \leq (n + 1)!.$$

Damit ist die Behauptung bewiesen. Denn wenn sie für n gilt, dann auch für $n + 1$. Sie gilt für $n = 4$ laut Induktionsanfang und damit nach Induktionsschluss auch für $n = 5$. Nochmalige Anwendung des Induktionsschlusses liefert die Aussage für $n = 6$ usw. ■

Beispiel 1.30 (Korrektheit von Programmen *)

Bei der Entwicklung von Software testet man, um Fehler zu finden. Dabei wird man bei größeren Programmen nie alle Fehler finden. Wenn, wie in der Raumfahrt oder Medizin, Fehler erhebliche Konsequenzen haben können, dann kann es notwendig werden, die Korrektheit von Programmen formal zu beweisen (wobei man dann leider auch wieder Fehler machen kann). Hier hilft ein ähnliches Vorgehen wie bei der Vollständigen Induktion. Wir betrachten ein in der Programmiersprache C geschriebenes Programmfragment, das den größten Zahlenwert aller in einem Feld `a[]` gespeicherten $n+1$ Zahlen `a[0]`, `a[1]`, ..., `a[n]` finden soll:

```
max=a[0];
for(k=1;k<=n;k=k+1)
    if(a[k]>max) max=a[k]; /*a[k] ist bislang maximaler Wert*/
```

Die Beweisidee besteht in der Verwendung einer Schleifeninvariante. Das ist eine Bedingung, die nach jedem Schleifendurchlauf erfüllt ist und die wir schon beim Euklid'schen Divisionsalgorithmus auf Seite 31 angesprochen haben. Wir wählen hier die Invariante: „Nach dem k -ten Schleifendurchlauf ist `max` die größte der Zahlen `a[0]`, `a[1]`, ..., `a[k]`.“. Wenn wir diese Bedingung zeigen können, dann liefert das Programm nach den n programmierten Schleifendurchläufen tatsächlich das gesuchte Maximum. Die Schleifeninvariante zeigt man nun analog zur Vollständigen Induktion. Dem Induktionsanfang entspricht die Initialisierung `max=a[0]`. Nach 0 Schleifendurchläufen ist die Bedingung also erfüllt. Jetzt müssen wir zeigen: Ist die Bedingung nach k Durchläufen erfüllt (Induktionsannahme), dann auch nach $k + 1$ (Induktionsschluss). Das ist hier aber offensichtlich, da das bisherige Maximum mit dem $k + 1$ -ten Wert verglichen und entsprechend modifiziert wird. ■

Wir werden die Vollständige Induktion später zum Beweis einiger Summenformeln und insbesondere des Binomischen Lehrsatzes einsetzen.

1.3.3 Reelle Zahlen

1.3.3.1 Von den rationalen zu den reellen Zahlen

Nicht-periodische Dezimalbrüche mit unendlich vielen Nachkommastellen sind keine Brüche. Das ist der Grund, warum die bereits mächtige Menge $\mathbb{Q}$ nicht ausreicht, wenn wir später Grenzwerte diskutieren werden. Hier wird es um Fragestellungen gehen wie beispielsweise die Angabe der Ladungen pro Zeit, die einen Leiterquerschnitt passieren, wenn die Zeit beliebig klein wird. Der Grenzwert ist die Stromstärke. Man benötigt dazu eine Zahlenmenge, in der alle möglichen Grenzwerte als Zahl erklärt sind. Solche Zahlenmengen heißen **vollständig**. Im Gegensatz zu $\mathbb{Q}$ sind die reellen Zahlen $\mathbb{R}$ vollständig. Man erreicht das, indem man zusätzlich zu den endlichen und periodischen Dezimalbrüchen auch unendliche, nicht-periodische als Zahlen verwendet. Für die Ingenieurmathematik reicht das als Definition schon völlig aus. Wir wollen aber etwas genauer vorgehen und verwenden das Vollständigkeitsaxiom. Die Darstellung reeller Zahlen als Dezimalbrüche ist dann eine Folgerung. Wenn es Ihnen keinen Spaß macht, dann müssen Sie diese Ausführungen nicht im Detail nachvollziehen. Einige Begriffe wie Körper, obere und untere Schranke sollten Sie sich aber dennoch kurz ansehen.

Man muss keine Grenzwerte bemühen, um zu sehen, dass $\mathbb{Q}$ für viele Zwecke noch nicht groß genug ist. Nach dem Satz von Pythagoras ist die Länge x der Diagonale in einem Quadrat mit Kantenlänge 1 gegeben durch die Gleichung $x^2 = 1^2 + 1^2$, d. h. als Lösung der Gleichung $x^2 = 2$. Diese hat aber keine Lösung in $\mathbb{Q}$. Bereits die antiken Griechen wussten, dass es Streckenlängen gibt, die keine Brüche sind.

Satz 1.9 (Beschränkung der rationalen Zahlen)

Die Lösung der Gleichung $x^2 = 2$ liegt nicht in $\mathbb{Q}$.

Beweis: Gäbe es einen Bruch $\frac{p}{q}$, $p, q \in \mathbb{N}$, mit $p^2/q^2 = 2$, also $p^2 = 2q^2$, so wäre dies ein Widerspruch zur eindeutig existierenden Primzahlzerlegung von p und q : In p^2 und q^2 kommt jeder Primfaktor wegen des Quadrats geradzahlig oft vor, also 0-mal oder 2-mal oder 4-mal oder 6-mal... Insbesondere kommt 2 in p^2 geradzahlig oft vor, in $2q^2$ wegen des zusätzlichen Faktors jedoch ungeradzahlig oft. $\square$

Man benötigt aber auch die positive Streckenlänge x , die Lösung der Gleichung $x^2 = 2$ ist, als reelle Zahl $\sqrt{2}$, ebenso sollten alle anderen denkbaren Streckenlängen als reelle Zahlen schreibbar sein. Daher werden die reellen Zahlen so definiert, dass sie sich auf der **Zahlengeraden** anordnen lassen. Jede reelle Zahl entspricht dabei einem Punkt auf der Geraden. Beispielsweise lässt sich der Punkt für $\sqrt{2}$ über die Länge der Diagonale des Einheitsquadrats geometrisch konstruieren (siehe Abbildung 1.9).

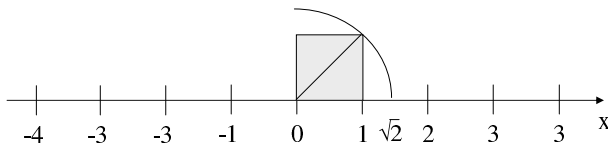


Abb. 1.9: $\sqrt{2}$ auf der Zahlengeraden

Um einen Zahlbereich zu erhalten, in dem auch die Lösungen $\pm\sqrt{2}$ liegen, muss $\mathbb{Q}$ so erweitert werden, dass die Lücken auf der Zahlengeraden geschlossen werden.

Zunächst sollen im erweiterten Zahlenbereich die gleichen Rechenregeln wie in $\mathbb{Q}$ gelten. Diese definiert man über Körperaxiome:

Definition 1.18 (Körperaxiome)

Sei $\mathbb{K}$ eine Menge (von Zahlen), so dass zwischen zwei Elementen Verknüpfungen „+“ und „ $\cdot$ “ als Abbildung von $\mathbb{K} \times \mathbb{K} \rightarrow \mathbb{K}$ erklärt sind. $\mathbb{K}$ zusammen mit diesen Verknüpfungen heißt ein **Körper** genau dann, wenn für alle Elemente $a, b, c \in \mathbb{K}$ die folgenden Axiome gelten:

Axiome zur Addition:

- **Kommutativgesetz:** $a + b = b + a$,
- **Assoziativgesetz:** $(a + b) + c = a + (b + c)$,
- *Existenz eines neutralen Elements* $0 \in \mathbb{K}$: $a + 0 = a$.
- *Jedes a hat ein inverses Element* $-a \in \mathbb{K}$, so dass $a + (-a) = a - a = 0$.

Axiome zur Multiplikation:

- **Kommutativgesetz:** $a \cdot b = b \cdot a$,
- **Assoziativgesetz:** $(a \cdot b) \cdot c = a \cdot (b \cdot c)$,
- *Existenz eines neutralen Elements* $1 \in \mathbb{K}$: $1 \cdot a = a$.
- *Jedes $a \neq 0$ hat ein inverses Element* $a^{-1} \in \mathbb{K}$, so dass $a \cdot a^{-1} = 1$.

Distributivgesetz (zwischen Addition und Multiplikation): $(a + b) \cdot c = a \cdot c + b \cdot c$.

Einige dieser Regeln haben wir bereits für Mengenoperationen und Logik-Verknüpfungen kennengelernt.

Hintergrund: Gruppen und Ringe

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 Ver-

knüpfung eine **Gruppe**. Ist zudem ein Kommutativgesetz erfüllt, so spricht man von einer **kommutativen Gruppe** oder **Abel'schen Gruppe** nach dem Mathematiker Niels 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 1.31 (Restklassen)

Definiert man auf der Menge $\mathbb{Z}_n := \{0, 1, 2, \dots, n-1\}$ 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 $0, \dots, n-1$ Vertreter (Repräsentanten) von **Restklassen**. So steht beispielsweise die Zahl 2 für die Restklasse $\{2, 2+n, 2-n, 2+2n, 2-2n, \dots\}$. 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 $0, \dots, n-1$ können wir auch direkt mit Restklassen rechnen: Zwei Restklassen werden addiert, indem man aus beiden Restklassen je 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 gelingt aber nur bei Primzahlen n . Die allgemeine Struktur, die entsteht, wenn man nicht nur Primzahlen nimmt, nennt man einen **Restklassenring**. Verzichtet man in Definition 1.18 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.

Beispiel 1.32 (RSA-Public-Key-Verschlüsselung)

Beim **RSA-Verschlüsselungsverfahren** wird eine als Zahl aus $\mathbb{Z}_n = \{0, 1, 2, \dots, n-1\}$ 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 irgendein 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) oder (Goebbels

und Rethmann, 2014, Kap. 3.2.3). 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 (Goebbels und Rethmann, 2014, Kap. 3.5). ■

1.3.3.2 Vollständigkeit und Einführung der reellen Zahlen

Zusätzlich zu diesen Rechenregeln soll die Zahlerweiterung die totale Ordnung der natürlichen und rationalen Zahlen erhalten, also eine entsprechende Ordnungsrelation $\leq$ besitzen (siehe Seite 27). Die totale Ordnung ist eine besondere Eigenschaft, die bei der nächsten Erweiterung zu den komplexen Zahlen verloren geht.

Definition 1.19 (Supremum)

Sei F eine Menge (von Zahlen) mit einer Ordnungsrelation $\leq$ und $E \subset F$ eine Teilmenge.

- E heißt genau dann **beschränkt nach oben**, falls es ein $M \in F$ gibt mit $x \leq M$ für alle $x \in E$. M heißt dann eine **obere Schranke** von E .
- M heißt eine **kleinste obere Schranke** oder **Supremum** von E (Notation: $M = \sup E$) genau dann, wenn M eine obere Schranke ist und es keine kleinere obere Schranke von E gibt, d. h., wenn es kein $M' \in F$ gibt, mit $M' < M$ und $x \leq M'$ für alle $x \in E$ (siehe Abbildung 1.10).
- M heißt ein **Maximum** von E (Notation: $M = \max E$) genau dann, wenn M eine kleinste obere Schranke ist und zusätzlich $M \in E$ gilt.

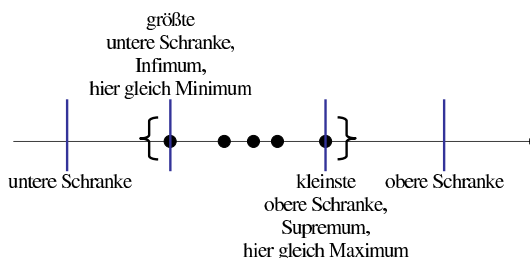


Abb. 1.10: Zur Definition des Supremums und Infimums

Beispiel 1.33

$F := \mathbb{N}$, $E := \{3, 5, 7\} \subset F$ ist nach oben beschränkt, z. B. mit $M = 20$, da $3 \leq 20$ und $5 \leq 20$ und $7 \leq 20$. E hat das Supremum 7 und auch das Maximum 7, da das Supremum 7 zur Menge E gehört. ■

Analog zur kleinsten oberen wird eine größte untere Schranke definiert:

Definition 1.20 (Infimum)

Sei F eine Menge (von Zahlen) mit einer Ordnungsrelation $\leq$ und $E \subset F$ eine Teilmenge.

- E heißt **beschränkt nach unten** genau dann, wenn es eine **untere Schranke** $m \in F$ von E gibt mit $x \geq m$ für alle $x \in E$.
- m heißt genau dann eine **größte untere Schranke** oder **Infimum** von E (Notation: $M = \inf E$), wenn m eine untere Schranke ist und es keine größere untere Schranke von E gibt, d. h., falls es kein $m' \in F$ gibt, mit $m' > m$ und $x \geq m'$ für alle $x \in E$.
- m heißt ein **Minimum** von E (Notation: $m = \min E$) genau dann, wenn m eine größte untere Schranke ist und zusätzlich $m \in E$ gilt.

Beispiel 1.34

- a) $E := \{3, 5, 7\} \subset \mathbb{N}$ ist nach unten beschränkt mit dem Infimum und Minimum 3.
- b) Die Menge $\{x \in \mathbb{Q} : x < 2\} \subset \mathbb{Q}$ ist zwar durch 2 nach oben beschränkt, und 2 ist auch die kleinste obere Schranke (das Supremum), aber 2 gehört nicht zur Menge, so dass es kein Maximum gibt.
- c) Michael Ende beschreibt in einem Abenteuer von Jim Knopf und Lukas dem Lokomotivführer die Einwohner von Mandala so, dass jeder einen kleineren an der Hand hält. Damit gibt es keinen oder unendlich viele Einwohner. Die Menge der Größen der Einwohner ist nach unten beschränkt mit 0. Es gibt aber kein Minimum, denn ein Einwohner, der so groß wäre wie das Minimum, müsste wieder einen kleineren an der Hand halten. ■

Die Existenz eines Minimums oder Maximums ist also nicht selbstverständlich.

Eine kleinste obere Schranke ist eindeutig, d. h., man kann von **der** kleinsten oberen Schranke und damit von **dem** Supremum reden, analog: **das** Infimum. Dies folgt direkt aus der Eigenschaft der Ordnungsrelation: Sind M und M' kleinste obere Schranken, dann gilt nach Definition weder $M < M'$ noch $M' < M$. Damit bleibt nur $M' = M$.

Definition 1.21 (Ordnungsvollständiger geordneter Körper*)

Ein Körper $\mathbb{K}$, der als Menge total geordnet ist und für den die Ordnung verträglich mit Addition und Multiplikation ist, d. h., für alle $a, b, c \in \mathbb{K}$ gilt

$$a > b \implies a + c > b + c,$$

$$a > 0 \wedge b > 0 \implies a \cdot b > 0,$$

und für den das sogenannte **Vollständigkeitsaxiom** gilt, heißt ein **ordnungsvollständiger geordneter Körper**. Das **Vollständigkeitsaxiom** lautet:

$$\begin{aligned} \text{Jede nach oben beschränkte, nicht-leere Teilmenge } E \subset \mathbb{K} \\ \text{hat ein Supremum in } \mathbb{K}. \end{aligned} \quad (1.5)$$

Man kann zeigen, dass es (bis auf Umbenennung der Zahlen) genau einen ordnungsvollständigen geordneten Körper gibt und dass dieser den Körper $\mathbb{Q}$ umfasst, siehe (Ebbinghaus et al., 1992, S. 42). Damit können wir für den ordnungsvollständigen geordneten Körper den Namen **reelle Zahlen** $\mathbb{R}$ vergeben, $\mathbb{R}$ ist durch das vorangehende Axiomensystem festgelegt.

Zugegeben, das Vollständigkeitsaxiom wirkt zunächst unnatürlich kompliziert. Die hier gewählte Formulierung eignet sich aber sehr für alle Beweise, in denen man die Existenz einer reellen Zahl zeigen möchte (z. B. später bei Grenzwerten). Anschaulich ist klar, dass $\sqrt{2}$ als Streckenlänge in einem rechtwinkligen Dreieck existiert, siehe Abbildung 1.9. Diese Anschauung muss man mit einem Axiom formulieren, damit man sich für weitere Rechnungen von der Anschauung lösen kann. Das Axiom erfüllt seinen Zweck: $\sqrt{2} := \sup\{x \in \mathbb{Q} : x^2 \leq 2\} \in \mathbb{R}$ existiert. In $\mathbb{Q}$ hat diese Menge dagegen kein Supremum.

Hintergrund: Alternativen zum Vollständigkeitsaxiom

Die antiken Griechen haben sich auf rationale Zahlen beschränkt, die aber für ihre Geometrie nicht ausreichten. Eudoxos (4. Jh. 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 an der Nahtstelle der 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 zu (1.5) ist, dass jede Cauchy-Folge (siehe Definition 2.8 auf Seite 235) von reellen Zahlen einen Grenzwert in $\mathbb{R}$ besitzt. Da wir im Buch die Zugehörigkeit des Supremums zu $\mathbb{R}$ verlangen, ergibt sich diese Alternative als Satz 2.8, Seite 235.

Man erhält durch Multiplikation der Elemente einer nicht-leeren, nach unten beschränkten Menge reeller Zahlen mit -1 eine nicht-leere, nach oben beschränkte Menge. Mit dem Vollständigkeitsaxiom ergibt sich damit auch, dass jede nicht-leere, nach unten beschränkte Menge reeller Zahlen ein Infimum besitzt.

Sollte Mandala Bewohner haben (siehe Beispiel 1.34), so existiert das Infimum ihrer Größen.

Lemma 1.3 (Eigenschaft eines Supremums*)

Eine wichtige Eigenschaft des Supremums $s \in \mathbb{R}$ einer nicht-leeren, nach oben beschränkten Menge $E \subset \mathbb{R}$ ist, dass es zu jeder noch so kleinen Zahl $\varepsilon > 0$ ein $x \in E$ gibt mit

$$s - x < \varepsilon. \quad (1.6)$$

Man kann dem Supremum mit Elementen der Menge $E \subset \mathbb{R}$ also beliebig nahekommen. Dabei ist natürlich nicht ausgeschlossen, dass $x = s$ gewählt werden kann, falls $s \in E$.

Der griechische Buchstabe ε (epsilon) wird in der Mathematik immer dann verwendet, wenn es um kleine positive Zahlen geht. Daher lautet der kürzeste Mathematiker-Witz: „ $\varepsilon < 0$ “.

Beweis: Wir beweisen (1.6) indirekt: Würde die Eigenschaft (1.6) nämlich nicht gelten, gäbe es zu einem $\varepsilon > 0$ kein $x \in E$ mit $s - x < \varepsilon$, d. h., für alle $x \in E$ gilt $s - x \geq \varepsilon$, also $x \leq s - \varepsilon$. Dann ist aber auch bereits $s - \varepsilon$ eine obere Schranke – im Widerspruch dazu, dass s die kleinste obere Schranke ist. $\square$

Immer wenn es um grundlegende Aussagen zu Grenzwerten geht, steckt später in letzter Konsequenz das Vollständigkeitsaxiom dahinter. Allerdings benötigt man es nur für Beweise der mathematischen Sätze und nicht für die Anwendung der Mathematik in den Ingenieurwissenschaften. Anschaulicher ist eine Charakterisierung der reellen Zahlen als Dezimalbrüche. Wir haben eingangs gesehen, dass die rationalen Zahlen über endliche oder periodische Dezimalbrüche dargestellt werden können. Der Übergang von $\mathbb{Q}$ zu $\mathbb{R}$ besteht nun genau darin, dass man auch alle unendlichen, nicht-periodischen Dezimalbrüche als Zahlen zulässt. Diese heißen **irrational**, da sie keine rationalen Zahlen sind. $\sqrt{2}$ ist eine irrationale Zahl, die als unendlicher, nicht-periodischer Dezimalbruch geschrieben werden kann: $\sqrt{2} = 1,4142135623730950\dots$

Hat man einen positiven Dezimalbruch mit unendlich vielen Nachkommastellen, so ist darüber eindeutig eine reelle Zahl festgelegt: Zu jedem $n \in \mathbb{N}$ schneiden wir die Zahlendarstellung hinter der n -ten Nachkommastelle ab und bilden aus den sich ergebenden rationalen Zahlen eine Menge M , also beispielsweise $M = \{1, 1,4, 1,41, 1,414, 1,4142, \dots\}$. Diese ist nicht-leer und z. B. nach oben beschränkt durch den Vorkommateil $+1$, also im Beispiel nach oben beschränkt durch die Zahl 2. Die beschriebene reelle Zahl ist dann $\sup M$. Bei einem negativen Dezimalbruch führt das Bilden des Infimums zur zugehörigen reellen Zahl.

Hat man umgekehrt eine reelle Zahl, so kann man aus dieser durch Einschachteln eine Dezimalbruchdarstellung gewinnen. Im Rahmen von Reihen beschäftigen wir uns auf Seite 248 noch einmal genauer mit der Dezimalbruchdarstellung.

1.3.3.3 Intervalle

Wichtige Teilmengen der reellen Zahlen sind die **Intervalle**.

Definition 1.22 (Intervalle)

Es seien $a, b \in \mathbb{R}$ mit $a \leq b$. Dann unterscheiden wir folgende endliche Intervalle:

- $[a, b] := \{x \in \mathbb{R} : a \leq x \leq b\}$ heißt **abgeschlossenes Intervall**.
- $]a, b[:= \{x \in \mathbb{R} : a < x < b\}$ heißt **offenes Intervall**. Üblich ist auch die Schreibweise $(a, b) :=]a, b[$.
- $[a, b[:= \{x \in \mathbb{R} : a \leq x < b\}$ und $]a, b] := \{x \in \mathbb{R} : a < x \leq b\}$ heißen **halboffene Intervalle**. Alternative Schreibweisen sind $[a, b) := [a, b[$ und $(a, b] :=]a, b]$.

Unendliche Intervalle sind von der Bauart

- $[a, \infty[:= \{x \in \mathbb{R} : x \geq a\}$ und $] - \infty, a] := \{x \in \mathbb{R} : x \leq a\}$
- $]a, \infty[:= \{x \in \mathbb{R} : x > a\}$ und $] - \infty, a[:= \{x \in \mathbb{R} : x < a\}$.

Das Symbol ∞ bedeutet „unendlich“, ∞ und $-\infty$ sind **keine** reellen Zahlen, sondern stets im Zusammenhang mit den entsprechenden Intervallen (und später mit Grenzwerten) zu verstehen. d. h., Mengen der Bauart $[-\infty, a[$, $[-\infty, a]$, $]a, \infty[$, $[a, \infty]$ sind nicht definiert, da ∞ oder $-\infty$ zur Menge gehören würde.

Der Begriff „abgeschlossen“ deutet darauf hin, dass Supremum und Infimum jeder nicht-leeren Teilmenge $E \subset [a, b]$ im Intervall enthalten sind.

$$\sup E \in [a, b], \quad \inf E \in [a, b]. \quad (1.7)$$

Eine Abfolge von Intervallen mit reellen Intervallgrenzen nennt man eine **Intervallschachtelung**, wenn jedes Intervall ganz im vorhergehenden Intervall enthalten ist und wenn die Intervalllängen gegen 0 gehen. Aufgrund der Vollständigkeit von $\mathbb{R}$ führt

eine Intervallschachtelung auf ein Element aus $\mathbb{R}$ (siehe Satz 2.6 auf Seite 233, dort wird mittels Intervallschachtelung als Beispiel $\sqrt{2}$ näherungsweise berechnet).

1.3.3.4 Die Zahlen e und π

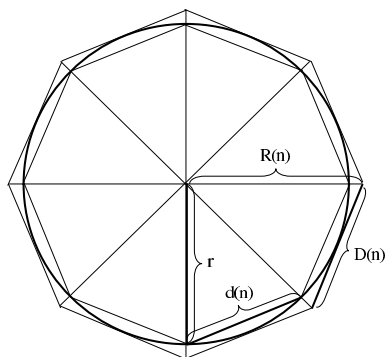


Abb. 1.11: Zur Berechnung des Kreisumfangs

Die beiden wichtigsten irrationalen Zahlen (unendliche, nicht-periodische Dezimalbrüche) sind die Euler'sche Zahl (Leonhard Euler, 1707–1783)

$$e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \dots = 2,7182818\dots, \quad (1.8)$$

die für die Differenzialrechnung wichtig ist (später mehr, siehe Beispiele 2.12 und 2.13 auf Seite 226), und die Kreiszahl $\pi = 3,14159265\dots$, die den Flächeninhalt eines Kreises mit Radius 1 angibt. Für einen Kreis mit Radius r ist der Flächeninhalt πr^2 (vgl. Beispiel 2.6 auf Seite 216).

Beispiel 1.35

Aus dem Flächeninhalt kann man den Kreisumfang $2\pi r$ ableiten (siehe Abbildung 1.11):

- Berechnet man annähernd den Flächeninhalt eines Kreises mit Radius r mittels eines n -Ecks I_n mit Umfang $nd(n)$, das innerhalb des Kreises liegt, so ist

$$\pi r^2 \geq n \frac{d(n) \sqrt{r^2 - \frac{d(n)^2}{4}}}{2}.$$

In der Mathematik nennt man eine annähernde Berechnung eine **Approximation**. Durch geeignet große Wahl von n kann man den Umfang u des Kreises beliebig genau durch $nd(n)$ annähern. Für große n strebt der Term $\sqrt{r^2 - \frac{d(n)^2}{4}}$ gegen r . Damit ist $\pi r^2 \geq u \frac{r}{2}$, also $u \leq 2\pi r$.

- Approximiert man den Flächeninhalt mittels eines n -Ecks A_n mit dem Umfang $nD(n)$, das außerhalb des Kreises liegt, so ist

$$\pi r^2 \leq n \frac{D(n) \sqrt{R(n)^2 - \frac{D(n)^2}{4}}}{2}.$$

Durch geeignet große Wahl von n kann man den Umfang u des Kreises beliebig genau durch $nD(n)$ und den Radius r beliebig genau durch $R(n)$ annähern. Für große n strebt $\sqrt{R(n)^2 - \frac{D(n)^2}{4}}$ gegen r . Damit ist $\pi r^2 \leq u \frac{r}{2}$, also $u \geq 2\pi r$.

Zusammen erhalten wir aus a) und b) den Kreisumfang $u = 2\pi r$. ■

1.3.3.5 Überabzählbarkeit der reellen Zahlen*

Wer den Abschnitt über die Abzählbarkeit der Mengen $\mathbb{N}$, $\mathbb{Z}$ und $\mathbb{Q}$ (siehe Kapitel 1.3.2.3) gelesen hat, sieht hier, dass die Elemente von $\mathbb{R}$ nicht mehr abzählbar sind.

Die Abzählbarkeit von $\mathbb{Q}$ haben wir mit der Cantor-Diagonalisierung bewiesen. Auf Cantor geht ein zweites Diagonalargument zurück, mit dem gezeigt wird, dass bereits die reellen Zahlen zwischen 0 und 1 nicht abzählbar sind. Dazu nehmen wir das Gegenteil an, d. h., es gibt eine Abzählreihenfolge von reellen Zahlen

$$\begin{aligned} x_1 &= 0, \boxed{x_{1,1}} x_{1,2} x_{1,3} \dots \\ x_2 &= 0, x_{2,1} \boxed{x_{2,2}} x_{2,3} \dots \\ x_3 &= 0, x_{3,1} x_{3,2} \boxed{x_{3,3}} \dots \\ &\vdots \\ x_k &= 0, x_{k,1} x_{k,2} x_{k,3} \dots \boxed{x_{k,k}} \dots, \\ &\vdots \end{aligned}$$

wobei $x_{k,i}$ die i -te Nachkommastelle einer Dezimalbruchdarstellung der k -ten reellen Zahl ist, also $x_{k,i} \in \{0, 1, \dots, 9\}$. Damit konstruieren wir eine neue Zahl, die sogenannte Diagonalszahl $y = 0, y_1 y_2 y_3 \dots$, deren i -te Nachkommastelle definiert ist über die Hauptdiagonale der obigen Auflistung der Nachkommastellen als

$$y_i := \begin{cases} 1 & \text{falls } x_{i,i} \neq 1, \\ 2 & \text{falls } x_{i,i} = 1. \end{cases}$$

Für die Konstruktion von y muss man wieder wissen, dass tatsächlich jeder Dezimalbruch mit beliebigen Nachkommastellen eine reelle Zahl ist (siehe Seite 237). Jetzt unterscheidet sich aber $0 < y < 1$ von der Darstellung jeder Zahl x_i , $i \in \mathbb{N}$, mindestens an der i -ten Nachkommastelle. Sie ist damit auch eine von allen x_i verschiedene

Zahl. So selbstverständlich ist das nicht, da z. B. $0,1 = 0,0\overline{9}$ zwei verschiedene Dezimalbruchdarstellungen hat. Diese Mehrdeutigkeiten treten aber nur bei endlich vielen Nachkommastellen und Neunerperioden auf, die bei unserer Konstruktion nicht auftreten. (Man kann zeigen, dass die Dezimalbruchdarstellung eindeutig wird, wenn man nur unendlich viele Nachkommastellen zulässt). Damit haben wir einen Widerspruch dazu, dass $\{x_1, x_2, \dots\}$ bereits alle reellen Zahlen zwischen 0 und 1 umfasst. $\mathbb{R}$ ist nicht abzählbar.

Daher nennt man $\mathbb{R}$ **überabzählbar**. $\mathbb{R}$ ist also deutlich mächtiger als $\mathbb{Q}$. Das wird uns bei Grenzwertbetrachtungen helfen.

In der Vorlesung eines Kollegen gab es an dieser Stelle den Einwand, dass man mit der Diagonalisierung auch die Überabzählbarkeit von $\mathbb{N}$ beweisen könne: Man schreibe die natürlichen Zahlen in einer unendlich langen Liste untereinander und konstruiere eine weitere Zahlendarstellung, die sich an der Einerstelle von der ersten Zahl, an der Zehnerstelle von der zweiten Zahl, usw. unterscheidet. Die entstehende Zahlendarstellung hat dann aber unendlich viele Dezimalstellen. Es kann sich dabei nicht um eine natürliche Zahl handeln, da es zwar unendlich viele natürliche Zahlen gibt, jede einzelne aber nur endlich viele Stellen hat. Schließlich ergibt sie sich durch endlich viele Additionen der Eins. Daran sieht man, dass es durchaus wichtig ist, darüber nachzudenken, dass das konstruierte Gegenbeispiel tatsächlich in der betrachteten Zahlenmenge enthalten ist.

Eine wichtige Fragestellung in der Entwicklung der Mathematik im 20. Jahrhundert war, ob es unendliche Teilmengen von $\mathbb{R}$ gibt, die mehr Elemente als $\mathbb{Q}$, aber weniger Elemente als $\mathbb{R}$ enthalten. Dabei sagt man, dass zwei Mengen genau dann die gleiche Anzahl von Elementen haben, wenn sie durch eine bijektive Abbildung ineinander überführt werden können. Die von Cantor formulierte Kontinuumshypothese besagt, dass jede unendliche Teilmenge von $\mathbb{R}$ entweder mit einer bijektiven Abbildung auf $\mathbb{N}$ oder auf $\mathbb{R}$ überführt werden kann. Cantor ist beim Versuch, diese Hypothese zu beweisen oder zu widerlegen, psychisch und körperlich zu Grunde gegangen.

Tatsächlich haben Kurt Gödel 1938 und Paul Cohen 1963 zusammen gezeigt, dass die Hypothese auf Basis der Axiome der Mengenlehre weder bewiesen noch verworfen werden kann. Nimmt man realistischerweise an, dass die üblichen Axiome der Mengenlehre widerspruchsfrei sind, so gilt:

- Die Axiome der Mengenlehre zusammen mit der Hypothese sind widerspruchsfrei.
- Die Axiome der Mengenlehre zusammen mit der negierten Hypothese sind ebenfalls widerspruchsfrei.

Das bedeutet: Man kann weder beweisen, dass die Kontinuumshypothese gilt, noch kann man das Gegenteil zeigen. Das ist vergleichbar mit der Unentscheidbarkeit des Halteproblems, die in Beispiel 1.19 auf Seite 24 gezeigt wird. Die Mathematik hat hier ihre Grenzen.

1.4 Rechnen mit reellen Zahlen

Beliebte Variablennamen für reelle Zahlen sind x, y, u, v, s, t . Da sehr schnell die für Variablen zur Verfügung stehenden Buchstaben erschöpft sind, unterscheidet man Variablen oft durch einen **Index**. Dabei schreibt man eine Zusatzinformation etwas kleiner unten rechts an die Variable (und nicht rechts daneben, da dies wie eine Multiplikation aussieht). So haben wir bereits x_1, x_2 und x_3 benutzt. Der Index kann auch durch eine andere Variable bestimmt werden: x_k . Mit $k = 2$ ist dann $x_k = x_2$. Auch das reicht noch nicht aus, so dass z. B. für Winkelangaben zusätzlich griechische Buchstaben zum Einsatz kommen (siehe Tabelle 1.5). Abstände werden häufig mit δ und sehr kleine positive Zahlen mit ε bezeichnet.

Tab. 1.5: Ausgewählte griechische Buchstaben

α	alpha	β	beta	γ	gamma	δ	delta	ε	epsilon
η	eta	ϑ	theta	κ	kappa	λ	lambda	μ	mü
ξ	xi	π	pi	ρ	rho	σ	sigma	τ	tau
φ	phi	χ	chi	ψ	psi	ω	omega	Δ	Delta
Γ	Gamma	Π	Pi	Σ	Sigma	Φ	Phi	Ω	Omega

1.4.1 Potenzen und Wurzeln

Für die reellen Zahlen können wir nun die Rechenregeln für Potenzen, die wir für Brüche kennengelernt haben und die auch für reelle Zahlen gelten, um die Wurzeln erweitern.

Eine zweite Wurzel $\sqrt{a}$ als nicht-negative Lösung der Gleichung $x^2 = a$, $a \geq 0$ benötigt man auch zur Lösung allgemeiner quadratischer Gleichungen der Form $x^2 + px + q = 0$.

Satz 1.10 (p-q-Formel)

Die Gleichung $x^2 + px + q = 0$ hat genau die reellen Lösungen $x_1 = -\frac{p}{2} + \sqrt{\left(\frac{p}{2}\right)^2 - q}$ und $x_2 = -\frac{p}{2} - \sqrt{\left(\frac{p}{2}\right)^2 - q}$, sofern die **Diskriminante** $D := \left(\frac{p}{2}\right)^2 - q$ nicht-negativ ist. Ist hingegen $D < 0$, so existiert keine reelle Lösung.

Beweis: Wir beweisen die Formel mittels quadratischer Ergänzung:

$$x^2 + px + q = x^2 + px + \left(\frac{p}{2}\right)^2 - \left(\frac{p}{2}\right)^2 + q = \left(x + \frac{p}{2}\right)^2 - \left(\frac{p}{2}\right)^2 + q.$$

Zu $x^2 + px + q = 0$ ist damit $\left(x + \frac{p}{2}\right)^2 = \left(\frac{p}{2}\right)^2 - q$ äquivalent. Die p-q-Formel ergibt sich nun durch Ziehen der Wurzel, sofern die Diskriminante dies zulässt. □

Achtung: Ein beliebter Fehler im Umgang mit der p-q-Formel ist das Weglassen des Minuszeichens von $-\frac{p}{2}$.

Die Diskriminante bestimmt die Anzahl der Lösungen. Bei $D = 0$ gibt es nur eine, bei $D > 0$ gibt es zwei. Bei mehreren Lösungen verwendet man Indizes: $x^2 = 1$ führt auf $x_1 = 1$ und $x_2 = -1$. Als abkürzende Schreibweise ist dafür auch $x_{1,2} = \pm 1$ gebräuchlich.

Beispiel 1.36

Zwei Strecken mit Längen a und b , $0 < a < b$ stehen im Verhältnis des goldenen Schnitts, wenn das Verhältnis der größeren b zur kleineren a dem Verhältnis der Gesamtlänge beider Strecken zur größeren entspricht, also

$$\frac{b}{a} = \frac{a+b}{b} \iff \frac{b}{a} = 1 + \frac{a}{b} = 1 + \frac{1}{\frac{b}{a}} \iff \left(\frac{b}{a}\right)^2 - \frac{b}{a} - 1 = 0.$$

Die p-q-Formel liefert nun die beiden Lösungen $\frac{b}{a} = \frac{1 \pm \sqrt{5}}{2}$. Da die Streckenlängen a und b positiv sind, muss $\frac{b}{a} = \frac{1 + \sqrt{5}}{2}$ sein. Die Zahl heißt der **goldene Schnitt**. Dieses Streckenverhältnis wird als schön empfunden und in der Architektur und Kunst verwendet. Außerdem findet man den goldenen Schnitt an vielen Stellen in der Natur (siehe Abbildung 1.12) und auch in der Mathematik. Er wird uns bei Zahlenfolgen (siehe Seite 231) und Eigenwerten (siehe Seite 551) wiederbegegnen. ■

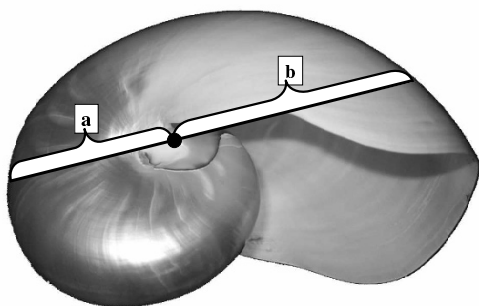


Abb. 1.12: Der goldene Schnitt $\frac{b}{a} = \frac{a+b}{b} = \frac{1+\sqrt{5}}{2}$

Analog zur zweiten Wurzel von $x \geq 0$ existieren auch n -te Wurzeln:

Definition 1.23 (Reelle Wurzeln)

Die n -te Wurzel ($n \in \mathbb{N}$) von $x \in \mathbb{R}$ mit $x \geq 0$ ist die eindeutige Zahl $y \in \mathbb{R}$, $y \geq 0$, mit $y^n = x$. Schreibweise: $y = \sqrt[n]{x} = x^{1/n}$.

Bemerkungen:

- Wegen $(-1)^n = \begin{cases} 1, & \text{falls } n \text{ gerade} \\ -1, & \text{falls } n \text{ ungerade ist,} \end{cases}$ gilt:

- Ist $x \geq 0$, so gilt für gerades n mit $y^n = x$ auch $(-y)^n = x$. Die Gleichung $y^n = x$ hat dann die Lösungen $\pm \sqrt[n]{x}$.
- Ist x negativ und n ungerade, so ist

$$y^n = x \iff (-1) \cdot y^n = -x \iff (-y)^n = -x,$$

so dass $-y = \sqrt[n]{-x}$ bzw. $y = -\sqrt[n]{-x}$. Bei ungeradem n benutzt man daher auch für negatives x die Schreibweise $\sqrt[n]{x} := -\sqrt[n]{-x}$.

Beispiel 1.37

- i) $\sqrt[3]{-27} = -3$. Bei geradem n und negativem x hat die Gleichung $y^n = x$ aber keine Lösung, so dass dafür $\sqrt[n]{x}$ nicht definiert ist.
- ii) $\sqrt[4]{-16}$ existiert nicht als reelle Zahl (aber als komplexe, siehe Kapitel 1.6). ■

- Die Schreibweise $x^{1/n}$ passt zu den Rechenregeln für Potenzen, denn

$$\left[x^{1/n}\right]^n = x^{\frac{1}{n} \cdot n} = x.$$

- Wir können nun auch mit rationalen Exponenten p/q , $p \in \mathbb{Z}, q \in \mathbb{N}$, rechnen:

$$x^{p/q} = [x^p]^{1/q} = \sqrt[q]{x^p}.$$

Exponenten, die bislang nur für Brüche definiert sind, können sogar beliebige reelle Zahlen sein. Dabei nutzt man aus, dass die rationalen Zahlen in $\mathbb{R}$ **dicht** liegen. Das bedeutet, dass man (z. B. durch Abschneiden von Nachkommastellen) zu einer Zahl $y \in \mathbb{R}$ rationale Zahlen $y_1, y_2, y_3, \dots \in \mathbb{Q}$ finden kann, die „gegen“ y streben, d. h., deren Abstand zu y immer kleiner wird und sich null annähert (eine exakte Formulierung holen wir nach, wenn wir Folgenkonvergenz eingeführt haben, siehe Satz 2.4 auf Seite 229). Die Zahlen $x^{y_1}, x^{y_2}, x^{y_3}, \dots$ streben gegen eine Zahl, die wir x^y nennen.

Die Rechenregeln für Potenzen gelten unverändert auch für reelle Exponenten.

1.4.2 Summen und Produkte, Binomischer Lehrsatz

1.4.2.1 Summenzeichen und Produktzeichen

Das Summen-Symbol $\sum$ (großes Sigma) wird verwendet, um kurz eine Summe mit vielen gleichartigen Summanden hinzuschreiben.

$$\sum_{k=1}^n a_k := a_1 + a_2 + \dots + a_n.$$

Dabei verwendet man eine Laufvariable (Index), die die ganzen Zahlen von einem Startwert (hier 1) bis zu einem Zielwert (hier n) durchläuft. Wenn die Laufvariable k von m bis n laufen soll, dann schreibt man $\sum_{k=m}^n$. Unter dem **Summenzeichen** steht also üblicherweise der Startwert für die Laufvariable (hier $k = m$) und darüber der Zielwert (hier n). Für jeden Wert der Laufvariable wird ein Summand berechnet und zur Summe addiert. Die Summanden dürfen damit von der Laufvariablen abhängen. In Kapitel 2.2 werden wir über die hier vorgestellten endlichen Summen hinausgehen und unendliche Summen (Reihen) einführen. Diese werden helfen, Funktionen als Summen einfacherer Funktionen darzustellen, mit denen man besser rechnen kann.

Beispiel 1.38

- a) $\sum_{k=1}^5 k^2 = 1 + 4 + 9 + 16 + 25$.
 b) $\sum_{k=-1}^{1001} k = 1\,001 + 1\,000 - 1\,000 + 999 - 999 + \dots + 1 - 1 + 0 = 1\,001$.
 c) $\sum_{k=-n}^n a = (2n+1)a$, wobei a eine Konstante ist. Man beachte, dass es hier $2n+1$ Summanden gibt, da auch der Index 0 mitgezählt werden muss. ■

Eine Summe kann in einer Programmiersprache mit einer Zählschleife berechnet werden. Die Berechnung der Summe $s := \sum_{k=m}^n a_k$ der Zahlen a_k mit einem C-Programm sieht z. B. so aus:

```
s=0;
for (k=m; k<=n; k=k+1)
    s=s+a[k];
```

Der Name der Zählvariablen (hier k) ist für den Wert der Summe irrelevant, z. B.

$$\sum_{k=0}^n a_k = \sum_{i=0}^n a_i = \sum_{l=1}^{n+1} a_{l-1} = \sum_{m=5}^{n+5} a_{m-5}.$$

Oft ist es nützlich, die Zählvariable zu transformieren.

Beispiel 1.39

- a) $\sum_{k=1}^n k^2 - \sum_{k=3}^{n+1} (k-1)^2 = \sum_{k=1}^n k^2 - \sum_{k=2}^n k^2 = 1^2 = 1$. Hier haben wir eine Index-Transformation bei der zweiten Summe vorgenommen. Dabei ändert sich der Summationsbereich, ohne dass sich der Wert der Summe ändert. Hier startet k nun bei 2 statt wie zuvor bei 3. Daher muss in den Summanden zu k der Wert 1 addiert werden, damit sie den gleichen Wert behalten. Die Index-Transformation ermöglicht häufig die Zusammenfassung von zwei oder mehr Summen zu einer.
 b) Gegeben sei die Summe $S = \sum_{k=0}^n a_k$. Wir kehren die Summationsreihenfolge um, indem wir $l = n - k$ bzw. $k = n - l$ setzen. Für $k = 0$ ist $l = n$, und für $k = n$ ist $l = 0$. Wir erhalten $S = \sum_{k=0}^n a_k = \sum_{l=0}^n a_{n-l}$. ■

Satz 1.11 (Rechenregeln für Summen)

Für Zahlen $a_1, a_2, \dots, a_n \in \mathbb{R}$, $b_1, b_2, \dots, b_n \in \mathbb{R}$ und $c \in \mathbb{R}$ gilt:

$$\text{a) } \sum_{k=1}^n (a_k + b_k) = \sum_{k=1}^n a_k + \sum_{k=1}^n b_k, \quad \text{b) } \sum_{k=1}^n c \cdot a_k = c \cdot \sum_{k=1}^n a_k.$$

Beweis: Die Regeln folgen direkt aus der Definition der Summe. Beispielsweise erhält man den Beweis zu a) so:

$$\begin{aligned} \sum_{k=1}^n (a_k + b_k) &= (a_1 + b_1) + (a_2 + b_2) + \dots + (a_n + b_n) \\ &= a_1 + \dots + a_n + b_1 + \dots + b_n = \sum_{k=1}^n a_k + \sum_{k=1}^n b_k \quad \square \end{aligned}$$

Formeln werden übersichtlicher, wenn wir auch leere Summen zulassen und ihren Wert als null definieren. Für $m > n$ setzen wir also $\sum_{k=m}^n a_k = 0$.

Beispiel 1.40

$$\begin{aligned} \text{a) } \sum_{k=1}^n \frac{1}{k(k+1)} &= \sum_{k=1}^n \left[\frac{k+1}{k(k+1)} - \frac{k}{k(k+1)} \right] = \sum_{k=1}^n \left(\frac{1}{k} - \frac{1}{k+1} \right) \\ &= \sum_{k=1}^n \frac{1}{k} - \sum_{k=2}^{n+1} \frac{1}{k} = 1 - \frac{1}{n+1}. \end{aligned} \quad (1.9)$$

Hier spricht man von einer **Teleskopsumme**, da sich aufeinander folgende Summanden paarweise aufheben. Es bleiben nur der erste und letzte Summand übrig. Man zieht bildlich diese Summe aus zwei Summanden durch Einfügen „konstruktiver Nullen“ wie ein Teleskop auseinander. Das gilt auch für die nächste Summe:

$$\text{b) } \sum_{k=1}^5 (-1)^k = -1 + 1 - 1 + 1 - 1 = -1. \quad \blacksquare$$

Beispiel 1.41 (Gauß'sche Summenformel)

Für alle $n \in \mathbb{N}$ gilt

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

Diese Formel hat der Legende nach der wohl bedeutendste deutsche Mathematiker Carl Friedrich Gauß (1777–1855) im Schulunterricht gefunden, als er die Zahlen von 1 bis $n := 100$ addieren sollte. Er bemerkte, dass $1 + 100 = 101$, $2 + 99 = 101$, $\dots$. Jede der $\frac{n}{2} = 50$ Additionen liefert also den Wert $n + 1 = 101$, und die Summe ist 5 050. Wir zeigen nun die Formel für allgemeines n mittels Vollständiger Induktion:

■ Induktionsanfang: $\sum_{k=1}^1 k = 1 = \frac{1 \cdot 2}{2}$ ist wahr.

- Induktionsannahme: Für ein $n \in \mathbb{N}$ gelte $\sum_{k=1}^n k = \frac{n(n+1)}{2}$.
- Induktionsschluss: Mit der Induktionsannahme ist

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

zu zeigen. Dazu spalten wir den letzten Summanden (für $k = n+1$) ab. Die Induktionsannahme kann dann für die Darstellung der verbleibenden Summe genutzt werden. Das ist ganz typisch für den Beweis von Summenformeln:

$$\begin{aligned} \sum_{k=1}^{n+1} k &= \left[\sum_{k=1}^n k \right] + (n+1) \stackrel{\text{Induktionsannahme}}{=} \frac{n(n+1)}{2} + (n+1) \\ &= (n+1) \cdot \left(\frac{n}{2} + 1 \right) = \frac{(n+1)(n+2)}{2}, \end{aligned}$$

d. h., die Aussage ist auch für $n+1$ wahr. Damit ist die Identität mittels Induktion für alle $n \in \mathbb{N}$ gezeigt, und wir können sie künftig benutzen. Auch der Name Gauß wird uns noch in verschiedenen anderen Zusammenhängen wiederbegegnen. ■

Analog zum Summen-Symbol gibt es eine Kurzschreibweise für das Produkt vieler gleichartiger Faktoren, bei der ein großes Pi ($\prod$) als **Produktzeichen** verwendet wird:

$$\prod_{k=1}^n a_k := a_1 \cdot a_2 \cdots a_n.$$

Beispiel 1.42

$$\text{a) } \prod_{k=1}^3 k = 1 \cdot 2 \cdot 3, \quad \text{b) } \prod_{k=-n}^m a = a^{n+m+1},$$

$$\text{c) } \prod_{k=1}^n \frac{k}{k+1} = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{n}{n+1} = \frac{1}{n+1} \quad (\text{Teleskopprodukt}). \quad \blacksquare$$

1.4.2.2 Geometrische Summenformel und Anwendungen

Mittels Vollständiger Induktion können wir nun eine Formel beweisen, auf der die Finanzmathematik beruht und die wir zudem später im Umgang mit Reihen benötigen.

Lemma 1.4 (Geometrische Summe)

Sei q eine reelle Zahl mit $0 \neq q \neq 1$. Für alle $n \in \mathbb{N}_0$ gilt:

$$\sum_{k=0}^n q^k = q^0 + q^1 + q^2 + q^3 + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q}. \quad (1.10)$$

Für die konkreten Zahlenwerte $q = \frac{1}{2}$ (vgl. Abbildung 2.8 auf Seite 238) und $n = 3$ verifizieren wir die Formel (1.10). Einsetzen liefert für die linke bzw. die rechte Seite in (1.10):

$$\sum_{k=0}^3 \left(\frac{1}{2}\right)^k = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{15}{8} \text{ bzw. } \frac{1 - \frac{1}{2^{3+1}}}{1 - \frac{1}{2}} = \frac{\frac{15}{16}}{\frac{1}{2}} = \frac{15}{8}.$$

Wir beweisen nun auch (1.10) mittels Vollständiger Induktion nach n .

Beweis: ■ Induktionsanfang für $n = 0$:

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

■ Induktionsannahme: Für ein $n \in \mathbb{N}$ gelte $\sum_{k=0}^n q^k = \frac{1-q^{n+1}}{1-q}$.

■ Induktionsschluss: Zu zeigen ist $\sum_{k=0}^{n+1} q^k = \frac{1-q^{n+2}}{1-q}$.

$$\begin{aligned} \sum_{k=0}^{n+1} q^k &= q^{n+1} + \sum_{k=0}^n q^k \stackrel{\text{Induktionsannahme}}{=} q^{n+1} + \frac{1-q^{n+1}}{1-q} \\ &= \frac{(1-q)q^{n+1} + 1 - q^{n+1}}{1-q} = \frac{q^{n+1} - q^{n+2} + 1 - q^{n+1}}{1-q} = \frac{1 - q^{n+2}}{1-q}. \quad \square \end{aligned}$$

Anwendungen der geometrischen Summenformel findet man in vielfältiger Art im Bereich der Finanz- und Versicherungsmathematik. Wir betrachten Beispiele zur Aufzinsung und zur Rentenrechnung.

Beispiel 1.43

Eine Geldsumme K_0 wird für n Jahre mit einer jährlichen Verzinsung von p % angelegt. Statt Jahre können im Folgenden auch beliebige andere Zeitspannen gewählt werden. Wir berechnen das Gesamtkapital K_n nach Ende des Anlagezeitraums. Für ein Anfangskapital von 5 000 Euro und einen jährlichen Zinssatz von 4,5 % interessiert uns insbesondere der Endwert des Kapitals nach 15 Jahren.

Am Ende des ersten Jahres erhält man zu K_0 einen Zinsgewinn von $K_0 \cdot \frac{p}{100}$, so dass zu Beginn des zweiten Jahres ein Kapital $K_1 = K_0 + \frac{p}{100} \cdot K_0 = K_0 \left(1 + \frac{p}{100}\right)$ als neues Anfangskapital vorliegt. Mit dem Zinsgewinn $K_1 \cdot \frac{p}{100}$ beginnt dann das dritte Jahr mit dem Kapital $K_2 = K_1 \left(1 + \frac{p}{100}\right) = K_0 \left(1 + \frac{p}{100}\right)^2$. Schließlich erhält man am Ende des n -ten Jahres den Geldbetrag

$$K_n = K_0 \left(1 + \frac{p}{100}\right)^n, \quad n = 0, 1, 2, \dots$$

Diese Gleichung heißt in der Finanzmathematik die **Leibniz'sche Zinseszinsformel**. Im konkreten Fall der Verzinsung eines Anfangskapitals $K_0 = 5\,000$ Euro mit einem jährlichen Zins p von 4,5 % erhält man am Ende von 15 Jahresperioden ein Endkapital $K_{15} = 5\,000 \cdot (1 + 0,045)^{15} = 9676,41$ Euro. ■

In diesem Beispiel haben wir noch keine Summen benötigt. Das ändert sich aber bei Ein- oder Auszahlungen.

Beispiel 1.44

Wird im Rahmen eines Sparplans regelmäßig zu Beginn oder alternativ zum Ende einer Zinsperiode eine feste Einzahlung E geleistet, so spricht man von einer vorschüssigen bzw. nachschüssigen Rentenzahlung. Der Name deutet an, dass es sich statt einer Einzahlung auch um eine Auszahlung (negative Einzahlung) handeln kann. Der Gesamtbetrag inklusive der Zahlungen und Zinsen wird als Rentenendwert bezeichnet. Wir berechnen den vorschüssigen Rentenendwert R_n nach n Zinsperioden bei einem Zinssatz von $p > 0$ Prozent. Im konkreten Fall sollen zu Beginn jedes Monats 100 Euro eingezahlt werden, die am Monatsende verzinst werden. Der monatliche Zinssatz betrage 0,5 %, und wir berechnen den vorschüssigen Rentenendwert nach 5 Jahren.

Die erste Einzahlung $K_0 = E$ ist das zu verzinsende Anfangskapital für die erste Periode, so dass vor der zweiten Einzahlung dieses Kapital auf $R_1 = K_0 \left(1 + \frac{p}{100}\right) = E \left(1 + \frac{p}{100}\right)$ angewachsen ist. Zusammen mit der zweiten Einzahlung E wird in der zweiten Periode ein Kapital $K_1 = E \left(1 + \frac{p}{100}\right) + E$ verzinst. Das ergibt am Ende der zweiten Periode einen Betrag von $R_2 = K_1 \left(1 + \frac{p}{100}\right) = E \left(1 + \frac{p}{100}\right)^2 + E \left(1 + \frac{p}{100}\right)$. Fortgesetzte Rechnung ergibt am Ende der n -ten Periode den Rentenendwert

$$R_n = E \left[\left(1 + \frac{p}{100}\right)^n + \left(1 + \frac{p}{100}\right)^{n-1} + \cdots + \left(1 + \frac{p}{100}\right) \right] = E \sum_{k=1}^n \left(1 + \frac{p}{100}\right)^k.$$

Setzt man $q = 1 + \frac{p}{100}$, so ist $q \neq 1$, und mit der Formel (1.10) folgt

$$\begin{aligned} R_n &= E \sum_{k=0}^{n-1} \left(1 + \frac{p}{100}\right)^{k+1} - E = E \cdot \frac{1 - q^{n+1}}{1 - q} - E \\ &= E \cdot \frac{(1 - q^{n+1}) - (1 - q)}{1 - q} = E \cdot \frac{q - q^{n+1}}{1 - q} = Eq \cdot \frac{q^n - 1}{q - 1}. \end{aligned}$$

Bei einem monatlichen Zinssatz $\frac{p}{100} = 0,005$, d. h. $q = 1 + \frac{p}{100} = 1,005$ erfolgen die Einzahlungen über $n = 60$ Monate. Damit ist bei Einzahlungen von $E = 100$ Euro nach 60 Monaten ein Rentenendwert angespart von

$$R_{60} = 100 \cdot 1,005 \cdot \frac{1,005^{60} - 1}{1,005 - 1} = 7011,89 \text{ Euro.}$$

■

Beispiel 1.45

Die Berechnung des Rentenendwerts aus dem vorangehenden Beispiel ist auch für Kredite über x Euro wichtig. Häufig vereinbart man für einen festen Zeitraum (z. B. 10 Jahre) einen festen Zinssatz p %, $p > 0$. Zum Zeitpunkt der Drucklegung war ca. $p = 3,5$ % realisierbar. Man muss dann in jeder Zinsperiode $x \cdot \frac{p}{100} + T$ Euro bezahlen, wobei T ein fester Tilgungsbetrag ist. Bei einer von den Banken angebotenen minimalen

Tilgung von 1 % ist $T = 0,01 \cdot x$. Die Höhe der Zahlung orientiert sich also stets an der ursprünglichen Kredithöhe. Tatsächlich reduzieren sich aber durch die Tilgung mit der Zeit die Schulden. Die Bank verlangt als Gebühr nur p % der tatsächlichen Restschuld. p % der Differenz der Restschuld zur ursprünglichen Kredithöhe gehen damit zusätzlich in die Tilgung. Die Höhe der ersten Tilgung ist T , die der zweiten ist $T + \frac{p}{100} \cdot T$ usw. Die Summe der Tilgungen ist identisch mit dem zuvor betrachteten Rentenendwert und beträgt nach n Zahlungsperioden

$$T \sum_{k=0}^{n-1} \left[1 + \frac{p}{100} \right]^k = T \frac{\left[1 + \frac{p}{100} \right]^n - 1}{\frac{p}{100}}.$$

Die Restschuld nach n Zahlungsperioden ist die Differenz aus x und diesem Wert. Die Höhe der festen Tilgung T ergibt sich damit im Allgemeinen aus einer geplanten Kreditlaufzeit von N Zahlungsperioden zu

$$T = \frac{x \cdot \frac{p}{100}}{\left[1 + \frac{p}{100} \right]^N - 1}.$$

Je höher der Zins p ist, desto niedriger kann die Tilgung gewählt werden, da die Zinsen für die Differenz zwischen Kredithöhe und Restschuld in die Tilgung fließen. ■

1.4.2.3 Binomischer Lehrsatz

Die Binomische Formel (vgl. Abbildung 1.13)

$$(a + b)^2 = a^2 + 2ab + b^2$$

lässt sich mit Binomialkoeffizienten etwas komplizierter schreiben:

$$(a+b)^2 = a^0 b^2 + 2a^1 b^1 + a^2 b^0 = \binom{2}{0} a^0 b^2 + \binom{2}{1} a^1 b^1 + \binom{2}{2} a^2 b^0 = \sum_{k=0}^2 \binom{2}{k} a^k b^{2-k}.$$

Allerdings ermöglicht die Schreibweise, die Formel allgemeiner zu fassen.

a^2	$a \cdot b$
$a \cdot b$	b^2

$\left. \begin{array}{l} \left. \begin{array}{l} \left. \left. \begin{array}{l} a^2 \\ a \cdot b \end{array} \right\} a \end{array} \right\} a+b \\ \left. \begin{array}{l} a \cdot b \\ b^2 \end{array} \right\} b \end{array} \right\}$

Abb. 1.13: Interpretation der Binomischen Formel $(a+b)^2 = a^2 + 2ab + b^2$ als Summe von vier Teilflächen eines Quadrats mit Kantenlänge $a + b$ und Flächeninhalt $(a + b)^2$

Satz 1.12 (Binomischer Satz)Für $n \in \mathbb{N}$ gilt:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}. \quad (1.11)$$

Für größere Werte von n rechnet man mit dieser Formel viel schneller als bei sukzessivem Ausmultiplizieren.

Beweis: ■ Induktionsanfang für $n = 1$:

$$\sum_{k=0}^1 \binom{1}{k} a^k b^{1-k} = \binom{1}{0} b + \binom{1}{1} a = a + b.$$

- Induktionsannahme: Für ein $n \in \mathbb{N}$ gelte $(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$.
- Induktionsschluss: Zu zeigen ist $(a + b)^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} a^k b^{n+1-k}$.

$$\begin{aligned}
 (a + b)^{n+1} &= (a + b)(a + b)^n = (a + b) \sum_{k=0}^n \binom{n}{k} a^k b^{n-k} \\
 &= \sum_{k=0}^n \binom{n}{k} a^{k+1} b^{n-k} + \sum_{k=0}^n \binom{n}{k} a^k b^{n+1-k} \\
 &= \sum_{k=1}^{n+1} \binom{n}{k-1} a^k b^{n+1-k} + \sum_{k=0}^n \binom{n}{k} a^k b^{n+1-k} \\
 &= \binom{n}{n} a^{n+1} + \sum_{k=1}^n \left[\binom{n}{k-1} + \binom{n}{k} \right] a^k b^{n+1-k} + \binom{n}{0} b^{n+1} \\
 &\stackrel{(1.2)}{=} \binom{n}{n} a^{n+1} + \sum_{k=1}^n \binom{n+1}{k} a^k b^{n+1-k} + \binom{n}{0} b^{n+1} \\
 &= \binom{n+1}{n+1} a^{n+1} b^0 + \sum_{k=1}^n \binom{n+1}{k} a^k b^{n+1-k} + \binom{n+1}{0} a^0 b^{n+1} \\
 &= \sum_{k=0}^{n+1} \binom{n+1}{k} a^k b^{n+1-k}.
 \end{aligned}$$

□

Neben $(a + b)^2 = a^2 + 2ab + b^2$ wird auch der Spezialfall für $-b$ und $n = 2$ als **Binomische Formel** bezeichnet:

$$(a - b)^2 = a^2 - 2ab + b^2.$$

Setzen wir im Binomischen Satz $n = 3$, so erhalten wir

$$\begin{aligned}
 (a+b)^3 &= \binom{3}{0} \cdot a^3 + \binom{3}{1} \cdot a^2b + \binom{3}{2} \cdot ab^2 + \binom{3}{3} \cdot b^3 \\
 &= a^3 + 3a^2b + 3ab^2 + b^3.
 \end{aligned}$$

Der Binomische Satz wird häufig nicht oder falsch angewendet: $625 = (2+3)^4 \neq 2^4 + 3^4 = 97$. Statt die Potenzen einzeln auszurechnen, muss man $(a+b)^4 = 1 \cdot a^4 + 4 \cdot a^3b + 6 \cdot a^2b^2 + 4 \cdot a \cdot b^3 + 1 \cdot b^4$ benutzen.

Mit dem Binomischen Satz ist

$$2^n = (1+1)^n = \sum_{k=0}^n \binom{n}{k} 1^k \cdot 1^{n-k} = \sum_{k=0}^n \binom{n}{k}.$$

Interpretation: $\binom{n}{k}$ ist die Anzahl der k -elementigen Teilmengen einer Menge mit n Elementen. $\sum_{k=0}^n \binom{n}{k}$ ist daher die Gesamtanzahl aller Teilmengen einer Menge mit n Elementen. Sie ist 2^n . Das hatten wir uns auch schon in Kapitel 1.1 anhand eines Entscheidungsbaums überlegt (siehe Abbildung 1.1).

Schließlich wird die folgende Regel als dritte Binomische Formel bezeichnet:

$$(a-b)(a+b) = a^2 + ab - ba - b^2 = a^2 - b^2.$$

1.4.3 Beträge und Ungleichungen

Um den Abstand zweier reeller Zahlen anzugeben, benötigen wir den **Betrag**. Vielfach sind wir aber gar nicht am exakten Abstand interessiert und möchten nur eine mehr oder weniger grobe Abschätzung machen. Dazu rechnen wir dann mit mit **Ungleichungen**. Das ist z. B. typisch für Grenzwertuntersuchungen in Kapitel 2.

1.4.3.1 Beträge

Definition 1.24 (Betrag)

Mit $|\cdot|$ bezeichnet man den **Absolut-Betrag** (oder kurz: *Betrag*) einer reellen Zahl:

$$|x| := \begin{cases} x, & \text{für } x \geq 0, \\ -x, & \text{für } x < 0. \end{cases} \quad (1.12)$$

Der Betrag ist also nicht-negativ, ein negatives Vorzeichen wird durch den Betrag entfernt. $|x|$ entspricht dem Abstand von x zum Nullpunkt auf der Zahlengeraden: $|-4| = 4$, $|3| = 3$. $|x - y|$ ist der Abstand zweier reeller Zahlen x und y .

Satz 1.13 (Rechenregeln für den Betrag)

Es seien $a, b, c \in \mathbb{R}$. Dann gilt

- $|x| \leq c$ ist gleichbedeutend mit $-c \leq x \leq c$.
- Der Betrag eines Produkts ist gleich dem Produkt der Beträge:

$$|a \cdot b| = |a| \cdot |b|, \quad \left| \frac{a}{b} \right| = \frac{|a|}{|b|}, \text{ falls } b \neq 0.$$

Da der Betrag über eine Fallunterscheidung definiert ist, führt die Auflösung eines Betrags automatisch zur Untersuchung von Fällen.

Beispiel 1.46

a) Für $x \neq -4$ ist

$$\left| \frac{x-3}{x+4} \right| = \begin{cases} \frac{x-3}{x+4}, & \text{für } x < -4, \\ -\frac{x-3}{x+4}, & \text{für } -4 < x < 3, \\ \frac{x-3}{x+4}, & \text{für } x \geq 3. \end{cases}$$

Hier wurde die Variable x in (1.12) durch $\frac{x-3}{x+4}$ ersetzt und überprüft, wo dieser Bruch negativ und wo positiv oder 0 ist.

b) $|x-2| \leq 3 \iff -3 \leq x-2 \leq 3 \iff -3+2 \leq x \leq 3+2 \iff -1 \leq x \leq 5$. ■

1.4.3.2 Ungleichungen

Ungleichungen mit Zahlen können äquivalent umgeformt werden, indem zu beiden Seiten ein gleicher Wert addiert wird oder beide Seiten mit einer positiven Zahl multipliziert werden. Die folgenden Ungleichungen sind für die gleichen Zahlen erfüllt und damit äquivalent:

$$5x + 1 > 2x \iff 3x + 1 > 0 \iff 3x > -1 \iff x > -\frac{1}{3}.$$

Achtung: Ein Fehler wird jedoch häufig bei der Multiplikation einer Ungleichung mit einer negativen Zahl gemacht: Hier wird aus „>“ nämlich „<“ und aus „≥“ wird „≤“ und umgekehrt. Dies wird schnell vergessen! Richtig ist: $-2 < 1 \implies 2 > -1$ und

$$5x + 1 > 2x \iff 1 > -3x \iff -\frac{1}{3} < x.$$

Ungleichungen treten bei praktischen Problemstellungen im Alltag häufig auf, wie das folgende Beispiel zeigt.

Beispiel 1.47

Der Preis für den bezogenen elektrischen Strom besteht aus einer monatlichen Grundgebühr und einem variablen Verbrauchsanteil. Die Stadtwerke bieten folgende (monatliche) Tarife an:

- Tarif A: Grundgebühr 6 Euro, Preis pro 1 kWh: 0,25 Euro,
- Tarif B: Grundgebühr 8 Euro, Preis pro 1 kWh: 0,21 Euro.

Für welche monatlichen Verbrauchsmengen ist Tarif A günstiger als Tarif B?

Zur Lösung bezeichne x den Verbrauch in kWh pro Monat. Dann erhalten wir die Kostenfunktionen

$$t_A(x) = 6 + 0,25 \cdot x, \quad t_B(x) = 8 + 0,21 \cdot x.$$

Für einen Verbrauch x ist Tarif A günstiger als B genau dann wenn $t_A(x) < t_B(x)$:

$$\begin{aligned} 6 + 0,25 \cdot x &< 8 + 0,21 \cdot x \\ \iff 0,04 \cdot x &< 2 \\ \iff x &< 50. \end{aligned}$$

Da der Stromverbrauch nie negativ werden kann, lautet die Lösungsmenge $\{x \in \mathbb{R} : 0 \leq x < 50\}$. Bei einem Monatsverbrauch bis 50 kWh ist Tarif A günstiger als B. ■

Satz 1.14 (Rechenregeln für Ungleichungen)

Für Zahlen a, b, c gilt (vgl. Definition 1.21 auf Seite 48):

- a) Aus $a < b$ und $b < c$ folgt $a < c$ (Transitivität der Ordnungsrelation).
 b) Die Addition einer Zahl auf beiden Seiten ändert eine Ungleichung nicht:

$$a < b \implies a + c < b + c.$$

- c) Die Multiplikation mit einer positiven Zahl ändert die Ungleichung nicht:

$$a < b \wedge c > 0 \implies a \cdot c < b \cdot c.$$

- d) **Achtung:** Bei einer Multiplikation mit einer negativen Zahl c kehrt sich die Ungleichung um, z. B. wird aus $a < b$ die Ungleichung $a \cdot c > b \cdot c$.
 e) Ist das Produkt zweier reeller Zahlen positiv, dann sind entweder beide Zahlen positiv, oder beide Zahlen sind negativ.
 f) Ist das Produkt zweier reeller Zahlen negativ, dann haben sie ein unterschiedliches Vorzeichen.
 g) **Achtung:** Sind $a > 0$ und $b > 0$, dann folgt aus $a < b$, dass $\frac{a}{b} < 1$ und $\frac{1}{b} < \frac{1}{a}$, also $\frac{1}{a} > \frac{1}{b}$. Beim Übergang zu Kehrwerten kehrt sich die Ungleichung um.

Häufig interessieren nicht tatsächliche Größen, sondern man ist nur an einer Abschätzung nach oben interessiert. Bei der Untersuchung von Grenzwerten in Kapitel 2 muss man z. B. komplizierte Ausdrücke so vergrößern, dass man sieht, dass sie unter einer vorgegebenen Schranke bleiben. Dabei hilft häufig die Regel: Ein Bruch positiver Zahlen wird vergrößert, indem der Zähler vergrößert wird und/oder der Nenner verkleinert wird. Teilt man durch weniger, so wird der Betrag einer Zahl größer.

Beispiel 1.48

- a) $1 < 2 \implies 1 \cdot 3 < 2 \cdot 3 \implies 3 < 6$ und $1 < 2 \implies 1 \cdot (-4) > 2 \cdot (-4) \implies -4 > -8$.
 b) $(2 > 0) \wedge (9 > 0) \implies 2 \cdot 9 = 18 > 0$,
 $(-2 < 0) \wedge (-9 < 0) \implies (-2) \cdot (-9) = 18 > 0$,
 $(-2 < 0) \wedge (9 > 0) \implies (-2) \cdot 9 = -18 < 0$.
 c) $0 < 2 < 8 \implies \frac{1}{2} > \frac{1}{8}$. ■

Die vielleicht wichtigste Beziehung der Analysis (das ist das Teilgebiet der Mathematik, mit dem wir uns in Kapitel 2 beschäftigen) ist die Dreiecksungleichung. Sie ist das elementare Werkzeug, das z. B. bei Stetigkeits- und Differenzierbarkeitsfragen verwendet wird. Im Gegensatz zur Algebra, bei der die Gleichheit beim Rechnen im Mittelpunkt steht, basiert die Analysis auf Abschätzungen.

Lemma 1.5 (Dreiecksungleichung)

Für $a, b \in \mathbb{R}$ gilt:

$$||a| - |b|| \leq |a - b|, \quad (1.13)$$

$$|a + b| \leq |a| + |b|. \quad (1.14)$$

Die Abschätzung (1.13) heißt **Dreiecksungleichung nach unten**, die Abschätzung (1.14) ist die eigentliche Dreiecksungleichung. Ihr Name wird verständlich, wenn wir diese Gleichungen für komplexe Zahlen über Längen im Dreieck erläutern (siehe Lemma 1.11 auf Seite 136). Das Lemma zeigt man, indem man die vier Kombinationen aus $a < 0$, $a \geq 0$, $b < 0$ und $b \geq 0$ diskutiert.

Beispiel 1.49

- a) $11 = |4 - (-7)| \geq ||4| - |-7|| = |4 - 7| = 3$ (Dreiecksungleichung nach unten),
 b) $9 = |3 + 6| \leq |3| + |6| = 9$, $3 = |3 - 6| \leq |3| + |-6| = 9$ (Dreiecksungleichung). ■

Bei der Abschätzung der Größe von Potenzen hilft der folgende Satz.

Satz 1.15 (Bernoulli'sche Ungleichung)

Für jede Zahl $x > -1$ und jedes $n \in \mathbb{N}_0$ gilt die **Bernoulli'sche Ungleichung**

$$(1 + x)^n \geq 1 + n \cdot x. \quad (1.15)$$

Beweis: Wir führen eine Vollständige Induktion nach der Variablen $n \in \mathbb{N}_0$ durch. Ungewohnt ist vielleicht, dass mit $x > -1$ eine weitere Variable vorkommt. Da x überabzählbar viele reelle Zahlen annehmen kann, ist eine Induktion nach x nicht möglich. Zu einem Wert von x gibt es nicht den Nachfolgewert, denn dazu müsste man eine beliebig kleine Zahl zu x addieren: $x + 0,1$ ist nicht der Nachfolger, $x + 0,0001$ schon eher, aber was ist mit $x + 0,000001$? Für einen bestimmten Wert von n lautet die zu beweisende Aussage: „Für alle $x \in \mathbb{R}$ mit $x > -1$ gilt: $(1+x)^n \geq 1+n \cdot x$ “. Wir müssen in jedem Schritt der Induktion alle möglichen Werte von x bedenken.

- Induktionsanfang: Für $n = 0$ ist $(1+x)^0 = 1 = 1 + 0 \cdot x$. Ist also $n = 0$, so gilt die Ungleichung für jeden Wert $x > -1$.
- Induktionsannahme: Für ein $n \in \mathbb{N}_0$ gelte die Ungleichung $(1+x)^n \geq 1+n \cdot x$ für jedes $x > -1$.
- Induktionsschluss: Zu zeigen ist, dass unter der Induktionsannahme auch $(1+x)^{n+1} \geq 1+(n+1) \cdot x$ für jedes $x > -1$ gilt:

$$\begin{aligned} (1+x)^{n+1} &= (1+x)^n \cdot (1+x) \stackrel{\text{Induktionsannahme und } 1+x>0}{\geq} (1+n \cdot x) \cdot (1+x) \\ &= 1 + n \cdot x + x + n \cdot x^2 \geq 1 + (n+1) \cdot x, \end{aligned}$$

wobei wir die nicht-negative Zahl nx^2 weggelassen haben. Wenn Sie bislang überwiegend mit Gleichungen gearbeitet haben, dann ist das Weglassen (oder Hinzufügen) eines Terms sehr ungewohnt. Aber es steht ja auch nicht „ $=$ “ zwischen den Rechenschritten. □

Wenn wir uns den Induktionsschluss im vorangehenden Beweis anschauen, dann sehen wir dort eine Kette von Gleichungen und Ungleichungen, die man von links nach rechts lesen und dann zu $(1+x)^{n+1} \geq 1+(n+1) \cdot x$ zusammenfassen kann. Denn in der Kette finden sich nur die Vergleiche „ $=$ “ und „ $\geq$ “, die zusammen „ $\geq$ “ ergeben. Vielfach ist das Hinschreiben solcher Ungleichungsketten einfacher als das Arbeiten mit äquivalenten Umformungen einer Ungleichung.

An dieser Stelle machen wir einen Vorgriff auf zwei Begriffe der beschreibenden Statistik (vgl. Kapitel 7.1.3.1).

Satz 1.16 (Ungleichung vom arithmetischen und geometrischen Mittel)

Es seien $x_1, x_2, \dots, x_n \geq 0$. Dann gilt

$$\sqrt[n]{x_1 x_2 \cdots x_n} \leq \frac{1}{n} \sum_{k=1}^n x_k.$$

Die linke Seite der Ungleichung ist das **geometrische Mittel** und die rechte Seite das **arithmetische Mittel** der Zahlen $x_1, x_2, \dots, x_n$.

Beweis: Wir verwenden wieder die Vollständige Induktion.

- Induktionsanfang für $n = 1$: $x_1 = x_1$ ist wahr.
- Induktionsannahme: Für ein $n \in \mathbb{N}$ gelte $x_1 x_2 \cdots x_n \leq \left[\frac{1}{n} \sum_{k=1}^n x_k \right]^n$.
- Induktionsschluss: Zu zeigen ist

$$x_1 x_2 \cdots x_{n+1} \leq \left[\frac{1}{n+1} \sum_{k=1}^{n+1} x_k \right]^{n+1},$$

o. B. d. A. sei x_{n+1} das maximale Element von $\{x_1, \dots, x_{n+1}\}$ und $\bar{x} := \frac{1}{n} \sum_{k=1}^n x_k$ das arithmetische Mittel der n Zahlen $x_1, \dots, x_n$. Dann gilt $x_{n+1} - \bar{x} \geq 0$, und mit der Bernoulli'schen Ungleichung folgt

$$\begin{aligned} \left(\frac{x_1 + \cdots + x_{n+1}}{(n+1)\bar{x}} \right)^{n+1} &= \left(\frac{n \cdot \bar{x} + x_{n+1}}{(n+1)\bar{x}} \right)^{n+1} = \left(\frac{(n+1) \cdot \bar{x} + x_{n+1} - \bar{x}}{(n+1)\bar{x}} \right)^{n+1} \\ &= \left(1 + \frac{x_{n+1} - \bar{x}}{(n+1)\bar{x}} \right)^{n+1} \stackrel{(1.15)}{\geq} 1 + (n+1) \frac{x_{n+1} - \bar{x}}{(n+1)\bar{x}} = \frac{x_{n+1}}{\bar{x}}. \end{aligned}$$

Mit der Induktionsannahme folgt dann

$$\left(\frac{x_1 + \cdots + x_{n+1}}{n+1} \right)^{n+1} \geq \bar{x}^{n+1} \frac{x_{n+1}}{\bar{x}} = \bar{x}^n x_{n+1} \geq x_1 \cdots x_n \cdot x_{n+1},$$

also genau die Behauptung. □

Für $x_1 = x_2 = \cdots = x_n$ sind beide Mittelwerte gleich.

Das arithmetische Mittel ist der gängigste Mittelwert. Ein Durchschnittspreis eines Produkts entsteht, indem die Einzelpreise addiert werden und durch die Anzahl der Angebote geteilt wird. Die Durchschnittsgeschwindigkeit eines Fahrzeugs, das z. B. eine Stunde mit 100 km/h, danach 0,5 Stunden mit 90 km/h und anschließend eine Stunde mit 80 km/h fährt, hat bei einer Fahrtdauer von fünf halben Stunden eine Durchschnittsgeschwindigkeit von

$$\frac{1}{5} (100 + 100 + 90 + 80 + 80) = 90 \text{ km/h}.$$

Das geometrische Mittel benutzt man im Zusammenhang mit Wachstumsprozessen, um eine mittlere Wachstumsrate zu berechnen. Verzinst sich beispielsweise ein Kapital im Rahmen eines Zuwachssparvertrags im ersten Jahr mit 5 % (Wachstumsfaktor 1,05), im zweiten und dritten Jahr mit 7 % und im vierten und fünften Jahr mit 10 %, so lautet der über das geometrische Mittel berechnete mittlere Wachstumsfaktor

$$\sqrt[5]{1,05 \cdot 1,07^2 \cdot 1,1^2} = 1,0778.$$

Dies entspricht einem durchschnittlichen Zinssatz von 7,78 %. Man bekommt den gleichen Zinsgewinn, wenn man das Kapital zu einem festen Zinssatz von 7,78 % für fünf Jahre anlegt, da man die fünf jährlichen Verzinsungen durch Multiplikation mit $(1,0778)^5 = 1,05 \cdot 1,07^2 \cdot 1,1^2$ durchführt.

1.4.4 Über das Lösen von Gleichungen und Ungleichungen

Neben dem Ausrechnen von exakten Werten reicht es häufig bereits aus, Abschätzungen für die exakten Werte zu ermitteln. Geht es beispielsweise um Toleranzen in technischen Systemen, so stellt sich die Aufgabe direkt als Ungleichung. Während das Lösen von Gleichungen in der Regel gut beherrscht wird, bereiten Ungleichungen unseren Studierenden häufig Schwierigkeiten. In diesem Abschnitt stellen wir daher einige typische Aufgabenstellungen zusammen und üben das Rechnen. Das Kapitel enthält keine neuen Begriffe oder Sätze und soll als Training dienen, bevor wir anschließend mit reellen Funktionen weitermachen.

Weitere Beispiele für das Lösen von Gleichungen mit Bezug zu Ingenieur Anwendungen findet man z. B. in (Ritter und Voß, 2015, S.79 ff). In späteren Kapiteln gehen wir auf das Lösen speziellerer Probleme mit Logarithmen und trigonometrischen Gleichungen ein.

1.4.4.1 Rationale Gleichungen, Wurzelgleichungen, Betragsgleichungen

Ganzrationale Gleichungen

Mit Hilfe von Äquivalenzumformungen versucht man, die Gleichung der Form $a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0$ zu vereinfachen und nach der Unbekannten x aufzulösen. Das gelingt gut für lineare ($n = 1$) und quadratische ($n = 2$) Gleichungen sowie mit Einschränkungen auch für Gleichungen höherer Ordnung. Das Lösen von ganzrationalen Gleichungen entspricht der Nullstellensuche von Polynomen, auf die wir noch detaillierter in Kapitel 1.5.6.1 eingehen werden. Im Folgenden seien $a, b, c \in \mathbb{R}$, $a \neq 0$, feste Zahlen.

Lineare Gleichungen $a \cdot x + b = 0$ sind unproblematisch. Die Lösung ist $x = -\frac{b}{a}$. Bei der **quadratischen Gleichung** $ax^2 + bx + c = 0$ betrachtet man die Normalform

$$x^2 + px + q = 0 \text{ mit } p = \frac{b}{a}, q = \frac{c}{a}$$

und erhält die Lösungen mit der p-q-Formel Satz 1.10 (Seite 55).

Beispiel 1.50

- a) Die Gleichung $-2x^2 - 4x + 6 = 0$ wird durch -2 dividiert und geht über in die quadratische Gleichung in Normalform $x^2 + 2x - 3 = 0$. Wegen der Diskriminante $D = \frac{4}{4} + 3 > 0$ gibt es zwei Lösungen, und zwar $x_{1,2} = -1 \pm \sqrt{\frac{4}{4} + 3}$ bzw. $x_1 = 1$ und $x_2 = -3$. Also ist die Lösungsmenge $L = \{1, -3\}$.
- b) Die Gleichung $x^2 - 4x + 13 = 0$ besitzt die Diskriminante $D = \frac{16}{4} - 13 < 0$ und ist deshalb reell nicht lösbar, $L = \emptyset$. ■

Die Lösung von **ganzrationalen Gleichungen höherer Ordnung** gelingt in der Regel höchstens bis zur Ordnung 4. Häufig kann man aber eine Lösung raten und

den Grad der Gleichung durch Polynomdivision reduzieren. Das lernen wir in Kapitel 1.5.6.3.

Ein weiterer Spezialfall ist die **biquadratische Gleichung**: $a \cdot x^4 + b \cdot x^2 + c = 0$. Die Substitution $z = x^2$ liefert eine quadratische Gleichung in z : $a \cdot z^2 + b \cdot z + c = 0$.

Beispiel 1.51

Die Gleichung $x^4 - 12x^2 + 11 = 0$ lässt sich durch die Substitution $z = x^2$ in $z^2 - 12z + 11 = 0$ überführen, also in eine quadratische Gleichung mit der Variable z . Diese lässt sich direkt mit der p-q-Formel lösen: $z_{1,2} = 6 \pm \sqrt{\frac{144}{4} - 11} = 6 \pm 5$, also $z_1 = 11$ und $z_2 = 1$. Für die ursprüngliche Gleichung ergibt sich $x^2 = 11 \Rightarrow x_{1,2} = \pm\sqrt{11}$ und $x^2 = 1 \Rightarrow x_{3,4} = \pm 1$. Wir erhalten die Lösungsmenge $L = \{\sqrt{11}, -\sqrt{11}, 1, -1\}$. ■

Alle Gleichungen, die man (bis auf Einschränkungen im Definitionsbereich) in eine ganzrationale Gleichung umformen kann, heißen **algebraische Gleichungen**. Dazu gehören auch gebrochen-rationale Gleichungen und Wurzelgleichungen.

Bei den **gebrochen-rationalen Gleichungen**, z. B. $\frac{ax^2+bx+c}{dx^3+ex^2+f} = g$, ist zunächst zu beachten, dass nicht durch null dividiert wird. Anschließend formt man in eine ganzrationale Gleichung um.

Beispiel 1.52

Wir bestimmen die Lösungsmenge der Gleichung

$$\frac{4x+1}{x+2} - \frac{2x^2-x-5}{x^2+4x+4} = 2 \quad \Longleftrightarrow \quad \frac{4x+1}{x+2} - \frac{2x^2-x-5}{(x+2)^2} = 2.$$

Die Gleichung ist definiert für $x \neq -2$. Multiplikation mit dem Hauptnenner $(x+2)^2$ liefert eine ganzrationale Gleichung:

$$\begin{aligned} (4x+1) \cdot (x+2) - (2x^2-x-5) &= 2(x+2)^2 \\ \Longleftrightarrow 4x^2+9x+2-2x^2+x+5 &= 2x^2+8x+8 \\ &\Longleftrightarrow 2x = 1 \\ &\Longleftrightarrow x = \frac{1}{2}. \end{aligned} \quad \blacksquare$$

Wurzelgleichungen

Möchte man Gleichungen, in denen eine zweite Wurzel vorkommt, lösen, so versucht man, die Wurzel auf eine Seite der Gleichung zu isolieren. Danach muss man quadrieren, d. h. nicht-äquivalent umformen (siehe Beispiel 1.14 auf Seite 21), und nach der Variable auflösen. Deshalb sind die erhaltenen Lösungen immer durch Einsetzen in die Ausgangsgleichung zu kontrollieren.

Beispiel 1.53

Wir lösen $\sqrt{2x+3} = x$. Durch Quadrieren erhält man

$$\sqrt{2x+3} = x \Rightarrow 2x+3 = x^2 \Rightarrow x^2 - 2x - 3 = 0.$$

Die quadratische Gleichung hat die Lösungen $x = 1 \pm \sqrt{1+3} = 1 \pm 2$, also 3 und -1 . Da quadriert wurde, können gegenüber der Ausgangsgleichung Lösungen hinzugekommen sein. Daher müssen wir die Lösungen durch Einsetzen verifizieren: Für $x = 3$ ist $\sqrt{6+3} = 3$ eine wahre Aussage. Aber für $x = -1$ ergibt sich: $\sqrt{-2+3} = 1 \neq -1$, -1 ist keine Lösung der Wurzelgleichung. Diese hat nur die Lösung 3.

Wir betrachten nochmals die Problematik beim Quadrieren: Die Gleichung $x-2 = 1$ besitzt die eindeutige Lösung $x = 3$. Durch Quadrieren der Gleichung folgt $(x-2)^2 = 1$, und diese Gleichung besitzt zwei Lösungen $x = 1$ und $x = 3$. Jedoch ist 1 keine Lösung der Ausgangsgleichung! ■

Beispiel 1.54

Wir lösen $\sqrt{2x-3} + 5 - 3x = 0$. Isolieren der Wurzel liefert: $\sqrt{2x-3} = 3x - 5$. Quadrieren ergibt $2x - 3 = 9x^2 - 30x + 25$ bzw. $x^2 - \frac{32}{9}x + \frac{28}{9} = 0$ mit Lösungen

$$x = \frac{16}{9} \pm \sqrt{\frac{256}{81} - \frac{28}{9}} = \frac{16}{9} \pm \sqrt{\frac{256 - 252}{81}} = \frac{16}{9} \pm \frac{2}{9},$$

also $x = 2$ oder $x = \frac{14}{9}$. Die Probe liefert für $x = 2$: $\sqrt{2 \cdot 2 - 3} + 5 - 3 \cdot 2 = 0$, also ist 2 Lösung der Ausgangsgleichung. Für $x = \frac{14}{9}$ ergibt sich $\sqrt{\frac{28}{9} - 3} + 5 - \frac{14}{3} = \frac{1}{3} + 5 - \frac{14}{3} \neq 0$, also ist $\frac{14}{9}$ keine Lösung der Ausgangsgleichung. ■

Betragsgleichungen

Betragsgleichungen löst man zumeist durch Fallunterscheidung (siehe Kapitel 1.14):

Beispiel 1.55

Wir bestimmen die Lösungsmenge von $|3x - 1| = -2x + 2$.

- Fall 1: $3x - 1 \geq 0$ bzw. $x \geq \frac{1}{3}$: In diesem Fall ist $|3x - 1| = 3x - 1$:

$$|3x - 1| = -2x + 2 \iff 3x - 1 = -2x + 2 \iff 5x = 3 \iff x = \frac{3}{5}.$$

Wegen $\frac{3}{5} \geq \frac{1}{3}$ handelt es sich dabei in diesem Fall tatsächlich um eine Lösung.

- Fall 2: $3x - 1 < 0$ bzw. $x < \frac{1}{3}$: In diesem Fall ist $|3x - 1| = -3x + 1$:

$$|3x - 1| = -2x + 2 \iff -3x + 1 = -2x + 2 \iff x = -1.$$

Auch diese Lösung passt zum Fall, daher ist die Lösungsmenge $L = \{\frac{3}{5}, -1\}$. ■

Mitunter kann man die Lösung grafisch durch Zeichnen der Funktionsgraphen bestimmen oder durch Quadrieren die Lösung berechnen:

Beispiel 1.56

$$\begin{aligned} |3x - 1| &= -2x + 2 \xrightarrow{\text{Quadrieren}} 9x^2 - 6x + 1 = 4x^2 - 8x + 4 \implies 5x^2 + 2x - 3 = 0 \\ \implies x^2 + \frac{2}{5}x - \frac{3}{5} &= 0 \implies x_{1,2} = -\frac{1}{5} \pm \sqrt{\frac{1}{25} + \frac{15}{25}} \implies x_1 = -1, x_2 = \frac{3}{5}. \end{aligned}$$

Einsetzen bestätigt, dass beide Kandidaten Lösungen sind. ■

1.4.4.2 Rationale Ungleichungen und Betragsungleichungen

Ungleichungen lassen sich wie Gleichungen äquivalent umformen. Dabei sind aber die Rechenregeln für Ungleichungen aus Satz 1.14 auf Seite 67 zu beachten. Häufige äquivalente Umformungen sind:

- Addition/Subtraktion einer Zahl c auf beiden Seiten der Gleichung,
- Multiplikation der Ungleichung mit einem Faktor $c \neq 0$:
 - $c > 0$: Ungleichungszeichen bleibt bestehen,
 - $c < 0$: Ungleichungszeichen dreht sich herum.

Zusätzlich kann man Folgerungen benutzen, indem man Terme unter Beachtung des Ungleichungszeichens vergrößert oder verkleinert. Zum Beispiel gilt: $x^2 - 1 > x \implies x^2 > x$. So kann man mit Gleichungen nicht verfahren.

Lineare Ungleichungen

Beispiel 1.57

Wir bestimmen die Lösungsmenge der Ungleichung $x + 3 \leq 7 - x$. Elementare Umformungen ergeben:

$$x + 3 \leq 7 - x \quad \xLeftrightarrow{+x-7} \quad 2x - 4 \leq 0 \quad \xLeftrightarrow{+4} \quad 2x \leq 4 \quad \xLeftrightarrow{/2} \quad x \leq 2.$$

Damit haben wir die Lösungsmenge $L = \{x \in \mathbb{R} : x \leq 2\} =]-\infty, 2]$ bestimmt. ■

Quadratische Ungleichungen

Hier betrachten wir Ungleichungen der Form $x^2 + a \cdot x + b < 0$ (> 0). Die Lösungsmenge kann bestehen aus

- der leeren Menge,
- allen reellen Zahlen,
- einem beschränkten Intervall,
- der Vereinigung zweier einseitig unbeschränkter Intervalle wie z. B. $] -\infty, c] \cup [d, \infty[$.

Beispiel 1.58

Wir lösen quadratische Ungleichungen mittels quadratischer Ergänzung. Alternativ kann man mit der p-q-Formel Nullstellen bestimmen. Nur dort kann sich das Vorzeichen der betrachteten Terme ändern.

a) Wir bestimmen alle Lösungen von $x^2 + 4x - 5 < 0$. Quadratische Ergänzung liefert:

$$\begin{aligned} x^2 + 4x + 4 - 9 < 0 &\iff (x + 2)^2 < 9 &\iff |x + 2| < 3 \\ &\iff -3 < x + 2 < 3 &\iff -5 < x < 1, \end{aligned}$$

also ist die Lösungsmenge $\{x \in \mathbb{R} : -5 < x < 1\}$. Da $x^2 + 4x - 5$ das Vorzeichen nur an den Nullstellen -5 und 1 ändert, erhält man diese Lösungsmenge auch durch Einsetzen von Zahlen, die zwischen den Nullstellen liegen.

b) Wir suchen alle Lösungen von $x^2 + 4x - 5 \geq 0$. Quadratische Ergänzung liefert:

$$\begin{aligned} x^2 + 4x + 4 - 9 \geq 0 &\iff (x+2)^2 \geq 9 \iff |x+2| \geq 3 \\ &\iff x \leq -5 \quad \vee \quad x \geq 1, \end{aligned}$$

also ist die Lösungsmenge $] -\infty, -5] \cup [1, \infty[$.

c) Für die Ungleichung $x^2 - 3x - 4 > 0$ erhalten wir

$$x^2 - 3x - 4 > 0 \iff \left(x^2 - 3x + \frac{9}{4}\right) - \frac{25}{4} > 0 \iff \left(x - \frac{3}{2}\right)^2 > \frac{25}{4}.$$

- Fall 1: $x \geq \frac{3}{2}$: Ziehen der Quadratwurzel ergibt $x - \frac{3}{2} > \frac{5}{2} \iff x > 4$.
- Fall 2: $x < \frac{3}{2}$: Unter Berücksichtigung des Vorzeichens erhalten wir

$$x - \frac{3}{2} < -\frac{5}{2} \iff x < -1.$$

Damit lautet die Lösungsmenge $L =] -\infty, -1[\cup] 4, \infty[$.

d) Die Ungleichung $x^2 + 1 > 0$ ist für alle $x \in \mathbb{R}$ erfüllt.

e) Die Ungleichung $x^2 + 1 < 0$ ist für kein $x \in \mathbb{R}$ erfüllt, also ist $L = \emptyset$. ■

Gebrochen-rationale Ungleichungen

Bei gebrochen-rationale Ungleichungen sind die interessanten Stellen die Nullstellen des Zählers und des Nenners. Hier kann sich das Vorzeichen der Funktionswerte ändern. Das muss bedacht werden, wenn man die Ungleichung mit Zähler oder Nenner multipliziert, so dass man es wieder mit Fallunterscheidungen zu tun hat.

Beispiel 1.59

Wir bestimmen die Lösungsmenge von $\frac{x+7}{x+1} \leq 4, x \in \mathbb{R} \setminus \{-1\}$.

- Fall 1: $x+1 > 0$, d. h. $x > -1$ und $x+7 \leq 4 \cdot (x+1) \iff 3 \leq 3x \iff x \geq 1$.
Damit spielt die Einschränkung $x > -1$ keine Rolle, und wir erhalten den Anteil $L_1 = [1, \infty[$ der gesamten Lösungsmenge.
- Fall 2: $x+1 < 0$, d. h. $x < -1$ und $x+7 \geq 4 \cdot (x+1) \iff 3 \geq 3x \iff x \leq 1$. Hier müssen wir berücksichtigen, dass wir uns im Fall $x < -1$ befinden und erhalten damit den Lösungsmengenanteil $L_2 =] -\infty, -1[$.

Insgesamt lautet die Lösungsmenge $L = L_1 \cup L_2 =] -\infty, -1[\cup [1, \infty[$. ■

Betragsungleichungen

Auch zur Auflösung des Betrags sind Fallunterscheidungen durchzuführen.

Beispiel 1.60

a) Wir bestimmen die Menge L aller Lösungen der Ungleichung $|x-1| > 3$.

- Fall 1: $x-1 \geq 0$, d. h. $x \geq 1$: $x-1 > 3 \iff x > 4$.

- Fall 2: $x - 1 < 0$, d. h. $x < 1$: $-x + 1 > 3 \iff -2 > x$.

Insgesamt ist die Lösungsmenge $L =]-\infty, -2[\cup]4, \infty[$.

- b) Wir suchen die Lösungen der Ungleichung $\frac{1-|x-2|}{|x-3|} < \frac{1}{2}$ für $x \neq 3$. Die Auflösung des Betrags ergibt:

$$|x-2| = \begin{cases} x-2, & x \geq 2 \\ -(x-2), & x < 2 \end{cases}, \quad |x-3| = \begin{cases} x-3, & x \geq 3 \\ -(x-3), & x < 3 \end{cases}.$$

Wir erhalten drei Fälle:

- Fall 1: $x > 3$: Hier ist $\frac{1-|x-2|}{|x-3|} = \frac{1-(x-2)}{x-3} = \frac{-x+3}{x-3} = -1$, d. h., die Ungleichung ist erfüllt.
- Fall 2: $2 \leq x < 3$: Hier ist $\frac{1-|x-2|}{|x-3|} = \frac{1-(x-2)}{-(x-3)} = \frac{-x+3}{-x+3} = 1$, d. h., die Ungleichung ist nicht erfüllt.
- Fall 3: $x < 2$: Hier ist $\frac{1-|x-2|}{|x-3|} = \frac{1-(-(x-2))}{-(x-3)} = \frac{x-1}{-x+3}$. Die Ungleichung lautet

$$\frac{x-1}{-x+3} < \frac{1}{2} \stackrel{2 \cdot (3-x) > 0}{\iff} 2(x-1) < 3-x \iff 3x < 5 \iff x < \frac{5}{3}.$$

Wegen $\frac{5}{3} < 2$ erhalten wir $] -\infty, \frac{5}{3}[$ als Bestandteil der Lösung.

Insgesamt haben wir die Lösungsmenge $L =]-\infty, \frac{5}{3}[\cup]3, \infty[$. ■

Allgemeine Ungleichungen

Ungleichungen treten in vielfältiger Form auf. Nicht für jeden Typ lässt sich ein Standardverfahren angeben. Hier ist – mal wieder – Kreativität gefragt.

Beispiel 1.61

Wir zeigen für alle $n \in \mathbb{N}$ die Ungleichung $\binom{n}{k} \frac{1}{n^k} \leq \frac{1}{2^{k-1}}$, $1 \leq k \leq n$. Es ist

$$\begin{aligned} \binom{n}{k} \frac{1}{n^k} &= \frac{\overbrace{n(n-1)(n-2) \cdots (n-k+1)}^{k \text{ Faktoren}}}{k!} \cdot \frac{1}{n^k} \\ &= \frac{1}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} \leq \frac{1}{k!}. \end{aligned}$$

Weiter folgt

$$\frac{1}{k!} = \frac{1}{1 \cdot 2 \cdot 3 \cdots (k-1) \cdot k} \leq \frac{1}{1 \cdot \underbrace{2 \cdot 2 \cdots 2}_{k-1 \text{ Faktoren}}} = \frac{1}{2^{k-1}}.$$

Wie bereits gesehen, kann man mitunter Ungleichungen mit natürlichen Zahlen mit Hilfe der Vollständigen Induktion nachweisen. ■

Beispiel 1.62

Wir beweisen durch Vollständige Induktion die Ungleichung

$$\frac{(2n)!}{n!n!} < 2^{2n-1}, \quad n = 2, 3, 4, \dots$$

- Induktionsanfang für $n = 2$: $\frac{4!}{2!2!} = 6 < 2^{4-1} = 8$ ist wahr.
- Induktionsannahme: Für ein $n \in \mathbb{N}$, $n \geq 2$, gelte $\frac{(2n)!}{n!n!} < 2^{2n-1}$.
- Induktionsschluss: Zu zeigen ist, dass aus der Induktionsannahme

$$\frac{(2(n+1))!}{(n+1)!(n+1)!} < 2^{2(n+1)-1}$$

folgt. Es gilt mit der Definition der Fakultät:

$$\frac{(2(n+1))!}{(n+1)!(n+1)!} = \frac{(2n)!(2n+1)(2n+2)}{n!(n+1)n!(n+1)} = \frac{(2n)!(2n+1)2(n+1)}{n!(n+1)n!(n+1)},$$

und durch Kürzen ergibt sich

$$\frac{(2n)!(2n+1)2}{n!(n+1)n!} = \frac{(2n)!}{n!n!} \cdot \frac{2(2n+1)}{n+1}.$$

Nach Einsetzen der Induktionsannahme folgt

$$\begin{aligned} \frac{(2n)!}{n!n!} \cdot \frac{2(2n+1)}{n+1} &< 2^{2n-1} \cdot \frac{2(2n+1)}{n+1} = 2^{2 \cdot n} \frac{2n+1}{n+1} \\ &< 2^{2n} \cdot \frac{2n+1+1}{n+1} = 2^{2n} \cdot \frac{2(n+1)}{n+1} = 2^{2n+1} = 2^{2(n+1)-1}. \end{aligned}$$

Damit ist gezeigt, dass die Behauptung für alle $n \in \mathbb{N}$, $n \geq 2$ richtig ist. ■

1.5 Reelle Funktionen

1.5.1 Notation reeller Funktionen

Abbildungen (siehe Definition 1.4 auf Seite 8) von Teilmengen D der reellen Zahlen auf Teilmengen $W \subset \mathbb{R}$ heißen **reellwertige Funktionen** bzw. **reelle Funktionen** mit einer reellen Variable. Sie bilden zusammen mit den komplexen Funktionen den Hauptgegenstand der Analysis, siehe Kapitel 2. Funktion ist nur ein anderer Name für Abbildung. Entsprechend heißt – sofern sie existiert – die zugehörige Umkehrabbildung f^{-1} **Umkehrfunktion**.

Reelle Funktionen kann man durch **Funktionsgraphen** darstellen. Dabei zeichnet man in einem kartesischen Koordinatensystem mit x - und y -Achse zu jedem x aus dem Definitionsbereich $D(f)$ von f den eindeutigen Punkt (x, y) mit $y = f(x)$ ein. Die x -Achse heißt auch **Abszisse**, die y -Achse **Ordinate** (siehe Abbildung 1.14). Durch die x - und die y -Achse wird die Zahlenebene in vier **Quadranten** eingeteilt, siehe Abbildung 1.15. Der erste Quadrant ist die Menge $\{(x, y) : x > 0, y > 0\}$ (oben rechts). Der zweite Quadrant ist $\{(x, y) : x < 0, y > 0\}$ (oben links), der dritte ist $\{(x, y) : x < 0, y < 0\}$ (unten links) und der vierte ist $\{(x, y) : x > 0, y < 0\}$ (unten rechts).

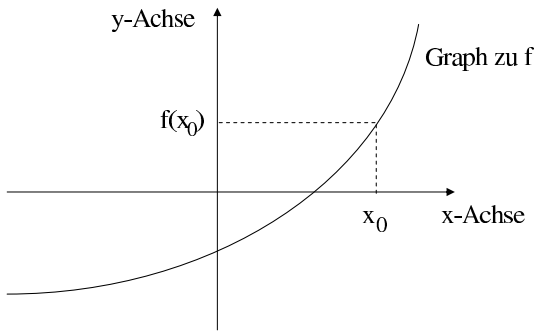


Abb. 1.14: Darstellung eines Funktionsgraphen

Beispiel 1.63

a) Eine **affin-lineare Funktion** ist gegeben durch $(a, b \in \mathbb{R})$

$$y = f(x) = a \cdot x + b.$$

Der Definitionsbereich ist $D(f) = \mathbb{R}$, und der Wertebereich ist $W(f) = \mathbb{R}$. Der Funktionsgraph ist eine Gerade.

b) Die **Normalparabel** besitzt die Funktionsgleichung

$$y = g(x) = x^2$$

mit Definitionsbereich $D(g) = \mathbb{R}$ und Wertebereich $W(g) = [0, \infty[$. Affin-lineare Funktionen und Parabeln sind Spezialfälle von Polynomen. Mit dieser Klasse von elementaren Funktionen werden wir uns noch eingehend in Kapitel 1.5.6.1 beschäftigen.

c) Die **stückweise lineare Funktion**

$$y = h(x) = \begin{cases} 1, & x \leq 1, \\ 2, & 1 < x \leq 2, \\ x, & x > 2 \end{cases}$$

hat den Definitionsbereich $D(h) = \mathbb{R}$ und den Wertebereich $W(h) = \{1\} \cup [2, \infty[$. Die Funktionsgraphen zu f , g und h sind in Abbildung 1.16 dargestellt. ■

Bis jetzt kennen wir Funktionen in **expliziter** Darstellung $y = f(x)$, wie z. B. $y = \sqrt{1 - x^2}$, $x \in [-1, 1]$. Ist die Funktion in der Form $F(x, y) = 0$, wie z. B.

$$x^2 + y^2 - 1 = 0, \quad x \in [-1, 1], \quad y \in [0, 1]$$

II	I
III	IV

Abb. 1.15: Quadranten des Koordinatensystems

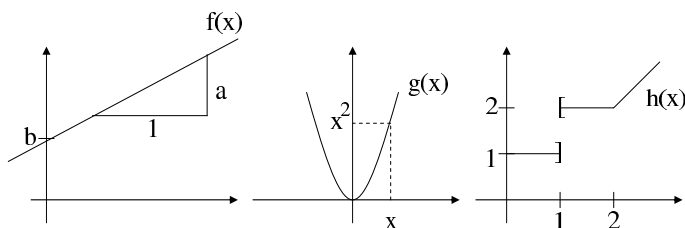


Abb. 1.16: Funktionsgraphen zu Beispiel 1.63

gegeben, so spricht man von **impliziter** Darstellung.

In der Ingenieurpraxis sind Funktionen häufig nur unvollständig als **Messreihen** der Form $(x_i, y_i = f(x_i))$, $i = 1, \dots, n$, d.h. als endliche Menge von Zahlenpaaren, gegeben.

Es ist durchaus üblich, dass der Definitionsbereich $D(f)$ (und damit der Wertebereich $W(f)$) einer Funktion nicht angegeben werden, obwohl die Angabe wichtig ist.

Beispiel 1.64

- Für $y = f(x) = \sqrt{x^2 - 2x}$ ist $D(f)$ so zu wählen, dass $x^2 - 2x$ nicht-negativ ist, d.h. $x^2 \geq 2x$. Im Fall $x \geq 0$ heißt dies $x \geq 2$, und im Fall $x \leq 0$ bedeutet dies $x \leq 2$. Also ist $D(f) =]-\infty, 0] \cup [2, \infty[$.
- Die Funktion $y = f(x) = \frac{2x}{x(x-3)}$ ist nur dann definiert, wenn der Nenner ungleich 0 ist, d.h. $D(f) = \mathbb{R} \setminus \{0, 3\}$.
- Durch die Gleichung $x^2 + y^2 = 1$ wird der Einheitskreis beschrieben, wobei nur $x \in [-1, 1]$ und $y \in [-1, 1]$ sinnvoll ist. Durch die Gleichung wird aber kein funktionaler Zusammenhang $y = f(x)$ erklärt. Beispielsweise sind dem Wert $x = 0$ die beiden y -Werte ± 1 zugeordnet. Schränkt man allerdings die y -Werte auf $[0, 1]$ ein, wird jedem $x \in [-1, 1]$ durch $x^2 + y^2 = 1$ genau ein y -Wert $y = f_1(x) = \sqrt{1 - x^2}$ zugeordnet. Entsprechend erhält man $y = f_2(x) = -\sqrt{1 - x^2}$, wenn man y aus $[-1, 0]$ nimmt. Somit werden durch die Kreisgleichung implizit zwei Funktionen definiert:
 - $y = f_1(x) = \sqrt{1 - x^2}$ mit $x \in [-1, 1]$ und $y \in [0, 1]$ beschreibt den oberen Halbkreis und
 - $y = f_2(x) = -\sqrt{1 - x^2}$ mit $x \in [-1, 1]$ und $y \in [-1, 0]$ beschreibt den unteren Halbkreis. ■

Sind f und g reelle Funktionen mit gleichem Definitionsbereich D , so kann man auf D die neuen Funktionen $f + g$, $f \cdot g$ und $\frac{f}{g}$ definieren über

$$f + g : x \mapsto f(x) + g(x), \quad f \cdot g : x \mapsto f(x) \cdot g(x) \quad \text{und} \quad \frac{f}{g} : x \mapsto \frac{f(x)}{g(x)}, \quad g(x) \neq 0,$$

oder mit anderen Worten

$$(f + g)(x) := f(x) + g(x), \quad (f \cdot g)(x) := f(x) \cdot g(x), \quad \left(\frac{f}{g}\right)(x) := \frac{f(x)}{g(x)}.$$

Die Verknüpfung der Funktionen ist also definiert über die entsprechende Verknüpfung der Funktionswerte. Mit $f(x)$ bezeichnet man streng genommen einen Funktionswert von f an der Stelle x und mit f (ohne Argument) die Funktion. So streng wird das aber nicht unterschieden. Im Gegenteil: Häufig meint man mit $f(x)$ die Funktion und möchte ausdrücken, dass man für diese Funktion die Variable x verwendet. Wenn wir später betonen wollen, dass es sich um einen konkreten Funktionswert an einer Stelle handelt, so fügen wir dem Variablennamen einen Index hinzu. $f(x_0)$ ist demnach der Funktionswert an der Stelle x_0 .

Besonders ausgezeichnet sind Stellen x_0 , an denen der Funktionswert null ist, sie heißen **Nullstellen** der Funktion.

Zum Abschluss des Abschnitts sehen wir uns an, wie man eine Funktion erhält, die einen gegenüber dem Graphen der Ausgangsfunktion verschobenen Funktionsgraphen besitzt. Eine Verschiebung in x -Richtung kann man durch die Addition eines konstanten Werts (eines Offsets) zum Argument x einer Funktion erreichen, eine Verschiebung in y -Richtung gewinnt man durch Addition einer Konstante zum Funktionswert:

- Der Graph von $f(x + c)$ ist um c gegenüber dem Graphen von f in Richtung der x -Achse verschoben. Ist $c > 0$, so liegt eine Verschiebung nach links vor, bei $c < 0$ handelt es sich um eine Verschiebung nach rechts.
- Der Graph von $f(x) + c$ ist um c gegenüber dem Graphen von f in Richtung der y -Achse verschoben. Ist $c > 0$, so handelt es sich um eine Verschiebung nach oben, ist $c < 0$, so ist der Graph nach unten verschoben.

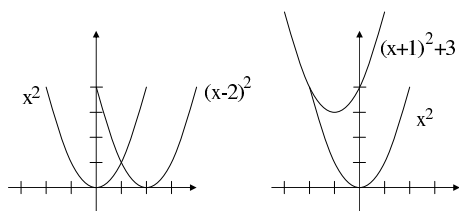


Abb. 1.17: Verschiebungen von Funktionsgraphen

Beispiel 1.65

- a) Die Funktion $\tilde{f}(x) = (x - 2)^2$ ist gegenüber der Normalparabel $f(x) = x^2$ um $c = 2$ nach rechts verschoben (siehe Abbildung 1.17).
- b) Die Funktion $\tilde{f}(x) = x^2 + 2x + 4 = x^2 + 2x + 1 + 3 = (x + 1)^2 + 3$ ist gegenüber der Normalparabel $f(x) = x^2$ um 3 nach oben und um 1 nach links verschoben. ■

1.5.2 Eigenschaften von reellen Funktionen

Für die Beschreibung von Funktionen sind einige qualitative Eigenschaften von Bedeutung, die wir bereits hier einführen und mit denen wir intensiv in Kapitel 2 arbeiten.

Definition 1.25 (Monotonie)

Eine Funktion $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ heißt auf dem Intervall $I \subseteq D$ genau dann

- **monoton wachsend** (*monoton steigend*), wenn $f(x) \leq f(y)$ für alle $x, y \in I$ mit $x < y$ (vgl. Abbildung 1.18).
- **streng monoton wachsend** (*streng monoton steigend*), wenn $f(x) < f(y)$ für alle $x, y \in I$ mit $x < y$.
- **monoton fallend**, wenn $f(x) \geq f(y)$ für alle $x, y \in I$ mit $x < y$.
- **streng monoton fallend**, wenn $f(x) > f(y)$ für alle $x, y \in I$ mit $x < y$.

f heißt **(streng) monoton** auf I , sofern f auf I (streng) monoton wächst oder fällt.

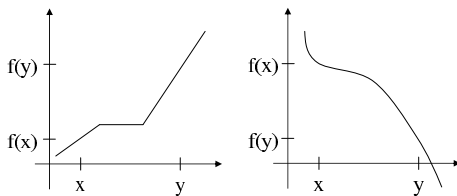


Abb. 1.18: Monoton wachsende und streng monoton fallende Funktion

Bislang haben Sie vermutlich nur Graphen zu Funktionen gezeichnet, die auf Teilintervallen des Definitionsbereichs monoton sind. Wenn man sich aber etwas anstrengt, dann findet man auch Funktionen, die auf keinem Teilintervall monoton steigend oder fallend sind.

Neben dem Monotonieverhalten ist die Symmetrie ein wichtiges Merkmal einer Funktion (siehe Abbildung 1.19).

Definition 1.26 (Symmetrie)

Eine Funktion $y = f(x)$ heißt im Intervall $I = [-a, a] \subseteq D$ genau dann

- **symmetrisch zur y -Achse oder gerade**, wenn für jedes $x \in I$ gilt

$$f(-x) = f(x),$$

- **punktsymmetrisch zum Ursprung oder ungerade**, wenn für jedes $x \in I$ gilt

$$f(-x) = -f(x).$$

Der Graph einer geraden Funktion ist an der y -Achse gespiegelt. Der Graph einer ungeraden Funktion geht bei Drehung um 180° um den Koordinatenursprung in sich selbst über (siehe Abbildung 1.19).

Beispiel 1.66

Ist $f : \mathbb{R} \rightarrow \mathbb{R}$, dann ist $g(x) := f(x) + f(-x)$ eine gerade und $h(x) := f(x) - f(-x)$ eine ungerade Funktion. Jedes f ist Summe einer geraden und einer ungeraden Funktion: $f(x) = \frac{1}{2}g(x) + \frac{1}{2}h(x)$. ■

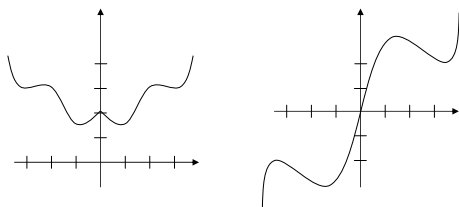


Abb. 1.19: Eine gerade und eine ungerade Funktion

Bei der Multiplikation symmetrischer Funktionen „vererben“ sich Symmetrien:

Lemma 1.6 (Produkt gerader und ungerader Funktionen)

- a) Das Produkt zweier gerader Funktionen ist gerade.
- b) Das Produkt zweier ungerader Funktionen ist gerade.
- c) Das Produkt einer geraden und einer ungeraden Funktion ist ungerade.

Achtung: Im Gegensatz zu b) und c) ist das Produkt zweier ungerader Zahlen ungerade und das Produkt einer geraden mit einer ungeraden Zahl gerade. Die Begriffe für Funktionen unterscheiden sich also von den entsprechenden Begriffen für Zahlen.

- Beweis:** a) Seien f und g gerade, dann ist $(f \cdot g)(-x) = f(-x)g(-x) = f(x)g(x) = (f \cdot g)(x)$, d. h., $f \cdot g$ ist gerade.
- b) Seien f und g ungerade, dann ist $(f \cdot g)(-x) = f(-x)g(-x) = [-f(x)] \cdot [-g(x)] = (f \cdot g)(x)$, d. h., $f \cdot g$ ist gerade.
- c) Ist f gerade und g ungerade, dann ist $(f \cdot g)(-x) = f(-x)g(-x) = f(x)[-g(x)] = -(f \cdot g)(x)$, d. h., $f \cdot g$ ist ungerade. □

Definition 1.27 (Beschränktheit)

Eine Funktion $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ heißt auf dem Intervall $I \subseteq D$ genau dann

- **beschränkt nach unten**, wenn es eine Konstante $m \in \mathbb{R}$ gibt mit $f(x) \geq m$ für alle $x \in I$ und
- **beschränkt nach oben**, wenn es eine Konstante M gibt mit $f(x) \leq M$ für alle $x \in I$ bzw.
- **beschränkt**, wenn es eine Konstante M gibt mit $|f(x)| \leq M$ für alle $x \in I$ (vgl. Abbildung 1.20).

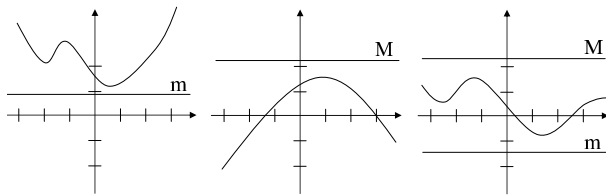


Abb. 1.20: Beschränkte Funktionen, von links nach rechts: beschränkt nach unten, nach oben und beides

Eine Funktion ist beschränkt genau dann, wenn ihr Wertebereich (als Menge reeller Zahlen) beschränkt ist. Sie ist nach oben (unten) beschränkt genau dann, wenn ihr Wertebereich nach oben (unten) beschränkt ist (siehe Definition 1.19 auf Seite 47).

In der Technik spielen periodische Funktionen bei der Beschreibung von Schwingungsvorgängen (mechanische und elektrische Schwingungen) eine zentrale Rolle.

Definition 1.28 (Periodizität)

Eine Funktion $y = f(x)$ mit Definitionsbereich $D = \mathbb{R}$ heißt genau dann **periodische Funktion** mit der **Periode** $p > 0$, wenn für jedes $x \in \mathbb{R}$ gilt (vgl. Abbildung 1.21)

$$f(x + p) = f(x).$$

Besitzt f die Periode p , dann ist auch $k \cdot p$ mit $k \in \mathbb{N}$ eine Periode von f . Die kleinste Periode von f heißt **primitive Periode** von f .

Bei einer periodischen Funktion wiederholen sich die Funktionswerte immer wieder. Jede Zahl des Wertebereichs wird unendlich oft angenommen.

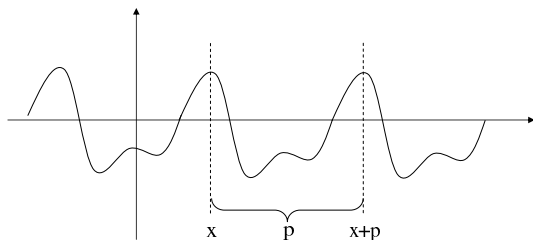


Abb. 1.21: Periodische Funktion mit Periode p

Beispiel 1.67

a) Für die Funktion $f(x) = x^2 - 1$ mit $D(f) = [0, 1]$ und $W(f) = [-1, 0]$ gilt:

- f ist streng monoton steigend: Für $0 \leq x_1 < x_2$ folgt $x_1 \cdot x_1 \leq x_2 \cdot x_1 < x_2 \cdot x_2$, d. h. $x_1^2 < x_2^2$. Subtraktion von -1 auf beiden Seiten liefert $x_1^2 - 1 < x_2^2 - 1$, d. h., es gilt: $x_1 < x_2 \implies f(x_1) < f(x_2)$.
- f ist beschränkt: Für $x \in [0, 1]$ folgt $-1 \leq x^2 - 1 \leq 0$.
- f besitzt die Nullstellen: $x^2 - 1 = 0 \iff x^2 = 1 \iff x = \pm 1$, aber $-1 \notin D(f)$. Also ist $x = 1$ die einzige Nullstelle von f .

b) Die Funktion $y = f(x) = a \cdot x + b$ mit $a, b \in \mathbb{R}$ beschreibt eine Gerade (vgl. Kapitel 1.5.6.1). Es ist $D(f) = \mathbb{R}$, und für $a \neq 0$ ist $W(f) = \mathbb{R}$.

- Monotonie: Für $a > 0$ ist f streng monoton steigend:

$$x_1 < x_2 \implies a \cdot x_1 < a \cdot x_2 \implies a \cdot x_1 + b < a \cdot x_2 + b.$$

Für $a < 0$ ist f streng monoton fallend:

$$x_1 < x_2 \implies a \cdot x_1 > a \cdot x_2 \implies a \cdot x_1 + b > a \cdot x_2 + b.$$

Im Fall $a = 0$ ist f die konstante Funktion $y = f(x) = b$.

- Symmetrie: Eine Gerade kann nur dann achsensymmetrisch sein, wenn sie parallel zur x -Achse verläuft, also die Steigung $a = 0$ hat. Punktsymmetrie liegt genau dann vor, wenn die Gerade durch den Koordinatenursprung geht, also bei $b = 0$.
- Beschränktheit: $y = f(x) = a \cdot x + b$ ist nur für $a = 0$ beschränkt.
- Achsenschnittpunkte: $f(0) = a \cdot 0 + b = b$ und $f(x) = 0 \iff a \cdot x + b = 0 \iff x = -\frac{b}{a}$ für $a \neq 0$. f besitzt für $a \neq 0$ genau eine Nullstelle. Für $a = 0$ besitzt f nur für $b = 0$ Nullstellen.
- Periodizität: $f(x+p) = f(x) \iff a(x+p) + b = ax + b \iff ap = 0 \iff a = 0$. f ist nur für $a = 0$ als konstante Funktion periodisch im Sinne der Definition.

c) Die Funktion $y = f(x)$ mit $D(f) = [-2, 2]$ und $W(f) = [0, 4]$ sei definiert durch

$$f(x) = \begin{cases} -x^2, & -2 \leq x < 0, \\ x^2, & 0 \leq x \leq 2. \end{cases}$$

- Monotonie: Wir untersuchen jeden Zweig der Funktion separat:

- * Für $x_1 < x_2 < 0$ folgt $-x_1^2 < -x_2^2$, d. h. $f(x_1) < f(x_2)$.
- * Für $x_1 < 0 \leq x_2$ folgt $-x_1^2 < x_2^2$, d. h. $f(x_1) < f(x_2)$.
- * Für $0 \leq x_1 < x_2$ folgt $x_1^2 < x_2^2$, d. h. $f(x_1) < f(x_2)$.

Also ist f streng monoton steigend.

- Nullstellen: $f(x) = 0 \implies x = 0$ ist einzige Nullstelle.
- Symmetrien: Es ist $f(-x) = -f(x)$, denn
 - * für $x \leq 0$ gilt: $f(-x) = (-x)^2 = x^2 = -f(x)$ und
 - * für $x \geq 0$ gilt: $f(-x) = -(-x)^2 = -x^2 = -f(x)$.

Also ist f punktsymmetrisch zum Ursprung.

d) Eine **Sägezahnfunktion** wird in Röhrenfernsehern als „Zeilenkipp“ verwendet, und wir werden uns später noch intensiv mit entsprechenden Wechselspannungen beschäftigen (vgl. Seite 819). Ein Vertreter dieser Funktionsklasse ist $y = f(x)$ mit $D(f) = \mathbb{R}$ und (siehe Abbildung 1.22)

$$f(x) := x + k, \text{ für } -k < x \leq -k + 1, \quad k \in \mathbb{Z}.$$

Liegt beispielsweise x im Intervall $]4, 5]$, so ist der zugehörige Funktionswert $f(x) = x - 4$. Die Funktion f ist periodisch mit Periode $p = 1$, denn wir zeigen $f(x + 1) =$

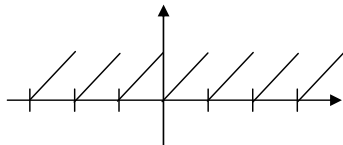


Abb. 1.22: Sägezahnfunktion

$f(x)$ für alle x : Für beliebiges $x \in \mathbb{R}$ gibt es ein $k \in \mathbb{Z}$ mit $x \in]-k, -k + 1]$. Dort ist $f(x) = x + k$. Das Argument $x + 1$ liegt in $] -k + 1, -k + 2] =] - (k - 1), -(k - 1) + 1]$, und folglich ist $f(x + 1) = x + 1 + k - 1 = x - k$. ■

1.5.3 Umkehrfunktion

Die Umkehrfunktion f^{-1} einer Funktion f wurde bereits mit Satz 1.3 auf Seite 10 eingeführt. Während wir es dort noch mit Abbildungen zwischen beliebigen Mengen zu tun hatten, gehen wir nun auf den wichtigen Spezialfall reeller Funktionen ein.

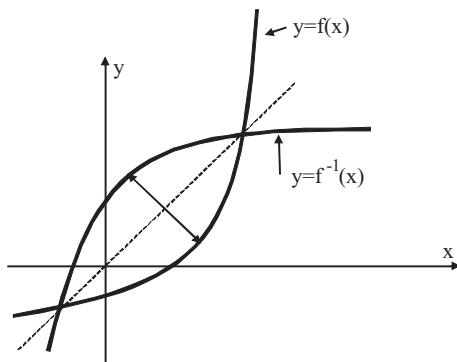


Abb. 1.23: Die Umkehrfunktion entsteht durch Spiegelung an der Diagonalen.

Beispiel 1.68

Das Weg-Zeit-Gesetz des freien Falls eines Steins lautet

$$s(t) = \frac{1}{2}gt^2$$

mit der Erdbeschleunigung $g = 9,81 \frac{\text{m}}{\text{s}^2}$. Wir berechnen die Zeitspanne, bis der Stein aus 10 m Höhe auf dem Boden aufschlägt.

Gegeben ist also der Funktionswert $s(t) = 10$ m, wir fragen nach dem zugehörigen Argument t . Wir haben die Gleichung $10 = \frac{1}{2}gt^2$ nach t aufzulösen:

$$t = s^{-1}(10) = \sqrt{\frac{2 \cdot 10}{g}} \quad s \approx 1,43 \quad s.$$

Im Beispiel suchen wir zum Bild der Funktion $s(t)$ das zugehörige Urbild. Das ist nur möglich, wenn es zum Bild keine zwei unterschiedlichen Urbilder gibt, d. h., wenn die Funktion injektiv ist. ■

Damit zu einer Funktion f eine Umkehrfunktion existiert, muss f sowohl injektiv als auch surjektiv sein. Die Surjektivität erhält man aber stets dadurch, dass man den Wertebereich $W(f)$ als Zielmenge der Abbildung verwendet: Statt $f : D \rightarrow E$ betrachten wir $f : D \rightarrow W(f) \subset E$. Dieser Wertebereich wird zum Definitionsbereich von f^{-1} : $D(f^{-1}) = W(f)$.

Beispiel 1.69

Die Funktion $f(x) = x^2$, $f : [0, \infty[\rightarrow \mathbb{R}$ ist nicht surjektiv, da keine negativen Zahlen angenommen werden. Dagegen ist $f : [0, \infty[\rightarrow [0, \infty[$ mit der gleichen Abbildungsvorschrift surjektiv. ■

Die Umkehrung der Umkehrfunktion ergibt die Ausgangsfunktion:

$$f^{-1}(f(x)) = x, \quad x \in D(f), \quad f(f^{-1}(y)) = y, \quad y \in W(f).$$

Bemerkungen:

- Ist f streng monoton, so ist f injektiv und umkehrbar. Zur Bestimmung der Umkehrfunktion für nicht-monotone Funktionen muss gegebenenfalls der Definitionsbereich der Funktion eingeschränkt werden. Die Existenz einer Umkehrfunktion ist bezüglich jedes Intervalls $I \subseteq D(f)$ gesichert, in welchem die Ausgangsfunktion f streng monoton und damit injektiv ist.
- Die Funktion $y = f(x)$ und ihre Umkehrfunktion $x = f^{-1}(y)$ besitzen bei dieser Bezeichnung der Variablen in einem Koordinatensystem den gleichen Funktionsgraphen (siehe Abbildung 1.23). Durch die Auflösung $x = f^{-1}(y)$ nach x und Vertauschen von x und y erhält man die Umkehrfunktion $y = f^{-1}(x)$ mit x als der unabhängigen Variablen:
 - 1. Schritt: Löse $y = f(x)$ nach x auf: $x = f^{-1}(y)$.
 - 2. Schritt: Vertausche die Variablennamen x und y .
- Die Funktionsgraphen von $y = f(x)$ und $y = f^{-1}(x)$ liegen im kartesischen Koordinatensystem spiegelbildlich zur Achse $y = x$.
- An der Spiegelung der Graphen kann man z. B. ablesen, dass f^{-1} die gleichen Monotonieeigenschaften wie f hat:

Lemma 1.7 (Monotonie der Umkehrfunktion)

Ist eine Funktion $f : D \rightarrow \mathbb{R}$ streng monoton wachsend (fallend) auf D , so ist die auf dem Wertebereich $W(f)$ existierende Umkehrfunktion $f^{-1} : W(f) \rightarrow D$ ebenfalls streng monoton wachsend (fallend) auf $W(f)$.

Beispiel 1.70

a) Die Funktion $y = f(x) = 2x + 1$ mit $D(f) = W(f) = \mathbb{R}$ ist streng monoton steigend und damit umkehrbar. Wir berechnen die Umkehrfunktion:

- 1. Auflösung nach x : $y = 2x + 1 \implies x = \frac{y-1}{2}$.
- 2. Umbenennung: $y = f^{-1}(x) = \frac{x-1}{2}$ mit $D(f^{-1}) = W(f) = \mathbb{R}$ und $W(f^{-1}) = D(f) = \mathbb{R}$.

b) Die Funktion $y = f(x) = 1 + x^2$ mit $D(f) = [0, \infty[$ und $W(f) = [1, \infty[$ ist streng monoton steigend.

- 1. Auflösung nach x : $y = 1 + x^2 \implies x = \sqrt{y-1}$, da $x \geq 0$ für $x \in D(f)$ gilt.
- 2. Umbenennung: $y = f^{-1}(x) = \sqrt{x-1}$ mit $D(f^{-1}) = [1, \infty[$ und $W(f^{-1}) = [0, \infty[$.

c) Die Funktion $y = f(x) = x^2$ mit $D(f) = \mathbb{R}$ und $W(f) = [0, \infty[$ ist nicht streng monoton, sie ist wegen $f(-1) = f(1) = 1$ auch nicht injektiv und somit nicht umkehrbar. Die Einschränkung des Definitionsbereichs auf $[0, \infty[$ führt auf die bereits oben betrachtete Funktion $y = \tilde{f}(x) = x^2$ mit $D(\tilde{f}) = W(\tilde{f}) = [0, \infty[$, die streng monoton steigend und somit umkehrbar ist. Die Umkehrfunktion lautet $y = \tilde{f}^{-1}(x) = \sqrt{x}$ mit $D(\tilde{f}^{-1}) = W(\tilde{f}^{-1}) = [0, \infty[$. ■

1.5.4 Verkettung von Funktionen

Häufig wendet man zwei Funktionen nacheinander an. Dabei setzt man das Ergebnis der ersten Funktion als Argument in die zweite ein.

Beispielsweise kann man bei einem festen Stundenlohn $S > 0$ ein Bruttogehalt als Funktion $g(x) = S \cdot x$ der geleisteten Arbeitsstunden x berechnen. Das zugehörige Nettogehalt ergibt sich nun über eine Funktion des Bruttogehalts, die vom Bruttogehalt z.B. einen festen Abgabebetrag a abzieht und den Rest durch Multiplikation mit einem Faktor $0 < b < 1$ versteuert. Dabei handelt es sich um die Funktion $f(u) = (u - a) \cdot b$ des Bruttogehalts u . Es entsteht eine **Verkettung** zweier Funktionen:

$$\text{Arbeitsstunden} \xrightarrow{g} \text{Bruttogehalt} \xrightarrow{f} \text{Nettogehalt}.$$

Zunächst wird $x \in D(g)$ durch g auf $u \in D(f)$ abgebildet. Dieses Element wird dann durch die zweite Funktion f auf ein weiteres Element $y \in W(f)$ abgebildet. Durch die Hintereinanderausführung beider Funktionen ergibt sich eine neue Funktion „ f verkettet mit g “ als

$$f \circ g : D(f \circ g) \rightarrow W(f), \quad y = (f \circ g)(x) := f(g(x))$$

mit Definitionsbereich $D(f \circ g) = \{x \in D(g) : g(x) \in D(f)\}$ (siehe Abbildung 1.24).

Beispiel 1.71

- a) Die Berechnung des Nettogehalts geschieht über $f(g(x)) = (g(x)-a) \cdot b = (S \cdot x - a) \cdot b$. Damit das Gehalt nicht negativ wird, wählen wir $D(f) = [a, \infty[$ und $D(f \circ g) = [\frac{a}{S}, \infty[\subset D(g) = [0, \infty[$.
- b) Die Funktion $y = f(x) = \sqrt{2x+4}$ mit $D(f) = [-2, \infty[$, $W(f) = [0, \infty[$ soll mit der Funktion $y = g(x) = x^2 - 1$ mit $D(g) = \mathbb{R}$, $W(g) = [-1, \infty[$ verkettet werden. Wegen $W(g) \subseteq D(f)$ ist die Verkettung für alle $x \in D(g)$ erklärt. Es ergibt sich die Abbildung:

$$y = (f \circ g)(x) = f(g(x)) = f(x^2 - 1) = \sqrt{2(x^2 - 1) + 4} = \sqrt{2x^2 + 2}$$

mit $D(f \circ g) = D(g) = \mathbb{R}$ und $W(f \circ g) = [\sqrt{2}, \infty[\subseteq W(f)$.

- c) Zur Temperaturmessung verwendet man neben der Celsius-Skala die (wissenschaftliche) Kelvin-Skala, und in angelsächsischen Ländern ist die Fahrenheit-Skala üblich. Bezeichnet man die Temperaturen auf der Kelvin-Skala mit x , die der Celsius-Skala mit u und die der Fahrenheit-Skala mit y , dann gelten folgende Umrechnungen:

$$u = x - 273,15 \quad \text{und} \quad y = 32 + 1,8 \cdot u.$$

Die Fahrenheit-Werte lassen sich durch Einsetzen auch unmittelbar aus den Kelvin-Graden gewinnen:

$$y = 32 + 1,8 \cdot (x - 273,15) = 1,8 \cdot x - 459,67.$$

Betrachtet man die Gleichungen als Definitionen von Funktionen, so wird daraus

$$g(x) = x - 273,15; \quad f(u) = 32 + 1,8 \cdot u; \quad h(x) = f(g(x)) = 1,8 \cdot x - 459,67.$$

- d) Die Verkettung der Funktionen $f(x) = \frac{1+x}{1-x}$ mit $D(f) = \mathbb{R} \setminus \{1\}$ und $g(x) = x^2 + 2$ mit $D(g) = \mathbb{R}$ und $W(g) = [2, \infty[$ ist gegeben durch

$$y = (f \circ g)(x) = f(g(x)) = f(x^2 + 2) = \frac{1 + (x^2 + 2)}{1 - (x^2 + 2)} = -\frac{3 + x^2}{1 + x^2}$$

mit $D(f \circ g) = \mathbb{R}$ und $W(f \circ g) = [-3, -1[$. ■

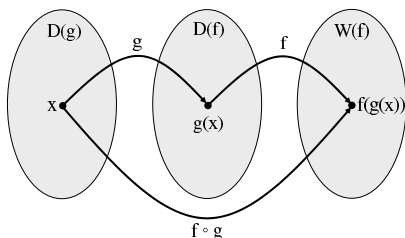


Abb. 1.24: Verkettung der Funktionen f und g

1.5.5 Signum- und Betragsfunktion

Mittels Signum- und Betragsfunktion können wir explizite Fallunterscheidungen vermeiden.

Definition 1.29 (Signumfunktion)

Die Funktion $f(x) = \text{sign}(x)$, die jeder reellen Zahl x ihr Vorzeichen zuordnet, heißt **Signumfunktion** oder **Vorzeichenfunktion** (siehe Abbildung 1.25).

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

Schon bei den reellen Zahlen haben wir den Betrag (1.12) auf Seite 65 eingeführt.

Definition 1.30 (Betragsfunktion)

Die Funktion $f(x) = |x|$, die jeder reellen Zahl x ihren Betrag zuordnet, heißt **Betragsfunktion**.

$$|x| := \begin{cases} x, & \text{für } x > 0, \\ 0, & \text{für } x = 0, \\ -x, & \text{für } x < 0. \end{cases}$$

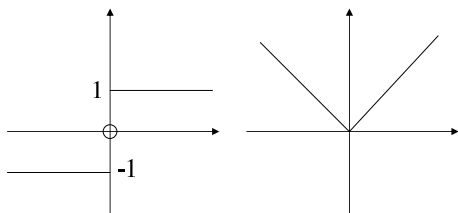


Abb. 1.25: Signum- und Betragsfunktion

Unter Verwendung der Signumfunktion lässt sich die Betragsfunktion schreiben in der Form

$$|x| = x \cdot \text{sign}(x).$$

Beispiel 1.72

Die Funktion $y = f(x) = |x^2 - 4|$ besitzt den Definitionsbereich $D(f) = \mathbb{R}$ und den Wertebereich $W(f) = [0, \infty[$. Zur Darstellung der Funktion betrachten wir zwei Fälle:

- Fall 1: $x^2 - 4 \geq 0$, d. h. $x^2 \geq 4$, also $x \in]-\infty, -2] \cup [2, \infty[$. Dort ist $f(x) = x^2 - 4$.
- Fall 2: $x^2 - 4 < 0$, d. h. $x^2 < 4$, also $x \in]-2, 2[$. Dort ist $f(x) = -x^2 + 4$.

Insgesamt gilt

$$f(x) = \begin{cases} x^2 - 4, & x \in]-\infty, -2] \cup [2, \infty[, \\ 4 - x^2, & x \in]-2, 2[. \end{cases}$$

■

1.5.6 Polynome und gebrochen-rationale Funktionen

Polynome und gebrochen-rationale Funktionen kennen Sie sicher schon aus der Schule. Funktionen dieses Typs lassen sich mit Hilfe endlich vieler rationaler Operationen (Addition, Subtraktion, Multiplikation und Division) berechnen und bilden damit die Klasse der **rationalen Funktionen**.

1.5.6.1 Polynome

Bei der Betrachtung von (Un-)Gleichungen haben wir bereits mit Polynomen gerechnet – auch wenn wir sie dort noch nicht so genannt haben. Später werden wir sehen, dass sich viele Funktionen durch Polynome approximieren lassen (vgl. Satz von Taylor auf Seite 384).

Definition 1.31 (Polynome)

Funktionen der Form $f(x) = x^n$ für ein $n \in \mathbb{N}_0$ nennt man **Monome**. Dabei verwenden wir $x^0 = 1$ (auch für $x = 0$). Aus Monomen durch Addition und Multiplikation mit Zahlen (Linearkombinationen) zusammengesetzte Funktionen heißen **Polynome** oder **ganzrationale Funktionen**. Der höchste tatsächlich auftretende Exponent heißt **Grad** des Polynoms. Die Menge aller Polynome mit Grad kleiner oder gleich n bezeichnen wir mit $\mathbb{P}_n$. Ein Polynom $p_n(x)$ vom Grad n hat die allgemeine Darstellung

$$p_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = \sum_{k=0}^n a_kx^k.$$

Dabei nennt man die Zahlen $a_k \in \mathbb{R}$ **Koeffizienten**. Ist der **Leitkoeffizient** $a_n = 1$, so spricht man von einem **normierten Polynom**.

Ganzrationale Gleichungen haben die Gestalt $p_n(x) = 0$ (vgl. Kapitel 1.4.4.1).

Beispiel 1.73

$f(x) = x^2 + 1$ hat Grad 2, $f(x) = x^3 + 2x^2 + 7x + 4$ hat Grad 3, $f(x) = 17$ hat Grad 0, $f(x) = (x + 2)^2$ hat Grad 2. ■

Addiert man zwei Polynome vom Grad kleiner oder gleich n , so entsteht wieder ein Polynom vom Grad kleiner oder gleich n . Multipliziert man zwei Polynome vom Grad

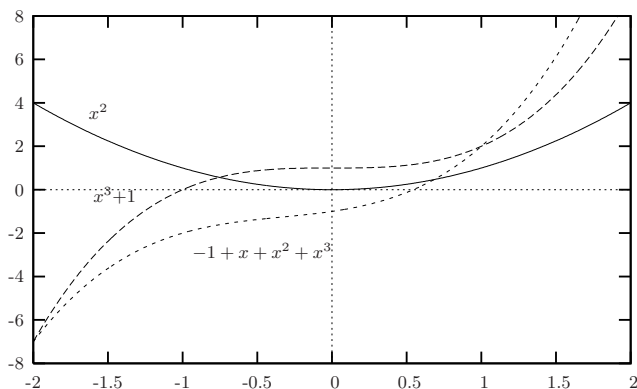


Abb. 1.26: Polynome $x^2, x^3 + 1, -1 + x + x^2 + x^3$

n und m , so entsteht ein Polynom vom Grad $n + m$. Die Menge aller Polynome bildet einen Ring (siehe Kasten auf Seite 46).

Ein wichtiger Spezialfall sind Polynome vom Grad 1: $p(x) = mx + b$. Der Funktionsgraph ist eine **Gerade**, die die y -Achse im Punkt $(0, b)$ und die x -Achse in $(-b/m, 0)$ schneidet (falls $m \neq 0$). Die Steigung (= Höhenzuwachs durch Horizontaländerung) der Geraden ist m (siehe Abbildung 1.27).

Beispiel 1.74

Wird an einer Straße eine Steigung von 30 % angegeben, dann steigt die Straße 30 Meter auf 100 Meter. Hier ist $m = 0,3$. Eine Steigung von 100 % entspricht $m = 1$ und einem Anstiegswinkel von 45 Grad. ■

Berechnung einer Geradengleichung:

- Hat man zwei verschiedene Punkte (x_1, y_1) und (x_2, y_2) , so gibt es genau eine Gerade, die durch diese beiden Punkte geht. Falls $x_1 \neq x_2$, ist diese Gerade über eine Funktion darstellbar:

$$y = \underbrace{\frac{y_2 - y_1}{x_2 - x_1}}_m x + \underbrace{y_1 - x_1 \frac{y_2 - y_1}{x_2 - x_1}}_b.$$

Diese Darstellung nennt man die **Zwei-Punkte-Form** der Geradengleichung. Dass die damit beschriebene Gerade tatsächlich durch die Punkte (x_1, y_1) und (x_2, y_2) geht, sieht man, indem man x_1 und x_2 einsetzt.

- Hat man einen Punkt (x_1, y_1) und eine Steigung m , so gibt es genau eine Gerade mit dieser Steigung durch den Punkt:

$$y = mx + \underbrace{y_1 - x_1 m}_b.$$

Diese Darstellung nennt man die **Punkt-Steigungsform** der Geradengleichung. Der Wert für b ergibt sich aus der Gleichung $y_1 = mx_1 + b$.

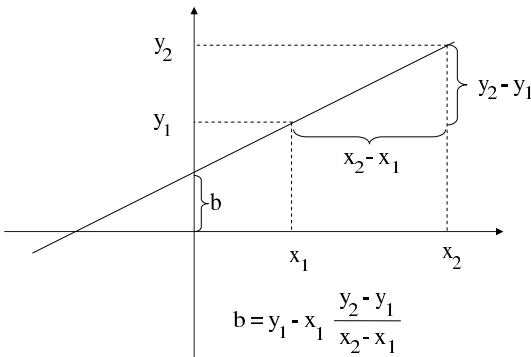


Abb. 1.27: Gerade mit Steigung $m = \frac{y_2 - y_1}{x_2 - x_1}$

1.5.6.2 Interpolation

Gibt man zwei Punkte mit unterschiedlichem x -Wert vor, so findet man genau eine Gerade, also ein Polynom vom maximalen Grad $2 - 1 = 1$, dessen Graph durch beide Punkte geht. Das lässt sich verallgemeinern: In der Ingenieurpraxis sind häufig Datenpaare (x_k, y_k) durch Messungen gegeben. Gibt man so n Punkte mit unterschiedlicher erster Komponente x_k vor, so gibt es genau ein Polynom mit maximalem Grad $n - 1$, dessen Graph durch alle Punkte geht. (Das kann man über ein Gleichungssystem für die Koeffizienten nachrechnen, mit Gleichungssystemen beschäftigen wir uns in Kapitel 1.7.1.) Man spricht von einem **Interpolationspolynom**.

Beispiel 1.75

Wir betrachten die vier Datenpaare (x_k, y_k) :

k	0	1	2	3
x_k	0	2	3	4
y_k	48	0	3	8

Gesucht ist das eindeutige Interpolationspolynom $p_3(x)$ vom Grad höchstens 3, das an den Stellen x_0, x_1, x_2 und x_3 die Werte y_0, y_1, y_2 und y_3 annimmt (also $p_3(x_k) = y_k$, für $k = 0, 1, 2, 3$). Der Graph von p_3 soll also durch die vorgegebenen vier Punkte verlaufen.

Zur Lösung der Aufgabe konstruieren wir zu den Stellen $x_0, \dots, x_3$ spezielle Polynome, die an jeweils genau einer der Stellen den Wert 1 und an den anderen den Wert 0 annehmen:

$$q_0(x) = \frac{(x-2) \cdot (x-3) \cdot (x-4)}{(0-2) \cdot (0-3) \cdot (0-4)}, \text{ so dass } q_0(0) = 1, q_0(x_k) = 0, k \neq 0,$$

$$q_1(x) = \frac{x \cdot (x-3) \cdot (x-4)}{(2-0) \cdot (2-3) \cdot (2-4)}, \text{ so dass } q_1(2) = 1, q_1(x_k) = 0, k \neq 1,$$

$$q_2(x) = \frac{x \cdot (x-2) \cdot (x-4)}{(3-0) \cdot (3-2) \cdot (3-4)}, \text{ so dass } q_2(3) = 1, q_2(x_k) = 0, k \neq 2,$$

$$q_3(x) = \frac{x \cdot (x-2) \cdot (x-3)}{(4-0) \cdot (4-2) \cdot (4-3)}, \text{ so dass } q_3(4) = 1, q_3(x_k) = 0, k \neq 3,$$

oder kurz

$$q_i(x_k) = \delta_i^k \quad \text{mit} \quad \delta_i^k := \begin{cases} 1, & k = i, \\ 0, & k \neq i. \end{cases}$$

Das Polynom $q_i(x)$ heißt i -tes **Lagrange'sches Knotenpolynom**, es nimmt in x_i den Wert 1 und an den anderen Stellen den Wert 0 an. Mit den Knotenpolynomen bilden wir das **Lagrange'sche Interpolationspolynom**

$$p_3(x) = \sum_{i=0}^3 y_i \cdot q_i(x).$$

Es gilt

$$p_3(x_k) = \sum_{i=0}^3 y_i \cdot \underbrace{q_i(x_k)}_{=\delta_i^k} = y_k, \quad k = 0, 1, 2, 3,$$

d. h., $p_3(x)$ **interpoliert** die gegebenen Daten. Speziell für die angegebenen Werte erhält man

$$\begin{aligned} p_3(x) &= 48 \cdot \frac{(x-2)(x-3)(x-4)}{-24} + 0 + 3 \cdot \frac{x(x-2)(x-4)}{-3} + 8 \cdot \frac{x(x-2)(x-3)}{8} \\ &= (x-2) [(-2x^2 + 14x - 24) - (x^2 - 4x) + (x^2 - 3x)] \\ &= (x-2)[-2x^2 + 15x - 24] = -2x^3 + 19x^2 - 54x + 48. \end{aligned}$$

Mittels Differenzialrechnung werden wir auf Seite 321 (Satz 2.39) später die Abweichungen zwischen Interpolationspolynom p_n und gegebener Funktion f analysieren. ■

Zur Interpolation großer Datenmengen teilt man die Daten in kleine Einheiten, die man jeweils mit einem Polynom niedrigen Grades n ($n \leq 3$) interpoliert. Diese Polynome setzt man dann zu einem stückweise definierten Interpolationspolynom (**Spline**) zusammen, siehe Beispiel 2.74 auf Seite 312.

1.5.6.3 Faktorzerlegung und Polynomdivision

Polynome lassen sich als Produkt einfacher Faktoren schreiben, so dass man an dieser Darstellung sofort ihre Nullstellen ablesen kann.

Satz 1.17 (Faktorzerlegung reeller Polynome)

Ein normiertes Polynom $p_n(z)$ mit reellen Koeffizienten kann stets in ein Produkt aus Linearfaktoren $(x - x_k)$ und/oder quadratischen Faktoren $(x - \alpha)^2 + \beta$ mit $\beta > 0$ zerlegt werden. Die Faktoren sind bis auf ihre Reihenfolge eindeutig. Die Zahlen x_k sind genau die reellen Nullstellen des Polynoms. Haben genau m Linearfaktoren den gleichen Zahlenwert x_k , so nennt man x_k eine **m -fache Nullstelle** von p_n .

Dieser Satz ist eine direkte Folgerung aus einer allgemeineren Aussage für komplexe Polynome und wird so später auf Seite 149 bewiesen. Die hier etwas unschönen quadratischen Faktoren werden unter Verwendung komplexer Zahlen ebenfalls zu Produkten von Linearfaktoren. In dieser Form heißt der Satz der Fundamentalsatz der Algebra (siehe Satz 1.24 auf Seite 147). Dieser Name unterstreicht die ganz außergewöhnliche Bedeutung des Satzes, den wir vielfältig anwenden werden, z. B. direkt für die nächste Aussage:

Satz 1.18 (Gleichheit von Polynomen)

Seien $f(x)$ und $g(x)$ Polynome (mit $D(f) = D(g) = \mathbb{R}$). Diese sind gleich genau dann, wenn ihr Grad gleich ist und alle Koeffizienten übereinstimmen.

Beweis: Bei gleichem Grad und gleichen Koeffizienten stimmen natürlich die Funktionswerte überein. Umgekehrt nehmen wir an, dass $f(x)$ und $g(x)$ gleiche Polynome mit unterschiedlichem Grad oder unterschiedlichen Koeffizienten sind. Gleichheit bedeutet, dass alle Funktionswerte $f(x) = g(x)$ für $x \in \mathbb{R}$ gleich sind. $f(x) - g(x)$ ist nun ein Polynom, das Koeffizienten ungleich null und damit einen Grad größer 0 hat. Zugleich sind alle Funktionswerte null. Aber wegen des Satzes 1.17 hat ein Polynom mit Grad größer 0 nur endlich viele Nullstellen x_k , wird also nicht überall 0. Aufgrund des Widerspruchs muss der Satz gelten. $\square$

Statt unendlich viele Funktionswerte auf Gleichheit untersuchen zu müssen, können wir nun wenige Koeffizienten vergleichen. Man nutzt dies häufig aus, um unbekannte Parameter zu bestimmen (**Koeffizientenvergleich**). Weiß man z. B., dass $ax^2 + bx + c = 4x^2 + 2x$, dann ist $a = 4$, $b = 2$ und $c = 0$.

Bei Polynomen bis zum Grad 2 kann man Nullstellen (z. B. mittels p-q-Formel) ohne Probleme berechnen. Für den Grad 3 und 4 ist das mit komplizierteren Formeln auch noch möglich. Für einen Grad ab 5 ist im Allgemeinen eine exakte Berechnung gar nicht mehr möglich. Wie man Nullstellen bei Polynomen dritten Grades berechnen kann, finden Sie nach der Einführung der komplexen Zahlen als Hintergrundinformation auf Seite 150.

In Klausuren ist die Welt oft 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

$$0 = p(x_0) = a_0 + a_1x_0 + a_2x_0^2 + \cdots + a_nx_0^n \implies a_0 = x_0(-a_1 - a_2x_0 - \cdots - a_nx_0^{n-1}).$$

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

Beispiel 1.76

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 . ■

In Anwendungssituationen kann man sich bei Polynomen $p(x)$ höherer Ordnung n damit behelfen, zunächst eine Nullstelle x_0 am Funktionsgraphen abzulesen oder über ein Näherungsverfahren (siehe Anmerkung zu Satz 2.25 auf Seite 278 und das Newton-Verfahren (2.23) auf Seite 306) auszurechnen. Hat man eine Nullstelle x_0 gefunden, so lässt sich das Polynom wegen Satz 1.17 schreiben als $p(x) = (x - x_0)s(x)$, wobei $s(x)$ ein Polynom vom Grad $n - 1$ ist. Die restlichen Nullstellen gehören nun zum Polynom $s(x)$. Ist beispielsweise $n = 3$, so hat $s(x)$ den Grad 2, und man kann die weiteren Nullstellen mittels p-q-Formel bestimmen.

Man erhält $s(x)$, indem man $\frac{p(x)}{x-x_0}$ berechnet. Dazu wendet man die **Polynomdivision** an. Allgemein betrachtet man dabei gebrochen-rationale Funktionen $p(x)/q(x)$ (siehe Seite 100), wobei der Grad des Zählerpolynoms p größer oder gleich dem Grad des Nennerpolynoms ist. Gesucht ist eine Darstellung

$$\frac{p(x)}{q(x)} = s(x) + \frac{r(x)}{q(x)}, \quad (1.16)$$

wobei s und r Polynome sind. r ist der Rest der Division, dessen Grad kleiner als der Grad von $q(x)$ ist. Ist speziell $q(x) = (x - x_0)$ für eine Nullstelle x_0 von $p(x)$, so ist $r(x) = 0$.

Man erhält (1.16), indem man die Polynome nach absteigenden Potenzen sortiert und darauf den von der schriftlichen Division bekannten Algorithmus anwendet.

Beispiel 1.77

$$\begin{array}{rcl}
 \begin{pmatrix} x^3 & +2x^2 & & -2 \end{pmatrix} & : & (x-1) = x^2 + 3x + 3 + \frac{1}{x-1} \\
 -[\begin{pmatrix} x^3 & & -x^2 & \end{pmatrix}] & & \\
 \hline
 & & 3x^2 & -2 \\
 -[& & 3x^2 & -3x \end{pmatrix}] & & \\
 \hline
 & & & 3x & -2 \\
 -[& & & 3x & -3 \end{pmatrix}] & & \\
 \hline
 & & & & 1
 \end{array}$$

Im ersten Schritt fragt man sich, womit man x multiplizieren muss, um x^3 zu erhalten. Das ist x^2 . Damit:

$$x^3 + 2x^2 - 2 = x^2(x - 1) + \text{Rest}, \quad \text{Rest} = x^3 + 2x^2 - 2 - [x^3 - x^2] = 3x^2 - 2.$$

Im zweiten Schritt teilt man nun den Rest durch $(x - 1)$, erhält $3x$ und den neuen Rest $(3x - 2)$. Division durch $(x - 1)$ liefert 3 und den Rest 1. ■

Wir benutzen jetzt die Polynomdivision, um bei einer gegebenen Nullstelle weitere zu finden.

Beispiel 1.78

- a) Das Polynom 4. Grades $p_4(x) = x^4 - 16$ lässt sich mit der Substitution $u = x^2$ in das quadratische Polynom $u^2 - 16$ mit den Nullstellen 4 und -4 überführen. Damit hat das Ausgangspolynom die Nullstellen $x = \pm 2$. Polynomdivision liefert: $(x^4 - 16) : ((x - 2)(x + 2)) = (x^4 - 16) : (x^2 - 4) = x^2 + 4$. Die Zerlegung von $p_4(x)$ lautet $p_4(x) = (x - 2) \cdot (x + 2) \cdot (x^2 + 4)$.
- b) Das **kubische** (Grad $n = 3$) Polynom $p_3(x) = x^3 - 8$ hat die reelle Nullstelle $x = 2$. Abspalten des Linearfaktors $(x - 2)$ durch Polynomdivision liefert:

$$\begin{array}{r}
 \begin{array}{rcccc}
 (& x^3 & & & -8 &) \\
 -[& x^3 & -2x^2 & & &] \\
 \hline
 & & 2x^2 & & -8 & \\
 -[& & 2x^2 & -4x & &] \\
 \hline
 & & & 4x & -8 & \\
 -[& & & 4x & -8 &] \\
 \hline
 & & & & 0 &
 \end{array}
 \quad : \quad (x - 2) = x^2 + 2x + 4.
 \end{array}$$

Der Term $x^2 + 2x + 4$ besitzt keine (reelle) Nullstelle, denn es gilt

$$x^2 + 2x + 4 = (x + 1)^2 + 3 \neq 0, \quad x \in \mathbb{R}.$$

Wir erhalten die reelle Zerlegung $p_3(x) = (x - 2) \cdot ((x + 1)^2 + 3)$.

- c) Das Polynom $p_3(x) = x^3 - 6x^2 + 11x - 6$ hat die Nullstelle $x_1 = 1$. Polynomdivision liefert $x^3 - 6x^2 + 11x - 6 : (x - 1) = x^2 - 5x + 6$, und man erhält die Faktordarstellung $p_3(x) = (x - 1) \cdot (x^2 - 5x + 6)$. Das Polynom $s_2(x) = x^2 - 5x + 6$ hat die Nullstellen

$$x_{2,3} = \frac{5}{2} \pm \sqrt{\frac{25}{4} - 6} = \frac{5}{2} \pm \frac{1}{2}, \quad x_2 = 3, \quad x_3 = 2.$$

Also lautet die Faktorzerlegung: $p_3(x) = x^3 - 6x^2 + 11x - 6 = (x - 1) \cdot (x - 2) \cdot (x - 3)$. ■

Die Nullstellen sind eindeutig durch die Koeffizienten des Polynoms bestimmt. Umgekehrt lassen sich die Koeffizienten aus den Nullstellen berechnen. Das ist der Satz von Vieta, für den wir ein normiertes Polynom

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0, \quad a_k \in \mathbb{R},$$

mit den Nullstellen $x_1, x_2, \dots, x_n \in \mathbb{R}$ betrachten. Nach Satz 1.17 gilt:

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = (x - x_1)(x - x_2) \cdots (x - x_n). \quad (1.17)$$

Satz 1.19 (Wurzelsatz von Vieta (1540–1603))

Es seien x_k , $k = 1, 2, \dots, n$, die Nullstellen des normierten Polynoms

$$p_n(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0.$$

Dann gelten die Gleichungen

$$\begin{aligned} x_1 + x_2 + \cdots + x_n &= (-1)^1 a_{n-1} \\ x_1x_2 + x_1x_3 + \cdots + x_1x_n + x_2x_3 + \cdots + x_2x_n + \cdots + x_{n-1}x_n &= (-1)^2 a_{n-2} \\ &\vdots \\ x_1x_2 \cdots x_n &= (-1)^n a_0. \end{aligned}$$

Beweis: Der Beweis besteht darin, die rechte Seite von (1.17) auszumultiplizieren und dann die Koeffizienten zu vergleichen. $\square$

1.5.6.4 Horner-Schema

Das **Horner-Schema** dient der effizienten Berechnung von Funktionswerten eines Polynoms. Es reduziert die Anzahl der dabei vorzunehmenden Multiplikationen:

Beispiel 1.79

$p(x) := 1 + 2x + 3x^2 + 2x^3$. Wir möchten $p(2)$ berechnen. Naheliegend ist dieser Ansatz: $x = 2$, $x^2 = 4$ (erste Multiplikation), $x^3 = 8$ (zweite Multiplikation), $2x = 4$ (dritte Multiplikation), $3x^2 = 12$ (vierte Multiplikation), $2x^3 = 16$ (fünfte Multiplikation). Dann müssen nur noch Zwischenergebnisse addiert werden: $p(2) = 33$. Wir können aber auch geschickt die Variable x ausklammern:

$$p(x) = 2x^3 + 3x^2 + 2x + 1 = [(2x + 3)x + 2]x + 1.$$

In dieser Form benötigt man nur noch drei Multiplikationen! $\blacksquare$

Wir verallgemeinern diese Beobachtung und erhalten das Horner-Schema: Gegeben sei das Polynom

$$\begin{aligned} p(x) &= a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \cdots + a_1 x + a_0 \\ &= (a_n x + a_{n-1}) x^{n-1} + a_{n-2} x^{n-2} + \cdots + a_1 x + a_0 = \dots \\ &= (\dots ((a_n x + a_{n-1}) x + a_{n-2}) x + \cdots + a_1) x + a_0. \end{aligned}$$

Um $p(x_0)$ für ein x_0 auszurechnen, wertet man nun die Klammern von innen nach außen aus. Die Zwischenergebnisse bezeichnen wir mit b_k , $0 \leq k \leq n-1$:

$$p(x_0) = (\dots (\underbrace{\underbrace{\underbrace{(a_n x_0 + a_{n-1})}_{b_{n-1}} x + a_{n-2}}_{b_{n-2}}}_{b_{n-3}} x + \cdots + a_1) x + a_0.$$

Damit ergibt sich das Rechenschema

zu	a_n	a_{n-1}	a_{n-2}	$\dots$	a_2	a_1	a_0	
addiere	0	$b_{n-1} \cdot x_0$	$b_{n-2} \cdot x_0$	$\dots$	$b_2 \cdot x_0$	$b_1 \cdot x_0$	$b_0 \cdot x_0$	(1.18)
erhalte	b_{n-1}	b_{n-2}	b_{n-3}	$\dots$	b_1	b_0	$p(x_0)$	

Beispiel 1.80

Für $x_0 = 2$ und $p(x) = 2x^3 + 3x^2 + 2x + 1$ erhalten wir $p(2)$ über

zu	2	3	2	1
addiere	0	$2 \cdot 2$	$7 \cdot 2$	$16 \cdot 2$
erhalte	2	7	16	33.

■

Achtung: Im Horner-Schema müssen auch Koeffizienten des Polynoms berücksichtigt werden, die null sind. Wir berechnen den Funktionswert an der Stelle $x = 2$ von $x^2 + 2x + 0 = [1x + 2]x + 0$:

zu	1	2	0
addiere	0	$1 \cdot 2$	$4 \cdot 2$
erhalte	1	4	8.

Hätten wir die letzte Spalte weggelassen, so hätten wir den falschen Wert 4 erhalten.

Zahlen, die mit einer anderen Basis als 10 dargestellt sind (siehe Kapitel 1.3.1.2), lassen sich mit dem Horner-Schema elegant in das Zehnersystem überführen. Sind $a_n a_{n-1} \dots a_1 a_0$ die Ziffern der Zahl zur Basis b , so muss für die Umrechnung die Summe

$$a_n b^n + a_{n-1} b^{n-1} + \cdots + a_2 b^2 + a_1 b + a_0 = p(b)$$

mit $p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x + a_0$ berechnet werden. Dies ist also nichts anderes als die Berechnung eines Funktionswerts eines Polynoms für $x = b$.

Beispiel 1.81

Wir stellen die Hexadezimalzahl $(4e20b)_{16}$ im Dezimalsystem dar:

zu	4	14	2	0	11
addiere	0	64	1248	20 000	320 000
erhalte	4	78	1250	20 000	320 011

Also gilt: $(4e20b)_{16} = (320\,011)_{10}$. ■

Das Horner-Schema kann nicht nur zum Ausrechnen von Funktionswerten, sondern auch zur Polynomdivision genutzt werden. Mit den Bezeichnungen aus (1.18) ist

$$\begin{aligned}
 p(x) &= a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \dots + a_1 x + a_0 \\
 &= \underbrace{(b_{n-1} x^{n-1} + b_{n-2} x^{n-2} + \dots + b_1 x + b_0)}_{=: q(x)} (x - x_0) + p(x_0).
 \end{aligned}$$

Diese Gleichheit rechnen wir nach:

$$\begin{aligned}
 &(b_{n-1} x^{n-1} + b_{n-2} x^{n-2} + \dots + b_1 x + b_0)(x - x_0) + p(x_0) \\
 &= b_{n-1} x^n + (-x_0 b_{n-1} + b_{n-2}) x^{n-1} + (-x_0 b_{n-2} + b_{n-3}) x^{n-2} + \dots \\
 &\quad + (-x_0 b_0 + p(x_0)) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \dots + a_0,
 \end{aligned}$$

dabei haben wir im letzten Schritt das Horner-Schema (1.18) verwendet.

Wir können das Ergebnis einer Polynomdivision durch $(x - x_0)$ direkt im Horner-Schema ablesen:

$$\frac{p(x)}{x - x_0} = b_{n-1} x^{n-1} + b_{n-2} x^{n-2} + \dots + b_1 x + b_0 + \frac{p(x_0)}{x - x_0}.$$

Ist x_0 eine Nullstelle des Polynoms, also $p(x_0) = 0$, so entfällt der Rest $\frac{p(x_0)}{x - x_0}$.

Beispiel 1.82

a) Für $x_0 = -1$ und $p(x) = 2x^3 + 3x^2 + 2x + 1$ ergibt das Horner-Schema:

zu	2	3	2	1
addiere	0	$2 \cdot (-1)$	$1 \cdot (-1)$	$1 \cdot (-1)$
erhalte	2	1	1	0.

Damit ist $\frac{p(x)}{x + 1} = 2x^2 + x + 1$.

b) Ist x_0 dagegen keine Nullstelle des Polynoms, so ergibt sich ein Rest bei der Division durch $(x - x_0)$. Wir betrachten $x_0 = 0$ und $p(x) = x^3 + 2x + 1$ und erhalten

zu	1	0	2	1
addiere	0	$1 \cdot 0$	$0 \cdot 0$	$2 \cdot 0$
erhalte	1	0	2	1.

Damit ist $\frac{p(x)}{x - 0} = x^2 + 2 + \frac{1}{x - 0}$.

■

1.5.6.5 Gebrochen-rationale Funktionen

Definition 1.32 (Gebrochen-rationale Funktionen)

Eine **gebrochen-rationale Funktion** ist der Quotient zweier Polynome

$$f(x) = \frac{q_m(x)}{p_n(x)} = \frac{b_m x^m + b_{m-1} x^{m-1} + \dots + b_1 x + b_0}{a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0}.$$

Im Fall $m < n$ spricht man von einer **echt gebrochen-rationale Funktion** und im Fall $m \geq n$ von einer **unecht gebrochen-rationale Funktion**.

Sucht man Nullstellen von gebrochen-rationale Funktionen, so genügt es, das Zählerpolynom $q_m(x)$ zu betrachten. Die Nullstellen des Nennerpolynoms $p_n(x)$ hingegen dürfen nicht im Definitionsbereich liegen. Man bezeichnet sie als **kritische Stellen** von f . Es gilt $D(f) = \mathbb{R} \setminus \{x : p_n(x) = 0\}$. Wir werden später kritische Stellen mit

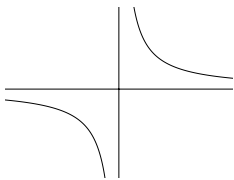


Abb. 1.28: Die **Kehrwertfunktion** $f(x) = \frac{1}{x}$ ist auf $\mathbb{R} \setminus \{0\}$ definiert. Sie hat die kritische Stelle 0. In Prüfungen wird diese Funktion oft fälschlich für eine Gerade gehalten, da die Variable vermeintlich den Exponenten 1 hat. Tatsächlich ist er aber -1 .

Hilfe von Grenzwerten weiter klassifizieren (siehe Definition 2.21 auf Seite 285).

Jede unecht gebrochen-rationale Funktion $f(x)$ kann mit Hilfe der Polynomdivision in die Summe einer ganzrationalen und einer echt gebrochen-rationale Funktion zerlegt werden. Beispielsweise erhält man mittels Polynomdivision die Zerlegung

$$\frac{x^3 - 4x + 6}{x - 2} = x^2 + 2x + \frac{6}{x - 2}.$$

Diese Vorgehensweise werden wir später im Rahmen der Integralrechnung bei der Partialbruchzerlegung verwenden (siehe Kapitel 2.6.4.3).

1.5.7 Potenz- und Wurzelfunktionen

Die Klasse der **algebraischen Funktionen** umfasst die rationalen Funktionen, also die Polynome und die gebrochen-rationalen Funktionen. Neu kommen die Wurzelfunktionen und die aus Wurzeln und rationalen Funktionen gebildeten Funktionen hinzu.

Wir erinnern uns, dass wir nach den Überlegungen von Seite 57 mit beliebigen reellen Exponenten rechnen können. Außerdem haben wir uns in Kapitel 1.4.1 bereits intensiv mit Potenzen und Wurzeln beschäftigt.

Definition 1.33 (Potenz- und Wurzelfunktion)

Eine Funktion f mit

$$f(x) = x^a$$

und $a \in \mathbb{R} \setminus \{0\}$ heißt **Potenzfunktion**. Es ist $D(f) =]0, \infty[$ und $W(f) =]0, \infty[$. Jede Potenzfunktion ist streng monoton auf $D(f)$, und somit existiert die Umkehrfunktion

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

Für eine natürliche Zahl $a = n$, $n \in \mathbb{N}$ wird die Umkehrfunktion als **Wurzelfunktion** bzw. als **n -te Wurzel aus x** bezeichnet.

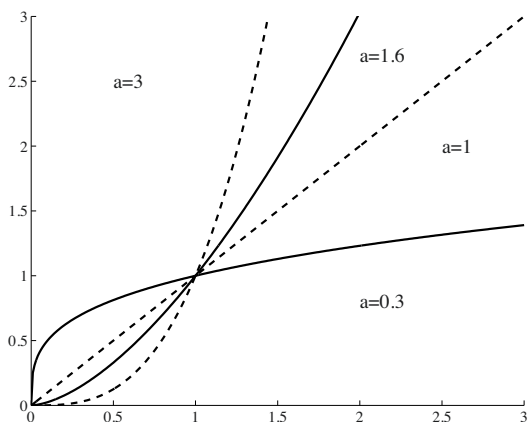


Abb. 1.29: Potenzfunktionen $y = x^a$ für verschiedene Exponenten a

- Für $a > 0$ ist $f(x) = x^a$ auch für $x = 0$ definiert und für $x \in D(f) = [0, \infty[$, $y \in W(f) = [0, \infty[$ streng monoton steigend.
- Für $a < 0$ ist $f(x) = x^a$ für $x \in D(f) =]0, \infty[$ streng monoton fallend und strebt für $x \rightarrow \infty$ gegen 0.

Für Potenzfunktionen mit ganzzahligen Exponenten gilt:

- Ist n gerade, so ist $f(x) = x^n$ gerade, da gradzahlig oft die ungerade Funktion x multipliziert wird (siehe Lemma 1.6 auf Seite 82).
- Ist n ungerade, so ist $f(x) = x^n$ ungerade, da eine ungerade Funktion ungradzahlig oft multipliziert wird.

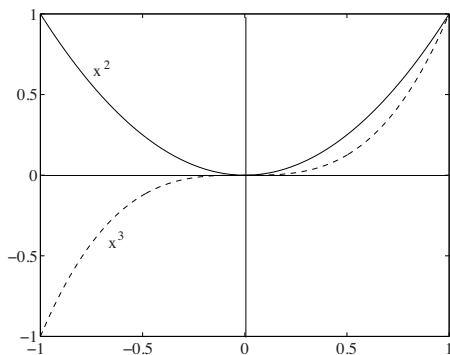


Abb. 1.30: Gerade und ungerade Potenzfunktionen

1.5.8 Exponentialfunktionen und Logarithmen

Funktionen, die nicht zur Klasse der algebraischen Funktionen gehören und damit nicht mittels der Grundrechenarten einschließlich Wurzeln berechnet werden können, heißen **transzendente Funktionen**. Dazu gehören die Exponentialfunktionen und ihre Umkehrfunktionen. Da sie allgemeine Wachstums- und Zerfallsprozesse beschreiben, trifft man sie überall in Naturwissenschaft und Technik an.

Beispiel 1.83

In Krefeld leben ca. 240 000 Menschen. Die Einwohnerzahl ist leicht rückläufig. Nach wie vielen Jahren wird die Einwohnerzahl nur noch 200 000 betragen, wenn die Bevölkerung jährlich um 1,5 % schrumpft? Nach einem Jahr leben nur noch 0,985-mal so viele Einwohner dort, ein weiteres Jahr später $(0,985)^2$ -mal so viele, und nach n Jahren ist die Population um den Faktor $(0,985)^n$ geschrumpft. Um die Gleichung $(0,985)^n \cdot 240\,000 = 200\,000$ nach n aufzulösen, beschäftigen wir uns mit der Exponentialfunktion und insbesondere ihrer Umkehrfunktion, dem natürlichen Logarithmus.

1.5.8.1 Exponentialfunktion und natürlicher Logarithmus

Bei den Polynomen und Potenzfunktionen waren die Exponenten unabhängig von der Variable. Das ändert sich jetzt:

Definition 1.34 (Exponentialfunktion)

Die Funktion

$$y = \exp(x) := e^x$$

mit der Basis $e = 2,7182818\dots$ (siehe (1.8) auf Seite 52) wird als **Exponentialfunktion** oder **e-Funktion** bezeichnet. Sie besitzt den Definitionsbereich $D = \mathbb{R}$ und den Wertebereich $W =]0, \infty[$.

Die Exponential- oder e-Funktion $\exp$ ist also die Funktion, die jedem $x \in \mathbb{R}$ die reelle Zahl e^x zuordnet. Warum gerade die komplizierte Basis e eine große Rolle spielt, wird erst im Rahmen der Differenzialrechnung verständlich, siehe Kapitel 3. Anschaulicher sind vielleicht die Basen 2 oder 10, die wir im Anschluss diskutieren und die sich beim Rechnen kaum von der Basis e unterscheiden.

Mathematiker sind schreibfaul und lassen bisweilen die Klammern um das Argument einer Funktion weg, wenn weiterhin eindeutig ist, was der Funktion übergeben wird, also z. B. $\exp x = \exp(x)$.

Beispiel 1.84

Wird ein Kondensator mit Kapazität C über einen Widerstand R entladen, so ergibt sich der Spannungsverlauf am Kondensator zu $u(t) = u(0) \exp\left(-\frac{t}{RC}\right)$. ■

Die Potenzregeln werden direkt zu den folgenden Rechenregeln:

$$\begin{aligned} \exp(x+y) &= \exp(x) \exp(y), & \exp(xy) &= [\exp(x)]^y & \text{und} & \exp(-x) = \frac{1}{\exp(x)}, \\ \exp(0) &= 1, & \exp(1) &= e. \end{aligned}$$

Beweis: Mit der Definition rechnet man dies unmittelbar nach: $\exp(x+y) = e^{x+y} = e^x e^y = \exp(x) \exp(y)$, $\exp(xy) = e^{xy} = [e^x]^y = [\exp(x)]^y$, sowie $\exp(-x) = e^{-x} = \frac{1}{e^x} = \frac{1}{\exp(x)}$. Einsetzen liefert $\exp(0) = e^0 = 1$, $\exp(1) = e^1 = e$. □

Man kann zeigen, dass $\exp : \mathbb{R} \rightarrow]0, \infty[$ streng monoton wächst und somit eine bijektive Abbildung ist, d. h., man kann die Gleichung $y = \exp(x)$ für jedes $y > 0$ nach x auflösen. Damit hat die Exponentialfunktion eine Umkehrfunktion, die nach Lemma 1.7 ebenfalls streng monoton wachsend ist:

Definition 1.35 (Natürlicher Logarithmus)

Die Umkehrfunktion der Exponentialfunktion ist der **natürliche Logarithmus** $\ln :]0, \infty[\rightarrow \mathbb{R}$. Es gilt

$$\ln(\exp(x)) = x, \text{ für } x \in \mathbb{R}, \quad \exp(\ln(x)) = x, \text{ für } x \in]0, \infty[.$$

Statt mit $\ln$ wird der natürliche Logarithmus auch mit $\log$ benannt. Die Funktion $f(x) := \exp(x)$ hat die Umkehrfunktion $f^{-1}(x) = \ln(x)$. Der Exponent -1 könnte suggerieren, dass dies gleich $\frac{1}{\exp(x)} = e^{-x}$ ist. Dem ist aber **nicht** so: $\ln(x)$ und e^{-x} sind grundverschiedene Funktionen. f^{-1} ist, wie bereits bei der Definition der Umkehrabbildung bemerkt, lediglich eine Schreibweise für die Umkehrfunktion und hat nichts mit einem Kehrwert zu tun.

Rechenregeln für den Logarithmus:

$$\ln(x) + \ln(y) = \ln(xy), \quad \ln(x) - \ln(y) = \ln\left(\frac{x}{y}\right).$$

Zum Nachweis der Regeln nutzen wir aus, dass die Exponentialfunktion bijektiv und damit injektiv ist. Sind also zwei Funktionswerte $\exp(a)$ und $\exp(b)$ der Exponentialfunktion gleich, so müssen auch die zugehörigen Argumente a und b gleich sein. Es genügt also, die beiden Gleichungen in einer Form nachzurechnen, bei der auf beiden Seiten die Exponentialfunktion angewendet ist:

$$\begin{aligned} \exp(\ln(x) + \ln(y)) &= \exp(\ln(x)) \exp(\ln(y)) = xy = \exp(\ln(xy)), \\ \exp(\ln(x) - \ln(y)) &= \frac{\exp(\ln(x))}{\exp(\ln(y))} = \frac{x}{y} = \exp\left(\ln\left(\frac{x}{y}\right)\right). \end{aligned}$$

Weitere Eigenschaften des Logarithmus:

$$\ln(1) = 0, \quad \ln(e) = 1, \quad -\ln(x) = \ln\left(\frac{1}{x}\right), \quad \ln(x^y) = y \ln(x),$$

denn $-\ln(x) = \ln(1) - \ln(x)$ und $\exp(\ln(x^y)) = x^y = [\exp(\ln(x))]^y = \exp(y \ln(x))$.

In Abbildung 1.31 sind $\exp(x)$ und $\ln(x)$ eingezeichnet. Man erkennt, dass der Funktionsgraph der Umkehrfunktion durch Spiegelung an der Hauptdiagonalen (Punkte mit $x = y$) entsteht (vgl. auch Abbildung 1.23 auf Seite 85).

1.5.8.2 Allgemeine Exponentialfunktionen und Logarithmen

Für jedes $a > 0$ kann man eine **Exponentialfunktion zur Basis a**

$$f : \mathbb{R} \rightarrow]0, \infty[, \quad f(x) := a^x,$$

definieren, deren Umkehrfunktion man als **Logarithmus zur Basis a** bezeichnet:

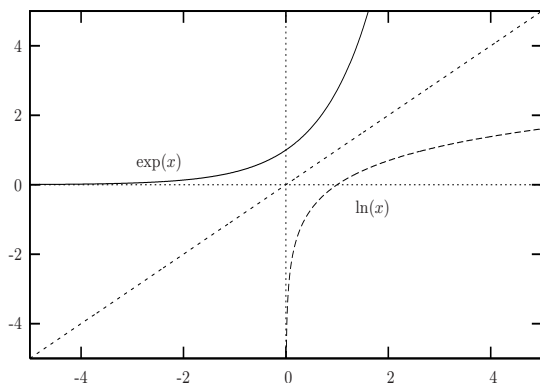


Abb. 1.31: $\exp(x)$ auf $] -\infty, \infty[$ und $\ln(x)$ auf $]0, \infty[$

$$f^{-1}:]0, \infty[\rightarrow \mathbb{R}, \quad f^{-1}(x) := \log_a(x).$$

Wichtig sind die Werte $a = 10$, $a = e$ und $a = 2$:

$$\begin{aligned} \lg(x) &:= \log_{10}(x) && \text{Dekadischer Logarithmus,} \\ \ln(x) &:= \log_e(x) && \text{Natürlicher Logarithmus,} \\ \text{ld}(x) &:= \log_2(x) && \text{Dualer Logarithmus.} \end{aligned}$$

Um mit a^x zu rechnen, führt man diese Funktion auf e^x zurück:

$$a^x = \exp(x \cdot \ln(a)), \quad \log_a(x) = \frac{\ln(x)}{\ln(a)},$$

denn es ist $a^x = [e^{\ln(a)}]^x = e^{x \ln(a)} = \exp(x \ln(a))$. Für $y = \log_a(x)$ gilt nach der Definition des Logarithmus $a^y = x$, und Anwendung des natürlichen Logarithmus auf beide Seiten liefert $\ln(a^y) = y \ln(a) = \ln(x)$ und somit die zweite Beziehung.

Beispiel 1.85 (Rechnen mit verschiedenen Basen)

a) $\lg(27) = \frac{\ln(27)}{\ln(10)} = \frac{3,2958...}{2,3025...} = 1,43136...; \text{ Probe: } 10^{1,43136...} = 27.$

b) $\log_{27}(123) = \frac{\ln(123)}{\ln(27)} = 1,46007...; \text{ Probe: } 27^{1,46007...} = 123.$

c) $\log_{17}(13) = \frac{\ln(13)}{\ln(17)} = 0,9053...; \text{ Probe: } 17^{0,9053...} = 13. \quad \blacksquare$

Durch die Rückführung auf die e -Funktion bzw. den natürlichen Logarithmus übertragen sich deren Rechengesetze und die strenge Monotonie, z. B.:

$$\log_a(x) + \log_a(y) = \log_a(xy),$$

$$\text{denn } \log_a(x) + \log_a(y) = \frac{1}{\ln(a)} [\ln(x) + \ln(y)] = \frac{\ln(xy)}{\ln(a)} = \log_a(xy).$$

Beispiel 1.86

a) $\log_{10}(100\,000) = 5$, denn $10^5 = 100\,000$,

b) $\log_5(125) = 3$, denn $5^3 = 125$,

c) $\log_a(a^x) = x$, denn $a^x = a^x$,

d) $\log_a(a) = 1$, denn $a^1 = a$; $\log_a 1 = 0$, denn $a^0 = 1$,

e) $\frac{2}{5} \cdot \log_a(u) + \log_a(v) = \log_a\left(u^{\frac{2}{5}}\right) + \log_a(v) = \log_a\left(u^{\frac{2}{5}} \cdot v\right),$

f) $\log_a\left(\frac{\sqrt{a} \cdot b \cdot c}{a}\right) = \log_a(\sqrt{a} \cdot b \cdot c) - \log_a(a) = \log_a(\sqrt{a}) + \log_a(b) + \log_a(c) - 1$
 $= \frac{1}{2} \cdot \log_a(a) + \log_a(b) + \log_a(c) - 1 = \log_a(b) + \log_a(c) - \frac{1}{2},$

g) $\log_3\left(\frac{81}{27}\right) = \log_3(81) - \log_3(27) = \log_3(3^4) - \log_3(3^3) = 4 - 3 = 1. \quad \blacksquare$

1.5.8.3 Anwendungen**Beispiel 1.87 (Wachstums- und Zerfallsprozesse)**

Die Exponentialfunktion beschreibt Wachstums- und Zerfallsprozesse. Bezeichnet $N(t)$ eine Population zum Zeitpunkt t , so gilt bei exponentiellem Wachstum/Zerfall die Gleichung

$$N(t) = N_0 \cdot e^{\lambda t}$$

mit der Anfangspopulation N_0 und der Wachstumsrate $\lambda > 0$ oder der Zerfallsrate $\lambda < 0$.

- Wir kommen zum Eingangsbeispiel zurück. Ausgehend von $N_0 = 240\,000$ Einwohnern zum Zeitpunkt $t = 0$ möge die Einwohnerzahl Krefelds (N) jährlich um 1,5 % sinken. Mit der Beziehung für exponentiellen Zerfall erhalten wir

$$\frac{N(t+1)}{N(t)} = \frac{0,985}{1} = \frac{N_0 \cdot e^{\lambda(t+1)}}{N_0 \cdot e^{\lambda t}} = e^{\lambda} \implies \lambda = \ln(0,985) \approx -0,0151.$$

Die Population nach t Jahren lässt sich direkt oder über die Beziehung für den Zerfall angeben:

$$N(t) = 240\,000 \cdot 0,985^t = 240\,000 \cdot e^{\ln(0,985)t} \approx 240\,000 \cdot e^{-0,0151 t}.$$

Nach wie viel Jahren T wird 200 000 unterschritten?

$$N(T) \approx 240\,000 \cdot e^{-0,0151 T} \leq 200\,000 \implies T \geq -\frac{1}{0,0151} \ln\left(\frac{5}{6}\right) \approx 12,07 \text{ Jahre.}$$

- Das Radon-Isotop ^{222}Rn hat eine Halbwertszeit von 3,8 Tagen. Wir ermitteln die Zeitspanne, nach der von 50 g Rn noch 0,5 g (also ein Prozent) übrig ist. Mit dem Zerfallsgesetz $N(t) = N_0 \cdot e^{\lambda t}$ und den angegebenen Werten erhalten wir:

$$N(3,8) = N_0 \cdot e^{\lambda \cdot 3,8} = \frac{1}{2} N_0 \implies 3,8 \cdot \lambda = \ln\left(\frac{1}{2}\right) \implies \lambda \approx -0,182.$$

Damit lautet für dieses Radon-Isotop das Zerfallsgesetz $N(t) = 50 \cdot e^{-0,182 t}$, mit dem wir die gesuchte Zeitspanne berechnen:

$$N(T) \approx 50 \cdot e^{-0,182 T} = 0,5 \implies e^{-0,182 T} = 0,01 \implies T = \frac{\ln(0,01)}{-0,182} \approx 25,3 \text{ Tage.}$$

Die Rechnung ist unabhängig von der konkreten Startmasse: Nach 25,3 Tagen sind von einer beliebigen Ausgangsmenge des Isotops nur noch ein Prozent übrig. Der Rest ist durch Strahlung verloren gegangen. ■

Beispiel 1.88 (Barometrische Höhenformel)

Die barometrische Höhenformel beschreibt den Luftdruck p in Abhängigkeit der Höhe h (über Meeresniveau $h = 0$):

$$p(h) = p_0 \cdot e^{-h/7\,991}, \quad p_0 = 1,013 \text{ bar.}$$

Der Luftdruck nimmt also mit zunehmender Höhe exponentiell ab. Das kann man sich bei der Konstruktion eines Höhenmessers zunutze machen:

$$p(h) = p_0 \cdot e^{-h/7\,991} \implies \frac{p}{p_0} = e^{-h/7\,991} \implies \ln\left(\frac{p}{p_0}\right) = -\frac{h}{7\,991},$$

und weiter folgt

$$h(p) = -7\,991 \cdot \ln\left(\frac{p}{p_0}\right).$$

Mit einem Wetterballon wird ein Luftdruck von 0,65 bar gemessen. Wie hoch schwebt der Ballon in diesem Moment?

$$h(0,65) = -7\,991 \cdot \ln\left(\frac{0,65}{1,013}\right) \approx 3\,546 \text{ m.}$$

Wie hoch ist der Ballon, wenn nur noch die Hälfte des Luftdrucks p_0 gemessen wird?

$$h\left(\frac{p_0}{2}\right) = -7\,991 \cdot \ln\left(\frac{p_0}{2p_0}\right) = -7\,991 \cdot \ln\frac{1}{2} \approx 5\,538,9 \text{ m.} \quad \blacksquare$$

Beispiel 1.89 (Mindestlaufzeit eines Sortierverfahrens*)

Bei Algorithmen interessieren die Laufzeiten in Abhängigkeit der Größe bzw. Menge der zu verarbeitenden Daten. Als Beispiel sollen n verschiedene Zahlen aufsteigend sortiert werden. Die Sortierung soll durch paarweisen Vergleich geschehen. Es werden also sukzessive Zahlen an zwei Stellen verglichen und in Abhängigkeit des Ergebnisses die Positionen der Zahlen geändert. Dann findet der nächste Vergleich und die nächste Umsortierung statt usw. Wir beginnen mit der unsortierten Auflistung und verfolgen einen Sortieralgorithmus bis zum sortierten Ergebnis. Bei jedem Vergleich können maximal zwei unterschiedliche Positionsreihenfolgen entstehen, wobei der Algorithmus mit einer von beiden weitermacht. Wir haben also nach dem ersten Vergleich die Ausgangspositionen und zwei weitere Positionsfolgen erreicht, insgesamt also $1 + 2$. Jede der beiden neuen Reihenfolgen kann beim zweiten Vergleich zu maximal wieder je zwei neuen Positionsfolgen führen. Nach zwei Vergleichen kann man also maximal $1 + 2 + 4$ verschiedene Positionsanordnungen erzeugen. Nach m paarweisen Vergleichen kann man bis zu $1 + 2 + 4 + \dots + 2^m = \sum_{k=0}^m 2^k$ mögliche Reihenfolgen erhalten. Das sind nach der Formel (1.10) von Seite 60 für die geometrische Summe $\frac{1-2^{m+1}}{1-2} = 2^{m+1} - 1$ Reihenfolgen.

Insgesamt gibt es $n!$ Permutationen (also unterschiedliche Reihenfolgen) der n Positionen (siehe Seite 32). Die Ausgangsreihenfolge lässt sich so wählen, dass jede vorgegebene Permutation als Positionsfolge der geordneten Zahlen auftritt. Um auf die sortierte Reihenfolge zu kommen, müssen also im schlechtesten Fall alle $n!$ Positionsreihenfolgen erreichbar sein, d. h., man benötigt mindestens m Vergleiche mit

$$2^{m+1} - 1 \geq n! \quad \Longleftrightarrow \quad 2^m \geq \frac{n! + 1}{2}.$$

Dabei schätzen wir die Faktultät für gerades n elementar mit $n! = 1 \cdot 2 \cdot 3 \cdots \frac{n}{2} \cdot (\frac{n}{2} + 1) \cdots (\frac{n}{2} + \frac{n}{2}) \geq (\frac{n}{2})^{\frac{n}{2}}$ ganz grob nach unten ab. Wir benutzen den streng monoton wachsenden Logarithmus, um daraus eine Abschätzung für m zu erhalten. Sei dazu $n \geq 4$.

$$\begin{aligned} m &\geq \lg\left(\frac{n! + 1}{2}\right) = \lg(n! + 1) - \lg(2) = \lg(n! + 1) - 1 \geq \lg\left(\left(\frac{n}{2}\right)^{\frac{n}{2}}\right) - 1 \\ &= \frac{n}{2} \lg\left(\frac{n}{2}\right) - 1 = \frac{n}{2} [\lg(n) - 1] - 1 \\ &\stackrel{n \geq 4}{\geq} \frac{n}{2} \left[\lg(n) - \frac{1}{2} \lg(n) \right] - \frac{n}{8} \lg(n) = \frac{1}{8} n \lg(n). \end{aligned}$$

Damit kann kein Sortieralgorithmus im ungünstigsten Fall mit weniger als $c \cdot n \lg(n)$ Vergleichen auskommen, wobei $c > 0$ eine von n unabhängige Konstante ist. Gute Sortieralgorithmen erreichen diese Größenordnung. Viele effiziente Algorithmen für andere Probleme als Sortierung haben ebenfalls $n \lg(n)$ -Laufzeiten, z. B. die schnelle Fourier-Transformation auf Seite 923. ■

Beispiel 1.90 (Dämpfung in Leitungen)

Der Wirkungsgrad η ist in der Energietechnik das Verhältnis von Ausgangsleistung P_2 zu Eingangsleistung P_1 . In der Nachrichtentechnik betrachtet man den umgekehrten Quotienten. Das Dämpfungsmaß a ist hier definiert als

$$a := \log_{10} \left(\frac{P_1}{P_2} \right)$$

und wird z. B. verwendet, um die Dämpfung in einer Leitung anzugeben. Die Einheit von a ist 1 B (1 Bel) = 10 dB (10 Dezibel).

Die Leistung P ist definiert als Produkt von Spannung und Stromstärke. Mit dem Ohm'schen Gesetz erhalten wir $P_1 = U_1 I_1 = \frac{U_1^2}{R_1}$, $P_2 = U_2 I_2 = \frac{U_2^2}{R_2}$. Ist $R_1 = R_2$, so gilt:

$$a = \lg \left(\frac{P_1}{P_2} \right) = \lg \left(\frac{U_1^2 R_2}{R_1 U_2^2} \right) = \lg \left(\left(\frac{U_1}{U_2} \right)^2 \right) = 2 \lg \left(\frac{U_1}{U_2} \right). \quad (1.19)$$

Bei Diagrammen, die Dämpfungen beschreiben, wird häufig eine logarithmische Darstellung verwendet. ■

1.5.8.4 Logarithmische Darstellungen

In Physik und Technik wird ein Zusammenhang $y = f(x)$ häufig anschaulicher, wenn man die x - und y -Achse einzeln oder gemeinsam logarithmisch einteilt. Folgende Darstellungen sind üblich (siehe Abbildung 1.32):

■ **Einfach-logarithmische Darstellungen**

- Die Funktion a^x wird für größer werdendes x sehr schnell sehr groß, so dass man beim Zeichnen des Funktionsgraphen schnell an den oberen Rand des Diagramms stößt. Abhilfe schafft hier eine andere Einteilung der y -Achse. Statt diese äquidistant mit $0, 1, 2, 3, \dots$ einzuteilen, benutzt man für die y -Achse eine 10^Y -Einteilung $10^0 = 1$, $10^1 = 10$, $10^2 = 100$, $10^3 = 1\,000$, $10^4 = 10\,000$, $\dots$, während die x -Achse „normal“ eingeteilt wird. Zeichnet man bei der logarithmischen Skalierung der y -Achse einen Funktionsgraphen zu $f(x)$, dann ergibt sich das gleiche Bild, das bei „normaler“ y -Achse zu $\lg(f(x))$ entsteht. Für die Funktion $f(x) = ca^{bx}$ zeichnet man so also den Graphen zu

$$\underbrace{\lg(f(x))}_Y = \lg(ca^{bx}) = \lg(c) + bx \lg(a) = \lg(c) + [b \lg(a)]x.$$

Wir sehen eine Gerade mit Steigung $b \lg(a)$. Das ist sehr hilfreich, wenn wir einen exponentiellen Zusammenhang vermuten und die Parameter mittels einer Ausgleichsgeraden durch Messwerte bestimmen möchten.

- Wählt man dagegen eine logarithmische Einteilung der x -Achse und eine „normale“ Einteilung der y -Achse, dann ergibt sich das gleiche Bild, das bei „normaler“ x -Achse zu $f(10^x)$ entsteht. Für die Funktion $f(x) = c \log_a(bx)$ zeichnet man so also den Graphen einer Funktion $g(X)$ mit

$$g(X) = f(10^X) = c \log_a(b \cdot 10^X) = c \log_a(b) + c \log_a(10^X) = c \log_a(b) + [c \log_a(10)]X,$$

also auch hier eine Gerade, jetzt mit Steigung $c \log_a(10)$.

Beispiel 1.91

Zeichnet man die Dämpfung $a = 2 \lg \left(\frac{U_1}{U_2} \right)$ (siehe (1.19)) als Funktion des Quotienten $x = \frac{U_1}{U_2}$, d. h. als Funktion $f(x) = 2 \lg(x)$, so entsteht bei einer logarithmisch unterteilten x -Achse der Graph der Geraden $2X$. ■

- Bei einer **doppelt-logarithmischen Darstellung** werden beide Achsen mit $10^0, 10^1, 10^2, 10^3, 10^4, \dots$ beschriftet. Hat man eine Potenzfunktion $f(x) = ax^b$, so erhält man durch Anwendung des Logarithmus

$$\underbrace{\lg(f(x))}_Y = \lg(a) + b \underbrace{\lg(x)}_X.$$

Trägt man also auf der „normalen“ x -Achse $\lg(x)$ und auf der y -Achse $\lg(f(x))$ an, so sieht man den Graphen zu $Y = \lg(a) + bX$ und damit eine Gerade. Der gleiche Graph entsteht bei der doppelt-logarithmischen Darstellung.

Beispiel 1.92

Setzen wir $x = 10^X$ in $f(x) := x^n$ ein, erhalten wir $[10^X]^n = 10^{nX}$, also $Y = nX$, und zeichnen eine Gerade mit Steigung n . ■

Für einfach- und doppelt-logarithmische Darstellungen gibt es Spezialpapier im Schreibwarenhandel.

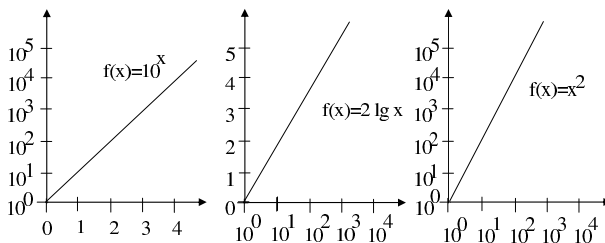


Abb. 1.32: Einfach- und doppelt-logarithmische Darstellung

Beispiel 1.93 (Dampfdruckkurve)

Der Druck, bei dem die flüssige und dampfförmige Phase eines Stoffes gleichzeitig auftreten, wird als **Dampfdruck** p bezeichnet. Für „reine“ Stoffe hängt er nur von der Temperatur T ab:

$$p(T) = a \cdot e^{-b/T}.$$

Dabei sind Parameter a (gemessen in Pa) und b (mit der Einheit K) Stoffkonstanten, und die Temperatur wird in K angegeben. Durch Division mit $p_0 = 1\,013$ mbar (mittlerer Druck auf Meereshöhe) erhält man eine Gleichung ohne Einheiten:

$$\frac{p(T)}{p_0} = \frac{a}{p_0} \cdot e^{-b/T}.$$

Mit dem Logarithmus zur Basis 10 erhalten wir

$$\lg\left(\frac{p(T)}{p_0}\right) = \lg\left(\frac{a}{p_0}\right) - \frac{b}{T} \cdot \lg(e).$$

Trägt man die Variable $y := \lg\left(\frac{p(T)}{p_0}\right)$ gegen $x := \frac{1}{T}$ auf, so erhält man eine Gerade, deren Steigung $-b \lg(e)$ und y -Achsenabschnitt $\lg\left(\frac{a}{p_0}\right)$ direkt mit je einem der beiden Stoffparameter a und b verbunden sind. ■

1.5.8.5 Exponential- und Logarithmusgleichungen

Einfache Exponential- und Logarithmusgleichungen lassen sich mit Hilfe der Definition der Potenz und des Logarithmus unter Verwendung der Rechenregeln umformen und gegebenenfalls lösen. Ansonsten hilft (nicht nur) hier Kreativität. Bevor man aber zu lange vergeblich nach Lösungen sucht, kann man mit Sätzen, die wir im Rahmen der Differenzialrechnung kennenlernen, zunächst die Existenz von Lösungen überprüfen. Danach kann man auch mit Näherungsverfahren arbeiten (vgl. Seite 306).

Beispiel 1.94

a) Wir lösen

$$2^x = 64.$$

Anwendung des Logarithmus zur Basis 2 auf beide Seiten der Gleichung liefert

$$2^x = 64 \implies \lg(2^x) = \lg(64) \implies x \cdot \lg(2) = \lg(64) \implies x = \frac{\lg(64)}{\lg(2)} = \frac{6}{1} = 6.$$

b) Zu bestimmen ist die Lösung von

$$\log_x\left(\frac{1}{4}\right) = -1.$$

Mit der Definition des Logarithmus ergibt sich sofort

$$\log_x\left(\frac{1}{4}\right) = -1 \implies x^{-1} = \frac{1}{4} \implies x = 4.$$

c) Zu lösen ist

$$\ln(2x + 1) = \frac{1}{4}.$$

Anwendung der Exponentialfunktion auf beide Seiten der Gleichung führt zu

$$\ln(2x + 1) = \frac{1}{4} \implies e^{\ln(2x+1)} = e^{\frac{1}{4}} \implies 2x + 1 = e^{\frac{1}{4}} \implies x = \frac{e^{\frac{1}{4}} - 1}{2}.$$

d) Gesucht sind die Lösungen von

$$-8e^{-x} + 2 \cdot e^{-2x} = -8.$$

Substitution $u = e^{-x}$ liefert die quadratische Gleichung $-8u + 2u^2 = -8$ bzw. in Normalform $u^2 - 4u + 4 = 0$ mit der Lösung $u = 2$. Wegen $u = e^{-x}$ ist $\ln(u) = -x$ und damit $x = -\ln(2)$.

e) Für $x > 1$ ist

$$\begin{aligned} \ln(x^2 - 1) - \ln(x) = 0 &\implies \ln\left(\frac{x^2 - 1}{x}\right) = 0 \implies \frac{x^2 - 1}{x} = e^0 = 1 \\ &\implies x^2 - 1 = x \implies x^2 - x - 1 = 0 \\ &\implies x = \frac{1}{2} \pm \sqrt{\frac{5}{4}}. \end{aligned}$$

Wegen $x > 1$ gibt es hier nur die Lösung $x = \frac{1}{2} + \sqrt{\frac{5}{4}}$. ■

1.5.9 Trigonometrische Funktionen

Wie kann man die von der Zeit t abhängende Spannung $u(t)$ beschreiben, die in einem Generator entsteht, bei dem sich eine Leiterschleife durch ein Magnetfeld dreht? Um diese Frage zu beantworten, benötigen wir trigonometrische Funktionen. Mit ihnen werden Schwingungen beschrieben.

Die trigonometrischen Funktionen gehören wie die Exponentialfunktion und der Logarithmus zur Klasse der transzendenten Funktionen. Sie besitzen für einen spitzen Winkel x eine anschauliche Interpretation als Verhältniszahlen von Seitenlängen in einem rechtwinkligen Dreieck bzw. am Einheitskreis.

1.5.9.1 Winkel und Bogenmaß

Aus historischen Gründen wird der Vollwinkel in 360° eingeteilt (Gradmaß, Taste „DEG“ auf dem Taschenrechner). Diese Einteilung ist für die Anwendung in der Mathematik wenig geeignet, hier wird hauptsächlich das Bogenmaß (Taschenrechner: „RAD“) verwendet.

Dabei nutzt man aus, dass ein Einheitskreis (ein Kreis mit Radius 1) den Umfang 2π besitzt. Ist α ein Winkel im Gradmaß, so erhält man über den Dreisatz

$$\frac{\alpha}{360} = \frac{x}{2\pi} \quad \text{bzw.} \quad \alpha = \frac{360}{2\pi} x \quad \text{und} \quad x = \frac{2\pi}{360} \alpha$$

den entsprechenden Winkel x im Bogenmaß. Dies ist genau die Länge des Bogens, den der Winkel aus dem Einheitskreis schneidet. Folgende weitere Eckdaten sollten Sie sich merken: $\frac{\pi}{2}$ entspricht 90° , π entspricht 180° , $\frac{\pi}{4}$ entspricht 45° .

1.5.9.2 Sinus, Kosinus und Tangens

Die Definition der trigonometrischen Funktionen erfolgt zunächst über die Längenverhältnisse im rechtwinkligen Dreieck.

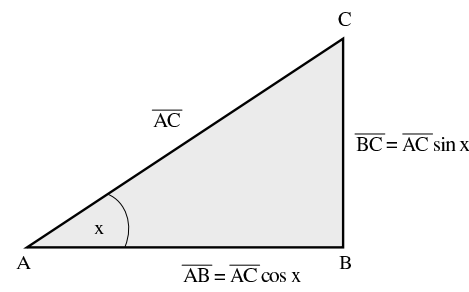


Abb. 1.33: Dreieck aus Definition 1.36

Definition 1.36 (Trigonometrische Funktionen)

Das Dreieck $\triangle ABC$ mit Seitenlängen $\overline{AB}, \overline{BC}, \overline{AC}$ sei rechtwinklig und $0 < x < \frac{\pi}{2}$ der Winkel zwischen $\overline{AB}$ und $\overline{AC}$ (siehe Abbildung 1.33). Dann setzen wir

$$\begin{aligned} \sin(x) &:= \frac{\overline{BC}}{\overline{AC}} \left[\frac{\text{Gegenkathete}}{\text{Hypotenuse}} \right], & \cos(x) &:= \frac{\overline{AB}}{\overline{AC}} \left[\frac{\text{Ankathete}}{\text{Hypotenuse}} \right], \\ \tan(x) &:= \frac{\overline{BC}}{\overline{AB}} \left[\frac{\text{Gegenkathete}}{\text{Ankathete}} \right], & \cot(x) &:= \frac{\overline{AB}}{\overline{BC}} \left[\frac{\text{Ankathete}}{\text{Gegenkathete}} \right]. \end{aligned}$$

Die Funktion $\sin$ heißt der **Sinus**, $\cos$ der **Kosinus**, $\tan$ der **Tangens** und $\cot$ der **Kotangens**.

Tab. 1.6: Wichtige Funktionswerte von Sinus und Kosinus

Winkel	0° 0	30° $\frac{\pi}{6}$	45° $\frac{\pi}{4}$	60° $\frac{\pi}{3}$	90° $\frac{\pi}{2}$	180° π
Sinus	$0 = \frac{1}{2}\sqrt{0}$	$\frac{1}{2} = \frac{1}{2}\sqrt{1}$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{3}$	$1 = \frac{1}{2}\sqrt{4}$	0
Kosinus	$1 = \frac{1}{2}\sqrt{4}$	$\frac{1}{2}\sqrt{3}$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2} = \frac{1}{2}\sqrt{1}$	$0 = \frac{1}{2}\sqrt{0}$	-1

Laut Definition ist

$$\tan(x) = \frac{\sin(x)}{\cos(x)} \quad \text{und} \quad \cot(x) = \frac{\cos(x)}{\sin(x)}.$$

Eigentlich müsste man sich überlegen, dass die Funktionswerte als Verhältnis von Seitenlängen tatsächlich nur vom Winkel abhängen, dass also das Seitenverhältnis unabhängig von der tatsächlichen Seitenlänge ist. Das folgt aber aus den Strahlensätzen (vgl. Seite 429).

Wir legen nun den Punkt A in den Koordinatenursprung. Für einen Punkt C auf dem Einheitskreis um A im ersten Quadranten (d. h., beide Koordinaten sind nicht-negativ, siehe Abbildung 1.15 auf Seite 78) ist $\sin(x) = \frac{\overline{BC}}{\overline{AC}} = \frac{\overline{BC}}{1} = \overline{BC}$, $\cos(x) = \overline{AB}$. Durchläuft C nun alle Punkte des Einheitskreises, so erhält man die Erweiterung des Sinus und Kosinus für beliebige Winkelargumente außerhalb $]0, \frac{\pi}{2}[$, siehe Abbildung 1.34. Ist B dabei links von A , so ist die Länge $\overline{AB}$ negativ zu verstehen, ebenso ist hier $\overline{BC}$ negativ, falls C unterhalb von B liegt.

Insbesondere ist

$$\sin(0) = \sin(2\pi) = 0 \quad \text{und} \quad \cos(0) = \cos(2\pi) = 1.$$

Damit sind Sinus und Kosinus aber nicht nur auf $[0, 2\pi]$ erklärt, sondern sie sind darüber hinaus 2π -periodisch fortgesetzt (vgl. Definition 1.28 auf Seite 83), d. h.

$$\sin(x + 2\pi) = \sin(x) \quad \text{und} \quad \cos(x + 2\pi) = \cos(x)$$

bzw.

$$\sin(x + k2\pi) = \sin(x) \quad \text{und} \quad \cos(x + k2\pi) = \cos(x), \quad k \in \mathbb{Z}.$$

Dabei haben wir eine Umlaufrichtung entgegen dem Uhrzeigersinn gewählt und positive Winkel benutzt. Der Gegenuhrzeigersinn heißt **mathematisch positiver Sinn**. Laufen wir aber im Uhrzeigersinn, so verwenden wir negative Winkel, z. B.: $\sin\left(\frac{\pi}{2}\right) = \sin\left(-\frac{3\pi}{2}\right)$.

Die Fortsetzung des Tangens bzw. Kotangens ist über den Quotienten des fortgesetzten Sinus und Kosinus bzw. über dessen Kehrwert erklärt. Im Gegensatz zu Sinus und Kosinus sind Tangens und Kotangens aber nicht an allen Stellen definiert. In Tabelle 1.7 sind Definitions- und Wertebereiche sowie die primitive (kleinste) Periode der Funktionen zusammengefasst.

Über die Konstruktion am Einheitskreis kann man auch sofort die Funktionsgraphen des Sinus und Kosinus ablesen (siehe Abbildung 1.35). Damit erhält man die Funktionsgraphen in Abbildung 1.36.

Man erkennt sofort am Einheitskreis, dass $|\sin(x)| \leq |x|$ ist, da die Länge der Gegenkathete $|\sin(x)|$ kürzer als die Länge des zugehörigen Bogens ist. In Abbildung 1.34 sieht man ebenso im Vergleich mit der Bogenlänge, dass $\tan(x) \geq x$ für $0 \leq x < \frac{\pi}{2}$ ist. Außerdem sieht man, dass

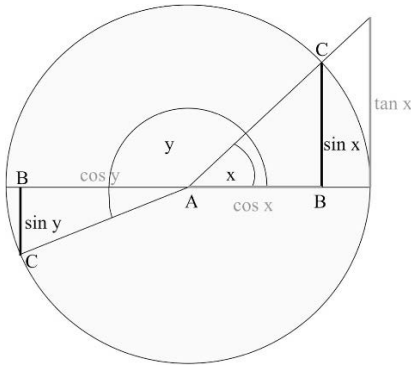


Abb. 1.34: Fortsetzung der trigonometrischen Funktionen am Einheitskreis

$$\sin(x + \pi) = -\sin(x) \quad \text{und} \quad \cos(x + \pi) = -\cos(x).$$

Am gestrichelten Dreieck in Abbildung 1.37 lesen wir ab, wie man $\sin(x)$ und $\cos(x)$ ineinander überführt. Sie sind um $\frac{\pi}{2}$ phasenverschoben:

$$\sin\left(x + \frac{\pi}{2}\right) = \cos(x), \quad \cos\left(x - \frac{\pi}{2}\right) = \sin(x), \quad x \in \mathbb{R}.$$

Die folgenden Symmetrieeigenschaften lassen sich ebenfalls direkt aus der Konstruktion am Einheitskreis ablesen.

Lemma 1.8 (Symmetrie, vgl. Definition 1.26)

Die Sinus- und Tangensfunktion sind ungerade, Kosinus ist eine gerade Funktion, d. h.

$$\begin{aligned} \sin(-x) &= -\sin(x) \quad \text{für alle } x \in \mathbb{R}, \\ \cos(-x) &= \cos(x) \quad \text{für alle } x \in \mathbb{R}, \\ \tan(-x) &= -\tan(x) \quad \text{für alle } x \in]-\pi/2, \pi/2[. \end{aligned}$$

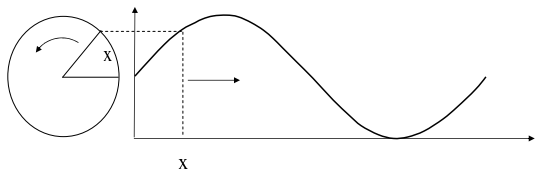


Abb. 1.35: Konstruktion des Funktionsgraphen von $\sin(x)$

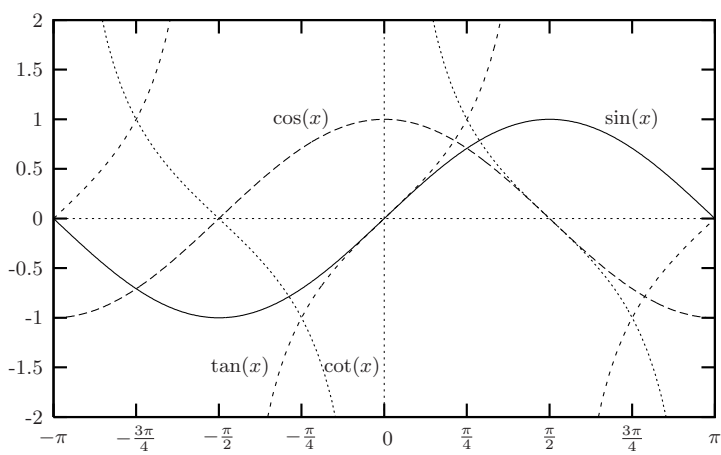


Abb. 1.36: $\sin(x)$ (durchgezogene Linie), $\cos(x)$ (Graph des Sinus um $\frac{\pi}{2}$ nach links verschoben), $\tan(x)$ (streng monoton steigende Äste), $\cot(x)$ (streng monoton fallende Äste) auf $[-\pi, \pi]$

Erlaubt man in der Definition 1.26 (Seite 81) der Punktsymmetrie eine Definitionslücke bei 0, so ist auch der Kotangens eine ungerade Funktion.

Häufig benötigt man die Nullstellen und die Extremstellen:

$$\begin{aligned} \sin(k\pi) &= 0 \quad \text{und} \quad \cos\left(\frac{\pi}{2} + k\pi\right) = 0, \quad k \in \mathbb{Z}, \\ \sin\left(\frac{\pi}{2} + k\pi\right) &= (-1)^k \quad \text{und} \quad \cos(k\pi) = (-1)^k, \quad k \in \mathbb{Z}. \end{aligned}$$

Beispiel 1.95

Dreht sich eine Leiterschleife in einem konstanten Magnetfeld, so wird eine von der Zeit abhängende Spannung induziert (siehe Abbildung 1.38). Die Spannung hängt ab von der Änderung des magnetischen Flusses durch die Schleife. Zunächst bestimmen wir den magnetischen Fluss in Abhängigkeit des Drehwinkels α . Für $\alpha = 0$ ist dieser

Tab. 1.7: Definitions-, Wertebereich und die jeweilige primitive Periode der Winkelfunktionen

Funktion	Definitionsbereich	Wertebereich	primitive Periode
$\sin(x)$	$\mathbb{R}$	$[-1, 1]$	2π
$\cos(x)$	$\mathbb{R}$	$[-1, 1]$	2π
$\tan(x)$	$\mathbb{R} \setminus \left\{ \frac{\pi}{2} + k\pi, k \in \mathbb{Z} \right\}$	$\mathbb{R}$	π
$\cot(x)$	$\mathbb{R} \setminus \{k\pi, k \in \mathbb{Z}\}$	$\mathbb{R}$	π

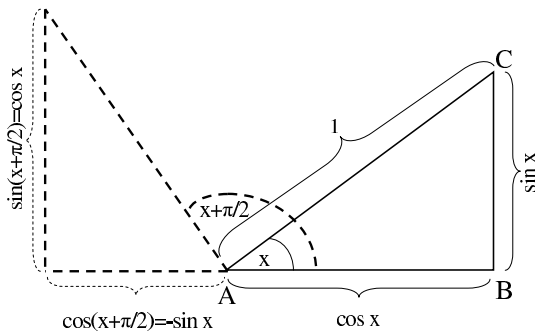


Abb. 1.37: $\sin(x + \pi/2) = \cos(x)$ und $\cos(x + \pi/2) = -\sin(x)$

maximal und sei $\Phi(0) := \Phi_0$. Damit ist $\cos(\alpha) = \frac{\Phi(\alpha)}{\Phi(0)}$, also $\Phi(\alpha) = \Phi_0 \cos(\alpha)$. Drehen wir mit konstanter Geschwindigkeit, so ist der Drehwinkel $\alpha(t)$ zum Zeitpunkt t bestimmt über $\alpha(t) = ct$. Der zugehörige magnetische Fluss ist $\Phi(t) = \Phi_0 \cos(ct)$. Die induzierte Spannung berechnet sich über die **Lenz'sche Regel** als negative momentane Änderung des magnetischen Flusses zum Zeitpunkt t . Hier muss eine Ableitung berechnet werden, die die momentane Änderung angibt. Wir verschieben diese Aufgabe, bis wir mit Ableitungen rechnen können (siehe Seite 299, $u(t) = c\Phi_0 \sin(ct)$). ■

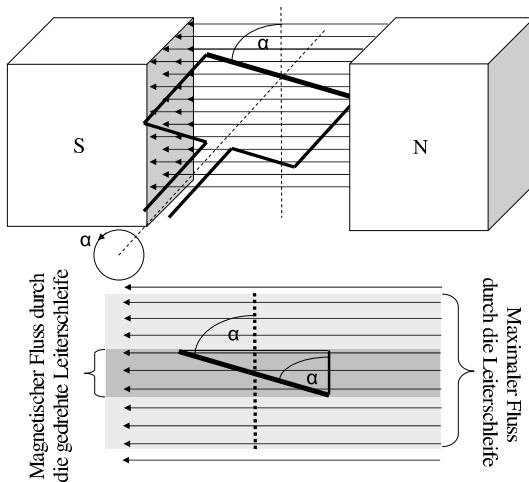


Abb. 1.38: Wechselstromgenerator mit magnetischem Fluss $\Phi(\alpha) = \Phi \cos(\alpha)$

1.5.9.3 Trigonometrische Funktionen in der Geometrie

Der Satz von Pythagoras, nach dem in einem rechtwinkligen Dreieck das Quadrat der Gegenkathete plus das Quadrat der Ankathete gleich dem Quadrat der Hypotenuse ist, lautet mittels Sinus und Kosinus ausgedrückt:

Lemma 1.9 (Trigonometrische Form des Satzes von Pythagoras)

Für alle $x \in \mathbb{R}$ gilt:

$$\sin^2(x) + \cos^2(x) = 1. \quad (1.20)$$

Denn hat die Hypotenuse die Länge 1, so ist $\sin(x)$ die Länge der Gegen- und $\cos(x)$ die Länge der Ankathete. Beim Satz von Pythagoras betrachtet man einen rechten Winkel. Mittels der Kosinusfunktion kann der Satz auf beliebige Winkel erweitert werden:

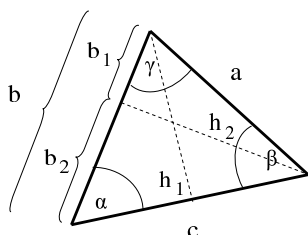


Abb. 1.39: Bezeichnungen für den Sinus- und Kosinus-Satz

Satz 1.20 (Kosinus-Satz)

In einem Dreieck mit Seitenlängen a, b und c sei γ der Winkel zwischen den Seiten mit Längen a und b (siehe Abbildung 1.39). Dann gilt:

$$c^2 = a^2 + b^2 - 2ab \cos(\gamma).$$

Beweis: Wir wenden den Satz von Pythagoras an und erhalten im Dreieck aus Abbildung 1.39: $h_2^2 + b_2^2 = c^2$ sowie $b_1^2 + h_2^2 = a^2$. Setzen wir $h_2^2 = a^2 - b_1^2$ in die erste Gleichung ein, erhalten wir

$$a^2 - b_1^2 + b_2^2 = c^2. \quad (1.21)$$

Da $\cos(\gamma) = \frac{b_1}{a}$, ist $b_1 = a \cos(\gamma)$ und $b_2 = b - b_1 = b - a \cos(\gamma)$. Eingesetzt in (1.21):

$$\begin{aligned} a^2 - a^2 \cos^2(\gamma) + (b - a \cos(\gamma))^2 &= c^2 \\ \iff a^2 - a^2 \cos^2(\gamma) + b^2 - 2ab \cos(\gamma) + a^2 \cos^2(\gamma) &= c^2 \\ \iff a^2 + b^2 - 2ab \cos(\gamma) &= c^2. \end{aligned}$$

□

Neben dem Kosinus-Satz ist der Sinus-Satz eine wichtige Aussage für allgemeine Dreiecke, die man direkt über die Definition der Sinusfunktion ablesen kann:

Satz 1.21 (Sinus-Satz)

In einem Dreieck mit Seitenlängen a, b und c sei α der Winkel, der der Seite mit Länge a gegenüberliegt, der Winkel β liege b und γ der Seite mit Länge c gegenüber (siehe Abbildung 1.39). Dann gilt:

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

Beweis: Wegen $\sin \alpha = \frac{h_1}{b}$, $\sin \alpha = \frac{h_2}{c}$, $\sin \beta = \frac{h_1}{a}$ und $\sin \gamma = \frac{h_2}{a}$ ist

$$\frac{a}{\sin(\alpha)} = \frac{a}{\frac{h_1}{b}} = \frac{ab}{h_1} = \frac{b}{\frac{h_1}{a}} = \frac{b}{\sin(\beta)}, \quad \frac{a}{\sin(\alpha)} = \frac{a}{\frac{h_2}{c}} = \frac{ac}{h_2} = \frac{c}{\frac{h_2}{a}} = \frac{c}{\sin(\gamma)}.$$

□

Beispiel 1.96

Die Breite eines Flusses soll bestimmt werden, ohne dass man dafür den Fluss überqueren muss. Dazu sucht man am gegenüberliegenden Ufer einen Bezugspunkt B und einen Punkt A direkt gegenüber von B am eigenen Ufer (siehe Abbildung 1.40). Außerdem benötigt man einen weiteren Punkt C am eigenen Ufer. Gesucht ist der Abstand $\overline{AB}$. Den Abstand $\overline{AC}$ von A und C kann man messen, ebenso durch Anpeilen des Punktes B die Winkel γ zwischen den Strecken AC und AB sowie β zwischen den Strecken CA und CB . Damit ist

$$\frac{\overline{AB}}{\sin(\beta)} = \frac{\overline{AC}}{\sin(\pi - \beta - \gamma)},$$

so dass die Breite $\overline{AB}$ des Flusses aus den Messungen berechnet werden kann. ■

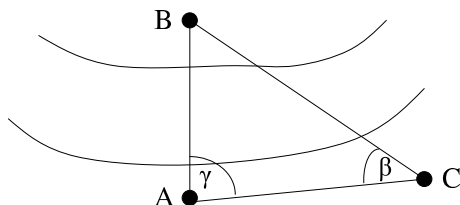


Abb. 1.40: Zur Anwendung des Sinus-Satzes im Beispiel 1.96

1.5.9.4 Additionstheoreme

Für die Funktionswerte von Winkelsummen gelten Rechenregeln, die man **Additionstheoreme** nennt. Unter Verwendung komplexer Zahlen werden diese Regeln später zu Potenzregeln, die sich wesentlich leichter merken lassen (siehe Kapitel 1.6.4.3).

Satz 1.22 (Additionstheoreme für Summen im Argument)

Für alle $x, y \in \mathbb{R}$ gilt:

$$\cos(x + y) = \cos(x) \cos(y) - \sin(x) \sin(y), \quad (1.22)$$

$$\sin(x + y) = \sin(x) \cos(y) + \cos(x) \sin(y). \quad (1.23)$$

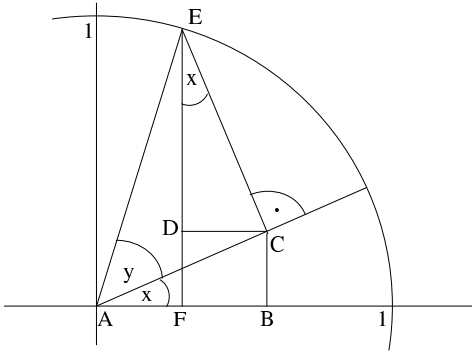


Abb. 1.41: Herleitung der Additionstheoreme am Einheitskreis

Beweis: (Skizze) Wir beschränken uns auf den anschaulichen ersten Quadranten, wie in Abbildung 1.41 dargestellt. Für $y = 0$ oder $y = \pi/2$ erhält man die Additionstheoreme unmittelbar. Für alle Werte ungleich $k\frac{\pi}{2}$ sind $\sin(y) \neq 0$ und $\cos(y) \neq 0$. Hier erhalten wir (siehe Abbildung 1.41):

$$\sin(x) = \frac{\overline{BC}}{\overline{AC}} = \frac{\overline{BC}}{\cos(y)} \implies \overline{BC} = \sin(x) \cos(y),$$

$$\sin(x) = \frac{\overline{DC}}{\overline{EC}} = \frac{\overline{DC}}{\sin(y)} \implies \overline{DC} = \sin(x) \sin(y),$$

$$\cos(x) = \frac{\overline{AB}}{\overline{AC}} = \frac{\overline{AB}}{\cos(y)} \implies \overline{AB} = \cos(x) \cos(y),$$

$$\cos(x) = \frac{\overline{ED}}{\overline{EC}} = \frac{\overline{ED}}{\sin(y)} \implies \overline{ED} = \cos(x) \sin(y).$$

Damit: $\cos(x + y) = \overline{AF} = \overline{AB} - \overline{DC} = \cos(x) \cos(y) - \sin(x) \sin(y),$

$\sin(x + y) = \overline{EF} = \overline{ED} + \overline{BC} = \cos(x) \sin(y) + \sin(x) \cos(y). \quad \square$

Damit hat man auch ein entsprechendes Additionstheorem für den Tangens bewiesen:

Folgerung 1.1 (Summe im Argument des Tangens)

Für $x, y \in \mathbb{R}$ mit $x, y, x + y \neq (2k + 1)\frac{\pi}{2}$, $k \in \mathbb{Z}$, gilt:

$$\tan(x + y) = \frac{\tan(x) + \tan(y)}{1 - \tan(x) \tan(y)}. \quad (1.24)$$

Beweis:

$$\tan(x+y) = \frac{\sin(x+y)}{\cos(x+y)} = \frac{\sin(x)\cos(y) + \cos(x)\sin(y)}{\cos(x)\cos(y) - \sin(x)\sin(y)} = \frac{\tan(x) + \tan(y)}{1 - \tan(x)\tan(y)}.$$

Im letzten Schritt haben wir mit $1/[\cos(x)\cos(y)]$ erweitert. $\square$

Setzt man in (1.22) und (1.23) $y = x$, so ergibt sich zusammen mit (1.20)

Folgerung 1.2 (Doppelter Winkel)

Für alle $x, y \in \mathbb{R}$ gilt:

$$\cos(2x) = \cos^2(x) - \sin^2(x) = 2\cos^2(x) - 1, \quad (1.25)$$

$$\sin(2x) = 2\sin(x)\cos(x). \quad (1.26)$$

Insbesondere ist also $\cos^2(x) = \frac{1}{2}[\cos(2x) + 1]$ und $\sin^2(x) = \cos^2(x) - \cos(2x) = \frac{1}{2}[1 - \cos(2x)]$. Ersetzt man in (1.22) und (1.23) y durch $-y$, so erhält man

Folgerung 1.3 (Additionstheoreme für Differenzen im Argument)

Für alle $x, y \in \mathbb{R}$ gilt:

$$\cos(x-y) = \cos(x)\cos(y) + \sin(x)\sin(y), \quad (1.27)$$

$$\sin(x-y) = \sin(x)\cos(y) - \cos(x)\sin(y). \quad (1.28)$$

Addieren wir (1.23) und (1.28), so erhalten wir

$$\sin(u+v) + \sin(u-v) = 2\sin(u)\cos(v).$$

Mit $u = \frac{x-y}{2}$ und $v = \frac{x+y}{2}$ ergibt sich die Beziehung

Folgerung 1.4 (Additionstheorem für die Differenz zweier Sinus-Werte)

Für alle $x, y \in \mathbb{R}$ gilt:

$$\sin(x) - \sin(y) = 2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right). \quad (1.29)$$

Berücksichtigen wir, dass $\sin(x)$ eine ungerade Funktion ist, ergibt sich daraus

Folgerung 1.5 (Additionstheorem für die Summe von Funktionswerten)

Für alle $x, y \in \mathbb{R}$ gilt:

$$\sin(x) + \sin(y) = 2\cos\left(\frac{x-y}{2}\right)\sin\left(\frac{x+y}{2}\right), \quad (1.30)$$

$$\cos(x) + \cos(y) = 2 \cos\left(\frac{x-y}{2}\right) \cos\left(\frac{x+y}{2}\right). \quad (1.31)$$

Beweis: Nach (1.29) ist

$$\sin(x) + \sin(y) = \sin(x) - \sin(-y) = 2 \cos\left(\frac{x-y}{2}\right) \sin\left(\frac{x+y}{2}\right).$$

Die Gleichung (1.31) erhält man, indem man Kosinus über den Sinus schreibt:

$$\begin{aligned} \cos(x) + \cos(y) &= \sin\left(x + \frac{\pi}{2}\right) + \sin\left(y + \frac{\pi}{2}\right) \\ &= 2 \cos\left(\frac{x + \frac{\pi}{2} - y - \frac{\pi}{2}}{2}\right) \sin\left(\frac{x + y + \pi}{2}\right) = 2 \cos\left(\frac{x-y}{2}\right) \cos\left(\frac{x+y}{2}\right). \quad \square \end{aligned}$$

1.5.9.5 Harmonische Schwingungen und Zeigerdiagramme

In der Wechselstromtechnik betrachtet man Überlagerungen von Spannungen und Strömen, die die gleiche Frequenz (z. B. 50 Hz) haben. Eine entsprechende Spannung lässt sich schreiben als

$$u(t) = \hat{u} \cos(\omega t + \varphi_u). \quad (1.32)$$

Die Funktion (1.32) beschreibt die allgemeine **harmonische Schwingung**. Sie ist eine der wichtigsten Funktionen in der Elektrotechnik und im Maschinenbau überhaupt.

Die **Amplitude** $\hat{u}$ der Schwingung ist die maximale Auslenkung, ω ist die **Winkelgeschwindigkeit** oder **Kreisfrequenz** (in rad/s), bzw. $\frac{\omega}{2\pi}$ ist die **Frequenz** der Schwingung (in 1/s, also Hertz, Hz). Man nennt $\omega t + \varphi_u$ den **Phasenwinkel** (die **Phase**). Dabei heißt φ_u **Nullphasenwinkel**, da zum Zeitpunkt $t = 0$ der Kosinus zu diesem Wert berechnet werden muss. Wegen $\sin(x + \frac{\pi}{2}) = \cos(x)$ erhält man für (1.32) auch die Darstellung

$$u(t) = \hat{u} \sin\left(\omega t + \left[\varphi_u + \frac{\pi}{2}\right]\right)$$

mit der gleichen Frequenz $\frac{\omega}{2\pi}$. Bezüglich des Sinus hat man den neuen Nullphasenwinkel $\varphi_u + \frac{\pi}{2}$. Die primitive Periode (vgl. Definition 1.28) oder **Schwingungsdauer** $T > 0$ der Funktion ergibt sich aus

$$\hat{u} \cos(\omega(t+T) + \varphi_u) \stackrel{!}{=} \hat{u} \cos(\omega t + \varphi_u) \iff \omega T = k \cdot 2\pi, \quad k \in \mathbb{Z},$$

zu $T = \frac{2\pi}{\omega}$.

Überlagert man zwei Sinus- oder Kosinus-Funktionen gleicher Frequenz, so hat auch die Summenfunktion diese Frequenz und lässt sich in der Form (1.32) schreiben. Man erkennt dies, indem man die Summe über ein **Zeigerdiagramm** bildet (siehe Abbildung 1.42, vgl. auch Abbildung 1.35). Man addiert die Zeiger, die die beiden Ausgangsfunktionen beschreiben, indem man einen an die Spitze des Zweiten setzt, und erhält einen Zeiger, der die Summenfunktion beschreibt. Diese Zeigerarithmetik (Vektorrechnung) werden wir noch ausführlich in der Linearen Algebra (Kapitel 3) untersuchen. Statt mit den Funktionen rechnet man in der Elektrotechnik mit Zeigerdiagrammen. Jeder Zeiger ist eindeutig durch Länge und Winkel charakterisiert. Man kann den Zeiger aber auch eindeutig mit den beiden Koordinaten des Punktes an der Spitze des Pfeils beschreiben. Eine Koordinate ist der zum Winkel gehörende Funktionswert, die andere Koordinate sieht man zwar nicht direkt im Funktionsgraphen, aber in der Tat vereinfacht sich der Umgang mit Funktionen des Typs (1.32), wenn man diese (imaginäre) Koordinate bei Rechnungen einbezieht. Geeignete Rechenregeln für Punkte in der Ebene führen uns zu den komplexen Zahlen in Kapitel 1.6.

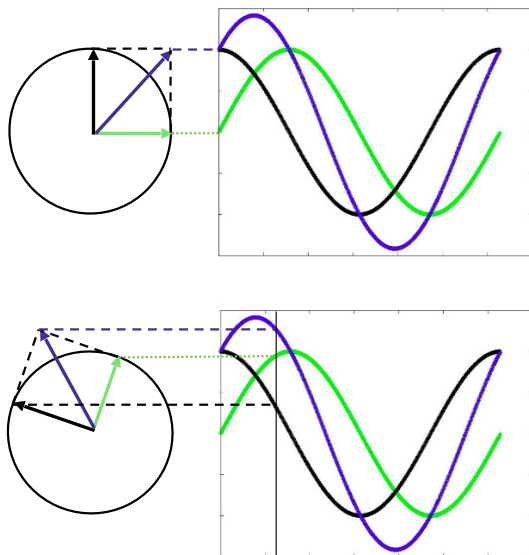


Abb. 1.42: Addition von trigonometrischen Funktionen mittels Zeigerdiagramm

Was wir mit Hilfe der Zeiger anschaulich erklärt haben, vollziehen wir nun rechnerisch nach. Gegeben sind die Schwingungen

$$u_1(t) = a \cdot \cos(\omega t) \quad \text{und} \quad u_2(t) = b \cdot \cos(\omega t + \varphi).$$

Das Additionstheorem (1.22) für den Kosinus liefert

$$\begin{aligned} u_1(t) + u_2(t) &= a \cdot \cos(\omega t) + b \cdot [\cos(\omega t) \cos(\varphi) - \sin(\omega t) \sin(\varphi)] \\ &= [a + b \cos(\varphi)] \cdot \cos(\omega t) - b \sin(\varphi) \cdot \sin(\omega t). \end{aligned} \quad (1.33)$$

Mit

$$A := \sqrt{(a + b \cos(\varphi))^2 + b^2 \sin^2(\varphi)}$$

folgt

$$u_1(t) + u_2(t) = A \cdot \left(\frac{a + b \cos(\varphi)}{A} \cdot \cos(\omega t) - \frac{b \sin(\varphi)}{A} \cdot \sin(\omega t) \right).$$

Die Koeffizienten von $\cos(\omega t)$ und $\sin(\omega t)$ in der Klammer liegen im Intervall $[-1, 1]$, und es gilt nach Definition von A

$$\left(\frac{a + b \cos(\varphi)}{A} \right)^2 + \left(\frac{b \sin(\varphi)}{A} \right)^2 = 1.$$

Damit können wir $\frac{a+b \cos(\varphi)}{A}$ und $\frac{b \sin(\varphi)}{A}$ als Seitenlängen in einem rechtwinkligen Dreieck mit Hypotenusenlänge 1 ansehen, d. h., die Koeffizienten können als Kosinus und Sinus eines Winkels φ_0 geschrieben werden, der eindeutig über

$$\cos(\varphi_0) = \frac{a + b \cos(\varphi)}{A}, \quad \sin(\varphi_0) = \frac{b \sin(\varphi)}{A}$$

festgelegt ist. Wir erhalten die Darstellung als **harmonische Schwingung**

$$u_1(t) + u_2(t) = A \cdot (\cos(\varphi_0) \cdot \cos(\omega t) - \sin(\varphi_0) \cdot \sin(\omega t)) \stackrel{(1.22)}{=} A \cdot \cos(\omega t + \varphi_0).$$

Beispiel 1.97 (Schwebungen)

Ein weiterer interessanter Schwingungstyp ist die **Schwebung**. Sie tritt auf, wenn sich zwei harmonische Schwingungen fast gleicher Frequenz überlagern. Wir betrachten die Schwingung

$$x(t) = A \cdot (\sin(\omega_0 t) - \sin(\omega t)) \quad \text{mit} \quad \omega \approx \omega_0.$$

Diese Schwingung kann beispielsweise entstehen, wenn sich die Töne zweier leicht verstimmt Orgelpfeifen überlagern. Mit den Additionstheoremen (1.23) und (1.28) des Sinus

$$\begin{aligned} \sin(\omega_0 t) &= \sin\left(\frac{\omega_0 + \omega}{2}t + \frac{\omega_0 - \omega}{2}t\right) \\ &= \sin\left(\frac{\omega_0 + \omega}{2}t\right) \cdot \cos\left(\frac{\omega_0 - \omega}{2}t\right) + \cos\left(\frac{\omega_0 + \omega}{2}t\right) \cdot \sin\left(\frac{\omega_0 - \omega}{2}t\right), \\ \sin(\omega t) &= \sin\left(\frac{\omega_0 + \omega}{2}t - \frac{\omega_0 - \omega}{2}t\right) \\ &= \sin\left(\frac{\omega_0 + \omega}{2}t\right) \cdot \cos\left(\frac{\omega_0 - \omega}{2}t\right) - \cos\left(\frac{\omega_0 + \omega}{2}t\right) \cdot \sin\left(\frac{\omega_0 - \omega}{2}t\right) \end{aligned}$$

erhalten wir

$$\sin(\omega_0 t) - \sin(\omega t) = 2 \sin\left(\frac{\omega_0 - \omega}{2}t\right) \cdot \cos\left(\frac{\omega_0 + \omega}{2}t\right),$$

d. h.

$$x(t) = 2A \cdot \sin\left(\frac{\omega_0 - \omega}{2}t\right) \cdot \cos\left(\frac{\omega_0 + \omega}{2}t\right),$$

bzw.

$$x(t) = C(\omega) \cdot \cos\left(\frac{\omega_0 + \omega}{2}t\right) \text{ mit } C(\omega) := 2A \cdot \sin\left(\frac{\omega_0 - \omega}{2}t\right).$$

Wegen $\omega_0 \approx \omega$ ist $\frac{\omega_0 - \omega}{2} \approx 0$ und $\frac{\omega_0 + \omega}{2} \approx \omega_0$, so dass $x(t)$ als hochfrequente Schwingung mit Kreisfrequenz $\approx \omega_0$ interpretiert werden kann, deren Amplitude $C(\omega)$ niederfrequent schwingt. Dieses Phänomen heißt **Schwebung** (siehe Abbildung 1.43). ■

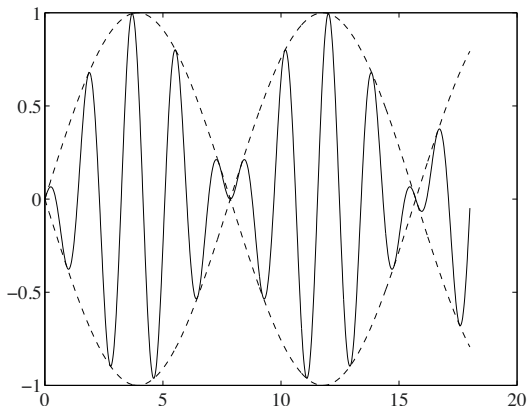


Abb. 1.43: Schwebung

1.5.9.6 Arkus-Funktionen

Da die trigonometrischen Funktionen 2π -periodisch sind, nehmen sie ihre Funktionswerte an unendlich vielen Stellen an. Sie sind also nicht injektiv und damit nicht umkehrbar. Schränkt man aber den Definitionsbereich so ein, dass sie auf dem neuen Definitionsbereich streng monoton und damit injektiv werden, nennt man die dann existierenden Umkehrfunktionen die **Arkus-Funktionen**. Das Monotonieintervall der Ausgangsfunktion entspricht dem Wertebereich der Umkehrfunktion. Da man zwischen unendlich vielen Monotonieintervallen wählen kann, gibt es auch entsprechend viele verschiedene Umkehrfunktionen. Wählt man die in Tabelle 1.8 und im Folgenden angegebenen Intervalle, so erhält man Umkehrfunktionen, die man jeweils als den **Hauptwert** der Arkus-Funktion bezeichnet. Umkehrfunktionen zu anderen Monotonieintervallen der Winkelfunktionen können dann mit Hilfe der Hauptwerte bestimmt werden.

- $\arcsin(x)$: Da die auf das Intervall $[-\frac{\pi}{2}, \frac{\pi}{2}]$ eingeschränkte Sinusfunktion $f(x) = \sin(x)$, $f : [-\frac{\pi}{2}, \frac{\pi}{2}] \rightarrow [-1, 1]$, streng monoton steigend ist, existiert die zugehörige Umkehrfunktion (siehe Abbildung 1.44). Diese Funktion f^{-1} heißt **Arkussinus**:

$$f^{-1} = \arcsin : [-1, 1] \rightarrow \left[-\frac{\pi}{2}, \frac{\pi}{2}\right].$$

Tab. 1.8: Arkus-Funktionen

Funktion	Monotonieintervall	Umkehrfunktion	zugehöriger Definitionsbereich
$\sin(x)$	$[-\frac{\pi}{2}, \frac{\pi}{2}]$	$\arcsin(x)$	$[-1, 1]$
$\cos(x)$	$[0, \pi]$	$\arccos(x)$	$[-1, 1]$
$\tan(x)$	$]-\frac{\pi}{2}, \frac{\pi}{2}[$	$\arctan(x)$	$] -\infty, \infty[$
$\cot(x)$	$]0, \pi[$	$\operatorname{arccot}(x)$	$] -\infty, \infty[$

Es gilt also $\sin(\arcsin(x)) = x$ für $x \in [-1, 1]$ und $\arcsin(\sin(x)) = x$ für $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Sucht man alle Winkel x , an denen der Sinus einen konkreten Wert $y \in [-1, 1]$ annimmt, dann erhält man diese als Menge $\{x \in \mathbb{R} : x = \arcsin(y) + k2\pi \text{ oder } x = \pi - \arcsin(y) + k2\pi, k \in \mathbb{Z}\}$.

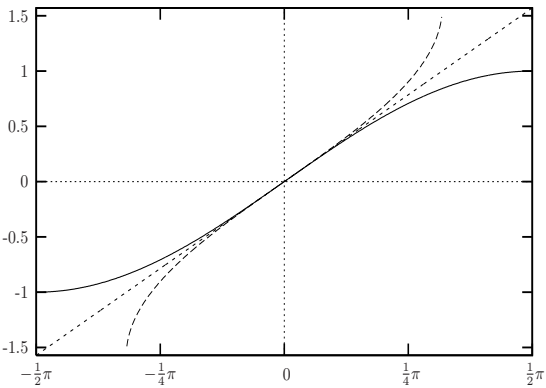


Abb. 1.44: $\sin(x)$ und $\arcsin(x)$

- $\arccos(x)$: Die auf $[0, \pi]$ eingeschränkte Kosinusfunktion $f(x) = \cos(x)$, $f : [0, \pi] \rightarrow [-1, 1]$, ist streng monoton fallend und somit umkehrbar (siehe Abbildung 1.45). Die Umkehrfunktion f^{-1} heißt **Arkuskosinus**:

$$f^{-1} = \arccos : [-1, 1] \rightarrow [0, \pi].$$

Sucht man alle Winkel x , an denen der Kosinus einen konkreten Wert $y \in [-1, 1]$ annimmt, dann erhält man diese als Menge $\{x \in \mathbb{R} : x = \arccos(y) + 2k\pi \text{ oder } x = -\arccos(y) + 2k\pi, k \in \mathbb{Z}\}$.

- $\arctan(x)$: Indem man die Tangensfunktion auf das Intervall $]-\frac{\pi}{2}, \frac{\pi}{2}[$ einschränkt, erhält man einen streng monoton steigenden Zweig $f(x) = \tan(x)$, $f :]-\frac{\pi}{2}, \frac{\pi}{2}[\rightarrow \mathbb{R}$ (siehe Abbildung 1.46). Die Umkehrfunktion f^{-1} heißt der **Arkustangens**:

$$f^{-1} = \arctan : \mathbb{R} \rightarrow]-\frac{\pi}{2}, \frac{\pi}{2}[.$$

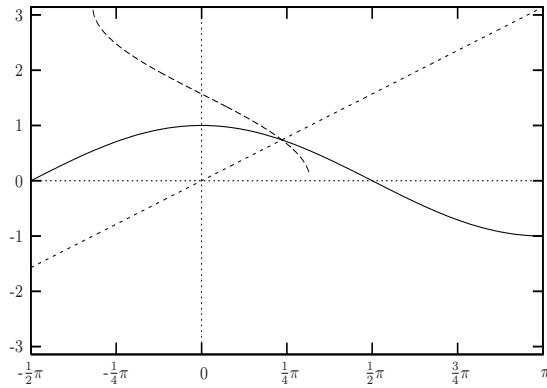


Abb. 1.45: $\cos(x)$ und $\arccos(x)$

Sucht man alle Winkel x , an denen der Tangens einen Wert $y \in \mathbb{R}$ annimmt, dann erhält man diese über $x = \arctan(y) + k\pi$, $k \in \mathbb{Z}$.

Der Arkustangens ist auf ganz $\mathbb{R}$ definiert, und wir werden noch ausnutzen, dass es sich um eine „glatte“ Funktion mit schönen Eigenschaften wie strenge Monotonie und Beschränktheit handelt. Sie wird uns insbesondere bei der Integration gebrochen-rationaler Funktionen helfen. Auch sehen wir, dass über den Arkustangens jeder reellen Zahl bijektiv ein Element aus dem beschränkten Intervall $]-\frac{\pi}{2}, \frac{\pi}{2}[$ zugeordnet ist. Allgemeiner können alle reellen Zahlen durch Werte aus $]a, b[$ mit der bijektiven Abbildung $g(x) := a + (b - a) \cdot \left[\frac{1}{2} + \frac{\arctan(x)}{\pi} \right]$ kodiert werden. Man sagt, das Intervall hat die gleiche **Mächtigkeit** wie $\mathbb{R}$ (vgl. Kapitel 1.3.3.5).

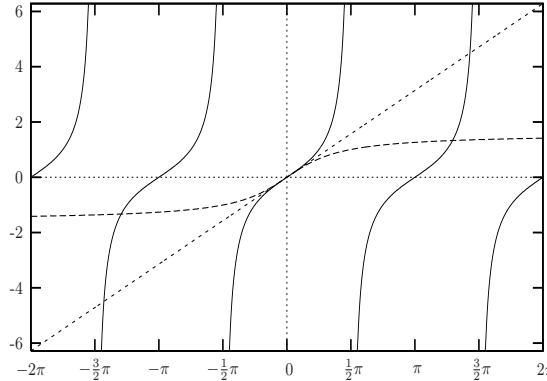


Abb. 1.46: $\tan(x)$ und $\arctan(x)$

Achtung: Bei der Umkehrung der trigonometrischen Funktionen werden häufig Fehler gemacht. Die Umkehrfunktion des Sinus ist nicht der Kosinus, auch nicht $-\sin(x)$ oder $\frac{1}{\sin(x)} = (\sin(x))^{-1}$. Die Arkus-Funktionen können nicht so elementar über $\sin(x)$ und $\cos(x)$ dargestellt werden.

1.5.9.7 Trigonometrische Gleichungen

Es gibt kein standardisiertes Lösungsverfahren für Gleichungen, in denen trigonometrische Funktionen vorkommen. Oft ist es aber hilfreich, in den trigonometrischen Ausdrücken zunächst die Variable so durch einen Term in einer anderen Variable zu ersetzen (zu substituieren), dass man eine Gleichung mit nur einer trigonometrischen Funktion erhält.

Beispiel 1.98

a) Für $x \in [0, 1]$ lösen wir die Gleichung

$$\sin^2(\arccos x) + 1 = 2\sqrt{1 - x^2},$$

indem wir $x = \cos(y)$ (d. h. $y = \arccos x$, $y \in [0, \frac{\pi}{2}]$) substituieren. Denn damit erhalten wir eine Gleichung mit nur einer trigonometrischen Funktion: Mit $x = \cos(y)$ ist $\sin^2(\arccos(x)) = \sin^2(\arccos(\cos y)) = \sin^2(y)$ und

$$\sqrt{1 - x^2} = \sqrt{1 - \cos^2(y)} = \sqrt{\sin^2(y)} = |\sin(y)| = \sin(y),$$

so dass die Gleichung lautet

$$\sin^2(y) - 2\sin(y) + 1 = 0$$

mit der Lösung $\sin(y) = 1$, also $y = \frac{\pi}{2}$. Die Rücksubstitution ergibt $x = \cos(\frac{\pi}{2}) = 0$.

b) Ebenfalls für $x \in [0, 1]$ lösen wir

$$\arcsin(x) + \arccos(x) = 2x.$$

Auch hier ersetzen wir zunächst die Variable x – nun durch $\sin(y) = \cos(\frac{\pi}{2} - y)$, $y \in [0, \frac{\pi}{2}]$ und erhalten

$$y + \frac{\pi}{2} - y = 2\sin(y).$$

Die Lösung ist $y = \arcsin(\frac{\pi}{4})$ und damit $x = \sin(y) = \frac{\pi}{4}$. ■

1.5.10 Hyperbel- und Areafunktionen

Die **Hyperbelfunktionen** (oder hyperbolische Funktionen) benötigt man beispielsweise, wenn man die Kurve eines an den Enden befestigten Seils beschreiben möchte (Kettenlinie, siehe Seite 738). In der Elektrotechnik werden Hyperbelfunktionen z. B. in digitalen Filtern gebraucht. Sie sind eng verwandt mit den trigonometrischen Funktionen (vgl. Seite 141).

1.5.10.1 Hyperbelfunktionen

Definition 1.37 (Hyperbelfunktionen)

Der **Hyperbelsinus (Sinushyperbolikus)** ist definiert durch

$$\sinh(x) := \frac{1}{2} (e^x - e^{-x}), \quad D(\sinh) = \mathbb{R}, \quad W(\sinh) = \mathbb{R}.$$

Der **Hyperbelkosinus (Kosinushyperbolikus)** ist erklärt durch

$$\cosh(x) := \frac{1}{2} (e^x + e^{-x}), \quad D(\cosh) = \mathbb{R}, \quad W(\cosh) = [1, \infty[.$$

Hyperbeltangens (Tangenshyperbolikus) und **Hyperbelkotangens (Kotangenshyperbolikus)** sind erklärt durch

$$\tanh(x) := \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad D(\tanh) = \mathbb{R}, \quad W(\tanh) =]-1, 1[,$$

$$\coth(x) := \frac{\cosh(x)}{\sinh(x)} = \frac{e^x + e^{-x}}{e^x - e^{-x}}, \quad D(\coth) = \mathbb{R} \setminus \{0\}, \quad W(\coth) = \mathbb{R} \setminus [-1, 1].$$

Die Funktionen werden als **Hyperbelfunktionen** bezeichnet (siehe Abbildung 1.47).

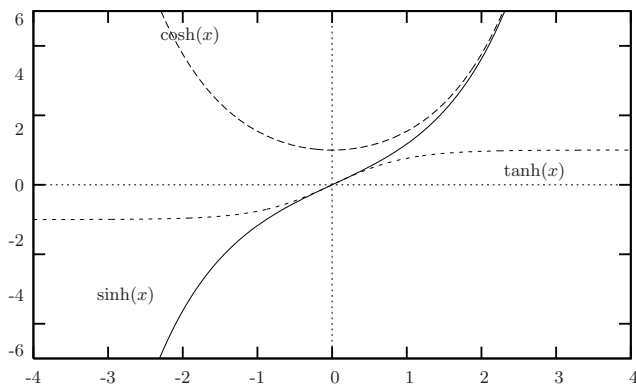


Abb. 1.47: Hyperbelfunktionen

Die Eigenschaften dieser Funktionen erinnern an die trigonometrischen Funktionen:

Satz 1.23 (Eigenschaften der Hyperbelfunktionen)

a) Es gelten die **Symmetrien**

$$\sinh(-x) = -\sinh(x), \quad \cosh(-x) = \cosh(x), \quad \tanh(-x) = -\tanh(x).$$

Der Hyperbelsinus und der Hyperbeltangens sind ungerade Funktionen, der Hyperbelkosinus ist eine gerade Funktion.

b) Es gilt die Gleichung

$$\cosh^2(x) - \sinh^2(x) = 1.$$

Die Lösungsmenge der Gleichung $x^2 - y^2 = 1$ stellt eine Hyperbel dar (vgl. (1.20) auf Seite 118).

c) Es gelten die **Additionstheoreme** (vgl. (1.23) und (1.22) auf Seite 120)

$$\begin{aligned}\sinh(x+y) &= \sinh(x) \cosh(y) + \cosh(x) \sinh(y) \quad \text{und} \\ \cosh(x+y) &= \cosh(x) \cosh(y) + \sinh(x) \sinh(y).\end{aligned}$$

Beweis: Die Eigenschaften ergeben sich direkt aus den Rechenregeln für die Exponentialfunktion. Beispielsweise erhalten wir das Additionstheorem für den Sinushyperbolikus so:

$$\begin{aligned}\sinh(x) \cosh(y) + \cosh(x) \sinh(y) &= \frac{e^x - e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x + e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2} \\ &= \frac{e^{x+y} + e^{x-y} - e^{y-x} - e^{-x-y}}{4} + \frac{e^{x+y} - e^{x-y} + e^{y-x} - e^{-x-y}}{4} \\ &= \frac{e^{x+y} - e^{-x-y}}{2} = \sinh(x+y).\end{aligned}$$

□

1.5.10.2 Areafunktionen als Umkehrfunktionen der Hyperbelfunktionen

Exemplarisch betrachten wir die Umkehrung des Hyperbelsinus auf $\mathbb{R}$. Da $y = \sinh(x)$ auf $\mathbb{R}$ streng monoton steigt (siehe Abbildung 1.47) existiert die Umkehrfunktion. Für

$$y = f(x) = \sinh(x) = \frac{1}{2}(e^x - e^{-x})$$

liefert die Substitution $z = e^x$: $y = \frac{1}{2}(z - \frac{1}{z})$ bzw. nach Multiplikation mit $z \neq 0$ die quadratische Gleichung $z^2 - 2yz - 1 = 0$ mit den beiden Lösungen $z_{1,2} = y \pm \sqrt{y^2 + 1}$. Der Term $y - \sqrt{y^2 + 1} < 0$ kann wegen $z = e^x > 0$ nicht Lösung sein. Wir erhalten

$$e^x = y + \sqrt{y^2 + 1} \iff x = \ln\left(y + \sqrt{y^2 + 1}\right)$$

bzw.

$$y = f^{-1}(x) = \ln\left(x + \sqrt{x^2 + 1}\right).$$

Diese Funktion heißt **Area-Sinushyperbolikus**:

$$y = \operatorname{arsinh}(x) = \ln \left(x + \sqrt{x^2 + 1} \right).$$

Ähnlich erhält man die Umkehrung der weiteren Hyperbelfunktionen, wobei beim Kosinushyperbolicus der ursprüngliche Definitionsbereich eingeschränkt werden muss. Der **Area-Kosinushyperbolicus** ist auf $[1, \infty[$ definiert mit Werten in $[0, \infty[$:

$$\operatorname{arcosh}(x) = \ln(x + \sqrt{x^2 - 1}).$$

Die Umkehrfunktion des Tangenshyperbolicus heißt **Area-Tangenshyperbolicus**

$$\operatorname{artanh}(x) = \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right)$$

und ist auf $] -1, 1[$ definiert mit Werten in $\mathbb{R}$. Schließlich ist Umkehrfunktion des Kotangenshyperbolicus der auf $\mathbb{R} \setminus [-1, 1]$ definierte **Area-Kotangenshyperbolicus**

$$\operatorname{arcoth}(x) = \frac{1}{2} \ln \left(\frac{x+1}{x-1} \right)$$

mit Werten in $\mathbb{R}$.

Wir haben jetzt die wichtigsten reellen Funktionen behandelt. Sie sollten in der Lage sein, zumindest die Graphen aus Abbildung 1.48 aus dem Gedächtnis zu zeichnen.

1.6 Komplexe Zahlen

Bei der Überlagerung von harmonischen Schwingungen gleicher Frequenz konnte die Summenfunktion mit Hilfe der Zeigeraddition dargestellt werden. Der rechnerische Umgang mit Zeigerdiagrammen ist aber bisweilen nicht einfach. Insbesondere muss man die Additionstheoreme der trigonometrischen Funktionen verwenden (siehe Kapitel 1.5.9.5). Einfacher wird es, wenn die Zeiger über Zahlenpaare dargestellt werden, für die geeignete Rechenoperationen definiert sind. Diese Zahlenpaare heißen **komplexe Zahlen**.

Neben ihrer Bedeutung für die Schwingungsanalyse und die Wechselstromrechnung benötigen wir komplexe Zahlen auch in der Mathematik. So müssen wir beispielsweise bei der Integration gebrochen-rationaler Funktionen Partialbruchzerlegungen durchführen und dabei komplexe Nullstellen beachten. Ebenso werden wir sie bei Fourier-Entwicklungen und Integraltransformationen einsetzen.

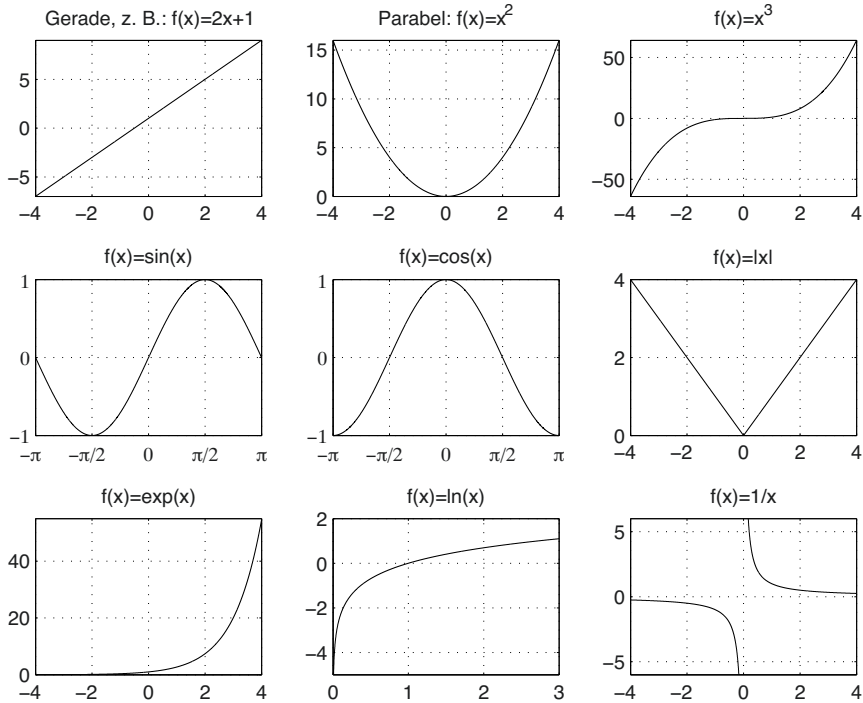


Abb. 1.48: Wichtige Funktionsgraphen

1.6.1 Erweiterung der reellen Zahlen um eine imaginäre Einheit

Um die Gleichung $x^2 = 2$ zu lösen, mussten wir die rationalen Zahlen $\mathbb{Q}$ zu den reellen Zahlen $\mathbb{R}$ erweitern, und wir erhielten die Lösungen $\sqrt{2}$ und $-\sqrt{2}$. Zur Lösung der Gleichung

$$x^2 = -1$$

muss man die reellen Zahlen $\mathbb{R}$ zu den komplexen Zahlen $\mathbb{C}$ erweitern. Eine Zahl j ($j \notin \mathbb{R}$), die diese Gleichung löst, heißt die **imaginäre Einheit** j .

$$j^2 := -1$$

j ist keine Variable, sondern ein neues Zahlsymbol, so wie auch „1“ und „2“ Symbole für Zahlen sind. Falls man mit j so rechnet wie mit den bislang verwendeten reellen

Zahlen, dann ist $(-j)^2 = (-1)^2 j^2 = j^2 = -1$ und $-j$ ebenfalls eine Lösung von $x^2 = -1$.

In der Mathematik wird statt j die Bezeichnung i verwendet. Allerdings ist i bereits für die Stromstärke von Wechselströmen reserviert, weshalb in der Ingenieur-Mathematik j üblich ist.

Die folgende Definition wurde zuerst von Raffaello Bombelli (1526–1572) in seinem im Todesjahr erschienenen Buch verwendet. Er hat nach Vorarbeiten von Girolamo Cardano (1501–1576) vorgeschlagen, Ausdrücke der Form $x + \sqrt{-1} \cdot y$ als Zahlen zu verstehen und korrekt damit gerechnet. Dieser Schritt war genauso revolutionär wie die Einführung der negativen Zahlen.

Definition 1.38 (Komplexe Zahlen)

$$\mathbb{C} := \{z : z = x + jy \text{ mit } x, y \in \mathbb{R}\}$$

ist die Menge der **komplexen Zahlen**. Für $z = x + jy$ ist x der **Realteil** von z , $x = \operatorname{Re}(z)$, und y der **Imaginärteil**, $y = \operatorname{Im}(z)$.

Achtung: Der Imaginärteil ist eine reelle Zahl: $\operatorname{Im}(x + jy) = y$, **nicht** $\operatorname{Im}(x + jy) = jy$. Es gilt

$$\mathbb{R} \subset \mathbb{C},$$

wobei die reellen Zahlen mit genau den komplexen Zahlen identifiziert werden, deren Imaginärteil 0 ist. Zwei komplexe Zahlen sind **gleich** genau dann, wenn Real- und Imaginärteil gleich sind.

1.6.2 Komplexe Arithmetik

Die Addition und Multiplikation definiert man so, dass man für den Spezialfall von reellen Zahlen die bislang verwendeten Operationen erhält und dass die Rechenregeln erhalten bleiben. Für $z_1 = (x_1 + jy_1)$ und $z_2 = (x_2 + jy_2)$ ist

$$\begin{aligned} z_1 + z_2 &:= (x_1 + x_2) + j(y_1 + y_2), \\ z_1 z_2 &= (x_1 + jy_1)(x_2 + jy_2) := (x_1 x_2 - y_1 y_2) + j(x_1 y_2 + y_1 x_2). \end{aligned}$$

Beispiel 1.99

- a) $(5 + 3j) \pm (7 - 4j) = 5 \pm 7 + j(3 \pm (-4)),$
 b) $(5 + 3j) \cdot (7 - 4j) = 35 - 20j + 21j - 12 \underbrace{j^2}_{=-1} = 47 + j.$ ■

Die zwei Zahlen des Real- und Imaginärteils erlauben eine neue Operation, die man z. B. im Rahmen der Bruchrechnung benötigt:

Definition 1.39 (Konjugation)

Zu $z = x + jy \in \mathbb{C}$ heißt $\bar{z} = x - jy$ die zu z **konjugiert komplexe Zahl**.

Statt der Schreibweise $\bar{z}$ ist auch z^* gängig.

Lemma 1.10 (Rechenregeln für die Konjugation)

$$\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2, \quad \overline{z_1 \cdot z_2} = \bar{z}_1 \cdot \bar{z}_2. \quad (1.34)$$

Um den Quotienten $\frac{x_1 + jy_1}{x_2 + jy_2}$ in der Form $x + jy$ darzustellen, erweitert man mit der konjugiert komplexen Zahl des Nenners. Zunächst ist

$$(x_2 + jy_2)\overline{(x_2 + jy_2)} = (x_2 + jy_2)(x_2 - jy_2) = x_2^2 - jx_2y_2 + jy_2x_2 - j^2y_2^2 = x_2^2 + y_2^2 \in \mathbb{R}.$$

So gelangt man zu einem reellen Nenner und sieht, dass auch der Quotient zweier komplexer Zahlen wieder eine komplexe Zahl mit Real- und Imaginärteil ist:

$$\begin{aligned} \frac{x_1 + jy_1}{x_2 + jy_2} &= \frac{(x_1 + jy_1)(x_2 - jy_2)}{(x_2 + jy_2)(x_2 - jy_2)} = \frac{(x_1 + jy_1)(x_2 - jy_2)}{x_2^2 + y_2^2} \\ &= \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + j \frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}. \end{aligned}$$

Beispiel 1.100

Wir berechnen Real- und Imaginärteil der komplexen Zahlen:

$$\begin{aligned} \text{a) } \frac{5 + 3j}{7 - 4j} &= \frac{5 + 3j}{7 - 4j} \cdot \frac{7 + 4j}{7 + 4j} = \frac{35 + 20j + 21j + 12j^2}{49 + 16} = \frac{23}{65} + j \cdot \frac{41}{65}, \\ \text{b) } \frac{3 + 2j}{4 + j} &= \frac{(3 + 2j)(4 - j)}{17} = \frac{12 + 2}{17} + j \frac{-3 + 8}{17}. \end{aligned}$$

■

Beispiel 1.101

Für welche $z \in \mathbb{C}$ gilt $(1 + j) \cdot z + (1 - j) \cdot \bar{z} = 0$?

Mit $z = x + jy$ folgt

$$\begin{aligned} (1 + j) \cdot (x + jy) + (1 - j) \cdot (x - jy) &= 0 \\ \iff (x - y) + j(x + y) + (x - y) + j(-x - y) &= 2(x - y) = 0 \iff x = y, \end{aligned}$$

d. h., $z = x + jx$, $x \in \mathbb{R}$ ist Lösung der Gleichung.

■

Die aus $\mathbb{R}$ bekannten Rechenregeln wie Kommutativität der Addition und Multiplikation sowie die Klammerregeln (Assoziativ- und Distributivgesetze) gelten ebenso in $\mathbb{C}$. Das **neutrale Element der Addition** ist $0 = 0 + j \cdot 0$, denn es gilt $z = z + 0 = 0 + z$ für $z \in \mathbb{C}$. Das **neutrale Element der Multiplikation** ist $1 = 1 + j \cdot 0$, denn es gilt

$z = z \cdot 1 = 1 \cdot z$ für $z \in \mathbb{C}$. Die Menge $\mathbb{C}$ zusammen mit der Addition und Multiplikation bildet einen Körper (siehe Definition 1.18 auf Seite 45), der den Körper $\mathbb{R}$ umfasst und damit neue Lösungen von Gleichungen bereithält.

Beispiel 1.102

In $\mathbb{C}$ können wir nun die quadratische Gleichung $x^2 + px + q = 0$ mit $p, q \in \mathbb{R}$ vollständig lösen. Es gilt

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} \in \mathbb{R} \quad \text{im Fall} \quad p^2 \geq 4q$$

und im Fall $p^2 \leq 4q$:

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} = -\frac{p}{2} \pm \sqrt{\left(q - \frac{p^2}{4}\right) \cdot j^2} = -\frac{p}{2} \pm j \cdot \underbrace{\sqrt{q - \frac{p^2}{4}}}_{\in \mathbb{R}}.$$

Ferner gilt $\bar{x}_1 = x_2$, d. h., die beiden Wurzeln sind konjugiert komplex. Zum Beispiel erhält man für die quadratische Gleichung $x^2 + 4x + 5 = 0$ die beiden komplexen Lösungen $x_{1,2} = -2 \pm j$. ■

1.6.3 Die Gauß'sche Zahlenebene

Man kann die komplexen Zahlen als Paare reeller Zahlen (Realteil, Imaginärteil) auffassen, die so verknüpft werden:

$$\begin{aligned}(x_1, y_1) + (x_2, y_2) &:= (x_1 + x_2, y_1 + y_2), \\ (x_1, y_1)(x_2, y_2) &:= (x_1x_2 - y_1y_2, x_1y_2 + y_1x_2).\end{aligned}$$

Damit können wir komplexe Zahlen $x + jy$ als Punkte (x, y) der Ebene oder als Pfeile auffassen, die vom Nullpunkt zum Punkt (x, y) führen. Die reellen Zahlen entsprechen dann der Punktmenge $\{(x, 0) : x \in \mathbb{R}\}$, die Zahl 1 z. B. dem Punkt $(1, 0)$. Die imaginäre Einheit j entspricht dem Punkt $(0, 1)$, $-j$ wird durch den Punkt $(0, -1)$ repräsentiert. Die Addition von zwei komplexen Zahlen wird zur Addition von Pfeilen: Man setzt die Pfeile der beiden Zahlen aneinander und gelangt so zum Punkt, der die komplexe Summe repräsentiert (siehe Abbildung 1.49 und vgl. mit der Vektoraddition auf Seite 469). Es ist ein großer Verdienst von Gauß, dass er den komplexen Zahlen das Mystische genommen und sie als Punkte der Ebene greifbar gemacht hat.

1.6.3.1 Betrag

Eine Ordnungsrelation „ $\leq$ “, die $\mathbb{C}$ (also die Zahlenebene) total ordnet, ist nicht möglich. Statt Zahlen zu vergleichen, kann man aber ihren Abstand zum Nullpunkt, ihren Betrag, vergleichen. Der **Betrag** einer komplexen Zahl wird über den Satz von Pythagoras definiert:

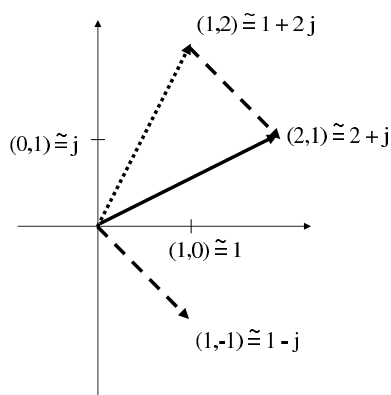


Abb. 1.49: Addition komplexer Zahlen als Addition von Pfeilen ($[1 + 2j] + [1 - j] = 2 + j$)

Definition 1.40 (Betrag einer komplexen Zahl $z = x + jy \in \mathbb{C}$)

$$|z| = |x + jy| := \sqrt{x^2 + y^2}.$$

Es gilt

$$|z| = \sqrt{(x + jy)(x - jy)} = \sqrt{(x + jy)\overline{(x + jy)}} = \sqrt{z\bar{z}}.$$

Man beachte, dass für reelle Zahlen diese Definition des Betrags identisch mit der zuvor gemachten (siehe (1.12) auf Seite 65) ist.

1.6.3.2 Rechnen mit Beträgen komplexer Zahlen

Für das Rechnen mit Beträgen gelten wie bei reellen Zahlen die folgende Regeln (vgl. (1.13), (1.14)):

Lemma 1.11 (Rechenregeln für Beträge komplexer Zahlen)

Für $z_1, z_2 \in \mathbb{C}$ gilt:

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|, \quad \left| \frac{z_1}{z_2} \right| = \frac{|z_1|}{|z_2|}, \quad (1.35)$$

$$|z_1 + z_2| \leq |z_1| + |z_2| \quad (\text{Dreiecksungleichung}), \quad (1.36)$$

$$|z_1 - z_2| \geq ||z_1| - |z_2|| \quad (\text{Dreiecksungleichung nach unten}). \quad (1.37)$$

Die Gültigkeit von (1.35) rechnet man elementar nach. Die beiden Dreiecksungleichungen lassen sich geometrisch veranschaulichen: Eine Dreiecksseite ist nie länger als die Summe der beiden anderen Dreiecksseiten; eine Dreiecksseite ist nicht kürzer als die Differenz der beiden anderen Dreiecksseiten.

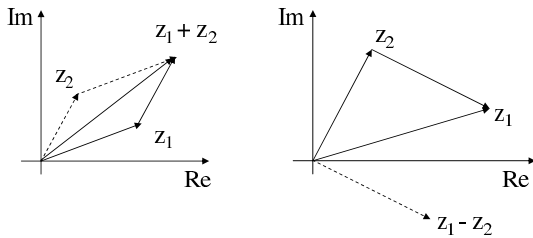


Abb. 1.50: Zu den Dreiecksungleichungen

Beispiel 1.103

a) Sei z_0 eine feste komplexe Zahl. Dann beschreibt die Menge aller $z \in \mathbb{C}$ mit

$$|z - z_0| = R \quad \text{mit} \quad R > 0$$

den Kreis um z_0 mit Radius R . $|z - z_0| \leq R$ beschreibt die Kreisscheibe (siehe Abbildung 1.51).

b) Welche Menge in $\mathbb{C}$ wird beschrieben durch $1 \leq |z - 1 - j| < 1,5$?

Mit $z_0 := 1 + j$ gilt $1 \leq |z - z_0| < 1,5$, d. h., die Menge beschreibt einen Kreisring mit Innenradius 1 und Außenradius 1,5 um z_0 . Der Innenkreis gehört zur Menge, der äußere Kreis gehört nicht zur Menge (siehe Abbildung 1.51).

c) Für welche $z \in \mathbb{C}$ gilt

$$\left| \frac{z - 1}{z + 2} \right| = 1?$$

Gesucht sind Zahlen z mit $|z - 1| = |z + 2|$. Mit $z = x + jy$ folgt:

$$\begin{aligned} |(x - 1) + jy| &= |(x + 2) + jy| \\ \iff \sqrt{(x - 1)^2 + y^2} &= \sqrt{(x + 2)^2 + y^2} \\ \iff (x - 1)^2 + y^2 &= (x + 2)^2 + y^2 \iff x^2 - 2x + 1 = x^2 + 4x + 4 \\ \iff -6x &= 3, \end{aligned}$$

d. h. $x = -\frac{1}{2}$. Die Gerade $z = -\frac{1}{2} + jy$, $y \in \mathbb{R}$, ist Lösung der Gleichung.

d) Wir berechnen die Lösungsmenge der Ungleichung $|z - 4j| < |z|$.

Mit $z = x + jy$ folgt:

$$\begin{aligned} |x + j(y - 4)| &< |x + jy| \\ \iff \sqrt{x^2 + (y - 4)^2} &< \sqrt{x^2 + y^2} \iff x^2 + y^2 - 8y + 16 < x^2 + y^2 \iff y > 2. \end{aligned}$$

Die Halbebene $\text{Im}(z) > 2$, d. h. die Menge $z = x + jy$ mit $x \in \mathbb{R}$ und $y > 2$, ist die gesuchte Lösungsmenge. ■

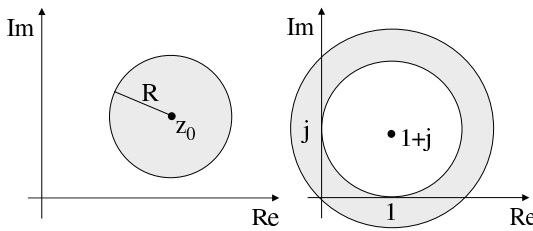


Abb. 1.51: Kreis $|z - z_0| = R$ und Kreissring $1 \leq |z - (1 + j)| < 1,5$

1.6.4 Euler'sche Gleichung und Polarform komplexer Zahlen

1.6.4.1 Polarform komplexer Zahlen

Statt der Euklid'schen Darstellung mittels eines kartesischen Koordinatensystems kann man Punkte der Ebene und damit auch komplexe Zahlen mittels **Polarkoordinaten** darstellen. Jeder Punkt (x, y) ist eindeutig über den Abstand $r := |x + jy| = \sqrt{x^2 + y^2} \geq 0$ zum Nullpunkt und einen Winkel $\varphi \in [0, 2\pi[$ zwischen positiver x -Achse und der Verbindungsstrecke vom Nullpunkt zu (x, y) gegeben (siehe Abbildung 1.52):

$$(x, y) = (r \cos(\varphi), r \sin(\varphi)), \quad x + jy = r [\cos(\varphi) + j \sin(\varphi)].$$

Diese Darstellung heißt auch **goniometrische Form** oder **Polarform**.

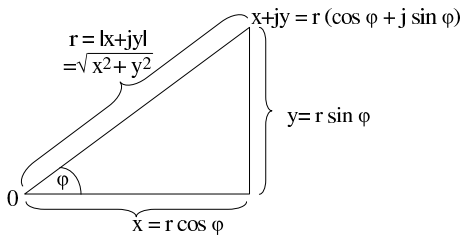


Abb. 1.52: Polardarstellung von $x + jy$

Um den Winkel φ aus z. B. $\tan(\varphi) = y/x$ zu bestimmen, kann man den Arkustangens (vgl. Kapitel 1.5.9.6) benutzen. Allerdings ist der Tangens nicht bijektiv. Um ihn umzukehren, wird der Definitionsbereich auf $]-\pi/2, \pi/2[$ eingeschränkt. Entsprechend erhält man nur Winkel $\arctan(x/y)$ aus diesem Intervall. Um Winkel $\varphi \in]-\pi, \pi]$ zu berechnen, kann man so vorgehen:

- Ist der Realteil $x = 0$, so liegt die Zahl auf der imaginären Achse, und bei positivem Imaginärteil ist $\varphi = \frac{\pi}{2}$, bei negativem Imaginärteil ist $\varphi = -\frac{\pi}{2}$.
- Ist $x \neq 0$, so können wir $\alpha := \arctan\left(\frac{y}{x}\right)$ berechnen. Für $x > 0$ ist $\varphi = \alpha$. Ist dagegen $x < 0$ und $y \geq 0$, so ist $\varphi = \alpha + \pi$. Für $x < 0$ und $y < 0$ erhalten wir $\varphi = \alpha - \pi$.

Diese Fallunterscheidung muss man nicht per Hand durchführen. In vielen Programmiersprachen gibt es die Funktion $\text{atan2}(y, x)$, mit der ein Winkel aus $] - \pi, \pi]$ berechnet wird. Im Gegensatz zum Arkustangens hat diese Funktion zwei Parameter und kann somit die Fallunterscheidung durchführen.

Beispiel 1.104

- a) $z = -2 + j$ hat einen negativen Realteil $x = -2$ und einen positiven Imaginärteil $y = 1$, also gilt $\frac{\pi}{2} < \varphi < \pi$. Aus $\tan(\varphi) = -\frac{1}{2}$ folgt $\varphi \approx -0,4636 + \pi \approx 2,6779$. $r = |z| = \sqrt{(-2)^2 + 1^2} = \sqrt{5}$. Somit hat z die Darstellung $z \approx -2 + j = \sqrt{5} \cdot [\cos(2,6779) + j \cdot \sin(2,6779)]$.
- b) $z = 1 - j$ hat den Realteil $x = 1$ und den Imaginärteil $y = -1$, also ist $\frac{3}{2}\pi < \varphi < 2\pi$, $\tan(\varphi) = -1$ und $\varphi = -\frac{\pi}{4} + 2\pi = \frac{7}{4}\pi$. $r = |z| = \sqrt{1^2 + (-1)^2} = \sqrt{2}$. Also: $z = \sqrt{2}e^{j\frac{7}{4}\pi} = \sqrt{2} \cdot \left(\cos\left(\frac{7}{4}\pi\right) + j \cdot \sin\left(\frac{7}{4}\pi\right) \right)$. ■

1.6.4.2 Euler'sche Gleichung

Mit der Euler'schen Gleichung lässt sich die Polarform nicht nur einfacher schreiben, auch Berechnungen, für die wir bislang Additionstheoreme benötigen, vereinfachen sich:

Definition 1.41 (Euler'sche Gleichung)

Für $\varphi \in \mathbb{R}$ setzen wir

$$e^{j\varphi} := \cos(\varphi) + j \cdot \sin(\varphi).$$

Mit der Euler'schen Gleichung erhalten wir für $z \in \mathbb{C}$ die **Exponentialform**

$$z = x + jy = r \cdot e^{j\varphi}.$$

Die Werte, die $e^{j\varphi}$ annehmen kann, haben alle den Betrag 1 und liegen auf dem Einheitskreis:

$$|e^{j\varphi}| = \sqrt{\cos^2(\varphi) + \sin^2(\varphi)} = 1.$$

In Abbildung 1.53 ist der Zusammenhang zwischen Zeigerdiagrammen und komplexen Zahlen unter Verwendung der Exponentialform dargestellt. Einige wichtige Werte von $e^{j\varphi}$ sind

$$\begin{aligned} e^{j0} = 1 + 0j = 1 &= e^{j2\pi}, & e^{j\frac{\pi}{2}} = 0 + 1j = j, \\ e^{j\pi} = -1 + 0j = -1, & e^{j\frac{3\pi}{2}} = 0 - 1j = -j, & e^{-j\frac{\pi}{2}} = -j. \end{aligned}$$

Wir verwenden $e^{j\varphi}$ als abkürzende Schreibweise für die komplexe Zahl $\cos(\varphi) + j \cdot \sin(\varphi)$. Tatsächlich kann man diese Definition auch als Satz verstehen, wenn man die Exponentialfunktion anhand ihrer Potenzreihenentwicklung von ihrem Definitionsbereich $\mathbb{R}$ auf den Definitionsbereich $\mathbb{C}$ erweitert. Das können wir hier aber noch nicht verstehen, so

dass wir das (viel) später in Kapitel 2.8.6 nachtragen. Dabei erhält man die komplexe Exponentialfunktion. In diesem Buch gehen wir (bis auf Hintergrundinformationen) nicht auf die Theorie von Funktionen auf $\mathbb{C}$ ein. Daher benutzen wir die komplexe Exponentialfunktion ebenfalls nur als Kurzschreibweise.

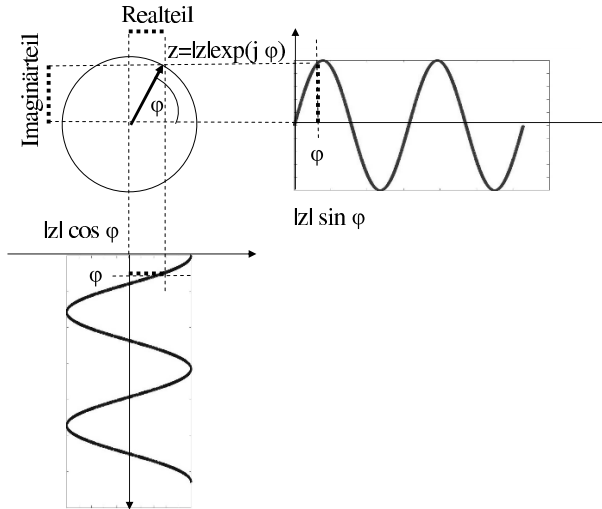


Abb. 1.53: Real- und Imaginärteil im Zeigerdiagramm

Definition 1.42 (Komplexe Exponentialfunktion)

Die komplex-wertige Funktion mit einer komplexen Variable $\exp : \mathbb{C} \rightarrow \mathbb{C}$ ist definiert durch

$$\exp(z) := e^z = e^{\operatorname{Re}(z)} e^{j \operatorname{Im}(z)} = e^{\operatorname{Re}(z)} [\cos(\operatorname{Im}(z)) + j \sin(\operatorname{Im}(z))]. \quad (1.38)$$

Diese Definition ist konsistent mit der reellen Funktion $\exp(x)$ und der Kurzschreibweise $e^{j\varphi}$, denn wenn wir $z = x$ in $\exp(z)$ einsetzen, erhalten wir genau den Wert der reellen Exponentialfunktion, und wenn wir $z = j\varphi$ setzen, erhalten wir $\cos(\varphi) + j \sin(\varphi)$.

Zunächst ist völlig unklar, warum die Exponentialfunktion im Komplexen so eng mit dem Sinus und dem Kosinus verknüpft ist. Es zeigt sich aber, dass $e^{j\varphi}$ Eigenschaften hat, die denen der reellen Exponentialfunktion entsprechen. Wegen der Additionstheoreme (1.22) und (1.23) gilt z. B.:

$$\begin{aligned} e^{j(\varphi+\vartheta)} &= \cos(\varphi+\vartheta) + j \sin(\varphi+\vartheta) \\ &= \cos(\varphi)\cos(\vartheta) - \sin(\varphi)\sin(\vartheta) + j[\sin(\varphi)\cos(\vartheta) + \cos(\varphi)\sin(\vartheta)] \\ &= [\cos(\varphi) + j \sin(\varphi)][\cos(\vartheta) + j \sin(\vartheta)] = e^{j\varphi} e^{j\vartheta}. \end{aligned} \quad (1.39)$$

Mit Hilfe der Euler'schen Gleichung und der Rechenregeln der Exponentialfunktion kann man umgekehrt die Additionstheoreme von Sinus und Kosinus leicht herleiten. Statt Additionstheoreme auswendig zu lernen, reicht es, wenn man die Definition von $e^{j\varphi}$ und die Regeln der Potenzrechnung beherrscht.

Jetzt können wir auch besser verstehen, warum man vom Sinus- und Kosinushyperbolikus spricht. Unter Verwendung der komplexen Exponentialfunktion werden die Funktionen auf $\mathbb{C}$ erweitert. Setzen wir zu $x \in \mathbb{R}$ speziell die komplexe Zahl jx in diese Funktionen ein, so ergibt sich

$$\sinh(jx) = \frac{1}{2} (e^{jx} - e^{-jx}) = j \sin(x) \quad \text{und} \quad \cosh(jx) = \frac{1}{2} (e^{jx} + e^{-jx}) = \cos(x).$$

1.6.4.3 Komplexe Potenzen und komplexe Wurzeln

Iteriert angewendet erhalten wir aus (1.39) eine Regel für Potenzen $n \in \mathbb{N}$, den **Satz von Moivre** (de Moivre 1667–1754):

$$e^{jn\varphi} = e^{j(\varphi+(n-1)\varphi)} = e^{j\varphi} e^{j(n-1)\varphi} = \dots = [e^{j\varphi}]^n.$$

Insbesondere gilt auch für negative Potenzen

$$e^{-jn\varphi} e^{jn\varphi} = e^{j0} = 1 \implies e^{-jn\varphi} = \frac{1}{e^{jn\varphi}} = \frac{1}{[e^{j\varphi}]^n} = [e^{j\varphi}]^{-n}.$$

Diese Regeln für $e^{j\varphi}$ sind viel eingängiger als die Additionstheoreme für Sinus und Kosinus. Insbesondere vereinfacht sich damit die Multiplikation, Division und die Berechnung von Potenzen komplexer Zahlen ganz erheblich:

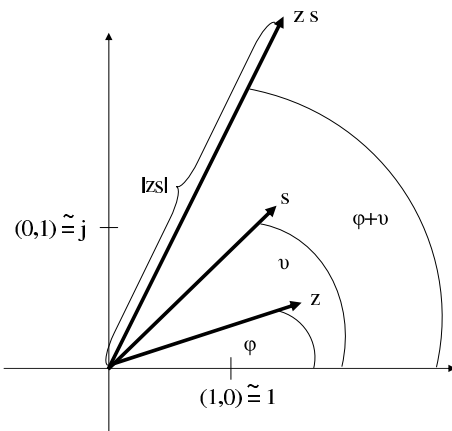


Abb. 1.54: Multiplikation komplexer Zahlen ist Addition der Winkel und Multiplikation der Beträge (hier: $z = |z|e^{j\varphi}$, $s = |s|e^{j\psi}$, $zs = |z||s|e^{j(\varphi+\psi)}$).

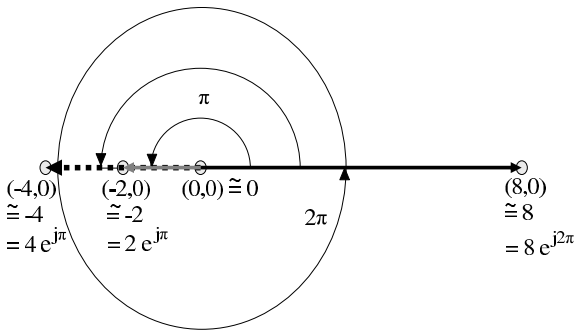


Abb. 1.55: Multiplikation reeller Zahlen in Polardarstellung am Beispiel $(-2)(-4) = 8$

Lemma 1.12 (Rechenregeln für die Polardarstellung)

Es seien $z_1 = r_1 e^{j\varphi_1}$, $z_2 = r_2 e^{j\varphi_2}$ und $z = r e^{j\varphi}$ (vgl. Abbildungen 1.54 und 1.55):

$$\begin{aligned} z_1 \cdot z_2 &= r_1 e^{j\varphi_1} \cdot r_2 e^{j\varphi_2} = r_1 \cdot r_2 \cdot e^{j(\varphi_1 + \varphi_2)}, \\ \frac{z_1}{z_2} &= \frac{r_1 e^{j\varphi_1}}{r_2 e^{j\varphi_2}} = \frac{r_1}{r_2} e^{j(\varphi_1 - \varphi_2)}, \\ z^n &= [r e^{j\varphi}]^n = r^n e^{jn\varphi}. \end{aligned}$$

Wir können also das Produkt zweier komplexer Zahlen in Polardarstellung berechnen, indem wir die Beträge multiplizieren und die Winkel addieren. Bei der Division werden umgekehrt die Beträge dividiert und die Winkel subtrahiert. Es wird mit einer natürlichen Zahl potenziert, indem der Radius damit potenziert und der Winkel damit multipliziert wird. Addition und Subtraktion sind dagegen in der Darstellung „Realteil + $j \cdot$ Imaginärteil“ einfacher.

Beispiel 1.105

a) $z_1 := 5 + j = \sqrt{26} \cdot e^{j \cdot \arctan \frac{1}{5}}$, $z_2 := j = e^{j \cdot \frac{\pi}{2}}$:

$$\begin{aligned} z_1 \cdot z_2 &= -1 + 5j = \sqrt{26} \cdot e^{j \cdot \arctan \frac{1}{5}} \cdot e^{j \cdot \frac{\pi}{2}} = \sqrt{26} \cdot e^{j \cdot (\arctan(\frac{1}{5}) + \frac{\pi}{2})}, \\ \frac{z_1}{z_2} &= \frac{5 + j}{j} = 1 - 5j = \sqrt{26} \cdot e^{j \cdot (\arctan(\frac{1}{5}) - \frac{\pi}{2})}. \end{aligned}$$

b) $z_1 := 2 \cdot e^{j \cdot \frac{\pi}{4}}$, $z_2 := 5 \cdot e^{j \cdot \frac{\pi}{2}}$:

$$\begin{aligned} z_1 \cdot z_2 &= 2 \cdot e^{j \cdot \frac{\pi}{4}} \cdot 5 \cdot e^{j \cdot \frac{\pi}{2}} = 10 \cdot e^{j \cdot (\frac{\pi}{2} + \frac{\pi}{4})} = 10 \cdot e^{j \cdot \frac{3\pi}{4}}, \\ \frac{z_2}{z_1} &= \frac{5 \cdot e^{j \cdot \frac{\pi}{2}}}{2 \cdot e^{j \cdot \frac{\pi}{4}}} = \frac{5}{2} \cdot e^{j \cdot (\frac{\pi}{2} - \frac{\pi}{4})} = \frac{5}{2} \cdot e^{j \cdot \frac{\pi}{4}}. \end{aligned}$$

c) $j = e^{j \frac{\pi}{2}}$, $j^2 = [e^{j \frac{\pi}{2}}]^2 = e^{j 2 \frac{\pi}{2}} = e^{j \pi} = -1$, $j^3 = e^{j 3 \frac{\pi}{2}} = -j$, $j^4 = e^{j 4 \frac{\pi}{2}} = 1$.

d) $(-1)^2 = [e^{j \pi}]^2 = e^{j 2 \pi} = 1$.

e) Die Zahl $-3e^{j\frac{\pi}{3}}$ hat den Betrag 3. Um als Faktor nicht den negativen Betrag zu schreiben, kann man das Minuszeichen über den Winkel ausdrücken:

$$-3e^{j\frac{\pi}{3}} = -1 \cdot 3e^{j\frac{\pi}{3}} = e^{j\pi} \cdot 3e^{j\frac{\pi}{3}} = 3e^{j\pi+j\frac{\pi}{3}} = 3e^{j\frac{4\pi}{3}}. \quad \blacksquare$$

Als Umkehrung der Potenzrechnung in Polarkoordinatenschreibweise erhalten wir zu $n \in \mathbb{N}$ genau n verschiedene **komplexe n -te Wurzeln** z_k von $re^{j\varphi}$ über

$$z_k = \sqrt[n]{r} e^{j\frac{\varphi+2k\pi}{n}}, \quad k \in \{0, 1, 2, \dots, n-1\}. \quad (1.40)$$

Da Sinus und Kosinus und damit auch $e^{j\varphi}$ 2π -periodisch sind, gilt (vgl. Abbildungen 1.56 und 1.57):

$$\left[\sqrt[n]{r} e^{j\frac{\varphi+2k\pi}{n}} \right]^n = r e^{j(\varphi+2k\pi)} = r e^{j\varphi}.$$

Unter den Wurzeln (1.40) kann es null bis zwei reelle n -te Wurzeln geben. Zwei reelle

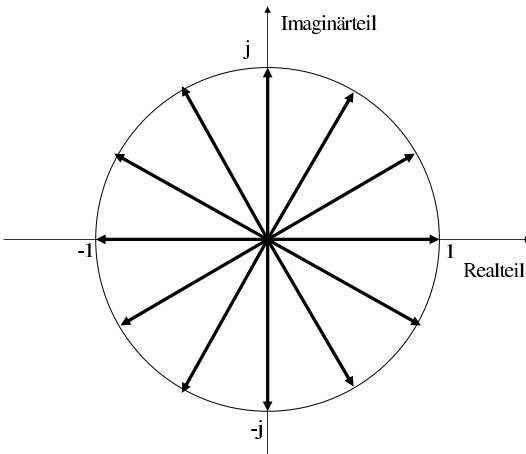


Abb. 1.56: Alle zwölften komplexen Wurzeln aus Eins (Einheitswurzeln)

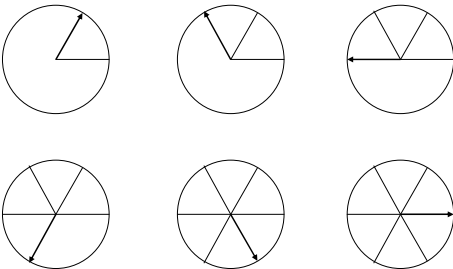


Abb. 1.57: Potenzen von $\exp(j \cdot 2 \cdot \frac{2\pi}{12})$ als eine zwölfte komplexe Wurzel von Eins: $[\exp(j\frac{2\pi}{6})]^k, 1 \leq k \leq 6$

Wurzeln unterscheiden sich nur im Vorzeichen (vgl. Bemerkungen zu Definition 1.23). Im Komplexen besitzt eine Zahl stets n komplexe n -te Wurzeln. Die Wurzel $\sqrt[n]{r} e^{j\frac{\varphi}{n}}$,

bei der $k = 0$ gewählt ist ($0 \leq \varphi < 2\pi$), bekommt dennoch einen besonderen Namen. Sie heißt der **Hauptwert** der n -ten Wurzel von $re^{j\varphi}$ und wird mit dem Symbol $\sqrt[n]{}$ benannt. Die Konvention passt zur Regel, dass im Reellen die positive Lösung der Gleichung $z^2 = x$ mit $\sqrt{x}$ bezeichnet wird, sofern sie existiert. Diese ist der Hauptwert bei komplexer Betrachtung.

Mit dem Symbol $\sqrt[n]{}$ sollte man im Komplexen vorsichtig umgehen, da aus dem Reellen bekannte Rechenregeln für den Hauptwert nicht mehr analog gelten müssen:

$$1 = \sqrt{1} = \sqrt{(-1) \cdot (-1)} \neq \sqrt{-1} \cdot \sqrt{-1} = j \cdot j = -1.$$

Hat man eine n -te Wurzel einer Zahl gefunden, so erhält man alle anderen durch fortgesetzte Multiplikation mit $e^{j\frac{2\pi}{n}}$. In diesem Sinne unterteilen also die n -ten Wurzeln den Vollwinkel 2π in n gleich große Abschnitte.

Beispiel 1.106

a) Die Lösungen der Gleichung $z^4 = 16$, also die vierten Wurzeln aus $16 = 16e^{j \cdot 0}$, sind $\sqrt[4]{16}e^{j\frac{0+2k\pi}{4}}$, $k \in \{0, 1, 2, 3\}$, also

$$2e^0 = 2, \quad 2e^{j\frac{2\pi}{4}} = 2j, \quad 2e^{j\frac{4\pi}{4}} = -2, \quad 2e^{j\frac{6\pi}{4}} = -2j.$$

b) Die dritten Wurzeln aus $27j$ sind $3e^{j\frac{\frac{\pi}{2}+2k\pi}{3}}$, $k \in \{0, 1, 2\}$, also $3e^{j\frac{\pi}{6}}$, $3e^{j\frac{5\pi}{6}}$, $3e^{j\frac{3\pi}{2}} = -3j$. ■

1.6.5 Komplexe Wechselstromrechnung*

Wir greifen das Beispiel (1.33), Seite 123, zur Überlagerung zweier harmonischer Schwingungen gleicher Frequenz nochmals auf. Wir betrachten die reellen Funktionen $u_1(t) = a \cdot \cos(\omega t)$ und $u_2(t) = b \cdot \cos(\omega t + \varphi)$ bzw. die komplexen Erweiterungen

$$\hat{u}_1(t) = a \cdot e^{j\omega t} \quad \text{und} \quad \hat{u}_2(t) = b \cdot e^{j(\omega t + \varphi)}.$$

Es gilt $u_i(t) = \text{Re}(\hat{u}_i(t))$, $i = 1, 2$. Für die Summe erhalten wir ohne Additionstheorem analog zu (1.33)

$$\hat{u}_1(t) + \hat{u}_2(t) = a \cdot e^{j\omega t} + b \cdot e^{j(\omega t + \varphi)} = (a + b \cdot e^{j\varphi}) e^{j\omega t}.$$

Die komplexe Zahl $\hat{u}_1(t) + \hat{u}_2(t)$ interpretiert als Zeiger ist die Summe der Zeiger $\hat{u}_1(t)$ und $\hat{u}_2(t)$.

Bei einer Gleichspannung kann man Ströme über das Ohm'sche Gesetz $R = \frac{U}{I}$ berechnen, wobei der Widerstand häufig in Ohm (Ω) angegeben wird. Ideale Kondensatoren sind eine Leitungsunterbrechung, ideale Spulen haben den Widerstand $R = 0$.

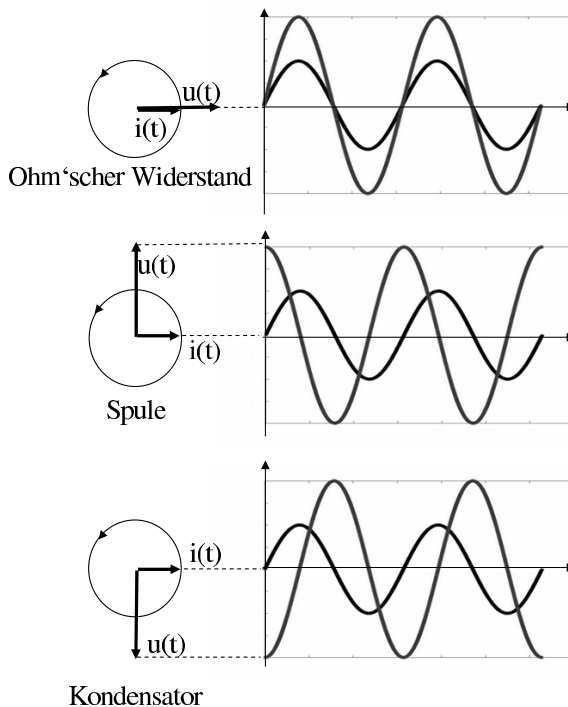


Abb. 1.58: Strom und Spannung an Wechselstromwiderständen

Das ändert sich, wenn man zeitlich nicht-konstante Spannungen und damit Wechselströme betrachtet. Die von Arthur Edwin Kennelly (1861–1939) eingeführte komplexe Wechselstromrechnung funktioniert für sinus- oder kosinusförmige Spannungen (1.32), also $u(t) = \hat{u} \cos(\omega t + \varphi_u)$. Die Idee ist nun, $u(t)$ als Realteil eines Zeigers, nämlich der komplexen Funktion

$$\underline{u}(t) := \hat{u} [\cos(\omega t + \varphi_u) + j \sin(\omega t + \varphi_u)] = \hat{u} e^{j(\omega t + \varphi_u)},$$

aufzufassen, also $u(t) = \operatorname{Re}(\underline{u}(t))$. Schreibt man eine Wechselspannung komplex, so ist der Realteil die Spannung, die man tatsächlich (also real) messen kann, der Imaginärteil dient dem Rechnen mit dieser Spannung. Dieser Anteil ist also nur gedanklich (imaginär) vorhanden.

Im Zeigerdiagramm (siehe Abbildung 1.53) ist der Funktionsgraph des Realteils unten abgebildet, während rechts der Imaginärteil als Funktion des Winkels $\varphi = \omega t + \varphi_u$ dargestellt ist. Betrachtet man entsprechend komplex erweiterte Ströme $\underline{i}(t)$ für eine feste Kreisfrequenz ω , so gilt am Ohm'schen Widerstand, einer Spule und einem Kondensator das Ohm'sche Gesetz im komplexen Bereich:

Der **komplexe Widerstand** (die **Impedanz**)

$$Z = \frac{u(t)}{i(t)} \quad (1.41)$$

ist eine komplexe Konstante (also unabhängig von t). Sie kann aber von der festen Kreisfrequenz ω abhängen. Man bezeichnet $\operatorname{Re}(Z)$ als **Wirkwiderstand** und $\operatorname{Im}(Z)$ als **Blindwiderstand**.

- Ohm'scher Widerstand: Man beobachtet, dass an einem Ohm'schen Widerstand R Strom und Spannung „in Phase“ sind. Im Zeigerdiagramm (siehe Abbildung 1.58) zeigen die Zeiger in die gleiche Richtung. Damit ist

$$Z_R = \frac{\hat{u}e^{j(\omega t + \varphi_u)}}{\hat{i}e^{j(\omega t + \varphi_u)}} = \frac{\hat{u}}{\hat{i}} = R \in \mathbb{R}, \quad \underline{i}(t) = \frac{u(t)}{R} = \underbrace{\frac{\hat{u}}{R}}_{=\hat{i}} e^{j(\omega t + \varphi_u)}.$$

- Spule: Hier werden wir später mittels Differenzialrechnung zeigen, dass für den komplexen Widerstand einer Spule mit Induktivität L gilt (siehe (2.21) auf Seite 300):

$$Z_L = j\omega L, \quad \underline{i}(t) = \frac{u(t)}{Z_L} = \frac{1}{j\omega L} \hat{u}e^{j(\omega t + \varphi_u)} = \underbrace{\frac{1}{\omega L} \hat{u}}_{=\hat{i}} \underbrace{e^{-j\frac{\pi}{2}}}_{=-j = \frac{-j^2}{j} = \frac{1}{j}} e^{j(\omega t + \varphi_u)} = \frac{\hat{u}}{\omega L} e^{j(\omega t + \varphi_u - \frac{\pi}{2})}.$$

Hier eilt demnach die Spannung dem Strom um $\frac{\pi}{2}$ voraus. Auch dies sieht man am Zeigerdiagramm (siehe Abbildung 1.58).

- Kondensator: Auch hier werden wir später mittels Differenzialrechnung ausrechnen (siehe (2.20) auf Seite 300), dass $Z_C = -\frac{j}{\omega C} = \frac{1}{j\omega C}$ ist, wobei C die Kapazität des Kondensators ist:

$$\underline{i}(t) = \frac{u(t)}{Z_C} = j\omega C \hat{u}e^{j(\omega t + \varphi_u)} = \underbrace{\omega C \hat{u}}_{=\hat{i}} \underbrace{e^{j\frac{\pi}{2}}}_{=j} e^{j(\omega t + \varphi_u)} = \hat{i}e^{j(\omega t + \varphi_u + \frac{\pi}{2})}.$$

Der Strom eilt der Spannung um 90 Grad voraus (siehe Abbildung 1.58).

Bei einer Reihenschaltung addieren sich komplexe Widerstände, bei einer Parallelschaltung bildet man den Kehrwert der Summe der Kehrwerte der Einzelwiderstände.

Man beachte, dass die komplexen Widerstände von der Kreisfrequenz ω abhängen. Im Kapitel 6.1 werden wir sehen, dass man periodische Spannungsverläufe häufig als unendliche Summe von kosinusförmigen Spannungen unterschiedlicher Kreisfrequenzen darstellen kann (Fourier-Reihe), so dass für jeden einzelnen Summanden mit komplexen Widerständen gerechnet werden kann (Überlagerung der Einzelergebnisse). Bei nicht-periodischen Spannungsverläufen versagt aber die Methode. In der Realität findet man häufig erst nach einer Einschwingzeit periodische Spannungen an den Bauteilen. Möchte man auch das Verhalten während der Einschwingzeit berechnen, muss man Differenzialgleichungen lösen. Auf Seite 733 tun wir das für einen R/L-Kreis.

Beispiel 1.107

Die komplexen Widerstände $Z_1 = (1-j) \Omega$, $Z_2 = (1+j) \Omega$ und $Z_3 = 1 \Omega$ seien parallel geschaltet. Dann ergibt sich der Gesamtwiderstand (in Ohm) zu

$$Z = \frac{1}{\frac{1}{1-j} + \frac{1}{1+j} + \frac{1}{1}} = \frac{1}{\frac{(1+j)+(1-j)+2}{2}} = \frac{1}{2}. \quad \blacksquare$$

1.6.6 Fundamentalsatz der Algebra

Wir betrachten nun Polynome mit komplexen Koeffizienten und einer komplexen Variable z . Dafür suchen wir die komplexen Nullstellen. Bei der Nullstellensuche von Polynomen zweiten Grades im Reellen mit der p-q-Formel kann man auf negative Diskriminanten stoßen. In einem solchen Fall existieren keine reellen Nullstellen. Wir haben bereits gesehen, dass es dieses Problem mit komplexen Zahlen nicht gibt. Es existieren immer zwei zweite Wurzeln. Dadurch wird das Rechnen im Komplexen einfacher als im Reellen.

Satz 1.24 (Fundamentalsatz der Algebra, D'Alembert (1717–1783))

Jedes normierte Polynom $p(z) = z^n + a_{n-1}z^{n-1} + \dots + a_1z + a_0$ mit $n \in \mathbb{N}$ und Koeffizienten $a_k \in \mathbb{C}$ (also insbesondere auch für $a_k \in \mathbb{R}$) ist darstellbar als Produkt $p(z) = \prod_{k=0}^{n-1} (z - b_k)$ von n (bis auf die Reihenfolge eindeutig bestimmten, irreduziblen) **Linearfaktoren** $(z - b_k)$, $b_k \in \mathbb{C}$. Die Zahlen b_k sind genau die Nullstellen des Polynoms.

Der Beweis des Satzes ist aufwändig und wird hier ausgelassen. In (de Jong, 2012, S. 158 f.) wird z. B. folgende Variante gezeigt: „Jedes Polynom $z^n + a_{n-1}z^{n-1} + \dots + a_1z + a_0$ besitzt eine komplexe Nullstelle z_0 “. Man erhält daraus die angegebene Form, indem man das Polynom durch $(z - z_0)$ dividiert (siehe Kapitel 1.5.6.3) und anschließend den Satz wieder auf das Ergebnis der Division anwendet usw.

Im Satz ist der Koeffizient zur höchsten Potenz z^n gleich 1, d. h., das Polynom ist normiert. Ein beliebiges Polynom $p(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_1 z + a_0$ vom Grad n (also mit $a_n \neq 0$) hat eine Darstellung

$$p(z) = a_n \cdot \prod_{k=0}^{n-1} (z - b_k) \quad (1.42)$$

mit den gleichen Nullstellen wie das normierte Polynom $z^n + \frac{a_{n-1}}{a_n} z^{n-1} + \dots + \frac{a_1}{a_n} z + \frac{a_0}{a_n}$.

Der Fundamentalsatz besagt also, dass jedes Polynom vom Grad n auch n komplexe Nullstellen hat. Dabei kann aber die gleiche Nullstelle mehrfach, z. B. m -mal vorkommen. Dann hat man m -mal den gleichen Linearfaktor und spricht auch jetzt von einer **Nullstelle mit Vielfachheit m** . Liegt ein Polynom schon als Produkt der Linearfaktoren vor, so kann man die Nullstellen direkt ablesen. Beispielsweise hat $(z-1)^3(z-2)$ die dreifache Nullstelle 1 und die einfache Nullstelle 2.

Beispiel 1.108

Das Polynom dritten Grades $z^3 + 8$ hat drei Nullstellen $2 \cdot e^{\frac{1}{3}j \cdot (\pi + 2k\pi)}$, $k \in \{0, 1, 2\}$:

$$z_1 = 2 \cdot e^{j \cdot \frac{\pi}{3}}, \quad z_2 = 2 \cdot e^{j \cdot \pi} = -2, \quad z_3 = 2 \cdot e^{j \cdot \frac{5}{3} \cdot \pi} = 2 \cdot e^{-j \cdot \frac{\pi}{3}}.$$

Es gilt $z_3 = \bar{z}_1$ und $z^3 + 8 = (z + 2) \cdot (z - 2e^{j \frac{\pi}{3}}) \cdot (z - 2e^{-j \frac{\pi}{3}})$. ■

Mittels Koeffizientenvergleich können die Koeffizienten a_k dargestellt werden über Terme, die aus den Nullstellen b_k gebildet werden. Der **Satz von Vieta** (vgl. Satz 1.19 auf Seite 97) gilt entsprechend im Komplexen.

Wir hatten bereits gesehen, dass zwei Polynome gleich sind genau dann, wenn ihr Grad übereinstimmt und sie die gleichen Koeffizienten besitzen. Mit dem Fundamentalsatz erhalten wir außerdem:

Folgerung 1.6 (Gleichheit von Polynomen)

Zwei Polynome $p(z)$ und $q(z)$ mit $\text{Grad} \leq n$ sind gleich genau dann, wenn sie an $n + 1$ Stellen gleiche Funktionswerte besitzen.

Beweis: Sind die Polynome gleich, so stimmen die Funktionswerte an allen Stellen überein. Stimmen umgekehrt p und q an $n + 1$ Stellen überein, so ist $p(z) - q(z)$ ein Polynom vom $\text{Grad} \leq n$, das $n + 1$ verschiedene Nullstellen besitzt. Da es zudem die Darstellung (1.42) besitzt, muss es das Nullpolynom sein ($a_n = 0$). Also ist $p = q$. □

Lemma 1.13 (Konjugiert komplexe Nullstellen)

Sind die Koeffizienten $a_{n-1}, a_{n-2}, \dots, a_1, a_0$ eines Polynoms $p(z) = z^n + a_{n-1}z^{n-1} + \dots + a_1z + a_0$ alle reell (wie z.B. bei $x^2 + 1$), so ist mit jeder komplexen Nullstelle z_0 auch die konjugiert komplexe Zahl $\bar{z}_0$ eine Nullstelle (wie z.B. mit j auch $-j$).

Beweis: Mit den Rechenregeln für das Konjugieren (siehe Seite 134) ergibt sich:

$$\begin{aligned} p(z_0) = 0 &\implies 0 = \bar{0} = \overline{p(z_0)} = \overline{z_0^n + a_{n-1}z_0^{n-1} + \dots + a_1z_0 + a_0} \\ &\stackrel{(1.34)}{=} \overline{z_0^n} + \overline{a_{n-1}z_0^{n-1}} + \dots + \overline{a_1z_0} + \overline{a_0} \stackrel{(1.34)}{=} \bar{z}_0^n + \bar{a}_{n-1} \cdot \bar{z}_0^{n-1} + \dots + \bar{a}_0 \\ &\stackrel{a_k \text{ reell}}{=} \bar{z}_0^n + a_{n-1} \cdot \bar{z}_0^{n-1} + \dots + a_1 \cdot \bar{z}_0 + a_0 = p(\bar{z}_0). \end{aligned}$$

□

Sind nicht alle Koeffizienten reell, so können natürlich auch nicht-konjugiert-komplexe Nullstellen auftreten, z.B. j und j bei $(x - j)^2$.

Lemma 1.14 (Vielfachheit konjugiert komplexer Nullstellen)

Ist z_0 eine echt komplexe (also nicht-reelle) Nullstelle der Vielfachheit m eines Polynoms mit reellen Koeffizienten, so ist auch $\overline{z_0}$ Nullstelle der Vielfachheit m .

Beweis: Mit z_0 ist nach dem soeben Gezeigten auch $\overline{z_0}$ eine Nullstelle: $p(z) = (z - z_0)(z - \overline{z_0})(z - z_0)^{m-1}r_1(z)$ mit einem Polynom r_1 . Da $(z - z_0)(z - \overline{z_0})$ ein reelles Polynom ist, sind auch die Koeffizienten von $(z - z_0)^{m-1}r_1(z)$ reell, und das Polynom hat mit z_0 ebenfalls die konjugiert komplexe Nullstelle $\overline{z_0}$. Damit ist $(z - z_0)^{m-1}r_1(z) = (z - z_0)(z - \overline{z_0})(z - z_0)^{m-2}r_2(z)$ für ein Polynom r_2 . Das kann man fortsetzen, bis man m Linearfaktoren $(z - \overline{z_0})$ erhält. Mehr kann es aber nicht geben, da z_0 die konjugiert komplexe Zahl zu $\overline{z_0}$ ist und man daher auch zu jedem Faktor $(z - \overline{z_0})$ einen Faktor $(z - z_0)$ findet. $\square$

Schon im Reellen haben wir Polynome in lineare und quadratische Faktoren zerlegt. Wir holen nun dazu den Beweis zu Satz 1.17 von Seite 94 nach, nach dem jedes normiertes Polynom $p(x)$ mit reellen Koeffizienten stets zerlegt werden kann in ein Produkt aus Linearfaktoren $(x - x_k)$ und/oder quadratischen Faktoren $(x - \alpha)^2 + \beta$ mit $\beta > 0$.

Beweis: Eine reelle Nullstelle x_k von $p(x)$ liefert den Faktor $(x - x_k)$. Eine echt komplexe Nullstelle $x = z_0$ von $p(x)$ tritt immer konjugiert komplex auf, d. h., mit z_0 ist auch $\overline{z_0}$ Nullstelle von $p(x)$. Für die komplexen Faktoren $(x - z_0)(x - \overline{z_0})$ erhalten wir

$$\begin{aligned}(x - z_0)(x - \overline{z_0}) &= (x - z_0)\overline{(x - z_0)} \\ &= [(x - \operatorname{Re}(z_0)) - j \operatorname{Im}(z_0)][(x - \operatorname{Re}(z_0)) + j \operatorname{Im}(z_0)] \\ &= (x - \operatorname{Re}(z_0))^2 + (\operatorname{Im}(z_0))^2.\end{aligned}$$

Das ist ein quadratischer Faktor mit $\alpha = \operatorname{Re}(z_0)$ und $\beta = (\operatorname{Im}(z_0))^2 > 0$. $\square$

Als Beispiel hat das reelle Polynom $x^3 + x = x(x^2 + 1) = (x - 0)(x - j)(x + j)$ zu den konjugiert komplexen Nullstellen j und $-j$ den reellen Faktor $x^2 + 1$.

Beispiel 1.109

a) Das kubische Polynom $p_3(z) = z^3 - 8$ hat die reelle Nullstelle $z_1 = 2$. Um die restlichen Nullstellen zu erhalten, haben wir bereits auf Seite 96 den Linearfaktor $(z - 2)$ durch Polynomdivision abgespaltet: $z^3 - 8 : z - 2 = z^2 + 2z + 4$. Jetzt können wir aber die weiteren Nullstellen berechnen:

$$z^2 + 2z + 4 = 0 \implies z_{2,3} = -1 \pm \sqrt{1 - 4} = -1 \pm \sqrt{3}j.$$

Damit erhalten wir die Zerlegung in komplexe Linearfaktoren:

$$p_3(x) = (z - 2) \cdot (z + 1 - \sqrt{3}j) \cdot (z + 1 + \sqrt{3}j).$$

Damit erhalten wir auch wieder die Zerlegung von $p_3(x)$ in irreduzible reelle Faktoren:

$$p_3(x) = (z - 2) \cdot ((z + 1)^2 + 3) = (z - 2) \cdot (z^2 + 2z + 4).$$

- b) Das Polynom vierten Grades $p_4(z) = z^4 - 2z^2 - 3$ lässt sich mit der Substitution $u = z^2$ in das quadratische Polynom $u^2 - 2u - 3$ überführen.

Mit der p-q-Formel erhalten wir

$$u_{1,2} = 1 \pm \sqrt{1^2 + 3} = 1 \pm 2, \quad \text{d. h. } u_1 = 3, \quad u_2 = -1.$$

Weiter folgt

$$z_{1,2} = \pm\sqrt{u_1} = \pm\sqrt{3} \quad \text{und} \quad z_{3,4} = \pm\sqrt{u_2} = \pm j.$$

Insgesamt ergibt sich

$$p_4(z) = (z - \sqrt{3}) \cdot (z + \sqrt{3}) \cdot (z - j) \cdot (z + j).$$

Die Zerlegung von $p_4(x)$ in irreduzible reelle Faktoren lautet $p_4(x) = (z - \sqrt{3}) \cdot (z + \sqrt{3}) \cdot (z^2 + 1)$. ■

Im Folgenden schreiben wir Polynome sowohl mit einer reellen Variable x als auch mit der komplexen Variable z . Beim Rechnen gibt es in der Regel keine Unterschiede.

Hintergrund: Eine „p-q-Formel“ für Nullstellen von Polynomen dritten Grades

Im 16. Jahrhundert gab es Rechenwettkämpfe, bei denen die Nullstellen von Polynomen zu bestimmen waren. Aus diesem Umfeld heraus sind Rechenverfahren entwickelt worden, um die Nullstellen von Polynomen dritten und vierten Grades zu bestimmen. Viel später wurde dann von Niels Abel (1802–1829) 1826 vollständig bewiesen, dass es solche (algebraischen) Verfahren für beliebige Polynome vom Grad fünf und höher nicht geben kann. In seinem Buch *Ars Magna* beschreibt Girolamo Cardano 1545 das im Folgenden angegebene Verfahren von Niccolo Tartaglia (1500–1557) für Polynome dritten Grades (**Cardan'sche Formeln**). Nullstellen von Polynomen vierten Grades können mit einem ähnlichen Algorithmus von Lodovico Ferrari (1522–1565), vgl. (Freud, 1990, S. 65 f.), berechnet werden.

Wir suchen mindestens eine Nullstelle der Gleichung

$$x^3 + bx^2 + cx + d = 0.$$

Wenn wir diese berechnet haben, können wir nach einer Polynomdivision die Nullstellen eines Polynoms zweiter Ordnung mit der p-q-Formel finden. Die Gleichung ist äquivalent mit

$$\underbrace{\left(x + \frac{b}{3}\right)^3}_{=x^3 + x^2b + x\frac{b^2}{3} + \frac{b^3}{27}} - \left(\frac{b^2}{3} - c\right)x - \frac{b^3}{27} + d = 0$$

$$\Leftrightarrow \left(x + \frac{b}{3}\right)^3 - \underbrace{\left(\frac{b^2}{3} - c\right)}_{=:p} \left(x + \frac{b}{3}\right) + \underbrace{\frac{b}{3} \left(\frac{b^2}{3} - c\right) - \frac{b^3}{27} + d}_{=:q} = 0.$$

Statt eine Lösung für x zu suchen, setzen wir $z = x + \frac{b}{3}$ und suchen eine Lösung der Gleichung

$$z^3 + pz + q = 0.$$

In dieser Gleichung gibt es keinen Summanden für z^2 , was die weitere Rechnung vereinfachen wird. Haben wir eine Lösung z gefunden, so kennen wir auch eine Lösung $x = z - \frac{b}{3}$ der Ausgangsgleichung. Jetzt wird ein Trick angewendet. Wir schreiben z als Summe von zwei neuen Variablen: $z = u + v$. Nun müssen wir Werte für u und v finden, so dass

$$\begin{aligned} (u+v)^3 + p(u+v) + q &= 0 \\ \Leftrightarrow u^3 + v^3 + 3u^2v + 3uv^2 + pu + pv + q &= 0 \\ \Leftrightarrow u^3 + v^3 + q + (u+v)(3uv + p) &= 0. \end{aligned} \quad (1.43)$$

Jetzt geht es nicht mehr mit Äquivalenzumformungen weiter. Wenn wir Werte für u und v finden können, die das (nicht-lineare) Gleichungssystem

$$\begin{aligned} u^3 + v^3 + q &= 0 \\ \wedge \quad 3uv + p &= 0 \end{aligned} \quad (1.44)$$

erfüllen, so gilt auch (1.43), und wir haben eine Nullstelle. Umgekehrt könnte es aber auch noch andere Werte für u und v geben, die zwar (1.43) erfüllen, nicht aber das Gleichungssystem, z. B. bei $q = 0$ und $p \neq 0$ die Lösung $u = v = 0$. Über das Gleichungssystem finden wir also evtl. nicht alle Nullstellen, das ist aber auch gar nicht nötig.

- Für $u \neq 0$ ergibt sich aus der zweiten Gleichung $v = -\frac{p}{3u}$ und aus der ersten

$$u^3 + q - \left(\frac{p}{3u}\right)^3 = 0 \Leftrightarrow u^6 + qu^3 - \frac{p^3}{27} = 0.$$

Nun lässt sich u^3 mit der p-q-Formel bestimmen:

$$u^3 = -\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}.$$

- Eine Lösung $u = 0$ ist nur möglich, falls $p = 0$ ist. Dann ist $v^3 = -q$.

Jetzt sind Werte für u und v und damit für $z = u + v$ bestimmt. Damit erhalten wir Nullstellen $x = u + v - \frac{b}{3}$.

Beispiel 1.110

- a) Wir berechnen die Nullstellen des Polynoms $x^3 + 1$. Dabei ist $b = 0$, $c = 0$, $d = 1$,
 $p = -\frac{b^2}{3} + c = 0$ und $q = \frac{b}{3} \left(\frac{b^2}{3} - c\right) - \frac{b^3}{27} + d = 1$.

Da $p = 0$ ist, können wir mit $u = 0$ und $v^3 = -q = -1$ die Lösungen $x = u + v - \frac{b}{3} = v$ finden. In diesem einfachen Fall vereinfacht sich die Rechnung nicht, wir hätten direkt die drei dritten komplexen Wurzeln -1 , $-e^{j\frac{2\pi}{3}}$ und $-e^{j\frac{4\pi}{3}}$ angeben können.

- b) Nun berechnen wir eine der Nullstellen 0, 1 und -1 von $x(x+1)(x-1) = x^3 - x$. Dabei ist $b = 0$, $c = -1$ und $d = 0$, $p = -\frac{b^2}{3} + c = -1$ und $q = \frac{b}{3} \left(\frac{b^2}{3} - c \right) - \frac{b^3}{27} + d = 0$. Da $p \neq 0$ ist, benutzen wir

$$u^3 = -\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} = \pm j \sqrt{\frac{1}{27}} = \pm \sqrt{\frac{1}{27}} e^{j\frac{\pi}{2}},$$

also z. B. $u = \frac{1}{\sqrt{3}} e^{j\frac{\pi}{6}}$. Dafür ist $v = \frac{-p}{3u} = \frac{1}{\sqrt{3}} e^{-j\frac{\pi}{6}}$. Damit erhalten wir $x = u + v - \frac{b}{3} = \sqrt{\frac{1}{3}} [e^{j\frac{\pi}{6}} + e^{-j\frac{\pi}{6}}] = \frac{2 \cos \frac{\pi}{6}}{\sqrt{3}} = 1$. Für $u = -\frac{1}{\sqrt{3}} e^{j\frac{\pi}{6}}$ bekommen wir analog die Lösung -1 . Die Lösung 0 können wir mit dem Verfahren nicht direkt finden. Denn u und v haben das gleiche Vorzeichen, so dass sie nur bei $u = v = 0$ entsteht und die zweite Gleichung (1.44) des Gleichungssystems wegen $p \neq 0$ nicht erfüllbar ist. ■

1.7 Lineare Gleichungssysteme und Matrizen

In diesem Abschnitt lösen wir lineare Gleichungssysteme und führen Matrizen ein. Später werden wir im Kapitel 3 (Lineare Algebra) diese Objekte in eine umfassendere Theorie einbetten.

1.7.1 Lineare Gleichungssysteme

Lineare Gleichungssysteme treten in vielfältiger Form bei der Lösung von mathematischen, physikalischen und technischen Problemstellungen auf. Beispiele sind die Berechnung der Koeffizienten eines Näherungspolynoms für eine unbekannte Funktion, die Anwendung von Rand- und Anfangsbedingungen auf einen linearen Schwinger oder die Berechnung eines elektrischen Netzwerks mit linearen Bauelementen.

Für die Behandlung vieler mathematischer Probleme der Technik, etwa bei der Berechnung von Fachwerken in der Statik, bei der Beschreibung dynamischer Systeme und bei Netzwerkberechnungen in der Elektrotechnik bedient man sich der Matrizenrechnung.

Beispiel 1.111

Wir wenden die Kirchhoff'schen Regeln auf den Gleichstrom-Schaltkreis in Abbildung 1.59 an, um die Stromstärken I_1 , I_2 und I_3 zu bestimmen. Die Knotenregel führt zur Gleichung $I_1 = I_2 + I_3$. Die Maschenregel liefert die Gleichungen $U_1 = R_1 I_2$ und

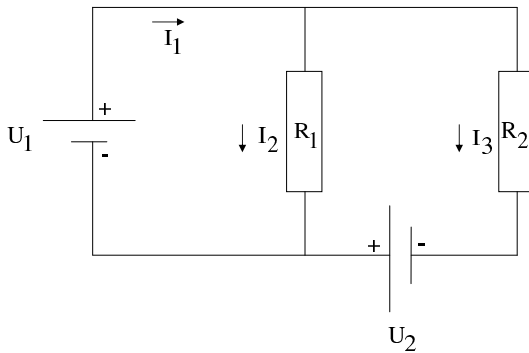


Abb. 1.59: Gleichstrom-Schaltung mit Widerständen

$U_2 = -R_1 I_2 + R_2 I_3$ (wobei die Spannungen vom Plus- zum Minuspol gerichtet sind). Wir erhalten also ein System von drei Gleichungen, wobei $R_1 \neq 0$ und $R_2 \neq 0$ seien.

$$\begin{array}{rclcl}
 I_1 & -I_2 & -I_3 & = & 0 \\
 \wedge & R_1 I_2 & & = & U_1 \\
 \wedge & -R_1 I_2 & +R_2 I_3 & = & U_2
 \end{array} \quad (1.45)$$

Bei Wechselstromkreisen mit sinusförmigen Spannungen ergeben sich bei komplexer Betrachtung komplexe lineare Gleichungssysteme. ■

Definition 1.43 (Lineares Gleichungssystem)

Ein **lineares Gleichungssystem** mit m Gleichungen in n Unbekannten $x_1, x_2, \dots, x_n$ hat die Form

$$\begin{array}{rclcl}
 a_{1,1} & x_1 & + & a_{1,2} & x_2 & + & \dots & + & a_{1,n} & x_n & = & b_1 \\
 \wedge & a_{2,1} & x_1 & + & a_{2,2} & x_2 & + & \dots & + & a_{2,n} & x_n & = & b_2 \\
 & \vdots & & & & & & & & & & \\
 \wedge & a_{m,1} & x_1 & + & a_{m,2} & x_2 & + & \dots & + & a_{m,n} & x_n & = & b_m,
 \end{array} \quad (1.46)$$

wobei $a_{l,k}, b_l \in \mathbb{C}$, $1 \leq l \leq m$, $1 \leq k \leq n$. Hier sind also für die Variablen $x_1, \dots, x_n$ Zahlenwerte so zu bestimmen, dass alle m Gleichungen erfüllt sind. Ist $b_1 = b_2 = \dots = b_m = 0$, so heißt das lineare Gleichungssystem **homogen**, anderenfalls **inhomogen**.

Das Gleichungssystem heißt linear, da die gesuchten Unbekannten $x_1, \dots, x_n$ nur in der ersten Potenz auftreten und nur über das Produkt mit Konstanten und Summation verknüpft sind. Außerdem lässt sich eine lineare Gleichung für zwei Unbekannte x_1 und x_2 der Form $a_1 x_1 + a_2 x_2 = b$ mit $a_2 \neq 0$ als Geradengleichung $x_2 = -\frac{a_1}{a_2} x_1 + \frac{b}{a_2}$ auffassen. Die Punkte auf der Geraden entsprechen den unendlich vielen Lösungen der

Gleichung. Die Lösungsmenge kann also mit einem Lineal gezeichnet werden. Hat man ein System aus zwei solchen linearen Gleichungen, so ist die Lösungsmenge die Menge der Punkte, die zugleich auf beiden Geraden liegen.

Ein homogenes lineares Gleichungssystem hat die offensichtliche (triviale) Lösung $x_1 = x_2 = \dots = x_n = 0$.

Man kann ein lineares Gleichungssystem nach verschiedenen elementaren Methoden lösen. Hier soll ein systematisches Verfahren betrachtet werden. Dieses Verfahren besteht darin, dass schrittweise eine Unbekannte nach der anderen aus dem Gleichungssystem eliminiert wird, bis sich die Lösung für die letzte verbleibende Variable unmittelbar ablesen lässt.

1.7.2 Matrizen, Zeilen- und Spaltenvektoren

Es bedeutet viel Schreibarbeit, ein Gleichungssystem in der Form (1.46) umzuformen. Die Information steckt ausschließlich in den Koeffizienten $a_{k,l}$ und den rechten Seiten b_k . Möchte man Gleichungssysteme mit Computerprogrammen bearbeiten, muss man nur diese Daten speichern. Das führt zu den folgenden Begriffen:

Definition 1.44 (Matrix)

Das rechteckige Zahlenschema

$$\mathbf{A} := \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix},$$

bei dem $m \cdot n$ Zahlen in m Zeilen und n Spalten angeordnet sind, heißt eine $(m \times n)$ -**Matrix**. Jedes **Element** $a_{i,k}$ der Matrix ist mit zwei Indizes versehen; der erste Index i ist der **Zeilenindex** $i = 1, \dots, m$, und der zweite Index ist der **Spaltenindex** $k = 1, \dots, n$. Es ist üblich, erst den Zeilen- und dann den Spaltenindex anzugeben.

Die **Zeilen** von $\mathbf{A}$ sind die horizontalen Reihen $\mathbf{A}_{i\cdot} = (a_{i,1}, a_{i,2}, \dots, a_{i,n})$, $i = 1, \dots, m$. Die **Spalten** von $\mathbf{A}$ sind die vertikalen Reihen

$$\mathbf{A}_{\cdot k} = \begin{pmatrix} a_{1,k} \\ \vdots \\ a_{m,k} \end{pmatrix}, \quad k = 1, \dots, n.$$

Das **Format** von $\mathbf{A}$ ist das Paar $(m \times n)$ aus Zeilenanzahl m und Spaltenanzahl n .

Die Menge aller $(m \times n)$ -Matrizen mit komplexen Elementen wird mit $\mathbb{C}^{m \times n}$ bezeichnet, entsprechend nennt man die Menge der $(m \times n)$ -Matrizen mit reellen Elementen $\mathbb{R}^{m \times n}$. Matrizen werden wir mit Großbuchstaben benennen (z. B. $\mathbf{A}, \mathbf{B}, \mathbf{C}$).

Wir nennen die Einträge einer Matrix Elemente. Dies ist im Zusammenhang nicht zu verwechseln mit den Elementen einer Menge. Eine andere gängige Bezeichnung für die Elemente einer Matrix ist **Komponenten**.

Beispiel 1.112

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \\ 0 & 1 & 2 & 3 \end{bmatrix}$$

hat das Format (3×4) , d. h. 3 Zeilen und 4 Spalten. Die Zeilen lauten $\mathbf{A}_{1.} = (1, 2, 3, 4)$, $\mathbf{A}_{2.} = (2, 1, 3, 4)$, $\mathbf{A}_{3.} = (0, 1, 2, 3)$, und die Spalten sind

$$\mathbf{A}_{.1} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \quad \mathbf{A}_{.2} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{A}_{.3} = \begin{pmatrix} 3 \\ 3 \\ 2 \end{pmatrix}, \quad \mathbf{A}_{.4} = \begin{pmatrix} 4 \\ 4 \\ 3 \end{pmatrix}. \quad \blacksquare$$

Definition 1.45 (Spezielle Matrizen)

- Eine **quadratische Matrix** besitzt gleiche Zeilen- und Spaltenzahl, d. h., sie hat das Format $(n \times n)$. Die Elemente $a_{1,1}, a_{2,2}, \dots, a_{n,n}$, bei denen Zeilen- und Spaltenindex übereinstimmen, heißen Elemente der **Hauptdiagonale** der Matrix.
- Eine **Zeilenmatrix** ist eine Matrix mit einer Zeile, d. h., sie hat das Format $(1 \times n)$. Statt von einer Zeilenmatrix spricht man auch von einem **Zeilenvektor** oder kurz **Vektor**.
- Eine **Spaltenmatrix** ist eine Matrix mit einer Spalte, d. h., sie hat das Format $(m \times 1)$. Statt von einer Spaltenmatrix spricht man auch von einem **Spaltenvektor** oder kurz **Vektor**.
- Bei einer $(n \times n)$ -**Nullmatrix** $\mathbf{O}$ sind alle Elemente null.
- Eine **Diagonalmatrix** $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ ist eine Matrix mit lauter Nullen außerhalb der Hauptdiagonalen:

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{bmatrix}.$$

- Die $(n \times n)$ -**Einheitsmatrix** $\mathbf{E}_n$ (oder kurz $\mathbf{E}$) ist eine Diagonalmatrix, bei der alle Diagonalelemente gleich eins sind:

$$\mathbf{E}_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Der Begriff Vektor ist damit in der Mathematik nicht eindeutig, er wird später sogar noch allgemeiner verwendet. Sowohl die Menge aller Zeilen- als auch die Menge aller Spalten-Vektoren mit m Komponenten wird mit $\mathbb{C}^m$ bzw. $\mathbb{R}^m$ bezeichnet, die Vektoren auch als **m -Tupel**. Aus dem Zusammenhang ergibt sich, ob Zeilen oder Spalten geschrieben werden. Für Vektoren werden wir Variablen verwenden, die mit einem Vektorpfeil markiert sind (z. B. $\vec{a}$, $\vec{b}$). In der Literatur gibt es hier vielfältige andere Bezeichnungen. Während wir Matrizen als Zahlenschema generell mit eckigen Klammern schreiben, werden wir Zeilen- und Spaltenmatrizen in der Regel mit runden Klammern angeben. Wir verzichten auf den zweiten, hier nicht benötigten Index. Bei Zeilenmatrizen trennen wir zur Übersichtlichkeit die einzelnen Einträge durch ein Komma, z. B.: $(1, 2, 3, 4)$.

Definition 1.46 (Gleichheit von Matrizen)

Zwei Matrizen $\mathbf{A}$ und $\mathbf{B}$ sind genau dann **gleich**, wenn sie gleiches Format ($m \times n$) haben und elementweise übereinstimmen,

$$\mathbf{A} = \mathbf{B} \iff a_{i,k} = b_{i,k}, \quad i = 1, \dots, m, \quad k = 1, \dots, n.$$

Definition 1.47 (Summe und skalar Vielfaches von Matrizen)

Es seien $\mathbf{A}$ und $\mathbf{B}$ zwei $(m \times n)$ -Matrizen und $\lambda \in \mathbb{R}$ oder $\lambda \in \mathbb{C}$. Wir nennen Zahlen λ zur Abgrenzung gegen Matrizen **Skalare**.

- Die **Summe** $\mathbf{A} + \mathbf{B}$ von zwei Matrizen $\mathbf{A}$ und $\mathbf{B}$ mit übereinstimmender Zeilenzahl und gleicher Spaltenzahl ist erklärt durch

$$\begin{bmatrix} a_{1,1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,n} \\ \vdots & & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{bmatrix} + \begin{bmatrix} b_{1,1} & \dots & b_{1,n} \\ b_{2,1} & \dots & b_{2,n} \\ \vdots & & \vdots \\ b_{m,1} & \dots & b_{m,n} \end{bmatrix} := \begin{bmatrix} a_{1,1} + b_{1,1} & \dots & a_{1,n} + b_{1,n} \\ a_{2,1} + b_{2,1} & \dots & a_{2,n} + b_{2,n} \\ \vdots & & \vdots \\ a_{m,1} + b_{m,1} & \dots & a_{m,n} + b_{m,n} \end{bmatrix}.$$

- Das **skalare Vielfache** $\lambda \cdot \mathbf{A}$ ist erklärt durch

$$\lambda \cdot \begin{bmatrix} a_{1,1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,n} \\ \vdots & & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{bmatrix} := \begin{bmatrix} \lambda a_{1,1} & \dots & \lambda a_{1,n} \\ \lambda a_{2,1} & \dots & \lambda a_{2,n} \\ \vdots & & \vdots \\ \lambda a_{m,1} & \dots & \lambda a_{m,n} \end{bmatrix}.$$

Statt $(-1) \cdot \mathbf{A}$ schreiben wir kurz $-\mathbf{A}$ und $\mathbf{A} - \mathbf{B} := \mathbf{A} + (-\mathbf{B})$.

- Für Matrizen $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k$ mit gleichem Format und Skalare $\lambda_1, \lambda_2, \dots, \lambda_k$ heißt der Ausdruck

$$\lambda_1 \cdot \mathbf{A}_1 + \lambda_2 \cdot \mathbf{A}_2 + \dots + \lambda_k \cdot \mathbf{A}_k$$

eine **Linearkombination** von $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k$ mit den **Koeffizienten** $\lambda_1, \lambda_2, \dots, \lambda_k$.

- Entsprechend sind Summe, skalaras Vielfaches und Linearkombination von Vektoren (Zeilen- bzw. Spaltenmatrizen) erklärt.

Der Begriff „Skalar“ für eine Zahl entspricht dem Sprachgebrauch in der Physik. Wir gehen darauf detaillierter in Kapitel 3.1 auf Seite 427 ein.

Für die Addition und skalare Vielfache von Matrizen gelten folgende Rechenregeln, die man sofort mit der vorangehenden Definition nachrechnet:

Satz 1.25 (Matrixaddition und Multiplikation mit einem Skalar)

Für $(m \times n)$ -Matrizen $\mathbf{A}, \mathbf{B}, \mathbf{C}$ und Skalare λ, μ gilt:

- *Kommutativgesetz:* $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$,
- *Assoziativgesetz:* $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + \mathbf{B} + \mathbf{C}$,
- *Die Nullmatrix ist neutral bezüglich der Addition:* $\mathbf{A} + \mathbf{O} = \mathbf{A}$, $\mathbf{A} + (-\mathbf{A}) = \mathbf{O}$,
- *Distributivgesetze:* $\lambda(\mathbf{A} + \mathbf{B}) = \lambda\mathbf{A} + \lambda\mathbf{B}$, $(\lambda + \mu)\mathbf{A} = \lambda\mathbf{A} + \mu\mathbf{A}$.

Beispiel 1.113

a) Für die Summe zweier (3×3) -Matrizen erhalten wir

$$\begin{bmatrix} 3 & -2 & 1 \\ -4 & 5 & -1 \\ 1 & 0 & 4 \end{bmatrix} + \begin{bmatrix} -1 & 6 & 4 \\ 2 & -3 & 2 \\ 1 & 3 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 5 \\ -2 & 2 & 1 \\ 2 & 3 & 3 \end{bmatrix}.$$

b) Ein skalaras Vielfaches einer (3×3) -Matrix:

$$3 \cdot \begin{bmatrix} 3 & -2 & 1 \\ -4 & 5 & -1 \\ 1 & 0 & 4 \end{bmatrix} = \begin{bmatrix} 9 & -6 & 3 \\ -12 & 15 & -3 \\ 3 & 0 & 12 \end{bmatrix}.$$

c) Die Linearkombination der Vektoren $\vec{a} = (1, 0, 2)$, $\vec{b} = (2, 1, 0)$, $\vec{c} = (0, 0, 1)$ mit den Koeffizienten $\lambda_1 = 1$, $\lambda_2 = -1$ und $\lambda_3 = 2$ lautet

$$\lambda_1 \cdot \vec{a} + \lambda_2 \cdot \vec{b} + \lambda_3 \cdot \vec{c} = 1 \cdot (1, 0, 2) - 1 \cdot (2, 1, 0) + 2 \cdot (0, 0, 1) = (-1, -1, 4). \quad \blacksquare$$

Definition 1.48 (Matrixmultiplikation)

Gegeben sei die $(l \times m)$ -Matrix $\mathbf{A}$ und die $(m \times n)$ -Matrix $\mathbf{B}$. Das Produkt der Matrizen $\mathbf{A}$ und $\mathbf{B}$ ist eine $(l \times n)$ -Matrix, die entsteht, wenn man die Elemente der Zeilen von $\mathbf{A}$ mit den Elementen der Spalten von $\mathbf{B}$ multipliziert und die Produkte aufaddiert („Zeile $\cdot$ Spalte“), genauer:

$$\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \vdots & \vdots & & \vdots \\ a_{l,1} & a_{l,2} & \dots & a_{l,m} \end{bmatrix} \cdot \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,n} \\ b_{2,1} & b_{2,2} & \dots & b_{2,n} \\ \vdots & \vdots & & \vdots \\ b_{m,1} & b_{m,2} & \dots & b_{m,n} \end{bmatrix} = \mathbf{C}$$

$$:= \begin{bmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} + \dots + a_{1,m}b_{m,1} & \dots & a_{1,1}b_{1,n} + a_{1,2}b_{2,n} + \dots + a_{1,m}b_{m,n} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} + \dots + a_{2,m}b_{m,1} & \dots & a_{2,1}b_{1,n} + a_{2,2}b_{2,n} + \dots + a_{2,m}b_{m,n} \\ \vdots & & \vdots \\ a_{l,1}b_{1,1} + a_{l,2}b_{2,1} + \dots + a_{l,m}b_{m,1} & \dots & a_{l,1}b_{1,n} + a_{l,2}b_{2,n} + \dots + a_{l,m}b_{m,n} \end{bmatrix}$$

mit den Elementen $c_{r,s} = \sum_{k=1}^m a_{r,k}b_{k,s}$, $1 \leq r \leq l$, $1 \leq s \leq n$.

Man beachte, dass die Anzahl m der Spalten der ersten Matrix $\mathbf{A}$ mit der Anzahl der Zeilen m der zweiten Matrix $\mathbf{B}$ übereinstimmen muss. Sonst kann man die Summen nicht bilden. Insbesondere ist $\mathbf{B} \cdot \mathbf{A}$ für $n \neq l$ nicht definiert.

Beispiel 1.114

a) Wir berechnen das Produkt zweier konkreter (2×2) -Matrizen:

$$\begin{bmatrix} 1 & 0 \\ 3 & 2 \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 0 \cdot 0 & 1 \cdot 1 + 0 \cdot 1 \\ 3 \cdot 2 + 2 \cdot 0 & 3 \cdot 1 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 6 & 5 \end{bmatrix}.$$

b) Für das allgemeine Produkt von zwei (2×2) -Matrizen $\mathbf{A}$ und $\mathbf{B}$ erhält man

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \cdot \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} = \begin{bmatrix} a_{1,1} \cdot b_{1,1} + a_{1,2} \cdot b_{2,1} & a_{1,1} \cdot b_{1,2} + a_{1,2} \cdot b_{2,2} \\ a_{2,1} \cdot b_{1,1} + a_{2,2} \cdot b_{2,1} & a_{2,1} \cdot b_{1,2} + a_{2,2} \cdot b_{2,2} \end{bmatrix} \\ = [\mathbf{A} \cdot \mathbf{B}_{\cdot 1}, \mathbf{A} \cdot \mathbf{B}_{\cdot 2}].$$

c) Entsprechend berechnet man bei (3×3) -Matrizen das Produkt über die Regel „Zeile mal Spalte“:

$$\begin{bmatrix} 3 & -2 & 1 \\ -4 & 5 & -1 \\ 1 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} -1 & 6 & 4 \\ 2 & -3 & 2 \\ 1 & 3 & -1 \end{bmatrix} = \begin{bmatrix} -6 & 27 & 7 \\ 13 & -42 & -5 \\ 3 & 18 & 0 \end{bmatrix}. \quad (1.47)$$

d) Wir multiplizieren eine (2×3) mit einer (3×2) -Matrix und erhalten eine (2×2) -Matrix. Außerdem betrachten wir die Multiplikation mit einem Zeilen- und einem Spaltenvektor:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 22 & 28 \\ 49 & 64 \end{bmatrix}, \quad (1, 2, 3) \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} = (22, 28),$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 14 \\ 32 \end{pmatrix}.$$

■

Mitunter hilft das **Falk-Schema**, eine versetzte Anordnung der Faktoren, bei der übersichtlichen Durchführung der Matrixmultiplikation. Das Produkt

$$\begin{bmatrix} 1 & 3 & 5 \\ -2 & 3 & -1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 3 & 1 & 0 \\ 3 & 5 & 0 & 0 \\ -2 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 23 & 6 & 5 \\ 7 & 8 & -3 & -1 \end{bmatrix}$$

lässt sich so berechnen:

$$\begin{array}{ccc|cccc} & & & 2 & 3 & 1 & 0 \\ & & & 3 & 5 & 0 & 0 \\ & & & -2 & 1 & 1 & 1 \\ \hline 1 & 3 & 5 & 1 & 23 & 6 & 5 \\ -2 & 3 & -1 & 7 & 8 & -3 & -1 \end{array}$$

Dabei ergibt sich jedes Element der Ergebnismatrix als Produkt des links angeordneten Zeilenvektors mit dem darüber angeordneten Spaltenvektor.

Satz 1.26 (Eigenschaften des Matrixprodukts)

a) Es gilt das **Assoziativgesetz**

$$(\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C})$$

für alle Matrizen $\mathbf{A}, \mathbf{B}, \mathbf{C}$, deren Formate die auftretenden Produkte gestatten.

b) Es gelten die **Distributivgesetze**

$$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C},$$

wenn das Format von $\mathbf{B}$ und $\mathbf{C}$ gleich ist und die Anzahl der Spalten von $\mathbf{A}$ mit der Anzahl der Zeilen von $\mathbf{B}$ und $\mathbf{C}$ übereinstimmt bzw.

$$(\mathbf{A} + \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot \mathbf{C} + \mathbf{B} \cdot \mathbf{C},$$

wenn $\mathbf{A}$ und $\mathbf{B}$ das gleiche Format haben und die Anzahl der Zeilen von $\mathbf{C}$ mit der Anzahl der Spalten von $\mathbf{A}$ und $\mathbf{B}$ übereinstimmt.

c) **Achtung:** Das Matrixprodukt ist **nicht kommutativ**, d. h., es ist i. Allg.

$$\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}.$$

d) Mit der $(n \times n)$ -Einheitsmatrix $\mathbf{E} = \mathbf{E}_n$ gilt

$$\begin{aligned} \mathbf{E}_n \cdot \mathbf{A} &= \mathbf{A} && \text{für } (n \times m) - \text{Matrizen } \mathbf{A}, \\ \mathbf{A} \cdot \mathbf{E}_n &= \mathbf{A} && \text{für } (m \times n) - \text{Matrizen } \mathbf{A}. \end{aligned}$$

Beweis: a), b) und d) lassen sich elementar nachrechnen. Für c) geben wir ein Gegenbeispiel an:

$$\begin{bmatrix} -1 & 6 & 4 \\ 2 & -3 & 2 \\ 1 & 3 & -1 \end{bmatrix} \cdot \begin{bmatrix} 3 & -2 & 1 \\ -4 & 5 & -1 \\ 1 & 0 & 4 \end{bmatrix} = \begin{bmatrix} -23 & 32 & 9 \\ 20 & -19 & 13 \\ -10 & 13 & -6 \end{bmatrix}.$$

Damit ergibt sich ein anderes Ergebnis als in (1.47). Selbst wenn $n = m$ ist, ist also eine Vertauschung i. Allg. nicht möglich. $\square$

Da beim Matrixprodukt die Reihenfolge der Faktoren wesentlich ist, sind folgende Bezeichnungen üblich. Beim Produkt $\mathbf{A} \cdot \mathbf{B}$ sagt man:

- $\mathbf{B}$ wird mit $\mathbf{A}$ **von links** multipliziert,
- $\mathbf{A}$ wird mit $\mathbf{B}$ **von rechts** multipliziert.

Beispiel 1.115 (Drehmatrix)

Die Multiplikation mit einer reellen (2×2) -Matrix ist eine Abbildung, die einen Vektor aus $\mathbb{R}^2$ auf einen anderen Vektor aus $\mathbb{R}^2$ abbildet. Fassen wir die Vektoren als Punkte der Ebene auf, so dreht die Multiplikation mit

$$\mathbf{A} := \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \quad (1.48)$$

einen Punkt um den Winkel φ im mathematisch positiven Sinn (Gegenuhreigersinn) um den Ursprung. Schreiben wir einen Punkt (x, y) als komplexe Zahl $x + jy = re^{j\vartheta}$, können wir diesen Punkt mittels $e^{j\varphi}(x + jy) = e^{j\varphi}re^{j\vartheta} = re^{j(\varphi+\vartheta)}$ um den Winkel φ drehen:

$$e^{j\varphi}(x + jy) = (\cos(\varphi) + j\sin(\varphi))(x + jy) = x\cos(\varphi) - y\sin(\varphi) + j[x\sin(\varphi) + y\cos(\varphi)],$$

also

$$\begin{pmatrix} \operatorname{Re}(e^{j\varphi}(x + jy)) \\ \operatorname{Im}(e^{j\varphi}(x + jy)) \end{pmatrix} = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \quad \blacksquare$$

Große Matrizen kann man blockweise multiplizieren, dabei müssen lediglich jeweils die Spalten- und Zeilenanzahlen der zu multiplizierenden Blöcke (Teilmatrizen) für die Multiplikation zusammenpassen. Seien $\mathbf{A}, \mathbf{B}, \dots, \mathbf{H}$ Matrizen, die zu zwei größeren

Matrizen zusammengesetzt werden. Die Spaltenzahl von $\mathbf{A}$ sei gleich der Zeilenzahl von $\mathbf{E}$ usw. Dann ergibt sich direkt aus der Definition der Matrixmultiplikation:

$$\left[\begin{array}{cccc|cccc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \mathbf{A} & \cdot & \cdot & \cdot & \mathbf{B} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \mathbf{C} & \cdot & \cdot & \cdot & \mathbf{D} & \cdot \end{array} \right] \cdot \left[\begin{array}{ccc|ccc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \mathbf{E} & \cdot & \mathbf{F} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \mathbf{G} & \cdot & \mathbf{H} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array} \right] = \left[\begin{array}{cccc|cccc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \mathbf{AE} + \mathbf{BG} & \cdot & \mathbf{AF} + \mathbf{BH} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \mathbf{CE} + \mathbf{DG} & \cdot & \mathbf{CF} + \mathbf{DH} & \cdot & \cdot & \cdot & \cdot \end{array} \right].$$

Beispiel 1.116

Das blockweise Produkt

$$\left[\begin{array}{ccc|ccc} 0 & 1 & 0 & \cdot & \cdot & \cdot \\ 1 & 0 & 2 & \cdot & \cdot & \cdot \\ \hline 1 & 1 & 2 & \cdot & \cdot & \cdot \end{array} \right] \cdot \left[\begin{array}{cc|cc} 1 & 2 & 1 & \cdot \\ 0 & 1 & 1 & \cdot \\ \hline 1 & 0 & 0 & \cdot \end{array} \right]$$

ist gleich

$$\begin{aligned} & \left[\begin{array}{ccc|ccc} \left[\begin{array}{c} 0 \\ 1 \end{array} \right] \cdot [1 \ 2] + \left[\begin{array}{cc} 1 & 0 \\ 0 & 2 \end{array} \right] \cdot \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right] & & \left[\begin{array}{c} 0 \\ 1 \end{array} \right] \cdot [1] + \left[\begin{array}{cc} 1 & 0 \\ 0 & 2 \end{array} \right] \cdot \left[\begin{array}{c} 1 \\ 0 \end{array} \right] \\ \hline [1] \cdot [1 \ 2] + [1 \ 2] \cdot \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right] & & [1] \cdot [1] + [1 \ 2] \cdot \left[\begin{array}{c} 1 \\ 0 \end{array} \right] \end{array} \right] \\ &= \left[\begin{array}{cc|cc} \left[\begin{array}{cc} 0 & 0 \\ 1 & 2 \end{array} \right] + \left[\begin{array}{cc} 0 & 1 \\ 2 & 0 \end{array} \right] & & \left[\begin{array}{c} 0 \\ 1 \end{array} \right] + \left[\begin{array}{c} 1 \\ 0 \end{array} \right] \\ \hline [1 \ 2] + [2 \ 1] & & [1] + [1] \end{array} \right] = \left[\begin{array}{cc|cc} 0 & 1 & 1 & \cdot \\ 3 & 2 & 1 & \cdot \\ \hline 3 & 3 & 2 & \cdot \end{array} \right]. \quad \blacksquare \end{aligned}$$

1.7.3 Lösen linearer Gleichungssysteme

Wir können nun das eingangs definierte lineare Gleichungssystem (1.46) über die Matrixmultiplikation formulieren:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_m \end{pmatrix}, \quad (1.49)$$

also $\mathbf{A}\vec{x} = \vec{b}$.

Definition 1.49 (Koeffizientenmatrix)

Die $(m \times n)$ -Matrix **A** in (1.49) heißt **Matrix** oder **Koeffizientenmatrix** des linearen Gleichungssystems (1.46).

Beispiel 1.117

Das Gleichungssystem (1.45) des Eingangsbeispiels sieht nun so aus:

$$\begin{bmatrix} 1 & -1 & -1 \\ 0 & R_1 & 0 \\ 0 & -R_1 & R_2 \end{bmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 0 \\ U_1 \\ U_2 \end{pmatrix}.$$

■

1.7.3.1 Gauß-Algorithmus

Wir haben jetzt viele Begriffe eingeführt, sind der Lösung eines linearen Gleichungssystems (1.46) bzw. (1.49) dadurch aber noch nicht näher gekommen. Zur Lösung entscheidend ist die Beobachtung, dass sich die Menge der Lösungen durch die folgenden Umformungen nicht ändert:

- Multiplikation einer Gleichung mit einer Zahl $c \neq 0$,
- Austausch einer Gleichung durch die Summe dieser Gleichung und einer anderen Gleichung,
- Vertauschung zweier Gleichungen.

Diese Umformungen können kombiniert und iteriert angewendet werden. Es resultiert der **Gauß-Algorithmus** zum Lösen linearer Gleichungssysteme:

a) Sorge dafür, dass x_1 nur in der ersten Gleichung auftritt:

- i) Tausche eine Gleichung nach oben, in der x_1 mit dem Faktor $a_{k,1} \neq 0$ vorkommt.
- ii) Multipliziere diese Gleichung mit $\frac{1}{a_{k,1}}$.
- iii) Subtrahiere von allen folgenden Gleichungen die mit $a_{i,1}$ multiplizierte neue erste Gleichung, wobei $a_{i,1}$ der Faktor vor x_1 in der entsprechenden Gleichung sei.

b) Sorge dafür, dass x_2 nur maximal in den ersten beiden Gleichungen auftritt. Falls x_2 nur in der ersten Gleichung auftritt, mache mit der nächsten Variable weiter.

- i) Tausche eine zweite bis m -te Gleichung nach oben, in der x_2 mit dem Faktor $a_{k,2} \neq 0$ vorkommt.
- ii) Multipliziere diese Gleichung mit $\frac{1}{a_{k,2}}$.
- iii) Subtrahiere von allen folgenden Gleichungen die mit $a_{i,2}$ multiplizierte neue zweite Gleichung, wobei $a_{i,2}$ der Faktor vor x_2 in der entsprechenden Gleichung sei. (Dann resultiert später eine Matrix in Dreiecksgestalt.) Zusätzlich kann man dies auch mit den vorangehenden Gleichungen, also hier mit der ersten, tun.

(Dann resultiert später eine Matrix in Diagonalgestalt. Der Algorithmus wird dann auch **Gauß-Jordan-Verfahren** genannt.)

- c) Verfahre nun analog für die restlichen Variablen, die in den nachfolgenden Gleichungen auftreten.

Für die effiziente Lösung eines linearen Gleichungssystems verwenden wir eine Kurzschreibweise ohne die gesuchten Variablen. Man beachte, dass sich durch die Gauß-Umformungen der Variablenvektor in (1.49) nicht verändert, denn sie sind Zeilenumformungen. Die Reihenfolge der Variablen in diesem Vektor würden sich nur bei Vertauschung von Spalten in der Matrix des Gleichungssystems ändern. Spaltenumformungen wollen wir hier aber nicht vornehmen. Daher können wir uns den Schreibaufwand für den Variablenvektor während der Umformungen ersparen.

Definition 1.50 (Erweiterte Matrix eines linearen Gleichungssystems)

Das $(m \times [n + 1])$ -Schema

$$\left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \dots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} & b_m \end{array} \right] \quad (1.50)$$

heißt **erweiterte Matrix** von (1.46).

Beispiel 1.118

Wir rechnen nun die Ströme des Eingangsbeispiels aus. Die erweiterte Matrix lautet:

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 0 \\ 0 & R_1 & 0 & U_1 \\ 0 & -R_1 & R_2 & U_2 \end{array} \right].$$

Der erste Schritt des Gauß-Algorithmus muss hier nicht ausgeführt werden, da in der ersten Spalte bereits nur eine 1 in der ersten Zeile steht. Im zweiten Schritt multiplizieren wir die zweite Gleichung mit $1/R_1$ und addieren sie anschließend zur ersten sowie das R_1 -Fache zur dritten:

$$\left[\begin{array}{ccc|c} 1 & 0 & -1 & \frac{U_1}{R_1} \\ 0 & 1 & 0 & \frac{U_1}{R_1} \\ 0 & 0 & R_2 & U_1 + U_2 \end{array} \right].$$

Jetzt multiplizieren wir die dritte Gleichung mit $\frac{1}{R_2}$ und addieren sie anschließend zur ersten:

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & \frac{U_1}{R_1} + \frac{U_1+U_2}{R_2} \\ 0 & 1 & 0 & \frac{U_1}{R_1} \\ 0 & 0 & 1 & \frac{U_1+U_2}{R_2} \end{array} \right].$$

Wir haben damit die Ströme berechnet:

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} \frac{U_1}{R_1} + \frac{U_1+U_2}{R_2} \\ \frac{U_1}{R_1} \\ \frac{U_1+U_2}{R_2} \end{pmatrix}. \quad \blacksquare$$

Achtung: Beim Rechnen mit dem erweiterten Zahlenschema darf man nur die Gauß-Umformungen anwenden. Umformungen wie die Multiplikation zweier Gleichungen oder das Potenzieren einer Gleichung führen (abgesehen von anderen Lösungsmengen) zu Produkten von Variablen und damit zu einem nicht-linearen Gleichungssystem. Das ist aber über die Matrix-Schreibweise nicht darstellbar. Falsch wäre daher eine Multiplikation der Einträge zweier Zeilen oder das Potenzieren der Komponenten einer Zeile.

Im Idealfall einer eindeutigen Lösung hat man mindestens so viele Gleichungen wie Variablen ($m \geq n$). Dann führt das Gauß-Verfahren zu einer der beiden folgenden Situationen:

■ Die Diagonalgestalt

$$\begin{array}{rcl} & 1 x_1 & = \tilde{b}_1 \\ \wedge & & \\ & 1 x_2 & = \tilde{b}_2 \\ \wedge & \dots & \\ \wedge & & \\ & 1 x_n & = \tilde{b}_n \\ \wedge & & \\ & 0 & = 0 \\ \wedge & \dots & \\ \wedge & & \\ & 0 & = 0 \end{array}$$

erhält man, wenn man auch Variablen in den vorangehenden Gleichungen eliminiert.

- Verzichtet man auf diese Elimination „nach oben“, dann kommt man zu dieser Dreiecksgestalt:

$$\begin{array}{rclcl}
 1 x_1 & + & \tilde{a}_{1,2} x_2 & + \dots + \tilde{a}_{1,n} x_n & = & \tilde{b}_1 \\
 \wedge & & 1 x_2 & + \dots + \tilde{a}_{2,n} x_n & = & \tilde{b}_2 \\
 \wedge & & \dots & & & \\
 \wedge & & & & 1 x_n & = & \tilde{b}_n \\
 \wedge & & & & 0 & = & 0 \\
 \wedge & & \dots & & & & \\
 \wedge & & & & 0 & = & 0.
 \end{array}$$

Hier kann man dann die Lösung entweder direkt ablesen oder von unten nach oben durch Einsetzen bestimmen.

Beispiel 1.119

Mit dem Gauß-Verfahren lösen wir das Gleichungssystem

$$\begin{bmatrix} 1 & j & 1 \\ 2 & j & -j \\ 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} j \\ 2+2j \\ 0 \end{pmatrix}, \quad \left[\begin{array}{ccc|c} 1 & j & 1 & j \\ 2 & j & -j & 2+2j \\ 1 & 0 & 1 & 0 \end{array} \right].$$

Wir eliminieren mit der ersten Zeile die erste Spalte in der zweiten und dritten Zeile:

$$\left[\begin{array}{ccc|c} 1 & j & 1 & j \\ 0 & -j & -2-j & 2 \\ 0 & -j & 0 & -j \end{array} \right].$$

Wir addieren die zweite Zeile zur ersten Zeile, subtrahieren sie von der dritten Zeile und multiplizieren sie dann mit j :

$$\left[\begin{array}{ccc|c} 1 & 0 & -1-j & 2+j \\ 0 & 1 & 1-2j & 2j \\ 0 & 0 & 2+j & -2-j \end{array} \right].$$

Damit ist $x_3 = -1$, $x_2 = 2j - (1-2j)(-1) = 1$, $x_1 = 2+j - (-1-j)(-1) = 1$. ■

Tipps zur Durchführung des Gauß-Algorithmus:

- Berechnet man die Lösung eines Gleichungssystems manuell, so helfen möglichst einfache Zahlenwerte. Daher ist es ratsam, Brüche durch geeignete Multiplikation der Gleichungen zu eliminieren.

- Bei der Lösung mit dem Computer oder Taschenrechner kann es wegen der endlichen Zahlendarstellung zu großen Rundungsfehlern kommen. Das betrifft insbesondere die zur Normierung verwendete Division durch den ersten Koeffizienten (Punkt ii) im Algorithmus), wenn dabei durch sehr kleine Zahlen geteilt wird. Man vermeidet dieses Problem, indem man die zu normierende Gleichung mit derjenigen Gleichung vertauscht, die in der betrachteten Spalte den betragsmäßig größten Koeffizienten besitzt. Dieses Verfahren heißt **Pivot-Suche** (engl. Pivot = Drehpunkt, Angelpunkt).

1.7.3.2 Lösungstheorie linearer Gleichungssysteme

Das Lösungsverhalten linearer Gleichungssysteme kann man anhand einfacher Beispiele in $\mathbb{R}^2$ diskutieren.

Beispiel 1.120

Das lineare System

$$\begin{array}{rcl} x_1 & +x_2 & = 2 \\ x_1 & -2x_2 & = 1 \end{array}, \quad \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

hat genau eine Lösung. Subtraktion der zweiten Gleichung von der ersten Gleichung liefert $3x_2 = 1$, also $x_2 = \frac{1}{3}$, und Einsetzen in die zweite Gleichung ergibt $x_1 = \frac{5}{3}$. Beide Geraden $x_2 = 2 - x_1$ und $x_2 = -\frac{1}{2} + \frac{1}{2}x_1$ schneiden sich in dem Punkt $(\frac{5}{3}, \frac{1}{3})$. ■

Beispiel 1.121

Das lineare System

$$\begin{array}{rcl} x_1 & +x_2 & = 2 \\ 2x_1 & +2x_2 & = 4 \end{array}, \quad \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

hat unendlich viele Lösungen. Subtraktion des doppelten der ersten Gleichung von der zweiten Gleichung liefert die Gleichung $0 \cdot x_1 + 0 \cdot x_2 = 0$, stellt also keine Bedingung an x_1, x_2 . Beide Gleichungen beschreiben dieselbe Gerade.

Setzen wir z. B. $x_2 = t \in \mathbb{R}$ beliebig, so folgt mit der ersten Gleichung $x_1 = 2 - t$. Für jedes $t \in \mathbb{R}$ erhalten wir die Lösungen $x_1 = 2 - t$, $x_2 = t$, $t \in \mathbb{R}$. ■

Beispiel 1.122

Das lineare System

$$\begin{array}{rcl} x_1 & +x_2 & = 2 \\ x_1 & +x_2 & = 1 \end{array}, \quad \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad (1.51)$$

besitzt keine Lösung. Subtraktion der zweiten Gleichung von der ersten Gleichung liefert die Gleichung $0 \cdot x_1 + 0 \cdot x_2 = 1$, also $0 = 1$, die kein x_1, x_2 erfüllen kann. Die beiden Geraden $x_2 = 2 - x_1$ und $x_2 = 1 - x_1$ sind parallel. ■

Die drei Beispiele zeigen bereits alle Situationen, die beim Lösen linearer Gleichungssysteme auftreten können: Ein lineares Gleichungssystem besitzt entweder

- eine eindeutige Lösung oder
- unendlich viele Lösungen oder
- keine Lösung.

Gibt es unendlich viele Lösungen, so spricht man von einem **unterbestimmten** Gleichungssystem. Man kann dann den Wert mindestens einer Variable frei wählen, so dass zu jeder Wahl eine Lösung existiert. Führt man Gauß-Umformungen durch, so gelangt man bei einem solchen Gleichungssystem zu einer Struktur wie

$$\begin{bmatrix} 1 & \tilde{a}_{1,2} & \dots & \tilde{a}_{1,n-2} & \tilde{a}_{1,n-1} & \tilde{a}_{1,n} \\ 0 & 1 & \dots & \tilde{a}_{2,n-2} & \tilde{a}_{2,n-1} & \tilde{a}_{2,n} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \tilde{a}_{n-2,n-1} & \tilde{a}_{n-2,n} \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-2} \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \vdots \\ \tilde{b}_{n-2} \\ 0 \\ 0 \end{pmatrix}. \quad (1.52)$$

In diesem Beispiel hat man für jede Wahl von x_n und x_{n-1} eine eindeutige Lösung für die restlichen Variablen:

$$\begin{bmatrix} 1 & \tilde{a}_{1,2} & \dots & \tilde{a}_{1,n-2} \\ 0 & 1 & \dots & \tilde{a}_{2,n-2} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-2} \end{pmatrix} = \begin{pmatrix} \tilde{b}_1 - \tilde{a}_{1,n-1}x_{n-1} - \tilde{a}_{1,n}x_n \\ \tilde{b}_2 - \tilde{a}_{2,n-1}x_{n-1} - \tilde{a}_{2,n}x_n \\ \vdots \\ \tilde{b}_{n-2} - \tilde{a}_{n-2,n-1}x_{n-1} - \tilde{a}_{n-2,n}x_n \end{pmatrix}.$$

Beispiel 1.123

$$\begin{bmatrix} 1 & 2 & 4 & 7 \\ 2 & 4 & 6 & 8 \\ 1 & 2 & 0 & -5 \\ 1 & 2 & 2 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 6 \\ 10 \\ 2 \\ 4 \end{pmatrix}, \quad \left[\begin{array}{cccc|c} 1 & 2 & 4 & 7 & 6 \\ 2 & 4 & 6 & 8 & 10 \\ 1 & 2 & 0 & -5 & 2 \\ 1 & 2 & 2 & 1 & 4 \end{array} \right] \quad (1.53)$$

Wir subtrahieren die erste Gleichung zweimal von der zweiten sowie von der dritten und vierten:

$$\left[\begin{array}{cccc|c} 1 & 2 & 4 & 7 & 6 \\ 0 & 0 & -2 & -6 & -2 \\ 0 & 0 & -4 & -12 & -4 \\ 0 & 0 & -2 & -6 & -2 \end{array} \right].$$

Jetzt sehen wir, dass das Gleichungssystem unterbestimmt ist. Wir subtrahieren die zweite Gleichung zweimal von der dritten und einmal von der vierten, anschließend multiplizieren wir sie mit $-\frac{1}{2}$:

$$\left[\begin{array}{cccc|c} 1 & 2 & 4 & 7 & 6 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right].$$

Jetzt können wir noch die zweite Gleichung viermal von der ersten abziehen:

$$\left[\begin{array}{cccc|c} 1 & 2 & 0 & -5 & 2 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right].$$

Die Lösungsmenge des linearen Gleichungssystems ist

$$\{\vec{x} \in \mathbb{R}^4 : x_2, x_4 \in \mathbb{R}, x_1 = 2 - 2x_2 + 5x_4, x_3 = 1 - 3x_4\}.$$

■

Wie bereits zuvor für (1.51) gesehen, kann es aber bei einem inhomogenen Gleichungssystem auch passieren, dass es gar keine Lösung gibt, dass sich also Gleichungen widersprechen. Beim Gauß-Algorithmus bleibt dann eine Gleichung der Form $0 = \tilde{b}_k$ (mit $\tilde{b}_k \neq 0$) übrig. Das Gleichungssystem heißt **überbestimmt**. Im homogenen Fall ist das nicht möglich, da stets $\vec{0} := (0, 0, 0, \dots, 0)$ eine Lösung ist.

1.7.4 Inverse Matrix und transponierte Matrix

Einen etwas anderen Weg zum Lösen eines Gleichungssystems $\mathbf{A}\vec{x} = \vec{b}$ hätte man, wenn man beide Seiten mit dem „Kehrwert“ von $\mathbf{A}$ multiplizieren könnte, um so die Gleichung nach $\vec{x}$ aufzulösen. Auf diese Weise müsste man außerdem nur einmal den Kehrwert, der „inverse Matrix“ heißt, berechnen und hätte dann zu jeder beliebigen Inhomogenität die Lösung über eine Multiplikation.

Wir benötigen also etwas Analoges zum Kehrwert $\frac{1}{a} = a^{-1}$ einer reellen Zahl a ungleich null. Analog zu $a \cdot a^{-1} = a^{-1} \cdot a = 1$ definieren wir:

Definition 1.51 (Inverse Matrix)

Sei $\mathbf{A} \in \mathbb{C}^{n \times n}$. Falls eine Matrix $\mathbf{B} \in \mathbb{C}^{n \times n}$ existiert mit $\mathbf{A} \cdot \mathbf{B} = \mathbf{E}$, so heißt $\mathbf{B}$ die **inverse Matrix** zu $\mathbf{A}$ und wird mit $\mathbf{A}^{-1}$ bezeichnet. In diesem Fall nennt man $\mathbf{A}$ **regulär** oder **invertierbar**. Ist $\mathbf{A}$ nicht invertierbar, existiert also keine inverse Matrix zu $\mathbf{A}$, so heißt $\mathbf{A}$ **singulär**.

Satz 1.27 (Eindeutigkeit der inversen Matrix)

- a) Mit $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{E}$ ist auch $\mathbf{A}$ die Inverse von $\mathbf{A}^{-1}$: $\mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{E}$. Man erhält also die Einheitsmatrix, egal, ob man $\mathbf{A}^{-1}$ von links oder von rechts gegen $\mathbf{A}$ multipliziert.
- b) $\mathbf{A}^{-1}$ ist eindeutig bestimmt, zu einer Matrix $\mathbf{A}$ gibt es also nicht mehrere inverse Matrizen.
- c) Sind $\mathbf{A}$ und $\mathbf{B}$ invertierbare $(n \times n)$ -Matrizen, dann ist das Produkt $\mathbf{A} \cdot \mathbf{B}$ invertierbar mit

$$(\mathbf{A} \cdot \mathbf{B})^{-1} = \mathbf{B}^{-1} \cdot \mathbf{A}^{-1}.$$

Beweis: a) Die Eigenschaft ist nicht selbstverständlich, da es normalerweise bei der Multiplikation von Matrizen auf ihre Reihenfolge ankommt. Wir werden weiter unten sehen, wie die Inverse durch simultanes Lösen von Gleichungssystemen mit dem Gauß-Verfahren berechnet werden kann. Liest man diesen Rechenweg rückwärts, so erkennt man, dass $\mathbf{A}$ mit dem gleichen Verfahren aus $\mathbf{A}^{-1}$ entsteht, so dass tatsächlich $\mathbf{A}$ die Inverse von $\mathbf{A}^{-1}$ ist.

b) Die Inverse einer Matrix ist eindeutig. Angenommen, es gibt zwei inverse Matrizen $\mathbf{B}$ und $\mathbf{C}$ von $\mathbf{A}$, so gilt mit a) $\mathbf{B} \cdot \mathbf{A} = \mathbf{E}$, und Multiplikation von rechts mit $\mathbf{C}$ liefert $(\mathbf{B} \cdot \mathbf{A}) \cdot \mathbf{C} = \mathbf{E} \cdot \mathbf{C} = \mathbf{C}$. Andererseits gilt $(\mathbf{B} \cdot \mathbf{A}) \cdot \mathbf{C} = \mathbf{B} \cdot (\mathbf{A} \cdot \mathbf{C}) = \mathbf{B} \cdot \mathbf{E} = \mathbf{B}$, woraus $\mathbf{B} = \mathbf{C}$ folgt.

c) $(\mathbf{A} \cdot \mathbf{B}) \cdot (\mathbf{B}^{-1} \cdot \mathbf{A}^{-1}) = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{B}^{-1}) \cdot \mathbf{A}^{-1} = \mathbf{A} \cdot \mathbf{E} \cdot \mathbf{A}^{-1} = \mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{E}$.

□

Beispiel 1.124

a) Wir zeigen, dass die Matrix

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ -\frac{3}{8} & \frac{1}{8} \end{bmatrix} \text{ die Inverse von } \mathbf{A} = \begin{bmatrix} 1 & -2 \\ 3 & 2 \end{bmatrix} \text{ ist.}$$

Es gilt

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \begin{bmatrix} 1 & -2 \\ 3 & 2 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ -\frac{3}{8} & \frac{1}{8} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

und

$$\mathbf{A}^{-1} \cdot \mathbf{A} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ -\frac{3}{8} & \frac{1}{8} \end{bmatrix} \cdot \begin{bmatrix} 1 & -2 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

b) Die (2×2) -Matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ ist für $ad - bc \neq 0$ invertierbar. Es ist

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix}, \quad (1.54)$$

wie man durch Multiplikation mit $\mathbf{A}$ zeigt. ■

Kennt man $\mathbf{A}^{-1}$, kann man die Lösung des Gleichungssystems $\mathbf{A}\vec{x} = \vec{b}$ direkt mit $\vec{x} = \mathbf{A}^{-1}\vec{b}$ angeben.

Zur Berechnung von $\mathbf{A}^{-1}$: Um die n Spaltenvektoren von $\mathbf{A}^{-1}$ zu bestimmen, muss man n Gleichungssysteme

$$\mathbf{A}\vec{y}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{A}\vec{y}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \mathbf{A}\vec{y}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad (1.55)$$

lösen, wobei die Inhomogenitäten genau die n Spalten der Einheitsmatrix $\mathbf{E}_n$ sind. $\mathbf{A}^{-1}$ hat nun die n Spalten $\vec{y}_1, \dots, \vec{y}_n$. Das sieht zunächst nach wesentlich mehr Arbeit als das direkte Lösen eines Gleichungssystems $\mathbf{A}\vec{x} = \vec{b}$ aus, aber dafür kann man mit $\mathbf{A}^{-1}$ sofort eine Lösung für jede Inhomogenität $\vec{b}$ angeben. Außerdem hat man nicht den n -fachen Aufwand:

Die n Gleichungssysteme bei der Matrix-Invertierung können gleichzeitig mittels des Gauß-Algorithmus gelöst werden. Dazu schreibt man statt eines Vektors rechts die Einheitsmatrix auf und macht so lange Gauß-Umformungen, bis links die Einheitsmatrix und damit rechts die Inverse der Ausgangsmatrix steht.

Beispiel 1.125

$$\left[\begin{array}{ccc|ccc} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & R_1 & 0 & 0 & 1 & 0 \\ 0 & -R_1 & R_2 & 0 & 0 & 1 \end{array} \right]$$

Wie zuvor wird die zweite Gleichung mit $1/R_1$ multipliziert, dann wird damit die zweite Spalte der übrigen Gleichungen eliminiert:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & \frac{1}{R_1} & 0 \\ 0 & 1 & 0 & 0 & \frac{1}{R_1} & 0 \\ 0 & 0 & R_2 & 0 & 1 & 1 \end{array} \right].$$

Multiplikation der letzten Gleichung mit $\frac{1}{R_2}$ und Elimination der dritten Variable in der ersten Gleichung führt zum Ziel:

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \cdot \frac{1}{R_1} + \frac{1}{R_2} \cdot \frac{1}{R_2} \\ 0 & 1 & 0 & 0 \cdot \frac{1}{R_1} + 0 \\ 0 & 0 & 1 & 0 \cdot \frac{1}{R_2} + \frac{1}{R_2} \end{array} \right]. \quad \blacksquare$$

Beispiel 1.126

$$\mathbf{A} = \left[\begin{array}{cccc} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{array} \right], \quad \left[\begin{array}{cccc|cccc} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right].$$

Wir sortieren die Zeilen um. Dann multiplizieren wir die zweite Zeile mit $1/2$ und eliminieren die übrigen Einträge der zweiten Spalte:

$$\left[\begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 & 0 & 0 \end{array} \right], \quad \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 & -\frac{1}{2} & 0 \end{array} \right].$$

Schließlich multiplizieren wir die vierte Zeile mit $\frac{1}{2}$ und erhalten die inverse Matrix:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \frac{1}{2} & -\frac{1}{4} & 0 \end{array} \right], \quad \mathbf{A}^{-1} = \left[\begin{array}{cccc} 0 & 0 & -\frac{1}{2} & 1 \\ 0 & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{1}{4} & 0 \end{array} \right]. \quad \blacksquare$$

Oft ist es praktisch (weil platzsparend), einen Spaltenvektor als Zeilenvektor zu schreiben. Um zu sagen, dass es sich doch um einen Spaltenvektor handelt, kennzeichnet man ihn mit „ $\top$ “ für „transponiert“:

Definition 1.52 (Transponierte Matrix)

Die Transponierte $\mathbf{A}^\top$ einer $(m \times n)$ -Matrix $\mathbf{A}$ ist die $(n \times m)$ -Matrix, die durch Austausch der Zeilen mit den Spalten von $\mathbf{A}$ entsteht, also durch Spiegelung an einer Diagonalen. Transponiert man einen Zeilenvektor, so entsteht ein Spaltenvektor und umgekehrt.

Beispiel 1.127

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^\top = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}, \quad (1, 2, 3)^\top = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 2 \end{pmatrix}^\top = (1, 2). \quad \blacksquare$$

Direkt aus den Definitionen ergeben sich die folgenden Rechenregeln:

Satz 1.28 (Transpositionsregeln)

Für $(m \times n)$ -Matrizen $\mathbf{A}, \mathbf{A}_1, \mathbf{A}_2$, die $(n \times k)$ -Matrix $\mathbf{B}$ und den Skalar λ gilt:

- Zweimaliges Transponieren führt zur Ausgangsmatrix: $(\mathbf{A}^\top)^\top = \mathbf{A}$.
- **Additivität:** $(\mathbf{A}_1 + \mathbf{A}_2)^\top = \mathbf{A}_1^\top + \mathbf{A}_2^\top$.
- **Homogenität:** $(\lambda \mathbf{A})^\top = \lambda \mathbf{A}^\top$.
- Die Transponierte des Produkts ist gleich dem Produkt der Transponierten in umgekehrter Reihenfolge: $(\mathbf{A} \cdot \mathbf{B})^\top = \mathbf{B}^\top \cdot \mathbf{A}^\top$.

Bei der Multiplikation von Matrizen und Matrizen mit Vektoren muss die Spaltenzahl der ersten Matrix mit der Zeilenzahl der zweiten übereinstimmen. Es macht also durchaus einen Unterschied, ob man mit einem Zeilen- oder Spaltenvektor multipliziert. Mittels Transposition kann man das jeweils richtige Format wählen.

Satz 1.29 (Inverse der Transponierten)

Für eine invertierbare $(n \times n)$ -Matrix $\mathbf{A}$ gilt

$$(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top.$$

Beweis: Für eine invertierbare $(n \times n)$ -Matrix $\mathbf{A}$ gilt: $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{E}_n \implies (\mathbf{A} \cdot \mathbf{A}^{-1})^\top = \mathbf{E}_n^\top \implies (\mathbf{A}^{-1})^\top \cdot \mathbf{A}^\top = \mathbf{E}_n$. □

Ein lineares Gleichungssystem $\mathbf{A}\vec{x} = \vec{b}$, wobei $\vec{x}$ und $\vec{b}$ Spaltenvektoren sind, kann man mittels Transposition nun auch mit Zeilenvektoren schreiben: $\vec{x}^\top \mathbf{A}^\top = \vec{b}^\top$, d. h.

$$(x_1, x_2, \dots, x_n) \begin{bmatrix} a_{1,1} & a_{2,1} & \dots & a_{m,1} \\ a_{1,2} & a_{2,2} & \dots & a_{m,2} \\ \vdots & \vdots & & \vdots \\ a_{1,n} & a_{2,n} & \dots & a_{m,n} \end{bmatrix} = (b_1, b_2, \dots, b_n).$$

Beispiel 1.128

- a) Die Matrix $\mathbf{A}$ aus (1.48), die eine Drehung beschreibt, hat eine besonders einfache Inverse: ihre Transponierte.

$$\begin{aligned}\mathbf{A}^\top \cdot \mathbf{A} &= \begin{bmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{bmatrix} \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \\ &= \begin{bmatrix} \cos^2(\varphi) + \sin^2(\varphi) & 0 \\ 0 & \sin^2(\varphi) + \cos^2(\varphi) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.\end{aligned}$$

- b) Die Matrix $\mathbf{B} := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ beschreibt keine Drehung. Hier gilt sogar $\mathbf{B} = \mathbf{B}^\top = \mathbf{B}^{-1}$. Multiplikation mit $\mathbf{B}$ vertauscht die beiden Elemente eines Vektors. Das ist eine Spiegelung an der Hauptdiagonalen des Koordinatensystems. Matrizen mit Eigenschaften wie $\mathbf{B}$ bzw. $\mathbf{A}$ erhalten spezielle Adjektive: symmetrisch bzw. orthogonal. ■

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

Definition 1.53 (Adjungierte Matrix)

Die **Adjungierte (adjungierte Matrix)** $\mathbf{A}^*$ 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. Die Regeln aus den Sätzen 1.28 und 1.29 gelten entsprechend für das Adjungieren. Lediglich bei der Homogenität muss auch der Skalar konjugiert werden: Für $\lambda \in \mathbb{C}$ gilt: $(\lambda \mathbf{A})^* = \bar{\lambda} \mathbf{A}^*$.

1.7.5 Symmetrische und orthogonale Matrizen

Einige Matrizen sind schön in dem Sinne, dass man gut mit ihnen rechnen kann und man zugehörige Probleme effizient mit dem Computer lösen kann.

Definition 1.54 (Symmetrische und schiefsymmetrische Matrix)

Eine $(n \times n)$ -Matrix $\mathbf{A}$ heißt **symmetrisch** genau dann, wenn gilt

$$\mathbf{A} = \mathbf{A}^\top \iff a_{i,k} = a_{k,i}, \quad i, k = 1, \dots, n.$$

$\mathbf{A}$ heißt **schiefsymmetrisch**, wenn gilt

$$\mathbf{A}^\top = -\mathbf{A} \iff a_{i,k} = -a_{k,i}, \quad i, k = 1, \dots, n.$$

$\mathbf{A}$ heißt **selbstadjungiert** oder **hermitesch** genau dann, wenn $\mathbf{A}$ mit der Adjungierten übereinstimmt: $\mathbf{A} = \mathbf{A}^*$.

- Bei reellen Matrizen fallen die Begriffe symmetrisch und selbstadjungiert zusammen.
- Bei einer schiefsymmetrischen Matrix müssen wegen $a_{k,k} = -a_{k,k}$ die Elemente auf der Hauptdiagonalen null sein.
- Bei einer symmetrischen $(n \times n)$ -Matrix muss man nur die n Elemente der Hauptdiagonalen und die $(n^2 - n)/2$ Elemente unterhalb (oder alternativ oberhalb) der Hauptdiagonalen kennen (und für einen Computeralgorithmus speichern), bei einer schiefsymmetrischen entfallen die n Hauptdiagonalwerte.

Haben wir später z. B. symmetrische Matrizen bei Differenzialgleichungssystemen, so fällt uns die Lösung (mit Lemma 5.3) besonders leicht, da es genügend reelle Eigenwerte (siehe Satz 3.36) gibt.

Satz 1.30 (Zerlegungssatz für quadratische Matrizen*)

Jede reelle $(n \times n)$ -Matrix $\mathbf{A}$ ist die Summe einer symmetrischen Matrix $\mathbf{R}$ und einer schiefsymmetrischen Matrix $\mathbf{S}$ mit

$$\mathbf{R} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top), \quad \mathbf{S} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top).$$

Beweis: Offensichtlich ist $\mathbf{R} + \mathbf{S} = \mathbf{A}$. Weiter gilt $\mathbf{R}^\top = \frac{1}{2}(\mathbf{A}^\top + (\mathbf{A}^\top)^\top) = \mathbf{R}$, d. h., $\mathbf{R}$ ist symmetrisch. $\mathbf{S}^\top = \frac{1}{2}(\mathbf{A}^\top - (\mathbf{A}^\top)^\top) = -\mathbf{S}$, d. h., $\mathbf{S}$ ist schiefsymmetrisch. $\square$

Beispiel 1.129

a) Die Matrix $\mathbf{A}$ ist symmetrisch, und $\mathbf{B}$ ist schiefsymmetrisch:

$$\mathbf{A} = \begin{bmatrix} 8 & -3 & 7 & 0 \\ -3 & 9 & 2 & -1 \\ 7 & 2 & 5 & -6 \\ 0 & -1 & -6 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & -3 & -1 \\ 3 & 0 & -5 \\ 1 & 5 & 0 \end{bmatrix}.$$

b) Wir zerlegen die Matrix $\mathbf{A} = \begin{bmatrix} 5 & -3 \\ -1 & 4 \end{bmatrix}$ in eine Summe aus einer symmetrischen Matrix $\mathbf{R}$ und einer schiefsymmetrischen Matrix $\mathbf{S}$. Mit dem Zerlegungssatz ist

$$\begin{aligned}\mathbf{R} &= \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) = \frac{1}{2} \left(\begin{bmatrix} 5 & -3 \\ -1 & 4 \end{bmatrix} + \begin{bmatrix} 5 & -1 \\ -3 & 4 \end{bmatrix} \right) = \begin{bmatrix} 5 & -2 \\ -2 & 4 \end{bmatrix} \\ \mathbf{S} &= \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top) = \frac{1}{2} \left(\begin{bmatrix} 5 & -3 \\ -1 & 4 \end{bmatrix} - \begin{bmatrix} 5 & -1 \\ -3 & 4 \end{bmatrix} \right) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.\end{aligned}$$

Wir erhalten also

$$\begin{bmatrix} 5 & -3 \\ -1 & 4 \end{bmatrix} = \begin{bmatrix} 5 & -2 \\ -2 & 4 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

■

Definition 1.55 (Orthogonale Matrix)

Eine reelle $(n \times n)$ -Matrix $\mathbf{A}$ heißt **orthogonal** genau dann, wenn $\mathbf{A}^{-1} = \mathbf{A}^\top$:

$$\mathbf{A} \cdot \mathbf{A}^\top = \mathbf{A}^\top \cdot \mathbf{A} = \mathbf{E}_n.$$

Eine orthogonale Matrix lässt sich also ganz leicht invertieren. Bei orthogonalen Matrizen stehen die Spaltenvektoren ebenso wie die Zeilenvektoren senkrecht zueinander. Was das bedeutet, lernen wir später in Kapitel 3.1.2, in dem wir das Skalarprodukt behandeln (siehe Seite 499). Daraus leitet sich aber der Name dieses Matrixtyps ab.

Beispiel 1.130

a) Die Matrix $\mathbf{A} = \begin{bmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{bmatrix}$ aus (1.48) ist eine orthogonale Matrix.

b) Die Matrix $\mathbf{C} = \frac{1}{\sqrt{2}} \cdot \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ ist ebenfalls orthogonal. Während $\mathbf{C}$ auch symmetrisch ist, gilt dies für $\mathbf{A}$ nur im Fall $\sin \varphi = 0$, d. h. für $\mathbf{A} = \pm \mathbf{E}_2$.

■

Satz 1.31 (Symmetrische und orthogonale Matrizen)

Ist $\mathbf{A}$ symmetrisch, so ist auch $\mathbf{A}^\top$ symmetrisch. Ist eine reelle Matrix $\mathbf{A}$ orthogonal, so ist auch $\mathbf{A}^\top$ orthogonal. Für eine symmetrische und orthogonale reelle Matrix $\mathbf{A}$ gilt:

$$\mathbf{A} \cdot \mathbf{A} = \mathbf{E}.$$

Hintergrund: Unitäre Matrizen

Arbeitet man mit komplexen Matrizen, so übernehmen die unitären Matrizen die Rolle der orthogonalen Matrizen. Eine komplexe $(n \times n)$ -Matrix $\mathbf{A}$ heißt **unitär** genau dann, wenn $\mathbf{A}^{-1} = \mathbf{A}^*$ mit $\mathbf{A}^* := \overline{\mathbf{A}^\top}$:

$$\mathbf{A} \cdot \mathbf{A}^* = \mathbf{A}^* \cdot \mathbf{A} = \mathbf{E}_n.$$

Für reelle Matrizen sind die Begriffe „orthogonal“ und „unitär“ identisch.

Die Notwendigkeit, bei komplexen Matrizen in Anwendungen zusätzlich eine komplexe Konjugation durchzuführen, ergibt sich daraus, dass die Definition des Skalarprodukts auf komplexen Vektorräumen eine komplexe Konjugation aufweist, siehe (3.22) auf Seite 486 und Bemerkung zu Folgerung 3.2 auf Seite 497.

1.7.6 Dreiecksmatrizen, Bandmatrizen und LR-Zerlegung*

Bei der mathematischen Simulation von technischen Vorgängen treten häufig sehr große Gleichungssysteme auf. Dabei achtet man bei der Modellierung darauf, dass die Matrizen viele Nullen enthalten. Bei der Verarbeitung mit Computerprogrammen werden dann nur die von null verschiedenen Einträge gespeichert.

1.7.6.1 Dreiecksmatrizen und Bandmatrizen

Folgende Bezeichnungen sind üblich:

Definition 1.56 (Dreiecksmatrizen und Bandmatrizen)

Eine quadratische Matrix, die oberhalb (unterhalb) der Hauptdiagonalen nur Nullen enthält, heißt **obere (untere) oder rechte (linke) Dreiecksmatrix**. Matrizen, die nur in der Umgebung der Hauptdiagonale Elemente ungleich null besitzen, heißen **Matrizen mit Bandstruktur oder Bandmatrizen**.

Beispiel 1.131

a) Bei den Matrizen

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 4 & 3 \\ 0 & -2 & 1 & 1 \\ 0 & 0 & 2 & 5 \\ 0 & 0 & 0 & 3 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & 0 & 0 \\ 7 & 8 & 0 \\ 1 & 2 & 3 \end{bmatrix}$$

handelt es sich um Dreiecksmatrizen. $\mathbf{A}$ ist eine rechte Dreiecksmatrix, $\mathbf{B}$ ist eine linke Dreiecksmatrix.

b) Da bei der Matrix

$$\mathbf{A} = \begin{bmatrix} a_1 & b_1 & 0 & 0 & 0 \\ c_1 & a_2 & b_2 & 0 & 0 \\ 0 & c_2 & a_3 & b_3 & 0 \\ 0 & 0 & c_3 & a_4 & b_4 \\ 0 & 0 & 0 & c_4 & a_5 \end{bmatrix}$$

nur drei „Diagonalen“ wesentlich sind, nennt man diese spezielle Bandmatrix auch **Tridiagonalmatrix**. ■

Zur Speicherung aller Elemente einer $(1\,000 \times 1\,000)$ -Matrix würde man eine Million Speicherplätze benötigen, für eine Tridiagonalmatrix mit 1 000 Reihen sind nur 2 998 Elemente wesentlich.

1.7.6.2 LR-Zerlegung

Können wir eine Matrix $\mathbf{A}$ als Produkt einer linken (unteren) Dreiecksmatrix $\mathbf{L}$ mit einer rechten (oberen) Dreiecksmatrix $\mathbf{R}$ schreiben, so wird das Lösen eines Gleichungssystems

$$\mathbf{A}\vec{x} = \vec{b} \iff \mathbf{L} \cdot \mathbf{R} \cdot \vec{x} = \vec{b}$$

ganz einfach. Denn zunächst können wir die Lösungen $\vec{y}$ von $\mathbf{L}\vec{y} = \vec{b}$ ablesen. Da $\mathbf{L}$ eine Dreiecksmatrix ist, geht das ganz einfach durch fortgesetztes Einsetzen der bereits abgelesenen Elemente von $\vec{y}$. Genauso wird für jedes $\vec{y}$ dann $\mathbf{R}\vec{x} = \vec{y}$ gelöst, und man hat die Lösungen des Ausgangssystems.

Die Zerlegung einer Matrix $\mathbf{A}$ als Produkt einer linken und einer rechten Dreiecksmatrix heißt eine **LR-Zerlegung** (oder auch **LU-Zerlegung**).

Eine LR-Zerlegung kann unter Umständen mit dem Gauß-Algorithmus berechnet werden. Beim Lösen von Gleichungssystemen haben wir Matrizen bereits in eine rechte Dreiecksmatrix gebracht. Die $\mathbf{L}$ -Matrix muss nun synchron so aufgebaut werden, dass sie die Gauß'schen Zeilenumformungen wieder rückgängig macht. Das gelingt, wenn man auf dem Weg zur rechten Dreiecksmatrix zu Zeilen nur Vielfache von vorangehenden Zeilen addieren muss. Addiert man zur i -ten Zeile das r -Fache der k -ten Zeile, so lässt sich das durch die Multiplikation einer Matrix von links ausdrücken. Diese ist die Einheitsmatrix, bei der zusätzlich an der Stelle der i -ten Zeile und k -ten Spalte der Wert r steht. Ist $k < i$, so handelt es sich um eine linke (untere) Dreiecksmatrix, ist $k > i$, so handelt es sich um eine rechte (obere) Dreiecksmatrix. Rückgängig macht man die Zeilenumformung, indem man das r -Fache der k -ten Zeile wieder abzieht. Das geschieht durch Multiplikation mit fast der gleichen Matrix. Man muss lediglich r durch $-r$ ersetzen. Im folgenden Beispiel machen wir Zeilenumformungen, die direkt durch Multiplikation mit einer unteren Dreiecksmatrix wieder rückgängig gemacht werden.

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 3 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & -3 \\ 1 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & -3 \\ 0 & -2 & -2 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & -3 \\ 0 & 0 & 4 \end{bmatrix}.
\end{aligned}$$

Im ersten Schritt haben wir die erste Zeile zweimal von der zweiten abgezogen und das durch Multiplikation mit der linken Dreiecksmatrix wieder rückgängig gemacht. Dann haben wir die erste einmal von der dritten Zeile abgezogen und schließlich das (-2) -Fache der zweiten Zeile zur dritten addiert. Diese Schritte haben wir durch eine entsprechende Matrixmultiplikation von links ebenfalls kompensiert.

Multipliziert man linke (rechte) Dreiecksmatrizen, so entstehen wieder linke (rechte) Dreiecksmatrizen. Multipliziert man speziell die hier auftretenden linken Dreiecksmatrizen, so entsteht eine linke Dreiecksmatrix, bei der die Faktoren der Gauß-Operationen mit -1 multipliziert unterhalb der Hauptdiagonalen hingeschrieben werden können: $\mathbf{A} = \mathbf{L} \cdot \mathbf{R}$ mit

$$\begin{aligned}
\mathbf{L} &= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \text{ und} \\
\mathbf{R} &= \begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & -3 \\ 0 & 0 & 4 \end{bmatrix}.
\end{aligned}$$

Dieses Verfahren zur Berechnung einer LR-Zerlegung gelingt nur, wenn man Vielfache von vorangehenden Zeilen addiert. Es gelingt nicht, wenn man dazu in einem Gauß-Schritt Zeilen vertauschen muss. Das Vertauschen von Zeilen lässt sich durch Multiplikation mit einer Matrix von links schreiben, die aus der Einheitsmatrix durch Vertauschen der betreffenden Zeilen entsteht. Diese ist aber keine Dreiecksmatrix.

Im Kasten ab Seite 627 ist für eine ganz spezielle Klasse von Matrizen ein noch effizienterer Algorithmus zur Berechnung einer LR-Zerlegung beschrieben (Cholesky-Zerlegung).

Ist $\mathbf{A}$ invertierbar, so kann man das Gleichungssystem $\mathbf{A}\vec{y} = \vec{b}$ für jede Inhomogenität $\vec{b}$ sofort lösen zu $\vec{x} = \mathbf{A}^{-1}\vec{b}$. Die LR-Zerlegung funktioniert dagegen auch für nicht-invertierbare (und nicht-quadratische) Matrizen, wenn man zunächst die Zeilen so umsortiert, dass das oben beschriebene Verfahren funktioniert. Startet man mit einem Gleichungssystem, so muss man also gegebenenfalls erst einmal die Reihenfolge der Gleichungen ändern. Hat man dann eine LR-Zerlegung von $\mathbf{A}$, so kann man das Gleichungssystem ebenfalls für jede Inhomogenität $\vec{b}$ schnell lösen. Mit der Matrix $\mathbf{L}$ hat man sich letztlich gemerkt, welche Operationen man bei einer Lösung durch das Gauß-Verfahren mit der Inhomogenität durchzuführen hat.

Hintergrund: Iterative Lösung linearer Gleichungssysteme

Wir haben bislang nur mit kleinen Matrizen gearbeitet. Bei Simulationen physikalischer Vorgänge können dagegen extrem große Matrizen mit Millionen von Zeilen und Spalten auftreten. Die Durchführung des Gauß-Algorithmus mit dem Computer kostet dann sehr viel Zeit und Speicherplatz. Günstiger sind iterative Verfahren, bei denen man sukzessive einer eindeutigen Lösung näher kommt. Wir betrachten dazu das Gleichungssystem

$$\begin{bmatrix} 2 & 0 & -1 \\ 1 & 4 & 1 \\ 0 & -2 & 4 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 8 \end{pmatrix}$$

und lösen die i -te Gleichung nach der i -ten Variable auf:

$$\begin{aligned} x &= \frac{1}{2}z + 1 \\ y &= -\frac{1}{4}x - \frac{1}{4}z + 1 \\ z &= \frac{1}{2}y + 2. \end{aligned}$$

Nun gewinnen wir daraus eine Iterationsvorschrift (**Jacobi-Verfahren**, **Gesamt-schritt-Verfahren**):

$$\begin{aligned} x_{k+1} &:= \frac{1}{2}z_k + 1 \\ y_{k+1} &:= -\frac{1}{4}x_k - \frac{1}{4}z_k + 1 \\ z_{k+1} &:= \frac{1}{2}y_k + 2. \end{aligned}$$

Startet man mit zufälligen Werten für x_0 , y_0 und z_0 , z. B. $x_0 = y_0 = z_0 = 0$, und berechnet dann sukzessive über dieses Schema neue Werte, so kommt man der tatsächlichen Lösung $x = 2$, $y = 0$, $z = 2$ immer näher. Die ersten Iterationen liefern: $(x_1, y_1, z_1) = (1, 1, 2)$, $(x_2, y_2, z_2) = (2, \frac{1}{4}, \frac{5}{2})$, $(x_3, y_3, z_3) = (\frac{9}{4}, -\frac{1}{8}, \frac{17}{8})$ usw.

Man erhält eine Verbesserung des Jacobi-Verfahrens zum **Gauß-Seidel-Verfahren (Einzelschritt-Verfahren)**, indem man bei jedem Iterationsschritt bereits auf die in diesem Schritt zuvor berechneten Variablen (einzeln) zurückgreift. Man geht dabei davon aus, dass diese Werte im Allgemeinen der Lösung näher sind als die Werte aus dem vorangehenden Schritt.

$$\begin{aligned}x_{k+1} &:= \frac{1}{2}z_k + 1 \\y_{k+1} &:= -\frac{1}{4}x_{k+1} - \frac{1}{4}z_k + 1 \\z_{k+1} &:= \frac{1}{2}y_{k+1} + 2.\end{aligned}$$

Der Name Einzelschrittverfahren weist auf die Verwendung einzelner Variablenwerte als Zwischenergebnisse hin. Beim Gesamtschrittverfahren wird dagegen nur die gesamte und unveränderte, im Vorgängerschritt berechnete Näherungslösung verwendet.

Wir schreiben die beiden Verfahren etwas formaler mittels Matrizen auf. Dazu zerlegen wir die Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ des linearen Gleichungssystems $\mathbf{A}\vec{x} = \vec{b}$ in die Summe

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$$

gemäß

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ a_{2,1} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n-1,1} & a_{n-1,2} & \dots & 0 & 0 \\ a_{n,1} & a_{n,2} & \dots & a_{n,n-1} & 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} a_{1,1} & 0 & \dots & 0 \\ 0 & a_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{n,n} \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} 0 & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ 0 & 0 & a_{2,3} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & a_{n-1,n} \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

Dann lässt sich das Jacobi-Verfahren schreiben in der Form

$$\mathbf{D} \cdot \vec{x}^{(k+1)} = -(\mathbf{L} + \mathbf{U}) \cdot \vec{x}^{(k)} + \vec{b},$$

wobei $\vec{x}^{(k)}$ der Ergebnisvektor des k -ten Rechenschritts ist. Entsprechend lautet das Gauß-Seidel-Verfahren

$$\begin{aligned}\mathbf{D} \cdot \vec{x}^{(k+1)} &= -\mathbf{L} \cdot \vec{x}^{(k+1)} - \mathbf{U} \cdot \vec{x}^{(k)} + \vec{b} \\ \iff (\mathbf{D} + \mathbf{L}) \cdot \vec{x}^{(k+1)} &= -\mathbf{U} \cdot \vec{x}^{(k)} + \vec{b}.\end{aligned}$$

Beide Verfahren besitzen damit die gleiche Struktur

$$\vec{x}^{(k+1)} = \mathbf{B} \cdot \vec{x}^{(k)} + \vec{c} \quad (1.56)$$

mit (falls die Inversen existieren)

$$\mathbf{B} = -\mathbf{D}^{-1} \cdot (\mathbf{L} + \mathbf{U}), \quad \vec{c} = \mathbf{D}^{-1} \vec{b} \quad \text{bzw.} \quad \mathbf{B} = -(\mathbf{D} + \mathbf{L})^{-1} \cdot \mathbf{U}, \quad \vec{c} = (\mathbf{D} + \mathbf{L})^{-1} \vec{b}.$$

Die exakte Lösung $\vec{x}$ des Gleichungssystems erfüllt dann die Fixpunktgleichung

$$\vec{x} = \mathbf{B} \cdot \vec{x} + \vec{c}, \quad (1.57)$$

so dass der Fehler des $k + 1$ -ten Iterationschritts darstellbar ist als

$$\vec{x}^{(k+1)} - \vec{x} = \mathbf{B} \cdot (\vec{x}^{(k)} - \vec{x}) = \mathbf{B}^2 \cdot (\vec{x}^{(k-1)} - \vec{x}) = \dots = \mathbf{B}^{(k+1)} (\vec{x}^{(0)} - \vec{x}). \quad (1.58)$$

Dabei ist $\mathbf{B}^2 := \mathbf{B} \cdot \mathbf{B}$, usw. Die Verfahren funktionieren z. B., wenn die Potenzen der Matrix $\mathbf{B}$ sich der Nullmatrix annähern. Wir schauen uns das später etwas genauer an.

Ein Trick, um möglicher Weise schneller zu einer guten Näherungslösung zu kommen, besteht in der Verwendung eines Relaxationsparameters $0 < \omega < 2$. Das **SOR-Verfahren** (Successive-Overrelaxation) kombiniert das Ergebnis eines Gauß-Seidel-Schritts mit dem Resultat des vorausgehenden Schritts:

$$\vec{x}^{(k+1)} = \omega(\mathbf{B} \cdot \vec{x}^{(k)} + \vec{c}) + (1 - \omega)\vec{x}^{(k)} = \omega(-(\mathbf{D} + \mathbf{L})^{-1} \cdot \mathbf{U} \cdot \vec{x}^{(k)} + (\mathbf{D} + \mathbf{L})^{-1} \vec{b}) + (1 - \omega)\vec{x}^{(k)}.$$

Für $\omega = 1$ ist dies das Gauß-Seidel-Verfahren.

Mittels mehrdimensionaler Differenzialrechnung lassen sich weitere iterative Lösungsverfahren gewinnen, siehe Beispiel 4.23 auf Seite 623.

1.8 Determinanten

Der Name Determinante kommt von ihrer Rolle als entscheidende (determinierende) Größe, welche bestimmt, ob ein lineares Gleichungssystem eindeutig lösbar ist oder nicht. Ferner trifft man auf Determinanten in der Flächen- und Volumenberechnung. Die eindeutige Lösung eines linearen Gleichungssystems mit zwei Unbekannten

$$\begin{aligned} a_{1,1} x_1 + a_{1,2} x_2 &= b_1 \\ a_{2,1} x_1 + a_{2,2} x_2 &= b_2 \end{aligned}$$

lautet bei Existenz der inversen Matrix (1.54):

$$x_1 = \frac{a_{2,2} b_1 - a_{1,2} b_2}{a_{1,1} a_{2,2} - a_{2,1} a_{1,2}}, \quad x_2 = \frac{a_{1,1} b_2 - a_{2,1} b_1}{a_{1,1} a_{2,2} - a_{2,1} a_{1,2}}.$$

Offenbar ist diese Lösung nur dann definiert, wenn man nicht durch die Null teilt. Das ist genau dann der Fall, wenn die Determinante von $\mathbf{A}$ von null verschieden ist:

$$\det \mathbf{A} = \begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} := a_{1,1} a_{2,2} - a_{2,1} a_{1,2}.$$

Beispiel 1.132

$$\begin{vmatrix} 1 & 3 \\ 2 & 4 \end{vmatrix} = 1 \cdot 4 - 3 \cdot 2 = -2. \quad \blacksquare$$

1.8.1 Definition und elementare Eigenschaften von Determinanten

Wir führen nun die Determinante etwas allgemeiner ein, so dass wir mit ihr für beliebige Gleichungssysteme mit $(n \times n)$ -Matrizen berechnen können, ob sie eindeutig lösbar sind. Außerdem können wir dann mit Determinanten die Lösung ähnlich wie im obigen Fall einer (2×2) -Matrix sogar direkt angeben. Dies geschieht mit der Cramer'schen Regel (Satz 1.39). Wir werden sehen, dass die allgemeinere Definition für (2×2) -Matrizen genau der Größe aus dem vorangehenden Beispiel entspricht.

Definition 1.57 (Determinante)

Gegeben sei die quadratische reelle (oder komplexe) $(n \times n)$ -Matrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}.$$

Die **Determinante** von $\mathbf{A}$ ist eine reelle (oder komplexe) Zahl, die mit $\det \mathbf{A}$ bezeichnet wird. Sie ist über die folgende Rekursion (Rechenvorschrift, die analog einer Vollständigen Induktion aufgebaut ist) definiert:

- Ist $n = 1$ und $\mathbf{A} = [a_{1,1}]$, so ist $\det[a_{1,1}] := a_{1,1}$.
- Ist $n > 1$ und bezeichne $\mathbf{A}_{l,k}$ die $((n-1) \times (n-1))$ -Matrix, die aus $\mathbf{A}$ durch Weglassen der l -ten Zeile und k -ten Spalte entsteht, so ist

$$\begin{aligned} \det \mathbf{A} &:= a_{1,1} \det \mathbf{A}_{1,1} - a_{1,2} \det \mathbf{A}_{1,2} + a_{1,3} \det \mathbf{A}_{1,3} - a_{1,4} \det \mathbf{A}_{1,4} + \dots \\ &= a_{1,1} \det \begin{bmatrix} a_{2,2} & a_{2,3} & \dots & a_{2,n} \\ a_{3,2} & a_{3,3} & \dots & a_{3,n} \\ \vdots & \vdots & & \vdots \\ a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{bmatrix} - a_{1,2} \det \begin{bmatrix} a_{2,1} & a_{2,3} & \dots & a_{2,n} \\ a_{3,1} & a_{3,3} & \dots & a_{3,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,3} & \dots & a_{n,n} \end{bmatrix} \end{aligned}$$

$$+ - \cdots = \sum_{k=1}^n (-1)^{k+1} a_{1,k} \det \mathbf{A}_{1,k}.$$

Eine andere Schreibweise für die Determinante nutzt Betragsstriche:

$$\det \mathbf{A} = \begin{vmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{vmatrix}.$$

Achtung: Die Verwendung der Betragsstriche als Kurzschreibweise bedeutet nicht, dass die entstehende Zahl größer oder gleich null ist. Determinanten können sehr wohl negative Zahlen sein.

Häufig benutzt man Determinanten von (2×2) -Matrizen. Man beachte, dass dafür die Definition 1.57 zur Zahl aus dem einführenden Beispiel führt:

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1} \det[a_{2,2}] - a_{1,2} \det[a_{2,1}] = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}.$$

Bei einer reellen (2×2) -Matrix ist der Betrag der Determinante gleich dem Flächeninhalt eines Parallelogramms mit den vier Eckpunkten $(0,0)$, $(a_{1,1}, a_{1,2})$, $(a_{1,1} + a_{2,1}, a_{1,2} + a_{2,2})$ und $(a_{2,1}, a_{2,2})$. Um das zu sehen, betrachten wir das Parallelogramm in Abbildung 1.60. Wir erhalten den Flächeninhalt, indem wir vom Inhalt

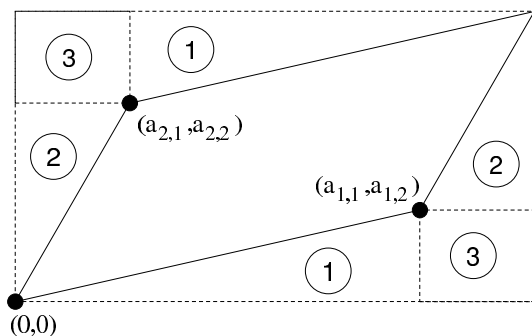


Abb. 1.60: Determinante als Flächeninhalt eines Parallelogramms

$(a_{1,1} + a_{2,1})(a_{1,2} + a_{2,2})$ des umschließenden Rechtecks die nummerierten Flächen abziehen. Die Dreiecke 1 haben jeweils den Flächeninhalt $\frac{1}{2} \cdot a_{1,1} \cdot a_{1,2}$, die Dreiecke 2 haben den Flächeninhalt $\frac{1}{2} \cdot a_{2,1} \cdot a_{2,2}$ und die Rechtecke 3 die Größe $a_{2,1} \cdot a_{1,2}$. Insgesamt ist der Flächeninhalt des Parallelogramms gleich

$$(a_{1,1} + a_{2,1})(a_{1,2} + a_{2,2}) - 2 \cdot \frac{1}{2} a_{1,1} a_{1,2} - 2 \cdot \frac{1}{2} a_{2,1} a_{2,2} - 2 a_{2,1} a_{1,2}$$

$$\begin{aligned}
&= a_{1,1}a_{1,2} + a_{1,1}a_{2,2} + a_{2,1}a_{1,2} + a_{2,1}a_{2,2} - a_{1,1}a_{1,2} - a_{2,1}a_{2,2} - 2a_{2,1}a_{1,2} \\
&= a_{1,1}a_{2,2} - a_{2,1}a_{1,2} = \det \mathbf{A}.
\end{aligned} \tag{1.59}$$

Für den Spezialfall von (3×3) -Matrizen kann man die Berechnung etwas einfacher als in der Definition gestalten:

Satz 1.32 (Sarrus'sche Regel)

Sei $\mathbf{A} \in \mathbb{R}^{3 \times 3}$ (oder $\mathbb{C}^{3 \times 3}$).

$$\begin{aligned}
\det \mathbf{A} = & +a_{1,1}a_{2,2}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2} \\
& - a_{1,3}a_{2,2}a_{3,1} - a_{1,1}a_{2,3}a_{3,2} - a_{1,2}a_{2,1}a_{3,3}.
\end{aligned}$$

Diese Regeln kann man direkt mit der Definition der Determinante nachrechnen. Sie lässt sich über das folgende Schema merken, bei dem die Matrix zyklisch fortgesetzt wird und man die Diagonalen multipliziert und die Produkte geeignet aufaddiert:

$$\begin{array}{ccccc}
& \mathbf{a}_{1,1} & \mathbf{a}_{1,2} & \mathbf{a}_{1,3} & \\
+ & a_{2,1} & \mathbf{a}_{2,2} & \mathbf{a}_{2,3} & \\
& a_{3,1} & a_{3,2} & \mathbf{a}_{3,3} & \\
\hline
& a_{1,1} & a_{1,2} & & \\
& \mathbf{a}_{2,1} & a_{2,2} & & \\
& \mathbf{a}_{3,1} & \mathbf{a}_{3,2} & & \\
\hline
& a_{1,1} & a_{1,2} & \mathbf{a}_{1,3} & \\
- & a_{2,1} & \mathbf{a}_{2,2} & \mathbf{a}_{2,3} & \\
& \mathbf{a}_{3,1} & \mathbf{a}_{3,2} & \mathbf{a}_{3,3} & \\
\hline
& \mathbf{a}_{1,1} & \mathbf{a}_{1,2} & & \\
& \mathbf{a}_{2,1} & a_{2,2} & & \\
& a_{3,1} & a_{3,2} & &
\end{array}$$

Beispiel 1.133

Anwendung der Sarrus'schen Regel liefert

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = 1 \cdot 5 \cdot 9 + 2 \cdot 6 \cdot 7 + 3 \cdot 4 \cdot 8 - 3 \cdot 5 \cdot 7 - 1 \cdot 6 \cdot 8 - 2 \cdot 4 \cdot 9 = 0.$$

Achtung: Für $n > 3$ werden solche Regeln zu komplex. Insbesondere gibt es **keine** einfache Erweiterung der Sarrus'schen Regel für den Fall $n > 3$. Diese kann nur für (3×3) -Matrizen eingesetzt werden. Ein häufig gemachter Fehler besteht in der analogen Anwendung auf größere Matrizen. Auch die sehr ähnliche Rechenregel für (2×2) -Matrizen $\mathbf{A}$ entspricht nicht einer Übertragung der Sarrus'schen Regel. Hier wird nur eine positiv und eine negativ gewichtete Diagonale betrachtet: $\det \mathbf{A} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$. Das entspricht dem Schema

$$\begin{array}{cc}
+ & \mathbf{a}_{1,1} & a_{1,2} \\
& a_{2,1} & \mathbf{a}_{2,2}
\end{array}
-
\begin{array}{cc}
a_{1,1} & \mathbf{a}_{1,2} \\
\mathbf{a}_{2,1} & a_{2,2}
\end{array}$$

und **nicht**

$$\begin{array}{cc}
+ & \mathbf{a}_{1,1} & \mathbf{a}_{1,2} \\
& a_{2,1} & \mathbf{a}_{2,2}
\end{array}
\begin{vmatrix} a_{1,1} & & \\ & a_{1,2} & \\ & \mathbf{a}_{2,1} & \end{vmatrix}
-
\begin{array}{cc}
a_{1,1} & \mathbf{a}_{1,2} \\
\mathbf{a}_{2,1} & \mathbf{a}_{2,2}
\end{array}
\begin{vmatrix} \mathbf{a}_{1,1} & & \\ & \mathbf{a}_{1,2} & \\ & a_{2,1} & \end{vmatrix}.$$

Für Matrizen beliebiger Größe ist hingegen der folgende Satz anwendbar:

Satz 1.33 (Laplace'scher Determinanten-Entwicklungssatz)

Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ (oder $\mathbb{C}^{n \times n}$). Die Determinante von $\mathbf{A}$ kann nach jeder Zeile oder jeder Spalte entwickelt werden:

- Entwicklung nach der i -ten Zeile (für $i = 1$ ist das die Definition):

$$\det \mathbf{A} = \sum_{k=1}^n (-1)^{k+i} a_{i,k} \det \mathbf{A}_{i,k}.$$

- Entwicklung nach der k -ten Spalte:

$$\det \mathbf{A} = \sum_{i=1}^n (-1)^{k+i} a_{i,k} \det \mathbf{A}_{i,k}.$$

Determinanten können mit Hilfe von Permutationen (vgl. Seite 32) geschrieben werden. Die entsprechende Aussage heißt Leibniz-Formel. Diese wird hier nicht behandelt. Der Entwicklungssatz basiert direkt auf der Leibniz-Formel.

Die Vorzeichen $(-1)^{k+i}$ der Summanden der Zeile oder Spalte, nach der entwickelt wird, kann man dem folgenden Schema entnehmen:

$$\begin{bmatrix} + & - & + & - & + & - & \dots \\ - & + & - & + & - & + & \dots \\ + & - & + & - & + & - & \dots \\ - & + & - & + & - & + & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Beispiel 1.134

Wir entwickeln die Determinante nach der zweiten Zeile:

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = -4 \begin{vmatrix} 2 & 3 \\ 8 & 9 \end{vmatrix} + 5 \begin{vmatrix} 1 & 3 \\ 7 & 9 \end{vmatrix} - 6 \begin{vmatrix} 1 & 2 \\ 7 & 8 \end{vmatrix} = -4(18-24) + 5(9-21) - 6(8-14)$$

$= 24 - 60 + 36 = 0$. Es bietet sich dabei insbesondere die Entwicklung nach Zeilen oder Spalten an, in denen viele Nullen stehen, hier nach der ersten Spalte:

$$\begin{vmatrix} 0 & 2 & 3 \\ 4 & 5 & 6 \\ 0 & 8 & 9 \end{vmatrix} = -4 \begin{vmatrix} 2 & 3 \\ 8 & 9 \end{vmatrix} = -4(18-24) = 24. \quad \blacksquare$$

Folgerung 1.7 (Eigenschaften der Determinante)

a) Die Determinante ist linear in jeder Spalte der Matrix, d. h., für $\mathbf{A} \in \mathbb{C}^{n \times n}$, $\vec{b} \in \mathbb{C}^n$ und $c \in \mathbb{C}$ gilt:

$$\begin{vmatrix} a_{1,1} & \dots & a_{1,k-1} & c \cdot a_{1,k} + b_1 & a_{1,k+1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,k-1} & c \cdot a_{2,k} + b_2 & a_{2,k+1} & \dots & a_{2,n} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{n,1} & \dots & a_{n,k-1} & c \cdot a_{n,k} + b_n & a_{n,k+1} & \dots & a_{n,n} \end{vmatrix} \\ = c \cdot \det \mathbf{A} + \begin{vmatrix} a_{1,1} & \dots & a_{1,k-1} & b_1 & a_{1,k+1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,k-1} & b_2 & a_{2,k+1} & \dots & a_{2,n} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{n,1} & \dots & a_{n,k-1} & b_n & a_{n,k+1} & \dots & a_{n,n} \end{vmatrix}.$$

- b) Die Determinante ist linear in jeder Zeile der Matrix.
- c) Der Wert der Determinante ändert sich nicht, wenn man Zeilen und Spalten der Matrix vertauscht, d. h. $\det \mathbf{A} = \det \mathbf{A}^\top$. Führt man zusätzlich zur Transposition eine komplexe Konjugation aller Komponenten der Matrix durch, so erhält man die Adjungierte $\mathbf{A}^*$. Für diese gilt: $\det \mathbf{A}^* = \overline{\det \mathbf{A}}$.
- d) Vertauscht man genau zwei Zeilen oder zwei Spalten, so ändert sich das Vorzeichen der Determinante.
- e) Stehen in einer Zeile oder einer Spalte lauter Nullen, so ist die Determinante gleich null.
- f) Sind zwei Zeilen oder Spalten gleich, so ist die Determinante gleich null.
- g) Addiert man zu einer Zeile (bzw. Spalte) ein Vielfaches einer anderen Zeile (bzw. Spalte), so ändert sich der Wert der Determinante nicht.

Beweis: a) Die Linearität ergibt sich sofort, wenn wir die Determinante nach der k -ten Spalte entwickeln. Sie hat aber weitreichende Konsequenzen für die Bedeutung der Determinante. Man kann sie auch über diese Eigenschaft definieren.

b) ergibt sich analog bei Entwicklung nach der entsprechenden Zeile.

c) $\det \mathbf{A} = \det \mathbf{A}^\top$ folgt ebenfalls direkt aus dem Entwicklungssatz, indem man von einer Entwicklung nach Zeilen zu einer Entwicklung nach Spalten übergeht. Die folgenden Aussagen müssen wir also nur entweder für Spalten oder für Zeilen zeigen. Ist $\bar{\mathbf{A}}$ die Matrix, die aus $\mathbf{A}$ durch komplexe Konjugation aller Komponenten hervorgeht, so folgt mit den Rechenregeln für das komplexe Konjugieren sofort $\det \bar{\mathbf{A}} = \overline{\det \mathbf{A}}$. Damit ist

$$\det \mathbf{A}^* = \det \bar{\mathbf{A}}^\top = \overline{\det \mathbf{A}^\top} = \overline{\det \mathbf{A}}.$$

- d) folgt ebenfalls aus dem Entwicklungssatz: Vertauscht man beispielsweise zwei benachbarte Spalten und entwickelt jeweils nach der vor der Verschiebung ersten Spalte, so ändern sich für diese Spalte alle Vorzeichen, aber die Unterdeterminanten bleiben gleich. Man kann nun zwei beliebige Spalten $k < l$ vertauschen, indem man $l - k + l - k - 1$ benachbarte Spalten paarweise vertauscht (wie beim Sortieralgorithmus „Bubblesort“): Mit $l - k$ Vertauschungen bringt man die rechte Spalte zur Position der zuvor linken. Diese steht danach an der Position $k + 1$, so dass man diese mit $l - k + 1$ Vertauschungen an die Position l bringt. $2(l - k) - 1$ ist eine ungerade Anzahl, so dass sich das Vorzeichen der Determinante ungradzahlig oft ändert, es kehrt sich also insgesamt um.
- e) Man entwickelt nach dieser Zeile bzw. Spalte und erhält null.
- f) Vertauscht man die beiden Zeilen oder Spalten, so ändert sich das Vorzeichen. Da die Matrix gleich bleibt, muss auch die Determinante gleich bleiben – das geht nur, wenn sie null ist.
- g) Wegen a) erhält man die Summe zweier Determinanten, wobei eine für eine Matrix mit zwei gleichen Zeilen (bzw. Spalten) gebildet wird und damit null ist. $\square$

Diese Regeln können nun genutzt werden, um Determinanten zu berechnen. Dabei kann man sowohl die vom Gauß-Algorithmus bekannten Zeilenumformungen durchführen als auch entsprechende Umformungen der Spalten. Ohne Änderung des Determinanten-Werts darf man

- Vielfache von Zeilen zu anderen Zeilen addieren,
- Vielfache von Spalten zu anderen Spalten addieren.

Der Wert der Determinante ändert sich bei folgenden Operationen:

- Multiplikation einer Zeile oder Spalte mit einem Skalar c . Dadurch wird auch der Wert der Determinante mit c multipliziert.
- Vertauschung zweier Zeilen oder zweier Spalten. Dadurch ändert sich das Vorzeichen der Determinante.

Beispiel 1.135

- a) Wir berechnen die folgende Determinante, indem wir von links nach rechts elementare Umformungen durchführen:

- Doppelte Addition der ersten Zeile zur dritten Zeile,
- Subtraktion des Doppelten der ersten Zeile von der zweiten Zeile,
- Subtraktion des Doppelten der zweiten Zeile von der dritten Zeile:

$$\begin{vmatrix} 1 & 4 & 0 \\ 2 & 10 & -1 \\ -2 & -4 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 4 & 0 \\ 2 & 10 & -1 \\ 0 & 4 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 4 & 0 \\ 0 & 2 & -1 \\ 0 & 4 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 4 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{vmatrix}.$$

Die Entwicklung nach der ersten, zweiten und dritten Spalte ergibt für die Determinante den Wert $1 \cdot 2 \cdot 3 = 6$.

- b) Bei der folgenden Determinante subtrahieren wir zunächst die zweite Spalte von der ersten, danach die dritte Zeile von der zweiten und wenden anschließend die Sarrus'sche Regel an:

$$\begin{vmatrix} 3 & 1 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \end{vmatrix} = \begin{vmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 1 & 3 \end{vmatrix} = \begin{vmatrix} 2 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 3 \end{vmatrix} = -(-1) \cdot 1 \cdot 2 = 2.$$

- c) Zur Berechnung der folgenden Determinante liefern

- Addition von Vielfachen der letzten Spalte zur ersten und dritten Spalte sowie
- Entwicklung nach der ersten Zeile:

$$\begin{vmatrix} 3 & 0 & -4 & -1 \\ 2 & 6 & 1 & 1 \\ 4 & -13 & 12 & -1 \\ 1 & 3 & 3 & 1 \end{vmatrix} = \begin{vmatrix} 0 & 0 & 0 & -1 \\ 5 & 6 & -3 & 1 \\ 1 & -13 & 16 & -1 \\ 4 & 3 & -1 & 1 \end{vmatrix} = -(-1) \begin{vmatrix} 5 & 6 & -3 \\ 1 & -13 & 16 \\ 4 & 3 & -1 \end{vmatrix}$$

Addition entsprechender Vielfacher der letzten Zeile zu den ersten beiden Zeilen ergibt weiter

$$\begin{vmatrix} -7 & -3 & 0 \\ 65 & 35 & 0 \\ 4 & 3 & -1 \end{vmatrix} = (-1) \begin{vmatrix} -7 & -3 \\ 65 & 35 \end{vmatrix} = 50.$$

- d) Da Dreiecksmatrizen z. B. beim Gauß-Algorithmus auftreten, interessieren deren Determinanten.

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & 0 & 0 & \dots & 0 \\ a_{2,1} & a_{2,2} & 0 & \dots & 0 \\ a_{3,1} & a_{3,2} & a_{3,3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{bmatrix} \text{ oder } \mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ 0 & a_{2,2} & a_{2,3} & \dots & a_{2,n} \\ 0 & 0 & a_{3,3} & \dots & a_{3,n} \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{n,n} \end{bmatrix}$$

Nach dem Entwicklungssatz für Determinanten folgt sofort in beiden Fällen:

$$\det \mathbf{A} = a_{1,1} \cdot a_{2,2} \cdot \dots \cdot a_{n,n}.$$

Demnach gilt: Die Determinante von Dreiecksmatrizen ist gleich dem Produkt der Diagonalelemente. Insbesondere gilt $\det \mathbf{E}_n = 1$.

e) Wir wenden Folgerung 1.7 d) bei der Berechnung der folgenden Determinante an:

$$\begin{aligned}
 \det \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 1 & 2 & 3 & 4 \end{bmatrix} &= (-1) \begin{vmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 1 & 2 & 3 & 4 \end{vmatrix} \\
 &= (-1)^2 \begin{vmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 1 & 2 & 3 \end{vmatrix} = (-1)^3 \begin{vmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{vmatrix} = -[6-2] \\
 &= -4. \quad \blacksquare
 \end{aligned}$$

Satz 1.34 (Multiplikationssatz)

Seien $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ (oder $\mathbb{C}^{n \times n}$), so gilt:

$$\det(\mathbf{A} \cdot \mathbf{B}) = \det(\mathbf{A}) \det(\mathbf{B}).$$

Der Beweis ist recht aufwändig, wird aber zumindest im Kasten auf Seite 191 motiviert. Ist $\mathbf{A}$ invertierbar, so erhalten wir als direkte Konsequenz:

$$1 = \det \mathbf{E} = \det(\mathbf{A} \cdot \mathbf{A}^{-1}) = \det(\mathbf{A}) \det(\mathbf{A}^{-1}).$$

Damit ist $\det \mathbf{A} \neq 0$ und $\det(\mathbf{A}^{-1}) = \frac{1}{\det \mathbf{A}}$.

Folgerung 1.8 (Determinante orthogonaler und unitärer Matrizen)

Ist $\mathbf{A} \in \mathbb{R}^{n \times n}$ eine orthogonale Matrix ($\mathbf{A}^{-1} = \mathbf{A}^\top$) oder $\mathbf{A} \in \mathbb{C}^{n \times n}$ eine unitäre Matrix ($\mathbf{A}^{-1} = \mathbf{A}^* = \overline{\mathbf{A}}^\top$), so ist $|\det \mathbf{A}| = 1$.

Beweis: Da für eine reelle Matrix $\mathbf{A}^\top = \mathbf{A}^*$ ist, müssen wir die Aussage nur für unitäre Matrizen zeigen.

$$\overline{\det \mathbf{A}} = \overline{\det \mathbf{A}^\top} = \det \overline{\mathbf{A}^\top} = \det \mathbf{A}^* = \det(\mathbf{A}^{-1}) = \frac{1}{\det \mathbf{A}}.$$

Damit ist $1 = \overline{\det \mathbf{A}} \cdot \det \mathbf{A} = |\det \mathbf{A}|^2$. □

Matrizen kann man blockweise multiplizieren, wenn die Blöcke für die Multiplikation zusammenpassen, siehe Seite 161. Daraus lassen sich zusammen mit dem Multiplikationssatz Rechenregeln für die Determinante gewinnen.

Satz 1.35 (Kästchensatz)

Sei $\mathbf{A}$ eine (komplexe) $n \times n$ -Matrix, die aus Matrizen $\mathbf{B}$, $\mathbf{C}$ und $\mathbf{D}$ wie folgt zusammengesetzt ist:

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{B} & \mathbf{C} \\ \hline \mathbf{O} & \mathbf{D} \end{array} \right].$$

Dabei seien $\mathbf{B}$ und $\mathbf{D}$ quadratische $m \times m$ - bzw. $k \times k$ -Matrizen, wobei $m + k = n$ ist. Damit ist $\mathbf{C}$ eine $m \times k$ -Matrix und die Nullmatrix $\mathbf{O}$ unten links ist eine $k \times m$ -Matrix. Dann gilt:

$$\det \mathbf{A} = \det \left[\begin{array}{c|c} \mathbf{B} & \mathbf{C} \\ \hline \mathbf{O} & \mathbf{D} \end{array} \right] = \det(\mathbf{B}) \det(\mathbf{D}).$$

Die gleiche Aussage gilt, wenn die Nullmatrix oben rechts steht.

Beweis: Ist $\det \mathbf{D} = 0$, dann sind die letzten k Zeilen linear abhängig, und es gilt auch $\det \mathbf{A} = 0$, womit in diesem Fall die Aussage gezeigt ist. Im Folgenden sei also $\det \mathbf{D} \neq 0$, d. h., $\mathbf{D}$ sei invertierbar zu $\mathbf{D}^{-1} \in \mathbb{C}^{k \times k}$. Dann gilt mit der $m \times m$ -Einheitsmatrix $\mathbf{E}_m$ und der $k \times k$ -Einheitsmatrix $\mathbf{E}_k$ aufgrund blockweiser Multiplikation und des Multiplikationssatzes (Satz 1.34) für Determinanten:

$$\begin{aligned} \det \left[\begin{array}{c|c} \mathbf{B} & \mathbf{C} \\ \hline \mathbf{O} & \mathbf{D} \end{array} \right] &= \det \left(\left[\begin{array}{c|c} \mathbf{B} & \mathbf{CD}^{-1} \\ \hline \mathbf{O} & \mathbf{E}_k \end{array} \right] \cdot \left[\begin{array}{c|c} \mathbf{E}_m & \mathbf{O} \\ \hline \mathbf{O} & \mathbf{D} \end{array} \right] \right) \\ &= \det \left(\left[\begin{array}{c|c} \mathbf{B} & \mathbf{CD}^{-1} \\ \hline \mathbf{O} & \mathbf{E}_k \end{array} \right] \right) \cdot \det(\mathbf{D}) \\ &= \det \left(\left[\begin{array}{c|c} \mathbf{E}_m & \mathbf{CD}^{-1} \\ \hline \mathbf{O} & \mathbf{E}_k \end{array} \right] \cdot \left[\begin{array}{c|c} \mathbf{B} & \mathbf{O} \\ \hline \mathbf{O} & \mathbf{E}_k \end{array} \right] \right) \cdot \det(\mathbf{D}) = 1 \cdot \det(\mathbf{B}) \cdot \det(\mathbf{D}). \end{aligned}$$

Man beachte, dass die Determinante einer Dreiecksmatrix gleich dem Produkt der Diagonalelemente ist, so entsteht hier der Wert 1. $\square$

Beispiel 1.136

$$\det \begin{bmatrix} 1 & 2 & 0 & 2 & 3 \\ 0 & 1 & 2 & 1 & 1 \\ 2 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 & 4 \end{bmatrix} = \det \left[\begin{array}{ccc|cc} 1 & 2 & 0 & 2 & 3 \\ 0 & 1 & 2 & 1 & 1 \\ 2 & 0 & 0 & 0 & 2 \\ \hline 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 & 4 \end{array} \right] = 8 \cdot (-2) = -16.$$

Analog ist

$$\det \left[\begin{array}{ccc|cc} 1 & 2 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ \hline 1 & 2 & 3 & 1 & 2 \\ 0 & 2 & 1 & 3 & 4 \end{array} \right] = 8 \cdot (-2) = -16.$$

■

Der Kästchensatz kann indirekt auch angewendet werden, wenn ein Kästchen oben links oder unten rechts gleich der Nullmatrix ist. Hier muss man aber durch Vertauschen von Zeilen oder Spalten die Matrix zunächst in die im Satz geforderte Struktur bringen (insbesondere Nullmatrix oben rechts oder unten links). Beim Vertauschen von zwei Zeilen oder Spalten ändert sich aber das Vorzeichen der Determinante!

Hintergrund: Alternative Determinantendefinitionen

Statt die Determinante über den Entwicklungssatz (siehe Definition 1.57 und Satz 1.33) einzuführen, kann man sie auch ganz anders (aber natürlich äquivalent) definieren.

So kann die Determinante axiomatisch als Abbildung von $n \times n$ -Matrizen in die komplexen Zahlen über ihre Eigenschaften als normierte alternierende Multilinearform definiert werden. Die Determinante ist dabei eindeutig festgelegt über folgende Eigenschaften aus Folgerung 1.7:

- Die Determinante erfüllt die Bedingung a), d. h., sie ist eine **Multilinearform**. Eine Linearkombination in einer Spalte führt zur entsprechenden Linearkombination der Determinante.
- Die Determinante ist **alternierend**, d. h. Bedingung d) ist (für Spalten) erfüllt, beim Tausch zweier Spalten kehrt sich das Vorzeichen der Determinante um.
- Zusätzlich ist die Determinante **normiert**, d. h., die Determinante der Einheitsmatrix ist eins.

Diese Definition nutzt man, wenn man eine Determinante über Zeilen- und Spaltenumformungen ausrechnet. So kann man auch zeigen, dass es sich hier tatsächlich um eine äquivalente Definition handelt: Zunächst gibt es Multilinearformen, die die obigen Bedingungen erfüllen – nämlich z. B. die über den Entwicklungssatz definierte Determinante.

Nun kann man jede invertierbare Matrix mittels Gauß'scher Spaltenumformungen in eine Einheitsmatrix überführen. Aufgrund der Rechenregeln und der Normierung führen dabei beide Definitionen zum gleichen Zahlenwert.

Es gibt also genau eine normierte alternierende Multilinearform – und diese ist die bekannte Determinante. Mit dieser Erkenntnis kann man vergleichsweise leicht den Multiplikationssatz (Satz 1.34)

$$\det(\mathbf{A} \cdot \mathbf{B}) = \det(\mathbf{A}) \det(\mathbf{B})$$

beweisen: Ist $\det(\mathbf{B}) = 0$, so ist $\mathbf{B}$ und mit $\mathbf{B}$ auch $\mathbf{A} \cdot \mathbf{B}$ singulär. Damit sind beide Seiten null. Anderenfalls können wir eine Funktion $f : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}$ definieren über

$$f(\mathbf{A}) = f(\vec{a}_1, \dots, \vec{a}_n) := \frac{\det(\mathbf{B} \cdot \mathbf{A})}{\det(\mathbf{B})}$$

für $\mathbf{A} = [\vec{a}_1, \dots, \vec{a}_n] \in \mathbb{C}^{n \times n}$.

Die Beweisidee besteht nun darin zu zeigen, dass f eine normierte alternierende Bilinearform ist. Haben wir das gezeigt, dann wissen wir

$$\det(\mathbf{A}) = f(\mathbf{A}) = \frac{\det(\mathbf{B} \cdot \mathbf{A})}{\det(\mathbf{B})},$$

also $\det(\mathbf{B}) \det(\mathbf{A}) = \det(\mathbf{B} \cdot \mathbf{A})$. Die Eigenschaften erhält man aber unmittelbar: Die Normierung folgt aus $f(\mathbf{E}) = \frac{\det(\mathbf{B})}{\det(\mathbf{B})} = 1$. Vertauscht man zwei Spalten in $\mathbf{A}$, dann vertauschen sich auch die entsprechenden Spalten in $\mathbf{B} \cdot \mathbf{A}$, so dass sich das Vorzeichen von f ändert. Schließlich ist f eine Multilinearform. Das ergibt sich direkt aus der Matrixmultiplikation und der entsprechenden Eigenschaft der Determinante.

Häufig findet man die folgende, ebenfalls alternative Definition der Determinante über Permutationen. Sei S_n die Menge der Permutationen der Zahlen $1, \dots, n$, d. h.

$$S_n := \{(x_1, \dots, x_n) \in \{1, \dots, n\}^n : \text{die Zahlen } x_1, \dots, x_n \text{ sind paarweise verschieden}\}.$$

Die Menge S_n hat genau $n!$ Elemente, da es für die erste Stelle n Möglichkeiten, dann jeweils für die zweite $n - 1$ usw. gibt. Alle Permutationen können durch sukzessiven paarweisen Tausch von jeweils zwei Komponenten aus $(1, 2, 3, \dots, n)$ gewonnen werden. Dabei kann eine bestimmte Permutation nur entweder über eine gerade oder über eine ungerade Anzahl von Vertauschungen erreicht werden. Die Anzahl der Vertauschungen ist allerdings nicht eindeutig, da man Vertauschungen ja auch wieder rückgängig machen kann. Wohldefiniert ist aber für $\vec{x} \in S_n$ die Signum-Funktion:

$$\text{sgn}(\vec{x}) = \begin{cases} 1 : \vec{x} \text{ lässt sich durch eine gerade Anzahl} \\ -1 : \vec{x} \text{ lässt sich durch eine ungerade Anzahl} \end{cases}$$

von Vertauschungen aus $(1, 2, 3, \dots, n)$ gewinnen.

Beispielsweise ist $\text{sgn}((2, 3, 1, 4)) = 1$, da zwei paarweise Vertauschungen ergeben: $(2, 1, 3, 4), (1, 2, 3, 4)$.

Satz 1.36 (Leibniz-Formel für Determinanten)

Sei $\mathbf{A} = [a_{k,i}]_{k,i=1}^n \in \mathbb{C}^{n \times n}$. Dann ist

$$\det \mathbf{A} = \sum_{\vec{x} \in S_n} \operatorname{sgn}(\vec{x}) \prod_{k=1}^n a_{k,x_k}. \quad (1.60)$$

Es ist nicht schwer nachzurechnen, dass über die rechte Seite von (1.60) eine normierte alternierende Multilinearform definiert ist, die daher mit der Determinante übereinstimmt:

- Eine Linearkombination in einer Spalte der Matrix führt zu einer Linearkombination in genau einem Faktor jedes Produkts, das entsprechend aufgeteilt werden kann. Damit handelt es sich bei der rechten Seite in (1.60) um eine Multilinearform.
- Vertauscht man zwei Spalten, so entsteht der gleiche Ausdruck, wenn man in den Permutationen die entsprechenden Spaltenpositionen paarweise tauscht. Allerdings führt der Tausch zu einem Vorzeichenwechsel der Signum-Funktion und damit insgesamt zum Vorzeichenwechsel.
- Bei der Einheitsmatrix ist nur das Produkt zu $\vec{x} = (1, 2, \dots, n)$ von null verschieden und gleich eins. Außerdem ist $\operatorname{sgn}((1, 2, \dots, n)) = 1$.

1.8.2 Determinanten und lineare Gleichungssysteme

Wir betrachten das lineare Gleichungssystem

$$\begin{array}{ccccccc} a_{1,1} x_1 & + a_{1,2} x_2 & \dots & + a_{1,n} x_n & = & b_1 \\ a_{2,1} x_1 & + a_{2,2} x_2 & \dots & + a_{2,n} x_n & = & b_2 \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{n,1} x_1 & + a_{n,2} x_2 & \dots & + a_{n,n} x_n & = & b_n, \end{array}$$

mit $a_{i,k} \in \mathbb{C}$ und $b_i \in \mathbb{C}$, oder in Kurzform

$$\mathbf{A} \cdot \vec{x} = \vec{b} \quad (1.61)$$

mit der $(n \times n)$ -Koeffizientenmatrix $\mathbf{A}$, dem gesuchten Lösungsvektor $\vec{x} \in \mathbb{C}^n$ sowie der Inhomogenität $\vec{b} \in \mathbb{C}^n$.

Determinanten charakterisieren das Lösungsverhalten linearer Gleichungssysteme:

Satz 1.37 (Determinanten und lineare Gleichungssysteme)

Das lineare Gleichungssystem (1.61) ist genau dann eindeutig lösbar, d. h. **regulär**, falls $\det \mathbf{A} \neq 0$. Für $\det \mathbf{A} = 0$ ist das System dagegen **singulär**, d. h., es existiert entweder keine Lösung, oder es gibt unendlich viele Lösungen.

Beweis: Das System ist genau dann eindeutig lösbar, wenn der Gauß-Algorithmus das Dreieckssystem

$$\left[\begin{array}{cccc|c} a_{1,1}^* & a_{1,2}^* & \dots & a_{1,n}^* & b_1^* \\ 0 & a_{2,2}^* & \dots & a_{2,n}^* & b_2^* \\ 0 & 0 & a_{3,3}^* & \dots & a_{3,n}^* & b_3^* \\ \vdots & & & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & a_{n,n}^* & b_n^* \end{array} \right]$$

mit $a_{i,i}^* \neq 0$, $i = 1, \dots, n$ liefert. Es ist genau dann nicht eindeutig lösbar, wenn der Gauß-Algorithmus auf ein Dreieckssystem mit einem $a_{i,i}^* = 0$ führt.

Wie in Beispiel 1.135 d) ist die Determinante der Dreiecksmatrix das Produkt der Diagonalelemente:

$$\begin{vmatrix} a_{1,1}^* & a_{1,2}^* & a_{1,3}^* & \dots & a_{1,n}^* \\ 0 & a_{2,2}^* & a_{2,3}^* & \dots & a_{2,n}^* \\ 0 & 0 & a_{3,3}^* & \dots & a_{3,n}^* \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{n,n}^* \end{vmatrix} = a_{1,1}^* a_{2,2}^* \dots a_{n,n}^*.$$

Die Determinante dieser Matrix ist also genau dann ungleich null, wenn das Ausgangssystem eindeutig lösbar ist. Jetzt müssen wir nur noch den Wert dieser Determinanten mit der Determinante $\det \mathbf{A}$ der Ausgangsmatrix in Beziehung bringen.

Die Gauß-Operationen sind Zeilenumformungen im Sinne von Folgerung 1.7 d). Da sich bei Addition von Vielfachen von Gleichungen die Determinante nicht ändert, sie bei Multiplikation einer Gleichung mit einer von null verschiedenen Zahl auch mit dieser Zahl multipliziert wird und sie nur das Vorzeichen bei Vertauschung von Gleichungen ändert, sind entweder die Determinanten aller Matrizen des Gauß-Verfahrens gemeinsam gleich 0 oder alle gemeinsam von 0 verschieden. Damit ist die eindeutige Lösbarkeit äquivalent zu $\det \mathbf{A} \neq 0$. $\square$

Eine direkte Anwendung von Satz 1.37 ist die Frage nach der Existenz und Eindeutigkeit von Lagrange-Interpolationspolynomen $p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_0$, die an $n+1$ vorgegebenen, verschiedenen Stellen $x_0, x_1, \dots, x_n$ die gleichen Funktionswerte wie eine gegebene Funktion f haben. Die Existenz eines solchen Polynoms haben wir in Kapitel 1.5.6.2 durch explizite Angabe geklärt. Aber wie sieht es mit der Eindeutigkeit

aus? Beide Fragen lassen sich gemeinsam durch die Berechnung einer Determinante beantworten. Die Matrix

$$\mathbf{V}(x_0, \dots, x_n) := \begin{bmatrix} x_0^0 & x_0^1 & x_0^2 & \dots & x_0^n \\ x_1^0 & x_1^1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ x_n^0 & x_n^1 & x_n^2 & \dots & x_n^n \end{bmatrix} = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix}$$

heißt eine **Vandermonde-Matrix**. Die Suche nach einem Lagrange-Interpolationspolynom ist äquivalent mit der Lösung des Gleichungssystems

$$\mathbf{V}(x_0, \dots, x_n) \cdot (a_0, a_1, \dots, a_n)^\top = (f(x_0), f(x_1), \dots, f(x_n))^\top.$$

Das ist eindeutig lösbar genau dann, wenn die Determinante ungleich null ist.

Satz 1.38 (Vandermonde-Determinante)

Die Determinante der Vandermonde-Matrix lautet für $n \in \mathbb{N}$:

$$\det \mathbf{V}(x_0, \dots, x_n) = \prod_{k=0}^{n-1} \left[\prod_{i=k+1}^n (x_i - x_k) \right].$$

Sind die Zahlen x_k verschieden, so ist die Determinante ungleich null, und das Interpolationsproblem hat tatsächlich eine eindeutige Lösung.

Beweis: Sind zwei der Zahlen x_k gleich, dann sind zwei Zeilen von $\mathbf{V}$ gleich, und die Determinante ist null, ebenso das Produkt. Daher seien im Folgenden alle Zahlen x_k verschieden. Wir benutzen Vollständige Induktion nach n .

■ Induktionsanfang für $n = 1$:

$$\det \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \end{bmatrix} = x_1 - x_0 = \prod_{k=0}^0 \left[\prod_{i=1}^1 (x_i - x_k) \right].$$

■ Induktionsannahme: Die Aussage gelte für ein $n \geq 1$ und beliebige, paarweise verschiedene Zahlen $x_0, \dots, x_n$.

■ Induktionsschritt: Wir subtrahieren zunächst die erste Zeile von allen weiteren und entwickeln dann nach der ersten Spalte:

$$\begin{aligned} & \det \mathbf{V}(x_0, \dots, x_n, x_{n+1}) \\ &= \det \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^{n+1} \\ 0 & x_1 - x_0 & x_1^2 - x_0^2 & \dots & x_1^{n+1} - x_0^{n+1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & x_{n+1} - x_0 & x_{n+1}^2 - x_0^2 & \dots & x_{n+1}^{n+1} - x_0^{n+1} \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
&= \det \begin{bmatrix} x_1 - x_0 & x_1^2 - x_0^2 & \dots & x_1^{n+1} - x_0^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n+1} - x_0 & x_{n+1}^2 - x_0^2 & \dots & x_{n+1}^{n+1} - x_0^{n+1} \end{bmatrix} \\
&= \left[\prod_{i=1}^{n+1} (x_i - x_0) \right] \det \begin{bmatrix} 1 & \frac{x_1^2 - x_0^2}{x_1 - x_0} & \dots & \frac{x_1^{n+1} - x_0^{n+1}}{x_1 - x_0} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{x_{n+1}^2 - x_0^2}{x_{n+1} - x_0} & \dots & \frac{x_{n+1}^{n+1} - x_0^{n+1}}{x_{n+1} - x_0} \end{bmatrix}.
\end{aligned}$$

Nun benutzen wir die Polynomdivision

$$\frac{a^m - b^m}{a - b} = a^{m-1} + ba^{m-2} + b^2 a^{m-3} + \dots + b^{m-1} = \sum_{k=0}^{m-1} a^{m-1-k} b^k$$

und erhalten

$$\begin{aligned}
\det \mathbf{V}(x_0, \dots, x_n, x_{n+1}) &= \left[\prod_{i=1}^{n+1} (x_i - x_0) \right] \cdot \det \begin{bmatrix} 1 & x_1 + x_0 & x_1^2 + x_1 x_0 + x_0^2 & \dots & \sum_{k=0}^n x_1^{n-k} x_0^k \\ 1 & x_2 + x_0 & x_2^2 + x_2 x_0 + x_0^2 & \dots & \sum_{k=0}^n x_2^{n-k} x_0^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n+1} + x_0 & x_{n+1}^2 + x_{n+1} x_0 + x_0^2 & \dots & \sum_{k=0}^n x_{n+1}^{n-k} x_0^k \end{bmatrix} \\
&= \left[\prod_{i=1}^{n+1} (x_i - x_0) \right] \det \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n+1} & x_{n+1}^2 & \dots & x_{n+1}^n \end{bmatrix},
\end{aligned}$$

wobei wir im letzten Schritt von jeder Spalte das x_0 -fache der vorangehenden Spalte abgezogen haben. Auf die verbleibende Determinante kann schließlich die Induktionsannahme angewendet werden:

$$\begin{aligned}
&\det \mathbf{V}(x_0, \dots, x_n, x_{n+1}) \\
&= \left[\prod_{i=1}^{n+1} (x_i - x_0) \right] \prod_{k=1}^n \left[\prod_{i=k+1}^{n+1} (x_i - x_k) \right] = \prod_{k=0}^n \left[\prod_{i=k+1}^{n+1} (x_i - x_k) \right]. \quad \square
\end{aligned}$$

Für reguläre Matrizen kann sogar die Lösung eines Gleichungssystems mit Determinanten berechnet werden:

Satz 1.39 (Cramer'sche Regel)

Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ (oder $\mathbb{C}^{n \times n}$) mit $\det \mathbf{A} \neq 0$ und $\vec{b} \in \mathbb{R}^n$ (oder $\mathbb{C}^n$). Die (eindeutige) Lösung $\vec{x} \in \mathbb{R}^n$ (oder $\mathbb{C}^n$) des Gleichungssystems

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

ist gegeben durch

$$x_k = \frac{\det \mathbf{A}_k}{\det \mathbf{A}} \text{ mit } \mathbf{A}_k := \begin{bmatrix} a_{1,1} & \dots & a_{1,k-1} & b_1 & a_{1,k+1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,k-1} & b_2 & a_{2,k+1} & \dots & a_{2,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & \dots & a_{n,k-1} & b_n & a_{n,k+1} & \dots & a_{n,n} \end{bmatrix}.$$

Beweis: Mit den Spaltenvektoren $\vec{a}_k := \mathbf{A}_{\cdot k} = (a_{1,k}, \dots, a_{n,k})^\top$ lautet die Matrix $\mathbf{A}$ des linearen Gleichungssystems

$$\mathbf{A} = [\vec{a}_1, \dots, \vec{a}_{k-1}, \vec{a}_k, \vec{a}_{k+1}, \dots, \vec{a}_n].$$

Ersetzt man in $\mathbf{A}$ die k -te Spalte durch $\vec{b} \in \mathbb{R}^n$, so erhält man die n Matrizen

$$\mathbf{A}_k = [\vec{a}_1, \dots, \vec{a}_{k-1}, \vec{b}, \vec{a}_{k+1}, \dots, \vec{a}_n], \quad k = 1, 2, \dots, n.$$

Der Vektor $\vec{b}$ ist die Inhomogenität des Gleichungssystems, das wegen $\det \mathbf{A} \neq 0$ nach Satz 1.37 eine eindeutige Lösung $\vec{x}$ besitzt. $\vec{b}$ hat daher die Darstellung $\vec{b} = \vec{a}_1 \cdot x_1 + \dots + \vec{a}_n \cdot x_n$. Damit erhalten wir für die Determinante der Matrix $\mathbf{A}_k$:

$$\det \mathbf{A}_k = \det[\vec{a}_1, \dots, \underbrace{\vec{a}_{k-1}, x_1 \cdot \vec{a}_1 + \dots + x_n \cdot \vec{a}_n}_{\vec{b}}, \vec{a}_{k+1}, \dots, \vec{a}_n],$$

für $k = 1, 2, \dots, n$. Die k -te Spalte von $\mathbf{A}_k$ ist die Summe von $x_k \vec{a}_k$ mit Vielfachen der restlichen Spalten. Mit Folgerung 1.7 g) und a) erhalten wir daher

$$\det \mathbf{A}_k = x_k \det \mathbf{A}.$$

Wegen $\det \mathbf{A} \neq 0$ (reguläres System) folgt

$$x_k = \frac{\det \mathbf{A}_k}{\det \mathbf{A}}, \quad k = 1, \dots, n.$$

□

In Worten: Man erhält das k -te Element x_k des Lösungsvektors, indem man die k -te Spalte der Koeffizientenmatrix durch den Vektor $\vec{b}$ ersetzt, die Determinante dieser Matrix berechnet und das Ergebnis durch die Determinante der Koeffizientenmatrix $\mathbf{A}$ dividiert.

Ist $\det \mathbf{A} \neq 0$, so kann man den Satz auch zur Berechnung von $\mathbf{A}^{-1}$ verwenden, indem man mit ihm die n Gleichungssysteme (1.55) löst. Aus $\det \mathbf{A} \neq 0$ folgt also die Existenz von $\mathbf{A}^{-1}$. Existiert umgekehrt die Inverse, so folgt aus dem Multiplikationssatz (Satz 1.34), dass $\det \mathbf{A} \neq 0$ ist:

$$\mathbf{A}^{-1} \text{ existiert} \iff \det \mathbf{A} \neq 0,$$

und das ist nach Satz 1.37 äquivalent dazu, dass $\mathbf{A}\vec{x} = \vec{b}$ eindeutig lösbar ist.

Beispiel 1.137

a) Wir berechnen die Lösung mittels der Cramer'schen Regel:

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

hat die Lösung

$$x_1 = \frac{1}{\begin{vmatrix} 1 & 2 \\ 0 & 3 \end{vmatrix}} \begin{vmatrix} 1 & 2 \\ 3 & 3 \end{vmatrix} = \frac{-3}{3} = -1, \quad x_2 = \frac{1}{\begin{vmatrix} 1 & 2 \\ 0 & 3 \end{vmatrix}} \begin{vmatrix} 1 & 1 \\ 0 & 3 \end{vmatrix} = \frac{3}{3} = 1.$$

Schon auf Seite 170 haben wir die hier verwendete Cramer'sche Regel im Spezialfall einer (2×2) -Matrix kennengelernt, siehe (1.54).

b) Auch bei einer (3×3) -Matrix ist die Cramer'sche Regel noch praktikabel:

$$\begin{array}{rrcr} 3x_1 & -2x_2 & +x_3 & = & 2 \\ 2x_1 & +x_2 & +x_3 & = & 7 \\ 6x_1 & -3x_2 & -x_3 & = & -3. \end{array}$$

Mit

$$\det \mathbf{A} = \begin{vmatrix} 3 & -2 & 1 \\ 2 & 1 & 1 \\ 6 & -3 & -1 \end{vmatrix} = -22$$

ergibt sich:

$$x_1 = \frac{\begin{vmatrix} 2 & -2 & 1 \\ 7 & 1 & 1 \\ -3 & -3 & -1 \end{vmatrix}}{-22} = 1, \quad x_2 = \frac{\begin{vmatrix} 3 & 2 & 1 \\ 2 & 7 & 1 \\ 6 & -3 & -1 \end{vmatrix}}{-22} = 2, \quad x_3 = \frac{\begin{vmatrix} 3 & -2 & 2 \\ 2 & 1 & 7 \\ 6 & -3 & -3 \end{vmatrix}}{-22} = 3.$$

■

Die Lösung eines linearen Gleichungssystems nach der Cramer'schen Regel bietet sich nur bei wenigen Variablen an. Bei zwei oder drei Unbekannten ist sie allerdings sehr praktisch und wird gerne in Beispielaufgaben der Ingenieurwissenschaften verwendet. Im Studium kann sie damit überlebenswichtig werden. Für mittelgroße Gleichungssysteme (etwa $n \leq 50$) verwendet man aber den Gauß-Algorithmus. Der Grund liegt in der Anzahl der durchzuführenden Rechenoperationen. Für große Systeme kommen iterative Verfahren zum Einsatz (siehe Kapitel 1.7.6.2).

Wir fassen den Abschnitt zu Determinanten zusammen:

- Eine Determinante ist eine Zahl.
- Determinanten können nur für quadratische $(n \times n)$ -Matrizen $\mathbf{A}$ berechnet werden.
- Eine Determinante ist bei $n = 2$ der Flächeninhalt eines Parallelogramms.
- Mit der Determinante kann man feststellen, ob das Gleichungssystem $\mathbf{A}\vec{x} = \vec{b}$ eine eindeutige Lösung besitzt (nämlich wenn $\det \mathbf{A} \neq 0$).
- Mit Determinanten kann ein Gleichungssystem gelöst werden (Cramer'sche Regel).

1.9 Aufgaben

Lösungen aller Aufgaben finden Sie auf der Internetseite zum Buch:

<http://www.springer-spektrum.de/978-3-8274-3007-6>

Aufgabe 1.1

- a) *Aus einem Wasserhahn fließen in 4 Minuten 16 Liter Wasser. Wie viele Liter Wasser laufen an einem Tag aus dem Hahn, wie viele in einem Jahr?*
- b) *Die Grube für ein Fundament wird von 4 Arbeitern in 9 Stunden ausgehoben und betoniert. Wie lange brauchen 3 Arbeiter für diese Arbeit? Wie lange würden 10 Arbeiter dafür benötigen?*
- c) *Fünf Wölfe reißen in 5 Minuten 5 Schafe. Wie viele Schafe werden von 100 Wölfen in 100 Minuten erlegt?*

Aufgabe 1.2

Das erste Semester Maschinenbau bestehe aus 200 Studierenden. 50 Personen belegen das Praktikum A, 60 Personen belegen die Vorlesung B und 30 Personen belegen das Projektfach C. 15 Studierende haben Praktikum A und Vorlesung B belegt, 10 Studierende belegen Praktikum A und das Projektfach C. Von den Studierenden, die das Projektfach C besuchen, haben 5 die Vorlesung B belegt. Kein Studierender hat alle Veranstaltungen gebucht. Beantworten Sie die Fragen mittels eines Venn-Diagramms:

- a) *Wie viele Studierende des Jahrgangs haben genau eine Veranstaltung gebucht?*
- b) *Wie viele Studierende belegen ausschließlich andere Veranstaltungen als die hier aufgezählten?*

Aufgabe 1.3

Gegeben sind die drei Mengen $M_1 = \{a, b, c, 1, 2\}$, $M_2 = \{2, 3, 4, 5, 6\}$ und $M_3 = \{a, c, 2, 4, 6\}$. Bilden Sie die Mengen $M_1 \cap M_2$, $M_1 \cup M_2$, $M_1 \cap M_3$, $M_1 \cup M_3$, $M_2 \cap M_3$ und $M_2 \cup M_3$ sowie $M_1 \setminus M_2$, $M_2 \setminus M_1$, $\bigcap_{n=1}^3 M_n := M_1 \cap M_2 \cap M_3$ und $\bigcup_{n=1}^3 M_n := M_1 \cup M_2 \cup M_3$.

Aufgabe 1.4

Negieren Sie folgende Aussagen:

- a) $A = \text{„Maria ist Italienerin und sehr freundlich.“}$, b) $B = \text{„} -1 \leq x \leq 4 \text{“}$,
 c) $C = \text{„Der Bus kommt pünktlich, oder er ist zu spät.“}$.

Aufgabe 1.5

- a) Beweisen Sie durch direkte Folgerungen (Modus Ponens): Sind die natürlichen Zahlen a und b ungerade, so ist $a + b$ gerade und $a \cdot b$ ungerade.
 b) Beim Rechnen mit ganzen Zahlen gelten die Regeln $1 \cdot n = n \cdot 1 = n$, $0 \cdot n = n \cdot 0 = 0$, es darf ausmultipliziert werden, und $-n$ ist die eindeutige Zahl mit $n + (-n) = (-n) + n = 0$ (Eindeutigkeit der Inversen). Zeigen Sie durch einen direkten Schluss, dass aus diesen Regeln bereits $(-1) \cdot (-1) = 1$ folgt.

Aufgabe 1.6

Beweisen Sie die folgenden Aussagen indirekt mittels Widerspruch:

- a) Für $x \in \mathbb{R}$, $x > 0$ gilt: $x + \frac{4}{x} \geq 4$. b) Für jedes $n \in \mathbb{N}$ gilt: $\frac{n^4 + 2}{n^6 + n} > \frac{1}{n^2}$.

Aufgabe 1.7

- a) Schreiben Sie die Dualzahl $a = 101\ 101\ 101$ als Dezimalzahl.
 b) Man stelle die hexadezimale Zahl $z = 4f6c$ als Summe von Potenzen der Basis 16 dar und bestimme die zugehörige Dezimalzahl.
 c) Schreiben Sie die periodische Dezimalzahl $r = 0, \overline{345}$ als Bruch $\frac{p}{q}$ mit $p, q \in \mathbb{N}$.

Aufgabe 1.8

Man berechne die Binomialkoeffizienten

- a) $\binom{13}{3}$, b) $\binom{10}{5}$, c) $\binom{13}{10}$, d) $\binom{2n}{n} \frac{(n!)^2}{(2n)!}$.

Aufgabe 1.9

Berechnen Sie die folgenden Summen:

- a) $\sum_{k=1}^6 (-1)^k k^2$, b) $\sum_{k=0}^{100} ((k+1)^2 - k^2)$, c) $\sum_{k=1}^n (3k-1)$, d) $\sum_{k=0}^{10\ 000} \binom{10\ 000}{k} 2^k$.

Aufgabe 1.10

Zeigen Sie für $n \in \mathbb{N}$: $\sum_{k=1}^n \frac{n+1}{k^2 + k} = n$.

Aufgabe 1.11

Zeigen Sie durch Vollständige Induktion die folgenden Aussagen für die angegebenen Werte $n \in \mathbb{N}$:

- a) $\prod_{k=1}^n k^k = 1^1 \cdot 2^2 \cdot 3^3 \cdots n^n < n^{\frac{n}{2}(n+1)}$ für $n \geq 2$, b) $n^2 \leq 2^n$ für $n \neq 3$,
 c) $\sum_{k=2}^{n+1} (k-1) \cdot k = \frac{n(n+1)(n+2)}{3}$.

Aufgabe 1.12

Zeigen Sie mittels Vollständiger Induktion, dass für jede Wahl von $n \in \mathbb{N}$ die Zahl $(13+1)^n - 1$ durch 13 teilbar ist.

Aufgabe 1.13

Bestimmen Sie mit dem Binomischen Lehrsatz jeweils den Koeffizienten der Potenz x^4 in der binomischen Entwicklung von a) $(1-4x)^8$, b) $(x+0,5a)^{12}$.

Aufgabe 1.14

Zeigen Sie mit Hilfe der Binomischen Formel für $n \in \mathbb{N}$ die Identität:

$$\left(1 + \frac{1}{n}\right)^n = 1 + \sum_{k=1}^n \frac{\left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right)}{k!}.$$

Für großes n stimmen die ersten Summanden nahezu mit denen der Summe $e = \sum_{k=0}^{\infty} \frac{1}{k!}$ überein. Damit ist es nicht verwunderlich, dass die Zahlen sich für großes n dem Wert e annähern (siehe Kapitel 2.1.5).

Aufgabe 1.15

Welche der folgenden Funktionen f sind bijektiv (also injektiv und surjektiv)? Berechnen Sie die Umkehrfunktionen f^{-1} der bijektiven Funktionen f .

- a) $f_1: \mathbb{R} \rightarrow \mathbb{R}$, $f_1(x) = 2x^6 + 3x^4 - 2x^2 + 1$. b) $f_2: \mathbb{R} \rightarrow \mathbb{R}$, $f_2(x) = 3x^5 + 5$.
 c) $f_3: [-\pi/2, \pi/2] \rightarrow [0, 1]$, $f_3(x) = [\sin(x)]^2$.
 d) $f_4: [-\pi/2, \pi/2] \rightarrow [-1, 1]$, $f_4(x) = \sin(x)$. e) $f_5: \mathbb{R} \rightarrow]0, \infty[$, $f_5(x) = e^{x+4}$.

Aufgabe 1.16

Vereinfachen Sie so weit wie möglich (mit $a > 0$):

- a) $\sqrt{\sqrt[3]{a^6 b^{24}}}$, b) $\frac{\sqrt[6]{a^5 \sqrt[3]{a^2}}}{\sqrt[3]{a^2 \sqrt[6]{a^2}}}$, c) $\left(\frac{x^2}{2y} - \frac{y^2}{2x}\right) : \left(\frac{1}{x} - \frac{1}{y}\right)$, d) $\frac{\frac{1}{x} + x}{\frac{1}{x} - x}$.

Aufgabe 1.17

Berechnen Sie: a) $\log_3(2^4)$, b) $\ln(\sqrt{e^4})$, c) $\ln \sqrt{e^{3[\ln(e^2) + \ln(e^8)]}}$, d) $\log_8 \sqrt{12}$.

Aufgabe 1.18

Bestimmen Sie die Lösungsmenge folgender Gleichungen:

- a) $16x^2 + 32x - 240 = 0$, b) $x^6 - 3x^4 + x^2 = 0$, c) $\sqrt{x^2 + x - 1} = \sqrt{x^2 + x + 1}$,

- d) $\sqrt{x^2 + 4} = x - 2$, e) $|x + 1| = |x - 1|$, f) $|x^2 + 4x + 4| = |x|$,
 g) $2^x + 4 \cdot 2^{-x} - 5 = 0$, h) $\lg(4x - 4) = \frac{3}{2}$, i) $3 + 2e^{-4t} - 5e^{-2t} = 0$.

Aufgabe 1.19

Bestimmen Sie alle reellen Zahlen, die die folgenden Gleichungen erfüllen:

- a) $|2x - 6| < 10$, b) $\left| \frac{x-3}{x+2} \right| < 2$, c) $|x + 1| + |x + 2| \leq 3$.

Aufgabe 1.20

Bestimmen Sie die Lösungsmenge folgender Wurzelgleichungen:

- a) $\sqrt{x+5} + \sqrt{x-3} + 13 = 0$, b) $\sqrt{-6x+8} - \sqrt{36+4x} = \sqrt{4x+46}$,
 c) $\sqrt{3x} + \sqrt{2x+5} = \sqrt{5x-1}$.

Aufgabe 1.21

Es soll ein Rechteck mit der Fläche 400 cm^2 erzeugt werden. In welchen Bereichen dürfen die beiden Seitenlängen liegen, wenn der Umfang 100 cm betragen soll?

Aufgabe 1.22

Lösen Sie die Gleichungen nach x auf:

- a) $y = \ln\left(14 - \frac{x}{2}\right)$ für $x < 28$, b) $y = \frac{1}{2}(e^{2x} - 4)$, c) $y = \frac{e^x}{1 + 7e^x}$,
 d) $y = \ln(x+2) + \ln(x-2)$ ($x > 2$).

Aufgabe 1.23

Ein Mensch kann in der Regel bis zu einer Schallintensität von $I_0 = 10^{-12} \text{ W/m}^2$ hören (Hörbarkeitsschwelle). Die Lautstärke eines Tons mit der Intensität I berechnet sich damit über (Dürschnabel, 2004, S. 91)

$$L = 10 \cdot \lg\left(\frac{I}{I_0}\right) \text{ dB}.$$

- a) Berechnen Sie die Lautstärke der Hörbarkeitsschwelle I_0 .
 b) Ein startender Düsenjet hat eine Lautstärke von ca. 130 dB , laute Rockmusik bis zu 120 dB . Um wie viel höher ist die Schallintensität I_F des Flugzeugs gegenüber der der Musik I_M ?

Aufgabe 1.24

Man bestimme den Definitionsbereich, Wertebereich und die Umkehrfunktion von

- a) $f(x) = 8e^{x-\frac{1}{2}}$, b) $f(x) = \sqrt{4x}$, c) $f(x) = \frac{x-2}{x+2}$.

Aufgabe 1.25

Auf welchen Intervallen ist $f(x) = |2+x| - |2-x|$ monoton?

Aufgabe 1.26

Zeigen Sie, dass die Funktion $f(x) = 3 + \sqrt{9-x}$ auf $] -\infty, 9]$ streng monoton fallend und damit injektiv ist. Berechnen Sie die Umkehrfunktion.

Aufgabe 1.27

Diskutieren Sie das Symmetrieverhalten der Funktionen

- a) $f(x) = \frac{\sin(x)}{x} = \text{sinc}(x)$, b) $f(x) = \frac{\exp(x) - 1}{\exp(x) + 1}$, c) $f(x) = \exp(x) + \frac{1}{\exp(x)}$,
 d) $f(x) = x^5 \sin^4(x)$, e) $f(x) = 1 + x + x^2$.

Aufgabe 1.28

Bilden Sie die verkettete Funktion $f \circ g : x \mapsto f(g(x))$ für

- a) $g(x) = \sin(x + \pi)$, $f(x) = -4x$, b) $g(x) = -4x$, $f(x) = \sqrt{x+5}$,
 c) $g(x) = x^2$, $f(x) = \frac{1}{1+x}$,

und geben Sie jeweils Definitions- und Wertebereich an.

Aufgabe 1.29

Wir mischen Wasser unterschiedlicher Temperatur. Wasser der Masse $m_1 > 0$ mit Temperatur T_1 wird mit Wasser der Masse $m_2 > 0$ und der Temperatur T_2 gemischt. Dann gilt für die Mischungstemperatur T

$$T = \frac{m_1 T_1 + m_2 T_2}{m_1 + m_2}.$$

- a) Berechnen Sie die Mischungstemperatur, wenn 1 000 g Wasser (also ein Liter) der Temperatur 20°C mit 500 g Wasser der Temperatur 80°C gemischt werden.
 b) In einem Eimer, der maximal fünf Liter (entspricht 5 000 g) Wasser aufnimmt, befinden sich bereits 1 000 g Wasser der Temperatur 20°C . Dazu lässt man Wasser der Temperatur 80°C fließen, das sich mit dem vorhandenen Wasser vermischt. Geben Sie die Wassertemperatur $T(x)$ als Funktion der zugeflossenen Wassermasse x (in g) an. Welchen Wert erhält man bei vollem Behälter?

Aufgabe 1.30

Der Kolben eines Stoßdämpfers lege beim Einschieben einen Weg nach dem Zeitgesetz $x(t) = 30(1 - e^{(-\frac{t}{0.5})})$ cm zurück, wobei $t \geq 0$ die Zeit in Sekunden bezeichnet, vgl. (Papula, 2008, Band 1, S. 321). Nach welcher Zeit ist der Kolben um 12 cm bzw. 15 cm eingeschoben?

Aufgabe 1.31

Heißer Kaffee mit der Anfangstemperatur T_0 wird durch die niedrigere Umgebungstemperatur T_1 gekühlt. Die Temperaturabnahme verläuft dabei exponentiell nach der Gleichung $T(t) = (T_0 - T_1)e^{-k \cdot t} + T_1$ ($t \geq 0$). Dabei ist $T(t)$ die Temperatur der Flüssigkeit zum Zeitpunkt t . Bei Zimmertemperatur $T_1 = 20^\circ \text{C}$ werden folgende Werte gemessen: Nach 5 Minuten beträgt die Kaffeetemperatur 50°C , nach 10 Minuten dagegen nur noch 30°C . Bestimmen Sie T_0 und k . Wie interpretieren Sie das Ergebnis?

Aufgabe 1.32

Berechnen Sie ohne Taschenrechner:

a) $\sin\left(25 \cdot \frac{\pi}{4}\right)$, b) $\sin\left(\frac{\pi}{4}\right) \cos\left(\frac{3}{4}\pi\right)$, c) $\sin\left(6 + \frac{\pi}{4}\right) - \cos\left(6 - \frac{\pi}{4}\right)$.

Aufgabe 1.33

Wir betrachten ein rechtwinkliges Dreieck mit den Kanten a , b und der Hypotenuse c . Der Winkel gegenüber von a heißt α , und der Winkel gegenüber von b heißt β . Vervollständigen Sie die folgende Tabelle:

a	b	c	α	β
2			$\frac{\pi}{6}$	
	3		$\frac{\pi}{4}$	
		5	$\frac{\pi}{6}$	
3	4			

Aufgabe 1.34

Eine Funktion

$$f(t) = a \cos(\omega t + \varphi_0) = a \cos\left(\omega \left(t + \frac{\varphi_0}{\omega}\right)\right)$$

hat die Amplitude a , Periode $p = \frac{2\pi}{\omega}$ und den Nullphasenwinkel φ_0 . Der Funktionsgraph ist um $\frac{\varphi_0}{\omega}$ gegenüber dem Graphen zu $\varphi_0 = 0$ nach links verschoben. Die gleichen Begriffe verwendet man für den Sinus. Bestimmen Sie Periode und Verschiebung gegenüber $\varphi_0 = 0$ für die folgenden Funktionen:

a) $f(t) = 2 \sin(3t - \frac{\pi}{3})$, b) $f(t) = 4 \cos(2t - \frac{\pi}{2})$, c) $f(t) = 10 \sin(\pi t - 3\pi)$.

Aufgabe 1.35

Bestimmen Sie alle reellen Lösungen der folgenden trigonometrischen Gleichungen:

a) $\tan(x) = \cos(x)$, b) $\tan(\arcsin x) = 1$.

Aufgabe 1.36

Bestimmen Sie die Parameter u und v der Funktion $f(x) = 2ue^{-vx}$ so, dass die Punkte $(0, 2)$ und $(1, 2e)$ auf dem Funktionsgraphen liegen.

Aufgabe 1.37

Berechnen Sie Real- und Imaginärteile der folgenden komplexen Zahlen:

$$(1+j)(2+j), \quad \frac{1+j}{2+j}, \quad e^{j\frac{\pi}{2}}.$$

Aufgabe 1.38

Schreiben Sie die folgenden komplexen Zahlen in der Form $re^{j\varphi}$ mit $r \in \mathbb{R}$, $r \geq 0$:

a) $5 - 5j$, b) $64j$, c) $-12(\cos(4) + j \sin(4))$, d) j^3 .

Aufgabe 1.39

Es seien folgende komplexe Zahlen gegeben:

$$z_1 = 20 + 6j, \quad z_2 = -4 + 3j, \quad z_3 = 2 - 4j, \quad z_4 = -3 - 2j.$$

- a) Zeichnen Sie die Zahlen als Punkte in der Gauß'schen Ebene, und berechnen Sie die Beträge.
 b) Bilden Sie die konjugiert komplexen Zahlen, und zeichnen Sie diese ebenfalls ein.
 c) Berechnen Sie die folgenden Zahlen:

$$\frac{z_1 + \overline{z_2}}{z_3}, \quad z_3^2, \quad |z_3|^2, \quad |z_1 - z_2|^2, \quad \frac{(z_3 + z_4) \cdot z_1}{z_2}.$$

- d) Bringen Sie z_1, z_2, z_3 und z_4 in die Polarform $re^{j\varphi}$ mit $r \geq 0$.

Aufgabe 1.40

Für welche $z \in \mathbb{C}$ gilt:

- a) $z^2 \cdot (1 + j) = 2z \cdot (1 - j)$, b) $\frac{4 - 2j}{z} = 5 + j$,
 c) $\frac{1 + j}{z} + \frac{20}{4 + 3j} = 3 - j$, d) $2 - 9j = (1 - 2j) \cdot (z - 5 + 4j)$?

Aufgabe 1.41

Berechnen Sie mit $z_1 = 1 + j$, $z_2 = 1 - j$, $z_3 = -3j$ die folgenden Zahlen:

- a) $z_1^5 \cdot z_2^3$, b) $\sqrt[4]{\frac{z_3}{z_2}}$, c) $\sqrt[3]{z_1 \cdot \overline{z_2}}$, d) $\frac{\sqrt[3]{z_1}}{z_2^2}$, e) $\sqrt[7]{z_2^3}$.

Aufgabe 1.42

Lösen Sie die quadratischen Gleichungen a) $8z^2 - 4z + 2 = 0$ und b) $z^2 + 16 = 0$.

Aufgabe 1.43

Zerlegen Sie die folgenden Polynome in komplexe Linearfaktoren:

- a) $p_1(x) = 2x^4 - 10x^3 + 30x^2 - 10x - 52$, b) $p_2(x) = 10x^4 - 160$,
 c) $p_3(x) = x^4 - 2x^3 + x^2 + 2x - 2$.

Hinweis: Bei c) verwende man $p_3(1 \pm j) = 0$.

Aufgabe 1.44

Bestimmen Sie die Lösung folgender Gleichungen, so dass man Real- und Imaginärteil ablesen kann:

- a) $\frac{3+j}{2-j}z = 1$, b) $z = \frac{1}{3-j} + \frac{4}{3+j}$, c) $z = (2 + 2j)^9$, d) $z^4 = 16$.

Aufgabe 1.45

Berechnen Sie mit dem Horner-Schema zu $p(x) = x^4 + 2x^3 - 45x$ den Funktionswert $p(3)$ und führen Sie mittels Horner-Schema die Polynomdivision $p(x)/(x - 3)$ durch.

Aufgabe 1.46

Bestimmen Sie die Lösungen des linearen Gleichungssystems

$$\begin{array}{rcl} jx_1 & +jx_2 & = j \\ & x_2 & +x_3 = 3 \\ jx_1 & +x_2 & = 1. \end{array}$$

Bringen Sie dazu das Gleichungssystem in eine Dreiecksgestalt. Bringen Sie die Spalten der zugehörigen Matrix von links nach rechts in die richtige Form.

Aufgabe 1.47

Für welche Werte des Parameters s existieren Lösungen des Gleichungssystems

$$\begin{array}{rcl} 3x & -2y & = s \\ -6x & +4y & = 1? \end{array}$$

Aufgabe 1.48

Bestimmen Sie die Lösungen der linearen Gleichungssysteme

$$\begin{array}{lll} \text{a)} & 6x_1 & -8x_2 & +10x_3 & = 2 & \text{b)} & 3x_1 & -4x_2 & +5x_3 & = 1 & \text{c)} & 3x_1 & -4x_2 & +5x_3 & = 1 \\ & x_1 & -3x_2 & +2x_3 & = 1 & & x_1 & -6x_2 & +4x_3 & = 1 & & x_1 & +8x_2 & -3x_3 & = 1 \\ & 2x_1 & +2x_2 & +x_3 & = 1 & & 2x_1 & +2x_2 & +x_3 & = 3 & & 5x_1 & +12x_2 & -x_3 & = 3. \end{array}$$

Aufgabe 1.49

Lösen Sie folgendes homogene lineare Gleichungssystem

$$\begin{array}{rcl} & x_4 & +x_5 & = 0 \\ & x_2 & +x_3 & = 0 \\ x_1 & +x_2 & & = 0 \\ x_1 & & & -x_5 = 0 \\ & x_3 & +x_4 & = 0. \end{array}$$

Aufgabe 1.50

a) Welche Bedingung muss der Vektor $\vec{b} = (b_1, b_2, b_3)^\top$ erfüllen, damit das Gleichungssystem

$$\mathbf{A} \cdot \vec{x} = \vec{b} \text{ mit } \mathbf{A} = \begin{bmatrix} 2 & -1 & 3 \\ -2 & 4 & -4 \\ 1 & 4 & 0 \end{bmatrix}$$

lösbar ist?

b) Welcher Wert ergibt sich für b_3 , wenn man $b_1 = 1$ und $b_2 = 1$ setzt? Wie viele Lösungen hat dann das Gleichungssystem?

Aufgabe 1.51

Für welche Werte von s ist das folgende lineare Gleichungssystem lösbar?

$$\mathbf{A} \cdot \vec{x} = \vec{0} \text{ mit } \mathbf{A} = \begin{bmatrix} 4 & 3 & 2 \\ 1 & 1 & 1 \\ 1 & 0 & s-1 \end{bmatrix}$$

Aufgabe 1.52

Berechnen Sie jeweils die Matrix $\mathbf{X}$:

$$\text{a) } \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - 2 \left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} - \mathbf{X} \right) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{b) } 3 \cdot \mathbf{X} - \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Aufgabe 1.53

Ermitteln Sie die Matrizen $\mathbf{X}$ und $\mathbf{Y}$ aus folgendem Gleichungssystem:

$$7 \cdot \mathbf{X} - \mathbf{Y} = \begin{bmatrix} 3 & 3 \\ -1 & 1 \end{bmatrix}, \quad 4 \cdot \mathbf{X} + \mathbf{Y} = \begin{bmatrix} 2 & 3 \\ -1 & 0 \end{bmatrix}.$$

Aufgabe 1.54

Gegeben sind $\mathbf{A} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix}$.

Berechnen Sie – falls möglich – folgende Matrixprodukte:

a) $\mathbf{A} \cdot \mathbf{B}$, b) $\mathbf{B} \cdot \mathbf{A}$, c) $\mathbf{A} \cdot \mathbf{C}$, d) $\mathbf{C} \cdot \mathbf{A}$, e) $\mathbf{B} \cdot \mathbf{C}$, f) $\mathbf{C} \cdot \mathbf{B}$, g) $\mathbf{C}^2$.

Aufgabe 1.55

a) Gegeben sind die Matrizen

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} -1 & 0 \\ 1 & 2 \end{bmatrix}.$$

Bestätigen Sie die Gültigkeit des Assoziativgesetzes $\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C}$.

b) Gegeben sind die Matrizen $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$.

Bestätigen Sie die Gültigkeit des Gesetzes $(\mathbf{A} \cdot \mathbf{B})^\top = \mathbf{B}^\top \cdot \mathbf{A}^\top$.

Aufgabe 1.56

Gegeben ist die Matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix}$. Bestimmen Sie $\mathbf{A} \cdot \vec{x}$ für $\vec{x}$ gleich

a) $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, b) $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, c) $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, d) $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, e) $\begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}$.

Aufgabe 1.57

Wir können eine $(n \times n)$ -Matrix nutzen, um damit durch Multiplikation die Zeilen und Spalten einer anderen Matrix zu permutieren, d. h., um deren Reihenfolge zu ändern. Bestimmen Sie eine solche (Permutations-)Matrix $\mathbf{P}$, so dass sich $\mathbf{P} \cdot \mathbf{A}$ von $\mathbf{A}$ nur durch Vertauschung der i -ten und k -ten Zeile unterscheidet. Wie wird eine Vertauschung der i -ten und k -ten Spalte von $\mathbf{A}$ erreicht? Betrachten Sie zur Vereinfachung (3×3) -Matrizen.

Aufgabe 1.58

Lösen Sie die folgenden Gleichungen mittels Matrixinvertierung:

a) $\mathbf{A} \cdot \mathbf{X} = \mathbf{B}$, b) $\mathbf{Y} \cdot \mathbf{C} = \mathbf{D}$.

Dabei ist

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \text{ und } \mathbf{C} = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ 1 & 2 & 3 \end{bmatrix}.$$

Aufgabe 1.59

In einer Produktionsanlage werden aus vier Rohstoffen R_1, R_2, R_3, R_4 drei Zwischenprodukte Z_1, Z_2, Z_3 hergestellt, aus diesen Zwischenprodukten werden schließlich fünf Endprodukte E_1, E_2, E_3, E_4, E_5 gefertigt. In den Tabellen ist angegeben, wie viel Rohstoffe zur Produktion eines Zwischenprodukts und wie viele Zwischenprodukte zur Produktion eines Endprodukts E_i benötigt werden.

	Z_1	Z_2	Z_3		E_1	E_2	E_3	E_4	E_5
R_1	2	1	3	Z_1	2	0	1	1	2
R_2	0	0	1	Z_2	1	1	0	2	1
R_3	1	0	2	Z_3	0	1	1	0	0
R_4	0	1	0						

Wie viele Einheiten von R_1, R_2, R_3, R_4 sind bereitzustellen, wenn der Betrieb 20 Einheiten von E_1 , 10 Stück E_2 , 20 Stück E_3 , 30 Stück E_4 und 10 Einheiten von E_5 herstellen soll? **Hinweis:** Schreiben Sie die Tabellen als Matrizen.

Aufgabe 1.60 (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 $R - Y$. Schreiben Sie diese Umrechnung des Vektors $(R, G, B)^\top$ in den Vektor $(Y, U, V)^\top$ mit einer Matrix.

Aufgabe 1.61

a) Berechnen Sie die Lösungen mit dem Gauß'schen Eliminationsverfahren:

$$\begin{bmatrix} 1 & j & 1 \\ 2 & j & -j \\ 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} j \\ 2 + 2j \\ 0 \end{pmatrix}, \quad \begin{bmatrix} 1 & j & -1 \\ 2 & 5 & -j \\ j & 0 & 2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 + 4j \\ 2 + j \end{pmatrix}.$$

b) Mittels des Gauß'schen Eliminationsverfahrens berechne man die Inverse von

$$\mathbf{A} := \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 4 \\ 8 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}.$$

Aufgabe 1.62

Zeigen Sie, dass die Inverse $\mathbf{A}^{-1}$ der oberen rechten Dreiecksmatrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & a_{2,2} & a_{2,3} \\ 0 & 0 & a_{3,3} \end{bmatrix} \quad \text{mit } a_{1,1} \cdot a_{2,2} \cdot a_{3,3} \neq 0$$

ebenfalls eine obere rechte Dreiecksmatrix ist, in deren Hauptdiagonale die Kehrwerte $\frac{1}{a_{1,1}}$, $\frac{1}{a_{2,2}}$ und $\frac{1}{a_{3,3}}$ stehen.

Aufgabe 1.63

Zeigen Sie, dass die Matrizen $\mathbf{A}^\top \cdot \mathbf{A}$ und $\mathbf{A} \cdot \mathbf{A}^\top$ für jede Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ symmetrisch sind.

Aufgabe 1.64

Berechnen Sie die Determinanten der Matrizen

$$A := \begin{bmatrix} 0 & 4 & 0 \\ 3 & 1 & 2 \\ 2 & 1 & 1 \end{bmatrix}, \quad B := \begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 2 \\ 0 & 0 & 5 \end{bmatrix}, \quad C := \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 3 & 4 \end{bmatrix}.$$

Aufgabe 1.65

Gegeben ist die Matrix $\mathbf{A}$ in Abhängigkeit vom Parameter s :

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 \\ 5 & -4 & 1 \\ 6 & -6 & 3 \end{bmatrix} + s\mathbf{E} = \begin{bmatrix} 2+s & -1 & 1 \\ 5 & -4+s & 1 \\ 6 & -6 & 3+s \end{bmatrix}.$$

- a) Bestimmen Sie mittels der Determinante, für welche s die Matrix $\mathbf{A}$ keine Inverse $\mathbf{A}^{-1}$ besitzt. Diese Problemstellung führt zur Bestimmung von Eigenwerten $-s$.
 b) Wie lautet $\mathbf{A}^{-1}$ für $s = -2$?

Aufgabe 1.66

Lösen Sie das Gleichungssystem **a)** mit der Cramer'schen Regel, **b)** mit dem Gauß'schen Algorithmus, **c)** durch Multiplikation mit der inversen Matrix:

$$\begin{bmatrix} 5 & 2 & 1 \\ -1 & 1 & 1 \\ -3 & -2 & 2 \end{bmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 8 \\ 1 \\ -3 \end{pmatrix}.$$

Mit welchem Verfahren rechnen Sie am schnellsten?

Aufgabe 1.67

Berechnen Sie die folgenden Determinanten:

$$\text{a) } \begin{vmatrix} \sin(\alpha) & \cos(\alpha) \\ -\cos(\alpha) & \sin(\alpha) \end{vmatrix}, \quad \text{b) } \begin{vmatrix} 1 & x & x^2 \\ 1 & y & y^2 \\ 1 & z & z^2 \end{vmatrix}.$$



2 Differenzial- und Integralrechnung

Übersicht

2.1	Folgen	211
2.2	Zahlen-Reihen	236
2.3	Grenzwerte von Funktionen und Stetigkeit	255
2.4	Differenzierbarkeit und Ableitungen	288
2.5	Zentrale Sätze der Differenzialrechnung	314
2.6	Integralrechnung	328
2.7	Satz von Taylor, Kurvendiskussion und Extremalprobleme	382
2.8	Potenzreihen	398
2.9	Aufgaben	420

In diesem und den nächsten Kapiteln beschäftigen wir uns mit der Analysis. Kern dieser Disziplin ist der Umgang mit Näherungswerten, die sich beliebig genau machen lassen. Das führt zu Grenzwertaussagen wie die Berechnung von Ableitungen und Integralen. Damit können wir dann z. B. aus der momentanen Änderung des magnetischen Flusses eine induzierte Spannung berechnen. In diesem Kapitel wagen wir damit den Sprung ins unendlich Kleine und unendlich Große. Heute ist das ungefährlich, aber vor gut 400 Jahren war es das Todesurteil für den Philosophen Giordano Bruno, der für seine Thesen zur Unendlichkeit des Weltalls und der Zeit auf dem Scheiterhaufen landete.

Letztlich basieren alle Grenzwertaussagen auf den Eigenschaften von Folgen, mit denen wir das Kapitel beginnen.

2.1 Folgen

Beispiel 2.1 (nach Fonfara (2008), S. 366)

Ein Baggersee von $1\,500\text{ m}^2$ Größe wird so ausgebaggert, dass er jede Woche um 200 m^2 wächst. Gleichzeitig breiten sich Algen aus. Zu Beginn der Baggararbeiten

bedecken sie einen Quadratmeter der Wasseroberfläche. Die Algenfläche verdreifacht sich wöchentlich:

Wochenzahl n	0	1	2	3	4	5	6	7	8
Seefläche in m^2	1 500	1 700	1 900	2 100	2 300	2 500	2 700	2 900	3 100
Algenfläche in m^2	1	3	9	27	81	243	729	2 187	6 561

Nach knapp 8 Wochen bedecken die Algen also den ganzen See. Die Flächen des Sees und der Algen in Abhängigkeit der betrachteten Wochen sind Beispiele für Folgen, d. h. Abbildungen der natürlichen Zahlen auf reelle Zahlen. Beide Folgen lassen sich mittels einer Funktion darstellen, die jeder Woche n einen Funktionswert zuordnet:

- Folge der Seeflächen: $s(n) = s_n = 1\,500 + n \cdot 200$, $n \in \mathbb{N}_0$,
- Folge der Algenflächen: $a(n) = a_n = 3^n$, $n \in \mathbb{N}_0$. ■

2.1.1 Definition und Grundbegriffe von Folgen

Definition 2.1 (Folge)
*Eine Folge $(a_n)_{n=1}^\infty$ von reellen Zahlen (oder komplexen Zahlen) ist eine Abbildung von $\mathbb{N}$ nach $\mathbb{R}$ (oder $\mathbb{C}$), die jedem $n \in \mathbb{N}$ eindeutig ein Element $a_n \in \mathbb{R}$ (oder $a_n \in \mathbb{C}$) zuordnet. Die Zahlen a_n heißen **Folgenglieder**.*

Eine Folge kann als unendliches Tupel (als unendliche Auflistung) $(a_n)_{n=1}^\infty = (a_1, a_2, a_3, \dots)$ geschrieben werden, wobei an der k -ten Stelle der Liste das Folgenglied a_k steht, das die Folge als Abbildung der natürlichen Zahl k zuordnet.

Wenn Sie eine Programmiersprache kennen, in der es Felder (Arrays) gibt, dann können Sie eine Folge mit einem unendlich großen Feld vergleichen. Der Wert $a[k]$ der k -ten Position des Felds entspricht dem Folgenglied a_k .

Beispiel 2.2

- a) $(a_n)_{n=1}^\infty := (1, 1, 1, 1, 1, \dots)$, d. h. $a_n = 1$ für alle $n \in \mathbb{N}$.
- b) $(b_n)_{n=1}^\infty := (1, 0, 1, 0, 1, \dots)$, d. h. $b_n = 1$ für $n = 2k - 1$, $k \in \mathbb{N}$, und $b_n = 0$ für $n = 2k$, $k \in \mathbb{N}$ (gerade Indizes).
- c) Durch das Bildungsgesetz $c_n = 2 + \frac{1}{n}$, $n \in \mathbb{N}$, ist die Folge $(c_n)_{n=1}^\infty$ erklärt.
- d) Die Gleichung $d_{n+1} = \frac{1}{2} \left(d_n + \frac{1}{d_n} \right)$, $n = 1, 2, 3, \dots$, definiert die Folge $(d_n)_{n=1}^\infty$ rekursiv (d. h. über sich selbst, analog zur Vollständigen Induktion) in Abhängigkeit vom Anfangselement d_1 . Zum Beispiel erhalten wir für $d_1 = 1$ die Folge $(1, 1, 1, \dots)$, für $d_1 = 2$ erhalten wir $(2, \frac{5}{4}, \frac{41}{40}, \dots)$. ■

Eine Folge $(a_n)_{n=1}^{\infty}$ hat die **Wertemenge** $\{a_1, a_2, a_3, \dots\}$. Die ersten beiden Beispielfolgen haben die Wertemengen $W((a_n)_{n=1}^{\infty}) = \{1\}$ und $W((b_n)_{n=1}^{\infty}) = \{0, 1\}$. In diesen Mengen spielt die Reihenfolge der Folgenglieder keine Rolle mehr. Außerdem können bei einer Menge (im Gegensatz zu einer Folge) keine Elemente mehrfach vorkommen. Daher schreibt man eine Folge auch nicht als Menge und verwendet runde Klammern.

Die Menge reeller Zahlen $\{3, 1, 5, 4, 2\} = \{1, 2, 3, 4, 5\}$ kann man durch Striche auf einer Zahlengeraden darstellen. Dabei entsteht das gleiche Bild unabhängig von der Reihenfolge, in der die Elemente aufgelistet sind. Dagegen benötigt man zur Darstellung einer Folge zwei Achsen, und das Bild ändert sich bei einer Änderung der Reihenfolge (vgl. z. B. mit der Darstellung der Folgen in Abbildung 2.2).

Häufig sieht man auch Folgen, deren erstes Glied a_1 nicht den Index 1, sondern den Index 0 hat, also $a_0, (a_n)_{n=0}^{\infty}$. Dadurch ändern sich die Aussagen nicht.

Definition 2.2 (Besondere Folgen)

- Folgen $(a_n)_{n=1}^{\infty}$, bei denen alle Glieder die gleiche Zahl $a_n = a$ sind, heißen **konstante Folgen**.
- Bei einer **alternierenden Folge** $(a_n)_{n=1}^{\infty}$ ändern die Folgenglieder a_n fortlaufend das Vorzeichen, z. B. $a_n = (-1)^n$.
- Eine Folge $(a_n)_{n=0}^{\infty}$, die für eine feste Zahl d und alle $n \in \mathbb{N}_0$ gebildet wird über

$$a_n = a_0 + n \cdot d,$$

wird als **arithmetische Folge** bezeichnet. Die Differenz $d = a_{n+1} - a_n$ zweier aufeinander folgender Glieder ist konstant.

- Eine zu einer reellen Zahl $q \neq 0$ gebildete Folge $(a_n)_{n=0}^{\infty}$ mit $a_0 \neq 0$ und

$$a_n = a_0 \cdot q^n$$

wird als **geometrische Folge** bezeichnet, z. B. ist die Folge der Algenflächen aus Beispiel 2.1 eine geometrische Folge. Der Quotient zweier aufeinander folgender Glieder ist konstant: $\frac{a_{n+1}}{a_n} = q$.

Definition 2.3 (Beschränkte Folgen)

Eine reelle Folge $(a_n)_{n=1}^{\infty}$ heißt genau dann **nach oben beschränkt**, wenn ein $M \in \mathbb{R}$ (**obere Schranke**) existiert mit

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

Mit anderen Worten: Die Wertemenge $\{a_n : n \in \mathbb{N}\}$ ist nach oben beschränkt (siehe Definition 1.19 auf Seite 47, vgl. Definition 1.27 auf Seite 82). Die Folge heißt genau dann **nach unten beschränkt**, wenn ein $m \in \mathbb{R}$ (**untere Schranke**) existiert mit

$$a_n \geq m \text{ für alle } n \in \mathbb{N}.$$

Mit anderen Worten: Die Wertemenge $\{a_n : n \in \mathbb{N}\}$ ist nach unten beschränkt. Die Folge heißt **beschränkt** genau dann, wenn sie nach oben und unten beschränkt ist, d. h., wenn $(|a_n|)_{n=1}^\infty$ nach oben beschränkt ist.

Da für komplexe Zahlen $>$ und $<$ nicht definiert sind, kann man bei einer komplexen Folge nicht von Beschränktheit nach oben oder unten sprechen. Allerdings kann die Folge der Beträge beschränkt sein. In diesem Fall nennt man eine komplexe Folge **beschränkt**.

Beispiel 2.3

- a) $(a_n)_{n=1}^\infty := (1, 1, 1, 1, 1, \dots)$ ist nach oben und nach unten mit 1 beschränkt. Statt 1 könnten wir aber auch 4711 als obere Schranke wählen.
- b) $(b_n)_{n=1}^\infty := (1, 0, 1, 0, 1, \dots)$ ist nach unten durch -11 und nach oben durch 19 beschränkt. Genauso sind 0 eine untere und 1 eine obere Schranke.

■

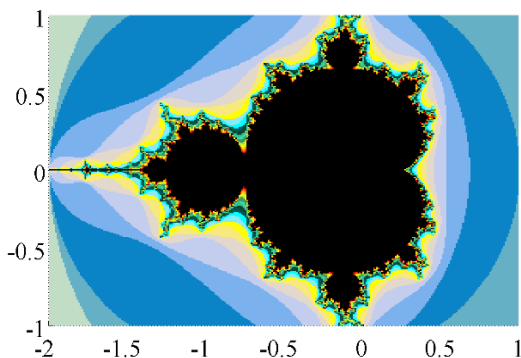


Abb. 2.1: Die Mandelbrotmenge als schwarze Fläche, $\operatorname{Re}(c) \in [-2, 1]$, $\operatorname{Im}(c) \in [-1, 1]$

Beispiel 2.4

Eine überraschende Struktur beobachtet man, wenn man zu jedem $c \in \mathbb{C}$ die komplexe Folge $(z_n)_{n=1}^\infty$ betrachtet, die per Induktion definiert ist über

$$z_1 := c \text{ und } z_{n+1} := z_n^2 + c.$$

Die Menge $\mathbb{M}$ der komplexen Zahlen c , für die die Folge beschränkt bleibt, heißt **Mandelbrot-Menge** nach dem Mathematiker Benoît Mandelbrot, der sie 1980 mit dem Computer visualisiert hat: Man fasst Bildschirmpunkte in der x-y-Ebene als komplexe Zahlen $c := x + jy$ auf und berechnet dazu die ersten Iterationen (30 in Abbildung 2.1) der jeweiligen Folge. Die Anzahl der Iterationen, die benötigt werden, damit die Beträge der Folgenglieder größer als zwei werden, wird als Grauton eingezeichnet. Die schwarze Fläche liegt näherungsweise über den Punkten c , für die die Folge beschränkt

ist und die damit die Mandelbrotmenge bilden. Genauer sind im Bild alle Punkte schwarz eingezeichnet, für die die Beträge der ersten 30 Folgenglieder den Wert 2 nicht überschreiten. Es lässt sich zeigen, dass bei einem Folgenglied mit Betrag größer 2 die Folge unbeschränkt ist. Da aber nur bis zum 30-ten Glied getestet wird, erhält man nur annähernd die Mandelbrotmenge (nämlich eine Obermenge).

- $c = 0$: $z_1 = 0$, $z_2 = \dots = z_{30} = \dots = 0$: Die Folge ist beschränkt, der Punkt ist schwarz.
- $c = j$: $z_1 = j$, $z_2 = j^2 + j = -1 + j$, $z_3 = (-1 + j)^2 + j = -2j + j = -j$, $z_4 = (-j)^2 + j = -1 + j$. Diesen Wert hatten wir bereits für z_2 , ab jetzt wiederholen sich die Werte, die Folge ist beschränkt, der Punkt ist schwarz.
- $c = 1$: $z_1 = 1$, $z_2 = 1 + 1 = 2$, $z_3 = 2^2 + 1 = 5 > 2$: Der Punkt erhält einen Grauton, der drei Iterationen zugeordnet ist. Die Folge ist hier nicht beschränkt.

Es entstehen wunderschöne Bilder (auch als „Apfelmännchen“ bekannt), die selbstähnlich sind: Vergrößert man Teilausschnitte des Randes der Mandelbrot-Menge immer weiter, so findet man immer wieder die gleichen Formen. ■

Wir betrachten im Folgenden zur Vereinfachung reelle Folgen. Alle Definitionen und Sätze, die keinen direkten Größenvergleich ($<$, $>$) der Folgenglieder ohne Betragsbildung beinhalten, gelten direkt auch für komplexe Folgen.

Die Monotoniebegriffe für Funktionen (siehe Seite 81) übertragen sich unmittelbar auf reelle Folgen:

Definition 2.4 (Monotonie)

Sei $(a_n)_{n=1}^{\infty}$ eine Folge reeller Zahlen. Sie heißt genau dann

- **monoton wachsend**, wenn $a_n \leq a_{n+1}$ für alle $n \in \mathbb{N}$ ist.
- **streng monoton wachsend**, wenn $a_n < a_{n+1}$ für alle $n \in \mathbb{N}$ gilt.
- **monoton fallend**, wenn $a_n \geq a_{n+1}$ für alle $n \in \mathbb{N}$ ist.
- **streng monoton fallend**, wenn $a_n > a_{n+1}$ für alle $n \in \mathbb{N}$ gilt.
- **monoton**, wenn die Folge monoton wächst oder fällt.

Beispiel 2.5

- a) Die Folge $(1/n)_{n=1}^{\infty}$ ist (streng) monoton fallend.
- b) Die konstante Folge $(1, 1, 1, \dots)$ ist sowohl monoton wachsend als auch monoton fallend.
- c) Die geometrische Folge mit Bildungsgesetz $a_n := q^n$, $n \in \mathbb{N}$, ist streng monoton fallend für $0 < q < 1$, denn es ist $q^{n+1} = q \cdot q^n < q^n$. Für $q > 1$ ist die Folge streng monoton wachsend.

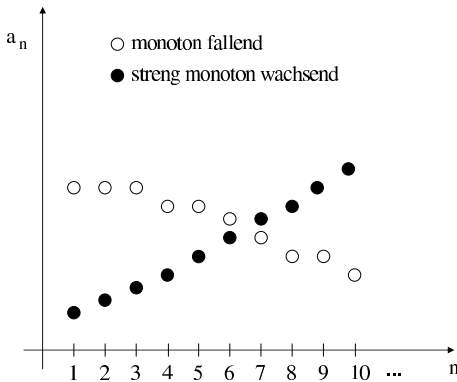


Abb. 2.2: Eine streng monoton wachsende und eine monoton fallende Folge

d) Die Folge $(a_n)_{n=1}^{\infty}$ mit $a_n = \sqrt[n]{c}$ und $c > 0$ ist monoton. Es gilt

$$\frac{a_{n+1}}{a_n} = \frac{c^{\frac{1}{n+1}}}{c^{\frac{1}{n}}} = c^{-\frac{1}{n(n+1)}}.$$

Für $0 < c < 1$ ist die Folge streng monoton wachsend, denn es gilt $a_{n+1} > a_n$. Für $c > 1$ ist die Folge streng monoton fallend.

■

2.1.2 Konvergenz und Divergenz von Folgen

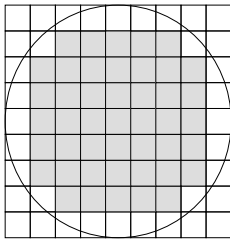


Abb. 2.3: Flächenberechnung durch Ausschöpfung mit Quadraten: Von den n^2 Quadraten liegen b_n vollständig innerhalb des Kreises mit Radius r . Zu erwarten ist, dass die Zahlen $b_n \cdot \left(\frac{2r}{n}\right)^2$ gegen πr^2 streben.

Beispiel 2.6

Wir wollen den Flächeninhalt eines Kreises mit Radius r berechnen. Dazu legen wir um den Kreis ein Quadrat mit Kantenlänge $2r$.

- **Monte-Carlo-Methode:** Wir wählen rein zufällig einen Punkt des Quadrats aus. Die Wahrscheinlichkeit, dass dieser im Kreis liegt, ist anschaulich unabhängig vom Radius, da Quadrat und Kreis in gleichem Maße mit r wachsen oder schrumpfen. Demnach gibt es eine Konstante c unabhängig von r , so dass der Kreisinhalt gleich $c \cdot 4r^2$ ist. Die Zahl π ist definiert als $\pi := 4 \cdot c$. Die Wahrscheinlichkeit, den Kreis

zu treffen, ist $\frac{c4r^2}{4r^2} = c = \frac{\pi}{4}$. Wir wiederholen das Experiment n -mal und zählen, wie oft wir im Kreis lagen. Dividieren wir diese Anzahl durch n , so erhalten wir eine Zahl a_n , die anschaulich eine Approximation an $\frac{\pi}{4}$ ist. So entsteht eine Folge $(a_n)_{n=1}^\infty$. Was passiert, wenn wir n immer größer wählen? Strebt a_n dann tatsächlich gegen $\frac{\pi}{4}$? Die Antwort darauf („ja“ bis auf pathologische Fälle) gibt das Gesetz der großen Zahlen, das wir im Rahmen der Wahrscheinlichkeitsrechnung als Satz 7.12 behandeln.

- Wir zerlegen das Quadrat in n^2 gleich große Quadrate und zählen, wie viele dieser Quadrate vollständig im Kreis liegen (siehe Abbildung 2.3). Die Anzahl sei b_n . Der Gesamtflächeninhalt dieser Quadrate ist $c_n := b_n \cdot \left(\frac{2r}{n}\right)^2$. Lässt man n gegen Unendlich streben, sollte die Folge $(c_n)_{n=1}^\infty$ gegen πr^2 streben. Dem ist auch so. Dahinter steckt die Integralrechnung, die wir in den Kapiteln 2.6 und 4.4 besprechen.

■

Jetzt konkretisieren wir den im Beispiel angedeuteten Übergang zum Unendlichen. Eine Folge $(a_n)_{n=1}^\infty$ strebt gegen eine Zahl a , wenn sich die Folgenglieder a_n mit wachsendem n immer mehr der Zahl a annähern, d. h., wenn der Abstand $|a_n - a|$ beliebig klein wird. Er muss kleiner als jede vorgegebene kleine Zahl $\varepsilon > 0$ werden. Präziser definiert man:

Definition 2.5 (Folgenkonvergenz)

Eine Folge $(a_n)_{n=1}^\infty$ heißt **konvergent** genau dann, wenn eine Zahl a existiert, so dass die folgende Bedingung erfüllt ist: Zu jedem (noch so kleinen) $\varepsilon > 0$ existiert ein $n_0 = n_0(\varepsilon) \in \mathbb{N}$, so dass ab diesem n_0 alle Folgenglieder in einem Streifen mit Radius ε um a liegen, d. h., wenn gilt:

$$|a_n - a| < \varepsilon \text{ für alle } n > n_0,$$

siehe Abbildung 2.4. Die Zahl a heißt dann **Grenzwert** oder **Limes** der Folge $(a_n)_{n=1}^\infty$, Schreibweise: $\lim_{n \rightarrow \infty} a_n = a$ (hier steht $\lim_{n \rightarrow \infty}$ für „Limes n gegen Unendlich“) oder auch $a_n \rightarrow a$ („ a_n strebt/geht gegen a “) für $n \rightarrow \infty$ („ n gegen Unendlich“).

Eine Folge $(a_n)_{n=1}^\infty$ heißt **divergent** genau dann, wenn sie nicht konvergent ist.

Also: $\lim_{n \rightarrow \infty} a_n = a$ gilt genau dann, wenn in jedem noch so schmalen Streifen $]a - \varepsilon, a + \varepsilon[$, $\varepsilon > 0$, alle bis auf endlich viele Folgenglieder liegen.

Der Umgang mit dieser ε - n_0 -Definition ist schwierig, wenn man sie zum ersten Mal sieht. Dabei löst sie aber das große Problem, das Streben gegen unendlich (oder später bei Funktionen das Streben gegen eine Stelle x_0) ohne anschauliche zeitliche Bewegungsabläufe und nur mit endlichen Mitteln darzustellen. Das ist wichtig, wenn man Konvergenz exakt beweisen möchte und sich nicht auf eine fehleranfällige Anschauung

verlassen möchte. Die bei Studierenden häufig nicht sehr beliebte „Epsilonantik“ ist ein Verdienst des großen westfälischen Mathematikers Karl Weierstraß (1815–1897).

Eine konvergente Folge $(a_n)_{n=1}^{\infty}$ mit dem Grenzwert $a = 0$ heißt **Nullfolge**. Zieht man von allen Gliedern einer konvergenten Folge ihren Grenzwert ab, so erhält man eine Nullfolge. Damit würde es prinzipiell ausreichen, nur die Eigenschaften von Nullfolgen zu untersuchen.

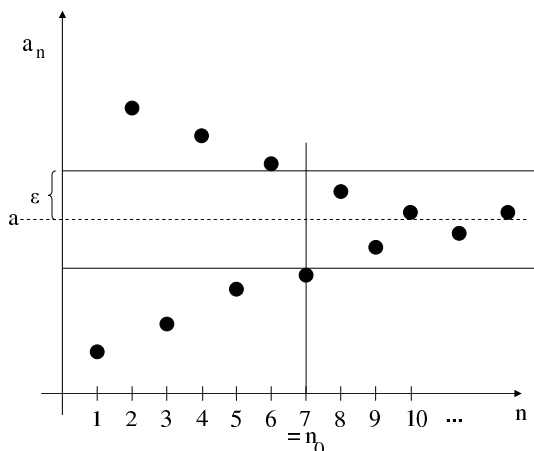


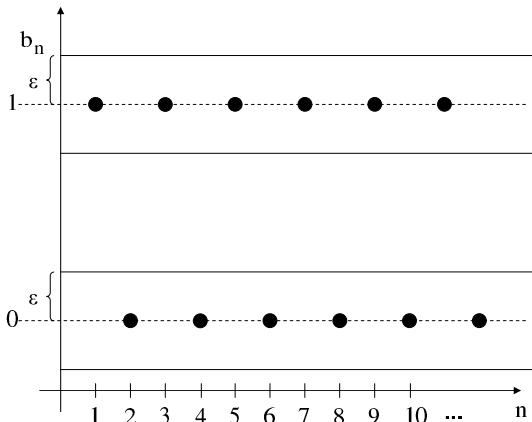
Abb. 2.4: ε - n_0 -Bedingung der Folgenkonvergenz

Beispiel 2.7

- a) Die Folge $(a_n)_{n=1}^{\infty} = (1, 1, 1, 1, 1, \dots)$ konvergiert gegen 1, $\lim_{n \rightarrow \infty} a_n = 1$.
- b) Die Folge $(b_n)_{n=1}^{\infty} := (1, 0, 1, 0, 1, \dots)$ konvergiert nicht, d. h., es gibt keinen Grenzwert. Um zu widerlegen, dass irgendein $b \in \mathbb{R}$ Grenzwert ist, wählen wir $\varepsilon = \frac{1}{4}$. Im Intervall $]b - \frac{1}{4}, b + \frac{1}{4}[$ können nicht sowohl 0 als auch 1 liegen (siehe Abbildung 2.5). Damit kann zu diesem $\varepsilon = \frac{1}{4}$ kein $n_0 \in \mathbb{N}$ gefunden werden, so dass für alle $n > n_0$ gilt: $|b_n - b| < \frac{1}{4}$, denn es gibt stets b_n , die den Wert 1 und andere, die den Wert 0 annehmen.
- c) Die Folge $(\frac{1}{n})_{n=1}^{\infty}$ konvergiert gegen 0, denn zu jedem $\varepsilon > 0$ gibt es ein $n_0 \in \mathbb{N}$ mit $n_0 > \frac{1}{\varepsilon}$. Damit ist $\frac{1}{n_0} < \varepsilon$ und auch $\frac{1}{n} < \varepsilon$ für alle $n > n_0$, d. h., ab n_0 liegen alle Folgenglieder in diesem ε -Streifen um den Grenzwert 0: $|\frac{1}{n} - 0| = \frac{1}{n} < \varepsilon$.
- d) Die Folgen $(\frac{1}{n^2})_{n=1}^{\infty}$, $(\frac{1}{n^3})_{n=1}^{\infty}$ usw. konvergieren ebenfalls gegen 0, da mit dem gleichen n_0 von oben zum Beispiel für $n > n_0$ gilt:

$$\left| \frac{1}{n^2} - 0 \right| = \frac{1}{n^2} < \frac{1}{n} < \varepsilon.$$

- e) Die Folge $(a_n)_{n=1}^{\infty}$ mit $a_n = 2 + \frac{1}{n}$ besitzt den Grenzwert $a = 2$. Zum Nachweis müssen wir zu $\varepsilon > 0$ die Stelle n_0 so wählen, dass für $n > n_0$ gilt: $|a_n - 2| = \frac{1}{n} < \varepsilon$. Diese Bedingung ist erfüllt, wenn $n > \frac{1}{\varepsilon}$ ist. Also können wir für n_0 irgendeine natürliche Zahl größer oder gleich $\frac{1}{\varepsilon}$ wählen.

Abb. 2.5: Divergente Folge $(b_n)_{n=1}^\infty$

- Für $\varepsilon = \frac{1}{10}$ wählt man $n_0(\varepsilon) \geq \frac{1}{\frac{1}{10}}$, also z. B. $n_0 = 10$. Dann ist $|a_n - 2| < \frac{1}{10}$ für $n > 10$.
 - Für $\varepsilon = \frac{1}{100}$ wählt man $n_0(\varepsilon) \geq \frac{1}{\frac{1}{100}}$, also z. B. $n_0 = 100$. Dann ist $|a_n - 2| < \frac{1}{100}$ für $n > 100$.
- f) 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.

- g) Die Folge mit Bildungsgesetz $a_n = \frac{n^2+2n}{n^2+1}$ besitzt den Grenzwert $a = 1$. Zum Nachweis dieser Vermutung bestimmen wir zu $\varepsilon > 0$ ein $n_0(\varepsilon)$ so, dass $|a_n - 1| < \varepsilon$ für $n > n_0(\varepsilon)$ gilt. Mit der Abschätzung

$$|a_n - 1| = \left| \frac{n^2 + 2n - (n^2 + 1)}{n^2 + 1} \right| = \left| \frac{2n - 1}{n^2 + 1} \right| \leq \frac{2n}{n^2} = \frac{2}{n}$$

wählen wir $n_0(\varepsilon)$ gemäß

$$|a_n - 1| \leq \frac{2}{n} < \varepsilon \implies n > \frac{2}{\varepsilon}$$

zu $n_0(\varepsilon) > \frac{2}{\varepsilon}$. ■

2.1.3 Rechnen mit konvergenten Folgen

Es wäre sehr mühsam, wenn man jede Folge mit der Definition der Konvergenz auf einen Grenzwert untersuchen müsste. Glücklicherweise nehmen einem die folgenden Grenzwertsätze viel Arbeit ab:

Satz 2.1 (Grenzwertsätze)

Seien $(a_n)_{n=1}^{\infty}$ und $(b_n)_{n=1}^{\infty}$ Folgen sowie $c \in \mathbb{R}$.

- a) Eine Folge besitzt höchstens einen Grenzwert.
- b) Jede konvergente Folge ist notwendigerweise beschränkt.
- c) Falls $\lim_{n \rightarrow \infty} a_n = a$ und $\lim_{n \rightarrow \infty} b_n = b$ ist, dann gilt:
 - i) $\lim_{n \rightarrow \infty} (a_n + b_n) = a + b$,
 - ii) $\lim_{n \rightarrow \infty} (c \cdot a_n) = c \cdot a$,
 - iii) $\lim_{n \rightarrow \infty} (a_n \cdot b_n) = a \cdot b$.
 - iv) Falls zusätzlich $b \neq 0$ gilt, so ist $b_n \neq 0$ für genügend große n und

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{a}{b}.$$

- d) Ist $(a_n)_{n=1}^{\infty}$ konvergent gegen a und $a_n \geq c$ für alle $n \in \mathbb{N}$, so ist auch $a \geq c$.
(Entsprechend ist für $a_n \leq c$ auch $a \leq c$.)

Die einzelnen Teilaussagen lassen sich alle direkt mit der Definition des Grenzwerts beweisen. Das Vorgehen basiert auf Abschätzungen mit der Dreiecksungleichung (1.14) und der Dreiecksungleichung nach unten (1.13) von Seite 68 und ist typisch für viele Beweise in der Analysis. Wenn Sie davon einen Eindruck bekommen möchten, dann vollziehen Sie ein paar Punkte nach, sonst können Sie direkt den Nutzen des Satzes am nachfolgenden Beispiel sehen.

Beweis: a) Gäbe es zwei Grenzwerte, so müssten alle Folgenglieder bis auf endlich viele in jedem Streifen um jeden der beiden Grenzwerte liegen. Wählt man die Streifen so klein, dass sie sich nicht überlappen, geht das aber nicht (vgl. Abbildung 2.5). Dieses Argument lautet in formaler Schreibweise so: Annahme: Die konvergente Folge $(a_n)_{n=1}^{\infty}$ habe (mindestens) zwei verschiedene Grenzwerte $a \neq a'$, d. h. $\lim_{n \rightarrow \infty} a_n = a$ und $\lim_{n \rightarrow \infty} a_n = a'$. Wähle $\varepsilon := |a - a'|/4 > 0$. Wegen $\lim_{n \rightarrow \infty} a_n = a$ existiert zu diesem $\varepsilon > 0$ ein n_0 , so dass für alle $n > n_0$ die Folgenglieder a_n näher bei a liegen als ε , d. h. $|a_n - a| < \varepsilon$. Wegen $\lim_{n \rightarrow \infty} a_n = a'$ existiert zu diesem $\varepsilon > 0$ ebenfalls ein n_1 , so dass die Folgenglieder für $n > n_1$ näher an a' liegen als ε : $|a_n - a'| < \varepsilon$. Für jedes $n > n_2 := \max\{n_0, n_1\}$ sind beide

Bedingungen erfüllt, und es ergibt sich der Widerspruch für den Abstand von a und a' :

$$|a - a'| = |a - a_n + a_n - a'| \leq |a - a_n| + |a_n - a'| < 2\varepsilon = \frac{1}{2}|a - a'|.$$

- b) Wir wählen einen Streifen um den Grenzwert. Dann liegen in diesem Streifen alle bis auf endlich viele Glieder. Die Beträge der Glieder im Streifen sind offensichtlich beschränkt. Bildet man nun das Maximum dieser oberen Schranke und der Beträge der übrigen endlich vielen Glieder, so erhält man eine obere Schranke für die gesamte Folge $(|a_n|)_{n=1}^{\infty}$. Formal sieht dieses Argument z. B. so aus: Es sei $\lim_{n \rightarrow \infty} a_n = a$. Zu $\varepsilon := 1$ existiert ein $n_0 \in \mathbb{N}$, so dass für alle $n > n_0$ gilt:

$$1 > |a_n - a| \stackrel{(1.13)}{\geq} ||a_n| - |a|| \geq |a_n| - |a|,$$

also $|a_n| < 1 + |a|$. Die Folge $(|a_n|)_{n=1}^{\infty}$ ist also nach oben beschränkt mit $\max\{\max\{|a_n| : 1 \leq n \leq n_0\}, 1 + |a|\}$.

- c) i) Nach Voraussetzung existieren zu jedem $\varepsilon > 0$ Stellen $n_1, n_2 \in \mathbb{N}$, die gemäß der Grenzwertdefinition für einen $\frac{\varepsilon}{2}$ -Streifen so gewählt werden können, dass $|a_n - a| < \frac{\varepsilon}{2}$ für $n > n_1$ und $|b_n - b| < \frac{\varepsilon}{2}$ für $n > n_2$ ist. Mit der Dreiecksungleichung (1.14) erhalten wir daraus für $n > n_0 := \max\{n_1, n_2\}$:

$$|(a_n + b_n) - (a + b)| \leq |a_n - a| + |b_n - b| < 2 \cdot \frac{\varepsilon}{2} = \varepsilon.$$

Damit ist die Konvergenz der Summenfolge gegen den Grenzwert $a + b$ gezeigt.

Bemerkung: Dass man mit $\frac{\varepsilon}{2}$ arbeiten muss, sieht man, wenn man die Abschätzung mit der Dreiecksungleichung macht. Dann kann man die Darstellung rückwirkend so anpassen, dass am Ende tatsächlich ε herauskommt. Dass man am Ende einer Abschätzung genau auf ε kommt, ist ästhetisch schön, aber für einen Beweis nicht nötig. Wenn man gegen $c \cdot \varepsilon$ für eine von n und n_0 unabhängige Konstante $c > 0$ abschätzen kann, kommt man durch Übergang von ε zu $\frac{\varepsilon}{c}$ wie oben im Fall $c = 2$ auf die Darstellung aus der Konvergenzdefinition.

- ii) Ist $c = 0$, so ist offensichtlich $\lim_{n \rightarrow \infty} c \cdot a_n = 0 = c \cdot a$. Sei nun $c \neq 0$. Nach Voraussetzung existiert zu jedem $\varepsilon > 0$ eine Stelle $n_0 \in \mathbb{N}$, die gemäß der Grenzwertdefinition für einen $\frac{\varepsilon}{|c|}$ -Streifen so gewählt werden kann, dass $|a_n - a| < \frac{\varepsilon}{|c|}$ für $n > n_0$ gilt. Damit ist aber auch

$$|c \cdot a_n - c \cdot a| = |c||a_n - a| < |c| \frac{\varepsilon}{|c|} = \varepsilon.$$

- iii) Da die Folge $(a_n)_{n=1}^{\infty}$ konvergent ist, ist sie nach b) beschränkt. Es gibt ein $M > 0$ mit $|a_n| \leq M$ für alle $n \in \mathbb{N}$. Nach Voraussetzung existieren zu jedem $\varepsilon > 0$ Stellen $n_1, n_2 \in \mathbb{N}$, die gemäß der Grenzwertdefinition für einen $\frac{\varepsilon}{M+|b|}$ -Streifen so gewählt werden können, dass $|a_n - a| < \frac{\varepsilon}{M+|b|}$ für $n > n_1$ und

$|b_n - b| < \frac{\varepsilon}{M+|b|}$ für $n > n_2$ ist. Wieder führt die Dreiecksungleichung (1.14) zum Ziel für $n > n_0 := \max\{n_1, n_2\}$:

$$\begin{aligned} |a_n b_n - ab| &= |a_n b_n - a_n b + a_n b - ab| \leq |a_n(b_n - b)| + |(a_n - a)b| \\ &\leq M|b_n - b| + |b||a_n - a| < (M + |b|) \frac{\varepsilon}{M + |b|} = \varepsilon. \end{aligned}$$

iv) Die Quotientenregel kann auf die Produktregel zurückgeführt werden. Dazu müssen wir zeigen, dass $\lim_{n \rightarrow \infty} \frac{1}{b_n} = \frac{1}{b}$ ist. Man beachte, dass in iv) nicht ausgeschlossen ist, dass $b_n = 0$ für einige $n \in \mathbb{N}$ ist. Da aber der Grenzwert $b \neq 0$ ist, können wir um b einen Streifen mit Radius $|b|/2$ legen, in dem 0 nicht enthalten ist, dafür aber alle bis auf endlich viele Folgenglieder. Es gibt also nur höchstens endlich viele Folgenglieder, die 0 sein können. Wir betrachten hier die Folge erst ab einem n_0 , so dass $|b_n| > |b|/2 > 0$ für alle $n \geq n_0$. Dafür ist der Quotient $\frac{1}{b_n}$ wohldefiniert. Da $\lim_{n \rightarrow \infty} b_n = b$ gefordert ist, existiert zu jedem $\varepsilon > 0$ eine Stelle n_1 , so dass ab dieser Position alle Folgenglieder innerhalb eines $|b|^2 \varepsilon/2$ -Streifens um b liegen, d. h.

$$|b_n - b| < \frac{|b|^2}{2} \varepsilon \quad (2.1)$$

für alle $n > n_1$. Damit gilt für alle $n > n_2 := \max\{n_0, n_1\}$:

$$\left| \frac{1}{b_n} - \frac{1}{b} \right| = \left| \frac{b - b_n}{b_n b} \right| \stackrel{|b_n| > |b|/2}{\leq} |b - b_n| \frac{1}{\frac{|b|^2}{2}} \stackrel{(2.1)}{<} \varepsilon.$$

Damit ist gezeigt, dass $1/b$ der Grenzwert der Folge $(1/b_n)_{n=1}^\infty$ ist. Mit iii) erhalten wir $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lim_{n \rightarrow \infty} a_n \cdot \lim_{n \rightarrow \infty} \frac{1}{b_n} = \frac{a}{b}$.

d) Wenn der Grenzwert $a < c$ wäre, gäbe es einen Mindestabstand zwischen Folgengliedern ($\geq c$) und Grenzwert, z. B. $c - a$. Wählt man dann einen Streifen kleinerer Breite um den Grenzwert, lägen in diesem keine Folgenglieder – im Widerspruch zur Definition des Grenzwerts.

□

Beispiel 2.8

Die Regeln des Satzes können genutzt werden, um Grenzwerte auszurechnen. Wir möchten mit der Quotientenregel c) iv) den Grenzwert $\lim_{n \rightarrow \infty} \frac{2n^2 - 2}{n^2 + 2n + 2}$ bestimmen. Man beachte, dass die Grenzwertregeln die Konvergenz der einzelnen Folgen voraussetzen. Dabei haben wir hier das Problem, dass die Grenzwerte des Zählers und des Nenners einzeln nicht existieren. Deren Werte werden mit n immer größer, ohne sich einer Zahl anzunähern. So darf die Quotientenregeln **keinesfalls** angewendet werden: $\lim_{n \rightarrow \infty} \frac{2n^2 - 2}{n^2 + 2n + 2} = \frac{\lim_{n \rightarrow \infty} (2n^2 - 2)}{\lim_{n \rightarrow \infty} (n^2 + 2n + 2)}$. Das Verhalten von Zähler und Nenner ändert sich aber, wenn man den Bruch mit $\frac{1}{n^2}$ erweitert. Nun bilden sowohl Zähler als auch Nenner konvergente Folgen, und c) iv) ist anwendbar:

$$\lim_{n \rightarrow \infty} \frac{2n^2 - 2}{n^2 + 2n + 2} = \lim_{n \rightarrow \infty} \frac{2 - \frac{2}{n^2}}{1 + \frac{2}{n} + \frac{2}{n^2}} \stackrel{\text{c), iv)}}{=} \frac{\lim_{n \rightarrow \infty} (2 - \frac{2}{n^2})}{\lim_{n \rightarrow \infty} (1 + \frac{2}{n} + \frac{2}{n^2})}$$

$$\stackrel{\text{c), i, ii)}}{=} \frac{2 - 2 \cdot \lim_{n \rightarrow \infty} \frac{1}{n^2}}{1 + 2 \cdot \lim_{n \rightarrow \infty} \frac{1}{n} + 2 \cdot \lim_{n \rightarrow \infty} \frac{1}{n^2}} = \frac{2 - 0}{1 + 0 + 0} = 2. \quad \blacksquare$$

Dieses Vorgehen lässt sich so verallgemeinern: Der höchste Exponent des Zählers und des Nenners setzt sich mit den zugehörigen Vorfaktoren $a_k, b_m \neq 0$ durch:

$$\lim_{n \rightarrow \infty} \frac{a_k n^k + a_{k-1} n^{k-1} + \cdots + a_1 n + a_0}{b_m n^m + b_{m-1} n^{m-1} + \cdots + b_1 n + b_0} = \begin{cases} \frac{a_k}{b_m} & \text{falls } k = m, \\ 0 & \text{falls } k < m. \end{cases} \quad (2.2)$$

Das ergibt sich wie im Beispiel, indem man mit $\frac{1}{n^m}$ erweitert und die Grenzwertsätze benutzt. Im Fall $k > m$ haben wir keine Konvergenz, damit werden wir uns später beschäftigen.

Mit den Regeln sieht man auch sofort: $\lim_{n \rightarrow \infty} a_n = a \iff \lim_{n \rightarrow \infty} a_n - a = 0$. Dazu ist wegen der Definition der Konvergenz auch $\lim_{n \rightarrow \infty} |a_n - a| = 0$ äquivalent.

2.1.4 Konvergenzkriterien

Lemma 2.1 (Arithmetisches Mittel einer Folge)

Sei $(a_n)_{n=1}^\infty$ eine konvergente Folge mit $\lim_{n \rightarrow \infty} a_n = a$. Dann ist auch die Folge $(b_n)_{n=1}^\infty$, die über die arithmetischen Mittel $b_n := \frac{1}{n} \sum_{k=1}^n a_k$ gebildet ist, konvergent mit $\lim_{n \rightarrow \infty} b_n = a$. In diesem Sinne ändert Mittelwertbildung nichts am Konvergenzverhalten.

Beweis: Da die Folge $(a_n)_{n=1}^\infty$ konvergent ist, existiert zu jedem $\varepsilon > 0$ eine Stelle $n_0 \in \mathbb{N}$, ab der alle Folgenglieder näher beim Grenzwert a liegen als $\frac{\varepsilon}{2}$, d. h., für $n > n_0$ gilt: $|a_n - a| < \frac{\varepsilon}{2}$. Bilden wir nun die arithmetischen Mittel b_n für $n > n_0$, so sind alle Summanden ab n_0 entsprechend nah bei a . Lediglich die ersten n_0 Summanden können stärker abweichen:

$$\begin{aligned} |b_n - a| &= \left| \left(\frac{1}{n} \sum_{k=1}^n a_k \right) - a \right| = \left| \frac{1}{n} \sum_{k=1}^n (a_k - a) \right| \leq \frac{1}{n} \sum_{k=1}^n |a_k - a| \\ &= \frac{1}{n} \sum_{k=1}^{n_0} |a_k - a| + \frac{1}{n} \sum_{k=n_0+1}^n |a_k - a| \\ &< \frac{1}{n} \sum_{k=1}^{n_0} |a_k - a| + \frac{1}{n} (n - n_0) \frac{\varepsilon}{2} \leq \frac{1}{n} \sum_{k=1}^{n_0} |a_k - a| + \frac{\varepsilon}{2}. \end{aligned}$$

Betrachtet man nun große Werte für n , so wird durch die Multiplikation mit $\frac{1}{n}$ auch die erste Summe für das feste n_0 sehr klein. Für $n > N_0 := \max \left\{ n_0, \frac{2}{\varepsilon} \sum_{k=1}^{n_0} |a_k - a| \right\}$ ist $\left| \left(\frac{1}{n} \sum_{k=1}^n a_k \right) - a \right| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$, und wir haben die Konvergenz der Folge der Mittelwerte bewiesen. Denn es gibt für das eingangs gewählte ε eine Stelle $N_0 \in \mathbb{N}$, so dass ab dieser alle Folgenwerte b_n näher bei a liegen als ε . $\square$

Satz 2.2 (Einschließungskriterium)

Es seien $(a_n)_{n=1}^\infty$, $(b_n)_{n=1}^\infty$ zwei konvergente Folgen mit dem gleichen Grenzwert a . Gilt für die Glieder der Folge $(c_n)_{n=1}^\infty$ für alle Indizes n , die größer als ein fester Index n_0 sind, die Einschließung

$$a_n \leq c_n \leq b_n,$$

so ist auch die Folge $(c_n)_{n=1}^\infty$ konvergent mit gleichem Grenzwert a .

Beweis: Nach Voraussetzung ist $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = a$, so dass zu jedem $\varepsilon > 0$ Stellen $n_1, n_2 \in \mathbb{N}$ existieren mit:

$$|a_n - a| < \varepsilon \quad \text{d. h.} \quad a - \varepsilon < a_n < a + \varepsilon \quad \text{für alle } n \geq n_1,$$

$$|b_n - a| < \varepsilon \quad \text{d. h.} \quad a - \varepsilon < b_n < a + \varepsilon \quad \text{für alle } n \geq n_2.$$

Für $n \geq n_3 := \max\{n_0, n_1, n_2\}$ gilt daher $a - \varepsilon < a_n \leq c_n \leq b_n < a + \varepsilon$, also $-\varepsilon < c_n - a < \varepsilon$ bzw. $|c_n - a| < \varepsilon$. Damit ist die Konvergenz von $(c_n)_{n=1}^\infty$ gegen a gezeigt. $\square$

Beispiel 2.9

a) Die Folge $a_n = \frac{n}{2^n}$ strebt gegen $a = 0$. Denn für $n \geq 4$ gilt $n^2 \leq 2^n$ (siehe Aufgabe 1.11), und weiter folgt

$$0 \leq \frac{n}{2^n} = \frac{1}{n} \cdot \frac{n^2}{2^n} \leq \frac{1}{n}.$$

Der Grenzwert $a = 0$ folgt mit dem Einschließungskriterium.

b) Die Folge $a_n = \sqrt[n]{4^n + 7^n}$ hat den Grenzwert $a = 7$, denn es gilt

$$\sqrt[n]{7^n} \leq \sqrt[n]{4^n + 7^n} \leq \sqrt[n]{7^n + 7^n}$$

und weiter

$$7 \leq \sqrt[n]{4^n + 7^n} \leq \sqrt[n]{2} \cdot 7.$$

Man kann zeigen, dass $\lim_{n \rightarrow \infty} \sqrt[n]{2} = \lim_{n \rightarrow \infty} \exp\left(\frac{1}{n} \ln 2\right) = 1$ ist. Dazu zieht man den Grenzwert in das Argument der Exponentialfunktion. Dass das geht, werden wir sehen, wenn wir uns mit Stetigkeit beschäftigen. Damit ergibt sich der Grenzwert $a = 7$ mit dem Einschließungskriterium. $\blacksquare$

Satz 2.3 (Monotoniekriterium)

a) Jede monoton wachsende, nach oben beschränkte Folge $(a_n)_{n=1}^\infty$ ist konvergent. Für jedes $k \in \mathbb{N}$ ist $a_k \leq \lim_{n \rightarrow \infty} a_n$.

b) Jede monoton fallende, nach unten beschränkte Folge $(a_n)_{n=1}^\infty$ ist konvergent. Für jedes $k \in \mathbb{N}$ ist $a_k \geq \lim_{n \rightarrow \infty} a_n$.

c) Jede monotone, beschränkte Folge ist konvergent.

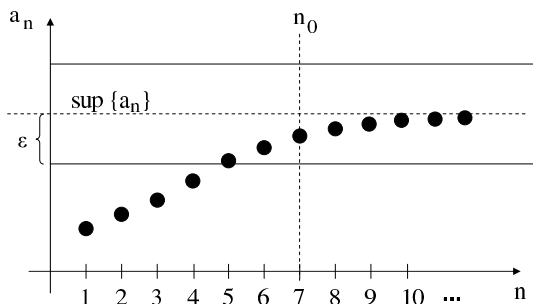


Abb. 2.6: Eine beschränkte, monoton wachsende Folge ist konvergent.

Ist eine Folge monoton wachsend und beschränkt, dann nähern sich die Folgenglieder zwangsläufig immer mehr der kleinsten oberen Schranke aller Folgenglieder an, die damit zum Grenzwert wird. Das benutzen wir im folgenden Beweis, wobei wir zusätzlich für die Existenz des Grenzwerts wissen müssen, dass es tatsächlich aufgrund der Voraussetzungen die kleinste obere Schranke (ein Supremum) der Wertemenge der Folge gibt. Das stellt aber das Vollständigkeitsaxiom (1.5) von Seite 49 sicher. Für Konvergenzfragen braucht man also mit der Vollständigkeit die zentrale Eigenschaft, die $\mathbb{R}$ im Gegensatz zu $\mathbb{Q}$ bietet.

Beweis: Wir zeigen die Aussage für monoton wachsende Folgen, für fallende ist der Beweis analog. Da die Folge $(a_n)_{n=1}^{\infty}$ nach oben beschränkt ist, gilt dies insbesondere für die nicht-leere Wertemenge $E := \{a_n : n \in \mathbb{N}\}$. Diese hat nach Axiom (1.5) ein Supremum $a \in \mathbb{R}$ (vgl. Abbildung 2.6). Da a das Supremum ist, gibt es wegen (1.6) auf Seite 50 zu jedem $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$, so dass $a - a_{n_0} < \varepsilon$. Da $(a_n)_{n=1}^{\infty}$ monoton steigend ist, gilt damit aber auch für alle $n \geq n_0$:

$$|a - a_n| = a - a_n \leq a - a_{n_0} < \varepsilon.$$

Damit ist die Konvergenz gegen a bewiesen. □

Beispiel 2.10

a) $(\frac{1}{n})_{n=1}^{\infty}$ ist nach unten beschränkt und monoton fallend. Wir haben bereits zuvor gesehen, dass diese Folge auch konvergent ist.

b) Wir untersuchen die Folge zu $a_n = \frac{9n}{n+3}$ auf Beschränktheit und Monotonie.

Durch Ausklammern und Kürzen von n in Zähler und Nenner erhält man $a_n = \frac{9}{1+\frac{3}{n}}$. $(a_n)_{n=1}^{\infty}$ ist nach oben durch $M = 9$ und wegen $a_n > 0$ nach unten durch $m = 0$ beschränkt.

Zur Untersuchung der Monotonie bilden wir die Differenz zweier aufeinander folgender Glieder

$$a_{n+1} - a_n = \frac{9(n+1)}{(n+1)+3} - \frac{9n}{n+3}$$

$$= \frac{9(n+1)(n+3) - 9n(n+4)}{(n+4)(n+3)} = \frac{27}{(n+4)(n+3)} > 0,$$

woraus $a_{n+1} > a_n$ für alle $n \in \mathbb{N}$ folgt, d. h. streng monotonen Wachstum der Folgenglieder. Damit ist die Folge $(a_n)_{n=1}^\infty$ konvergent. Das wussten wir bereits, da wir sogar Grenzwerte von Folgen dieses Typs mit den Grenzwertsätzen ausrechnen können: $\lim_{n \rightarrow \infty} \frac{9n}{n+3} = \frac{9}{1} = 9$. ■

Beispiel 2.11

Wir zeigen, dass die rekursiv definierte Folge $(a_n)_{n=1}^\infty$ mit

$$a_1 = 1, a_{n+1} := \frac{1}{2}(a_n + 7)$$

konvergent ist, indem wir nachrechnen, dass sie monoton wachsend und beschränkt nach oben (z. B. mit 7) ist. Wir zeigen beide Eigenschaften per Vollständiger Induktion:

- Zur Monotonie: Induktionsanfang für $n = 1$: $a_2 = \frac{1+7}{2} = 4 > 1 = a_1$.
Induktionsschritt: Sei $a_{n+1} > a_n$. Dann ist $a_{n+2} = \frac{1}{2}(a_{n+1} + 7) > \frac{1}{2}(a_n + 7) = a_{n+1}$.
- Zur Beschränktheit: Induktionsanfang für $n = 1$: $a_1 = 1 < 7$.
Induktionsschritt: Sei $a_n < 7$. Dann ist $a_{n+1} = \frac{1}{2}(a_n + 7) < \frac{1}{2}(7 + 7) = 7$. ■

2.1.5 Die Euler'sche Zahl e als Grenzwert von Folgen

Beispiel 2.12

Wir haben mit (1.8) auf Seite 52 halbformal die Zahl e eingeführt über die Summe $e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \dots$. Wir können diese Lücke nun schließen, indem wir e als Grenzwert der Folge

$$(a_n)_{n=1}^\infty := \left(\sum_{k=0}^n \frac{1}{k!} \right)_{n=1}^\infty \quad (2.3)$$

definieren. Bei jedem Folgenglied kommt hier ein weiterer Summand hinzu:

$$\begin{aligned} a_1 &:= \frac{1}{0!} + \frac{1}{1!} \\ a_2 &:= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} \\ a_3 &:= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} \\ &\vdots \end{aligned}$$

Man nennt Folgen dieses Typs auch **Reihen** (siehe Kapitel 2.2). Wir müssen zeigen, dass die Folge konvergiert. Dann haben wir einen Grenzwert zur Verfügung, den wir e nennen. Die Folge $(a_n)_{n=1}^{\infty}$ ist streng monoton wachsend, da beim Übergang von a_n zu a_{n+1} die positive Zahl $\frac{1}{(n+1)!}$ addiert wird:

$$a_{n+1} - a_n = \sum_{k=0}^{n+1} \frac{1}{k!} - \sum_{k=0}^n \frac{1}{k!} = \frac{1}{(n+1)!} > 0 \implies a_{n+1} > a_n.$$

Die Folge ist aber auch nach oben beschränkt und damit konvergent: Dazu benötigen wir zunächst eine Hilfsaussage:

Für $k \geq 2$ ist $k! = 1 \cdot \underbrace{2 \cdot 3 \cdot \dots \cdot k}_{k-1 \text{ Stück}} \geq 2^{k-1}$, und für $k = 1$ ist $1! = 1 \geq 1 = 2^0$. Es

gilt also für $k \in \mathbb{N}$:

$$k! \geq 2^{k-1}, \quad \frac{1}{k!} \leq \frac{1}{2^{k-1}}. \quad (2.4)$$

Damit zeigen wir die Beschränktheit: Für $n \geq 1$ ist

$$a_n = 1 + \sum_{k=1}^n \frac{1}{k!} \stackrel{(2.4)}{\leq} 1 + \sum_{k=1}^n \frac{1}{2^{k-1}} = 1 + \sum_{k=0}^{n-1} \frac{1}{2^k} \stackrel{(1.10)}{=} 1 + \frac{1 - \frac{1}{2^n}}{\frac{1}{2}} < 3. \quad (2.5)$$

Dabei haben wir die geometrische Summe mit der Formel (1.10) von Seite 60 ausgerechnet. ■

Beispiel 2.13

Die Folge $\left(1 + \frac{1}{n}\right)_{n=1}^{\infty}$ ist streng monoton wachsend und beschränkt und damit konvergent. Der Nachweis der Eigenschaften ist etwas mühsam. Zunächst rechnen wir die Potenz mit dem Binomischen Lehrsatz (1.11) von Seite 64 aus (für das Produktsymbol $\prod$ siehe Kapitel 1.4.2):

$$\begin{aligned} a_n &:= \left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^n \binom{n}{k} \frac{1}{n^k} = 1 + \sum_{k=1}^n \frac{n(n-1) \cdots (n-k+1)}{k! n^k} \\ &= 1 + \sum_{k=1}^n \frac{1}{k!} \prod_{l=0}^{k-1} \frac{n-l}{n} = 1 + \sum_{k=1}^n \frac{1}{k!} \prod_{l=0}^{k-1} \left(1 - \frac{l}{n}\right). \end{aligned} \quad (2.6)$$

Für $0 \leq l \leq n-1$ ist

$$0 \leq 1 - \frac{l}{n} \leq 1 - \frac{l}{n+1} \leq 1. \quad (2.7)$$

$(a_n)_{n=1}^{\infty}$ ist streng monoton wachsend:

$$\begin{aligned} a_{n+1} - a_n &\stackrel{(2.6)}{=} \sum_{k=1}^{n+1} \frac{1}{k!} \prod_{l=0}^{k-1} \left(1 - \frac{l}{n+1}\right) - \sum_{k=1}^n \frac{1}{k!} \prod_{l=0}^{k-1} \left(1 - \frac{l}{n}\right) \\ &= \frac{1}{(n+1)!} \prod_{l=0}^n \left(1 - \frac{l}{n+1}\right) + \sum_{k=1}^n \frac{1}{k!} \left[\prod_{l=0}^{k-1} \left(1 - \frac{l}{n+1}\right) - \prod_{l=0}^{k-1} \left(1 - \frac{l}{n}\right) \right] \end{aligned}$$

$$\stackrel{(2.7)}{\geq} \frac{1}{(n+1)!} \prod_{l=0}^n \left(1 - \frac{l}{n+1}\right) > 0.$$

$(a_n)_{n=1}^{\infty}$ ist beschränkt:

$$2 = a_1 < a_n \stackrel{(2.6)}{=} 1 + \sum_{k=1}^n \frac{1}{k!} \prod_{l=0}^{k-1} \left(1 - \frac{l}{n}\right) \stackrel{(2.7)}{\leq} \sum_{k=0}^n \frac{1}{k!} \stackrel{(2.5)}{<} 3.$$

Wegen des vorangehenden Beispiels können wir statt gegen 3 sogar gegen den Grenzwert e der streng monoton wachsenden Folge $(\sum_{k=0}^n \frac{1}{k!})$ abschätzen. ■

Auf ähnliche Weise kann man zeigen, dass nicht nur $\lim_{n \rightarrow \infty} a_n \leq e$, sondern sogar $\lim_{n \rightarrow \infty} a_n = e$ gilt (vgl. Aufgabe 1.14, siehe z. B. (Endl und Luh, 1989, S. 63); dort wird auch bewiesen, dass e keine rationale Zahl ist). Wir haben also:

$$e := \lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{1}{k!} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = \lim_{n \rightarrow \infty} (1 + a_n)^{\frac{1}{a_n}}, \quad (2.8)$$

dabei kann $a_n = 1/n$ ersetzt werden durch eine beliebige Nullfolge mit positiven Gliedern, siehe z. B. (Arens et al., 2012, S. 267).

Beispiel 2.14

Wir bestimmen mit (2.8) zwei Grenzwerte:

- a) $\lim_{n \rightarrow \infty} \left(1 + \frac{7}{6n}\right)^{6n} = \lim_{n \rightarrow \infty} \left[\left(1 + \frac{7}{6n}\right)^{\frac{6n}{7}}\right]^7 = \left[\lim_{n \rightarrow \infty} \left(1 + \frac{7}{6n}\right)^{\frac{6n}{7}}\right]^7 = e^7$, da nach den Grenzwertsätzen für eine konvergente Folge $(a_n)_{n=1}^{\infty}$ gilt:

$$\lim_{n \rightarrow \infty} a_n^7 = \lim_{n \rightarrow \infty} (a_n \cdot a_n \cdots a_n) = \left(\lim_{n \rightarrow \infty} a_n\right) \cdots \left(\lim_{n \rightarrow \infty} a_n\right) = \left(\lim_{n \rightarrow \infty} a_n\right)^7.$$

b)

$$\lim_{n \rightarrow \infty} \left(1 + \frac{7}{n}\right)^{-n} = \lim_{n \rightarrow \infty} \frac{1}{\left[\left(1 + \frac{7}{n}\right)^{\frac{n}{7}}\right]^7} = \frac{1}{\left[\lim_{n \rightarrow \infty} \left(1 + \frac{7}{n}\right)^{\frac{n}{7}}\right]^7} = \frac{1}{e^7}.$$

■

2.1.6 Approximation reeller Potenzen

Wir müssen noch eine weitere Lücke schließen. Allgemeine reelle Potenzen x^y haben wir eingeführt, indem wir y durch Brüche p/q angenähert und dann $x^{p/q}$ berechnet haben. Das können wir jetzt exakt ausdrücken.

Satz 2.4 (Approximation reeller Potenzen)

Sei $x > 0$ und $(a_n)_{n=1}^\infty$ eine Folge mit Folgengliedern $a_n \in \mathbb{Q}$, die gegen einen Grenzwert in $\mathbb{R}$ konvergiert. Dann konvergiert auch die Folge $(x^{a_n})_{n=1}^\infty$. Ist $(b_n)_{n=1}^\infty$ eine weitere Folge mit Folgengliedern $b_n \in \mathbb{Q}$, die den gleichen Grenzwert wie $(a_n)_{n=1}^\infty$ hat, also $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n$, dann ist auch $\lim_{n \rightarrow \infty} x^{a_n} = \lim_{n \rightarrow \infty} x^{b_n}$.

Den Beweis findet man z. B. in (Endl und Luh, 1989, S. 142). Er basiert darauf, dass man für jede Nullfolge rationaler Zahlen $(c_n)_{n=1}^\infty$ zeigen kann, dass $\lim_{n \rightarrow \infty} x^{c_n} = 1$ ist. Damit lässt sich sowohl die Konvergenz der Folge $(x^{a_n})_{n=1}^\infty$ (als Cauchy-Folge, vgl. Kapitel 2.1.9 unten) beweisen als auch im Fall $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n$ zeigen, dass

$$\lim_{n \rightarrow \infty} (x^{a_n} - x^{b_n}) = \lim_{n \rightarrow \infty} \underbrace{x^{a_n}}_{\text{beschränkt}} (1 - x^{b_n - a_n}) = 0.$$

Damit kann man tatsächlich für $x > 0$ definieren:

$$x^y := \lim_{n \rightarrow \infty} x^{a_n}, \quad (2.9)$$

wobei $(a_n)_{n=1}^\infty$ eine beliebige Folge rationaler Zahlen ist, die gegen y konvergiert.

Der Satz 2.4 bestätigt, dass der Grenzwert unabhängig von der Wahl der Folge $(a_n)_{n=1}^\infty$ ist. Da die rationalen Zahlen dicht in den reellen Zahlen liegen, findet man auch zu jedem $y \in \mathbb{R}$ eine entsprechende Folge $(a_n)_{n=1}^\infty$ aus $\mathbb{Q}$ mit $\lim_{n \rightarrow \infty} a_n = y$. Damit ist x^y wohldefiniert.

Man kann nun mittels Grenzwertaussagen nachrechnen, dass für reelle Exponenten tatsächlich die Rechenregeln für Potenzen gelten.

2.1.7 Bestimmte Divergenz

Neben den reellen Zahlen macht es Sinn, auch die Grenzwerte $\pm\infty$ zuzulassen. Folgen, die gegen diese Werte streben, nennt man bestimmt divergent. Wie aber definiert man das Streben gegen ∞ ? Man kann schließlich keinen ε -Streifen um ∞ legen, $\infty - \varepsilon$ kann nur ∞ sein. Stattdessen kann man aber einen Streifen $]M, \infty[$ betrachten, wobei das M beliebig groß werden darf:

Definition 2.6 (Bestimmte Divergenz)

Eine Folge $(a_n)_{n=1}^\infty$ heißt **bestimmt divergent** gegen ∞ genau dann, falls für jedes noch so große (positive) $M \in \mathbb{R}$ eine Stelle $n_0 = n_0(M) \in \mathbb{N}$ existiert, so dass für alle $n > n_0$, also für alle Folgenglieder nach dieser Stelle, gilt: $a_n > M$ (siehe Abbildung 2.7). Entsprechend heißt sie **bestimmt divergent** gegen $-\infty$, falls für jedes noch so

kleine (negative) $m \in \mathbb{R}$ eine Stelle $n_0 = n_0(m) \in \mathbb{N}$ existiert, so dass für alle $n > n_0$ gilt: $a_n < m$.

Achtung: Eine divergente Folge strebt nicht automatisch gegen $\pm\infty$. Hier haben wir bereits die beschränkte Folge mit abwechselnden Gliedern $+1$ und -1 betrachtet. Diese ist nicht bestimmt divergent. Aber auch die unbeschränkte Folge $((-1)^n n)_{n=1}^\infty$ ist nicht bestimmt divergent. Bestimmte Divergenz ist also mehr als nur Unbeschränktheit.

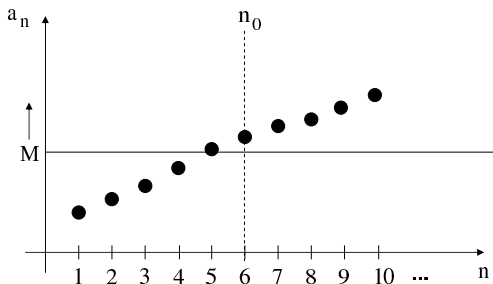


Abb. 2.7: $\lim_{n \rightarrow \infty} a_n = \infty$

Beispiel 2.15

$\lim_{n \rightarrow \infty} n^2 = \infty$, $\lim_{n \rightarrow \infty} 2^n = \infty$, $\lim_{n \rightarrow \infty} -n = -\infty$. ■

Solange man keine undefinierten Operationen wie $\infty - \infty$ durchführt, gelten viele Aussagen für reelle Grenzwerte auch für die Grenzwerte $\pm\infty$. Hat man beispielsweise zwei Folgen mit $\lim_{n \rightarrow \infty} a_n = a \in \mathbb{R}$ und $\lim_{n \rightarrow \infty} b_n = \infty$, so gilt $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$. Denn während sich der Zähler immer mehr der Zahl a annähert, wird der Nenner immer größer, der Betrag des Quotienten wird damit immer kleiner.

Beispiel 2.16

Bildet man die Folgenglieder über eine gebrochen-rationale Funktion, so haben wir bereits gesehen, dass sich der Grenzwert im Falle der Konvergenz an den Faktoren zu den größten Exponenten im Zähler und Nenner ablesen lässt. Dies gilt auch im Fall der bestimmten Divergenz. Wir erweitern wieder mit $1/n^m$, wobei m der größte Exponent des Nenners ist, z. B. $m = 2$:

$$\lim_{n \rightarrow \infty} \frac{4n^3 + 2n^2 + 1}{-2n^2 + 5} = \lim_{n \rightarrow \infty} \frac{4n + \overbrace{2 + n^{-2}}^{\rightarrow 2}}{\underbrace{-2 + 5n^{-2}}_{\rightarrow -2}} = -\infty.$$

Wir können damit (2.2) so komplettieren:

$$\lim_{n \rightarrow \infty} \frac{a_k n^k + a_{k-1} n^{k-1} + \cdots + a_1 n + a_0}{b_m n^m + b_{m-1} n^{m-1} + \cdots + b_1 n + b_0} = \begin{cases} \frac{a_k}{b_m} & \text{falls } k = m, \\ 0 & \text{falls } k < m, \\ +\infty & \text{falls } k > m, \frac{a_k}{b_m} > 0, \\ -\infty & \text{falls } k > m, \frac{a_k}{b_m} < 0. \end{cases} \quad (2.10) \quad \blacksquare$$

Beispiel 2.17

Nicht erst seit dem Buch „Sakrileg“ von Dan Brown ist die bestimmt divergente Folge der **Fibonacci-Zahlen** berühmt. Wir gehen davon aus, dass ein Kaninchenpaar nach seiner Geburt eine Zeiteinheit benötigt, um selbst fortpflanzungsfähig zu werden. Zur Vereinfachung sei die Austragungszeit ebenfalls eine Zeiteinheit, und die Kaninchen mögen ewig leben und sich vermehren. Weiterhin bringen sie immer Zwillingspärchen zur Welt. Die Anzahl der Kaninchenpaare nach der n -ten Zeiteinheit sei a_n . Zu Beginn $n = 0$ haben wir noch kein Kaninchenpaar. Das wächst gerade im Klon-Labor heran, also $a_0 = 0$. Nach einer Zeiteinheit ($n = 1$) wird das erste Pärchen geboren, also $a_1 = 1$. Nach zwei Zeiteinheiten ($n = 2$) ist das Pärchen geschlechtsreif, aber hat noch keine Jungen, also $a_2 = 1$. Bei $n = 3$ kommen die ersten Jungen zur Welt. Es gibt $a_3 = a_2 + a_1 = 1 + 1 = 2$ Pärchen. Nach vier Zeiteinheiten ($n = 4$) kommen die zweiten Jungen zur Welt, und die ersten Jungen sind fortpflanzungsfähig, also $a_4 = a_3 + a_2 = 2 + 1 = 3$ Pärchen. Bei $n = 5$ haben wir dann $a_5 = a_4 + a_3 = 3 + 2 = 5$ Pärchen usw. In jeder Zeiteinheit kommt immer die Anzahl der Kaninchen dazu, die vor zwei Zeiteinheiten schon gelebt hat. Wir erhalten das Bildungsgesetz der Fibonacci-Zahlen:

$$a_0 := 0, \quad a_1 := 1, \quad a_n := a_{n-1} + a_{n-2} \text{ für } n \geq 2.$$

Die Folgenglieder sind also 0, 1, 1, 2, 3, 5, 8, 13, 21, $\dots$, und die Folge ist bestimmt divergent gegen $+\infty$. Bildet man aber den Quotienten zweier benachbarter Folgenglieder, so erhält man eine konvergente Folge $\left(\frac{a_n}{a_{n-1}}\right)_{n=2}^{\infty}$, die gegen einen Grenzwert $\Phi \geq 1$ strebt, d. h. $\Phi = \lim_{n \rightarrow \infty} \frac{a_n}{a_{n-1}}$. Das werden wir später als Anwendung der Matrixdiagonalisierung zeigen (siehe Seite 551). Damit haben wir auch $\lim_{n \rightarrow \infty} \frac{a_{n-1}}{a_n} = \frac{1}{\Phi}$ und $\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \Phi$. Berücksichtigen wir noch die rekursive Definition der Fibonacci-Zahlen, gewinnen wir eine Bestimmungsgleichung für den Grenzwert:

$$\Phi = \lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \lim_{n \rightarrow \infty} \frac{a_n + a_{n-1}}{a_n} = 1 + \lim_{n \rightarrow \infty} \frac{a_{n-1}}{a_n} = 1 + \frac{1}{\Phi}.$$

Demnach muss Φ die Gleichung $\Phi^2 - \Phi - 1 = 0$ erfüllen, also $\Phi = \frac{1}{2} \pm \sqrt{\frac{1}{4} + 1}$. Da alle Folgenglieder positiv sind, kann Φ nicht-negativ sein. Damit:

$$\Phi = \lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \frac{1 + \sqrt{5}}{2}.$$

Diese Zahl haben wir bereits als goldenen Schnitt (siehe Seite 56) kennengelernt. $\blacksquare$

2.1.8 Häufungspunkte einer Folge*

Neben den Grenzwerten spielen ab und zu Folgehäufungspunkte eine Rolle. Während bei einem Grenzwert a in jedem Streifen um a alle bis auf endlich viele Folgenglieder liegen müssen, definiert man:

Definition 2.7 (Folgehäufungspunkt)

*Eine Folge besitzt einen **Folgehäufungspunkt** a genau dann, wenn in jedem Streifen $[a - \varepsilon, a + \varepsilon]$, $\varepsilon > 0$, unendlich viele Folgenglieder liegen.*

Unendlich viele Folgenglieder können deutlich weniger sein als „alle bis auf endlich viele“. Die Folge $(b_n)_{n=1}^\infty := (1, 0, 1, 0, 1, \dots)$ hat demnach die beiden Folgehäufungspunkte 0 und 1 (vgl. Abbildung 2.5). Ein Grenzwert ist insbesondere ein Folgehäufungspunkt. Gibt es aber mehrere Folgehäufungspunkte, kann es keinen Grenzwert geben.

Satz 2.5 (Konvergente Teilfolgen)

Zu jedem Folgehäufungspunkt findet man eine Teilfolge der Ausgangsfolge, die gegen diesen konvergiert.

Beweis: Wir wählen zu jedem $k \in \mathbb{N}$ einen Streifen mit dem Radius $\varepsilon := \frac{1}{k}$ um den Folgehäufungspunkt. In jedem dieser Streifen liegen unendlich viele Glieder der Folge. Damit lässt sich ein Folgenglied a_{n_1} auswählen, das im Streifen zu $\varepsilon = 1$ liegt. Als Nächstes wählen wir ein Folgenglied a_{n_2} mit $n_2 > n_1$, das im Streifen zu $\varepsilon = 1/2$ liegt usw. a_{n_k} liegt dann im Streifen zu $\varepsilon = \frac{1}{k}$ und $n_k > n_{k-1}$, wobei man für die letzte Bedingung braucht, dass auch ohne die bislang ausgewählten Folgenglieder immer noch weitere im Streifen liegen. Das ist aber der Fall, wenn es dort unendlich viele Folgenglieder gibt. Alle weiteren Glieder der so konstruierten Teilfolge liegen ebenfalls im Streifen zu $\varepsilon = \frac{1}{k}$, da sie sogar in noch kleineren Streifen enthalten sind.

So entsteht eine Teilfolge $(a_{n_k})_{k=1}^\infty$, die gegen den Folgehäufungspunkt konvergiert, er ist Grenzwert dieser Teilfolge. Denn zu jedem $\varepsilon > 0$ gibt es ein $k_0 \in \mathbb{N}$ mit $\frac{1}{k_0} < \varepsilon$, so dass alle a_{n_k} mit $k \geq k_0$ in einem ε -Streifen um den Folgehäufungspunkt liegen.

□

2.1.9 Folgenkompaktheit und Cauchy-Folgen*

In diesem Kapitel beschäftigen wir uns intensiver mit der Vollständigkeit der reellen Zahlen, die daraus erwächst, dass man neben den endlichen und periodischen Dezimalbrüchen nun auch nicht-periodische Dezimalbrüche zulässt. Für den mathematischen Aufbau der Analysis und ihre Verallgemeinerung zur Funktionalanalysis sind die Er-

gebnisse sehr wichtig, und wir werden sie auch an späterer Stelle noch in Beweisen benötigen. Aus Anwendersicht ist vor Allem der folgende Satz über die Intervallschachtelung noch interessant:

Satz 2.6 (Intervallschachtelung)

Seien $(a_n)_{n=1}^{\infty}$ und $(b_n)_{n=1}^{\infty}$ Folgen mit $a_n \leq b_n$, so dass für die Intervalle $[a_n, b_n]$ gilt: $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$ für alle $n \in \mathbb{N}$. Falls $\lim_{n \rightarrow \infty} b_n - a_n = 0$, der Durchmesser der Intervalle also gegen null geht, dann existiert genau ein Punkt $x_0 \in \mathbb{R}$, so dass $x_0 \in [a_n, b_n]$ für alle $n \in \mathbb{N}$. Insbesondere gilt:

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = x_0.$$

Bei einer Intervallschachtelung liegen die Intervalle also ineinander wie die russischen Matroschkas, bemalte ineinanderschachtelbare, eiförmige Puppen.

Beweis: Die Folge $(a_n)_{n=1}^{\infty}$ ist monoton wachsend und nach oben durch b_1 beschränkt. Die Folge $(b_n)_{n=1}^{\infty}$ ist monoton fallend und nach unten durch a_1 beschränkt. Damit sind nach Satz 2.3 a), b) beide Folgen konvergent: $\lim_{n \rightarrow \infty} a_n =: a$, $\lim_{n \rightarrow \infty} b_n =: b$ und $a_n \leq a$, $b \leq b_n$, $n \in \mathbb{N}$. Wegen $\lim_{n \rightarrow \infty} b_n - a_n = 0$ folgt $a = b$. Damit ist $x_0 := a = b$ in jedem Intervall enthalten. Ein zweiter Punkt $x_1 \neq x_0$ kann nicht in allen Intervallen enthalten sein. Denn da $|x_0 - x_1| > 0$, gibt es ein $n_0 \in \mathbb{N}$ mit $b_{n_0} - a_{n_0} < |x_0 - x_1|$. Da $x_0 \in [a_{n_0}, b_{n_0}]$, kann x_1 nicht in diesem Intervall liegen. $\square$

Die Intervallschachtelung wird häufig bei der näherungsweisen Berechnung von Problemen mit dem Computer benutzt. Man tastet sich sukzessive an eine exakte Lösung heran, indem man diese mit immer kleineren Intervallen einschachtelt. Ein Beispiel ist die Nullstellensuche für Funktionen mit dem Zwischenwertsatz, die wir später auf Seite 278 behandeln.

Beispiel 2.18

Wir konstruieren eine Intervallschachtelung für $x_0 = \sqrt{2}$, indem wir bereits in der Nähe von x_0 starten: $\sqrt{2} \in [a_1, b_1] := [1,4; 1,5]$, da $(1,4)^2 = 1,96 < 2 < (1,5)^2 = 2,25$.

Eine neue Intervallgrenze erhalten wir durch Halbierung des aktuellen Intervalls: 1,45. Die neue Grenze muss wegen $(1,45)^2 = 2,1025$ die rechte Grenze des neuen Intervalls $[a_2, b_2]$ sein: $a_2 = 1,4 < \sqrt{2} < 1,45 = b_2$, weil $(1,4)^2 = 1,96 < 2 < (1,45)^2 = 2,1025$. Durch fortgesetzte Halbierung erhalten wir immer kleinere Intervalle, in denen $\sqrt{2}$ liegt:

$$\begin{array}{ccccccc} 1,4 & < & \sqrt{2} & < & 1,425 \\ 1,4125 & < & \sqrt{2} & < & 1,425 \end{array}$$

$$\begin{array}{rclclcl}
1,4125 & < & \sqrt{2} & < & 1,41875 \\
1,4125 & < & \sqrt{2} & < & 1,415625 \\
1,4140625 & < & \sqrt{2} & < & 1,415625 \text{ usw.}
\end{array}$$

Mit diesem Algorithmus kann $\sqrt{2}$ durch rationale Zahlen beliebig genau angenähert werden. ■

Man beachte, dass wir bei der Intervallschachtelung über Satz 2.3 wieder die Vollständigkeit der reellen Zahlen ausgenutzt haben. Die Intervallschachtelung ist eine weitere anschauliche Deutung dieser Eigenschaft, mit der auch der folgende Satz bewiesen wird:

Satz 2.7 (Bolzano-Weierstraß, Folgenkompaktheit)

Sei $(a_n)_{n=1}^{\infty}$ eine beschränkte Folge. Dann existiert ein Folgenhäufungspunkt, d. h. eine konvergente Teilfolge $(a_{n_k})_{k=1}^{\infty}$.

Die Aussage ist zunächst verblüffend. Wenn alle Folgenglieder in einem Intervall $[m, M]$ liegen, so ballen sie sich an mindestens einer Stelle $a \in [m, M]$ – und das, obwohl es überabzählbar viele Zahlen in $[m, M]$ gibt.

Beweis: (Skizze) Nach Voraussetzung existieren $m, M \in \mathbb{R}$ mit $m \leq a_n \leq M$ für alle $n \in \mathbb{N}$. Wir unterteilen nun das Intervall $[m, M]$ sukzessive durch Halbierung in Teilintervalle und machen stets mit einem Teilintervall weiter, in dem unendlich viele Folgenglieder liegen. So entsteht eine Intervallschachtelung, die Satz 2.6 erfüllt. Die Teilfolge, die entsteht, wenn man bei jeder Intervallteilung das nächste Folgenglied auswählt, das im ausgewählten Teilintervall liegt, konvergiert gegen den Punkt x_0 , der in allen ausgewählten Teilintervallen liegt. □

Eine Anwendung dieses Satzes ist die

Folgerung 2.1 (Grenzwert und Folgenhäufungspunkt)

Sei $(a_n)_{n=1}^{\infty}$ eine beschränkte Folge. Dann sind äquivalent:

- a) *Die Folge ist konvergent.*
- b) *Die Folge besitzt genau einen Folgenhäufungspunkt.*

Beweis: Aus a) folgt b), da der Grenzwert der einzige Folgenhäufungspunkt ist. Aus b) folgt a): Der eindeutige Folgenhäufungspunkt a ist Kandidat für den Grenzwert. Sei $\varepsilon > 0$ beliebig. In dem Streifen $]a - \varepsilon, a + \varepsilon[$ liegen unendlich viele Folgenglieder – aber sind es auch alle bis auf endlich viele? Falls außerhalb des Streifens unendlich viele Glieder lägen, dann gäbe es unendlich viele Glieder im $[a + \varepsilon, M]$ oder in $[m, a -$

$\varepsilon]$ oder in beiden Intervallen, wobei M eine obere und m eine untere Schranke der nach Voraussetzung beschränkten Folge ist. Nach dem Satz von Bolzano-Weierstraß (Satz 2.7) gibt es dann aber einen Folgenhäufungspunkt in $[a + \varepsilon, M]$ oder $[m, a - \varepsilon]$, der damit von a verschieden ist. Dies ist ein Widerspruch dazu, dass a alleiniger Folgenhäufungspunkt ist. In $]a - \varepsilon, a + \varepsilon[$ liegen alle bis auf endlich viele Folgenglieder. Da das für jedes $\varepsilon > 0$ gilt, ist a Grenzwert. $\square$

Definition 2.8 (Cauchy-Folge)

Eine Folge $(a_n)_{n=1}^{\infty}$ heißt genau dann **Cauchy-Folge**, wenn zu jedem (noch so kleinen) $\varepsilon > 0$ ein $n_0 = n_0(\varepsilon) \in \mathbb{N}$ existiert, so dass für alle $n, m > n_0$ gilt:

$$|a_n - a_m| < \varepsilon.$$

Im Gegensatz zur Grenzwert-Definition werden hier die Folgenglieder nicht mit einem Grenzwert, sondern mit anderen Folgengliedern verglichen. Bei einer Cauchy-Folge gibt es zu jeder noch so kleinen Zahl $\varepsilon > 0$ eine Stelle, so dass ab dieser Stelle alle Folgenglieder (und nicht nur benachbarte) näher als ε zusammenliegen.

Vor Augustin Louis Cauchy (1789–1857) fehlte die klare Verwendung eines Grenzwertbegriffs in der Analysis. Er präziserte die vorhandene Theorie und entwickelte sie weiter. Cauchy hat damit die Analysis so umgebaut, wie wir sie heute kennen.

Die reellen Zahlen wurden genau so konstruiert, dass der folgende Satz gilt:

Satz 2.8 (Cauchy-Kriterium)

Eine Folge konvergiert genau dann, wenn sie eine Cauchy-Folge ist.

Wir haben damit in diesem Kontext keinen neuen, eigenständigen Begriff eingeführt, sondern lediglich eine alternative Beschreibung der Konvergenz. Damit dies so ist, benötigt man aber erneut die Vollständigkeit der reellen Zahlen, die in diesen Beweis über den Satz von Bolzano-Weierstraß einfließt:

Beweis: a) Wir zeigen, dass aus der Konvergenz die Cauchy-Bedingung folgt: Sei also die Folge $(a_n)_{n=1}^{\infty}$ konvergent gegen den Grenzwert a . Außerdem sei $\varepsilon > 0$ beliebig vorgegeben. Aufgrund der Konvergenz existiert insbesondere zu $\varepsilon/2$ ein $n_0 \in \mathbb{N}$, so dass für alle $n > n_0$ gilt: $|a_n - a| < \varepsilon/2$. Damit gilt für $n, m > n_0$ die Cauchy-Bedingung:

$$|a_n - a_m| = |a_n - a + a - a_m| \leq |a_n - a| + |a - a_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Dies ist wieder eine typische Situation, in der in der Analysis die Dreiecksungleichung angewendet wird (siehe Seite 68).

- b) Zu zeigen ist, dass aus der Cauchy-Bedingung die Konvergenz folgt. Für ein beliebiges $\varepsilon > 0$ liefert die Cauchy-Bedingung die Existenz eines $n_0 \in \mathbb{N}$, so dass $|a_n - a_{n_0+1}| < \varepsilon$ für alle $n > n_0$. Damit ist die Folge $(a_n)_{n=1}^\infty$ beschränkt:

$$|a_n| < \max\{\max\{|a_k| : 1 \leq k \leq n_0\}, |a_{n_0+1}| + \varepsilon\}.$$

Nach Satz 2.7 (Bolzano-Weierstraß) existiert damit eine konvergente Teilfolge $(a_{n_i})_{i=1}^\infty$, $\lim_{i \rightarrow \infty} a_{n_i} = a$ und $n_i \geq i$. Jetzt kann mittels der Cauchy-Bedingung gezeigt werden, dass die vollständige Ausgangsfolge $(a_n)_{n=1}^\infty$ gegen a konvergiert. Dazu beginnen wir erneut mit der Cauchy-Bedingung für ein $\varepsilon > 0$. Zu ε existiert wieder ein $n_0 = n_0(\varepsilon) \in \mathbb{N}$, so dass jetzt $|a_k - a_m| < \varepsilon$ für alle $k, m > n_0$ ist. Da die Teilfolge konvergiert, existiert außerdem ein $N_0 = N_0(\varepsilon) \in \mathbb{N}$, so dass $|a_{n_i} - a| < \varepsilon$ für alle $i > N_0$ ist. Mit $N_1 := \max\{n_0, N_0\}$ gilt für $k > N_1$:

$$|a_k - a| = |a_k - a_{n_k} + a_{n_k} - a| \leq |a_k - a_{n_k}| + |a_{n_k} - a| \leq 2\varepsilon.$$

Dabei haben wir für die Abschätzung von $|a_k - a_{n_k}|$ ausgenutzt, dass sowohl k als auch $n_k \geq k$ größer als n_0 sind.

□

2.2 Zahlen-Reihen

Hat man eine Folge $(a_k)_{k=1}^\infty$, so erhält man daraus eine neue Folge $(s_n)_{n=1}^\infty$, indem man jeweils die ersten n Glieder $a_1, a_2, \dots, a_n$ der Ausgangsfolge aufsummiert:

$$s_n = a_1 + a_2 + \dots + a_n = \sum_{k=1}^n a_k.$$

Man nennt $(s_n)_{n=1}^\infty$ eine **Reihe** (siehe Definition 2.9). Sind die Folgenglieder a_k alle positiv, so können wir sie uns als Höhen von Bauklötzchen vorstellen. Eine Frage, die wir untersuchen werden, ist: Hat der Turm, der durch Übereinanderlegen der unendlich vielen Klötzchen entsteht, eine endliche Höhe?

Eigentlich ist der Begriff der Reihe überflüssig, da jede Reihe eine Folge $(s_n)_{n=1}^\infty$ ist, aber umgekehrt auch jede Folge $(s_n)_{n=1}^\infty$ als Reihe zur Ausgangsfolge $a_1 := s_1$, $a_k := s_k - s_{k-1}$ für $k > 1$, aufgefasst werden kann. Als Teleskopsumme (siehe Seite 59) ist nämlich

$$\sum_{k=1}^n a_k = s_1 + \sum_{k=2}^n (-s_{k-1} + s_k) = s_1 - s_1 + s_2 - s_2 + \dots + s_{n-1} - s_{n-1} + s_n = s_n.$$

Da viele praktisch wichtigen Folgen über eine Summation aufgebaut sind, ist es aber dennoch sinnvoll, sich mit dem Begriff der Reihe auseinanderzusetzen. Er wird uns helfen, Funktionen durch wesentlich einfachere Funktionen anzunähern (z. B. über Potenz- oder Fourier-Reihen).

2.2.1 Definition und Konvergenz einer Reihe

Beispiel 2.19

Die Dezimalbruchdarstellung der reellen Zahlen basiert auf der Konvergenz einer unendlichen Summe:

$$3,12345 \cdots = 3 + \frac{1}{10} + \frac{2}{100} + \frac{3}{1\,000} + \frac{4}{10\,000} + \frac{5}{100\,000} + \cdots$$

Wir werden als Anwendung des Vergleichskriteriums (siehe Satz 2.15) sehen, dass die Folge $(3; 3,1; 3,12; 3,123; 3,1234; 3,12345; \dots)$ und die entsprechende Folge für jede andere Dezimalzahl konvergiert. Hier betrachten wir jetzt den wichtigen Spezialfall periodischer Dezimalbrüche. Wie wir mit (1.3) auf Seite 38 gesehen haben, handelt es sich dabei um rationale Zahlen.

Die reelle Zahl $3, \overline{14}$ ist beispielsweise erklärt als

$$\begin{aligned} 3 + \frac{14}{100} + \frac{14}{10\,000} + \frac{14}{1\,000\,000} + \cdots &= 3 + \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{14}{100^k} \\ &= 3 + 14 \left(-1 + \lim_{n \rightarrow \infty} \sum_{k=0}^n \left(\frac{1}{100} \right)^k \right). \end{aligned}$$

Wir summieren also über alle Glieder der Folge $(3, \frac{14}{100}, \frac{14}{10\,000}, \frac{14}{1\,000\,000}, \dots)$. Für $q \in \mathbb{R} \setminus \{0, 1\}$ und für alle $n \in \mathbb{N}_0$ haben wir auf Seite 60 die geometrische Summenformel (1.10) bewiesen: $\sum_{k=0}^n q^k = \frac{1-q^{n+1}}{1-q}$. Also gilt insbesondere für $q = \frac{1}{100}$:

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \left(\frac{1}{100} \right)^k = \lim_{n \rightarrow \infty} \frac{1 - \left(\frac{1}{100} \right)^{n+1}}{1 - \frac{1}{100}} = \frac{100}{99}.$$

Damit:

$$3, \overline{14} = 3 + 14 \left(-1 + \frac{100}{99} \right) = 3 + \frac{14}{99}.$$

Überraschend ist das nicht, schließlich haben wir diese Umrechnung bereits auf Seite 38 hergeleitet. Nun aber haben wir eine unendliche Summe dafür eingesetzt. ■

Wir haben in diesem Beispiel als Erweiterung der geometrischen Summe die **geometrische Reihe** $\lim_{n \rightarrow \infty} \sum_{k=0}^n q^k$ kennengelernt, die über die geometrische Folge $(q^k)_{k=0}^\infty$ gebildet wird. Für $q \neq 0$, $|q| < 1$ gilt:

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n q^k = \lim_{n \rightarrow \infty} \frac{1 - q^{n+1}}{1 - q} = \frac{1}{1 - q}. \quad (2.11)$$

Häufig beginnt man die Summe nicht mit $k = 0$, sondern mit $k = 1$ (siehe Abbildung 2.8):

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n q^k = -1 + \lim_{n \rightarrow \infty} \sum_{k=0}^n q^k = -1 + \frac{1}{1 - q} = \frac{q}{1 - q}. \quad (2.12)$$

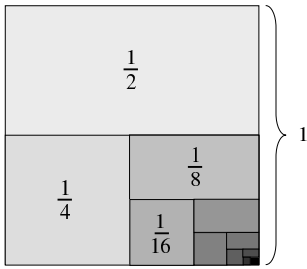


Abb. 2.8: Die geometrische Reihe $\lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{1}{2^k} = 1$ für $q = \frac{1}{2}$ ab dem Index $k = 1$

Beispiel 2.20 (Paradoxon von Zeno)

Achilles bestreitet einen Wettlauf gegen eine Schildkröte. Achilles legt 10 m/s, die Schildkröte 1 m/s zurück und bekommt einen Vorsprung von 10 m. Nach wie vielen Sekunden t holt Achilles die Schildkröte ein?

$$10 \cdot t = 10 + 1 \cdot t \implies t = \frac{10}{9}.$$

Aber andererseits: Wenn Achilles die Stelle erreicht hat, an der die Schildkröte beim Start war, dann ist sie schon ein Stück weiter. Wenn er dann an dieser neuen Stelle angekommen ist, dann ist sie wieder weiter usw. Er scheint sie also nie zu erreichen?

- Nach 1 s hat der Läufer den Startpunkt der Schildkröte bei 10 m erreicht. Diese ist dann bei 11 m.
- Nach weiteren $\frac{1}{10}$ s ist der Läufer dann dort, die Schildkröte ist bei 11,1 m.
- Nach weiteren $\frac{1}{100}$ s ist der Läufer dann dort usw.

Es vergehen also Zeiten von $\frac{1}{10^0}, \frac{1}{10^1}, \frac{1}{10^2}, \frac{1}{10^3}, \dots$ Sekunden. Addiert man diese Zeitspannen, erhält man wie oben eine konvergente geometrische Reihe für $q = 1/10$ (siehe (2.11)):

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{1}{10^k} = \frac{1}{1 - \frac{1}{10}} = \frac{10}{9}.$$

Der Läufer holt also sehr wohl die Schildkröte nach endlicher Zeit ein, da die Reihe konvergiert – und unsere erste Rechnung, bei der wir $\frac{10}{9}$ Sekunden erhalten haben, stimmt. ■

Definition 2.9 (Reihe)

Sei $(a_k)_{k=1}^\infty$ eine Folge reeller Zahlen. Die Folge $(s_n)_{n=1}^\infty$ mit

$$s_n := \sum_{k=1}^n a_k$$

heißt die **Folge der Partialsummen** (Teilsummen) von $(a_k)_{k=1}^\infty$ oder auch die (unendliche) **Reihe** der Zahlen a_k , $k \in \mathbb{N}$, also

$$s_1 := a_1$$

$$\begin{aligned}
 s_2 &:= a_1 + a_2 \\
 s_3 &:= a_1 + a_2 + a_3 \\
 s_n &:= \sum_{k=1}^n a_k = a_1 + a_2 + \cdots + a_n.
 \end{aligned}$$

Diese Reihe wird mit dem Symbol $\sum_{k=1}^{\infty} a_k$ bezeichnet.

- Die Reihe $\sum_{k=1}^{\infty} a_k$ heißt **konvergent** gegen ein $S \in \mathbb{R}$ genau dann, wenn

$$\lim_{n \rightarrow \infty} s_n = S.$$

S heißt die **Summe** (oder der **Grenzwert**) der Reihe. In diesem Fall wird S auch mit $\sum_{k=1}^{\infty} a_k$ bezeichnet. Mit der Schreibweise $\sum_{k=1}^{\infty} a_k < \infty$ wird ausgedrückt, dass die Reihe konvergiert.

- Die Reihe heißt genau dann **divergent**, wenn die Folge $(s_n)_{n=1}^{\infty}$ der Partialsummen divergiert.
- Die Reihe $\sum_{k=1}^n a_k$ heißt genau dann **bestimmt divergent**, wenn die Folge $(s_n)_{n=1}^{\infty}$ der Partialsummen bestimmt divergent ist (d. h. gegen ∞ oder $-\infty$ „konvergiert“). Dies drückt die Schreibweise $\sum_{k=1}^{\infty} a_k = \infty$ bzw. $\sum_{k=1}^{\infty} a_k = -\infty$ aus.
- Schließlich heißt die Reihe $\sum_{k=1}^n a_k$ **unbestimmt divergent** genau dann, wenn die Folge $(s_n)_{n=1}^{\infty}$ divergent, aber nicht bestimmt divergent ist.

Das Symbol $\sum_{k=1}^{\infty} a_k$ hat also zwei Bedeutungen. Zum einen bezeichnet es lediglich die Folge der Partialsummen. Falls die Reihe konvergiert, wird damit zum anderen auch der Grenzwert ausgedrückt. Aus dem Zusammenhang ist meistens klar, welche Bedeutung gemeint ist.

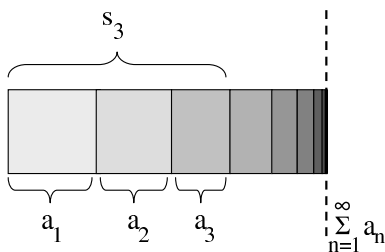


Abb. 2.9: Reihe der Zahlen $a_1, a_2, a_3, \dots$

Beispiel 2.21

- a) Einige Reihen, die so wichtig sind, dass sie einen Namen bekommen haben, sind
- die **geometrische Reihe** $\sum_{k=1}^{\infty} q^k$, die wir schon in den Eingangsbeispielen verwendet haben und die wegen (2.11) für $|q| < 1$ konvergent ist. In Abbildung 2.8 haben wir den Wert der Reihe für $q = \frac{1}{2}$ **geometrisch** konstruiert.

- die **harmonische Reihe** $\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$, die sich später als Anwendung des Cauchy-Kriteriums (Satz 2.10) als divergent erweisen wird. Der Name der Reihe leitet sich von Obertönen einer schwingenden Saite ab, bei denen die Wellenlängen $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$ so lang wie die Basiswellenlänge sind. Überlagert (summiert) man diese (stehenden) Wellen, so entsteht ein harmonischer Klang.
 - die **alternierende harmonische Reihe** $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k} = -1 + \frac{1}{2} - \frac{1}{3} \pm \dots$, von der wir mit dem Leibniz-Kriterium (Satz 2.19) sehen werden, dass sie konvergiert.
- b) $\sum_{k=1}^{\infty} (-1)^{k+1} = 1 - 1 + 1 - 1 \pm \dots$ ist eine unbestimmt divergente Reihe, da $s_n = 1$ für ungerades n und $s_n = 0$ für gerades n ist und damit $(s_n)_{n=1}^{\infty}$ eine unbestimmt divergente Folge ist. An diesem Beispiel sieht man, dass sich unendliche Summen anders verhalten als endliche. Klammert man links beginnend jeweils zwei Summanden, so erhält man nur noch die Werte der geraden Partialsummen und als Grenzwert 0.
- c) $\sum_{k=1}^{\infty} k = 1 + 2 + 3 + 4 + \dots = \infty$, da

$$\lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} \sum_{k=1}^n k = \lim_{n \rightarrow \infty} \frac{n(n+1)}{2} = \infty.$$

Dabei haben wir die Summenformel von Gauß benutzt, die wir auf Seite 59 mittels Vollständiger Induktion bewiesen haben.

- d) Die Reihe $\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$ ist konvergent mit Grenzwert 1, also $\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$, da man sie über eine Teleskopsumme ausrechnen kann (siehe (1.9) von Seite 59)

$$s_n = \sum_{k=1}^n \frac{1}{k(k+1)} = \sum_{k=1}^n \left(\frac{1}{k} - \frac{1}{k+1} \right) = 1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \dots - \frac{1}{n+1} = 1 - \frac{1}{n+1} \quad (2.13)$$

und damit $\lim_{n \rightarrow \infty} s_n = 1$. ■

2.2.2 Rechnen mit konvergenten Reihen

Die für die Partialsummen als Folgen geltenden Konvergenzregeln führen sofort zu

Satz 2.9 (Linearkombination konvergenter Reihen)

Seien $\sum_{k=1}^{\infty} a_k$ und $\sum_{k=1}^{\infty} b_k$ zwei konvergente Reihen mit $\sum_{k=1}^{\infty} a_k = A$ und $\sum_{k=1}^{\infty} b_k = B$ sowie $c_1, c_2 \in \mathbb{R}$. Dann konvergiert $\sum_{k=1}^{\infty} (c_1 a_k + c_2 b_k)$ gegen $c_1 A + c_2 B$, d. h.

$$\sum_{k=1}^{\infty} (c_1 a_k + c_2 b_k) = c_1 \sum_{k=1}^{\infty} a_k + c_2 \sum_{k=1}^{\infty} b_k.$$

Konvergente Reihen dürfen daher gliedweise addiert werden.

Beweis:

$$\begin{aligned} \sum_{k=1}^{\infty} (c_1 a_k + c_2 b_k) &= \lim_{n \rightarrow \infty} \sum_{k=1}^n (c_1 a_k + c_2 b_k) = \lim_{n \rightarrow \infty} \left[c_1 \sum_{k=1}^n a_k + c_2 \sum_{k=1}^n b_k \right] \\ &\stackrel{\text{Satz 2.1 c)i),ii)}}{=} c_1 \lim_{n \rightarrow \infty} \sum_{k=1}^n a_k + c_2 \lim_{n \rightarrow \infty} \sum_{k=1}^n b_k = c_1 \sum_{k=1}^{\infty} a_k + c_2 \sum_{k=1}^{\infty} b_k \quad \square \end{aligned}$$

In diesem Beweis haben wir lediglich die Definition des Symbols $\sum_{k=1}^{\infty}$ als Grenzwert der Partialsummen aufgelöst sowie die Rechenregeln für Folgen auf Partialsummen angewendet. Entsprechend kann man aus jedem Satz für Folgen einen entsprechenden Satz für Reihen ableiten.

Während konvergente Reihen gliedweise addiert werden dürfen, gilt die Umkehrung aber offensichtlich nicht, d. h., aus $\sum_{k=1}^{\infty} (a_k + b_k) < \infty$ folgt nicht, dass $\sum_{k=1}^{\infty} a_k$ und $\sum_{k=1}^{\infty} b_k$ konvergieren. Dazu betrachten wir ein Gegenbeispiel: $\sum_{k=1}^{\infty} 1$ ist bestimmt divergent gegen ∞ , und $\sum_{k=1}^{\infty} (-1)$ ist bestimmt divergent gegen $-\infty$, aber $\sum_{k=1}^{\infty} (1 - 1) = 0$ ist konvergent.

2.2.3 Alternativen zur Definition der Reihenkonvergenz

Wir beginnen mit einer äquivalenten Formulierung der Konvergenz:

Satz 2.10 (Cauchy-Kriterium für Reihen)

Eine Reihe $\sum_{k=1}^{\infty} a_k$ konvergiert genau dann, wenn die folgende **Cauchy-Bedingung** erfüllt ist: Zu jedem noch so kleinen $\varepsilon > 0$ existiert eine Stelle $n_0 = n_0(\varepsilon) \in \mathbb{N}$, so dass für alle $n, m \in \mathbb{N}$ mit $m \geq n > n_0$ gilt:

$$\left| \sum_{k=n}^m a_k \right| < \varepsilon.$$

Das entspricht der Anschauung: Die Konvergenz einer Reihe ist damit äquivalent, dass die Restsummen hinreichend klein werden, genauer: Beliebige Abschnitte der Summe müssen kleiner als jeder vorgegebene Wert $\varepsilon > 0$ sein, wenn sie nur weit genug hinten in der Summe liegen.

Beweis: Nach dem Cauchy-Kriterium für Folgen (Satz 2.8) ist die Konvergenz der Reihe äquivalent mit der Cauchy-Bedingung für die Partialsummen, d. h., zu jedem $\varepsilon > 0$ existiert ein $n_0 = n_0(\varepsilon) \in \mathbb{N}$, so dass für alle $m \geq n > n_0$ gilt:

$$|s_m - s_{n-1}| < \varepsilon.$$

Da $s_m - s_{n-1} = \sum_{k=n}^m a_k$, ist damit der Beweis erbracht. $\square$

Beispiel 2.22

Die harmonische Reihe $\sum_{k=1}^{\infty} \frac{1}{k}$ ist divergent, da die Cauchy-Bedingung nicht erfüllt ist: Wählen wir $m = 2n$, dann ist

$$\sum_{k=n}^m \frac{1}{k} = \sum_{k=n}^{2n} \frac{1}{k} > \underbrace{(n+1)}_{\text{Anzahl der Summanden}} \cdot \underbrace{\frac{1}{2n}}_{\text{kleinstes Glied}} > \frac{1}{2}$$

unabhängig von n . Das heißt, zu $\varepsilon := 1/4$ existiert kein $n_0 \in \mathbb{N}$, so dass für alle $n, m > n_0$ gilt: $|\sum_{k=n}^m \frac{1}{k}| < \varepsilon$. ■

Völlig klar ist, dass eine unendliche Summe nur dann einen endlichen Wert ergeben kann, wenn betragsmäßig immer weniger addiert wird:

Satz 2.11 (Notwendige Bedingung für die Konvergenz einer Reihe)

Ist eine Reihe $\sum_{k=1}^{\infty} a_k$ konvergent, dann ist $\lim_{k \rightarrow \infty} a_k = 0$.

Eine Reihe kann also nur konvergieren, wenn die Zahlen, die man aufsummiert, gegen null konvergieren.

Beweis: Die Aussage folgt direkt aus dem Cauchy-Kriterium, indem man dort $m = n$ wählt: Ist die Reihe konvergent, so existiert zu jedem $\varepsilon > 0$ eine Stelle $n_0 \in \mathbb{N}$, so dass für alle $n > n_0$ gilt: $|\sum_{k=n}^n a_k| = |a_n - 0| < \varepsilon$. Das ist aber gerade die Definition einer Nullfolge. □

Die geometrische Reihe $\sum_{k=0}^{\infty} q^k$ (siehe (2.11)) divergiert für $|q| \geq 1$, da $(q^k)_{k=1}^{\infty}$ keine Nullfolge ist.

Achtung: Die harmonische Reihe, die über die Nullfolge $(\frac{1}{k})_{k=1}^{\infty}$ gebildet ist, beweist, dass diese Bedingung nicht hinreichend ist. Es gibt also Nullfolgen, deren Reihen nicht konvergieren. Daher benötigt man geeignete Konvergenzkriterien, von denen wir einige im Folgenden diskutieren.

Satz 2.12 (Konvergenz bei nicht-negativen Summanden)

Ist $(a_k)_{k=1}^{\infty}$ eine nicht-negative Folge, d. h. $a_k \geq 0, k \in \mathbb{N}$. Dann sind äquivalent:

- a) Die Reihe $\sum_{k=1}^{\infty} a_k$ ist konvergent.
- b) Die Folge der Partialsummen $(s_n)_{n=1}^{\infty} = (\sum_{k=1}^n a_k)_{n=1}^{\infty}$ ist nach oben beschränkt, d. h., es existiert ein $M \in \mathbb{R}$, so dass $s_n \leq M$ für alle $n \in \mathbb{N}$.

Beweis: Aus a) folgt b), da konvergente Folgen beschränkt sind (siehe Satz 2.1 b)). Aus b) und der Voraussetzung $a_k \geq 0$ folgt a), da die Folge der Partialsummen monoton steigend und beschränkt – und damit nach Satz 2.3 konvergent – ist. □

Wir haben bereits über das Cauchy-Kriterium gezeigt, dass die harmonische Reihe $\sum_{k=1}^{\infty} \frac{1}{k}$ nicht konvergent ist. Da $\frac{1}{k} \geq 0$ ist, bleibt nach Satz 2.12 nur der Schluss, dass die Folge der Partialsummen nicht beschränkt ist. Wegen der Monotonie der Folge ist $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

2.2.4 Absolute Konvergenz

Reihen, die über eine Folge mit nicht-negativen Gliedern gebildet werden, verhalten sich besonders einfach, da es bei der Summation keine Auslöschungen durch wechselnde Vorzeichen gibt. Das macht man sich auch für Reihen über Folgen mit wechselndem Vorzeichen zunutze:

Definition 2.10 (Absolute Konvergenz)

Eine Reihe $\sum_{k=1}^{\infty} a_k$ heißt genau dann **absolut konvergent**, wenn die zugehörige Reihe $\sum_{k=1}^{\infty} |a_k|$ konvergiert.

Insbesondere ist jede konvergente Reihe mit nicht-negativen Gliedern absolut konvergent.

Satz 2.13 (Absolute Konvergenz bedeutet insbesondere Konvergenz)

Ist $\sum_{k=1}^{\infty} a_k$ eine absolut konvergente Reihe, so ist sie auch konvergent.

Beweis: Die absolute Konvergenz bedeutet, dass $\sum_{k=1}^{\infty} |a_k|$ konvergent ist, d. h., nach Cauchy-Kriterium existiert zu jedem $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$, so dass für alle $m \geq n > n_0$ gilt:

$$\sum_{k=n}^m |a_k| = \left| \sum_{k=n}^m |a_k| \right| < \varepsilon.$$

Damit ist aber auch die Cauchy-Bedingung

$$\left| \sum_{k=n}^m a_k \right| \leq \sum_{k=n}^m |a_k| < \varepsilon$$

für die Reihe $\sum_{k=1}^{\infty} a_k$ erfüllt, und die Reihe ist konvergent. $\square$

Die Umkehrung gilt nicht, die Reihe $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k}$ wird sich als konvergent erweisen (siehe Satz 2.19). Gehen wir aber zu den Beträgen über, erhalten wir die divergente harmonische Reihe, d. h., die Reihe ist nicht absolut konvergent. Positive und negative Summanden können sich geeignet auslöschen. Diesen Effekt verliert man beim Übergang zu Beträgen.

Absolute Konvergenz ist also ein strengerer Begriff als nur Konvergenz. Ist eine Reihe (sogar) absolut konvergent, so ist sie sehr umgänglich. Man kann z. B. ihre Glieder umsortieren, ohne dass sich der Grenzwert ändert. Es scheint etwas überraschend, dass sich ein Grenzwert durch Umsortieren ändern kann. Schließlich gilt das Kommutativgesetz für die Addition von endlich vielen Zahlen. Dieses gilt aber nicht mehr, wenn man unendlich viele Zahlen addiert. Das haben wir bereits bei der Reihe $\sum_{k=1}^{\infty} (-1)^{k+1}$ gesehen. Ein anderes Beispiel ist

Beispiel 2.23 (Umordnung einer nicht absolut konvergenten Reihe)

Die Reihe $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k}$ ist, wie wir später noch sehen werden, konvergent, aber nicht absolut konvergent. Es seien $a_k := -\frac{1}{2k-1}$ die Summanden mit ungradzahligem und $b_k := \frac{1}{2k}$ die mit gradzahligem Index. Als harmonische Reihe ist $\sum_{k=1}^{\infty} b_k = \frac{1}{2} \sum_{k=1}^{\infty} \frac{1}{k} = \infty$. Sei $M := 1 + \left| \lim_{n \rightarrow \infty} \sum_{k=1}^n (-1)^k \frac{1}{k} \right|$. Wenn wir nun die a_k und b_k zu einer neuen Reihe umsortieren, indem wir als Glieder zunächst so lange b_k fortlaufend auswählen, bis deren Summe größer als $M + 1$ ist, dann das nächste a_k wählen, so ist die nächste Partialsumme größer als M . Dann wählen wir so lange b_k , bis die Summe größer $M + 1$ ist und fügen dann das nächste a_k hinzu. Die Gesamtsumme ist nun immer noch größer als M . Da die Reihe der b_k gegen ∞ divergiert, kann man so unendlich fortfahren. Der Grenzwert der resultierenden Reihe, falls er existiert, ist größer oder gleich M . Somit hat die umsortierte Reihe ein anderes Grenzwertverhalten als die Ausgangsreihe, deren Grenzwert mindestens um den Wert 1 kleiner ist. ■

Solche erstaunlichen Phänomene gibt es mit absolut konvergenten Reihen nicht. Außerdem kann man zwei absolut konvergente Reihen in geeigneter Weise zu einer Produktreihe multiplizieren (**Cauchy-Produkt**), so dass der Grenzwert das Produkt der einzelnen Grenzwerte ist:

Satz 2.14 (Konvergenz des Cauchy-Produkts*)

Sind $\sum_{k=0}^{\infty} a_k$ und $\sum_{k=0}^{\infty} b_k$ absolut konvergente Reihen, dann ist auch die Produktreihe $\sum_{k=0}^{\infty} c_k$ mit

$$c_k := \sum_{l=0}^k a_l b_{k-l}$$

absolut konvergent. Ihr Grenzwert ist das Produkt der Grenzwerte der beiden Ausgangsreihen.

Das Cauchy-Produkt ist (mit Hinblick auf die später noch einzuführenden Potenzreihen) den Koeffizienten nachempfunden, die beim Produkt von Polynomen entstehen. Damit man die Summanden direkt den Koeffizienten der Polynome zuordnen kann, haben wir hier die Reihen mit dem Index 0 begonnen.

$$(a_0 + a_1x + a_2x^2) \cdot (b_0 + b_1x + b_2x^2)$$

$$\begin{aligned}
&= a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + R \\
&= \sum_{l=0}^0 a_l b_{0-l} + \left(\sum_{l=0}^1 a_l b_{1-l} \right) x + \left(\sum_{l=0}^2 a_l b_{2-l} \right) x^2 + R \\
&= c_0 + c_1 x + c_2 x^2 + R
\end{aligned}$$

mit $R = (a_1 b_2 + a_2 b_1)x^3 + a_2 b_2 x^4$. Man kann einzelne Koeffizienten des Produkts über die angegebenen Summen schnell ausrechnen, ohne das Produkt der Polynome vollständig auszurechnen.

Beispiel 2.24 (Cauchy-Produkt *)

Wir betrachten die beiden geometrischen Reihen $\sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^k = \frac{1}{1-\frac{1}{2}} = 2$ und

$$\sum_{k=0}^{\infty} \left(\frac{1}{4}\right)^k = \frac{1}{1-\frac{1}{4}} = \frac{4}{3}.$$

Mit $a_k = \left(\frac{1}{2}\right)^k$ und $b_k = \left(\frac{1}{4}\right)^k$ und dem ersten Summationsindex 0 statt 1 ergibt sich für die mit dem Cauchy-Produkt gebildete Produktreihe

$$\begin{aligned}
\sum_{n=0}^{\infty} \sum_{k=0}^n a_k b_{n-k} &= \sum_{n=0}^{\infty} \sum_{k=0}^n \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{4}\right)^{n-k} = \sum_{n=0}^{\infty} \sum_{k=0}^n \frac{1}{2^k \cdot 2^{2n-2k}} \\
&= \sum_{n=0}^{\infty} \frac{1}{2^{2n}} \sum_{k=0}^n 2^k = \sum_{n=0}^{\infty} \frac{1}{2^{2n}} \cdot \frac{1-2^{n+1}}{1-2} \\
&= - \sum_{n=0}^{\infty} \left(\left(\frac{1}{4}\right)^n - 2 \left(\frac{1}{2}\right)^n \right) = - \sum_{n=0}^{\infty} \left(\frac{1}{4}\right)^n + 2 \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n \\
&= -\frac{4}{3} + 2 \cdot 2 = \frac{8}{3}.
\end{aligned}$$

Dies ist aber genau das Produkt der Grenzwerte 2 und $\frac{4}{3}$ der beiden einzelnen geometrischen Reihen. ■

2.2.5 Konvergenzkriterien für Reihen

Oft ist es schwierig, die Summe einer Reihe exakt auszurechnen. Vielfach genügt es aber völlig, wenn man weiß, ob eine Reihe überhaupt konvergiert. Um einen Näherungswert zu berechnen, kann man dann zu einer endlichen Summe (mit ausreichend vielen Summanden) übergehen. Das gelingt natürlich nicht, wenn die Reihe gar nicht konvergiert. Daher beschäftigen wir uns jetzt mit einfachen Kriterien, mit denen Konvergenz oder Divergenz einer Reihe festgestellt werden kann.

Beispiel 2.25

Ist die Reihe $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}}$ konvergent? Wir wissen, dass $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$ ist, und vergleichen damit: Wegen $\frac{1}{\sqrt{k}} \geq \frac{1}{k}$ ist

$$\sum_{k=1}^n \frac{1}{\sqrt{k}} \geq \sum_{k=1}^n \frac{1}{k}.$$

Da $\lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{1}{k} = \infty$ ist, muss dies insbesondere auch für die Ausgangsreihe gelten: $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} = \infty$. Entsprechend ist $\sum_{k=1}^{\infty} \frac{1}{k^\alpha} = \infty$ für $\alpha \in \mathbb{R}$, $\alpha \leq 1$. ■

Mit einem Vergleich kann man nicht nur die Divergenz einer Reihe zeigen, sondern auch die Konvergenz:

Satz 2.15 (Vergleichs- oder Majoranten/Minoranten-Kriterium)

Gegeben sei eine Reihe $\sum_{k=1}^{\infty} a_k$.

- a) Falls es ein $n_0 \in \mathbb{N}$, eine Konstante $M > 0$ sowie eine konvergente Reihe (Majorante) $\sum_{k=1}^{\infty} c_k$ gibt, so dass für alle $k \geq n_0$ gilt

$$|a_k| \leq M c_k,$$

dann ist die Reihe $\sum_{k=1}^{\infty} a_k$ absolut konvergent und insbesondere konvergent. $\sum_{k=1}^{\infty} c_k$ heißt eine **Majorante** von $\sum_{k=1}^{\infty} a_k$.

- b) Falls es ein $n_0 \in \mathbb{N}$, eine Konstante $M > 0$ sowie eine divergente Reihe $\sum_{k=1}^{\infty} d_k$ gibt, so dass für alle $k \geq n_0$ gilt

$$|a_k| \geq M d_k \geq 0,$$

dann ist die Reihe $\sum_{k=1}^{\infty} |a_k|$ ebenfalls divergent, d. h., $\sum_{k=1}^{\infty} a_k$ ist nicht absolut konvergent. $\sum_{k=1}^{\infty} d_k$ heißt eine **Minorante** von $\sum_{k=1}^{\infty} a_k$.

Beweis: Der Beweis basiert auf der Cauchy-Bedingung und dem Vergleich mit der Majorante:

- a) Da die Reihe $\sum_{k=1}^{\infty} c_k$ nach Voraussetzung konvergent ist, erfüllt sie die Cauchy-Bedingung, d. h., zu jedem $\frac{\varepsilon}{M} > 0$ existiert ein $n_1 = n_1\left(\frac{\varepsilon}{M}\right) \in \mathbb{N}$, so dass für alle $m \geq n > n_1$ gilt: $\left|\sum_{k=n}^m c_k\right| = \sum_{k=n}^m c_k < \frac{\varepsilon}{M}$. Für $m \geq n > \max\{n_0, n_1\}$ gilt damit:

$$\sum_{k=n}^m |a_k| \leq M \sum_{k=n}^m c_k < M \frac{\varepsilon}{M} = \varepsilon.$$

Damit ist aber die Cauchy-Bedingung auch für die Reihe $\sum_{k=1}^{\infty} |a_k|$ erfüllt, die also konvergent ist, d. h., $\sum_{k=1}^{\infty} a_k$ ist absolut konvergent.

b) Nach Voraussetzung ist $0 \leq d_k \leq \frac{1}{M}|a_k|$. Wäre nun $\sum_{k=1}^{\infty} |a_k|$ konvergent, so wäre nach a) auch $\sum_{k=1}^{\infty} d_k$ konvergent – im Widerspruch zur Voraussetzung der Divergenz. Also muss $\sum_{k=1}^{\infty} |a_k|$ divergent sein. $\square$

Um mit diesem Satz die Konvergenz oder Divergenz einer Reihe zu zeigen, benötigt man geeignete Reihen zum Vergleich. Mit der harmonischen Reihe kann man gut Divergenz zeigen. Zum Nachweis von Konvergenz ist eine der prominentesten Reihen die geometrische Reihe $\sum_{k=1}^{\infty} q^k$ für $|q| < 1$ (siehe (2.12)).

Beispiel 2.26

Wir untersuchen die Konvergenz von $\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{4} + \frac{1}{9} + \dots$ mit dem Majorantenkriterium. Aus $k^2 \geq k(k-1) > 0$ für $k \geq 2$ folgt

$$\frac{1}{k^2} \leq \frac{1}{k(k-1)}.$$

Damit haben wir wegen

$$\sum_{k=2}^n \frac{1}{k(k-1)} = \sum_{k=1}^{n-1} \frac{1}{k(k+1)} = \sum_{k=1}^{n-1} \left(\frac{1}{k} - \frac{1}{k+1} \right) = 1 - \frac{1}{(n-1)+1}$$

eine konvergente Majorante mit Grenzwert 1 (vgl. 2.13). Somit ist auch $\sum_{k=1}^{\infty} \frac{1}{k^2}$ konvergent. Wir erhalten so zwar nicht den Grenzwert, können ihn aber abschätzen:

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

Für $m \geq 2$ haben wir wegen $\frac{1}{k^m} \leq \frac{1}{k^2}$ darüber hinaus $\sum_{k=1}^{\infty} \frac{1}{k^m} \leq 2$ gezeigt. Es gilt sogar für $\alpha \in \mathbb{R}$ mit $\alpha > 1$: $\sum_{k=1}^{\infty} \frac{1}{k^\alpha} \leq 1 + \frac{1}{\alpha-1}$, siehe (2.61) auf Seite 368. Dagegen ist die Reihe für $\alpha \leq 1$ divergent, siehe Beispiel 2.25. $\blacksquare$

Beispiel 2.27

Wegen $\frac{1}{(k+1)(k+3)} \leq \frac{1}{k^2}$ konvergiert die Zahlenreihe

$$\sum_{k=1}^{\infty} \frac{1}{(k+1)(k+3)} = \frac{1}{2 \cdot 4} + \frac{1}{3 \cdot 5} + \frac{1}{4 \cdot 6} + \dots$$

nach dem Vergleichskriterium.

Mit dem Vergleichskriterium können wir allenfalls den Grenzwert einschränken. Um ihn auszurechnen, hilft die Partialbruchzerlegung des Summanden a_k :

$$a_k = \frac{1}{(k+1)(k+3)} = \frac{1}{2} \cdot \frac{(k+3) - (k+1)}{(k+1)(k+3)} = \frac{1}{2} \left(\frac{1}{k+1} - \frac{1}{k+3} \right).$$

Mit dieser Partialbruchzerlegung können wir die Partialsumme s_n als Teleskopsumme schreiben und ausrechnen:

$$\begin{aligned} s_n &= \sum_{k=1}^n \frac{1}{(k+1)(k+3)} = \frac{1}{2} \sum_{k=1}^n \left(\frac{1}{k+1} - \frac{1}{k+3} \right) \\ &= \frac{1}{2} \left[\left(\frac{1}{2} - \frac{1}{4} \right) + \left(\frac{1}{3} - \frac{1}{5} \right) + \left(\frac{1}{4} - \frac{1}{6} \right) + \dots \right. \\ &\quad \left. \dots + \left(\frac{1}{n} - \frac{1}{n+2} \right) + \left(\frac{1}{n+1} - \frac{1}{n+3} \right) \right] = \frac{1}{2} \left[\frac{1}{2} + \frac{1}{3} - \frac{1}{n+2} - \frac{1}{n+3} \right]. \end{aligned}$$

Da sich fast alle Summanden aufheben, bleiben zu den Indizes $k = 1$ und $k = 2$ nur die Summanden $\frac{1}{2}$ und $\frac{1}{3}$ bzw. zu den Indizes $k = n - 1$ und $k = n$ nur die Summanden $-\frac{1}{n+2}$ und $-\frac{1}{n+3}$ übrig. Für $n \rightarrow \infty$ streben diese gegen null, so dass gilt

$$\sum_{k=1}^{\infty} \frac{1}{(k+1)(k+3)} = \lim_{n \rightarrow \infty} s_n = \frac{1}{2} \left[\frac{1}{2} + \frac{1}{3} \right] = \frac{5}{12}.$$

■

Beispiel 2.28 (Dezimalbruchdarstellung der reellen Zahlen)

Der Dezimalbruch $\pm x_0, x_1 x_2 x_3 x_4 \dots$ mit $x_0 \in \mathbb{N}_0$ und $x_k \in \{0, 1, 2, \dots, 9\}$ als Nachkommastellen ist als Grenzwert der Reihe $\pm \left(\sum_{k=0}^{\infty} \frac{x_k}{10^k} \right)$ tatsächlich eine reelle Zahl. Denn wegen $|x_k| \leq 9$ für $k \in \mathbb{N}$ hat die Reihe $\sum_{k=1}^{\infty} \frac{x_k}{10^k}$ die konvergente geometrische Reihe für $q = 10^{-1}$ als Majorante: $\left| \frac{x_k}{10^k} \right| \leq 9q^k$.

Streng genommen haben wir auf Seite 51 den Wert eines positiven Dezimalbruchs mittels Supremum über $\sup \left\{ \sum_{k=0}^n \frac{x_k}{10^k} : n \in \mathbb{N}_0 \right\}$ erklärt. Die Reihe ist nun aber eine monoton wachsende Folge. Da sie konvergiert, ist sie auch beschränkt. Ihr Grenzwert ist nach dem Beweis des Monotoniekriteriums Satz 2.3 das Supremum der Wertemenge der Folge der Partialsummen. Daher ist die Definition über das Supremum nur eine andere Schreibweise für den Grenzwert der Reihe.

Jeder Dezimalbruch ist eine reelle Zahl. Umgekehrt stellt sich noch die Frage, ob auch jede reelle Zahl x eine Darstellung als Dezimalbruch besitzt. Dazu verwenden wir eine Intervallschachtelung. Zur Vereinfachung der Schreibweise sei $x > 0$. Zunächst gibt es eine natürliche Zahl $x_0 \in \mathbb{N}_0$ mit $x_0 \leq x < x_0 + 1$, denn $\{n \in \mathbb{N}_0 : n \leq x\}$ ist beschränkt und damit endlich, so dass es ein Maximum x_0 gibt. Entsprechend findet man eine erste Nachkommastelle $x_1 \in \{0, 1, \dots, 9\}$ mit $x_0 + \frac{x_1}{10} \leq x < x_0 + \frac{x_1}{10} + 0,1$. Setzt man diese Konstruktion fort, so erhält man eine Intervallschachtelung

$$\left[\sum_{k=0}^{n+1} \frac{x_k}{10^k}, \sum_{k=0}^{n+1} \frac{x_k}{10^k} + \frac{1}{10^{n+1}} \right] \subset \left[\sum_{k=0}^n \frac{x_k}{10^k}, \sum_{k=0}^n \frac{x_k}{10^k} + \frac{1}{10^n} \right].$$

In jedem der Intervalle liegt die Zahl x . Da der Durchmesser der Intervalle gegen null strebt, ist diese nach Satz 2.6 auch die einzige gemeinsame Zahl in allen Intervallen, gegen die außerdem die Folge der unteren Intervallränder konvergiert. Das ist aber gerade die Reihe einer Dezimalbruchdarstellung. ■

Beispiel 2.29 (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 in Kapitel 6.3 behandelte Laplace-Transformation für Funktionen, siehe z. B. Goebbels (2014).

Wegen der Wachstumsbedingung ist $|a_k \frac{1}{z^k}| \leq C \left| \frac{\alpha}{z} \right|^k$. 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

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

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 $|z| > 2$:

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

Hier haben wir ausgenutzt, dass konvergente Reihen gliedweise addiert werden dürfen (Satz 2.9). Die Nullstellen $\frac{1 \pm \sqrt{5}}{2}$ des Nenners liegen in $[-2, 2]$, die positive ist der goldene Schnitt (siehe Seiten 56, 231). ■

Das Vergleichskriterium kann sich in den Anwendungen als etwas sperrig erweisen, da man eine konvergente Majorante finden muss. Daher werden aus diesem Kriterium nun einige weitere abgeleitet, die auf gängige Situationen besser zugeschnitten sind.

Durch Vergleich mit der geometrischen Reihe beweist man das wichtige Quotientenkriterium von D'Alembert, der uns bereits als Entdecker des Fundamentalsatzes der Algebra (Satz 1.24 auf Seite 147) begegnet ist:

Satz 2.16 (Quotientenkriterium für eine Reihe)

Sei $\sum_{k=1}^{\infty} a_k$ eine Reihe mit $a_k \neq 0$, $k \in \mathbb{N}$. Weiter möge

$$r := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|$$

existieren, dann gilt:

- a) Falls $r < 1$ ist, ist $\sum_{k=1}^{\infty} a_k$ absolut konvergent und insbesondere konvergent.
- b) Falls $r > 1$ ist, ist $\sum_{k=1}^{\infty} a_k$ divergent.

Damit eine Reihe konvergieren kann, muss sie über eine Nullfolge $(a_k)_{k=1}^{\infty}$ gebildet sein. Wir haben dann die (etwas unpräzise) Vorstellung, dass $|a_k|$ im Wesentlichen monoton fallend ist, dass also in der Regel $\left| \frac{a_{k+1}}{a_k} \right| < 1$ ist. Wenn nun der Grenzwert r dieses Quotienten nicht nur kleiner oder gleich eins, sondern sogar echt kleiner eins ist, dann fällt $|a_k|$ schnell genug, um die Konvergenz der Reihe zu erreichen. Wäre sogar ohne Grenzwert $r = \left| \frac{a_{k+1}}{a_k} \right|$ konstant für alle Werte $k \in \mathbb{N}$, so wäre unter Verwendung eines Teleskopprodukts

$$|a_k| = \left| \frac{a_k}{a_{k-1}} \right| \left| \frac{a_{k-1}}{a_{k-2}} \right| \cdots \left| \frac{a_2}{a_1} \right| |a_1| = r^{k-1} |a_1|,$$

und die Reihe wäre bis auf einen konstanten Faktor $|a_1|$ eine geometrische Reihe. Im Beweis müssen wir lediglich noch den Grenzwert berücksichtigen.

Achtung: Erfahrungsgemäß bereitet das Hinschreiben des Ausdrucks $\frac{a_{k+1}}{a_k}$ für eine konkrete Folge Schwierigkeiten. Im Zähler muss man jedes Auftreten von k bei der Berechnung des Folgenglieds durch $k+1$ ersetzen, ist z. B. $a_k = \frac{\sin k}{k^2}$, so ist $a_{k+1} = \frac{\sin(k+1)}{(k+1)^2}$.

Beweis: a) Falls $r < 1$ ist, gibt es nach Definition des Grenzwerts zu einem fest gewählten $\alpha \in]r, 1[$ ein $n_0 \in \mathbb{N}$, so dass für alle $k \geq n_0$ gilt: $\left| \frac{a_{k+1}}{a_k} \right| < \alpha$. Damit ist für $k \geq n_0$:

$$|a_k| = \left| \frac{a_k}{a_{k-1}} \right| \left| \frac{a_{k-1}}{a_{k-2}} \right| \cdots \left| \frac{a_{n_0+1}}{a_{n_0}} \right| |a_{n_0}| \leq \alpha^{k-n_0} |a_{n_0}| \leq \underbrace{\frac{|a_{n_0}|}{\alpha^{n_0}}}_{=: M} \alpha^k.$$

Da die geometrische Reihe $\sum_{k=1}^{\infty} \alpha^k$ für dieses $0 < \alpha < 1$ konvergiert, gilt dies laut Vergleichskriterium auch für die Reihe $\sum_{k=1}^{\infty} |a_k|$.

- b) Da $r > 1$ ist, existiert nach Definition des Grenzwerts ein $n_0 \in \mathbb{N}$, so dass für alle $k \geq n_0$ gilt:

$$\left| \frac{a_{k+1}}{a_k} \right| \geq 1.$$

Damit ist aber $|a_{k+1}| \geq |a_k| \geq \dots \geq |a_{n_0}|$. Nach Voraussetzung ist zudem $|a_{n_0}| \neq 0$. Damit ist $(|a_k|)_{k=1}^{\infty}$ und somit auch $(a_k)_{k=1}^{\infty}$ keine Nullfolge – die Reihe $\sum_{k=1}^{\infty} a_k$ divergiert wegen Satz 2.11.

□

Hätten wir den Satz nicht durch Vergleich mit der geometrischen Reihe $\sum_{k=1}^{\infty} q^k$ gezeigt, könnten wir umgekehrt mit dem Quotientenkriterium deren Konvergenz für $|q| < 1$ beweisen.

Beispiel 2.30

- a) Wir haben bereits auf Seite 226 mühsam gezeigt, dass die Reihe $\sum_{k=0}^{\infty} \frac{1}{k!}$ konvergiert, so dass wir die Zahl e über ihren Grenzwert definieren konnten. Mit dem Quotientenkriterium ist der Konvergenznachweis einfacher: Mit $a_k = \frac{1}{k!} > 0$ ist

$$r := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k \rightarrow \infty} \frac{\frac{1}{(k+1)!}}{\frac{1}{k!}} = \lim_{k \rightarrow \infty} \frac{k!}{(k+1)!} = \lim_{k \rightarrow \infty} \frac{1}{k+1} = 0 < 1.$$

Das Quotientenkriterium eignet sich besonders gut, wenn in der Reihe Fakultäten auftreten. Durch die Division kürzen sich diese weitgehend weg. Fakultäten hat man z. B. bei Potenzreihen (siehe Kapitel 2.8).

- b) Die Reihe $\sum_{k=1}^{\infty} \frac{2^k}{k}$ ist divergent, da $\lim_{k \rightarrow \infty} \left| \frac{\frac{2^{k+1}}{k+1}}{\frac{2^k}{k}} \right| = \lim_{k \rightarrow \infty} \frac{2^{k+1}k}{(k+1)2^k} = 2 \cdot \lim_{k \rightarrow \infty} \frac{k}{k+1} = 2 > 1$. Auch ohne das Quotientenkriterium erkennt man die Divergenz daran, dass $\left(\frac{2^k}{k}\right)_{k=1}^{\infty}$ keine Nullfolge ist.

- c) Wir untersuchen die Reihe $\sum_{k=1}^{\infty} \frac{k!}{k^k} = 1 + \frac{2!}{2^2} + \frac{3!}{3^3} + \frac{4!}{4^4} + \dots$ auf absolute Konvergenz.

Mit $a_k = \frac{k!}{k^k}$, $a_{k+1} = \frac{(k+1)!}{(k+1)^{k+1}}$ erhält man durch Kürzen der Fakultäten bei der Berechnung des Grenzwerts aus Satz 2.16 für $k \rightarrow \infty$

$$\left| \frac{a_{k+1}}{a_k} \right| = \frac{(k+1)!}{(k+1)^{k+1}} \frac{k^k}{k!} = \frac{(k+1)}{(k+1)^{k+1}} \frac{k^k}{1} = \left(\frac{k}{k+1} \right)^k = \frac{1}{\left(1 + \frac{1}{k}\right)^k} \xrightarrow{(2.8)} \frac{1}{e}.$$

Wegen $\frac{1}{e} < 1$ folgt die absolute Konvergenz der Reihe.

■

Im Fall $r = 1$ ist eine einfache Aussage zur Konvergenz oder Divergenz nicht möglich. Hier versagt das Quotientenkriterium. Während für die divergente harmonische Reihe $\sum_{k=1}^{\infty} \frac{1}{k}$ gilt:

$$r = \lim_{k \rightarrow \infty} \frac{\frac{1}{k+1}}{\frac{1}{k}} = \lim_{k \rightarrow \infty} \frac{k}{k+1} = 1,$$

ist die Reihe $\sum_{k=1}^{\infty} \frac{1}{k^2}$ konvergent bei $r = 1$. Das haben wir bereits mit einer konvergenten Majorante bewiesen, lässt sich aber auch mit dem Quotientenvergleichskriterium zeigen:

Satz 2.17 (Quotientenvergleichskriterium *)

Seien $a_k, b_k > 0$ für alle $k \geq n_0$.

a) Ist $\sum_{k=1}^{\infty} b_k$ konvergent und gilt für alle $k \geq n_0$

$$\frac{a_{k+1}}{a_k} \leq \frac{b_{k+1}}{b_k},$$

dann ist $\sum_{k=1}^{\infty} a_k$ konvergent.

b) Ist $\sum_{k=1}^{\infty} b_k$ divergent und gilt für alle $k \geq n_0$

$$\frac{a_{k+1}}{a_k} \geq \frac{b_{k+1}}{b_k},$$

dann ist $\sum_{k=1}^{\infty} a_k$ divergent.

Beweis: a) Für $k \geq n_0$ ist

$$0 < \frac{a_{k+1}}{b_{k+1}} \leq \frac{a_k}{b_k} \leq \frac{a_{k-1}}{b_{k-1}} \leq \dots \leq \frac{a_{n_0}}{b_{n_0}} =: M.$$

Damit ist $|a_k| = a_k \leq M b_k$ für alle $k \geq n_0$, so dass mit dem Vergleichskriterium die Konvergenz von $\sum_{k=1}^{\infty} a_k$ folgt.

b) Völlig analog zu a) gilt für $k \geq n_0$, dass

$$\frac{a_{k+1}}{b_{k+1}} \geq \frac{a_k}{b_k} \geq \frac{a_{k-1}}{b_{k-1}} \geq \dots \geq \frac{a_{n_0}}{b_{n_0}} =: M.$$

Damit ist nun $a_k \geq M b_k$, und die Divergenz folgt ebenfalls mit dem Vergleichskriterium. □

Beispiel 2.31

Wir können jetzt auch mit dem Quotientenvergleichskriterium die Konvergenz der Reihe $\sum_{k=1}^{\infty} \frac{1}{k^2}$ zeigen. Nach (2.13) ist $\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$. Wir setzen $a_k = \frac{1}{k^2}$, $b_k = \frac{1}{k(k+1)}$. Dann ist $a_k, b_k > 0$ und da $k(k+2) = k^2 + 2k < k^2 + 2k + 1 = (k+1)^2$:

$$\frac{a_{k+1}}{a_k} = \frac{k^2}{(k+1)^2} < \frac{k^2}{k(k+2)} = \frac{k}{k+2} = \frac{k(k+1)}{(k+1)(k+2)} = \frac{b_{k+1}}{b_k}. \quad \blacksquare$$

Ebenfalls auf einem Vergleich basiert das Wurzelkriterium:

Satz 2.18 (Wurzelkriterium)

Sei $\sum_{k=1}^{\infty} a_k$ eine Reihe, bei der $r := \lim_{k \rightarrow \infty} \sqrt[k]{|a_k|}$ existiert.

- a) Ist $r < 1$, so ist die Reihe $\sum_{k=1}^{\infty} a_k$ absolut konvergent und insbesondere konvergent.
- b) Ist $r > 1$, so ist die Reihe $\sum_{k=1}^{\infty} a_k$ divergent.

Auch dieses Kriterium prüft, ob die Folge $(|a_k|)_{k=1}^{\infty}$ schnell genug gegen null strebt. Wie das Quotientenkriterium wird es später bei Potenzreihen (siehe Kapitel 2.8) wichtig. Wäre ohne Grenzwert bereits $r = \sqrt[k]{|a_k|}$ konstant, also $|a_k| = r^k$, so würden wir unmittelbar eine geometrische Reihe untersuchen. Im Beweis berücksichtigen wir zusätzlich den Grenzwert.

Beweis: a) Wir beweisen auch diese Aussage wie angekündigt durch Vergleich mit der geometrischen Reihe: Zu einem beliebig gewählten $\beta \in]r, 1[$ existiert ein $n_0 \in \mathbb{N}$, so dass für alle $k > n_0$ gilt: $|a_k|^{1/k} < \beta$, d. h. $|a_k| < \beta^k$. Damit ist die geometrische Reihe $\sum_{k=1}^{\infty} \beta^k$ eine konvergente Majorante.

b) Wir zeigen, dass $(a_k)_{k=1}^{\infty}$ keine Nullfolge ist, so dass auch die Reihe nicht konvergieren kann: Da $\lim_{k \rightarrow \infty} |a_k|^{1/k} = r > 1$ ist, existiert ein $n_0 \in \mathbb{N}$, so dass für alle $k \geq n_0$ gilt: $|a_k|^{1/k} \geq 1$, also $|a_k| \geq 1$. $\square$

Sind sowohl das Quotientenkriterium (Satz 2.16) als auch das Wurzelkriterium anwendbar, so sind die in den Kriterien zu berechnenden Grenzwerte gleich. Das sieht man z. B. ohne zusätzlichen Beweis mittels der noch folgenden Formel von Cauchy-Hadamard (Satz 2.60, Seite 413) und der Eindeutigkeit des Konvergenzradius einer Potenzreihe (Satz 2.59, Seite 411). Allerdings handelt es sich bei diesen Grenzwerten um Hilfsgrößen, die in der Regel nicht dem Grenzwert der Reihe entsprechen!

Beispiel 2.32

Die Reihe $\sum_{k=1}^{\infty} \left(\frac{2}{k}\right)^k$ ist konvergent, denn $\lim_{k \rightarrow \infty} \sqrt[k]{\left(\frac{2}{k}\right)^k} = \lim_{k \rightarrow \infty} \frac{2}{k} = 0 < 1$. $\blacksquare$

Im Fall $r = 1$ ist keine allgemeine Aussage möglich. Auch hier betrachten wir wieder die divergente harmonische Reihe $\sum_{k=1}^{\infty} \frac{1}{k}$ und die konvergente Reihe $\sum_{k=1}^{\infty} \frac{1}{k^2}$. Bei beiden ist $r = 1$ (Nachrechnen gelingt mit dem Satz von L'Hospital, siehe Kapitel 2.5.3).

Während sich bei Anwendung des Quotientenkriteriums Fakultäten regelmäßig wegkürzen, ist das Ziehen der k -ten Wurzel aus einer Fakultät schwierig. Dazu benötigt man als Hilfsmittel die Stirling'sche Formel (2.62), die wir auf Seite 369 kennenlernen.

Quotienten- und Wurzelkriterium untersuchen eine Reihe auf absolute Konvergenz. Sie erfassen nicht, ob sich Summanden durch Vorzeichenwechsel aufheben. Das wichtigste Kriterium, das dies berücksichtigt, geht auf Gottfried Wilhelm Leibniz (1646–1716) zurück, der als einer der letzten Universalgelehrten gilt. Nach ihm ist auch der gleichnamige Keks benannt.

Satz 2.19 (Leibniz-Kriterium)

Ist $(a_k)_{k=1}^\infty$ eine monoton fallende Nullfolge (und damit $a_k \geq 0$ für alle $k \in \mathbb{N}$), dann ist die (alternierende) Reihe $\sum_{k=1}^\infty (-1)^k a_k$ konvergent.

Hier wird eine alternierende Folge $(-1)^k a_k$ aufsummiert. Dagegen wird häufig das Leibniz-Kriterium für Reihen $\sum_{k=1}^\infty (-1)^{k+1} a_k$ formuliert. Die Multiplikation der Reihe mit -1 ändert nichts an der Konvergenz der Reihe. Das Leibniz-Kriterium ist ein Spezialfall des allgemeineren Dirichlet-Kriteriums. Einer der Autoren möchte an dieser Stelle aus Lokalpatriotismus erwähnen, dass Dirichlet (1805–1859) aus Düren stammte und damit ein bedeutender Mathematiker aus dem Rheinland war.

Beweis: Sei $s_n := \sum_{k=1}^n (-1)^k a_k$. Wir betrachten Teilfolgen zu ungeraden und zu geraden Indizes:

- $(s_{2n-1})_{n=1}^\infty$ wächst monoton:

$$s_{2n+1} = s_{2n-1} + (-1)^{2n} a_{2n} + (-1)^{2n+1} a_{2n+1} = s_{2n-1} + a_{2n} - a_{2n+1} \geq s_{2n-1},$$

da $a_{2n} \geq a_{2n+1}$ aufgrund der Monotonie der Folge $(a_k)_{k=1}^\infty$ ist.

- $(s_{2n})_{n=1}^\infty$ fällt monoton: $s_{2n+2} = s_{2n} - a_{2n+1} + a_{2n+2} \leq s_{2n}$, da $a_{2n+1} \geq a_{2n+2}$ wieder aufgrund der Monotonie gilt.
- $s_{2n-1} \leq s_{2n}$ für alle $n \in \mathbb{N}$: $s_{2n} = s_{2n-1} + a_{2n} \geq s_{2n-1}$, da $a_{2n} \geq 0$ nach Voraussetzung.

Anders ausgedrückt: Die Intervalle $I_n := [s_{2n-1}, s_{2n}]$ bilden eine Intervallschachtelung $I_n \subset I_{n-1} \subset I_{n-2} \subset \cdots \subset I_1$.

Für die Länge der Intervalle gilt: $\lim_{n \rightarrow \infty} (s_{2n} - s_{2n-1}) = \lim_{n \rightarrow \infty} a_{2n} = 0$.

Aufgrund der Vollständigkeit der reellen Zahlen existiert ein $s \in \mathbb{R}$, das in allen Intervallen enthalten ist (siehe Satz 2.6) und für das insbesondere gilt:

$$\lim_{n \rightarrow \infty} s_{2n} = s = \lim_{n \rightarrow \infty} s_{2n-1}.$$

Gegen diesen Grenzwert s konvergiert dann aber auch die Folge $(s_n)_{n=1}^\infty$, die aus den beiden Teilfolgen gebildet wird; die Reihe ist also konvergent. $\square$

Insbesondere haben wir eine Fehlerabschätzung bewiesen. Da $s \in I_n$ bzw. $s_{2n-1} \leq s \leq s_{2n}$ gilt, ist $0 \leq s_{2n} - s \leq s_{2n} - s_{2n-1} = a_{2n}$ und $0 \leq s - s_{2n-1} \leq s_{2n} - s_{2n-1} = a_{2n} \leq a_{2n-1}$. Mit anderen „Worten“ gilt für die Restsummen:

$$\left| \sum_{k=m+1}^{\infty} (-1)^k a_k \right| = |s_m - s| \leq a_m.$$

Beispiel 2.33

- a) Die alternierende harmonische Reihe $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k}$ konvergiert, da die Nullfolge $\left(\frac{1}{k}\right)_{k=1}^{\infty}$ monoton fällt. Es gilt die Fehlerabschätzung

$$\left| \sum_{k=1}^n (-1)^k \frac{1}{k} - \sum_{k=1}^{\infty} (-1)^k \frac{1}{k} \right| \leq \frac{1}{n}.$$

- b) Wir zeigen, dass die alternierende Reihe $\sum_{k=1}^{\infty} (-1)^k \cdot \frac{k+7}{k^2}$ über eine monotone Nullfolge gebildet und damit konvergent ist. Zunächst ist $\lim_{k \rightarrow \infty} |a_k| = \lim_{k \rightarrow \infty} \frac{k+7}{k^2} = \lim_{k \rightarrow \infty} \frac{k(1+\frac{7}{k})}{k \cdot k} = \lim_{k \rightarrow \infty} \frac{1+\frac{7}{k}}{k} = 0$, und für die Beträge aufeinander folgender Glieder der Reihe erhält man

$$|a_{k+1}| = \frac{(k+1)+7}{(k+1)^2} = \frac{(k+1)\left(1+\frac{7}{k+1}\right)}{(k+1)^2} = \frac{1+\frac{7}{k+1}}{k+1} \leq \frac{1+\frac{7}{k}}{k+1} \leq \frac{1+\frac{7}{k}}{k} = |a_k|,$$

d. h., $(|a_k|)_{k=1}^{\infty}$ ist eine monotone Nullfolge, und die Reihe ist konvergent. ■

2.3 Grenzwerte von Funktionen und Stetigkeit

Bei Folgen $(a_n)_{n=1}^{\infty}$ und Reihen $(s_n)_{n=1}^{\infty}$ hat uns das Verhalten der Glieder für $n \rightarrow \infty$ interessiert. Bei Funktionen f kann man entsprechend diskutieren, was für $x \rightarrow \infty$ passiert. Ebenso kann man jetzt aber auch $x \rightarrow -\infty$ und sogar das Grenzwertverhalten an einzelnen Stellen des Definitionsbereichs analysieren. Dazu benötigen wir jedoch zunächst ein paar neue Begriffe.

2.3.1 Umgebungen und Überdeckungen

Um Eigenschaften von Funktionen an einer Stelle $x_0 \in \mathbb{R}$ ihres Definitionsbereichs untersuchen zu können, müssen die Funktionen auch noch ein Stückchen rechts und (oder) links davon definiert sein. Wenn wir z. B. wissen, wie weit zu jedem Zeitpunkt 11:00 Uhr $\leq t < 12:00$ Uhr ein Zug vom Bahnhof entfernt ist, dann können wir aus diesen Daten schließen, welche Distanz der Zug zum exakten Zeitpunkt 12:00 Uhr zum Bahnhof hat, obwohl der Funktionswert nicht angegeben ist. Beschreibt die Funktion

$s(t)$ das Weg-Zeit-Diagramm des Zuges, dann werden wir seine Entfernung um 12:00 Uhr als einen Grenzwert dieser Funktion für t gegen 12:00 Uhr (von links) erhalten.

Definition 2.11 (Umgebungen)

Für eine Zahl $\delta > 0$ bezeichnet man die Menge

$$U_\delta(x_0) :=]x_0 - \delta, x_0 + \delta[= \{x \in \mathbb{R} : x_0 - \delta < x < x_0 + \delta\}$$

als (offene) δ -Umgebung eines Punktes x_0 .

- Gibt es zu $x_0 \in D$ eine δ -Umgebung $U_\delta(x_0) \subset D$, so nennt man x_0 einen **inneren Punkt** von D . In diesem Fall gehört also nicht nur x_0 zu D , sondern auch noch ein Stückchen rechts und links von x_0 .
- Entfernt man den Punkt x_0 aus der Umgebung, so spricht man von einer **punktierten δ -Umgebung** $]x_0 - \delta, x_0[\cup]x_0, x_0 + \delta[$.
- $]x_0 - \delta, x_0[$ ist eine **linksseitige**, $[x_0, x_0 + \delta[$ eine **rechtsseitige Umgebung** von x_0 . $]x_0 - \delta, x_0[$ ist eine **linksseitige punktierte Umgebung** von x_0 , $]x_0, x_0 + \delta[$ eine **rechtsseitige punktierte Umgebung**.

Zu jedem Punkt $x \in]a, b[$ kann eine δ -Umgebung gefunden werden, die ganz in $]a, b[$ liegt. Für die Punkte a und b in $[a, b]$ ist dies nicht der Fall, hier gibt es nur einseitige Umgebungen.

In Verallgemeinerung zu offenen Intervallen sagt man, dass eine Menge $E \subset \mathbb{R}$ **offen** heißt, falls zu jedem $x \in E$ eine δ -Umgebung $U_\delta(x) \subset E$ existiert. Eine Menge $E \subset \mathbb{R}$ heißt **abgeschlossen**, falls ihr Komplement $C_{\mathbb{R}}E := \{x \in \mathbb{R} : x \notin E\}$ offen ist, z. B. ist $[a, b]$ abgeschlossen. Die **Topologie** beschäftigt sich als mathematische Disziplin allgemein mit offenen Mengen und führt diese über ein Axiomensystem ein. Damit löst sie sich von $\mathbb{R}$, siehe Kasten auf Seite 276.

Hintergrund: Der Satz von Heine-Borel

Wir behandeln in diesem Einschub einen Spezialfall des Satzes von Heine-Borel für abgeschlossene Intervalle. Dieser Satz ist für einige Beweise von Konvergenzaussagen sehr wichtig. Den hier betrachteten Spezialfall verwenden wir beispielsweise noch, um zu zeigen, dass auf einem abgeschlossenen Intervall $[a, b]$ stetige Funktionen (der Funktionsgraph lässt sich durchzeichnen) insbesondere auch integrierbar sind (die Fläche zwischen Graph und x -Achse ist berechenbar). Der Rest dieses Einschubs ist nur für Leser interessant, die sich mit den später angegebenen Beweisen auseinandersetzen möchten.

Satz 2.20 (Satz von Heine-Borel für abgeschlossene Intervalle)

Sei $[a, b] \subset \mathbb{R}$ und $\mathcal{F}$ eine (eventuell unendlich große) Menge von offenen Intervallen. $[a, b]$ möge in der Vereinigungsmenge aller dieser offenen Intervalle liegen. Man sagt,

dass $\mathcal{F}$ das Intervall $[a, b]$ überdeckt. Dann existiert eine endliche Teilmenge von $\mathcal{F}$, so dass $[a, b]$ auch in der Vereinigung dieser endlich vielen Intervalle liegt. Man spricht von einer **endlichen Teilüberdeckung** von $[a, b]$.

Man benutzt den Satz, wenn man eine Eigenschaft z. B. einer Funktion für **unendlich** viele kleine offene Intervalle kennt und damit auf die Eigenschaft für ein großes Intervall $[a, b]$ schließen möchte. Oft ist es dabei ausreichend, wenn man $[a, b]$ (mit einer endlichen Teilüberdeckung) in **endlich** viele Teile zerlegen kann, für die man die Eigenschaft kennt.

Der Satz ist alles andere als selbstverständlich, da $\mathcal{F}$ sogar mehr als abzählbar unendlich viele offene Intervalle enthalten kann. Deshalb geht der Beweis auch auf das für $\mathbb{R}$ grundlegende Vollständigkeitsaxiom (1.5) von Seite 49 zurück:

Beweis: Sei $E := \{x \in [a, b] : \text{Es existiert eine endliche Teilmenge von } \mathcal{F}, \text{ die } [a, x] \text{ überdeckt.}\}$. Der Satz ist bewiesen, wenn wir zeigen können, dass $b \in E$ ist, da es dann nach Definition von E für $[a, b]$ eine endliche Teilüberdeckung gibt.

Dazu verwenden wir die Vollständigkeit der reellen Zahlen. Die Menge E ist nicht-leer, da ein Element von $\mathcal{F}$ den Punkt a überdecken muss. Damit ist $a \in E$. Außerdem ist wegen $E \subset [a, b]$ die Menge E beschränkt. Damit existiert nach Axiom (1.5) $S := \sup E$ und wegen (1.7) ist $S \in [a, b]$ (siehe Seite 51).

Da das Supremum im Intervall $[a, b]$ enthalten ist, wird es von einem offenen Intervall aus $\mathcal{F}$ überdeckt, d. h., es existieren $\delta_1, \delta_2 > 0$ mit $]S - \delta_1, S + \delta_2[\in \mathcal{F}$. Dieses Intervall überdeckt aber nicht nur S , sondern auch noch Punkte aus einer Umgebung von S , so dass man auch für diese zu einer endlichen Überdeckung gelangt: Nach Definition von S gibt es eine endliche Teilmenge von $\mathcal{F}$, die $[a, S - \frac{\delta_1}{2}]$ überdeckt. Fügen wir zu dieser Überdeckung das Intervall $]S - \delta_1, S + \delta_2[$ hinzu, so erhalten wir eine endliche Überdeckung von $[a, \min\{b, S + \frac{\delta_2}{2}\}]$. Nach Definition von S kann kein größeres Intervall als $[a, S]$ endlich überdeckt werden. Damit kein Widerspruch entsteht, muss $S = b$ sein, und $[a, b]$ wird bereits von einer endlichen Teilmenge von $\mathcal{F}$ überdeckt. $\square$

Der Satz gilt nicht für (einseitig) offene Intervalle. Wir betrachten $]0, 1[$ und die Menge $\mathcal{F} := \{]\frac{1}{n}, 1[: n \geq 2\}$. Offensichtlich ist $]0, 1[= \bigcup_{n=2}^{\infty}]\frac{1}{n}, 1[$, aber jede endliche Vereinigung $\bigcup_{k=1}^m]\frac{1}{n_k}, 1[$ besitzt ein größtes Element $n_0 \in \{n_1, \dots, n_m\}$, so dass $\frac{1}{n_0+1} \notin \bigcup_{k=1}^m]\frac{1}{n_k}, 1[$. Das Intervall $]0, 1[$ wird also nicht vollständig überdeckt.

2.3.2 Grenzwerte von Funktionen

Wie verhält sich eine Funktion, wenn man immer größere Werte einsetzt, also Funktionswerte für Folgenglieder einer bestimmt divergenten Folge berechnet? Was passiert, wenn man in eine Funktion Werte einsetzt, die immer näher bei einer Stelle x_0 liegen? Diese Fragestellungen führen zum Begriff des Grenzwerts für Funktionen.

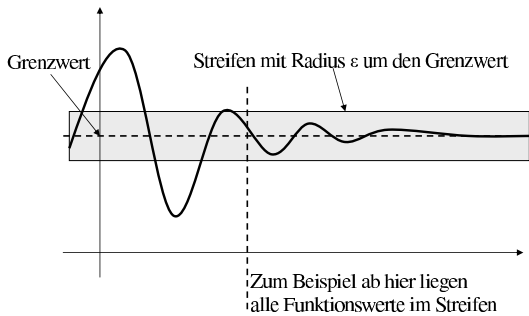


Abb. 2.10: Zur Definition des Grenzwerts $\lim_{x \rightarrow \infty} f(x) = L$

Definition 2.12 (Konvergenz einer Funktion im Unendlichen)

Sei f eine Funktion von $D \subset \mathbb{R}$ in $\mathbb{R}$.

- Es sei $]a, \infty[:= \{x \in \mathbb{R} : x > a\} \subset D$ für ein $a \in \mathbb{R}$, d. h., $f(x)$ ist für alle Argumente $x > a$ erklärt.

Die Funktion f heißt **konvergent** gegen $L \in \mathbb{R}$ für $x \rightarrow \infty$ genau dann, wenn zu jedem (noch so kleinen) $\varepsilon > 0$ ein $X_0 = X_0(\varepsilon) \in \mathbb{R}$, $X_0 \geq a$, existiert, so dass für alle $x > X_0$ gilt: $|f(x) - L| < \varepsilon$ (siehe Abbildung 2.10). L heißt dann der **Grenzwert** von f für $x \rightarrow \infty$. Schreibweise:

$$\lim_{x \rightarrow \infty} f(x) = L.$$

- Falls $f(x)$ für alle Argumente $x < a$ erklärt ist und zu jedem $\varepsilon > 0$ ein $X_0 \in \mathbb{R}$, $X_0 < a$, existiert, so dass $|f(x) - L| < \varepsilon$ für alle $x < X_0$ gilt, dann (und nur dann) heißt f **konvergent** gegen $L \in \mathbb{R}$ für $x \rightarrow -\infty$. Schreibweise:

$$\lim_{x \rightarrow -\infty} f(x) = L.$$

Ab einer Stelle X_0 müssen damit also alle Funktionswerte in einem Streifen mit Radius ε um den Grenzwert L liegen.

Dies entspricht genau der Definition der Folgenkonvergenz, wobei $n \rightarrow \infty$ durch $x \rightarrow \infty$ bzw. $x \rightarrow -\infty$ ersetzt ist und man „mehr“ Funktionswerte als Folgenglieder zu betrachten hat.

Beispiel 2.34

Die Funktion $f(x) = \frac{1}{x}$ strebt sowohl für $x \rightarrow \infty$ als auch für $x \rightarrow -\infty$ gegen 0:

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

Man schreibt hier auch kurz $\lim_{|x| \rightarrow \infty} \frac{1}{x} = \lim_{x \rightarrow \pm \infty} \frac{1}{x} = 0$.

Wir beweisen mit der Definition, dass $\lim_{x \rightarrow \infty} \frac{1}{x} = 0$ ist: Sei $\varepsilon > 0$ und $X_0 := \frac{1}{\varepsilon}$. Für $x > X_0$ ist

$$\left| \frac{1}{x} - 0 \right| = \frac{1}{x} < \frac{1}{X_0} = \varepsilon. \quad \blacksquare$$

Definition 2.13 (Bestimmte Divergenz im Unendlichen)

Es sei $]a, \infty[\subset D$ für ein $a \in \mathbb{R}$. Die Funktion $f : D \rightarrow \mathbb{R}$ heißt **bestimmt divergent gegen ∞** für $x \rightarrow \infty$ genau dann, wenn für jedes (noch so große) $M > 0$ ein $X_0 = X_0(M) \in \mathbb{R}$, $X_0 \geq a$, existiert, so dass für alle $x > X_0$ gilt: $f(x) > M$. Schreibweise:

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

Egal, wie groß man ein $M \in \mathbb{R}$ vorgibt – ab einer Stelle X_0 müssen alle Funktionswerte größer als M sein. Auch dieser Begriff ist so formuliert wie die bestimmte Divergenz bei Folgen.

Man definiert $\lim_{x \rightarrow \infty} f(x) = -\infty$, $\lim_{x \rightarrow -\infty} f(x) = \infty$ und $\lim_{x \rightarrow -\infty} f(x) = -\infty$ analog.

Beispiel 2.35

$\lim_{x \rightarrow \infty} x^2 = \infty$, denn zu $M > 0$ sei $X_0 := \sqrt{M}$. Für $x > X_0$ ist $x^2 > X_0^2 = M$. ■

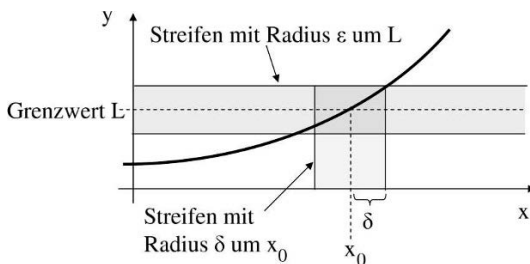


Abb. 2.11: Zur Definition des Grenzwerts $\lim_{x \rightarrow x_0} f(x)$

Definition 2.14 (Konvergenz einer Funktion an einer Stelle)

Sei $f : D \rightarrow \mathbb{R}$, wobei $]x_0 - \delta_0, x_0[\cup]x_0, x_0 + \delta_0[\subset D$, d. h., eine punktierte Umgebung von x_0 ist im Definitionsbereich D von f enthalten, so dass Funktionswerte links und rechts von x_0 erklärt sind.

Die Funktion f heißt **konvergent gegen $L \in \mathbb{R}$** für $x \rightarrow x_0$ genau dann, wenn zu jedem (noch so kleinen) $\varepsilon > 0$ ein $\delta = \delta(\varepsilon, x_0) > 0$, $\delta \leq \delta_0$, existiert, so dass für alle $x \neq x_0$ mit $|x - x_0| < \delta$ gilt: $|f(x) - L| < \varepsilon$. L heißt dann der **Grenzwert** von f für $x \rightarrow x_0$. Schreibweise: $\lim_{x \rightarrow x_0} f(x) = L$.

Man hat genau dann Konvergenz an der Stelle x_0 , wenn man für jeden vorgegebenen Abstand ε zum Grenzwert eine kleine punktierte Umgebung von x_0 findet, in der alle Funktionswerte höchstens den vorgegebenen Abstand zum Grenzwert haben (siehe Abbildung 2.11). Man kommt dem Grenzwert also beliebig nahe, wenn man sich der Stelle x_0 nähert. Mit ε gibt man eine Toleranz vor. Ein zugehöriges δ besagt dann, wie nah man bei x_0 sein muss, damit die Funktionswerte innerhalb der Toleranz vom Grenzwert liegen. Das ist für viele technische Fragestellungen wichtig.

In der Definition wird vermieden, von Zahlen x zu sprechen, die unendlich nah bei x_0 liegen (und die es in $\mathbb{R}$ nicht gibt). Statt eines „unendlich kleinen“ $x - x_0$ werden hier nur wohldefinierte Zahlen ε und δ verwendet, um das Streben gegen x_0 auszudrücken. Eine Alternative dazu bietet die Nichtstandardanalysis, in der eine unendlich kleine Zahl eingesetzt wird. So können ε - δ -Formulierungen vermieden werden. Dafür muss man aber einen nicht unerheblichen theoretischen Aufwand treiben, um die neue Zahl sinnvoll einzuführen.

Zu diesem und den folgenden Konvergenzbegriffen gibt es (abgesehen vom später formulierten Übertragungsprinzip) keine direkte Korrespondenz zur Grenzwertdefinition bei Folgen.

Falls man nicht fordert, dass die Funktion links und rechts von der betrachteten Stelle x_0 erklärt ist und nur verlangt, dass es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt mit $|f(x) - L| < \varepsilon$ für alle $x \in (]x_0 - \delta, x_0[\cup]x_0, x_0 + \delta[) \cap D$, so fordert man eventuell etwas für alle $x \in \emptyset$. Diese Bedingung ist dann aber für jedes L erfüllt, und der Grenzwert wäre nicht mehr eindeutig. Die Situation tritt für Funktionen auf, die in nur einem oder in einzelnen Punkten definiert sind (z. B. $f : \{1\} \rightarrow \mathbb{R}$). Hier macht eine Grenzwertbetrachtung keinen Sinn.

Beispiel 2.36

- a) Die Funktion $f(x) = 3x - 1$ hat an der Stelle $x = 1$ den Grenzwert $\lim_{x \rightarrow 1} f(x) = 2$. Um das zu beweisen, müssen wir zu (jedem beliebigen) $\varepsilon > 0$ ein $\delta(\varepsilon) > 0$ so bestimmen, dass für x mit $|x - 1| < \delta$ folgt $|f(x) - 2| < \varepsilon$. Wegen

$$|x - 1| < \delta \implies |f(x) - 2| = |3x - 1 - 2| = |3x - 3| = 3|x - 1| < 3\delta$$

erfüllt z. B. $\delta := \frac{\varepsilon}{3}$ diese Bedingung, und der Funktionengrenzwert ist nachgewiesen. Für $\varepsilon = \frac{1}{100}$ und damit $\delta = \frac{1}{300}$ wissen wir jetzt, dass

$$|x - 1| < \frac{1}{300} \implies |f(x) - 2| < \frac{1}{100}.$$

- b) Die Funktion $f(x) = x \cdot \sin\left(\frac{1}{x}\right)$ besitzt bei $x = 0$ den Grenzwert $\lim_{x \rightarrow 0} f(x) = 0$. Um zu einem $\varepsilon > 0$ ein δ wie in der Definition zu finden, betrachten wir

$$|f(x) - 0| = \left| x \cdot \sin\left(\frac{1}{x}\right) \right| = |x| \left| \sin\left(\frac{1}{x}\right) \right| \leq |x| \cdot 1 = |x - 0|.$$

Wenn wir also $\delta := \varepsilon$ wählen, so gilt $|x - 0| < \delta = \varepsilon \implies |f(x) - 0| < \varepsilon$, und der Grenzwert ist bewiesen.

- c) Die gebrochen-rationale Funktion $f(x) = \frac{2x^2-2x}{x-1}$, $x \in \mathbb{R} \setminus \{1\}$, besitzt an der Stelle 1 den Grenzwert $\lim_{x \rightarrow 1} f(x) = 2$. Um zu einem $\varepsilon > 0$ ein geeignetes δ zu finden, betrachten wir wieder die Differenz der Funktionswerte vom Grenzwert für Argumente nahe bei 1:

$$\begin{aligned} |x-1| < \delta \quad \implies \quad |f(x) - 2| &= \left| \frac{2x^2-2x}{x-1} - 2 \right| = \left| \frac{2x^2-2x-2x+2}{x-1} \right| \\ &= 2 \left| \frac{(x-1)^2}{x-1} \right| = 2|x-1| < 2\delta. \end{aligned}$$

Für $\delta = \frac{\varepsilon}{2}$ gilt demnach $|x-1| < \delta \implies |f(x) - 2| < \varepsilon$.

- d) Die Wurzelfunktion $f(x) = \sqrt{x}$ besitzt an jeder Stelle $x_0 > 0$ den Grenzwert $\lim_{x \rightarrow x_0} f(x) = \sqrt{x_0}$. Um hier zu $\varepsilon > 0$ ein geeignetes δ zu finden, wenden wir einen Trick an: Wir benutzen die dritte Binomische Formel. Damit erhalten wir für $x > 0$:

$$\begin{aligned} |x - x_0| < \delta \quad \implies \quad |f(x) - f(x_0)| &= |\sqrt{x} - \sqrt{x_0}| = |\sqrt{x} - \sqrt{x_0}| \cdot \frac{|\sqrt{x} + \sqrt{x_0}|}{|\sqrt{x} + \sqrt{x_0}|} \\ &= \frac{|x - x_0|}{\sqrt{x} + \sqrt{x_0}} < \frac{\delta}{\sqrt{x} + \sqrt{x_0}} < \frac{\delta}{\sqrt{x_0}}. \end{aligned}$$

Für $\delta := \min\{x_0, \varepsilon\sqrt{x_0}\} > 0$ folgt aus $|x - x_0| < \delta$ einerseits, dass $x > 0$ ist und andererseits $|\sqrt{x} - \sqrt{x_0}| < \varepsilon$. ■

Beispiel 2.37

Für $n \in \mathbb{N}_0$ ist

$$\lim_{x \rightarrow x_0} x^n = x_0^n.$$

Der Grenzwert an der Stelle x_0 ist $L = x_0^n$: Für $n = 0$ ist $|x^0 - x_0^0| = |1 - 1| = 0 < \varepsilon$ für alle $\varepsilon > 0$ unabhängig von x und x_0 , so dass die Grenzwertaussage trivial folgt. Für $n > 0$ verwenden wir die Darstellung $x = x_0 + h$ mit $|h| < 1$ für Zahlen $x \in]x_0 - 1, x_0 + 1[$:

$$\begin{aligned} |x_0^n - (x_0 + h)^n| &\stackrel{(1.11)}{=} \left| x_0^n - \sum_{k=0}^n \binom{n}{k} h^k x_0^{n-k} \right| = \left| \sum_{k=1}^n \binom{n}{k} h^k x_0^{n-k} \right| \\ &= |h| \left| \sum_{k=1}^n \binom{n}{k} h^{k-1} x_0^{n-k} \right| \stackrel{|h| \leq 1}{\leq} |h| \underbrace{\left| \sum_{k=1}^n \binom{n}{k} |x_0|^{n-k} \right|}_{\text{unabhängig von } h} \\ &= |h| \left[-|x_0|^n + \sum_{k=0}^n \binom{n}{k} 1^k |x_0|^{n-k} \right] \stackrel{(1.11)}{=} |h| \left((1 + |x_0|)^n - |x_0|^n \right). \end{aligned}$$

Zu $\varepsilon > 0$ sei (beachte $(1 + |x_0|)^n - |x_0|^n > 0$)

$$\delta := \min \left\{ 1, \frac{\varepsilon}{(1 + |x_0|)^n - |x_0|^n} \right\}.$$

Für $|x - x_0| < \delta$ ist $|h| = |x - x_0| < \delta$ und damit

$$|x_0^n - x^n| = |x_0^n - (x_0 + h)^n| < \delta \left((1 + |x_0|)^n - |x_0|^n \right) \leq \varepsilon,$$

so dass die Konvergenz gegen $L = x_0^n$ gezeigt ist. ■

Beispiel 2.38

$\lim_{x \rightarrow x_0} \sin x = \sin x_0$, denn wegen $|\sin t| \leq |t|$ ist für $|x - x_0| < \delta := \varepsilon$:

$$|\sin x - \sin x_0| \stackrel{(1.29)}{=} 2 \left| \cos \frac{x + x_0}{2} \right| \left| \sin \frac{x - x_0}{2} \right| \leq 2 \left| \sin \frac{x - x_0}{2} \right| \leq |x - x_0| < \delta = \varepsilon.$$

Damit erhalten wir auch

$$\lim_{x \rightarrow x_0} \cos x = \lim_{x \rightarrow x_0} \sin \left(x + \frac{\pi}{2} \right) = \lim_{x \rightarrow x_0 + \frac{\pi}{2}} \sin(x) = \sin \left(x_0 + \frac{\pi}{2} \right) = \cos(x_0). \quad \blacksquare$$

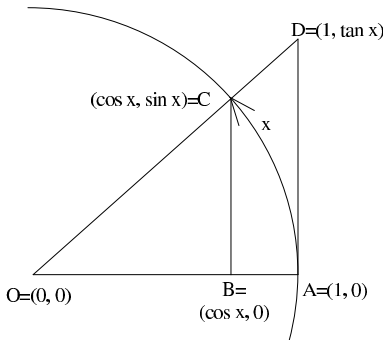


Abb. 2.12: Herleitung zu $\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$ am Kreis mit Radius 1

Beispiel 2.39

Für den Nachweis der Differenzierbarkeit von $\sin x$ und im Rahmen der Fourier-Analyse benötigen wir später den Grenzwert

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1. \quad (2.14)$$

Um diesen nachzurechnen, betrachten wir für $0 < |x| < \pi/2$ am Einheitskreis die Punkte $O = (0, 0)$, $A = (1, 0)$, $B = (\cos x, 0)$, $C = (\cos x, \sin x)$ und $D = (1, \tan x)$. Die Länge $|\sin x|$ der Gegenkathete zum Winkel x im Dreieck mit den Ecken O , B und C ist nicht größer als die Bogenlänge $|x|$, d. h., wie auf Seite 114 lesen wir $|\sin(x)| \leq |x|$ ab. Das Kreissegment zu den Punkten O , A und C hat mit Radius $r = 1$ den Flächeninhalt $\pi r^2 \frac{|x|}{2\pi} = \frac{|x|}{2}$. Dieses Segment liegt wiederum im Dreieck mit den Ecken O , A und D . (Dabei beachte man, dass der Punkt D auf der Geraden durch O und C liegt, die die Steigung $\frac{\sin x}{\cos x} = \frac{\tan x}{1}$ besitzt.) Das Dreieck mit den Ecken O , A und D hat den Flächeninhalt $\frac{|\tan x|}{2}$. Damit gilt also:

$$\frac{|x|}{2} \leq \frac{|\tan x|}{2} \iff |x| \leq \frac{|\sin x|}{\cos x} \iff \cos x \leq \frac{|\sin x|}{|x|} = \frac{\sin x}{x}.$$

Zusammen mit $|\sin(x)| \leq |x|$ wird daraus

$$\cos x \leq \frac{\sin x}{x} \leq 1.$$

Damit ist

$$\left| 1 - \frac{\sin x}{x} \right| = 1 - \frac{\sin x}{x} \leq 1 - \cos x.$$

Wir haben bereits gesehen, dass $\lim_{x \rightarrow 0} \cos x = \cos 0 = 1$. Zu jedem $\varepsilon > 0$ gibt es also ein $0 < \delta < \pi/2$, so dass $1 - \cos x = |1 - \cos x| < \varepsilon$ für alle $0 < |x| < \delta$. Dafür ist dann aber auch $\left| 1 - \frac{\sin x}{x} \right| < \varepsilon$. ■

Beispiel 2.40 (Dirichlet-Funktion)

Ein klassisches Beispiel für eine Funktion, bei der der Grenzwert an keiner Stelle existiert, ist die **Dirichlet-Funktion**. Diese Funktion hat für alle rationalen x den Funktionswert 1 und für alle irrationalen den Funktionswert 0. Wenn man versucht, den Funktionsgraphen zu zeichnen, dann entstehen zwei horizontale Linien (da man die unendlich vielen Unterbrechungen nicht sehen kann). In jeder Umgebung einer Stelle sind also unendlich viele Funktionswerte gleich 1 und unendlich viele gleich 0, die nicht gemeinsam in einem Streifen zu $\varepsilon \leq \frac{1}{2}$ liegen können. ■

Definition 2.15 (Bestimmte Divergenz einer Funktion an einer Stelle)

Sei $]x_0 - \delta_0, x_0[\cup]x_0, x_0 + \delta_0[\subset D$. Die Funktion $f : D \rightarrow \mathbb{R}$ heißt **bestimmt divergent** gegen ∞ für $x \rightarrow x_0$ genau dann, wenn für jedes (noch so große) $M > 0$ ein $\delta = \delta(M, x_0) > 0$, $\delta \leq \delta_0$, existiert, so dass für alle $x \neq x_0$ mit $|x - x_0| < \delta$ gilt: $f(x) > M$. Schreibweise: $\lim_{x \rightarrow x_0} f(x) = \infty$.

Analog definiert man $\lim_{x \rightarrow x_0} f(x) = -\infty$.

$\lim_{x \rightarrow x_0} f(x) = \infty$ bedeutet also, dass man mit den Funktionswerten über jede noch so große Zahl M kommt, wenn man nur nah genug bei der Stelle x_0 ist.

Beispiel 2.41

Während $\lim_{x \rightarrow 0} \frac{1}{|x|} = \infty$ ist, ist die Funktion $\frac{1}{x}$ für $x_0 = 0$ divergent, aber nicht bestimmt divergent. Beachte: Zwar ist $\lim_{x \rightarrow 8} \frac{1}{|8-x|} = \infty$, aber es gilt nicht $\lim_{x \rightarrow 4} \frac{1}{|4-x|} = \infty$. ■

Auch wenn der Grenzwert an einer Stelle x_0 nicht existiert, so ist es bisweilen sinnvoll, sogenannte einseitige Grenzwerte zu betrachten.

Definition 2.16 (Einseitige Konvergenz einer Funktion an einer Stelle)

- Es sei die linksseitige punktierte Umgebung $]x_0 - \delta_0, x_0[$ in D enthalten. Die Funktion $f : D \rightarrow \mathbb{R}$ heißt **linksseitig konvergent** gegen $L \in \mathbb{R}$ bzw. **linksseitig bestimmt divergent** gegen ∞ bzw. $-\infty$ für $x \rightarrow x_0 -$ genau dann, wenn die Bedingungen aus Definition 2.14 bzw. 2.15 gelten, wenn man dort die beidseitige punktierte Umgebung $]x_0 - \delta_0, x_0[\cup]x_0, x_0 + \delta_0[$ durch $]x_0 - \delta_0, x_0[$ austauscht und die Bedingung $|x - x_0| < \delta$ durch $0 < x_0 - x < \delta$ ersetzt. L heißt dann der **linksseitige Grenzwert** von f an der Stelle x_0 . Schreibweisen:

$$\lim_{x \rightarrow x_0 -} f(x) = L, \quad \lim_{x \rightarrow x_0 -} f(x) = \infty, \quad \lim_{x \rightarrow x_0 -} f(x) = -\infty.$$

- Es sei die rechtsseitige punktierte Umgebung $]x_0, x_0 + \delta_0[\subset D$. Die Funktion $f : D \rightarrow \mathbb{R}$ heißt genau dann **rechtsseitig konvergent** gegen $L \in \mathbb{R}$ bzw. **rechtsseitig bestimmt divergent** gegen ∞ bzw. $-\infty$ für $x \rightarrow x_0 +$, falls die Bedingungen aus Definition 2.14 bzw. 2.15 gelten, wenn man die beidseitige punktierte Umgebung $]x_0 - \delta_0, x_0[\cup]x_0, x_0 + \delta_0[$ durch $]x_0, x_0 + \delta_0[$ austauscht und die Bedingung $|x - x_0| < \delta$ durch $0 < x - x_0 < \delta$ ersetzt. L heißt dann der **rechtsseitige Grenzwert** von f an der Stelle x_0 . Schreibweisen:

$$\lim_{x \rightarrow x_0 +} f(x) = L, \quad \lim_{x \rightarrow x_0 +} f(x) = \infty, \quad \lim_{x \rightarrow x_0 +} f(x) = -\infty.$$

Hier nähert man sich ausschließlich von links oder von rechts der Stelle x_0 . Dabei gibt man mittels „+“ (rechts) oder „-“ (links) die Richtung an, aus der man kommt. Das entspricht der Windrichtung. Ein Westwind kommt aus Westen und weht nach Osten.

Beispiel 2.42

a) $\lim_{x \rightarrow 0+} \frac{1}{x} = \infty, \quad \lim_{x \rightarrow 0-} \frac{1}{x} = -\infty.$

b) Sei $f(x) := \begin{cases} 1 & \text{für } x \geq 5 \\ -1 & \text{für } x < 5. \end{cases}$

Dann ist $\lim_{x \rightarrow 5+} f(x) = 1$ und $\lim_{x \rightarrow 5-} f(x) = -1$.

c) Die Funktion $f(x) = \frac{2x^2 - 4x}{x - 2}$ besitzt bei $x = 2$ übereinstimmende einseitige Grenzwerte, da $\frac{2x^2 - 4x}{x - 2} = 2x$ und $\lim_{x \rightarrow 2-} 2x = \lim_{x \rightarrow 2+} 2x = \lim_{x \rightarrow 2} 2x = 4$.

d) Die Funktion

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

(siehe Seite 89) besitzt in $x = 0$ die einseitigen Grenzwerte $\lim_{x \rightarrow 0-} \text{sign}(x) = -1$ und $\lim_{x \rightarrow 0+} \text{sign}(x) = +1$. Ein Grenzwert existiert bei 0 nicht. ■

Lemma 2.2 (Konvergenz und einseitige Konvergenz)

Sei $f : D \rightarrow \mathbb{R}$ und $]x_0 - \delta_0, x_0[\cup]x_0, x_0 + \delta_0[\subset D$. Dann sind äquivalent:

- a) $\lim_{x \rightarrow x_0} f(x) = L$.
- b) $\lim_{x \rightarrow x_0-} f(x) = \lim_{x \rightarrow x_0+} f(x) = L$, d. h., die einseitigen Grenzwerte existieren und sind gleich.

Konvergenz an einer Stelle bedeutet also nicht nur, dass rechts- und linksseitige Konvergenz vorliegt, sondern zusätzlich, dass die einseitigen Grenzwerte auch übereinstimmen.

Beweis: Aus a) folgt sofort b), da mit dem nach Definition zu einem $\varepsilon > 0$ existierenden $\delta > 0$ sowohl die Bedingung des rechts- als auch des linksseitigen Grenzwerts erfüllt ist. Umgekehrt folgt aus b) die Aussage a), da zu einem $\varepsilon > 0$ aufgrund der einseitigen Grenzwerte ein $\delta_1 > 0$ und ein $\delta_2 > 0$ existiert mit

$$|f(x) - L| < \varepsilon$$

für alle $x \in D$ mit $0 < x_0 - x < \delta_1$ oder $0 < x - x_0 < \delta_2$. Damit gilt dies aber auch für alle $x \in D$ mit $|x - x_0| < \delta := \min\{\delta_1, \delta_2\}$ und $x \neq x_0$. $\square$

Bei Folgen haben wir mit Satz 2.3 gesehen, dass eine monotone, beschränkte Folge zwangsläufig konvergent sein muss. Denn ist sie z. B. nach oben beschränkt und legt man einen ε -Streifen um die kleinste obere Schranke, so liegen wegen der Monotonie alle Folgenglieder, die einem Glied im Streifen folgen, auch in diesem. Analog verhält es sich mit der Konvergenz einer monotonen Funktion an einer Stelle x_0 :

Satz 2.21 (Existenz der einseitigen Grenzwerte bei Monotonie *)

Ist f monoton (siehe Definition 1.25 auf Seite 81) auf einem Intervall I , so existiert in jedem Punkt $x_0 \in I$ (mit einer rechts- bzw. linksseitigen Umgebung in I) der rechts- bzw. linksseitige Grenzwert.

Die einseitigen Grenzwerte müssen aber nicht an jeder Stelle übereinstimmen, obwohl sie es sogar in der Regel tun (was wir auf Seite 286 sehen werden).

Beweis: Wir zeigen die Aussage für einen linksseitigen Grenzwert an einer Stelle x_0 mit $]x_0 - \delta, x_0[\subset I$ für eine monoton steigende Funktion. Die anderen Situationen verhalten sich analog. Da f monoton steigend ist, ist die nicht-leere Menge $\{f(x) : x \in]x_0 - \delta, x_0[\}$ nach oben beschränkt mit $f(x_0)$, so dass nach Axiom (1.5) das Supremum existiert. Wir zeigen, dass

$$\lim_{x \rightarrow x_0-} f(x) = L := \sup\{f(x) : x \in]x_0 - \delta, x_0[\},$$

indem wir erneut die Monotonie ausnutzen. Wegen (1.6) auf Seite 50 existiert zunächst zu jedem $\varepsilon > 0$ eine Stelle $x_1 \in]x_0 - \delta, x_0[$, deren Funktionswert näher als ε beim Supremum liegt, also $L - f(x_1) < \varepsilon$. Wegen der Monotonie gilt nun aber für alle $x \in [x_1, x_0[$ ebenfalls $|L - f(x)| = L - f(x) \leq L - f(x_1) < \varepsilon$. Zu ε gibt es also ein $\delta = x_0 - x_1$, so dass alle Funktionswerte der linksseitigen punktierten δ -Umgebung von x_0 näher als ε bei L liegen. Damit ist die linksseitige Konvergenz gegen L gezeigt. $\square$

Jetzt kommen wir zu **dem** Zusammenhang zwischen Folgenkonvergenz und Grenzwerten von Funktionen: Wir können den Grenzwert einer Funktion auch schreiben als Grenzwert von Folgen:

Satz 2.22 (Übertragungsprinzip)

Seien f eine Funktion von D in $\mathbb{R}$, $]x_0 - \delta_0, x_0[\cup]x_0, x_0 + \delta_0[\subset D$ und $L \in \mathbb{R}$. Dann sind äquivalent:

- a) $\lim_{x \rightarrow x_0} f(x) = L$.
- b) Für jede Folge $(x_n)_{n=1}^{\infty}$ mit $x_n \in D$ und $x_n \neq x_0$ für alle $n \in \mathbb{N}$ sowie $\lim_{n \rightarrow \infty} x_n = x_0$ gilt: $\lim_{n \rightarrow \infty} f(x_n) = L$.

Für Grenzwerte gegen $\pm\infty$, einseitige Grenzwerte und bestimmte Divergenz gelten entsprechende Aussagen.

Die Einschränkung $x_n \neq x_0$ im Satz wird nur für den Fall benötigt, dass f in x_0 nicht definiert ist oder dass $L \neq f(x_0)$ ist.

Konvergenz an einer Stelle x_0 hat man also genau dann, wenn man sich entlang jeder Folge der Stelle x_0 nähern kann und dabei die Funktionswerte gegen einen gemeinsamen Grenzwert streben.

Die Bedingung b) verlangt die Konvergenz für **jede** Folge mit Grenzwert x_0 . Es reicht nicht aus, wenn Konvergenz für eine spezielle Folge vorliegt. Dagegen ist die Bedingung verletzt, wenn man Divergenz der Funktionswerte zu einer Folge oder verschiedene Grenzwerte der Funktionswerte für zwei verschiedene Folgen nachrechnen kann.

Beweis: b) folgt aus a) durch Ineinanderschachteln der Definitionen. Um zu sehen, dass a) aus b) folgt, führen wir einen indirekten Beweis. Wir nehmen an, a) gelte nicht. Dann gibt es ein $\varepsilon > 0$, so dass für jedes $\delta = 1/n$ ein $x_n \in D \setminus \{x_0\}$ mit $|x_0 - x_n| < 1/n$ existiert, so dass $|f(x_n) - L| > \varepsilon$. Die Folge $(x_n)_{n=1}^{\infty}$ konvergiert nach Konstruktion gegen x_0 , aber $(f(x_n))_{n=1}^{\infty}$ konvergiert nicht gegen L . Damit gilt b) nicht. Wenn also b) zutrifft, dann muss auch a) gelten. $\square$

Wir können jetzt den Grenzwert einer Funktion durch Folgengrenzwerte ausdrücken.

Beispiel 2.43

Mit dem Übertragungsprinzip untersuchen wir die gebrochen-rationale Funktion $f(x) = \frac{(x-2)(x+2)}{(x-2)(x+3)}$ an den Nullstellen 2 und -3 des Nenners. Bei der Grenzwertberechnung für eine beliebige Folge $(x_n)_{n=1}^\infty$ mit $\lim_{n \rightarrow \infty} x_n = 2$ und $x_n \neq 2$ kann der Faktor $(x_n - 2)$ gekürzt werden, und man erhält

$$\lim_{n \rightarrow \infty} \frac{(x_n - 2)(x_n + 2)}{(x_n - 2)(x_n + 3)} = \lim_{n \rightarrow \infty} \frac{x_n + 2}{x_n + 3} = \frac{2 + 2}{2 + 3} = \frac{4}{5}.$$

Da dies für jede gegen 2 konvergente Folge gilt, ist $\frac{4}{5}$ auch der Grenzwert der Funktion an der Stelle 2. Bei $x = -3$ existiert kein reeller Grenzwert, da der Zähler für $x = -3$ nicht null wird. Links- und rechtsseitiger Grenzwert sind uneigentlich $-\infty$ bzw. $+\infty$. Wir zeigen dies nur für den linksseitigen Grenzwert $x \rightarrow -3-$. Dazu wählen wir eine Folge $(x_n)_{n=1}^\infty$ mit $\lim_{n \rightarrow \infty} x_n = -3$ und $x_n < -3$. Diese lässt sich auch schreiben als $x_n = -3 - \varepsilon_n$, wobei $(\varepsilon_n)_{n=1}^\infty$ eine Nullfolge positiver Zahlen ist. Es gilt dann

$$\lim_{n \rightarrow \infty} \frac{(x_n - 2)(x_n + 2)}{(x_n - 2)(x_n + 3)} = \lim_{n \rightarrow \infty} \frac{x_n + 2}{x_n + 3} = \lim_{n \rightarrow \infty} \frac{(-3 - \varepsilon_n) + 2}{(-3 - \varepsilon_n) + 3} = \lim_{n \rightarrow \infty} \frac{-1 - \varepsilon_n}{-\varepsilon_n} = \infty.$$

■

Beispiel 2.44

Mit dem Übertragungsprinzip zeigen wir, dass $\lim_{x \rightarrow 0+} \sin \frac{1}{x}$ nicht existiert (siehe Abbildung 2.13). Dazu betrachten wir die Nullfolge $a_n = \frac{1}{(2n+1)\frac{\pi}{2}}$, die sich der Null von rechts nähert:

$$\sin \frac{1}{a_n} = \sin \left((2n+1) \frac{\pi}{2} \right) = (-1)^n.$$

Die Folge der Funktionswerte $\left(\sin \frac{1}{a_n} \right)_{n=1}^\infty$ hat also (mindestens) zwei Folgehäufungspunkte -1 und 1 und ist damit nicht konvergent. Nach Übertragungsprinzip existiert somit auch $\lim_{x \rightarrow 0+} \sin \frac{1}{x}$ nicht. ■

Für Folgen haben wir bereits Konvergenzaussagen bewiesen, die sich mit dem Übertragungsprinzip auf Funktionen übertragen lassen. Dazu gehört zunächst, dass der Grenzwert eindeutig ist. Weiterhin erhalten wir direkt Regeln für Linearkombinationen, Produkte und (sofern definiert) Quotienten von Funktionen:

Folgerung 2.2 (Grenzwertsätze)

Unter den Voraussetzungen des Übertragungsprinzips gilt:

$$\lim_{x \rightarrow x_0} (cf(x) + dg(x)) = c \lim_{x \rightarrow x_0} f(x) + d \lim_{x \rightarrow x_0} g(x),$$

$$\lim_{x \rightarrow x_0} f(x)g(x) = \left[\lim_{x \rightarrow x_0} f(x) \right] \cdot \left[\lim_{x \rightarrow x_0} g(x) \right], \quad \lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \frac{\lim_{x \rightarrow x_0} f(x)}{\lim_{x \rightarrow x_0} g(x)}, \quad (2.15)$$

wobei für die letzte Gleichung der Grenzwert $\lim_{x \rightarrow x_0} g(x)$ ungleich null sein muss, damit nicht durch null dividiert wird.

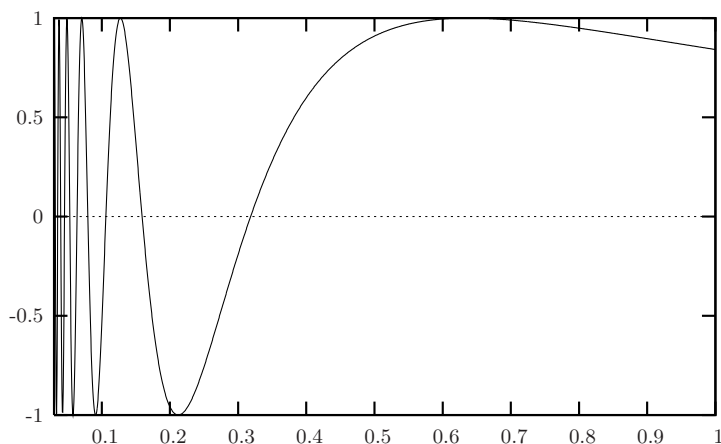


Abb. 2.13: $f(x) = \sin \frac{1}{x}$

Beispiel 2.45

$$\lim_{x \rightarrow 0} \frac{3x^2 - 2x + 8}{2 \cos x} = \frac{\lim_{x \rightarrow 0} (3x^2 - 2x + 8)}{\lim_{x \rightarrow 0} 2 \cos x} = \frac{3 \lim_{x \rightarrow 0} x^2 - 2 \lim_{x \rightarrow 0} x + \lim_{x \rightarrow 0} 8}{2 \lim_{x \rightarrow 0} \cos x} = \frac{8}{2} = 4.$$

■

Ebenso wie an einer Stelle x_0 gelten die Grenzwertsätze für Grenzwerte gegen $\pm\infty$ und für einseitige Grenzwerte.

Beispiel 2.46

a) Wir berechnen den Funktionengrenzwert $\lim_{x \rightarrow \infty} \frac{4x^2 + 4}{2x^2 - 1}$:

$$\lim_{x \rightarrow \infty} \frac{4x^2 + 4}{2x^2 - 1} = \lim_{x \rightarrow \infty} \frac{x^2 \cdot (4 + \frac{4}{x^2})}{x^2 \cdot (2 - \frac{1}{x^2})} = \frac{\lim_{x \rightarrow \infty} (4 + \frac{4}{x^2})}{\lim_{x \rightarrow \infty} (2 - \frac{1}{x^2})} = \frac{4}{2} = 2.$$

b) Für $f(x) = \frac{x^2 - 4}{x^2 - 8x + 12}$ gilt:

$$\begin{aligned} \lim_{x \rightarrow \pm\infty} f(x) &= \lim_{x \rightarrow \pm\infty} \frac{x^2(1 - \frac{4}{x^2})}{x^2(1 - \frac{8}{x} + \frac{12}{x^2})} = \frac{1 - \lim_{x \rightarrow \pm\infty} \frac{4}{x^2}}{1 - \lim_{x \rightarrow \pm\infty} \frac{8}{x} + \lim_{x \rightarrow \pm\infty} \frac{12}{x^2}} \\ &= \frac{1 - 0}{1 - 0 + 0} = 1. \end{aligned}$$

c) Wir berechnen den Grenzwert für $x \rightarrow -\infty$ für $g(x) = \frac{2|x-1|}{x^2 + 2x - 3}$.

$$\lim_{x \rightarrow -\infty} g(x) = \lim_{x \rightarrow -\infty} \frac{-x^2 (\frac{2}{x} - \frac{2}{x^2})}{x^2 (1 + \frac{2}{x} - \frac{3}{x^2})} = -\frac{\lim_{x \rightarrow -\infty} (\frac{2}{x} - \frac{2}{x^2})}{1 + \lim_{x \rightarrow -\infty} (\frac{2}{x} - \frac{3}{x^2})} = \frac{0}{1} = 0.$$

■

Wie bei Folgen kann man mit den folgenden Regeln direkt die Grenzwerte von ganz-rationalen Funktionen angeben (vgl. (2.10)):

$$\lim_{x \rightarrow \infty} \frac{a_k x^k + a_{k-1} x^{k-1} + \cdots + a_1 x + a_0}{b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0} = \begin{cases} \frac{a_k}{b_m} & \text{falls } k = m, \\ 0 & \text{falls } k < m, \\ +\infty & \text{falls } k > m \text{ und } \frac{a_k}{b_m} > 0, \\ -\infty & \text{falls } k > m \text{ und } \frac{a_k}{b_m} < 0. \end{cases}$$

Auch im Fall $\lim_{x \rightarrow x_0} g(x) = 0$ sind Aussagen für $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)}$ möglich. Dabei helfen die Differenzialrechnung (siehe Seite 323 ff.) oder geschickte Umformungen.

Mit dem Übertragungsprinzip können wir Satz 2.1 d) umformulieren:

Folgerung 2.3 (Abschätzung des Grenzwerts)

Ist $f(x) \leq C$ in einer Umgebung von x_0 und existiert $\lim_{x \rightarrow x_0} f(x)$, so ist $\lim_{x \rightarrow x_0} f(x) \leq C$, entsprechend ist $\lim_{x \rightarrow x_0} f(x) \geq c$, falls $f(x) \geq c$ in einer Umgebung von x_0 ist.

Diese Regeln gelten analog für einseitige Grenzwerte und Grenzwerte gegen $\pm\infty$.

Beispiel 2.47

Allgemeine reelle Potenzen haben wir mit (2.9) auf Seite 229 erklärt, indem wir den Exponenten durch Folgen rationaler Zahlen angenähert haben. Damit bietet sich das Übertragungsprinzip an, wenn wir Grenzwerte von Potenzfunktionen berechnen wollen: Wir zeigen damit, dass

$$\lim_{x \rightarrow x_0} \exp x = \exp x_0. \quad (2.16)$$

Sei dazu $(x_n)_{n=1}^\infty$ eine beliebige Folge reeller Zahlen mit $\lim_{n \rightarrow \infty} x_n = x_0$. Wegen des Übertragungsprinzips genügt es, zu zeigen, dass $\lim_{n \rightarrow \infty} \exp x_n = \exp x_0$ ist. Wäre $(x_n)_{n=1}^\infty$ eine Folge rationaler Zahlen, wäre das nach Definition der reellen Potenz bereits gezeigt. Nicht-rationale x_n ersetzen wir jetzt durch rationale Zahlen. Dazu betrachten wir nun ein fest gewähltes Folgenglied x_n .

Da $\mathbb{Q}$ dicht in $\mathbb{R}$ liegt, gibt es zu x_n eine Folge $(a_{n,k})_{k=1}^\infty$ in $\mathbb{Q}$ mit $\lim_{k \rightarrow \infty} a_{n,k} = x_n$, d. h., zu $\varepsilon = \frac{1}{n}$ gibt es ein $n_1 = n_1(x_n) \in \mathbb{N}$, so dass $|a_{n,k} - x_n| < \frac{1}{n}$ für alle $k \geq n_1$.

Aufgrund der Definition der reellen Potenz über eine Folge rationaler Exponenten (2.9) gilt: $\lim_{k \rightarrow \infty} e^{a_{n,k}} = e^{x_n}$. Es gibt also ein $n_2 = n_2(x_n) \in \mathbb{N}$, so dass $|e^{a_{n,k}} - e^{x_n}| < \frac{1}{n}$ für alle $k \geq n_2$. Wir setzen $a_n := a_{n, \max\{n_1, n_2\}} \in \mathbb{Q}$ und erhalten damit

$$|a_n - x_n| < \frac{1}{n}, \quad |e^{a_n} - e^{x_n}| < \frac{1}{n}.$$

Wegen $|a_n - x_0| = |a_n - x_n + x_n - x_0| \leq |a_n - x_n| + |x_n - x_0| < \frac{1}{n} + |x_n - x_0| \rightarrow 0$, $n \rightarrow \infty$, ist außerdem $\lim_{n \rightarrow \infty} a_n = x_0$ und wieder nach Definition der reellen Potenz $\lim_{n \rightarrow \infty} e^{a_n} = e^{x_0}$. Insgesamt erhalten wir:

$$|\exp x_n - \exp x_0| \leq |e^{x_n} - e^{a_n}| + |e^{a_n} - e^{x_0}| < \frac{1}{n} + |e^{a_n} - e^{x_0}| \rightarrow 0, \quad n \rightarrow \infty. \quad \blacksquare$$

2.3.3 Stetigkeit

Stetigkeit einer Funktion bedeutet, dass man den Funktionsgraphen auf einem Intervall des Definitionsbereichs ohne Absetzen des Stiftes durchzeichnen kann (vgl. Satz 2.25). Streng mathematisch ist dies nicht äquivalent zur Stetigkeit (sondern nur eine Folgerung aus dem Zwischenwertsatz), diese Interpretation reicht aber für die praktische Anwendung völlig aus.

Nähert man sich einer Stelle, müssen also die Funktionswerte gegen den Funktionswert an dieser Stelle streben. In diesem Sinne wird die Stetigkeit zunächst auch für eine einzelne Stelle (für einen einzelnen Punkt) definiert. Stetigkeit ist daher ein punktwiser Begriff:

Definition 2.17 (Stetigkeit)

Sei $f : D \rightarrow \mathbb{R}$ mit $D \subset \mathbb{R}$ und $x_0 \in D$.

- Sei x_0 ein innerer Punkt von D , d. h., es gibt eine Umgebung $]x_0 - \delta_0, x_0 + \delta_0[\subset D$. f heißt genau dann **stetig** im Punkt x_0 , wenn $\lim_{x \rightarrow x_0} f(x) = f(x_0)$.
- Zu x_0 sei die linksseitige Umgebung $]x_0 - \delta_0, x_0] \subset D$. f heißt genau dann **linksseitig stetig** im Punkt x_0 , wenn $\lim_{x \rightarrow x_0-} f(x) = f(x_0)$.
- Zu x_0 sei die rechtsseitige Umgebung $[x_0, x_0 + \delta_0[\subset D$. f heißt genau dann **rechtsseitig stetig** im Punkt x_0 , wenn $\lim_{x \rightarrow x_0+} f(x) = f(x_0)$.
- Ist D ein abgeschlossenes, offenes oder einseitig abgeschlossenes Intervall $[a, b]$, $]a, b[$, $[a, b[$ oder $]a, b]$, so heißt f genau dann **stetig** auf D , wenn f in jedem $x \in]a, b[$ stetig ist und, falls $a \in D$, f rechtsseitig stetig in a und falls $b \in D$, f linksseitig stetig in b ist. Entsprechend definiert man die Stetigkeit auf $[a, \infty[$, $]a, \infty[$, $] - \infty, b]$, $] - \infty, b[$ und $] - \infty, \infty[= \mathbb{R}$.

Ist f stetig in x_0 , so stimmt der Funktionswert $f(x_0)$ mit dem Grenzwert $\lim_{x \rightarrow x_0} f(x)$ überein. Das ist insbesondere hilfreich beim Berechnen von Grenzwerten. Weiß man, dass f stetig in x_0 ist, so gilt:

$$\lim_{x \rightarrow x_0} f(x) = f(x_0) = f \left(\lim_{x \rightarrow x_0} x \right). \quad (2.17)$$

Der Grenzwert und die Berechnung des Funktionswerts können vertauscht werden! Das werden wir noch häufig bei der Berechnung von Grenzwerten ausnutzen.

Beispiel 2.48

- a) Polynome $p_n : \mathbb{R} \rightarrow \mathbb{R}$ mit $p_n(x) = a_0 + a_1x + \cdots + a_nx^n$ und $a_0, a_1, \dots, a_n \in \mathbb{R}$ sind stetig auf $\mathbb{R}$. Denn für jede Stelle $x_0 \in \mathbb{R}$ gilt mit den Grenzwertregeln

$$\begin{aligned} \lim_{x \rightarrow x_0} (a_0 + a_1x + \cdots + a_nx^n) &= \lim_{x \rightarrow x_0} a_0 + \lim_{x \rightarrow x_0} a_1x + \cdots + \lim_{x \rightarrow x_0} a_nx^n \\ &= a_0 + a_1x_0 + \cdots + a_nx_0^n = p_n(x_0). \end{aligned}$$

- b) Die Betragsfunktion $f(x) = |x| = \begin{cases} x, & x \geq 0, \\ -x, & x < 0, \end{cases}$ ist stetig auf $\mathbb{R}$. Die Stetigkeit

an Stellen $x_0 \neq 0$ folgt aus der Stetigkeit des Polynoms $p_1(x) = x$ bzw. $p_1(x) = -x$. Für $x_0 = 0$ müssen wir den Grenzwert mit dem Funktionswert $f(0) = 0$ vergleichen: $\lim_{x \rightarrow 0+} |x| = \lim_{x \rightarrow 0+} x = 0$ und $\lim_{x \rightarrow 0-} |x| = \lim_{x \rightarrow 0-} (-x) = 0$, also $\lim_{x \rightarrow 0} |x| = 0 = |0| = f(0)$.

- c) Die **Heaviside-Funktion** $H : \mathbb{R} \rightarrow \mathbb{R}$ mit

$$H(x) = \begin{cases} 1, & x > 0, \\ \frac{1}{2}, & x = 0, \\ 0, & x < 0, \end{cases}$$

ist stetig in $x_0 \neq 0$. Für $x_0 = 0$ ist $H(x)$ nicht stetig, da wegen $\lim_{x \rightarrow 0+} H(x) = 1 \neq \lim_{x \rightarrow 0-} H(x) = 0$ der Grenzwert nicht existiert. Insbesondere weicht der Funktionswert an der Stelle 0 von beiden einseitigen Grenzwerten ab.

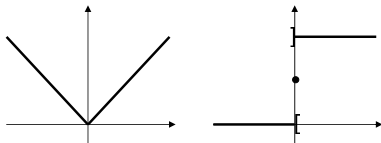


Abb. 2.14: Betragsfunktion und Heaviside-Funktion

- d) Da $\lim_{x \rightarrow x_0} \sqrt{x} = \sqrt{x_0}$ für jedes $x_0 > 0$ gilt (siehe Seite 261), ist die Wurzelfunktion stetig auf $]0, \infty[$ (zusätzlich ist sie rechtsseitig stetig im Nullpunkt).
- e) Wir haben bereits ebenfalls auf Seite 262 gezeigt, dass $\lim_{x \rightarrow x_0} \sin x = \sin x_0$ und $\lim_{x \rightarrow x_0} \cos x = \cos x_0$. Nach (2.16) ist außerdem $\lim_{x \rightarrow x_0} \exp x = \exp x_0$. Damit sind Sinus, Kosinus und die Exponentialfunktion auf $\mathbb{R}$ stetig.
- f) Wegen $\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$ (siehe (2.14)) ist die Funktion

$$f(x) = \operatorname{sinc} x := \begin{cases} \frac{\sin x}{x}, & x \neq 0, x \in \mathbb{R}, \\ 1, & x = 0, \end{cases} \quad (2.18)$$

stetig in 0. Diese Funktion wird später bei der Fourier-Transformation noch wichtig und heißt **Sinus Cardinalis** in der nicht-normierten Form, die wir ausschließlich

verwenden (siehe auch die Abbildung 6.12 auf Seite 864). Die normierte Form des Sinus Cardinalis wird über

$$g(x) := \begin{cases} \frac{\sin(\pi x)}{\pi x}, & x \neq 0, x \in \mathbb{R} \\ 1, & x = 0, \end{cases}$$

definiert und erhält zur Verwirrung den gleichen Funktionsnamen. ■

Wir haben die Stetigkeit über einen Funktionengrenzwert erklärt. Lösen wir diesen einerseits mit Definition 2.14 auf und benutzen wir andererseits das Übertragungsprinzip Satz 2.22, so erhalten wir für die Stetigkeit einer Funktion f an einer Stelle x_0 , die innerer Punkt des Definitionsbereichs D von f ist, die beiden folgenden alternativen Definitionen:

Satz 2.23 (Äquivalente Formulierungen der Stetigkeit)

Die Funktion f ist stetig im Punkt $x_0 \in D$ genau dann,

- wenn zu jedem (noch so kleinen) $\varepsilon > 0$ ein $\delta = \delta(\varepsilon) > 0$ existiert, so dass gilt (siehe Abbildung 2.11 mit $L = f(x_0)$):

$$|x - x_0| < \delta \implies |f(x) - f(x_0)| < \varepsilon.$$

- wenn zu jeder ε -Umgebung $]f(x_0) - \varepsilon, f(x_0) + \varepsilon[$ von $f(x_0)$ eine δ -Umgebung $]x_0 - \delta, x_0 + \delta[$ von x_0 existiert, so dass die Funktionswerte aller Zahlen der δ -Umgebung in der ε -Umgebung liegen.
- wenn für **jede** Folge $(x_n)_{n=1}^\infty$, $x_n \in D$, die gegen die Stelle x_0 konvergiert, gilt

$$\lim_{n \rightarrow \infty} f(x_n) = f(x_0), \text{ d. h. } \lim_{n \rightarrow \infty} f(x_n) = f\left(\lim_{n \rightarrow \infty} x_n\right).$$

Beispiel 2.49

- a) Die Funktion $f(x)$ aus Abbildung 2.15 (links) ist stetig in jedem Punkt $x_0 \in D(f)$: Zu jeder ε -Umgebung von $f(x_0)$ existiert eine δ -Umgebung von x_0 , die durch f ganz in die ε -Umgebung von $f(x_0)$ abgebildet wird.
- b) Die Funktion $f(x)$ aus dem Bild 2.15 (rechts) ist nicht stetig in $x_0 \in D(f)$: Es gibt eine ε -Umgebung von $f(x_0)$, zu der keine δ -Umgebung von x_0 existiert, die durch f ganz in die ε -Umgebung von $f(x_0)$ abgebildet wird, d. h., für die alle Funktionswerte in der ε -Umgebung liegen.
- c) Die Gerade $f(x) = a \cdot x + b$ mit $a, b \in \mathbb{R}$, $a \neq 0$, ist stetig für alle $x_0 \in \mathbb{R}$. Sei $x_0 \in \mathbb{R}$ beliebig. Zu $\varepsilon > 0$ bestimmen wir ein $\delta(\varepsilon) > 0$ so, dass für x mit $|x - x_0| < \delta$ folgt

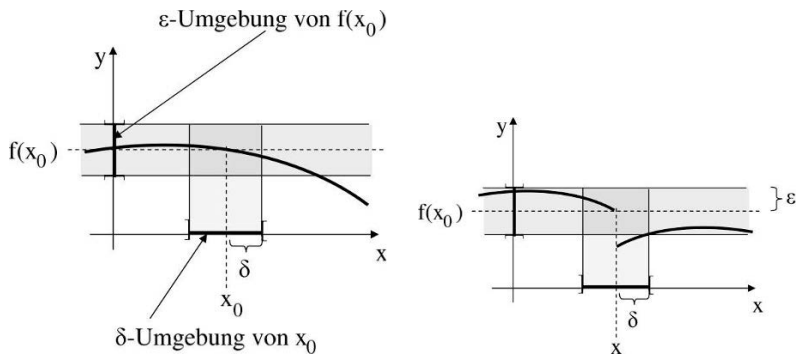


Abb. 2.15: Eine an der Stelle x_0 stetige (links) und eine unstetige Funktion (rechts); im rechten Bild findet man zur gewählten ε -Umgebung von $f(x_0)$ keine δ -Umgebung von x_0 , so dass alle ihre Funktionswerte im ε -Streifen liegen.

$|f(x) - f(x_0)| < \varepsilon$. Das haben wir in Beispiel 2.36 bereits für die Gerade $2x - 1$ an der konkreten Stelle $x_0 = 1$ getan. Für $|x - x_0| < \delta$ folgt allgemeiner

$$|f(x) - f(x_0)| = |a \cdot x + b - (a \cdot x_0 + b)| = |a| |x - x_0| < |a| \delta,$$

so dass mit der Wahl $\delta := \frac{\varepsilon}{|a|}$ für $|x - x_0| < \delta$ gilt: $|f(x) - f(x_0)| < \varepsilon$ (siehe Abbildung 2.16 (links)).

- d) Aufgrund der Grenzwertsätze wissen wir bereits, dass für $x_0 \neq 0$ gilt: $\lim_{x \rightarrow x_0} \frac{1}{x} = \frac{1}{x_0}$ und dass damit die Funktion $f(x) = \frac{1}{x}$ stetig auf $\mathbb{R} \setminus \{0\}$ ist. Das zeigen wir in diesem Beispiel noch einmal explizit mittels der ε - δ -Definition für $x_0 > 0$ (siehe Abbildung 2.16 (rechts)). Zu $\varepsilon > 0$ müssen wir also ein $\delta(\varepsilon) > 0$ so angeben, dass für $x \in \mathbb{R}$ mit $|x - x_0| < \delta$ folgt $|f(x) - f(x_0)| < \varepsilon$.

Für Zahlen $x > 0$ ist

$$|f(x) - f(x_0)| = \left| \frac{1}{x} - \frac{1}{x_0} \right| = \frac{|x_0 - x|}{x \cdot x_0}.$$

Das gesuchte δ darf sich für jedes x_0 unterscheiden. Falls wir ein δ mit $\delta \leq \frac{1}{2}x_0$ wählen, dann ist für $|x - x_0| < \delta$ insbesondere $x > x_0 - \delta \geq \frac{1}{2}x_0$, und es gilt

$$|f(x) - f(x_0)| = \frac{|x_0 - x|}{x \cdot x_0} < \frac{\delta}{\frac{1}{2}x_0 \cdot x_0} \leq \varepsilon,$$

falls zusätzlich $\delta \leq \frac{\varepsilon}{2} x_0^2$ ist. Damit ist die Stetigkeit mit $\delta := \min \left\{ \frac{1}{2}x_0, \frac{\varepsilon}{2} x_0^2 \right\}$ nachgewiesen. ■

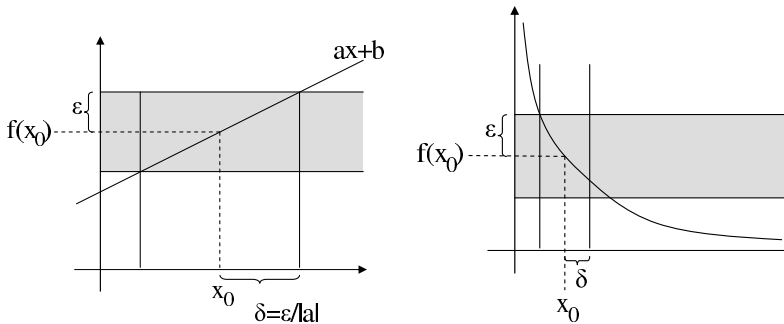


Abb. 2.16: Stetigkeit der Funktionen $f(x) = ax + b$ (links) und $f(x) = \frac{1}{x}$ (rechts)

Viele Vorgänge in der Natur lassen sich mit stetigen Funktionen beschreiben. Betrachtet man z. B. in einem Weg-Zeit-Diagramm die zurückgelegte Wegstrecke als Funktion der Zeit, so wird es hier keine Sprünge geben (es sei denn, man könnte beamen wie in Star Trek). Die Weg-Funktion ist stetig. Auf Englisch heißt Stetigkeit „continuity“ und stetig „continuous“. Stetige Funktionen sind also „kontinuierlich“.

Ändert man das Argument einer stetigen Funktion nur gering, so ändert sich auch der Funktionswert nur gering: Kleine Ursachen haben eine kleine Wirkung. Das gilt für viele Situationen in der Natur, anderenfalls wäre Vieles im täglichen Leben nicht beherrschbar. Rechenverfahren, die ein entsprechendes Verhalten zeigen, heißen **stabil**.

Bei der Dirichlet-Funktion aus Beispiel 2.40 auf Seite 263 ist das Verhalten anders, an keiner Stelle existiert der Grenzwert, sie ist also überall unstetig. Die Funktion springt überall zwischen 0 und 1, eine kleine Änderung kann also zu einer großen Änderung der Funktionswerte (um 1) führen.

Ist eine Funktion an einer Stelle stetig mit Funktionswert größer als null, so drückt sich die Stabilität dadurch aus, dass die Funktion auch auf einer kleinen Umgebung um diese Stelle herum größer als null ist:

Lemma 2.3 (Lokales Verhalten stetiger Funktionen)

Sei $f : D \rightarrow \mathbb{R}$ stetig im inneren Punkt $x_0 \in D$ mit $f(x_0) > 0$, dann gibt es eine Umgebung $]x_0 - \delta, x_0 + \delta[$, $\delta > 0$, auf der f positiv ist, also $f(x) > 0$ für alle $|x - x_0| < \delta$. Die analoge Aussage gilt für $f(x_0) < 0$.

Beweis: Wegen $\lim_{x \rightarrow x_0} f(x) = f(x_0)$ existiert zu $\varepsilon := \frac{f(x_0)}{2}$ ein $\delta > 0$, so dass $|f(x) - f(x_0)| < \frac{f(x_0)}{2}$ für alle $x \in]x_0 - \delta, x_0 + \delta[$. Damit ist in diesem Intervall $f(x) > f(x_0) - \frac{f(x_0)}{2} = \frac{f(x_0)}{2} > 0$. $\square$

Lemma 2.4 (Stetigkeit und einseitige Stetigkeit)

Seien $f : D \rightarrow \mathbb{R}$ und $x_0 \in D$ ein innerer Punkt. Dann sind äquivalent:

- a) f ist stetig in x_0 .
- b) f ist rechtsseitig und linksseitig stetig in x_0 .

Beweis: Die Aussage folgt direkt aus Lemma 2.2 mit $L = f(x_0)$. □

Beispiel 2.50

Die Funktion $f(x) = |x|$ ist stetig in 0, denn sie ist dort links- und rechtsseitig stetig: $\lim_{x \rightarrow 0-} f(x) = \lim_{x \rightarrow 0-} (-x) = 0 = f(0)$ und $\lim_{x \rightarrow 0+} f(x) = \lim_{x \rightarrow 0+} x = 0 = f(0)$. ■

Aus den Rechenregeln für Funktionengrenzwerte (Satz 2.2) ergibt sich direkt, dass Linearkombinationen (wie zuvor beim Stetigkeitsbeweis für Polynome), Produkte und (sofern definiert) Quotienten stetiger Funktionen stetig sind:

Lemma 2.5 (Verknüpfung stetiger Funktionen)

Seien die Funktionen $f, g : D \rightarrow \mathbb{R}$ stetig in einem inneren Punkt $x_0 \in D$. Dann sind $cf(x)$ ($c \in \mathbb{R}$), $f(x) + g(x)$, $f(x)g(x)$ und falls $g(x_0) \neq 0$ auch $\frac{f(x)}{g(x)}$ stetig in x_0 .

Entsprechendes gilt für einseitige Stetigkeit. Man beachte, dass wegen $g(x_0) \neq 0$ nach Lemma 2.3 die Funktionswerte von g sogar auf einer ganzen Umgebung von x_0 von null verschieden sind, so dass der Grenzwert $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)}$ berechnet werden kann.

Beispiel 2.51

- a) Wir haben bereits gezeigt, dass Polynome stetig sind. Damit sind auch alle gebrochen-rationalen Funktionen als Quotient zweier stetiger Polynome auf ihrem Definitionsbereich stetig.
- b) Mit $\sin x$ und $\cos x$ sind auch $\tan x = \frac{\sin x}{\cos x}$ und $\cot x = \frac{\cos x}{\sin x}$ in jedem Punkt ihres Definitionsbereichs stetig.
- c) Als Quotient stetiger Funktionen ist $\frac{\sin x}{x}$ in jedem Punkt $0 \neq x_0 \in \mathbb{R}$ stetig. Damit ist $\operatorname{sinc} x$ (siehe (2.18)) nicht nur in 0, sondern auf ganz $\mathbb{R}$ stetig. ■

Verkettet man stetige Funktionen, d. h., führt man stetige Funktionen hintereinander aus (vgl. Kapitel 1.5.4), so ergibt sich wieder eine stetige Funktion:

Satz 2.24 (Verkettung stetiger Funktionen)

Sei $g : D \rightarrow \mathbb{R}$ stetig in x_0 mit $]x_0 - \delta_0, x_0 + \delta_0[\subset D$, und sei $g[)x_0 - \delta_0, x_0 + \delta_0[\subset I$, wobei I ein Intervall ist. Weiterhin sei $f : I \rightarrow \mathbb{R}$ stetig in $g(x_0)$ (bzw. einseitig stetig, falls $g(x_0)$ ein Randpunkt von I ist). Dann ist die Verkettung $[f \circ g](x) := f(g(x))$ stetig in x_0 .

Beweis: Sei $(x_n)_{n=1}^\infty$ eine beliebige Folge mit Gliedern aus $]x_0 - \delta_0, x_0 + \delta_0[$, die gegen x_0 konvergiert. Aufgrund der Stetigkeit von g ist $\lim_{n \rightarrow \infty} g(x_n) = g(x_0)$ und damit insbesondere nach dem Übertragungsprinzip $\lim_{n \rightarrow \infty} g(x_n) = g(x_0)$, d. h., die Folge $(y_n)_{n=1}^\infty := (g(x_n))_{n=1}^\infty$ konvergiert gegen $y_0 := g(x_0)$, und ihre Glieder liegen in I . Aufgrund der Stetigkeit von f gilt damit: $\lim_{n \rightarrow \infty} f(g(x_n)) = \lim_{n \rightarrow \infty} f(y_n) = f(y_0) = f(g(x_0))$. Da $(x_n)_{n=1}^\infty$ eine beliebige gegen x_0 konvergente Folge aus $]x_0 - \delta_0, x_0 + \delta_0[$ ist, folgt aus dem Übertragungsprinzip $\lim_{x \rightarrow x_0} f(g(x)) = f(g(x_0))$ und damit die Stetigkeit von $f \circ g$ in x_0 . $\square$

Die Aussage von Satz 2.24 gilt bei analogem Beweis auch für einseitige Stetigkeit (von g und $f \circ g$).

Beispiel 2.52

- a) Als Verkettung stetiger Funktionen $f(x) = x^2$ und $g(x) = \sin x$ ist $(\sin x)^2 = \sin^2 x$ stetig auf $\mathbb{R}$.
- b) Die Funktion $h(x) = x^x$ lässt sich nicht unmittelbar naiv als Verkettung zweier Funktionen schreiben. Man könnte auf die Idee kommen, als innere Funktion $g(x) = x$ und als äußere Funktion $f(y) = x^y$ zu wählen. Man mache sich aber klar, dass das **nicht** zulässig ist, da die Funktion f so von den **beiden** Variablen x und y abhängt, also $f(x, y) = x^y$. Damit muss die innere Funktion g einen Vektor $g(x) = (x, x)$ liefern, siehe Seite 608. $\blacksquare$

Um die Stetigkeit der meisten gebräuchlichen Funktionen nachzuweisen, fehlt noch eine Aussage zur Stetigkeit der Umkehrfunktion, die wir untersuchen, wenn wir etwas mehr über stetige Funktionen wissen (siehe Seite 282).

Hintergrund: Topologie - offene Mengen und Stetigkeit

In Kapitel 2.3.1 haben wir definiert, dass eine Menge $E \subset \mathbb{R}$ offen heißt, falls zu jedem $x \in E$ ein offenes Intervall (eine offene Umgebung) $]x - \delta, x + \delta[\subset E$ existiert. Jedes Element $x \in E$ liegt dann also im Inneren von E . Offensichtlich sind sowohl $\mathbb{R}$ als auch $\emptyset$ offene Mengen. Ebenso können wir beliebig viele offene Mengen vereinigen und erhalten wieder eine offene Menge: Jedes Element x der Vereinigung liegt in mindestens einer der offenen Mengen, so dass auch ein offenes Intervall um x in dieser Menge und damit in der Vereinigung liegt. Wenn wir endlich viele (d. h. n) offenen Mengen schneiden, dann

entsteht ebenfalls eine offene Menge: Sei x ein Element der Schnittmenge. Dann liegt x in jeder der n offenen Mengen zusammen mit je einem offenen Intervall $]x - \delta_k, x + \delta_k[$, $1 \leq k \leq n$. Zu $\delta_0 := \min\{\delta_k : 1 \leq k \leq n\}$ ist das offene Intervall $]x - \delta_0, x + \delta_0[$ in der Schnittmenge enthalten. Die soeben nachgerechneten Eigenschaften sind geeignet, den Begriff der offenen Menge von $\mathbb{R}$ auf beliebige andere Mengen X auszudehnen:

Definition 2.18 (Topologie)

Eine Menge X zusammen mit einer Menge $\mathcal{T}$, deren Elemente Teilmengen von X sind, heißt **Topologischer Raum** genau dann, wenn die folgenden Bedingungen erfüllt sind:

- $X \in \mathcal{T}$ und $\emptyset \in \mathcal{T}$.
- Die Vereinigung beliebig vieler Elemente von $\mathcal{T}$ ist selbst auch in $\mathcal{T}$ enthalten.
- Die Schnittmenge endlich vieler Elemente von $\mathcal{T}$ ist ebenfalls in $\mathcal{T}$ enthalten.

Die Menge $\mathcal{T}$ heißt eine **Topologie**. Die Elemente von $\mathcal{T}$ heißen **offene Mengen**. Ist $E \in \mathcal{T}$ eine offene Menge, dann heißt ihr Komplement $C_X E$ eine **abgeschlossene Menge**.

Jetzt können wir auch den in der Analysis sehr zentralen Begriff der Stetigkeit allgemeiner formulieren:

Definition 2.19 (Stetigkeit)

Eine Abbildung von der Menge X eines Topologischen Raums in die entsprechende Menge eines anderen heißt **stetig** genau dann, wenn die Urbilder offener Mengen wieder offen sind.

Diese topologische Definition entspricht im Spezialfall $X = \mathbb{R}$, versehen mit der Topologie der offenen Mengen in $\mathbb{R}$, genau der klassischen Stetigkeit. Das folgt direkt aus dem zweiten Punkt von Satz 2.23 auf Seite 272.

Hat man eine bijektive stetige Abbildung, deren Umkehrfunktion ebenfalls stetig ist, dann nennt man sie einen **Homöomorphismus**. Solche Abbildungen ordnen offene Mengen der einen Topologie eindeutig offenen Menge der anderen Topologie zu.

In der Mathematik betrachtet man an vielen Stellen Mengen, die eine gewisse Struktur haben. Ein Beispiel dafür sind Vektorräume, auf die wir in Kapitel 3 intensiv eingehen werden. Von besonderem Interesse sind dann Abbildungen zwischen solchen Mengen, die die Struktur erhalten. Bei Vektorräumen sind das die linearen Abbildungen, die Linearkombinationen erhalten (Kapitel 3.4). Eine bijektive lineare Abbildung zwischen zwei Vektorräumen bedeutet, dass die Vektorräume in der Linearen Algebra genau die gleichen Eigenschaften haben und eigentlich nicht unterschieden werden müssen. In der Analysis gilt das für Topologische Räume, die durch einen Homöomorphismus aufeinander abgebildet werden können.

Es erscheint zunächst wie eine mathematische Spielerei, den Begriff der Stetigkeit auf Abbildungen zwischen Topologischen Räumen zu erweitern. Tatsächlich ist diese Verallgemeinerung aber sehr wichtig, da viele Lösungsverfahren als Abbildung aufgefasst werden können. Man hätte nun gerne, dass man bei geringfügigen Änderungen an den Daten (die man sich z. B. beim numerischen Lösen durch Rechenungenauigkeiten und einen endlichen Zahlenbereich einhandelt) auch nur geringfügige Änderungen der Lösung erhält. Das ist aber genau die Stetigkeit der Abbildung: Kleine Änderungen führen auch nur zu kleinen Wirkungen. In diesem Fall diskutiert man also Stetigkeit von Abbildungen, die - statt auf $\mathbb{R}$ - auf ganz unterschiedlichen Mengen erklärt sein können.

2.3.4 Eigenschaften stetiger Funktionen

Eine Begründung dafür, dass man den Funktionsgraphen einer auf einem Intervall stetigen Funktion ohne Absetzen des Stiftes von einem Punkt zu einem anderen zeichnen kann, ist, dass alle Funktionswerte zwischen den beiden Funktionswerten der Endpunkte angenommen werden:

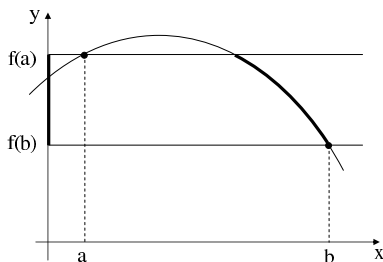


Abb. 2.17: Alle Werte zwischen $f(a)$ und $f(b)$ werden von einer stetigen Funktion f angenommen.

Satz 2.25 (Zwischenwertsatz)

Sei f eine auf $[a, b]$ stetige Funktion, und sei c eine Zahl zwischen (und ungleich) $f(a)$ und $f(b)$. Dann existiert mindestens eine Stelle x_0 echt zwischen a und b , also $x_0 \in]a, b[$, an der die Funktion f den Wert c annimmt: $f(x_0) = c$.

Lässt man für c auch die Randpunkte $f(a)$ und $f(b)$ zu, gibt es offensichtlich mindestens ein $x_0 \in [a, b]$ mit $f(x_0) = c$.

Die „Zwischenstelle“ x_0 wird in der Mathematik gerne mit dem griechischen Buchstaben ξ (sprich: „xi“) bezeichnet, der an „zwischen“ erinnert.

Beweis: Zum Beweis der Aussage benutzen wir wieder die Vollständigkeit der reellen Zahlen (1.5) von Seite 49 und kombinieren sie mit der Definition der Stetigkeit sowie dem Übertragungsprinzip. Sei $f(a) < f(b)$, und sei $c \in \mathbb{R}$ mit $f(a) < c < f(b)$. Der Fall

$f(b) < f(a)$ lässt sich analog beweisen. Wir zeigen, dass der Wert c angenommen wird. Dazu betrachten wir die Menge $A := \{x \in [a, b] : f(x) \leq c\}$. Da $f(a) \leq c$ gefordert ist, ist $a \in A$ und damit $A \neq \emptyset$. Außerdem ist $A \subset [a, b]$ und damit nach oben beschränkt durch b . Nach Axiom (1.5) existiert eine kleinste obere Schranke $x_0 := \sup A \in \mathbb{R}$. Aufgrund (1.7), siehe Seite 51, ist insbesondere sogar $x_0 \in [a, b]$. Wegen $f(a) \neq c \neq f(b)$ ist sogar $x_0 \in]a, b[$. Wir zeigen nun, dass an dieser Stelle der Funktionswert c angenommen wird. Wegen der Stetigkeit ist $\lim_{x \rightarrow x_0-} f(x) = \lim_{x \rightarrow x_0+} f(x) = f(x_0)$. Mit Folgerung 2.3 erhalten wir:

- Für $x < x_0$ ist $f(x) \leq c$ und damit $\lim_{x \rightarrow x_0-} f(x) = f(x_0) \leq c$.
- Für $x > x_0$ ist $f(x) > c$ und damit $\lim_{x \rightarrow x_0+} f(x) = f(x_0) \geq c$.

Damit ist $f(x_0) = c$. □

Den Zwischenwertsatz kann man zur Suche von Nullstellen bei stetigen Funktionen und damit insbesondere bei Polynomen nutzen. Hat man erst einmal Stellen $x_1 < x_2$ gefunden, an denen die Funktionswerte unterschiedliche Vorzeichen haben, muss dazwischen eine Nullstelle liegen (**Nullstellensatz von Bolzano**). Man kann das Intervall $[x_1, x_2]$ teilen in Intervalle $[x_1, x_3]$ und $[x_3, x_2]$ (siehe Abbildung 2.18). In Abhängigkeit des Vorzeichens bei x_3 weiß man, in welchem der Teilintervalle eine Nullstelle liegen muss und kann dieses seinerseits wieder unterteilen usw. Nach Satz 2.6 (Intervallschachtelung, Seite 233) gibt es eine gemeinsame Stelle, die in allen Intervallen liegt, und diese ist die gesuchte Nullstelle. So wird im Beispiel 2.18 auf Seite 233 mit der Intervallschachtelung eine Nullstelle $\sqrt{2}$ des Polynoms $x^2 - 2 = 0$ berechnet. Eine Verbesserung dieses **Bisaktionsverfahrens** lernen wir im Zuge der Differenzialrechnung kennen (siehe Seite 306).

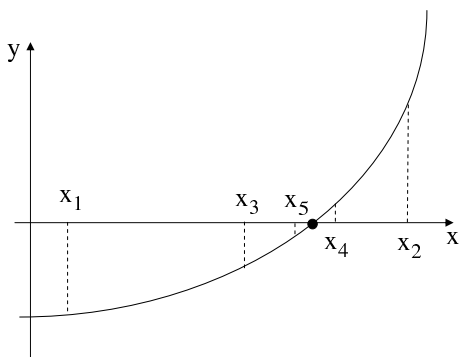


Abb. 2.18: Annäherung an die Nullstelle mittels Zwischenwertsatz

Besonders schön verhalten sich stetige Funktionen auf abgeschlossenen Intervallen $[a, b]$. Sie sind dort beschränkt und nehmen einen größten und einen kleinsten Funktionswert in mindestens einem Punkt des Intervalls an. Das gilt im Allgemeinen nicht für offene (oder einseitig offene) Intervalle $]a, b[$.

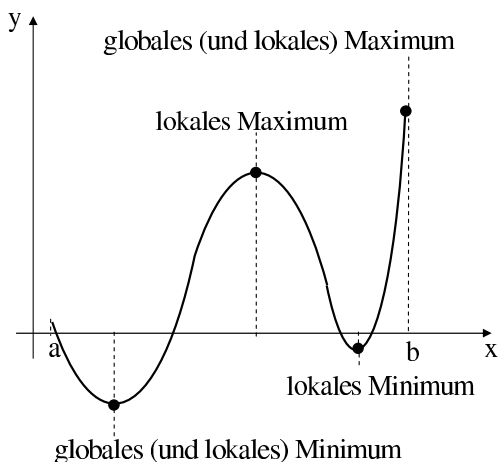


Abb. 2.19: Maxima und Minima

Definition 2.20 (Extrema)

Sei $f : D \rightarrow \mathbb{R}$ (vgl. Abbildung 2.19).

- f hat in $x_0 \in D$ ein **globales Maximum** genau dann, wenn $f(x) \leq f(x_0)$ für alle $x \in D$. Gilt stattdessen (für alle $x \neq x_0$) sogar $f(x) < f(x_0)$, spricht man von einem **strikten globalen Maximum**.
- f hat in $x_0 \in D$ ein **globales Minimum** genau dann, wenn $f(x) \geq f(x_0)$ für alle $x \in D$. Gilt stattdessen (für alle $x \neq x_0$) sogar $f(x) > f(x_0)$, spricht man von einem **strikten globalen Minimum**.
- f hat in $x_0 \in D$ ein **lokales Maximum** (relatives Maximum, Hochpunkt) genau dann, wenn es eine δ_0 -Umgebung $U_{\delta_0} :=]x_0 - \delta_0, x_0 + \delta_0[$ für ein $\delta_0 > 0$ gibt, so dass innerhalb dieser Umgebung alle Funktionswerte $f(x)$, $x \in U_{\delta_0} \cap D$, kleiner oder gleich $f(x_0)$ sind. Sind auf dieser Umgebung mit Ausnahme der Stelle x_0 (also auf einer punktierten Umgebung) sogar alle Funktionswerte echt kleiner als $f(x_0)$, so spricht man von einem **strikten lokalen Maximum**.
- f hat in $x_0 \in D$ ein **lokales Minimum** (relatives Minimum, Tiefpunkt) genau dann, wenn für eine δ_0 -Umgebung U_{δ_0} von x_0 gilt: $f(x) \geq f(x_0)$ für alle $x \in U_{\delta_0} \cap D$. Gilt stattdessen sogar $f(x) > f(x_0)$ für alle $x \in U_{\delta_0} \cap D$ mit $x \neq x_0$, spricht man von einem **strikten lokalen Minimum**.
- Ein **Extremum** ist ein Minimum oder Maximum.

Globale Extrema sind automatisch auch lokale Extrema. Die Umkehrung gilt natürlich nicht immer.

Satz 2.26 (Verhalten stetiger Funktionen auf Intervallen)

Sei f stetig auf $[a, b]$. Dann nimmt f auf $[a, b]$ (d. h. in mindestens einem Punkt $x_M \in [a, b]$) ein globales Maximum und (in mindestens einem Punkt $x_m \in [a, b]$) ein globales Minimum an. Insbesondere ist f auf $[a, b]$ beschränkt.

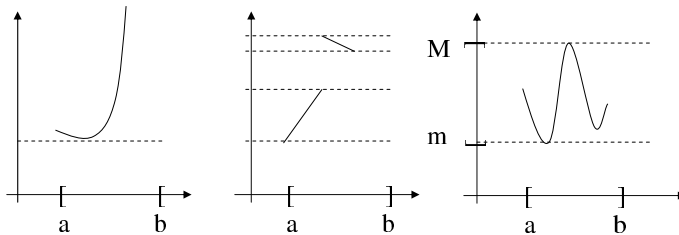


Abb. 2.20: Abbildung von Intervallen mit stetigen und unstetigen Funktionen, links: $[a, b]$ wird durch stetiges f nicht auf ein beschränktes Intervall abgebildet, mittig: Unstetiges f bildet $[a, b]$ auf kein Intervall ab, rechts: Stetiges f bildet $[a, b]$ auf $[m, M]$ ab.

Beweis: Wir zeigen zunächst, dass f auf $[a, b]$ beschränkt ist. Angenommen, f sei nicht beschränkt. Dann kann man eine Folge $(x_n)_{n=1}^{\infty}$ mit $x_n \in [a, b]$, $n \in \mathbb{N}$, konstruieren, so dass $|f(x_n)| > n$. Da die Folge $(x_n)_{n=1}^{\infty}$ nach unten mit a und oben mit b beschränkt ist, existiert nach dem Satz von Bolzano-Weierstraß (Satz 2.7, Seite 234) eine konvergente Teilfolge $(x_{n_k})_{k=1}^{\infty}$. Diese muss einen Grenzwert $x_0 \in [a, b]$ annehmen. Nach Konstruktion ist $\lim_{k \rightarrow \infty} |f(x_{n_k})| = \infty$. Aufgrund der Stetigkeit von f gilt aber im Widerspruch dazu $\lim_{k \rightarrow \infty} f(x_{n_k}) = f(x_0)$. f muss also auf $[a, b]$ beschränkt sein.

Damit existieren $m := \inf\{f(x) : x \in [a, b]\}$ und $M := \sup\{f(x) : x \in [a, b]\}$. Wir zeigen, dass das Supremum angenommen wird. Die Aussage für das Infimum folgt völlig analog.

Zu jedem $n \in \mathbb{N}$ existiert wegen (1.6) auf Seite 50 ein $x_n \in [a, b]$ mit $M - f(x_n) < 1/n$. Die Folge $(x_n)_{n=1}^{\infty}$ ist beschränkt und besitzt wieder wegen des Satzes von Bolzano-Weierstraß eine konvergente Teilfolge $(x_{n_k})_{k=1}^{\infty}$. Diese muss wieder einen Grenzwert $x_0 \in [a, b]$ annehmen. Nach Konstruktion ist $\lim_{k \rightarrow \infty} f(x_{n_k}) = M$. Aufgrund der Stetigkeit von f gilt aber auch $\lim_{k \rightarrow \infty} f(x_{n_k}) = f(x_0)$. Also nimmt f das Maximum an der Stelle x_0 an. $\square$

Dieser Satz zusammen mit dem Zwischenwertsatz, der sicherstellt, dass auch alle Werte zwischen Maximum und Minimum angenommen werden, führt zu (siehe Abbildung 2.20)

Satz 2.27 (Stetige Funktionen bilden Intervalle auf Intervalle ab)

Ist die Funktion f auf $[a, b]$ stetig, so ist $f([a, b])$ ein Intervall $[m, M]$, wobei $m := \min\{f(x) : x \in [a, b]\}$ und $M := \max\{f(x) : x \in [a, b]\}$.

Mit dieser Aussage lässt sich zeigen, dass sich nicht nur das Monotonieverhalten auf die Umkehrfunktion überträgt (siehe Lemma 1.7 auf Seite 86), sondern auch die Stetigkeit:

Satz 2.28 (Stetigkeit der Umkehrfunktion)

Sei f stetig auf $[a, b]$ und streng monoton (steigend oder fallend). Dann ist die existierende Umkehrfunktion $f^{-1} : f([a, b]) \rightarrow [a, b]$ stetig auf $f([a, b])$.

Kann man den Graphen von f ohne Absetzen eines Stiftes auf dem Intervall zeichnen, so geht das auch für den Graphen der Umkehrfunktion, der durch Spiegelung an der Geraden $y = x$ entsteht. Das ist aber noch kein formaler Beweis. Dieser ist etwas verwirrend, da man zwischen Definitions- und Wertebereich von f bzw. Werte- und Definitionsbereich von f^{-1} hin- und herspringen muss. Er kann mit dem Übertragungsprinzip geführt werden. Wir benutzen aber den eben gezeigten Satz 2.27.

Beweis: Nach Satz 2.27 ist $f([a, b])$ ein abgeschlossenes Intervall, so dass es zu jedem $y_0 \in f([a, b])$ eine in der Grenzwertdefinition geforderte (in Randpunkten einseitige) Umgebung von y_0 gibt.

Wir müssen nun für jedes $y_0 \in f([a, b])$ zeigen: $\lim_{y \rightarrow y_0} f^{-1}(y) = f^{-1}(y_0)$ (einseitig für die Intervallränder, die wegen der Monotonie $f(a)$ und $f(b)$ sind). Nach Definition des Grenzwerts ist also zu zeigen: Für alle $\varepsilon > 0$ existiert ein $\delta = \delta(\varepsilon, y_0) > 0$, so dass für alle $y \in f([a, b])$ mit $|y - y_0| < \delta$ gilt:

$$|f^{-1}(y_0) - f^{-1}(y)| < \varepsilon.$$

Wir zeigen dies damit, dass eine stetige Funktion ein abgeschlossenes, beschränktes Intervall auf ein ebensolches abbildet. Wir ersetzen dazu die ε -Umgebung $\{x : |x - f^{-1}(y_0)| < \varepsilon\}$, in der die Funktionswerte $f^{-1}(y)$ liegen sollen, durch eine kleinere abgeschlossene Umgebung $I_\varepsilon := [f^{-1}(y_0) - \varepsilon/2, f^{-1}(y_0) + \varepsilon/2] \cap [a, b]$. $f(I_\varepsilon)$ ist nach Satz 2.27 ein Intervall $[c, d] := f(I_\varepsilon) \subset f([a, b])$, in dem insbesondere y_0 liegt. Das Intervall $[c, d]$ wird von f^{-1} nach Definition auf I_ε und damit vollständig in die ε -Umgebung von $f^{-1}(y_0)$ abgebildet. Wenn wir innerhalb von $[c, d]$ eine δ -Umgebung von y_0 angeben können, dann werden alle Zahlen y aus dieser Umgebung durch f^{-1} auf einen Wert abgebildet, der näher an $f^{-1}(y_0)$ liegt als ε , und die Stetigkeit wäre über die Grenzwertdefinition gezeigt. Dazu unterscheiden wir die beiden Fälle $y_0 \in]c, d[$ und $y_0 \in \{c, d\}$:

■ Ist $y_0 \in]c, d[$, so folgt die Stetigkeit mit $\delta := \min\{|y_0 - c|, |y_0 - d|\} > 0$.

- Ist dagegen y_0 ein Randpunkt c oder d , so muss wegen der strengen Monotonie von f auch $f^{-1}(y_0)$ ein Randpunkt von I_ε sein. Das geht aber nur, wenn $f^{-1}(y_0) = a$ oder $f^{-1}(y_0) = b$ ist. Damit kann in diesem Fall $\delta := |c - d|$ für eine einseitige Umgebung gewählt werden.

□

Mit diesem Satz kann man zeigen, dass die Umkehrfunktionen der trigonometrischen Funktionen wie z. B. $\arctan x$ auf ihrem Definitionsbereich stetig sind. Dazu betrachtet man die trigonometrischen Funktionen auf abgeschlossenen Intervallen, auf denen sie streng monoton (und stetig) sind. Ebenfalls ist $\ln x$ als Umkehrfunktion der streng monoton steigenden Exponentialfunktion stetig auf $]0, \infty[$.

Jetzt erhalten wir die Stetigkeit weiterer wichtiger Funktionen mit der soeben bewiesenen Stetigkeit des Logarithmus.

Beispiel 2.53

Die Potenzfunktionen x^α sind für $\alpha \in \mathbb{R}$ und $x > 0$ stetig, da sie eine Verkettung stetiger Funktionen sind: $x^\alpha = \exp(\alpha \ln x)$.

Für $\alpha > 0$ ist x^α zudem rechtsseitig stetig in 0. Das zeigen wir explizit mit der Grenzwertdefinition: Zu $\varepsilon > 0$ sei $\delta := \varepsilon^{\frac{1}{\alpha}}$. Für $0 \leq x < \delta$ ist

$$|x^\alpha - 0^\alpha| = x^\alpha < \delta^\alpha = \varepsilon.$$

Wählen wir insbesondere $\alpha = \frac{1}{2}$, so erhalten wir die Stetigkeit von $\sqrt{x}$ auf $[0, \infty[$ (vgl. Seite 261). ■

Beispiel 2.54

Für $a > 0$ ist die allgemeine Exponentialfunktion $f(x) := a^x = \exp(x \ln a)$ ebenso als Verkettung des stetigen Polynoms $x \ln a$ und der stetigen Exponentialfunktion stetig. ■

Hat man eine stetige Funktion, dann ist die Frage nach der Existenz einer Umkehrfunktion vergleichsweise leicht zu beantworten. Als weitere Anwendung des Zwischenwertsatzes zeigen wir:

Satz 2.29 (Monotonie und Invertierbarkeit)

Für eine stetige Funktion $f : [a, b] \rightarrow \mathbb{R}$ sind äquivalent:

- a) f ist streng monoton auf $[a, b]$.
- b) f ist injektiv (jeder Funktionswert wird höchstens einmal angenommen) mit Umkehrfunktion $f^{-1} : f([a, b]) \rightarrow [a, b]$.

Beweis: Aus der strengen Monotonie folgt sofort, dass f injektiv ist. Schränken wir die Zielmenge von f auf den Wertebereich $f([a, b])$ ein (der wegen der Stetigkeit ein Intervall ist), dann wird f bijektiv, die Umkehrabbildung f^{-1} existiert.

Dass die erste Aussage aus der zweiten folgt, ist eine Konsequenz des Zwischenwertsatzes für stetige Funktionen. Zunächst ist vorausgesetzt, dass f nicht an zwei Stellen den gleichen Wert annehmen kann. Wir nehmen nun an, f sei nicht streng monoton. Also gibt es Stellen $x_1 < x_2 < x_3$ mit

$$f(x_1) < f(x_2) > f(x_3) \text{ oder } f(x_1) > f(x_2) < f(x_3).$$

Gleichheitszeichen sind nicht zu berücksichtigen, da f ja keinen Wert zweimal annimmt. Wir betrachten die erste Konstellation, für die zweite können wir analog argumentieren. Wegen des Zwischenwertsatzes werden auf $[x_1, x_2]$ alle Werte aus $[f(x_1), f(x_2)]$ angenommen, und auf $[x_2, x_3]$ alle Werte aus $[f(x_3), f(x_2)]$. Die Funktionswerte aus $[\max\{f(x_1), f(x_3)\}, f(x_2)] \neq \emptyset$ werden also im Widerspruch zur Injektivität zweimal angenommen. Die Annahme gilt nicht, es liegt strenge Monotonie vor. $\square$

Auf Intervallen $[a, b]$ verhalten sich stetige Funktionen besonders schön. Das geht so weit, dass sie auch eine strengere Definition der Stetigkeit erfüllen. Diese werden wir beweistechnisch verwenden (z. B. um zu zeigen, dass stetige Funktionen integrierbar sind). Für technische Anwendungen ist der nächste Satz aber nicht zentral:

Satz 2.30 (Gleichmäßige Stetigkeit*)

*Ist eine Funktion f auf $[a, b]$ stetig, so ist f hier sogar **gleichmäßig stetig**, d. h., für jedes noch so kleine $\varepsilon > 0$ existiert ein $\delta = \delta(\varepsilon)$, so dass für alle $x_1, x_2 \in [a, b]$ mit $|x_1 - x_2| < \delta$ gilt: $|f(x_1) - f(x_2)| < \varepsilon$.*

Die gleichmäßige Stetigkeit ist mehr als nur Stetigkeit. Aus der gleichmäßigen Stetigkeit folgt sofort die „normale“ Stetigkeit an jeder Stelle $x_1 \in [a, b]$. Das Besondere ist hier aber, dass δ nur von ε , nicht aber von einer Stelle $x_1 \in [a, b]$ abhängt. Für jede Stelle kann also beim Stetigkeitsnachweis über die Grenzwertdefinition bei einem vorgegebenen ε das gleiche δ gewählt werden. Das bedeutet eine zusätzliche Glattheit der Funktion. Noch glatter sind differenzierbare Funktionen im nächsten Kapitel.

Beweis: Sei $\varepsilon > 0$ beliebig vorgegeben. Da f stetig auf $[a, b]$ ist, existiert zu jedem $x \in [a, b]$ ein von x abhängiges $\delta = \delta(x)$, so dass $|f(x) - f(x_1)| < \varepsilon$ für alle $x_1 \in]x - \delta(x), x + \delta(x)[\cap [a, b]$. Damit erstellen wir eine offene Überdeckung von $[a, b]$: $\mathcal{F} := \{]x - \delta(x)/2, x + \delta(x)/2[: x \in [a, b]\}$. Hier können beliebig kleine $\delta(x)$ auftreten. Nach dem Satz von Heine-Borel (Satz 2.20, Seite 256) existiert jedoch bereits eine endliche Teilmenge von $\mathcal{F}$, mit der $[a, b]$ überdeckt werden kann. Diese hat die Gestalt $\{]x - \delta(x)/2, x + \delta(x)/2[: x \in E\}$ für eine endliche Teilmenge $E \subset [a, b]$. Sei nun $\delta := \min\{\delta(x)/2 : x \in E\}$. Als Minimum von endlich vielen positiven Zahlen ist $\delta > 0$. Mit diesem δ folgt die Aussage des Satzes:

Zu jedem $x_1 \in [a, b]$ gibt es ein $x \in E$, so dass $x_1 \in]x - \delta(x)/2, x + \delta(x)/2[$. Für jedes $x_2 \in [a, b]$ mit $|x_1 - x_2| < \delta/2$ ist $x_2 \in]x - \delta(x), x + \delta(x)[$. Nach Konstruktion der Umgebungen aus der Stetigkeit ist damit aber $|f(x_1) - f(x_2)| < \varepsilon$.

Der Trick des Beweises ist, dass die Überdeckung nicht mit $\delta(x)$ -, sondern mit $\delta(x)/2$ -Umgebungen realisiert wurde. $\square$

2.3.5 Unstetigkeitsstellen

Auch die Stellen, an denen eine Funktion nicht stetig ist, sind von Interesse. Häufig muss man ihnen besonders viel Aufmerksamkeit schenken, da dort tatsächlich etwas passiert, während die Funktionen sonst „harmlos“ sind.

Definition 2.21 (Unstetigkeitsstellen)

Sei $f : D \rightarrow \mathbb{R}$, so dass eine punktierte Umgebung $]x_0 - \delta_0, x_0[\cup]x_0, x_0 + \delta_0[$ von x_0 in D enthalten ist.

■ x_0 heißt genau dann eine **hebbare Unstetigkeitsstelle** von f , wenn

- i) $x_0 \in D$ und $\lim_{x \rightarrow x_0} f(x)$ existiert, aber $\lim_{x \rightarrow x_0} f(x) \neq f(x_0)$ oder
- ii) $x_0 \notin D$ und $\lim_{x \rightarrow x_0} f(x)$ existiert.

In diesem Fall heißt f **stetig ergänzbar** oder **stetig fortsetzbar** in x_0 mit Wert $\lim_{x \rightarrow x_0} f(x)$.

- x_0 heißt genau dann eine **Sprungstelle** von f , wenn die Grenzwerte $\lim_{x \rightarrow x_0-} f(x)$ und $\lim_{x \rightarrow x_0+} f(x) \in \mathbb{R}$ existieren, aber unterschiedlich sind.
- x_0 heißt genau dann eine **Unstetigkeitsstelle erster Art**, wenn x_0 eine hebbare Unstetigkeitsstelle oder eine Sprungstelle ist.
- x_0 heißt genau dann eine **Unstetigkeitsstelle zweiter Art**, wenn mindestens einer der Grenzwerte $\lim_{x \rightarrow x_0-} f(x)$ oder $\lim_{x \rightarrow x_0+} f(x)$ nicht existiert (dazu zählen auch die uneigentlichen Grenzwerte $\pm\infty$).
- x_0 heißt **Polstelle** oder **Pol** der Funktion f genau dann, wenn beide einseitigen Grenzwerte uneigentlich sind: $\lim_{x \rightarrow x_0-} f(x) = \pm\infty$ und $\lim_{x \rightarrow x_0+} f(x) = \pm\infty$, wobei auch $+\infty$ und $-\infty$ gemischt auftreten dürfen. Polstellen sind also insbesondere Unstetigkeitsstellen zweiter Art.

Beispiel 2.55

- a) Wir haben bereits gesehen, dass die auf $\mathbb{R} \setminus \{0\}$ stetige Funktion $\frac{\sin x}{x}$ stetig ergänzbar ist in 0 mit Wert 1 (und dann $\operatorname{sinc} x$ heißt).
- b) Dagegen hat $f(x) := \frac{1}{x}$, $f : \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R} \setminus \{0\}$ eine Unstetigkeitsstelle zweiter Art in 0, da $\lim_{x \rightarrow 0-} f(x) = -\infty$, $\lim_{x \rightarrow 0+} f(x) = \infty$. Insbesondere ist dies ein Pol.

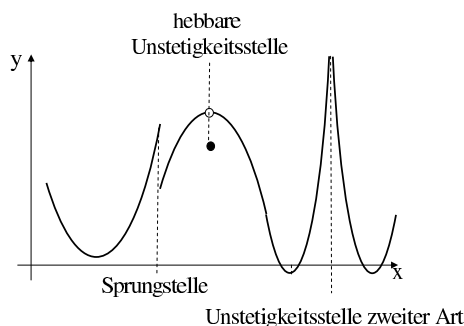


Abb. 2.21: Unstetigkeitsstellen erster und zweiter Art

- c) $\sin\left(\frac{1}{x}\right)$ hat eine Unstetigkeitsstelle zweiter Art in 0, da dort der Grenzwert nicht existiert. 0 ist kein Pol.
 d) Unstetigkeitsstellen der Funktion

$$f(x) := \begin{cases} \frac{2}{x}, & x < 0, \\ \frac{1}{x}, & 0 < x \leq 1, \\ 2x, & 1 < x < 2, \\ 4, & x > 2. \end{cases}$$

können nur bei 0, 1 und 2 vorliegen:

- i) $\lim_{x \rightarrow 0-} f(x) = \lim_{x \rightarrow 0-} \frac{2}{x} = -\infty$, $\lim_{x \rightarrow 0+} f(x) = \lim_{x \rightarrow 0+} \frac{1}{x} = +\infty$. Damit hat f bei 0 eine Unstetigkeitsstelle zweiter Art (einen Pol).
 ii) $\lim_{x \rightarrow 1-} f(x) = \lim_{x \rightarrow 1-} \frac{1}{x} = 1$, $\lim_{x \rightarrow 1+} f(x) = \lim_{x \rightarrow 1+} 2x = 2$. Damit hat f bei 1 eine Sprungstelle.
 iii) $\lim_{x \rightarrow 2-} f(x) = \lim_{x \rightarrow 2-} 2x = 4$, $\lim_{x \rightarrow 2+} f(x) = \lim_{x \rightarrow 2+} 4 = 4$. Damit hat f bei 2 eine hebbare Unstetigkeitsstelle (beachte: 2 ist nicht im Definitionsbereich). ■

Monotone Funktionen (wie auch monotone Folgen) sind besonders einfach zu analysieren. Ihnen bleibt fast nichts anderes übrig, als stetig zu sein. Sie können höchstens einige Sprungstellen, aber keine Unstetigkeitsstellen zweiter Art besitzen:

Lemma 2.6 (Stetigkeit monotoner Funktionen *)

Sei $f :]a, b[\rightarrow \mathbb{R}$ eine auf einem Intervall $]a, b[$ monotone Funktion. Dann ist f stetig auf $]a, b[$ mit Ausnahme von höchstens abzählbar vielen Sprungstellen.

Beweis: Nach Satz 2.21 existieren aufgrund der Monotonie zu jedem Punkt $x_0 \in]a, b[$ die beiden einseitigen Grenzwerte. Ebenfalls wegen der Monotonie gilt entweder $\lim_{x \rightarrow x_0-} f(x) \leq f(x_0) \leq \lim_{x \rightarrow x_0+} f(x)$ für alle $x_0 \in]a, b[$ oder $\lim_{x \rightarrow x_0-} f(x) \geq f(x_0) \geq \lim_{x \rightarrow x_0+} f(x)$ für alle $x_0 \in]a, b[$.

- Stimmen die beiden Grenzwerte überein, so sind sie gleich $f(x_0)$, und wir haben $\lim_{x \rightarrow x_0} f(x) = f(x_0)$ und Stetigkeit in x_0 .
- Stimmen die beiden Grenzwerte nicht überein, so liegt bei x_0 eine Sprungstelle vor. Da die rationalen Zahlen dicht in den reellen Zahlen liegen, gibt es einen Bruch $p_{x_0} \in \mathbb{Q}$, der echt zwischen den beiden einseitigen Grenzwerten liegt.

f ist also stetig auf $]a, b[$ bis auf Sprungstellen. Jeder Sprungstelle haben wir einen Bruch zugeordnet. Da f monoton ist, sind die Brüche zu verschiedenen Sprungstellen verschieden. Da $\mathbb{Q}$ aber abzählbar unendlich ist (siehe Kapitel 1.3.2.3), kann auch die Menge der Sprungstellen nicht größer sein. $\square$

Häufig untersucht man **kritische Stellen** von gebrochen-rationalen Funktionen, also Stellen, an denen der Nenner null wird:

Beispiel 2.56

a) Die gebrochen-rationale Funktion $f(x) = \frac{4x-11}{x-3}$ besitzt die Nennernullstelle $x_0 = 3$. Polynomdivision liefert $f(x) = 4 + \frac{1}{x-3}$. Wir untersuchen das Verhalten von $f(x)$ für $x \rightarrow 3$:

- Wir nähern uns der kritischen Stelle 3 von rechts, d. h., wir betrachten $f(3+h)$ für $h \rightarrow 0$, $h > 0$:

$$f(3+h) = 4 + \frac{1}{(3+h)-3} = 4 + \frac{1}{h} \rightarrow +\infty \quad (h \rightarrow 0+).$$

- Wir nähern uns der kritischen Stelle 3 von links, d. h., wir betrachten $f(3-h)$ für $h \rightarrow 0$, $h > 0$:

$$f(3-h) = 4 + \frac{1}{(3-h)-3} = 4 - \frac{1}{h} \rightarrow -\infty \quad (h \rightarrow 0+).$$

Somit erweist sich $x_0 = 3$ als Pol mit Vorzeichenwechsel.

b) Die gebrochen-rationale Funktion $f(x) = \frac{1}{(x+3)^2}$ hat die doppelte Nennernullstelle $x_0 = -3$. Für $f(-3+h)$ mit $h \rightarrow 0$ folgt:

$$f(-3+h) = \frac{1}{(-3+h+3)^2} = \frac{1}{h^2} \rightarrow +\infty.$$

Somit erweist sich $x_0 = -3$ als Pol ohne Vorzeichenwechsel. Da im Nenner zweimal der Faktor $(x+3)$ vorkommt, spricht man auch von einem Pol der Ordnung 2 (siehe das folgende Lemma 2.7). ■

Lemma 2.7 (Polstellen)

Ist x_0 eine k -fache Nullstelle des Nennerpolynoms $p_n(x)$ der gebrochen-rationalen Funktion

$$f(x) = \frac{q_m(x)}{p_n(x)} = \frac{b_mx^m + b_{m-1}x^{m-1} + \dots + b_1x + b_0}{a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0}$$

und ist x_0 keine Nullstelle des Zählerpolynoms $q_m(x)$, so besitzt $f(x)$ bei x_0 einen Pol. Ist

- k gerade, dann ist x_0 ein Pol ohne Vorzeichenwechsel, d. h. $\lim_{x \rightarrow x_0-} f(x) = \lim_{x \rightarrow x_0+} f(x) = \pm\infty$,
- k ungerade, dann ist x_0 ein Pol mit Vorzeichenwechsel, d. h. $\lim_{x \rightarrow x_0-} f(x) = -\lim_{x \rightarrow x_0+} f(x) = \pm\infty$.

k heißt **Ordnung des Pols**.

Jetzt betrachten wir kritische Stellen, die sowohl Nullstellen von $p_n(x)$ als auch Nullstellen von $q_m(x)$ sind. Durch sukzessives Wegdividieren der Linearfaktoren zur gemeinsamen Nullstelle x_0 erhalten wir

$$f(x) = \frac{q_m(x)}{p_n(x)} = \frac{(x - x_0) \cdot q_{m-1}(x)}{(x - x_0) \cdot p_{n-1}(x)} = \dots = \frac{(x - x_0)^l \cdot q_{m-l}(x)}{(x - x_0)^l \cdot p_{n-l}(x)},$$

bis schließlich $q_{m-l}(x_0) \neq 0$ oder $p_{n-l}(x_0) \neq 0$ gilt.

- Für $p_{n-l}(x_0) \neq 0$ und $q_{m-l}(x_0) = 0$ oder $q_{m-l}(x_0) \neq 0$ folgt: $f(x)$ kann bei x_0 stetig ergänzt werden.
- Für $p_{n-l}(x_0) = 0$ und $q_{m-l}(x_0) \neq 0$ folgt: $f(x)$ hat bei x_0 einen Pol.

Gebrochen-rationale Funktionen besitzen also ausschließlich Pole und hebbare Unstetigkeitsstellen. Andere Unstetigkeitsstellen (z. B. Sprungstellen) können nicht vorkommen.

Beispiel 2.57

- a) Für $f(x) = \frac{x^3 - x^2}{x - 1}$ mit $D(f) = \mathbb{R} \setminus \{1\}$ erhalten wir $f(x) = \frac{x^2 \cdot (x-1)}{x-1} = x^2$ und $\lim_{x \rightarrow 1} f(x) = 1$. Daher ist $f(x)$ durch den Wert $f(1) := 1$ stetig fortsetzbar.
- b) Für $f(x) = \frac{x^2 - 1}{2x^2 - 4x + 2} = \frac{(x-1)(x+1)}{2(x-1)^2}$ mit $D(f) = \mathbb{R} \setminus \{1\}$ erhalten wir $f(x) = \frac{x+1}{2(x-1)}$. Für $x = 1$ besitzt $f(x)$ einen Pol mit Vorzeichenwechsel. ■

2.4 Differenzierbarkeit und Ableitungen

Die Differenzialrechnung (das Rechnen mit Ableitungen) geht auf Arbeiten von Isaac Newton (1642–1727) und Gottfried Wilhelm Leibniz im 17. Jahrhundert zurück und war ein ganz entscheidender Schritt im Aufbau der modernen Naturwissenschaften.

2.4.1 Ableitung als Grenzwert des Differenzenquotienten

Die Stromstärke I ergibt sich als Quotient der Ladung ΔQ , die einen Leiterquerschnitt in der Zeit Δt passiert durch die zugehörige Zeitspanne Δt . Um die Stromstärke zu einem Zeitpunkt t zu erhalten, berechnet man den Grenzwert

$$I(t) := \lim_{\Delta t \rightarrow 0} \frac{Q(t + \Delta t) - Q(t)}{\Delta t}.$$

Dies ist gerade die Ableitung der Ladung $Q(t)$ nach der Zeit zum Zeitpunkt t und wird mit $Q'(t)$ bezeichnet. In der Physik ist es üblich, Ableitungen nach einer Variable, die die Zeit repräsentiert, mit einem Punkt statt mit einem Strich zu kennzeichnen: $\dot{Q}(t) := Q'(t)$. Der Punkt ist die von Newton verwendete Schreibweise der Ableitung. Hier verwenden wir aber ausschließlich den Strich.

Betrachten wir nun allgemein statt $Q(t)$ eine Funktion $f(x)$ an einer Stelle $x_0 \in D(f)$, siehe Abbildung 2.22. Die Sekante $s(x)$ durch die Punkte $(x_0, f(x_0))$ und $(x_0 + h, f(x_0 + h))$ mit $h > 0$ ist die Gerade

$$s(x) = m \cdot (x - x_0) + f(x_0) \quad \text{mit Steigung} \quad m := \frac{f(x_0 + h) - f(x_0)}{h}.$$

Im Zähler des Quotienten m steht eine Differenz der Funktionswerte (Höhenzuwachs) und im Nenner eine Differenz der zugehörigen Argumente (Horizontaldifferenz) $x + h - x = h$ (siehe auch Seite 91). Die Sekantensteigung m bezeichnet man auch als **Differenzenquotient**. Wenn nun h gegen 0 strebt, dann wird aus den Sekanten eine Tangente an den Funktionsgraphen von f an der Stelle x_0 . Darauf gehen wir gleich noch etwas genauer ein. Die Steigung der Tangente ist $f'(x_0)$.

Definition 2.22 (Ableitung)

Sei $f : D \rightarrow \mathbb{R}$, und sei $x_0 \in D$ ein innerer Punkt. Die Funktion f heißt **differenzierbar** an der Stelle x_0 genau dann, wenn der Grenzwert des **Differenzenquotienten**

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} =: f'(x_0)$$

existiert. Der Grenzwert $f'(x_0)$ heißt **Ableitung** (oder **Differenzialquotient**) von f an der Stelle x_0 . Eine alternative Schreibweise für $f'(x_0)$ ist $\frac{df}{dx}(x_0)$.

Bei der Schreibweise $\frac{df}{dx}(x_0)$ ist die Variable angegeben, nach der abgeleitet wird. Das wird später wichtig, wenn wir Funktionen diskutieren, die von mehreren Variablen abhängen. h ist der Abstand (die Differenz) zwischen den Punkten $x_0 + h$ und x_0 . Statt h findet man in der Literatur auch häufig wie oben die Variable Δx (Δ erinnert an „Differenz“).

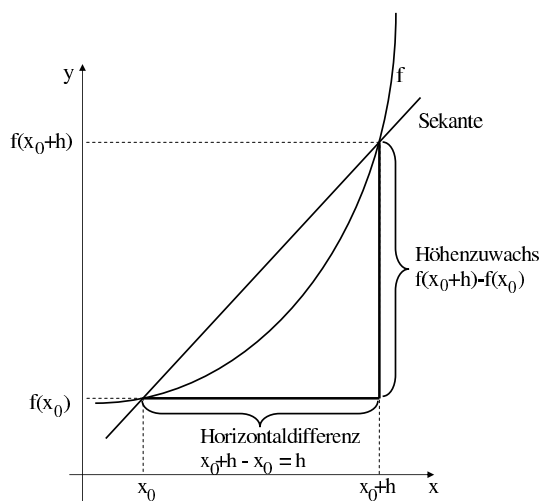


Abb. 2.22: Differenzenquotient zur Definition der Ableitung

Würde man eine Kugel den Funktionsgraphen entlang rollen und an der Stelle $(x_0, f(x_0))$ loslassen, so würde sie sich tangential, d. h. auf einer Geraden, die man Tangente nennt, weiterbewegen. Die Ableitung einer Funktion an einer Stelle x ist genau die Steigung, die die Tangente an den Funktionsgraphen an dieser Stelle hat.

Beispiel 2.58 (Physikalischer Zugang zur Ableitung)

Das Weg-Zeit-Gesetz der gleichförmig beschleunigten Bewegung eines Fahrzeugs lautet (vgl. Beispiel 1.68 auf Seite 85)

$$s(t) = \frac{1}{2}at^2 + v_0t + s_0.$$

Hier ist s_0 der Weg, der bereits zum Zeitpunkt $t = 0$ zurückgelegt ist (Anfangsweg), und v_0 ist die Anfangsgeschwindigkeit zum Zeitpunkt $t = 0$ und a die konstante Beschleunigung des Fahrzeugs. Die **mittlere Geschwindigkeit** $v_m(t, \Delta t)$ ist der Differenzenquotient aus Weg und Zeit

$$\begin{aligned} v_m(t, \Delta t) &= \frac{s(t + \Delta t) - s(t)}{\Delta t} = \frac{\frac{1}{2}a(t + \Delta t)^2 + v_0(t + \Delta t) + s_0 - (\frac{1}{2}at^2 + v_0t + s_0)}{\Delta t} \\ &= \frac{at\Delta t + \frac{1}{2}a\Delta t^2 + v_0\Delta t}{\Delta t} = at + \frac{1}{2}a\Delta t + v_0. \end{aligned}$$

Die **Momentangeschwindigkeit** $v(t)$ zum Zeitpunkt t ist definiert als Grenzwert

$$v(t) = \lim_{\Delta t \rightarrow 0} v_m(t, \Delta t) = \lim_{\Delta t \rightarrow 0} \frac{s(t + \Delta t) - s(t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \left(at + \frac{1}{2}a\Delta t + v_0 \right) = at + v_0,$$

also $v(t) = s'(t) = at + v_0$. Man betrachtet also die mittlere Geschwindigkeit zu immer kleiner werdenden Zeitintervallen und erhält schließlich die Geschwindigkeit zu einem Zeitpunkt. Bei der gleichförmig beschleunigten Bewegung ist die Geschwindigkeit eine lineare Funktion der Zeit. Die Beschleunigung ist die zeitliche Änderungsrate $v'(t)$ der Geschwindigkeit

$$v'(t) = \lim_{\Delta t \rightarrow 0} \frac{v(t + \Delta t) - v(t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{a(t + \Delta t) + v_0 - (at + v_0)}{\Delta t} = a.$$

■

In Anlehnung an den Differenzenquotienten $\frac{f(t+\Delta t)-f(t)}{\Delta t}$ wird die Schreibweise $\frac{d}{dt}f(t)$ verwendet. Dabei steht df für den Zähler und dt für den Nenner, wenn wir Δt „beliebig“ klein machen (siehe Kapitel 2.4.4). Außerdem wird die Variable angegeben, von der die Funktion abhängig ist. Bei einem Weg-Zeit-Diagramm ist der Weg $s(t)$ eine Funktion der Zeit t . Die Geschwindigkeit als Ableitung des Wegs ist $v(t) = \frac{d}{dt}s(t)$. Entsprechend schreibt man $f'(x) = \frac{d}{dx}f(x)$ oder $f'(u) = \frac{d}{du}f(u)$. Die von Leibniz eingeführte Notation $\frac{d}{dt}$ hat sich in Verbindung mit der Integralrechnung als äußerst geschickt erwiesen, da sie ein formales Rechnen mit Ableitungen erlaubt. Dem steht die zuerst von Lagrange 1797 verwendete kürzere Schreibweise f' entgegen, die uns bei Rechnungen aber nicht unterstützt.

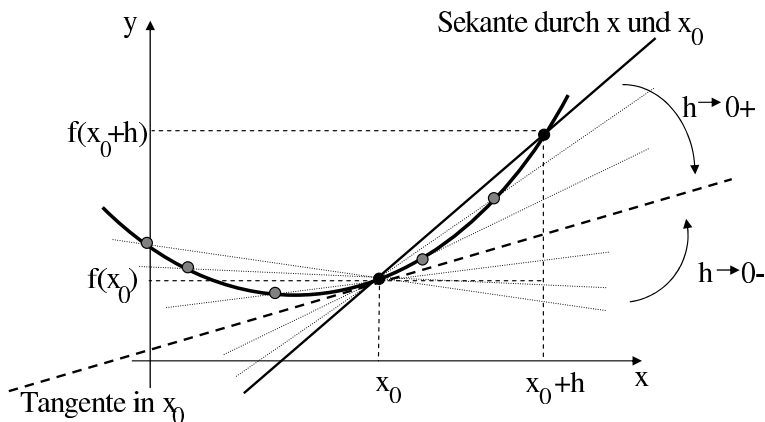


Abb. 2.23: Die Steigungen $\frac{f(x_0+h)-f(x_0)}{h}$ der Sekanten konvergieren für $h \rightarrow 0$ gegen die Steigung $f'(x_0)$ der Tangente.

Beispiel 2.59 (Ableitungen wichtiger Grundfunktionen)

- a) Für $f(x) = 1$ ist $f'(x) = 0$, denn $\lim_{h \rightarrow 0} \frac{1-1}{h} = 0$. Allgemeiner gilt für jede Konstante $c \in \mathbb{R}$:

$$\frac{d}{dx}c = 0.$$

b) Für $f(x) = x$ ist $f'(x) = 1$, denn $\lim_{h \rightarrow 0} \frac{x+h-x}{h} = 1$.

$$\frac{d}{dx}x = 1$$

c) Für $f(x) = x^n$ mit $n \in \mathbb{N}$ gilt mit dem Binomischen Satz

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} = \lim_{h \rightarrow 0} \frac{\left(\sum_{k=0}^n \binom{n}{k} x^{n-k} \cdot h^k \right) - x^n}{h} \\ &= \lim_{h \rightarrow 0} \frac{x^n + nx^{n-1}h + h^2(\dots) - x^n}{h} = nx^{n-1}. \end{aligned}$$

$$\frac{d}{dx}x^n = nx^{n-1}$$

d) Für $f(x) = \frac{1}{x}$, $x \neq 0$, folgt

$$f'(x) = \lim_{h \rightarrow 0} \frac{\frac{1}{x+h} - \frac{1}{x}}{h} = \lim_{h \rightarrow 0} \frac{x - (x+h)}{h(x+h)x} = \lim_{h \rightarrow 0} \frac{-1}{(x+h)x} = -\frac{1}{x^2}.$$

$$\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$$

e) Um die Ableitung des Sinus zu berechnen, benötigen wir die Stetigkeit des Kosinus, den Grenzwert $\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1$ (siehe (2.14) auf Seite 262) und das Additionstheorem (1.29) von Seite 121. Damit gilt:

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\sin(x+h) - \sin(x)}{h} &= \lim_{h \rightarrow 0} \frac{2}{h} \cos\left(\frac{2x+h}{2}\right) \sin\left(\frac{h}{2}\right) \\ &= \lim_{h \rightarrow 0} \cos\left(\frac{2x+h}{2}\right) \cdot \lim_{h \rightarrow 0} \frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}} = \cos(x). \end{aligned}$$

$$\frac{d}{dx} \sin x = \cos x$$

- f) Wir berechnen die Ableitung des natürlichen Logarithmus auf $]0, \infty[$. Dabei verwenden wir das Übertragungsprinzip, die Rechenregeln für den Logarithmus, die zuvor gezeigte Stetigkeit und die Definition von e über eine Folge: Sei $(h_n)_{n=1}^\infty$ eine beliebige Nullfolge mit $h_n \neq 0$, $n \in \mathbb{N}$. Wegen des Übertragungsprinzips genügt es, wenn wir den folgenden Grenzwert für den Differenzenquotienten diskutieren:

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \frac{\ln(x + h_n) - \ln x}{h_n} &= \lim_{n \rightarrow \infty} \frac{1}{h_n} \ln \left(\frac{x + h_n}{x} \right) = \lim_{n \rightarrow \infty} \ln \left(\left[\frac{x + h_n}{x} \right]^{\frac{1}{h_n}} \right) \\
 &= \lim_{n \rightarrow \infty} \ln \left(\left[1 + \frac{h_n}{x} \right]^{\frac{x}{h_n} \cdot \frac{1}{x}} \right) = \frac{1}{x} \cdot \lim_{n \rightarrow \infty} \ln \left(\left[1 + \frac{h_n}{x} \right]^{\frac{x}{h_n}} \right) \\
 &\stackrel{(2.17)}{=} \frac{1}{x} \cdot \ln \left(\lim_{n \rightarrow \infty} \left[1 + \frac{h_n}{x} \right]^{\frac{x}{h_n}} \right) \stackrel{(2.8)}{=} \frac{1}{x} \cdot \ln e = \frac{1}{x}.
 \end{aligned}$$

Dabei haben wir ausgenutzt, dass für jedes einmal gewählte $x > 0$ mit $(h_n)_{n=1}^\infty$ auch die Folge $(\frac{h_n}{x})_{n=1}^\infty$ eine Nullfolge ist.

$$\frac{d}{dx} \ln x = \frac{1}{x}$$

In Definition 2.22 nähern wir uns der Stelle x_0 sowohl von rechts (h positiv) als auch von links (h negativ). Betrachtet man Funktionen, deren Graphen an einer Stelle x_0 einen „Knick“ haben, so existiert dieser beidseitige Grenzwert nicht. Dennoch können die einseitigen Grenzwerte existieren:

Definition 2.23 (Einseitige Ableitungen)

Sei $f : D \rightarrow \mathbb{R}$, und sei $x_0 \in D$.

- Gibt es eine linksseitige Umgebung $]x_0 - \delta_0, x_0] \subset D$, dann heißt die Funktion f **linksseitig differenzierbar** an der Stelle x_0 genau dann, wenn der Grenzwert

$$\lim_{h \rightarrow 0^-} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{h \rightarrow 0^+} \frac{f(x_0 - h) - f(x_0)}{-h} =: f'_-(x_0)$$

existiert.

- Gibt es eine rechtsseitige Umgebung $[x_0, x_0 + \delta[\subset D$, dann heißt die Funktion f **rechtsseitig differenzierbar** an der Stelle x_0 genau dann, wenn der Grenzwert

$$\lim_{h \rightarrow 0^+} \frac{f(x_0 + h) - f(x_0)}{h} =: f'_+(x_0)$$

existiert. Ist f an einer Stelle x_0 rechts- oder linksseitig differenzierbar, so sagt man, f ist **einseitig differenzierbar** an der Stelle x_0 .

Beispiel 2.60

Die Funktion $f(x) = |x|$ ist an der Stelle $x_0 = 0$ rechts- und linksseitig differenzierbar, allerdings mit unterschiedlichen einseitigen Ableitungen, da sie dort einen „Knick“ hat:

$$\lim_{h \rightarrow 0-} \frac{f(h) - f(0)}{h} = \lim_{h \rightarrow 0-} \frac{-h}{h} = -1, \quad \lim_{h \rightarrow 0+} \frac{f(h) - f(0)}{h} = \lim_{h \rightarrow 0+} \frac{+h}{h} = 1. \quad \blacksquare$$

Lemma 2.8 (Zusammenhang Ableitung und einseitiger Ableitungen)

Eine Funktion f ist an einer Stelle x_0 rechtsseitig und linksseitig differenzierbar mit gleichen einseitigen Ableitungen genau dann, wenn f in x_0 differenzierbar ist.

Beweis: Die Aussage folgt direkt aus der entsprechenden Aussage für Grenzwerte (Lemma 2.2). $\square$

Differenzierbarkeit ist wie die Stetigkeit zunächst nur für einzelne Stellen $x_0 \in D$ erklärt.

Definition 2.24 (Differenzierbarkeit auf Intervallen)

Sei $f : D \rightarrow \mathbb{R}$.

- f heißt auf einem Intervall differenzierbar genau dann, wenn f in jedem inneren Punkt des Intervalls differenzierbar und an Randpunkten einseitig differenzierbar ist.
- Die Funktion, die entsteht, indem man einer Stelle x die Ableitung von f an dieser Stelle zuordnet, heißt die von f **abgeleitete Funktion** (kurz ebenfalls **Ableitung**). Diese Funktion wird mit f' oder auch mit $\frac{d}{dx}f(x)$ bezeichnet. Der Funktionswert $f'(x)$ ist also die Ableitung der Funktion f an der Stelle $x_0 := x$.
- f heißt genau dann auf einem Intervall **stetig differenzierbar**, wenn dort f' existiert und stetig ist.

Man beachte, dass aus der Differenzierbarkeit von f in x_0 sofort auch die Stetigkeit von f in x_0 folgt, da insbesondere $\lim_{h \rightarrow 0} f(x_0 + h) - f(x_0) = 0$ gelten muss, damit der Differenzenquotient konvergieren kann:

Satz 2.31 (Differenzierbare Funktionen sind stetig)

Ist $f : [a, b] \rightarrow \mathbb{R}$ differenzierbar in $x_0 \in]a, b[$ oder einseitig differenzierbar in $x_0 = a$ oder $x_0 = b$, so ist f stetig in x_0 .

Beweis: Wir zeigen für $x_0 \in]a, b[$, dass $\lim_{x \rightarrow x_0} f(x) = f(x_0)$ gilt, dass also $\lim_{h \rightarrow 0} f(x_0 + h) - f(x_0) = 0$ ist (analog mit einseitigen Grenzwerten für Randpunkte):

$$\lim_{h \rightarrow 0} f(x_0 + h) - f(x_0) = \lim_{h \rightarrow 0} h \cdot \frac{f(x_0 + h) - f(x_0)}{h}$$

$$= \left[\lim_{h \rightarrow 0} h \right] \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = 0 \cdot f'(x_0) = 0. \quad \square$$

Differenzierbare Funktionen sind also „glatter“ als nur stetige Funktionen. Die Funktionsgraphen stetiger Funktionen sind zwar durchgängig, können aber noch „Knicke“ wie beispielsweise $f(x) = |x|$ für $x = 0$ haben. Die Graphen differenzierbarer Funktionen haben dagegen keine Knicke, da sich an einem Knick rechts- und linksseitige Ableitung unterscheiden würde. Die Funktionsgraphen, die wir üblicherweise zeichnen, gehören zu Funktionen, die nur an wenigen Stellen nicht differenzierbar sind. Es gibt jedoch sogar auf $\mathbb{R}$ definierte stetige Funktionen, die an keiner einzigen Stelle differenzierbar sind (z. B. die Weierstraß-Funktion, siehe Seite 406).

Auch gibt es auf einem Intervall $[a, b]$ differenzierbare Funktionen, deren Ableitung nicht auf $[a, b]$ stetig ist, die also nicht stetig differenzierbar sind. Ein Beispiel ist $f(x) := x^2 \sin\left(\frac{1}{x}\right)$ für $x \in [-1, 1] \setminus \{0\}$ mit $f(0) := 0$. f' ist in 0 nicht stetig.

2.4.2 Ableitungsregeln

Die folgenden Regeln sind ganz fundamental im Umgang mit Funktionen und müssen in der Anwendung beherrscht werden.

Satz 2.32 (Ableitungsregeln)

Seien die Funktionen f und $g : D \rightarrow \mathbb{R}$ differenzierbar in einem inneren Punkt x_0 von D , so gelten die folgenden Rechenregeln:

■ **Linearität der Ableitung** ($c \in \mathbb{R}$):

$$(f + g)'(x_0) = f'(x_0) + g'(x_0), \quad (c \cdot f)'(x_0) = c \cdot f'(x_0),$$

■ **Produktregel:**

$$(f \cdot g)'(x_0) = f'(x_0) \cdot g(x_0) + f(x_0) \cdot g'(x_0),$$

■ **Quotientenregel** (falls $g(x_0) \neq 0$):

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0) \cdot g(x_0) - f(x_0) \cdot g'(x_0)}{g(x_0)^2}.$$

Die Regeln gelten entsprechend auch für einseitige Ableitungen. Aufgrund der Einfachheit der Regeln lässt sich leicht ein Programm schreiben, das Funktionen symbolisch differenziert. Darin unterscheidet sich die Differenzialrechnung von der noch zu behandelnden Integralrechnung.

Leibniz hat die Produkt- und die Quotientenregel 1677 gefunden.

Beweis: Die Linearität folgt sofort aus den Eigenschaften des Grenzwerts für Funktionen. Die Produktregel gilt unter Ausnutzung der Stetigkeit von f und g sowie der Ableitungen:

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{f(x_0 + h)g(x_0 + h) - f(x_0)g(x_0)}{h} \\ &= \lim_{h \rightarrow 0} \left(\frac{f(x_0 + h) - f(x_0)}{h} g(x_0 + h) + f(x_0) \frac{g(x_0 + h) - g(x_0)}{h} \right) \\ &= f'(x_0)g(x_0) + f(x_0)g'(x_0). \end{aligned}$$

Der Trick besteht hier darin, dass wir $0 = -\frac{f(x_0)}{h}g(x_0 + h) + f(x_0)\frac{g(x_0 + h)}{h}$ addiert haben. Diese Technik findet man häufiger in Beweisen. Man spricht von einer „konstruktiven Null“.

Die Quotientenregel lässt sich analog beweisen: Da $g(x_0) \neq 0$ und g stetig in x_0 ist, kann nach Lemma 2.3 o. B. d. A. h so klein gewählt werden, dass auch $g(x_0 + h) \neq 0$.

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{1}{h} \left(\frac{f(x_0 + h)}{g(x_0 + h)} - \frac{f(x_0)}{g(x_0)} \right) \\ &= \lim_{h \rightarrow 0} \frac{1}{g(x_0 + h)g(x_0)} \frac{g(x_0)f(x_0 + h) - g(x_0 + h)f(x_0)}{h} \\ &= \lim_{h \rightarrow 0} \frac{1}{g(x_0 + h)g(x_0)} \left(g(x_0) \frac{f(x_0 + h) - f(x_0)}{h} - f(x_0) \frac{g(x_0 + h) - g(x_0)}{h} \right). \end{aligned}$$

Unter Berücksichtigung der Stetigkeit von g in x_0 , also von $\lim_{h \rightarrow 0} g(x_0 + h) = g(x_0)$, folgt daraus die Aussage des Satzes. $\square$

Die Regel $(c \cdot f)'(x_0) = c \cdot f'(x_0)$ gilt insbesondere für negative Konstanten c . Merkwürdigerweise wird in Klausuren beim Ableiten ein negatives Vorzeichen oft vergessen. Es ist mit $c = -1$:

$$\frac{d}{dx}[-f(x)] = -f'(x), \text{ z. B. } \frac{d}{dx}[-\sin x] = -\cos x.$$

Aus der Quotientenregel erhalten wir sofort eine Ableitungsregel für den Kehrwert einer Funktion:

$$\left(\frac{1}{f} \right)'(x_0) = \frac{0 - 1 \cdot f'(x_0)}{f(x_0)^2} = -\frac{f'(x_0)}{f(x_0)^2}. \quad (2.19)$$

Beispiel 2.61

a) Mit der Linearität erhalten wir $(2x^3 + 5x)' = 6x^2 + 5$.

b) Die Produktregel liefert

$$((x^3 + 7x + 2) \cdot \sin(x))' = (3x^2 + 7) \cdot \sin(x) + (x^3 + 7x + 2) \cdot \cos(x).$$

- c) Als Anwendung der Produktregel betrachten wir nochmals die Funktionen $p_n(x) = x^n$. Mit $\frac{d}{dx}1 = 0$ und $\frac{d}{dx}x = 1$ nutzen wir die Produktregel und zeigen per Induktion, dass $\frac{d}{dx}x^n = nx^{n-1}$, $n \in \mathbb{N}_0$. Der Induktionsanfang für $n = 0$ ist wegen $\frac{d}{dx}x^0 = 0$ bereits gemacht. Unter der Annahme, dass die Regel für n gilt, müssen wir zeigen, dass sie auch für $n + 1$ richtig ist:

$$\frac{d}{dx}x^{n+1} = \frac{d}{dx}[x \cdot x^n] = 1 \cdot x^n + x \cdot nx^{n-1} = (n+1)x^n.$$

Die Regel gilt auch für negative Exponenten. Mit der Quotientenregel erhalten wir für $n \in \mathbb{Z}$, $n < 0$:

$$\frac{d}{dx}x^n = \frac{d}{dx} \frac{1}{x^{-n}} = \frac{0 - 1 \cdot (-nx^{-n-1})}{x^{-2n}} = nx^{n-1}.$$

- d) Mittels Linearität und Quotientenregel können wir nun jede gebrochen-rationale Funktion ableiten, z. B.: $\frac{d}{dx} \frac{x^2+4x+3}{3x+4} = \frac{(2x+4) \cdot (3x+4) - (x^2+4x+3) \cdot 3}{(3x+4)^2}$. ■

Satz 2.33 (Kettenregel)

Sei $f \circ g : D \rightarrow \mathbb{R}$, so dass g differenzierbar in x_0 und f differenzierbar in $g(x_0)$ ist. Dann ist die Funktion $f \circ g$ mit $[f \circ g](x) = f(g(x))$ differenzierbar in x_0 , und es gilt:

$$[f \circ g]'(x_0) = f'(g(x_0))g'(x_0).$$

Man merkt sich die Kettenregel in der Form: „äußere Ableitung mal innere Ableitung“.

Achtung: Bei der Anwendung der Kettenregel werden viele Fehler gemacht. Die Verkettung $[f \circ g](x)$ bedeutet, dass man zunächst den Wert $u = g(x)$ berechnen muss. Dann berechnet man f an der Stelle u . Ist z. B. $f(u) = u^2 + u + 1$ und $g(x) = \sin x$, so ist $[f \circ g](x) = f(g(x)) = \sin^2(x) + \sin(x) + 1$. Jedes Auftreten der Variable von f wird durch $g(x)$ ersetzt. Leitet man nun eine verkettete Funktion $[f \circ g](x)$ ab, so kann man f separat ableiten und diese Ableitung an der Stelle $g(x)$ betrachten: $f'(g(x))$. Das ist aber noch **nicht** die Ableitung der Verkettung. Zusätzlich muss man die innere Funktion g ableiten (man spricht vom Nachdifferenzieren der inneren Funktion) und diese Ableitung mit $f'(g(x))$ multiplizieren: $[f \circ g]'(x) = \frac{d}{dx} f(g(x)) = f'(g(x))g'(x)$. Im Allgemeinen ist $[f \circ g]'(x) \neq f'(g'(x))$, $[f \circ g]'(x) \neq f'(g(x))$, $[f \circ g]'(x) \neq f(g'(x))f'(x)$ usw.

Beweis: Es gilt

$$\begin{aligned} & \frac{f(g(x_0+h)) - f(g(x_0))}{h} \\ &= \frac{g(x_0+h) - g(x_0)}{h} \cdot \begin{cases} \frac{f(g(x_0+h)) - f(g(x_0))}{g(x_0+h) - g(x_0)}, & g(x_0+h) \neq g(x_0) \\ f'(g(x_0)), & g(x_0+h) = g(x_0). \end{cases} \end{aligned}$$

Man beachte, dass im Fall $g(x_0 + h) = g(x_0)$ auf beiden Seiten der Gleichung eine Null steht.

Aufgrund der Stetigkeit von g ist $\lim_{h \rightarrow 0} g(x_0 + h) = g(x_0)$. Mit $h \rightarrow 0$ strebt also auch $\tilde{h} := g(x_0 + h) - g(x_0)$ gegen 0. Ist $\tilde{h} \neq 0$, so lautet der Ausdruck hinter der geschweiften Klammer $\frac{f(g(x_0+h)) - f(g(x_0))}{g(x_0+h) - g(x_0)} = \frac{f(g(x_0)+\tilde{h}) - f(g(x_0))}{\tilde{h}}$, sonst ist er $f'(g(x_0))$. Insgesamt strebt er damit für $h \rightarrow 0$ (und damit $\tilde{h} \rightarrow 0$) gegen $f'(g(x_0))$, und es ist

$$\begin{aligned}(f \circ g)'(x_0) &= \lim_{h \rightarrow 0} \frac{f(g(x_0 + h)) - f(g(x_0))}{h} \\ &= \left(\lim_{h \rightarrow 0} \frac{g(x_0 + h) - g(x_0)}{h} \right) f'(g(x_0)) = g'(x_0) f'(g(x_0)).\end{aligned}$$

□

Auch die Kettenregel gilt unter sinnvollen Voraussetzungen für einseitige Ableitungen.

In Ingenieurbüchern findet man die Kettenregel bisweilen in der Schreibweise

$$\frac{d}{dx}[f(g)] = \frac{df}{dg} \cdot \frac{dg}{dx}.$$

Diese unterstreicht, dass man die äußere Ableitung bildet, indem man die innere Funktion $g(x)$ durch eine Variable (hier g) ersetzt und dann f danach ableitet. Anschließend muss man aber g wieder durch $g(x)$ ersetzen. Wir werden die ausführlichere Schreibweise unter Angabe der Variable x bzw. der Stelle x_0 verwenden.

Beispiel 2.62

a) Es gilt

$$\frac{d}{dx} \sin(x^2) = 2x \cos(x^2),$$

hier ist $g(x) = x^2$ die innere Funktion und $f(y) = \sin y$ die äußere Funktion.

b) $((x^2 - 2x + 3)^4)' = 4(x^2 - 2x + 3)^3 \cdot (2x - 2)$.

c) Wegen $\cos x = \sin(x + \pi/2)$ (siehe Lemma 1.5.9.2 auf Seite 115) ist

$$\begin{aligned}\frac{d}{dx} \cos x &= \frac{d}{dx} \sin\left(x + \frac{\pi}{2}\right) = \cos\left(x + \frac{\pi}{2}\right) \cdot \frac{d}{dx}\left(x + \frac{\pi}{2}\right) \\ &= \cos\left(x + \frac{\pi}{2}\right) = \sin(x + \pi) = -\sin x.\end{aligned}$$

Innere Funktion bei dieser Rechnung ist $g(x) = x + \frac{\pi}{2}$, die äußere ist $f(y) = \sin y$. Mit diesem Ergebnis können wir mittels der Quotientenregel auch die Ableitungen von Tangens und Kotangens ausrechnen:

$$\begin{aligned}\frac{d}{dx} \tan x &= \frac{d}{dx} \frac{\sin x}{\cos x} = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)}, \quad x \neq (2k+1)\frac{\pi}{2}, \\ \frac{d}{dx} \cot x &= \frac{d}{dx} \frac{\cos x}{\sin x} = \frac{-\sin^2(x) - \cos^2(x)}{\sin^2(x)} = -\frac{1}{\sin^2(x)}, \quad x \neq k\pi.\end{aligned}$$

- d) Als Anwendung der Kosinus-Funktion haben wir in Kapitel 1.5.9 einen Wechselstromgenerator diskutiert (siehe Abbildung 1.38 auf Seite 117), bei dem sich der magnetische Fluss durch eine Leiterschleife mit der Zeit t darstellt als $\Phi(t) = \Phi_0 \cdot \cos(ct)$. Nun können wir daraus mit der Kettenregel die induzierte Spannung mit der Lenz'schen Regel $u(t) = -\frac{d}{dt}\Phi(t)$ berechnen:
 $u(t) = -\Phi_0 \cdot [-c \sin(ct)] = \Phi_0 c \sin(ct)$. ■

Wir halten fest:

$$\frac{d}{dx} \cos x = -\sin x, \quad \frac{d}{dx} \tan x = \frac{1}{\cos^2 x}, \quad \frac{d}{dx} \cot x = -\frac{1}{\sin^2 x}.$$

Die Ableitung des ungeraden Sinus ist die gerade Kosinus-Funktion. Deren Ableitung ist die ungerade Funktion $-\sin(x)$. Das ist kein Zufall. Ist f eine gerade, auf $\mathbb{R}$ differenzierbare Funktion (also $f(x) = f(-x)$), so ist f' ungerade (also $f'(-x) = -f'(x)$):

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

Entsprechend ist die Ableitung einer ungeraden Funktion gerade.

Beispiel 2.63

Die aus der Quotientenregel hergeleitete Ableitungsregel (2.19) lässt sich wegen $\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$ auch mit der Kettenregel zeigen. Dazu setzen wir $h(x) := \frac{1}{x}$ und erhalten

$$\left(\frac{1}{f}\right)'(x) = \frac{d}{dx} h(f(x)) = h'(f(x)) f'(x) = -\frac{1}{f(x)^2} f'(x) = -\frac{f'(x)}{f(x)^2}. \quad \blacksquare$$

Beispiel 2.64

Strom $i(t)$ und Spannung $u(t)$ sind an einem Kondensator mit Kapazität C über die Differenzialgleichung $i(t) = C u'(t)$ miteinander verknüpft. Bei einer Wechselspannung $u(t) = \hat{u} \cos(\omega t + \varphi_u)$ ist damit

$$\begin{aligned} i(t) &= -C \hat{u} \omega \sin(\omega t + \varphi_u) = C \hat{u} \omega \sin(\omega t + \varphi_u + \pi) = C \hat{u} \omega \cos\left(\omega t + \varphi_u + \frac{\pi}{2}\right) \\ &= C \hat{u} \omega \cos(\omega t + \varphi_i), \quad \varphi_i := \varphi_u + \frac{\pi}{2}. \end{aligned}$$

Der Strom eilt der Spannung also um $\frac{\pi}{2}$ voraus. Außerdem gilt für die Amplituden $\hat{i}$ und $\hat{u}$ von Strom und Spannung:

$$\frac{\hat{u}}{\hat{i}} = \frac{\hat{u}}{C \hat{u} \omega} = \frac{1}{C \omega}.$$

Der komplexe Wechselstromwiderstand Z_C des Kondensators ist damit (siehe (1.41) auf Seite 146)

$$\begin{aligned} Z_C &= \frac{u(t)}{\dot{i}(t)} = \frac{\hat{u}e^{j[\omega t + \varphi_u]}}{\hat{i}e^{j[\omega t + \varphi_i]}} = \frac{\hat{u}}{\hat{i}}e^{j(\varphi_u - \varphi_i)} \\ &= \frac{1}{C\omega}e^{-j\frac{\pi}{2}} = \frac{1}{C\omega} \left[\cos\left(\frac{\pi}{2}\right) - j \sin\left(\frac{\pi}{2}\right) \right] = -\frac{j}{C\omega}. \end{aligned} \quad (2.20)$$

■

Beispiel 2.65

An einer Spule mit Induktivität L gilt für Spannung $u(t)$ und Strom $i(t)$: $u(t) = Li'(t)$. Bei $i(t) = \hat{i} \cos(\omega t + \varphi_i)$ ist damit

$$\begin{aligned} u(t) &= -L\hat{i}\omega \sin(\omega t + \varphi_i) = L\hat{i}\omega \sin(\omega t + \varphi_i + \pi) = L\hat{i}\omega \cos\left(\omega t + \varphi_i + \frac{\pi}{2}\right) \\ &= L\hat{i}\omega \cos(\omega t + \varphi_u), \quad \varphi_u := \varphi_i + \frac{\pi}{2}. \end{aligned}$$

Die Spannung eilt dem Strom um $\frac{\pi}{2}$ voraus. Außerdem ist

$$\frac{\hat{u}}{\hat{i}} = \frac{L\hat{i}\omega}{\hat{i}} = L\omega.$$

Der komplexe Wechselstromwiderstand Z_L der Spule ist damit (siehe (1.41) auf Seite 146)

$$Z_L = \frac{\hat{u}}{\hat{i}}e^{j(\varphi_u - \varphi_i)} = L\omega e^{j\frac{\pi}{2}} = L\omega \left[\cos\left(\frac{\pi}{2}\right) + j \sin\left(\frac{\pi}{2}\right) \right] = jL\omega. \quad (2.21)$$

■

Die Ableitungsregeln ergänzen sich. Mit der Produkt- und der Kettenregel können wir z. B. die Quotientenregel nachrechnen:

$$\begin{aligned} \frac{d}{dx} \frac{f(x)}{g(x)} &= \frac{d}{dx} [f(x) \cdot (g(x))^{-1}] = \frac{f'(x)}{g(x)} + f(x) \cdot \frac{d}{dx} (g(x))^{-1} \\ &= \frac{f'(x)}{g(x)} + f(x) \cdot [-(g(x))^{-2}] \cdot g'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}. \end{aligned}$$

Hier haben wir aber etwas geschummelt, da wir zuvor $\frac{d}{dx}x^{-1} = -x^{-2}$ mit der Quotientenregel ausgerechnet haben, die wir damit indirekt doch in der Rechnung verwendet haben.

Als weitere Anwendung der Kettenregel leiten wir implizit definierte Funktionen und Umkehrfunktionen ab.

Mit der Kettenregel lassen sich Funktionen $y = f(x)$, die nur in einer impliziten Darstellung $F(x, y) = 0$ gegeben sind, differenzieren. So erhalten wir beispielsweise mit der Kettenregel die Ableitung von $y^2 - x - y = 0$ (also von $f(x)^2 - x - f(x) = 0$) zu $2y \cdot y' - 1 - y' = 0 \iff y' = \frac{1}{2y-1}$ für $y \neq \frac{1}{2}$ (also $f'(x) = \frac{1}{2f(x)-1}$).

Die Ableitung einer Umkehrfunktion lässt sich, wie in Abbildung 2.24 dargestellt, anschaulich am Funktionsgraphen ablesen: Hat f an der Stelle x_0 eine Tangente mit Steigung $f'(x_0)$, dann wird durch Vertauschung der Bezeichnungen von x - und y -Achse diese Tangente zur Tangente an den Graphen von f^{-1} an der Stelle $f(x_0)$. Vertauscht wird auch Höhenzuwachs und Horizontaldifferenz des Steigungsdreiecks der Tangente, so dass $[f^{-1}]'(f(x_0)) = \frac{1}{f'(x_0)}$ der Kehrwert der Ableitung $f'(x_0)$ ist. Dazu muss natürlich $f'(x_0) \neq 0$ sein. Diese Formel erhalten wir auch weniger anschaulich mittels rein formaler Anwendung der Kettenregel:

$$1 = \frac{d}{dx} x = \frac{d}{dx} f^{-1}(f(x)) = [f^{-1}]'(f(x)) \cdot f'(x).$$

Mit $y = f(x)$ ist also

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

Damit man die Kettenregel in dieser Form anwenden darf, muss f differenzierbar an der Stelle x mit $f'(x) \neq 0$ und f^{-1} differenzierbar an der Stelle $y = f(x)$ sein. Eine dazu hinreichende Voraussetzung ist, dass f injektiv (d. h. f^{-1} existiert) und stetig auf einem Intervall $[a, b]$ ist. (Für eine auf einem Intervall stetige Funktion ist, wie Satz 2.29 auf Seite 283 zeigt, die Injektivität gleichbedeutend mit strenger Monotonie.) Ist f dann zusätzlich differenzierbar an einer Stelle $x_0 \in [a, b]$ mit $f'(x_0) \neq 0$, so ist f^{-1} differenzierbar an der Stelle $y_0 := f(x_0)$, und es gilt

$$[f^{-1}]'(f(x_0)) = \frac{1}{f'(x_0)}, \quad [f^{-1}]'(y_0) = \frac{1}{f'(f^{-1}(y_0))}.$$

Man beachte, dass man aus der formalen Anwendung der Kettenregel nicht auf die Differenzierbarkeit von f^{-1} schließen kann. Dazu bemühen wir die Definition der Ableitung:

Beweis: Wir berechnen $[f^{-1}]'(y_0)$ als Grenzwert des Differenzenquotienten mittels des Übertragungsprinzips (Satz 2.22 auf Seite 266). Dazu sei $(y_n)_{n=1}^{\infty}$ eine beliebige Folge im Wertebereich von f mit $\lim_{n \rightarrow \infty} y_n = y_0 = f(x_0)$ und $y_n \neq y_0$ für alle $n \in \mathbb{N}$. Da f injektiv ist, gilt auch $f^{-1}(y_n) \neq f^{-1}(y_0) = x_0$, und wir dürfen durch $f^{-1}(y_n) - f^{-1}(y_0)$ dividieren:

$$\lim_{n \rightarrow \infty} \frac{f^{-1}(y_n) - f^{-1}(y_0)}{y_n - y_0} = \lim_{n \rightarrow \infty} \frac{1}{\frac{y_n - y_0}{f^{-1}(y_n) - f^{-1}(y_0)}} = \frac{1}{\lim_{n \rightarrow \infty} \frac{f(f^{-1}(y_n)) - f(x_0)}{f^{-1}(y_n) - x_0}} = \frac{1}{f'(x_0)},$$

denn wegen der geforderten Stetigkeit von f ist nach Satz 2.28 auch f^{-1} stetig, so dass $\lim_{n \rightarrow \infty} f^{-1}(y_n) = f^{-1}(y_0) = x_0$ ist. Damit liefert die Richtung a) $\implies$ b) des

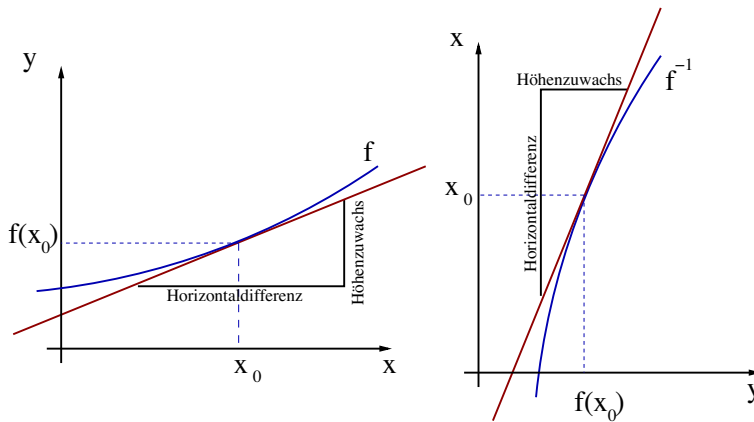


Abb. 2.24: Steigung der Umkehrfunktion an der Stelle $f(x_0)$

Übertragungsprinzips wegen der Differenzierbarkeit von f an der Stelle x_0 den angegebenen Grenzwert $\frac{1}{f'(x_0)}$. Da die Folge $(y_n)_{n=1}^\infty$ beliebig gewählt war, stellt die andere Richtung $b) \Rightarrow a)$ des Übertragungsprinzips sicher, dass

$$[f^{-1}]'(f(x_0)) = [f^{-1}](y_0) = \lim_{y \rightarrow y_0} \frac{f^{-1}(y) - f^{-1}(y_0)}{y - y_0} = \frac{1}{f'(x_0)} = \frac{1}{f'(f^{-1}(y_0))}.$$

□

Beispiel 2.66 (Ableitungen wichtiger Umkehrfunktionen)

a) Da $f^{-1}(x) := x^{\frac{1}{n}}$ für $x > 0$ die Umkehrfunktion zu $f(x) := x^n$ ist, gilt:

$$\frac{d}{dx} x^{\frac{1}{n}} = \frac{1}{f'(f^{-1}(x))} = \frac{1}{n[x^{\frac{1}{n}}]^{n-1}} = \frac{1}{n} \frac{1}{x^{\frac{n-1}{n}}} = \frac{1}{n} x^{\frac{1}{n}-1}, \quad n \in \mathbb{N}.$$

Insbesondere sollten Sie sich merken:

$$\frac{d}{dx} \sqrt{x} = \frac{d}{dx} x^{\frac{1}{2}} = \frac{1}{2\sqrt{x}}.$$

b) Für den Arkussinus erhalten wir

$$\frac{d}{dx} \arcsin x = \frac{1}{\sin'(\arcsin x)} = \frac{1}{\cos(\arcsin x)}, \quad -1 < x < 1.$$

Diesen Term können wir noch vereinfachen. Wegen $-\frac{\pi}{2} < \arcsin x < \frac{\pi}{2}$ ist $\cos(\arcsin x) > 0$, so dass wir die trigonometrische Fassung des Satzes von Pythagoras (1.20) ausnutzen können, um die trigonometrischen Funktionen zu eliminieren:

$$\cos(\arcsin x) = \sqrt{\cos^2(\arcsin x)} = \sqrt{1 - \sin^2(\arcsin x)} = \sqrt{1 - x^2}.$$

$$\frac{d}{dx} \arcsin x = \frac{1}{\sqrt{1 - x^2}}, \quad -1 < x < 1.$$

Entsprechend erhalten wir die Ableitung des Arkuskosinus. Wegen $0 < \arccos x < \pi$ für $x \in]-1, 1[$ ist $\sin(\arccos x) > 0$ und

$$\sin(\arccos x) = \sqrt{\sin^2(\arccos x)} = \sqrt{1 - \cos^2(\arccos x)} = \sqrt{1 - x^2}.$$

$$\frac{d}{dx} \arccos x = \frac{1}{\cos'(\arccos x)} = -\frac{1}{\sin(\arccos x)} = -\frac{1}{\sqrt{1 - x^2}}, \quad -1 < x < 1.$$

Für die Ableitung des Arkustangens nutzen wir für $x \in \mathbb{R}$

$$\frac{1}{\cos^2(\arctan x)} = \frac{\sin^2(\arctan x) + \cos^2(\arctan x)}{\cos^2(\arctan x)} = \tan^2(\arctan x) + 1 = 1 + x^2$$

aus und erhalten

$$\frac{d}{dx} \arctan x = \frac{1}{\tan'(\arctan x)} = \frac{1}{\frac{1}{\cos^2(\arctan x)}} = \frac{1}{1 + x^2}, \quad x \in \mathbb{R}.$$

c) Für die Exponentialfunktion als Umkehrfunktion zu $\ln x$ gilt wegen $\frac{d}{dx} \ln x = \frac{1}{x}$:

$$\frac{d}{dx} \exp(x) = \frac{1}{\ln'(\exp x)} = \frac{1}{\frac{1}{\exp x}} = \exp(x), \quad x \in \mathbb{R}.$$

Die Ableitung der Exponentialfunktion ist die Exponentialfunktion selbst! Dies zeichnet diese Funktion und damit auch die Zahl e , über die sie gebildet wird, aus. Wegen dieser Identität hat die Funktion ihre Bedeutung in den Naturwissenschaften. Einen anderen Weg, um diese wichtige Ableitung zu berechnen, sehen Sie später mit (2.84) auf Seite 416.

d) Damit können wir auch die folgenden Ableitungen berechnen:

$$\frac{d}{dx} a^x = \frac{d}{dx} \exp(x \ln(a)) = \ln(a) \exp(x \ln(a)) = a^x \ln(a), \quad a > 0, x \in \mathbb{R}.$$

$$\frac{d}{dx} x^x = \frac{d}{dx} \exp(x \ln(x)) = \left(\ln(x) + \frac{x}{x} \right) \exp(x \ln(x)) = [\ln(x) + 1] x^x, \quad x > 0. \quad (2.22)$$

Außerdem können wir die bereits bekannte Regel $\frac{d}{dx} x^n = nx^{n-1}$ über rationale Exponenten hinaus für allgemeine Exponenten $a \in \mathbb{R}$ zeigen ($x > 0$):

$$\frac{d}{dx} x^a = \frac{d}{dx} \exp(a \ln(x)) = \frac{a}{x} \exp(a \ln(x)) = \frac{a}{x} x^a = ax^{a-1}.$$

Man beachte, dass diese Regel nicht auf a^x angewendet werden darf, da hier die Variable im Exponenten steht. Das ist ein häufig gemachter Fehler!

■

Beispiel 2.67 (Logarithmische Ableitung)

Für differenzierbare, positive Funktionen f ist die **Logarithmische Ableitung** gegeben durch

$$[\ln f(x)]' = \frac{f'(x)}{f(x)}, \quad f(x) > 0.$$

Für $f(x) = \sin x$ und $x \in]0, \pi[$ erhält man speziell:

$$[\ln \sin x]' = \frac{\cos x}{\sin x} = \cot x.$$

■

Beispiel 2.68 (Ableitung der Hyperbelfunktionen)

Die Ableitungen der Hyperbelfunktionen sind:

$$\begin{aligned} \frac{d}{dx} \sinh x &= \frac{d}{dx} \frac{e^x - e^{-x}}{2} = \frac{e^x + e^{-x}}{2} = \cosh x, \\ \frac{d}{dx} \cosh x &= \frac{d}{dx} \frac{e^x + e^{-x}}{2} = \frac{e^x - e^{-x}}{2} = \sinh x, \\ \frac{d}{dx} \tanh x &= \frac{d}{dx} \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{[e^x + e^{-x}][e^x + e^{-x}] - [e^x - e^{-x}][e^x - e^{-x}]}{[e^x + e^{-x}]^2} \\ &= 1 - \frac{[e^x - e^{-x}]^2}{[e^x + e^{-x}]^2} = 1 - \tanh^2 x, \\ \frac{d}{dx} \coth x &= \frac{d}{dx} \frac{e^x + e^{-x}}{e^x - e^{-x}} = \frac{[e^x - e^{-x}][e^x - e^{-x}] - [e^x + e^{-x}][e^x + e^{-x}]}{[e^x - e^{-x}]^2} \\ &= 1 - \frac{[e^x + e^{-x}]^2}{[e^x - e^{-x}]^2} = 1 - \coth^2 x. \end{aligned}$$

Man beachte, dass sich die Hyperbelfunktionen ähnlich, aber nicht völlig analog zu den trigonometrischen Funktionen verhalten. Während $\frac{d}{dx} \sin x = \cos x$ und $\frac{d}{dx} \sinh x = \cosh x$ eine Parallele ist, weichen $\frac{d}{dx} \cos x = -\sin x$ und $\frac{d}{dx} \cosh x = \sinh x$ bezüglich des Vorzeichens voneinander ab.

■

Tab. 2.1: Häufig benötigte Ableitungen

Funktion f	Ableitung f'
$f(x) = x^n, \quad x \in \mathbb{R}, \quad n \in \mathbb{N}$	$f'(x) = n \cdot x^{n-1}$
$f(x) = x^a, \quad x > 0, \quad a \in \mathbb{R}$	$f'(x) = a \cdot x^{a-1}$
$f(x) = a^x, \quad a > 0, \quad x \in \mathbb{R}$	$f'(x) = a^x \cdot \ln a$
$f(x) = e^x, \quad x \in \mathbb{R}$	$f'(x) = e^x$
$f(x) = \ln x, \quad x \in]0, \infty[$	$f'(x) = \frac{1}{x}$
$f(x) = \sinh x, \quad x \in \mathbb{R}$	$f'(x) = \cosh x$
$f(x) = \cosh x, \quad x \in \mathbb{R}$	$f'(x) = \sinh x$
$f(x) = \tanh x, \quad x \in \mathbb{R}$	$f'(x) = 1 - \tanh^2 x$
$f(x) = \coth x, \quad x \in \mathbb{R}, \quad x \neq 0$	$f'(x) = 1 - \coth^2 x$
$f(x) = \sin x, \quad x \in \mathbb{R}$	$f'(x) = \cos x$
$f(x) = \cos x, \quad x \in \mathbb{R}$	$f'(x) = -\sin x$
$f(x) = \tan x, \quad x \in \mathbb{R}, \quad x \neq \frac{\pi}{2}(2k+1)$	$f'(x) = \frac{1}{\cos^2 x}$
$f(x) = \cot x, \quad x \in \mathbb{R}, \quad x \neq k\pi$	$f'(x) = -\frac{1}{\sin^2 x}$
$f(x) = \arcsin x, \quad x \in]-1, 1[$	$f'(x) = \frac{1}{\sqrt{1-x^2}}$
$f(x) = \arccos x, \quad x \in]-1, 1[$	$f'(x) = -\frac{1}{\sqrt{1-x^2}}$
$f(x) = \arctan x, \quad x \in \mathbb{R}$	$f'(x) = \frac{1}{1+x^2}$
$f(x) = \operatorname{arccot} x, \quad x \in \mathbb{R}$	$f'(x) = -\frac{1}{1+x^2}$

2.4.3 Newton-Verfahren

Mit Hilfe der Ableitung kann man das Intervallschachtelungsverfahren zur Bestimmung von Nullstellen von Seite 279 erheblich verbessern (in dem Sinne, dass man die Nullstelle wesentlich schneller eingrenzen kann). Dazu startet man an einer Stelle x_1 , die nahe bei der Nullstelle x_0 liegen sollte. Nun ersetzt man die Funktion f durch die Tangente g_1 an f im Punkt x_1 :

$$g_1(x) = f(x_1) + f'(x_1)(x - x_1).$$

Falls $f'(x_1) \neq 0$, hat diese Gerade genau die eine Nullstelle

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

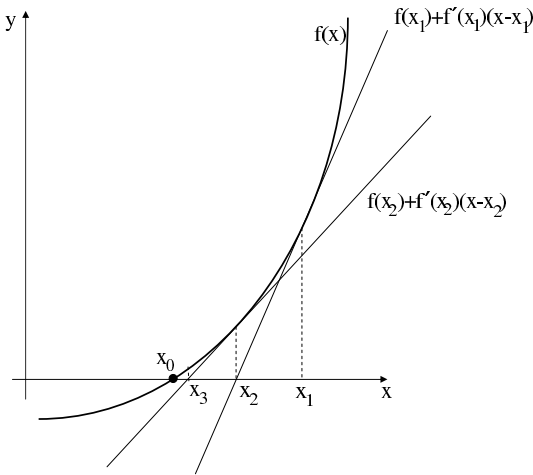


Abb. 2.25: Zwei Iterationen des Newton-Verfahrens

Wenn die Tangente g_1 die Funktion f genügend gut approximiert, dann wird der Punkt x_2 näher bei x_0 liegen als der Startpunkt x_1 . Jetzt setzt man das Verfahren für x_2 wie zuvor für x_1 fort, indem man die Tangente g_2 an f in x_2 bestimmt usw. (siehe Abbildung 2.25). Man erhält die Folge $(x_n)_{n=1}^{\infty}$ mit

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad (2.23)$$

die gegen die Nullstelle x_0 konvergieren soll. Man nennt dieses Iterationsverfahren das **Newton-Verfahren**. Die gute Nachricht ist: Wenn die Folge $(x_n)_{n=1}^{\infty}$ gegen eine Zahl x_0 konvergiert, dann ist der Grenzwert x_0 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' :

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

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

In einer Anwendung können wir also in ein 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.

Man kann mit Satz 2.38, den wir in einem Einschub auf Seite 318 vorstellen, zeigen, dass das Verfahren konvergiert, wenn man $x_1 \in]x_0 - \delta, x_0 + \delta[$ wählt und wenn für alle $x \in]x_0 - \delta, x_0 + \delta[$ sowohl $f'(x) \neq 0$ ist, als auch gilt:

$$\frac{|f(x)| |f''(x)|}{(f'(x))^2} < \lambda < 1. \quad (2.24)$$

Ist der Quotient $\frac{|f^{(2)}(x)|}{(f'(x))^2}$ beschränkt, so kann man die Bedingung erfüllen, indem man δ klein genug und damit den Startpunkt nahe bei x_0 wählt. Denn hier sind die Funktionswerte von f klein wegen $\lim_{x \rightarrow x_0} f(x) = 0$.

Beispiel 2.69

Mit dem Newton-Verfahren berechnen wir ausgehend von $x_0 = \frac{1}{2}$ die ersten zwei Iterationen, um eine Nullstelle von $f(x) = x^2 - 1$ zu finden: Hier ist

$$x_{n+1} = x_n - \frac{x_n^2 - 1}{2x_n}.$$

Wir starten bei $x_0 = \frac{1}{2}$:

$$x_1 = \frac{1}{2} + \frac{\frac{3}{4}}{1} = \frac{5}{4}, \quad x_2 = \frac{5}{4} - \frac{\frac{9}{16}}{\frac{5}{2}} = \frac{50}{40} - \frac{9}{40} = \frac{41}{40}.$$

Offensichtlich hat $f(x) = x^2 - 1 = (x - 1)(x + 1)$ die Nullstelle 1. Auch die Newton-Iterierten x_n , $n = 1, 2, \dots$ konvergieren gegen 1. ■

Im Newton-Verfahren wird eine Nullstelle einer Funktion f durch die Nullstellen von Tangenten angenähert, die lokal die Funktion f gut beschreiben. Falls man keine Ableitungen berechnen möchte, kann man statt der Tangenten auch Sekanten verwenden, die im Allgemeinen die Funktion aber nicht ganz so gut wie Tangenten lokal annähern. Das resultierende **Sekantenverfahren** sieht dann so aus: Wir starten mit zwei Stellen x_1 und x_2 in der Nähe einer vermuteten Nullstelle. Dann berechnen wir eine Folge mit Gliedern x_n für $n \geq 2$ als Nullstellen der Sekanten

$$g_{n+1}(x) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_n) + f(x_n)$$

durch die Punkte $(x_{n-1}, f(x_{n-1}))$ und $(x_n, f(x_n))$:

$$g_{n+1}(x_{n+1}) = 0 \iff x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}f(x_n).$$

Das gelingt, falls stets $f(x_n) \neq f(x_{n-1})$ ist. Falls die Folge konvergiert und f im betrachteten Intervall differenzierbar mit einer Ableitung ohne Nullstellen ist, strebt die Folge gegen eine Nullstelle x_0 von f . In der Regel nähert sich dann das Verfahren immer noch wesentlich schneller der Nullstelle als das Intervallschachtelungsverfahren. Allerdings besitzt das Sekantenverfahren (aber auch das Newton-Verfahren) nicht die Eigenschaft, dass x_0 stets zwischen x_n und x_{n+1} liegt.

2.4.4 Das Differenzial

In diesem Abschnitt schauen wir uns den Hintergrund der Schreibweise $\frac{df}{dx}$ näher an und gehen darauf ein, wie mit dem Symbol df in der Ingenieur-Mathematik gerechnet wird. Inhaltlich geschieht nichts Neues.

Ist f eine Funktion mit der Variable x und ist x_0 eine feste Stelle aus dem Definitionsbereich von f , so bezeichnet man zur Horizontaldifferenz

$$dx = \Delta x := x - x_0$$

den Höhenzuwachs (die **Wertedifferenz**) mit

$$\Delta f := f(x) - f(x_0) = f(x_0 + \Delta x) - f(x_0).$$

Δf gibt also die Änderung der Funktionswerte bei einer Änderung des Arguments x_0 um Δx an. Für kleine Werte von Δx erhalten wir die Näherung (vgl. mit dem Mittelwertsatz (Seite 315) und dem Satz von Taylor (Seite 384), die wir beide später behandeln)

$$\Delta f = \frac{f(x) - f(x_0)}{\Delta x} \cdot \Delta x \approx f'(x_0) \cdot \Delta x.$$

Definition 2.25 (Differenzial)

Das **Differenzial** df einer differenzierbaren Funktion f an einer Stelle x_0 ist eine Abbildung, die die Horizontaldifferenz $dx = \Delta x = x - x_0$ abbildet auf den Wert $f'(x_0) \cdot \Delta x$:

$$df = df(\Delta x) := f'(x_0) \cdot \Delta x.$$

Das Differenzial df hängt von der gewählten Stelle x_0 ab und beschreibt den linearen Anteil des Wertezuwachses Δf der Funktion in Abhängigkeit des Zuwachses $\Delta x = dx$. Mit ihm lässt sich näherungsweise der Funktionsgraph von f in einer kleinen Umgebung des **Arbeitspunktes** $(x_0, f(x_0))$ darstellen als Gerade $g(x) := f(x_0) + df(x - x_0)$, die eine Tangente an den Funktionsgraphen beschreibt, also

$$f(x_0 + \Delta x) \approx g(x_0 + \Delta x) = f(x_0) + df(\Delta x) = f(x_0) + df.$$

Mit dem Begriff des Differenzials wird die Schreibweise $\frac{df}{dx}$ für die Ableitung verständlich:

$$\frac{df}{dx} = \frac{df(dx)}{dx} = \frac{f'(x_0) \cdot \Delta x}{\Delta x} = f'(x_0).$$

Beispiel 2.70

a) Nahe bei der Stelle $x_0 = 0$ ersetzen wir $\sin x$ und $\cos x$ durch die Tangente an den Funktionsgraphen, die wir mit dem Differenzial schreiben:

- Für $f(x) = \sin x$ mit $f'(x) = \cos x$ gilt bei $x_0 = 0$: $f(0) = 0$, $f'(0) = 1$ und damit $df = 1 \cdot dx$. Die Tangente im Arbeitspunkt $(0, 0)$ hat die Geradengleichung $g(dx) = dx$, d. h., für kleine $|x|$ erhält man die gängige Approximation $\sin x \approx x$, die man auch am Einheitskreis ablesen kann (siehe Kapitel 1.5.9).

- Für $y = f(x) = \cos x$ mit $f'(x) = -\sin x$ gilt bei $x_0 = 0$: $f(0) = 1$, $f'(0) = 0$ und damit $df = 0$. Die Tangente im Arbeitspunkt $(0, 1)$ hat die Gleichung $g(dx) = 1$, d. h., für kleine Werte $|x|$ hat man $\cos x \approx 1$.

b) Für die Funktion $f(x) = \sin x$ vergleichen wir Wertedifferenz und Differenzial an der allgemeinen Stelle x_0 :

$$\begin{aligned}\Delta f &= \sin(x_0 + \Delta x) - \sin(x_0) \\ &\stackrel{(1.23)}{=} \sin(x_0) \cdot \cos(\Delta x) + \cos(x_0) \cdot \sin(\Delta x) - \sin(x_0) \\ &= \sin(x_0)(\cos(\Delta x) - 1) + \cos(x_0) \cdot \sin(\Delta x),\end{aligned}$$

und für kleine $|\Delta x|$ ist $\cos(\Delta x) - 1 \approx 0$ und $\sin(\Delta x) \approx \Delta x$, so dass aus der Wertedifferenz annähernd das Differenzial an der Stelle x_0 für das Argument Δx wird:

$$df(\Delta x) = \cos(x_0) \cdot \Delta x. \quad \blacksquare$$

Eine wichtige Anwendung des Differenzials ist die **Fehlerrechnung**. Ist $f(x)$ eine gegebene Funktion, dann ist in der Praxis die Frage von Bedeutung, welche Auswirkung eine Änderung (Fehler) der Argumente von x_0 zu $x_0 + \Delta x$ auf die zugehörigen Funktionswerte hat. Dabei heißt die Wertedifferenz $\Delta f = f(x_0 + \Delta x) - f(x_0)$ der **absolute Fehler**. Ist $f(x_0) \neq 0$, so nennt man $\frac{\Delta f}{f(x_0)}$ den **relativen Fehler**. Indem man die Wertedifferenz durch den Wert $df(\Delta x)$ des Differenzials zur Stelle x_0 ersetzt, erhält man Näherungswerte für den absoluten und relativen Fehler, vgl. (Dobner und Engelmann, 2002, Band 1, S. 150).

Beispiel 2.71

Aus einer Messung der Kantenlänge a eines Würfels wird sein Volumen $V = V(a) = a^3$ berechnet. Bei der Messung ist ein Messfehler unvermeidbar. Wenn bekannt ist, dass der maximale Messfehler $da = \Delta a$ beträgt, interessiert uns, wie groß der relative Fehler des berechneten Volumens maximal ist.

Mit dem Differenzial erhalten wir für den maximalen absoluten Fehler

$$\Delta V \approx dV = \frac{dV(a)}{da} \Delta a = 3a^2 \Delta a$$

und damit für den relativen Fehler

$$\frac{\Delta V}{V} \approx \frac{dV}{V} = \frac{3a^2 \Delta a}{a^3} = \frac{3\Delta a}{a}.$$

Wählen wir konkret den maximalen Messfehler $\Delta a = \pm 0,1$ cm und die gemessene Kantenlänge $a = 100$ cm, so erhalten wir für den (maximalen) relativen Fehler die Näherung $\frac{dV}{V} = \frac{3 \cdot (\pm 0,1)}{100} = \pm 0,003$, also einen relativen Fehler von maximal 0,3 %. Wir greifen das Thema Fehlerrechnung im Rahmen der mehrdimensionalen Analysis nochmals auf. ■

2.4.5 Höhere Ableitungen

Die Ableitung einer Funktion kann unter Umständen auch wieder abgeleitet werden. So kommt man zu höheren Ableitungen, die iterativ berechnet werden. Als Schreibweise für die n -te Ableitung verwenden wir:

$$f^{(n)}(x) = \frac{d^n}{dx^n} f(x) := (f^{(n-1)})'(x) = \frac{d}{dx} f^{(n-1)}(x),$$

wobei $f^{(1)}(x) = f'(x)$ und $f^{(0)}(x) = f(x)$. Zweite und dritte Ableitungen schreibt man auch als $f''(x) := f^{(2)}(x)$ und $f'''(x) := f^{(3)}(x)$. Die Angabe, wie oft abzuleiten ist, wird als Index oben an die Funktion geschrieben. Damit es zu keinen Verwechslungen mit Potenzen kommt, sind die Klammern erforderlich.

Achtung: Es ist häufig bequem, die Funktion als ihre nullte Ableitung zu bezeichnen: $f(x) = f^{(0)}(x)$. So werden künftig Fallunterscheidungen vermieden.

Definition 2.26 (Stetige Differenzierbarkeit)

Ist f auf einem Intervall n -mal differenzierbar und $f^{(n)}$ dort stetig, so heißt f **n -mal stetig differenzierbar** auf diesem Intervall.

Man beachte dabei, dass auch die Funktionen $f^{(k)}$, $0 \leq k < n$, stetig sind, da sie sogar differenzierbar sind. Je häufiger eine Funktion (stetig) differenzierbar ist, desto „glatter“ ist ihr Funktionsgraph. Stetige Funktionen können noch „Knicke“ haben, bei einmal stetig differenzierbaren Funktionen kann die erste Ableitung „Knicke“ haben. Bei einer Weg-Zeit-Funktion entspricht ein solcher Knick einem Sprung in der Beschleunigung.

Beispiel 2.72

$$\begin{aligned} \frac{d}{dx} \sin x &= \cos x, & \frac{d^2}{dx^2} \sin x &= \frac{d}{dx} \cos x = -\sin x, \\ \frac{d^3}{dx^3} \sin x &= -\frac{d}{dx} \sin x = -\cos x, & \frac{d^4}{dx^4} \sin x &= -\frac{d}{dx} \cos x = \sin x, \\ \frac{d^n}{dx^n} x^m &= m(m-1)(m-2) \cdots (m-n+1)x^{m-n} = \frac{m!}{(m-n)!} x^{m-n}, \quad n \leq m. \end{aligned}$$

■

Die Ableitungsregeln aus Satz 2.1 von Seite 220 lassen sich auf höhere Ableitungen übertragen, indem man sie mehrfach anwendet. So ist

$$\frac{d^n}{dx^n} [cf(x) + dg(x)] = c \frac{d^n}{dx^n} f(x) + d \frac{d^n}{dx^n} g(x).$$

Die Produktregel wird zur **Leibniz-Regel**:

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

Beweis: Wir müssen eine Aussage für jedes $n \in \mathbb{N}$ beweisen und tun dies mittels Vollständiger Induktion. Der Induktionsanfang für $n = 1$ ist genau die Produktregel. Zum Schritt von n nach $n + 1$ nutzen wir die Linearität der Ableitung und ebenfalls die Produktregel: Die Aussage gelte für n (Induktionsannahme). Wir zeigen sie damit für $n + 1$:

$$\begin{aligned} (f \cdot g)^{(n+1)}(x) &= \frac{d}{dx} \left[\sum_{k=0}^n \binom{n}{k} f^{(k)}(x) g^{(n-k)}(x) \right] \\ &= \sum_{k=0}^n \binom{n}{k} [f^{(k+1)}(x) g^{(n-k)}(x) + f^{(k)}(x) g^{(n+1-k)}(x)] \\ &= \sum_{k=1}^{n+1} \binom{n}{k-1} f^{(k)}(x) g^{(n+1-k)}(x) + \sum_{k=0}^n \binom{n}{k} f^{(k)}(x) g^{(n+1-k)}(x). \end{aligned}$$

Jetzt fassen wir die überlappenden Summanden beider Summen zusammen:

$$\begin{aligned} (f \cdot g)^{(n+1)}(x) &= \sum_{k=1}^n \left[\binom{n}{k-1} + \binom{n}{k} \right] f^{(k)}(x) g^{(n+1-k)}(x) \\ &\quad + \binom{n}{n} f^{(n+1)}(x) g^{(0)}(x) + \binom{n}{0} f^{(0)}(x) g^{(n+1)}(x) \\ &\stackrel{(1.2)}{=} \sum_{k=1}^n \binom{n+1}{k} f^{(k)}(x) g^{(n+1-k)}(x) \\ &\quad + \binom{n+1}{n+1} f^{(n+1)}(x) g^{(0)}(x) + \binom{n+1}{0} f^{(0)}(x) g^{(n+1)}(x) \\ &= \sum_{k=0}^{n+1} \binom{n+1}{k} f^{(k)}(x) g^{(n+1-k)}(x). \end{aligned}$$

□

Beispiel 2.73

Wir wenden die Leibniz-Regel an:

$$\begin{aligned} \frac{d^2}{dx^2}(x^2 \sin x) &= \sum_{k=0}^2 \binom{2}{k} (x^2)^{(k)} (\sin x)^{(2-k)} = x^2(-\sin x) + 2 \cdot 2x \cos(x) + 2 \sin(x) \\ &= (2 - x^2) \sin(x) + 4x \cos(x). \end{aligned}$$

■

Beispiel 2.74 (Kubische Spline-Interpolation)

In Kapitel 1.5.6.2 haben wir die Lagrange-Interpolation betrachtet, bei der ein Polynom gesucht wird, das an vorgegebenen Stützstellen ebenfalls vorgegebene Funktionswerte annimmt. Ein Polynom vom Grad n hat (unter Berücksichtigung der Vielfachheit) n Nullstellen. Je mehr Stützstellen vorliegen, desto größer wird n und desto „wilder“ oszilliert das Interpolationspolynom. Hier ist es oft günstiger, eine Interpolationsfunktion stückweise über Polynome kleinen Grades so zusammenzusetzen, dass die Übergänge möglichst glatt sind. Entsprechende Interpolationsfunktionen heißen **Splines**. Der Begriff stammt aus dem Schiffsbau, bei dem dünne Holzplanken über Stützstellen gebogen und aneinandergesetzt werden. Die Glattheitsbedingungen an den Übergängen können wir nun mittels Ableitungen formulieren. Dazu sehen wir uns in diesem Beispiel kubische Splines an, bei denen die aneinandergesetzten Polynome alle vom Grad drei sind.

Gegeben seien dazu $n + 1$ Stützstellen $a = x_0 < x_1 < \dots < x_n = b$ und Stützwerte $y_0, y_1, \dots, y_n$. Gesucht ist eine auf $[a, b]$ definierte **kubische Splinefunktion**

$$s(x) := \begin{cases} s_0(x), & x_0 \leq x < x_1, \\ s_1(x), & x_1 \leq x < x_2, \\ \vdots & \\ s_{n-1}(x), & x_{n-1} \leq x \leq x_n, \end{cases} \quad (2.25)$$

mit Polynomen

$$s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad i \in \{0, 1, \dots, n-1\} \quad (2.26)$$

vom Grad drei, die die folgenden Bedingungen erfüllen:

$$\begin{aligned} s_i(x_i) &= y_i \text{ für } i \in \{0, 1, \dots, n-1\} \text{ und } s_{n-1}(x_n) = y_n \text{ (Interpolation),} \\ s_i(x_{i+1}) &= s_{i+1}(x_{i+1}) = y_{i+1} \text{ für } i \in \{0, 1, \dots, n-2\} \text{ (Stetigkeit),} \\ s'_i(x_{i+1}) &= s'_{i+1}(x_{i+1}) \text{ für } i \in \{0, 1, \dots, n-2\} \text{ (Differenzierbarkeit),} \\ s''_i(x_{i+1}) &= s''_{i+1}(x_{i+1}) \text{ für } i \in \{0, 1, \dots, n-2\} \text{ (zweimalige Differenzierbarkeit).} \end{aligned}$$

Zusätzlich werden am linken und rechten Intervallrand Werte der ersten oder zweiten Ableitung vorgegeben. Wählt man $s''_0(x_0) = s''_{n-1}(x_n) = 0$, so spricht man von einem **natürlichen Spline**. Legt man dagegen hier die erste Ableitung mit vorgegebenen Werten fest, so ist ein **Hermite'scher-Spline** gesucht.

Wir zeigen jetzt, dass durch die Bedingungen und die Randbedingung eines natürlichen Splines genau ein kubischer Spline definiert ist. Außerdem leiten wir ein Gleichungssystem und Formeln her, über die die Koeffizienten a_i , b_i , c_i und d_i berechnet werden können.

Offensichtlich ist der erste Teil der Interpolationsbedingung äquivalent mit

$$a_i = y_i \text{ für } i \in \{0, 1, \dots, n-1\},$$

so dass die Koeffizienten a_i bereits festgelegt sind. Zur Abkürzung setzen wir

$$h_i := x_{i+1} - x_i \text{ und } \Delta y_i := y_{i+1} - y_i \text{ für } i \in \{0, 1, \dots, n-1\}.$$

Damit erhalten wir aus der Stetigkeitsbedingung, der Interpolation an der Stelle x_n , sowie den beiden Ableitungsbedingungen

$$a_i + b_i \cdot h_i + c_i \cdot h_i^2 + d_i \cdot h_i^3 = y_{i+1} \text{ für } i \in \{0, 1, \dots, n-1\}, \quad (2.27)$$

$$b_i + 2c_i \cdot h_i + 3d_i \cdot h_i^2 = b_{i+1} \text{ für } i \in \{0, 1, \dots, n-2\}, \quad (2.28)$$

$$2c_i + 6d_i \cdot h_i = 2c_{i+1} \text{ für } i \in \{0, 1, \dots, n-2\}. \quad (2.29)$$

Wegen $s''(x_0) = 0$ ist $c_0 = 0$, und wegen $s''_{n-1}(x_n) = 0$ folgt

$$2c_{n-1} + 6d_{n-1} \cdot h_{n-1} = 0. \quad (2.30)$$

Die Koeffizienten d_i und b_i lassen sich direkt aus den Werten c_i berechnen: Mit $c_n := 0$ folgt aus (2.29) und (2.30):

$$d_i = \frac{2c_{i+1} - 2c_i}{6h_i} = \frac{c_{i+1} - c_i}{3h_i} \text{ für } i \in \{0, 1, \dots, n-1\}. \quad (2.31)$$

Mit d_i werden bei bekannten c_i auch die b_i für $i \in \{0, 1, \dots, n-1\}$ mit (2.27) berechnet:

$$\begin{aligned} b_i &= \frac{1}{h_i} (y_{i+1} - a_i - c_i \cdot h_i^2 - d_i \cdot h_i^3) = \frac{1}{h_i} (\Delta y_i - c_i \cdot h_i^2 - d_i \cdot h_i^3) \\ &= \frac{1}{h_i} \left(\Delta y_i - c_i \cdot h_i^2 - \frac{c_{i+1} - c_i}{3} h_i^2 + \frac{c_i}{3} h_i^2 \right) = \frac{1}{h_i} \Delta y_i - \frac{2}{3} h_i \cdot c_i - \frac{1}{3} h_i \cdot c_{i+1}. \end{aligned} \quad (2.32)$$

Jetzt müssen wir nur noch die c_i finden. Dazu setzen wir (2.31) und (2.32) in (2.28) ein, $i \in \{0, 1, \dots, n-2\}$:

$$\begin{aligned} &\frac{1}{h_i} \Delta y_i - \frac{2}{3} h_i \cdot c_i - \frac{1}{3} h_i \cdot c_{i+1} + 2h_i \cdot c_i + h_i(c_{i+1} - c_i) \\ &= \frac{1}{h_{i+1}} \Delta y_{i+1} - \frac{2}{3} h_{i+1} \cdot c_{i+1} - \frac{1}{3} h_{i+1} \cdot c_{i+2} \\ \iff &\frac{1}{3} h_i \cdot c_i + \frac{2}{3} (h_i + h_{i+1}) \cdot c_{i+1} + \frac{1}{3} h_{i+1} \cdot c_{i+2} = \frac{1}{h_{i+1}} \Delta y_{i+1} - \frac{1}{h_i} \Delta y_i \\ \iff &h_i \cdot c_i + 2(h_i + h_{i+1}) \cdot c_{i+1} + h_{i+1} \cdot c_{i+2} = 3 \left(\frac{1}{h_{i+1}} \Delta y_{i+1} - \frac{1}{h_i} \Delta y_i \right). \end{aligned}$$

Damit erhalten wir ein lineares Gleichungssystem für $c_1, c_2, \dots, c_{n-1}$ mit einer symmetrischen Tridiagonalmatrix (beachte $c_0 = c_n = 0$):

$$\begin{bmatrix} 2(h_0 + h_1) & h_1 & 0 & \dots & 0 \\ h_1 & 2(h_1 + h_2) & h_2 & \ddots & \vdots \\ 0 & h_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & h_{n-2} \\ 0 & \dots & 0 & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{bmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{pmatrix} = 3 \begin{pmatrix} \frac{\Delta y_1}{h_1} - \frac{\Delta y_0}{h_0} \\ \frac{\Delta y_2}{h_2} - \frac{\Delta y_1}{h_1} \\ \vdots \\ \frac{\Delta y_{n-1}}{h_{n-1}} - \frac{\Delta y_{n-2}}{h_{n-2}} \end{pmatrix}. \quad \blacksquare$$

2.5 Zentrale Sätze der Differenzialrechnung

2.5.1 Satz von Fermat: notwendige Bedingung für lokale Extrema

An Stellen lokaler Extrema muss die Ableitung null sein: Hier geht es augenblicklich weder weiter bergauf noch bergab.

Satz 2.34 (Fermat, 1601–1665, hier 1630)

Sei $f : D \rightarrow \mathbb{R}$, $x_0 \in D$ ein innerer Punkt, und sei f differenzierbar in x_0 . Hat f in x_0 ein lokales Extremum, dann ist (notwendigerweise) $f'(x_0) = 0$.

Beweis: Wir beweisen die Aussage für ein lokales Maximum, für ein lokales Minimum verläuft der Beweis völlig analog. Da x_0 ein innerer Punkt ist, existiert wegen des lokalen Maximums ein $\delta > 0$, so dass

$$f(x_0 \pm h) \leq f(x_0)$$

für alle $0 \leq h < \delta$. Damit gilt für den Differenzenquotienten für $0 < h < \delta$

$$\frac{f(x_0 + h) - f(x_0)}{h} \leq 0 \leq \frac{f(x_0 - h) - f(x_0)}{-h}. \quad (2.33)$$

Da f differenzierbar in x_0 ist, ist f hier insbesondere rechts- und linksseitig differenzierbar mit $f'_-(x_0) = f'_+(x_0) = f'(x_0)$. Für $h \rightarrow 0+$ ist wegen (2.33) aber $f'_+(x_0) \leq 0 \leq f'_-(x_0)$ (siehe Lemma 2.3), so dass $f'(x_0) = 0$. $\square$

Dies ist eine notwendige Bedingung (siehe Seite 19) für die Existenz eines lokalen Extremums. Ist also $f'(x_0) \neq 0$, so kann an der Stelle x_0 kein Extremum liegen. Ist dagegen $f'(x_0) = 0$, so kann hier ein Extremum liegen, das ist aber nicht zwingend und muss weiter untersucht werden. Zum Beispiel ist für $f(x) = x^3$ zwar $f'(0) = 3 \cdot 0^2 = 0$, aber die Funktion hat hier weder ein lokales Minimum noch ein lokales Maximum. Ist $f'(x_0) = 0$, ohne dass ein Extremum vorliegt, so spricht man von einem **Sattelpunkt**.

Beispiel 2.75

Die Funktion $f(x) := (x + 2)^2$ ist eine Parabel mit der Nullstelle -2 . Da die Funktionswerte nicht negativ werden, hat f in -2 ein lokales und globales Minimum. Wir erwarten also $f'(-2) = 0$, was wegen $f'(x) = 2(x + 2)$ auch stimmt. $\blacksquare$

Nullstellensuche und die Suche nach lokalen Extrema hängen also sehr eng zusammen. Beispielsweise kann man das Newton-Verfahren (Seite 305) nutzen, um Nullstellen der Ableitung und damit Kandidaten für lokale Extremstellen zu finden. Umgekehrt lassen sich zumindest einige Nullstellen einer Ableitung f' über lokale Extrema von f finden.

2.5.2 Mittelwertsätze der Differenzialrechnung

Wir kommen nun zu den Mittelwertsätzen und beginnen mit der einfachsten Variante, auf der die anderen Formulierungen basieren:

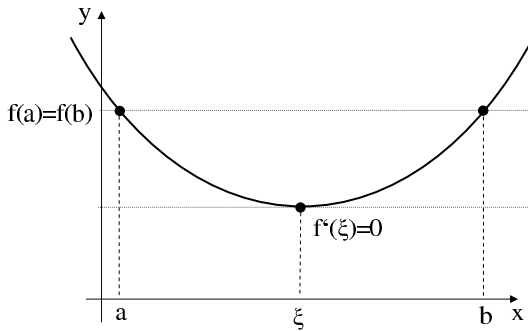


Abb. 2.26: Satz von Rolle

Satz 2.35 (Satz von Rolle, 1652–1719, hier 1680)

Sei f stetig auf $[a, b]$ und differenzierbar auf $]a, b[$ mit $f(a) = f(b)$. Dann existiert mindestens eine Stelle $\xi \in]a, b[$ mit $f'(\xi) = 0$.

Hat eine stetige Funktion an den Endpunkten eines Intervalls die gleichen Funktionswerte, so gibt es dazwischen eine horizontale Tangente an den Funktionsgraphen (siehe Abbildung 2.26).

Beweis: Da f stetig auf $[a, b]$ ist, nimmt f hier Maximum und Minimum an. Falls $f(x) = f(a) = f(b)$ für alle $x \in [a, b]$ vorliegt, ist die Aussage des Satzes trivial, da $f(x) = c$ und $f'(x) = 0$. Anderenfalls gibt es mindestens eine Stelle $\xi \in]a, b[$, an der f ein Maximum oder Minimum annimmt. Damit muss hier aber notwendigerweise $f'(\xi) = 0$ sein. $\square$

Jetzt können wir den eigentlichen Mittelwertsatz von Lagrange (1736–1813) beweisen:

Satz 2.36 (Mittelwertsatz der Differenzialrechnung)

Sei f stetig auf $[a, b]$ und differenzierbar auf $]a, b[$. Dann existiert mindestens eine Stelle $\xi \in]a, b[$, so dass

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

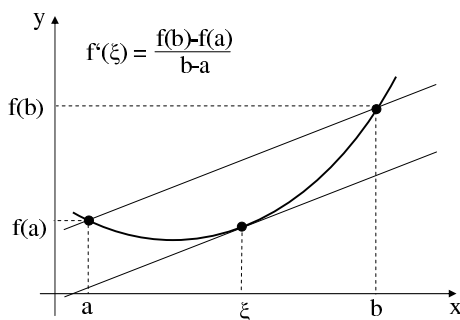


Abb. 2.27: Mittelwertsatz

Der Satz besagt also, dass es eine Stelle ξ gibt, an der die Tangente an den Funktionsgraphen genau die gleiche Steigung hat wie die Gerade

$$g(x) := f(a) + \frac{f(b) - f(a)}{b - a} \cdot (x - a)$$

durch die Punkte $(a, f(a))$ und $(b, f(b))$ (siehe Abbildung 2.27). Betrachten wir den Funktionsgraphen als Weg-Zeit-Diagramm, bei dem auf der x -Achse die Zeit und auf der y -Achse der Weg abgetragen ist, so ist der Differenzenquotient $\frac{f(b) - f(a)}{b - a}$ die Durchschnittsgeschwindigkeit zwischen den Zeitpunkten a und b . Der Satz besagt, dass die Durchschnittsgeschwindigkeit tatsächlich als Momentangeschwindigkeit in einem Zeitpunkt ξ angenommen wird. Das ist auch nicht verwunderlich, wenn sich die Geschwindigkeit nur stetig ändert: Die Geschwindigkeit kann auf $[a, b]$ nicht immer kleiner als die Durchschnittsgeschwindigkeit sein. Ebenso kann sie nicht immer größer als die Durchschnittsgeschwindigkeit sein, sonst würde es sich nicht um die Durchschnittsgeschwindigkeit handeln. Die Momentangeschwindigkeit stimmt also (wegen des Zwischenwertsatzes) irgendwo mit der Durchschnittsgeschwindigkeit überein. Der Satz gilt aber auch ohne die Voraussetzung der Stetigkeit der Ableitung. Im Beweis benutzen wir den Satz von Rolle. Er ist der Spezialfall $f(b) = f(a)$. Indem wir von f die Funktionswerte der Gerade g abziehen, können wir den allgemeineren Fall des Mittelwertsatzes in diesen Spezialfall überführen:

Beweis: Wir betrachten die aufgrund der Voraussetzungen an f auf $[a, b]$ stetige und auf $]a, b[$ differenzierbare Hilfsfunktion $h(x) := f(x) - g(x)$, also

$$h(x) := f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

Diese Funktion ist so konstruiert, dass $h(a) = h(b) = 0$. Nach dem Satz von Rolle gibt es nun aber ein $\xi \in]a, b[$ mit $0 = h'(\xi) = f'(\xi) - \frac{f(b) - f(a)}{b - a}$, und das ist die Aussage des Satzes. $\square$

Mittelwertsatz (für differenzierbare Funktionen) klingt so ähnlich wie Zwischenwertsatz (für stetige Funktionen). Der Zwischenwertsatz besagt aber, dass zwischen zwei Funktionswerten auch alle anderen Zahlen als Funktionswerte angenommen werden, während der Mittelwertsatz besagt, dass die Steigung einer Sekanten tatsächlich auch an einer Stelle angenommen wird.

Im Folgenden benötigen wir noch eine verallgemeinerte Form des Mittelwertsatzes:

Satz 2.37 (Verallgemeinerter Mittelwertsatz der Differenzialrechnung *)

Seien f, g stetig auf $[a, b]$ und differenzierbar auf $]a, b[$ mit $g'(x) \neq 0$, $x \in]a, b[$. Dann existiert mindestens eine (gemeinsame) Stelle $\xi \in]a, b[$, so dass

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$

Der Mittelwertsatz ist gerade der Spezialfall für $g(x) = x$. Wendet man andererseits den Mittelwertsatz separat auf f und g an, so erhält man zwei Stellen ξ und μ mit

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)(b - a)}{g'(\mu)(b - a)} = \frac{f'(\xi)}{g'(\mu)}.$$

Die Aussage des verallgemeinerten Mittelwertsatzes ist darüber hinaus, dass man eine gemeinsame Stelle $\xi = \mu$ finden kann.

Beweis: Zunächst ist der Bruch auf der linken Seite wohldefiniert, da wegen des Mittelwertsatzes angewendet auf g gilt:

$$g(b) - g(a) = (b - a)g'(\xi_g) \neq 0.$$

Zum Beweis des Satzes wenden wir analog zum Beweis des Mittelwertsatzes den Satz von Rolle wieder auf eine Hilfsfunktion $h(x)$ an:

$$h(x) := f(x) - f(a) - [g(x) - g(a)] \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Zusammen mit f und g ist h stetig auf $[a, b]$ und differenzierbar auf $]a, b[$. Nach Konstruktion ist $h(a) = h(b) = 0$. Damit existiert nach dem Satz von Rolle ein $\xi \in]a, b[$, so dass

$$0 = h'(\xi) = f'(\xi) - g'(\xi) \frac{f(b) - f(a)}{g(b) - g(a)}.$$

□

Mit Hilfe des Mittelwertsatzes können praktische Fehlerabschätzungen durchgeführt werden. Wertet man mittels eines Computers eine (stetig differenzierbare) Funktion an einer Stelle $x_0 \in [a, b]$ aus, so liegt häufig x_0 nur näherungsweise als $\tilde{x}_0 \in [a, b]$ vor (da man nur endlich viele Stellen speichern kann). Daher macht man einen Fehler

$$|f(x_0) - f(\tilde{x}_0)| = |f'(\xi)| |x_0 - \tilde{x}_0| \leq \max\{|f'(x)| : x \in [a, b]\} |x_0 - \tilde{x}_0|.$$

Eine kleine Abweichung zwischen $\tilde{x}_0$ und x_0 führt daher bei stetig differenzierbaren Funktionen auch nur zu einem kleinen Unterschied der Funktionswerte.

Weiß man sogar, dass für eine differenzierbare Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ für alle $x \in \mathbb{R}$ gilt: $|f'(x)| < \lambda < 1$, so ist für $x_1 \neq x_2$: $|f(x_1) - f(x_2)| = |f'(\xi_{x_1, x_2})||x_1 - x_2| < \lambda|x_1 - x_2|$. Die Funktionswerte liegen also näher zusammen als die Ausgangswerte. Man spricht von einer Kontraktion (siehe Kasten).

Hintergrund: Kontraktionen und der Banach'sche Fixpunktsatz

Definition 2.27 (Kontraktion)

Man nennt eine Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$, für die es eine Zahl $0 \leq \lambda < 1$ gibt mit $|f(x_1) - f(x_2)| < \lambda|x_1 - x_2|$ für alle $x_1, x_2 \in \mathbb{R}$, $x_1 \neq x_2$, eine **Kontraktion**.

Ist f eine Kontraktion, so ist f insbesondere (gleichmäßig) stetig auf $\mathbb{R}$ (vgl. Satz 2.30).

Kontraktionen spielen eine wichtige Rolle bei vielen numerischen Verfahren, mit denen man Lösungen mathematischer Probleme näherungsweise mit dem Computer berechnet. Das liegt am Banach'schen Fixpunktsatz.

Definition 2.28 (Fixpunkt)

Ein **Fixpunkt** einer Funktion f ist eine Stelle x_0 mit $f(x_0) = x_0$.

f hält also x_0 fix, bildet die Stelle x_0 auf sich selbst ab. Ein Fixpunkt von f ist eine Nullstelle von $f(x) - x$.

Satz 2.38 (Banach'scher Fixpunktsatz, Stefan Banach, 1892–1945)

Jede Kontraktion $f : \mathbb{R} \rightarrow \mathbb{R}$ besitzt genau einen **Fixpunkt** x_0 , und dieser ist der Grenzwert jeder Folge $(x_n)_{n=1}^{\infty}$ mit beliebig wählbarem $x_1 \in \mathbb{R}$ und $x_{n+1} := f(x_n)$.

Der Satz sagt also nicht nur etwas über die Existenz eines Fixpunktes aus, sondern liefert direkt ein programmierbares Verfahren, um den Fixpunkt zu bestimmen. Für die praktische Verwendung ist insbesondere die allgemeinere Fassung auf Seite 567 hilfreich.

Beweis: Für den Beweis erinnern wir uns an die Konvergenz von Reihen und das Vergleichskriterium Satz 2.15 von Seite 246: Wir schreiben die Folge $(x_n)_{n=1}^{\infty}$ als Teleskopsumme (siehe Kapitel 1.4.2):

$$x_{n+1} - x_1 = (x_{n+1} - x_n) + (x_n - x_{n-1}) + (x_{n-1} - x_{n-2}) + \cdots + (x_2 - x_1) = \sum_{k=1}^n (x_{k+1} - x_k).$$

Da f eine Kontraktion ist, gilt für den Abstand zweier Folgenglieder ($k > 1$)

$$\begin{aligned} |x_{k+1} - x_k| &= |f(x_k) - f(x_{k-1})| < \lambda|x_k - x_{k-1}| = \lambda|f(x_{k-1}) - f(x_{k-2})| \\ &< \lambda^2|x_{k-1} - x_{k-2}| < \cdots < \lambda^{k-1}|x_2 - x_1|. \end{aligned} \quad (2.34)$$

Damit können wir aber das Vergleichskriterium anwenden. Die Reihe $\sum_{k=1}^{\infty} (x_{k+1} - x_k)$ ist konvergent, da sie eine geometrische Reihe als konvergente Majorante

$$\sum_{k=1}^{\infty} \lambda^{k-1} |x_2 - x_1| = |x_2 - x_1| \sum_{k=0}^{\infty} \lambda^k \stackrel{(2.11)}{=} \frac{|x_2 - x_1|}{1 - \lambda}$$

besitzt. Also existiert der Grenzwert

$$\begin{aligned} x_0 := \lim_{n \rightarrow \infty} x_n &= \lim_{n \rightarrow \infty} x_{n+1} = x_1 + \lim_{n \rightarrow \infty} [x_{n+1} - x_1] \\ &= x_1 + \lim_{n \rightarrow \infty} \sum_{k=1}^n (x_{k+1} - x_k) = x_1 + \sum_{k=1}^{\infty} (x_{k+1} - x_k). \end{aligned}$$

Jetzt müssen wir noch zeigen, dass x_0 Fixpunkt und eindeutig ist. Da f stetig ist, gilt:

$$f(x_0) = f\left(\lim_{n \rightarrow \infty} x_n\right) = \lim_{n \rightarrow \infty} f(x_n) = \lim_{n \rightarrow \infty} x_{n+1} = \lim_{n \rightarrow \infty} x_n = x_0,$$

x_0 ist also ein Fixpunkt. Gäbe es einen weiteren Fixpunkt $\tilde{x}_0 \neq x_0$, so liefert die Kontraktionsbedingung einen Widerspruch:

$$|x_0 - \tilde{x}_0| = |f(x_0) - f(\tilde{x}_0)| < \lambda |x_0 - \tilde{x}_0| < |x_0 - \tilde{x}_0|.$$

□

Alternativ zum Grenzwertnachweis über das Vergleichskriterium können wir auch direkt zeigen, dass $(x_n)_{n=1}^{\infty}$ eine Cauchy-Folge und daher konvergent ist. Der folgende Ansatz hat den Vorteil, dass er direkt auch eine Fehlerabschätzung liefert. Mit $|x_{k+1} - x_k| < \lambda^{k-1} |x_2 - x_1|$ für $k \in \mathbb{N}$ (siehe (2.34)) erhalten wir für zwei Indizes $k > n \geq 1$:

$$\begin{aligned} |x_k - x_n| &= |x_k - x_{k-1} + x_{k-1} - x_{k-2} \pm \cdots + x_{n+1} - x_n| \\ &\leq |x_k - x_{k-1}| + \cdots + |x_{n+1} - x_n| \leq (\lambda^{k-2} + \cdots + \lambda^{n-1}) \cdot |x_2 - x_1| \\ &\leq \lambda^{n-1} (\lambda^{k-1-n} + \cdots + 1) \cdot |x_2 - x_1| \stackrel{(1.10)}{=} \lambda^{n-1} \frac{1 - \lambda^{k-n}}{1 - \lambda} |x_2 - x_1| \\ &\leq \frac{\lambda^{n-1}}{1 - \lambda} \cdot |x_2 - x_1|. \end{aligned} \tag{2.35}$$

Wegen $\lambda < 1$ kann zu jedem $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$ bestimmt werden mit

$$|x_k - x_n| < \varepsilon \text{ für alle } k > n > n_0,$$

d. h., $(x_n)_{n=1}^{\infty}$ ist eine Cauchy-Folge, die wegen der Vollständigkeit von $\mathbb{R}$ gegen einen Grenzwert $x_0 \in \mathbb{R}$ konvergiert, für den wir im Beweis die Fixpunkteigenschaft gezeigt haben. Die rechte Seite in (2.35) ist unabhängig von k , so dass

$$|x_0 - x_n| = \lim_{k \rightarrow \infty} |x_k - x_n| \leq \frac{\lambda^{n-1}}{1 - \lambda} \cdot |x_2 - x_1|. \tag{2.36}$$

Mit dieser Abschätzung kann man vor Beginn der Berechnung mit einer Iteration schon sagen, nach wie vielen Iterationen eine vorgegebene Distanz zum gesuchten Fixpunkt x_0

sicher unterschritten wird. Daher heißt sie **A-priori-Abschätzung**. Wählt man dagegen statt x_1 den Startwert x_{n-1} , so wird aus (2.36) die **A-posteriori-Abschätzung**

$$|x_0 - x_n| \leq \frac{\lambda}{1 - \lambda} \cdot |x_n - x_{n-1}|. \quad (2.37)$$

In Computeralgorithmen wird $x_{n+1} = f(x_n)$ iterativ angewendet, bis man eine genügend gute Näherungslösung hat. Zwei derartige Verfahren haben wir bereits kennengelernt:

- **Newton-Verfahren** (Seite 305): Eine Nullstelle einer differenzierbaren Funktion f berechnet sich als Fixpunkt der Funktion $g(x) := x - \frac{f(x)}{f'(x)}$, sofern $f'(x)$ nirgends null ist. Die Bedingung (2.24) bewirkt, dass g eine Kontraktion ist:

$$|g'(x)| = \left| 1 - \frac{(f'(x))^2 - f(x)f^{(2)}(x)}{(f'(x))^2} \right| = \frac{|f(x)||f^{(2)}(x)|}{(f'(x))^2} < \lambda < 1.$$

- **Jacobi- und Gauß-Seidel-Verfahren** zum Lösen linearer Gleichungssysteme (siehe Seite 179): Hier wird bei einem Gleichungssystem mit m Variablen und Gleichungen keine Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$, sondern eine Funktion $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ verwendet. Hier führt aber die allgemeinere Fassung des Fixpunktsatzes auf Seite 567 zu einer Konvergenzaussage, siehe Seite 575.

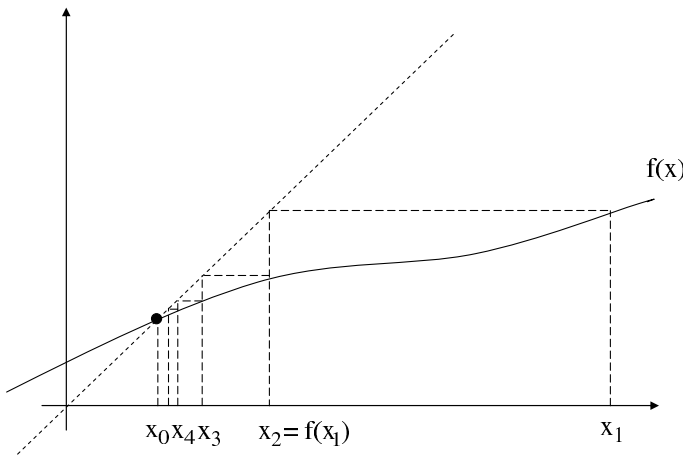


Abb. 2.28: Fixpunktiteration mit dem Satz von Banach

Beispiel 2.76

Wir betrachten eine differenzierbare Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$, deren Ableitung für große Werte von x größer als eine Zahl $c > 0$ ist. Anschaulich bedeutet das, dass die Funktion für große x streng monoton steigt und die Steigung größer als c ist. Wir erwarten also $\lim_{x \rightarrow \infty} f(x) = \infty$. Das rechnen wir mit dem Mittelwertsatz nach.

Wegen der Voraussetzung existiert eine Stelle x_0 , so dass für alle $x > x_0$ gilt $f'(x) > c > 0$. Nach dem Mittelwertsatz ist $\frac{f(x)-f(x_0)}{x-x_0} = f'(\xi)$ für eine Stelle $\xi \in]x_0, x[$. Wegen $f'(\xi) > c > 0$ folgt daraus

$$f(x) - f(x_0) > c(x - x_0), \text{ d. h. } f(x) > c \cdot x + f(x_0) - c \cdot x_0.$$

Gibt man eine Zahl $M > 0$ vor, so gilt für alle $x > \max \left\{ x_0, \frac{1}{c}[M + c \cdot x_0 - f(x_0)] \right\}$, dass $f(x) > M$. Damit ist $\lim_{x \rightarrow \infty} f(x) = \infty$ gezeigt. ■

Mit dem Mittelwertsatz können wir sofort den Fehler bei linearer Lagrange-Interpolation (siehe Kapitel 1.5.6.2) abschätzen. Eine auf $[a, b]$ stetige und auf $]a, b[$ zweimal differenzierbare Funktion f wird durch das Polynom

$$p_1(x) := \frac{x-b}{a-b}f(a) + \frac{x-a}{b-a}f(b) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$$

an den Stellen a und b interpoliert. Wegen des Mittelwertsatzes gibt es eine Stelle $\xi_1 \in]a, b[$, so dass $p_1(x) = f(a) + f'(\xi_1)(x-a)$. Ebenso können wir mit dem Mittelwertsatz die Funktion f mit einer (von x abhängenden) Stelle $\xi_2 \in]a, x[$ schreiben: $f(x) = f(a) + f'(\xi_2)(x-a)$. Damit erhalten wir für den Fehler

$$|f(x) - p_1(x)| = |f'(\xi_2) - f'(\xi_1)|(x-a).$$

Jetzt ist die Funktion f' auf dem Intervall zwischen ξ_1 und ξ_2 ebenfalls differenzierbar, so dass wir auch für f' den Mittelwertsatz verwenden können. Es gibt eine Stelle $\xi_3 \in]a, b[$ mit

$$|f(x) - p_1(x)| = |f''(\xi_3)| \cdot |\xi_2 - \xi_1| \cdot |x - a| \leq |f''(\xi_3)| \cdot |b - a|^2.$$

Allgemeiner können wir den Fehler der Lagrange-Interpolation mit $n+1$ verschiedenen Stützstellen $x_0, \dots, x_n$, an denen ein Polynom vom Grad n mit der gegebenen Funktion f übereinstimmt, ähnlich wie im Beweis des Mittelwertsatzes mit dem Satz von Rolle darstellen:

Satz 2.39 (Fehler der Lagrange-Interpolation)

Sei f auf $[a, b]$ eine $n+1$ -mal differenzierbare Funktion und p_n das Lagrange-Interpolationspolynom vom Grad n aus Kapitel 1.5.6.2, das f an den verschiedenen Stützstellen $x_0, x_1, \dots, x_n \in [a, b]$ interpoliert: $f(x_k) = p_n(x_k)$, $0 \leq k \leq n$. Dann gibt es für jedes $x \in [a, b]$ eine Stelle $\xi \in]a, b[$ (die von x abhängt), so dass

$$f(x) = p_n(x) + f^{(n+1)}(\xi) \frac{(x-x_0)(x-x_1) \cdots (x-x_n)}{(n+1)!}.$$

Beweis: Falls x eine Stützstelle ist, dann wird der letzte Term null, und die Gleichung ist offensichtlich erfüllt. Sei also x von den Stützstellen verschieden. Dann definieren wir

$$h(x) := \frac{f(x) - p_n(x)}{(x - x_0)(x - x_1) \cdots (x - x_n)}.$$

Für $t \in [a, b]$ betrachten wir die $n + 1$ -mal differenzierbare Hilfsfunktion

$$g(t) := f(t) - p_n(t) - h(x)(t - x_0)(t - x_1) \cdots (t - x_n),$$

die nach Konstruktion $n + 2$ Nullstellen $x, x_0, \dots, x_n$ besitzt. Zwischen je zwei Nullstellen liegt nach Satz von Rolle eine Nullstelle der Ableitung $g'(t)$, die damit $n + 1$ Nullstellen in $]a, b[$ hat. Jetzt kann wieder der Satz von Rolle genutzt werden um zu zeigen, dass g'' noch $n - 1$ Nullstellen besitzt, usw. Schließlich hat $g^{(n+1)}$ eine Nullstelle $\xi \in]a, b[$. Es lässt sich leicht nachrechnen, dass

$$g^{(n+1)}(t) = f^{(n+1)}(t) - 0 - h(x) \left[\frac{d^{n+1}}{dt^{n+1}} t^{n+1} + 0 \right] = f^{(n+1)}(t) - h(x)(n + 1)!$$

ist, so dass wir für die Stelle $t = \xi$ die Aussage des Satzes erhalten:

$$0 = f^{(n+1)}(\xi) - h(x)(n + 1)! = f^{(n+1)}(\xi) - \frac{f(x) - p_n(x)}{(x - x_0)(x - x_1) \cdots (x - x_n)}(n + 1)!.$$

□

2.5.3 Regeln von L'Hospital

Jetzt interessieren uns Situationen, in denen die Grenzwertregel (2.15) von Seite 267 nicht weiterhilft. $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \frac{\lim_{x \rightarrow x_0} f(x)}{\lim_{x \rightarrow x_0} g(x)}$ darf man nur rechnen, wenn die Grenzwerte des Zählers und Nenners als reelle Zahl existieren und zudem $\lim_{x \rightarrow x_0} g(x) \neq 0$ ist. Häufig sind Zähler und/oder Nenner aber bestimmt divergent (streben also gegen $+\infty$ oder $-\infty$) oder sind konvergent gegen 0. Solche Grenzwerte treten z.B. auf, wenn man Konvergenzradien von Potenzreihen bestimmt (siehe Kapitel 2.8) oder die Konvergenzgeschwindigkeiten von numerischen (näherungsweise) Verfahren (wie das Newton-Verfahren) miteinander vergleichen möchte. In der Regelungstechnik führen der Anfangs- und Endwertsatz auf Grenzwerte dieses Typs (siehe Kapitel 6.3.2.3).

Was ist z.B. $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)}$, wenn $\lim_{x \rightarrow x_0} f(x) = 0$ und $\lim_{x \rightarrow x_0} g(x) = 0$? Zur Motivation des folgenden Satzes sei $f(x_0) = g(x_0) = 0$ (und damit $\frac{f(x_0)}{g(x_0)}$ nicht definiert):

$$\frac{f(x_0 + h)}{g(x_0 + h)} = \frac{f(x_0 + h) - f(x_0)}{g(x_0 + h) - g(x_0)} = \frac{\frac{f(x_0 + h) - f(x_0)}{h}}{\frac{g(x_0 + h) - g(x_0)}{h}}.$$

Für $h \rightarrow 0$ erhalten wir, falls f und g in x_0 differenzierbar sind mit $g'(x_0) \neq 0$:

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{h \rightarrow 0} \frac{f(x_0 + h)}{g(x_0 + h)} = \lim_{h \rightarrow 0} \frac{\frac{f(x_0 + h) - f(x_0)}{h}}{\frac{g(x_0 + h) - g(x_0)}{h}} = \frac{f'(x_0)}{g'(x_0)}.$$

Sind f' und g' stetig in x_0 , so ist $\frac{f'(x_0)}{g'(x_0)} = \lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}$, und wir erhalten insgesamt

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}.$$

Wir können in dieser Situation bei der Grenzwertberechnung die Funktionen durch ihre Ableitungen ersetzen.

Beispiel 2.77

(2.14) von Seite 262 erhält man jetzt auch so: $\lim_{x \rightarrow 0} \frac{\sin x}{x} = \lim_{x \rightarrow 0} \frac{\cos x}{1} = \cos 0 = 1$. ■

Achtung: An diesem Beispiel sieht man insbesondere, dass $\frac{0}{0}$ nicht sinnvoll definiert werden kann, da für jede Zahl $c \in \mathbb{R}$ gilt: $\lim_{x \rightarrow 0} \frac{c \sin x}{x} = c$. Damit kann $\frac{0}{0}$ für jede Zahl stehen. Das muss auch so sein, denn der Grenzwert des Differenzenquotienten in der Definition der Ableitung ist auch vom Typ $\frac{0}{0}$, und als Steigung ist jede reelle Zahl denkbar.

Wir haben soeben einen Spezialfall des Satzes von L'Hospital kennengelernt. Der Grenzwert $\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}$ darf sogar $\pm\infty$ sein. Da man den Grenzwert gegen x_0 betrachtet, müssen die Funktionen an der Stelle x_0 selbst nicht definiert sein (was auch nicht der Fall ist, wenn eine Nullstelle im Nenner vorliegt). Daher wird diese Stelle in den Voraussetzungen ausgeklammert:

Satz 2.40 (Satz von L'Hospital (1661–1704, hier 1700) für $\frac{0}{0}$)

Seien die Funktionen f und g dort differenzierbar, wo sie für die folgenden Grenzwerte erklärt sein müssen. Weiterhin sei dort $g'(x) \neq 0$.

Ist $\lim_{x \rightarrow x_0} f(x) = \lim_{x \rightarrow x_0} g(x) = 0$, und existiert der Grenzwert $\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}$ als reelle Zahl oder uneigentlich ($\pm\infty$), so gilt:

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}.$$

Entsprechende Aussagen gelten für einseitige Grenzwerte und für Grenzwerte gegen $+\infty$ bzw. $-\infty$: Ist $\lim_{x \rightarrow \pm\infty} f(x) = \lim_{x \rightarrow \pm\infty} g(x) = 0$, so ist

$$\lim_{x \rightarrow \pm\infty} \frac{f(x)}{g(x)} = \lim_{x \rightarrow \pm\infty} \frac{f'(x)}{g'(x)},$$

sofern der rechte Grenzwert eigentlich oder uneigentlich existiert.

Achtung: Hier werden Grenzwerte von $\frac{f'(x)}{g'(x)}$ und nicht von $\left[\frac{f(x)}{g(x)}\right]'$ berechnet, also ohne Quotientenregel. Auch ist nicht $\frac{f(x)}{g(x)} = \frac{f'(x)}{g'(x)}$. Lediglich die Grenzwerte der Quotienten sind als Zahlen gleich.

Der Beweis benötigt nicht die Stetigkeit von f' und g' an der Stelle x_0 , die wir in der Motivation des Satzes verwendet haben. Die Funktionen f' und g' müssen nicht einmal in x_0 definiert sein. Um den Satz von L'Hospital zu zeigen, benötigen wir den verallgemeinerten Mittelwertsatz 2.37 von Seite 317, den wir genau deswegen hergeleitet haben und der schon die Quotientenstruktur des Satzes von L'Hospital nachbildet.

Beweis: ■ Wir beweisen exemplarisch den Fall eines linksseitigen Grenzwerts an der Stelle x_0 . Die Aussage für einen rechtsseitigen Grenzwert ergibt sich analog, so dass man damit auch die Aussage für den Grenzwert gegen x_0 erhält.

Wir ergänzen f und g stetig an der Stelle x_0 , die nicht zum Definitionsbereich gehören muss, mit Funktionswert 0. So haben wir auch in der Motivation des Satzes $f(x_0) = g(x_0) = 0$ gesetzt. Wir können voraussetzen, dass eine linksseitige punktierte Umgebung $]x_0 - \delta, x_0[$ zum Definitionsbereich von f und g gehört, so dass f und g dort differenzierbar sind und $g'(x) \neq 0$ ist. Außerdem sei δ so gewählt, dass die fortgesetzten Funktionen f und g stetig auf $[x_0 - \delta, x_0]$ seien. Nach dem Mittelwertsatz ist dann auf $]x_0 - \delta, x_0[$ auch $g(x) = g(x) - g(x_0) = g'(\xi) \neq 0$ für ein $\xi \in]x_0 - \delta, x_0[$. Wir dürfen also sowohl durch $g(x)$ als auch durch $g'(x)$ dividieren. Außerdem sind die Voraussetzungen des verallgemeinerten Mittelwertsatzes (Satz 2.37) erfüllt. Für einen beliebigen Punkt $y \in]x_0 - \delta, x_0[$ ergibt sich damit:

$$\frac{f(y)}{g(y)} = \frac{f(y) - f(x_0)}{g(y) - g(x_0)} = \frac{f'(\xi_y)}{g'(\xi_y)}$$

für eine Stelle $\xi_y \in]y, x_0[$. Nun betrachten wir für y eine beliebige Folge $(x_n)_{n=1}^\infty$ mit $x_n \in]x_0 - \delta, x_0[$ und $\lim_{n \rightarrow \infty} x_n = x_0$. Wegen $|x_0 - \xi_{x_n}| < |x_0 - x_n|$ konvergiert nach dem Einschließungskriterium Satz 2.2 (Seite 224) auch die Folge $(\xi_{x_n})_{n=1}^\infty$ gegen x_0 . Nach Konvergenzvoraussetzung und Übertragungsprinzip (Satz 2.22) ist

$$\lim_{n \rightarrow \infty} \frac{f(x_n)}{g(x_n)} = \lim_{n \rightarrow \infty} \frac{f'(\xi_{x_n})}{g'(\xi_{x_n})} = \lim_{x \rightarrow x_0 -} \frac{f'(x)}{g'(x)}.$$

Da die Folge $(x_n)_{n=1}^\infty$ beliebig gewählt war, folgt ebenfalls mit dem Übertragungsprinzip

$$\lim_{x \rightarrow x_0 -} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0 -} \frac{f'(x)}{g'(x)}.$$

- Für einen Grenzwert gegen $-\infty$ ($+\infty$ analog) erhalten wir die Aussage mit dem soeben Gezeigten und der Substitution $x = \frac{1}{t}$:

$$\lim_{x \rightarrow -\infty} \frac{f(x)}{g(x)} = \lim_{t \rightarrow 0 -} \frac{f\left(\frac{1}{t}\right)}{g\left(\frac{1}{t}\right)} = \lim_{t \rightarrow 0 -} \frac{-f'\left(\frac{1}{t}\right)/t^2}{-g'\left(\frac{1}{t}\right)/t^2} = \lim_{x \rightarrow -\infty} \frac{f'(x)}{g'(x)}.$$

Die erste und dritte Gleichheit müsste man streng genommen ebenfalls mit der Grenzwertdefinition oder dem Übertragungsprinzip zeigen.

□

Beispiel 2.78

Für die Funktion

$$f(s) := \frac{1}{s} (1 - e^{-s \cdot t})$$

bestimmen wir für ein festes $t \geq 0$ den Grenzwert für $s \rightarrow 0$. Dieser ist vom Typ $\left[\frac{0}{0}\right]$. Mit dem Satz von L'Hospital erhalten wir

$$\lim_{s \rightarrow 0} f(s) = \lim_{s \rightarrow 0} \frac{1 - e^{-s \cdot t}}{s} = \lim_{s \rightarrow 0} \frac{-e^{-s \cdot t} \cdot (-t)}{1} = t.$$

■

Ein analoger Grenzwertsatz gilt, wenn Zähler und Nenner nicht jeweils gegen null, sondern gegen $\pm\infty$ streben. Wir betrachten dazu differenzierbare Funktionen f und g mit $\lim_{x \rightarrow x_0} f(x) = \pm\infty$ und $\lim_{x \rightarrow x_0} g(x) = \pm\infty$. Durch Übergang zu den Kehrwerten können wir den Satz von L'Hospital für den Fall $\left[\frac{0}{0}\right]$ anwenden:

$$\begin{aligned} \lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} &= \lim_{x \rightarrow x_0} \frac{\frac{1}{g(x)}}{\frac{1}{f(x)}} = \lim_{x \rightarrow x_0} \frac{\frac{d}{dx} \frac{1}{g(x)}}{\frac{d}{dx} \frac{1}{f(x)}} = \lim_{x \rightarrow x_0} \frac{-\frac{g'(x)}{(g(x))^2}}{-\frac{f'(x)}{(f(x))^2}} \\ &= \left[\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} \right]^2 \frac{1}{\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}}, \end{aligned}$$

sofern alle einzelnen Grenzwerte existieren und wir nicht durch null dividieren. Falls auch $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} \neq 0$ ist, erhalten wir durch Umformen

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}.$$

Damit haben wir den Satz von L'Hospital für den Grenzwertfall $\left[\frac{\infty}{\infty}\right]$ hergeleitet. Die Voraussetzung, dass $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)}$ existiert, benötigt man aber gar nicht. Sie folgt bereits aus der Existenz des Grenzwerts $\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}$. Es gilt damit sogar:

Satz 2.41 (Satz von L'Hospital für Grenzwerte vom Typ $\frac{\infty}{\infty}$)

Seien die Funktionen f und g dort differenzierbar, wo sie für die folgenden Grenzwerte erklärt sein müssen. Weiterhin sei dort $g'(x) \neq 0$.

Ist $\lim_{x \rightarrow x_0} f(x) = \pm\infty$ und $\lim_{x \rightarrow x_0} g(x) = \pm\infty$ und existiert der Grenzwert $\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}$ als reelle Zahl oder uneigentlich ($\pm\infty$), so gilt:

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}.$$

Entsprechende Aussagen gelten für einseitige Grenzwerte und für Grenzwerte gegen $+\infty$ bzw. $-\infty$: Ist $\lim_{x \rightarrow \pm\infty} f(x) = \pm\infty$ und $\lim_{x \rightarrow \pm\infty} g(x) = \pm\infty$, so ist

$$\lim_{x \rightarrow \pm\infty} \frac{f(x)}{g(x)} = \lim_{x \rightarrow \pm\infty} \frac{f'(x)}{g'(x)},$$

sofern der rechte Grenzwert eigentlich oder uneigentlich existiert.

Der Beweis dieser Variante des Satzes ist ähnlich dem zuvor angegebenen. Allerdings kann man hier die Funktionen an der Stelle x_0 nicht stetig ergänzen, siehe z. B. (Hachenberger, 2005, S. 466).

Beispiel 2.79

a) Wir betrachten einen Grenzwert vom Typ $\left[\frac{\infty}{\infty}\right]$ ($n \in \mathbb{N}$):

$$\lim_{x \rightarrow \infty} \frac{\ln x}{x^n} = \lim_{x \rightarrow \infty} \frac{\frac{1}{x}}{nx^{n-1}} = \lim_{x \rightarrow \infty} \frac{1}{nx^n} = 0.$$

Damit wächst $\ln x$ für $x \rightarrow \infty$ langsamer als jedes Polynom.

b) Der Satz von L'Hospital kann auch mehrfach angewendet werden. Die Existenz der Grenzwerte folgt dann von „rechts nach links“: Der Satz von L'Hospital darf nur angewendet werden, wenn der Quotient der Ableitungen konvergiert. Kann man diese Konvergenz selbst wieder mit dem Satz zeigen, indem man den Quotienten der zweiten Ableitungen betrachtet, so folgt daraus, dass auch der ursprüngliche Grenzwert mit dem Satz von L'Hospital berechnet werden darf.

In jedem Schritt wird im folgenden Beispiel ein Grenzwert vom Typ $\left[\frac{\infty}{\infty}\right]$ bestimmt ($n \in \mathbb{N}$):

$$\lim_{x \rightarrow \infty} \frac{x^n}{e^x} = \lim_{x \rightarrow \infty} \frac{nx^{n-1}}{e^x} = \lim_{x \rightarrow \infty} \frac{n(n-1)x^{n-2}}{e^x} = \dots = \lim_{x \rightarrow \infty} \frac{n!}{e^x} = 0.$$

Damit wächst $\exp(x)$ für $x \rightarrow \infty$ schneller als jedes Polynom.

c) Wenn der Satz nicht anwendbar ist, heißt das nicht, dass der Grenzwert nicht existiert (oder dass er existiert): Wir betrachten $\lim_{x \rightarrow \infty} \frac{x(c + \sin x)}{x^2}$ mit $c > 1$, also einen Grenzwert vom Typ $\left[\frac{\infty}{\infty}\right]$. Der Grenzwert der Ableitungen $\lim_{x \rightarrow \infty} \frac{c + \sin(x) + x \cos x}{2x}$ existiert nicht, da $\lim_{x \rightarrow \infty} \frac{c + \sin x}{2x} = 0$ und $\lim_{x \rightarrow \infty} \frac{x \cos x}{2x} = \lim_{x \rightarrow \infty} \frac{\cos x}{2}$ nicht existiert. Der Satz ist hier also nicht anwendbar. Dennoch existiert der Grenzwert:

$$\lim_{x \rightarrow \infty} \frac{x(c + \sin x)}{x^2} = \lim_{x \rightarrow \infty} \frac{c + \sin x}{x} = 0. \quad \blacksquare$$

Mit dem Satz von L'Hospital können auch andere Typen von Grenzwerten berechnet werden:

- Typ $[0 \cdot \infty]$, d. h., bestimme $\lim_{x \rightarrow x_0} f(x) \cdot g(x)$, wobei $\lim_{x \rightarrow x_0} f(x) = 0$ und $\lim_{x \rightarrow x_0} g(x) = \pm\infty$. Hier kann man entweder den Ansatz $f(x)g(x) = \frac{f(x)}{\frac{1}{g(x)}}$ (Typ $\left[\frac{0}{0}\right]$) oder den Ansatz $f(x)g(x) = \frac{g(x)}{\frac{1}{f(x)}}$ (Typ $\left[\frac{\infty}{\infty}\right]$) wählen (so dass der Nenner keine störende Nullstelle hat).

Beispiel 2.80

Wir wandeln einen Grenzwert des Typs $[0 \cdot \infty]$ in den Typ $\left[\frac{\infty}{\infty}\right]$ um. Für $n \in \mathbb{N}$ gilt:

$$\lim_{x \rightarrow 0+} x^n \ln x = \lim_{x \rightarrow 0+} \frac{\ln x}{x^{-n}} = \lim_{x \rightarrow 0+} \frac{\frac{1}{x}}{-nx^{-n-1}} = \lim_{x \rightarrow 0+} -\frac{x^n}{n} = 0. \quad (2.38) \quad \blacksquare$$

- Typ $[\infty - \infty]$, d. h., bestimme $\lim_{x \rightarrow x_0} [f(x) - g(x)]$, wobei $\lim_{x \rightarrow x_0} f(x) = \infty$ und $\lim_{x \rightarrow x_0} g(x) = \infty$. Hier kann der Ansatz

$$f - g = \frac{\frac{1}{g} - \frac{1}{f}}{\frac{1}{f \cdot g}},$$

der zum Typ $\left[\frac{0}{0}\right]$ führt, weiterhelfen.

Beispiel 2.81

Eine Umwandlung in den Typ $\left[\frac{0}{0}\right]$ mittels doppelter Anwendung des Satzes von L'Hospital ergibt

$$\begin{aligned} \lim_{x \rightarrow 0+} \left(\frac{1}{x} - \frac{1}{\sin x} \right) &= \lim_{x \rightarrow 0+} \frac{\sin(x) - x}{x \sin x} = \lim_{x \rightarrow 0+} \frac{\cos(x) - 1}{\sin(x) + x \cos(x)} \\ &= \lim_{x \rightarrow 0+} \frac{-\sin x}{\cos(x) + \cos(x) - x \sin(x)} = 0. \end{aligned}$$

■

- Typen $[0^0]$, $[\infty^0]$, $[1^\infty]$: Falls er wohldefiniert ist, kann hier der Ansatz

$$f(x)^{g(x)} = \exp(g(x) \ln f(x))$$

gewählt werden, wobei dann der Grenzwert der Funktion $g(x) \ln f(x)$ zu bestimmen ist sowie die Stetigkeit der Exponentialfunktion ausgenutzt werden kann.

Beispiel 2.82

a) Wir berechnen einen Grenzwert vom Typ $[0^0]$:

$$\lim_{x \rightarrow 0+} x^x = \lim_{x \rightarrow 0+} \exp(x \ln x) \stackrel{(2.17)}{=} \exp \left(\lim_{x \rightarrow 0+} x \ln x \right) \stackrel{(2.38)}{=} \exp(0) = 1. \quad (2.39)$$

Auch wenn hier zufällig 1 herauskommt, lässt sich die Potenz 0^0 nicht sinnvoll als feste Zahl definieren. Für $x > 0$ ist

$$x^{\frac{c}{\ln x}} = \exp \left(c \cdot \frac{\ln x}{\ln x} \right) = e^c,$$

aber sowohl $\lim_{x \rightarrow 0+} x = 0$ als auch $\lim_{x \rightarrow 0+} \frac{c}{\ln x} = 0$.

b) Wir berechnen einen Grenzwert vom Typ $[1^\infty]$:

$$\begin{aligned} \lim_{x \rightarrow \infty} \left(1 + \frac{a}{x} \right)^x &\stackrel{(2.17)}{=} \exp \left(\lim_{x \rightarrow \infty} \frac{\ln(1 + ax^{-1})}{x^{-1}} \right) \\ &= \exp \left(\lim_{x \rightarrow \infty} \frac{-ax^{-2} \frac{1}{1+ax^{-1}}}{-x^{-2}} \right) = \exp \left(a \lim_{x \rightarrow \infty} \frac{1}{1 + \frac{a}{x}} \right) = e^a. \end{aligned} \quad (2.40)$$

Dies ist kein Beweis für (2.8) auf Seite 228 mit $a = 1$, da wir diese Identität zum Nachweis der Stetigkeit der Exponentialfunktion verwendet haben. Hier haben wir allerdings die Stetigkeit ausgenutzt.

■

- Man beachte, dass die Typen $[0^\infty]$ und $[\infty^\infty]$ direkt und ohne den Satz von L'Hospital gelöst werden können.

Ist im Fall $[0^\infty]$ $\lim_{x \rightarrow x_0} f(x) = 0$ und $\lim_{x \rightarrow x_0} g(x) = \infty$, so ist für $f(x) > 0$:

$$\lim_{x \rightarrow x_0} f(x)^{g(x)} = \lim_{x \rightarrow x_0} \underbrace{\exp(\underbrace{g(x)}_{\rightarrow \infty} \underbrace{\ln f(x)}_{\rightarrow -\infty})}_{\rightarrow -\infty} = 0.$$

Ist im Fall $[\infty^\infty]$ $\lim_{x \rightarrow x_0} f(x) = \lim_{x \rightarrow x_0} g(x) = \infty$, so ist für $f(x) > 0$:

$$\lim_{x \rightarrow x_0} f(x)^{g(x)} = \lim_{x \rightarrow x_0} \exp(\underbrace{g(x)}_{\rightarrow \infty} \underbrace{\ln f(x)}_{\rightarrow \infty})_{\rightarrow \infty} = \infty.$$

- Der Satz von L'Hospital kann auch zur Berechnung von Folggrenzwerten eingesetzt werden. Hier berechnet man den Grenzwert einer Funktion und nutzt dann das Übertragungsprinzip, um den Grenzwert einer Folge zu erhalten.

Beispiel 2.83 (Folggrenzwert über Satz von L'Hospital)

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = \lim_{x \rightarrow \infty} \left(1 + \frac{a}{x}\right)^x \stackrel{(2.40)}{=} e^a.$$

Die Anwendung des Satzes von L'Hospital geschieht hier über (2.40). ■

Tab. 2.2: Umformungen zur Anwendung des Satzes von L'Hospital

Typ des Grenzwerts	Funktion	umgeformte Funktion
$[0 \cdot \infty], \quad [0 \cdot (-\infty)]$	$f(x) \cdot g(x)$	$\frac{f(x)}{\frac{1}{g(x)}}$
$[\infty - \infty]$	$f(x) - g(x)$	$\left[\frac{1}{g(x)} - \frac{1}{f(x)}\right] / \frac{1}{f(x) \cdot g(x)}$
$[0^0], \quad [\infty^0], \quad [1^\infty]$	$f(x)^{g(x)}$	$\ln(f(x)^{g(x)}) = g(x) \cdot \ln(f(x))$

2.6 Integralrechnung

Die Berechnung von Flächen unter beliebig geformten Kurven stellt ein klassisches mathematisches Problem dar, man denke etwa an die Bestimmung der Fläche eines Kreises. Die Integralrechnung behandelt diese Aufgabenstellungen. Darüber hinaus kann man mit Integralen Volumina, Oberflächen, Schwerpunkte, Trägheitsmomente und viele andere Kenngrößen von Objekten berechnen, Differenzialgleichungen lösen

sowie viele stochastische Fragestellungen bearbeiten. In der Informatik und im Ingenieurbereich benötigt man Integralrechnung beispielsweise für die Computergrafik sowie bei der Simulation dynamischer Systeme.

2.6.1 Definition des Integrals

Das Integral einer Funktion $f : [a, b] \rightarrow \mathbb{R}$ mit nicht-negativen Funktionswerten $f(x) \geq 0$ über einem Intervall $[a, b]$ ist eine Zahl, die dem Flächeninhalt der durch Funktionsgraph von f , durch x -Achse und durch vertikale Linien bei $x = a$ und $x = b$ begrenzten Fläche entspricht. Wir definieren im Folgenden das Integral exakter in einer Weise, die auch für Funktionen mit negativen Funktionswerten Sinn macht.

Beispiel 2.84

Das Integral der Funktion $f(x) := 3$ über $[2, 6]$ ist $(6 - 2) \cdot 3 = 12$, nämlich der Flächeninhalt des Rechtecks mit Eckpunkten $(2, 0)$, $(2, 3)$, $(6, 3)$ und $(6, 0)$. ■

Damit tatsächlich eine umrandete Fläche entsteht, fordert man, dass f auf $[a, b]$ beschränkt ist, dass es also ein $M > 0$ gibt mit $|f(x)| \leq M$ für alle $x \in [a, b]$. Wir nähern den Flächeninhalt unter dem Funktionsgraphen an, indem wir einerseits die Fläche von innen heraus mittels Rechtecken ausfüllen und andererseits sie von außen mittels Rechtecken überdecken. Das führt zu den Begriffen Unter- und Obersumme (siehe Abbildung 2.29).

Definition 2.29 (Zerlegung)

Seien $[a, b] \subset \mathbb{R}$ und $x_0, \dots, x_m \in [a, b]$ mit

$$a = x_0 < x_1 < x_2 < \dots < x_{m-1} < x_m = b.$$

Dann heißt $Z := (x_0, x_1, \dots, x_m)$ eine **Zerlegung** des Intervalls $[a, b]$.

Die Zerlegung Z zerlegt $[a, b]$ in Teilintervalle $[x_0, x_1]$, $[x_1, x_2]$, $\dots$, $[x_{m-1}, x_m]$.

Definition 2.30 (Unter- und Obersummen)

Sei $f : [a, b] \rightarrow \mathbb{R}$ eine beschränkte Funktion und $Z := (x_0, x_1, \dots, x_m)$ eine Zerlegung von $[a, b]$.

$$s_Z := \sum_{k=1}^m \left[\inf \{f(x) : x \in [x_{k-1}, x_k]\} \right] (x_k - x_{k-1})$$

heißt eine **Untersumme** von f bezüglich der Zerlegung Z und

$$S_Z := \sum_{k=1}^m \left[\sup \{f(x) : x \in [x_{k-1}, x_k]\} \right] (x_k - x_{k-1})$$

heißt eine **Obersumme** von f bezüglich der Zerlegung Z .

In dieser Definition haben wir erneut das Infimum (inf) als größte untere und das Supremum (sup) als kleinste obere Schranke einer Menge von Funktionswerten verwendet. Ist f stetig, so nimmt f auf den Teilintervallen $[x_{k-1}, x_k]$ Supremum und Infimum an, d. h., f hat dort ein Maximum und ein Minimum. In diesem Fall können wir statt von $\inf \{f(x) : x \in [x_{k-1}, x_k]\}$ vom kleinsten Funktionswert von f im Intervall $[x_{k-1}, x_k]$ sprechen. Analog ist dann $\sup \{f(x) : x \in [x_{k-1}, x_k]\}$ der größte Funktionswert, den f auf diesem Intervall annimmt. Bei der Obersumme multiplizieren wir also den jeweils größten Funktionswert mit der Länge des Zerlegungsintervalls, bei der Untersumme den jeweils kleinsten. Dabei erhalten wir (bei nicht-negativwertigen Funktionen) einerseits die Fläche eines Rechtecks, das den Graphen überdeckt, und andererseits die Fläche eines Rechtecks unter dem Funktionsgraphen. Die Summe über alle so gewonnenen Flächen ist bei Verwendung der größten Funktionswerte die Obersumme der Zerlegung, bei Verwendung der kleinsten Funktionswerte die Untersumme.

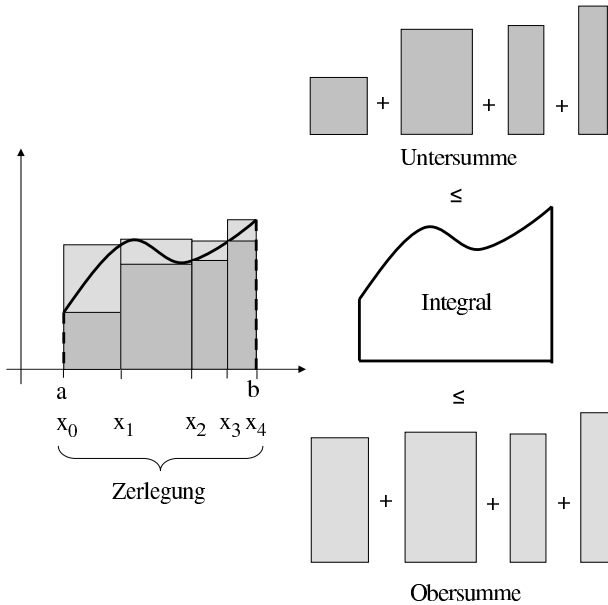


Abb. 2.29: Das Integral wird als Fläche unter dem Funktionsgraphen durch Ober- und Untersummen angenähert.

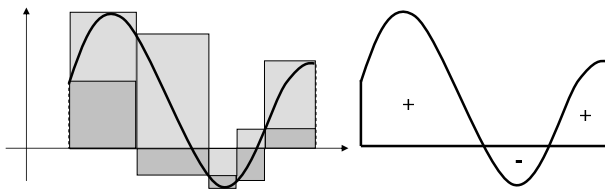


Abb. 2.30: Flächen unterhalb der x-Achse fließen negativ in das Integral ein.

Da f mit M beschränkt ist, gilt offensichtlich

$$-M(b-a) \leq s_Z \leq S_Z \leq M(b-a),$$

und beide Zahlen sind eine Approximation an den gesuchten Flächeninhalt, der zwischen s_Z und S_Z liegt. Unterteilt man eine Zerlegung weiter, so wird die Obersumme kleiner und die Untersumme größer.

Definition 2.31 (Unter- und Oberintegral)

Seien $f : [a, b] \rightarrow \mathbb{R}$ eine beschränkte Funktion und $\mathcal{Z}$ die Menge aller Zerlegungen des Intervalls $[a, b]$. Die Zahl

$$\underline{I} := \sup\{s_Z : Z \in \mathcal{Z}\},$$

also die kleinste Zahl, die größer oder gleich allen Untersummen ist, heißt **Riemann-Unterintegral** von f auf $[a, b]$. Die Zahl

$$\bar{I} := \inf\{S_Z : Z \in \mathcal{Z}\},$$

also die größte Zahl, die kleiner oder gleich allen Obersummen ist, heißt **Riemann-Oberintegral** von f auf $[a, b]$. Falls $\underline{I} = \bar{I}$, heißt f auf $[a, b]$ **Riemann-integrierbar** (kurz: **integrierbar**), und die Zahl $I := \underline{I} = \bar{I}$ heißt das **Riemann-Integral** (oder kurz **Integral**) von f auf $[a, b]$. Bezeichnung: $I = \int_a^b f(x) dx$.

Bemerkungen:

- Falls es eine größte Untersumme und eine kleinste Obersumme gibt, ist f integrierbar genau dann, wenn beide übereinstimmen. Die Annäherung an den Flächeninhalt von außen führt zum gleichen Wert wie die Annäherung von innen.
- Die Schreibweise $\int_a^b f(x) dx$ besteht aus zwei Teilen, die wie eine Klammer wirken: $\int_a^b$ und dx . Dazwischen steht die zu integrierende Funktion. Das ist anders als bei der Ableitung: Möchte man $f(x) + g(x)$ ableiten, so schreibt man $\frac{d}{dx}[f(x) + g(x)]$. Beim Integrieren können die Klammern aufgrund der Schreibweise weggelassen werden: $\int_a^b f(x) + g(x) dx$.
- Bei dx wird die Variable x benannt, die in der Funktion verwendet wird. Möchte man statt x einen anderen Variablennamen wie z.B. t verwenden, so schreibt man $\int_a^b f(t) dt$. Wenn wir später Funktionen benutzen, die von mehreren Variablen abhängig sind, wird es wichtig zu sagen, auf welche Variable sich ein Integral bezieht (siehe Kapitel 4.4).
- In der Definition wird nicht verlangt, dass f nicht-negativwertig ist. Hat f ein wechselndes Vorzeichen, so werden vom Funktionsgraphen Flächen oberhalb und unterhalb der x -Achse eingeschlossen. Das Integral ist die Summe aller dieser Flächeninhalte, wobei der Inhalt von Flächen unterhalb der x -Achse negativ gewertet wird (siehe Abbildung 2.30). Möchte man diese Flächen ebenfalls positiv werten, so muss man $\int_a^b |f(x)| dx$ berechnen.

Man kann zeigen: Ist f auf $[a, b]$ integrierbar, so sind auch die beiden nicht-negativwertigen Funktionen

$$f^+(x) := \begin{cases} f(x), & f(x) > 0 \\ 0, & f(x) \leq 0 \end{cases}, \quad f^-(x) := \begin{cases} -f(x), & f(x) < 0 \\ 0, & f(x) \geq 0 \end{cases} \quad (2.41)$$

integrierbar, und es gilt $\int_a^b f(x) dx = \int_a^b f^+(x) dx - \int_a^b f^-(x) dx$. Insbesondere ist auch $|f|$ integrierbar mit $\int_a^b |f(x)| dx = \int_a^b f^+(x) dx + \int_a^b f^-(x) dx$.

- Für am Aufbau der Mathematik Interessierte: Man beachte, dass Ober- und Unterintegral als ein Supremum bzw. Infimum einer nicht-leeren Menge definiert sind. Wegen des Vollständigkeitsaxioms (1.5) auf Seite 49 existiert dieses Supremum und Infimum. Das Integral ist im Prinzip ein Grenzwert von Ober- und Untersummen. Damit Grenzwerte existieren können, benötigt man die Vollständigkeit der reellen Zahlen.

Möchte man nun mit der Definition Integrale näherungsweise berechnen, so muss man auf jedem Zerlegungsintervall größte und kleinste Funktionswerte bestimmen. Das ist schwierig. Wir könnten die Extremwerte z. B. mit der Differenzialrechnung ermitteln. Einfacher wäre es aber, wenn wir irgendeinen Funktionswert auf dem Teilintervall verwenden könnten. Glücklicherweise kann man alternativ das Riemann-Integral auch über Zwischensummen einführen, bei denen statt des Supremums und Infimums tatsächlich in einem Zerlegungsintervall ein beliebiger Funktionswert verwendet wird. Dieser Zugang ist äquivalent und kann insbesondere für die numerische Berechnung von Integralen z. B. mit dem Computer benutzt werden.

Definition 2.32 (Zwischensumme)

Sei $Z = (x_0, x_1, \dots, x_m)$ eine Zerlegung des Intervalls $[a, b]$.

- $\|Z\|$ sei der größte Abstand zweier aufeinander folgender Zerlegungspunkte, also

$$\|Z\| := \max\{x_k - x_{k-1} : 1 \leq k \leq m\}.$$

- Sei $\xi := (\xi_1, \xi_2, \dots, \xi_m)$ mit $\xi_1 \in [x_0, x_1]$, $\xi_2 \in [x_1, x_2]$, $\dots$, $\xi_m \in [x_{m-1}, x_m]$. ξ heißt eine **Zwischenpunktwahl** zur Zerlegung Z mit den **Zwischenpunkten** $\xi_1, \xi_2, \dots, \xi_m$.
- Die Summe

$$S(Z, f, \xi) := \sum_{k=1}^m f(\xi_k)(x_k - x_{k-1})$$

heißt **Riemann-Zwischensumme** zur Zerlegung Z und Zwischenpunktwahl ξ .

Bei einer Riemann-Zwischensumme wird für jedes Zerlegungsintervall ein Funktionswert berechnet und mit der Intervalllänge multipliziert (Flächeninhalt eines Recht-

ecks). Die resultierenden Werte werden dann für alle Zerlegungsintervalle aufsummiert, so dass man auch hier eine Annäherung an die Fläche unter dem Funktionsgraphen erhält. Eine Riemann-Zwischensumme liegt **zwischen** der Unter- und der Obersumme zur gleichen Zerlegung, d. h. $s_Z \leq S(Z, f, \xi) \leq S_Z$. Dies folgt direkt aus den Definitionen wegen

$$\inf \{f(x) : x \in [x_{k-1}, x_k]\} \leq f(\xi_k) \leq \sup \{f(x) : x \in [x_{k-1}, x_k]\}.$$

Satz 2.42 (Riemann'sches Integrabilitätskriterium)

Sei $f : [a, b] \rightarrow \mathbb{R}$ beschränkt und $F \in \mathbb{R}$. Dann sind äquivalent

- a) f ist integrierbar mit Wert $F = \int_a^b f(x) dx$.
- b) Zu jedem (noch so kleinen) $\varepsilon > 0$ gibt es ein (von ε abhängendes) $\delta > 0$, so dass für alle Zerlegungen Z mit $\|Z\| \leq \delta$ und jede zu Z gehörende Zwischenpunktwahl ξ gilt:

$$|S(Z, f, \xi) - F| < \varepsilon.$$

Mit anderen Worten: Wählt man die Zerlegungen immer feiner, so dass die maximale Länge aller Zerlegungsintervalle gegen null strebt, so konvergieren die zugehörigen Riemann-Zwischensummen für jede denkbare Zwischenpunktwahl gegen die Zahl F .

Der Beweis des Satzes ist nicht schwierig, aber recht technisch, da man Eigenschaften des Supremums und Infimums ausnutzen muss und geeignete Zerlegungen baut. Er kann z. B. in (Heuser, 2009, S. 468 f.) nachgelesen werden.

Das Integral erhält man also auch über Riemann-Zwischensummen zu immer feiner werdenden Zerlegungen. Man verwendet den Satz daher oft in der Richtung a) $\implies$ b). Hat man beispielsweise eine auf $[0, 1]$ integrierbare Funktion f , so gilt bei konkreter Wahl der Zerlegung (m gleich große Teilintervalle) und Zwischenpunkte (jeweils der rechte Rand des Teilintervalls):

$$\int_0^1 f(x) dx = \lim_{m \rightarrow \infty} \left[\sum_{k=1}^m f\left(\frac{k}{m}\right) \frac{1}{m} \right]. \quad (2.42)$$

Mit (2.42) können wir ein Integral also tatsächlich als ein Folgen Grenzwert verstehen. Die Integral-Notation $\int_0^1 f(x) dx$ ist in Anlehnung an die Summe der rechten Seite gewählt: In der Summe werden Funktionswerte $f(x_k) = f\left(\frac{k}{m}\right)$ mit der Differenz $\Delta x_k := x_k - x_{k-1} = \frac{k}{m} - \frac{k-1}{m}$ der Zerlegungsstellen der x -Achse multipliziert, um die Fläche eines Rechtecks zu erhalten: $f(x_k) \cdot \Delta x_k$. Schreibt man statt Δx schlicht dx , so gelangen wir zur Integralschreibweise $f(x) dx$. Wir werden später sehen, dass diese auch in Verbindung mit der Ableitungs-Notation $\frac{d}{dx}$ sinnvoll ist (siehe Seite 341).

Beispiel 2.85

Wir berechnen das Integral $\int_0^1 x \, dx$ über Riemann-Zwischensummen. Dabei handelt es sich um die Fläche $\frac{1}{2}$ eines Dreiecks. Diesen Wert erhalten wir auch über den Grenzwert (2.42) unter Verwendung der Summenformel von Gauß.

$$\begin{aligned} \int_0^1 x \, dx &= \lim_{m \rightarrow \infty} \left[\sum_{k=1}^m \frac{k}{m} \frac{1}{m} \right] = \lim_{m \rightarrow \infty} \frac{1}{m^2} \left[\sum_{k=1}^m k \right] \\ &= \lim_{m \rightarrow \infty} \frac{m(m+1)}{2m^2} = \lim_{m \rightarrow \infty} \left[\frac{1}{2} + \frac{1}{2m} \right] = \frac{1}{2}. \quad \blacksquare \end{aligned}$$

2.6.2 Eigenschaften des Integrals

Für die exakte Berechnung von Integralen ist die Grenzwertbetrachtung von Zwischen-summen zu mühsam. Diese haben ihre Bedeutung bei der näherungsweisen Berechnung mittels Computern. Darauf gehen wir noch in Kapitel 2.6.5 ein. Für die exakte Berechnung helfen die folgenden Eigenschaften des Integrals (die man über die Definition nachrechnen kann) und der Hauptsatz (siehe Satz 2.48), der später behandelt wird.

Sind f und g auf $[a, b]$ integrierbar, so gilt dies auch für die Funktionen $f + g$, $c \cdot f$ für eine Konstante $c \in \mathbb{R}$ und $f \cdot g$. Für die Integrale gilt:

$$\int_a^b f(x) + g(x) \, dx = \int_a^b f(x) \, dx + \int_a^b g(x) \, dx, \quad \int_a^b c \cdot f(x) \, dx = c \cdot \int_a^b f(x) \, dx. \quad (2.43)$$

Man spricht hier auch von der **Linearität** des Integrals. Eine weitere wichtige Eigenschaft ist die Monotonie:

Sind f und g auf $[a, b]$ integrierbar, und ist $f(x) \leq g(x)$ für alle $x \in [a, b]$, so ist auch

$$\int_a^b f(x) \, dx \leq \int_a^b g(x) \, dx. \quad (2.44)$$

Es gilt eine Verallgemeinerung der Dreiecksungleichung (1.14) von Seite 68, wenn man die Integration als eine unendliche Summation versteht. Mit den Bezeichnungen f^+ und f^- aus (2.41) ist

$$\begin{aligned} \left| \int_a^b f(x) \, dx \right| &= \left| \int_a^b f^+(x) \, dx - \int_a^b f^-(x) \, dx \right| \\ &\leq \left| \int_a^b f^+(x) \, dx \right| + \left| \int_a^b f^-(x) \, dx \right| = \int_a^b f^+(x) \, dx + \int_a^b f^-(x) \, dx \\ &= \int_a^b f^+(x) + f^-(x) \, dx = \int_a^b |f(x)| \, dx, \end{aligned}$$

also

$$\left| \int_a^b f(x) dx \right| \leq \int_a^b |f(x)| dx. \quad (2.45)$$

Bislang haben wir keine Vorstellung davon, welche Eigenschaften eine beschränkte Funktion haben muss, damit Ober- und Unterintegral übereinstimmen. Das ändern wir:

Satz 2.43 (Hinreichende Bedingung für Integrierbarkeit)

Ist $f : [a, b] \rightarrow \mathbb{R}$ stetig auf $[a, b]$, so ist f auch integrierbar auf $[a, b]$.

Die Umkehrung des Satzes gilt aber nicht: Es gibt Riemann-integrierbare Funktionen, die nicht stetig sind, z. B. Funktionen mit einigen Sprungstellen. Riemann-integrierbare Funktionen sind aber „ziemlich“ stetig. Nur an wenigen Stellen können sie unstetig sein. Umgekehrt sind Funktionen, die fast überall stetig sind, integrierbar. Exakt formuliert ist das der folgende Satz, den wir ohne Beweis zitieren:

Satz 2.44 (Stetigkeit integrierbarer Funktionen*)

Eine beschränkte Funktion f ist genau dann integrierbar auf $[a, b]$, wenn die Menge $U \subset [a, b]$ der Unstetigkeitsstellen von f die folgende Bedingung erfüllt:

Zu jedem (noch so kleinen) $\varepsilon > 0$ gibt es eine Folge von offenen Intervallen $]a_k, b_k[$ mit $U \subset \bigcup_{k=0}^{\infty}]a_k, b_k[$ und $\sum_{k=0}^{\infty} (b_k - a_k) < \varepsilon$.

Die Menge U ist in diesem Sinne also beliebig klein.

Zum Beweis des Satzes 2.43 benötigen wir schwere Geschütze im Sinne des Satzes von Heine-Borel, der hier indirekt über den Satz zur gleichmäßigen Stetigkeit (siehe Satz 2.30, Seite 284) einfließt. Dahinter steckt wie bei vielen Konvergenzsätzen die Vollständigkeit der reellen Zahlen.

Beweis: (zu Satz 2.43) Da f stetig auf $[a, b]$ ist, ist f nach Satz 2.26 (Seite 281) auf $[a, b]$ insbesondere beschränkt, so dass eine Voraussetzung der Integrierbarkeit bereits erfüllt ist. Wir zeigen nun, dass zu jedem $\varepsilon > 0$ eine Zerlegung Z existiert mit

$$|S_Z - s_Z| = S_Z - s_Z < \varepsilon. \quad (2.46)$$

Dann ist

$$|\bar{I} - \underline{I}| = \bar{I} - \underline{I} \leq S_Z - s_Z < \varepsilon$$

und damit $\bar{I} = \underline{I}$. Es bleibt also (2.46) zu zeigen. Da f stetig auf $[a, b]$ ist, ist f nach Satz 2.30 dort sogar gleichmäßig stetig, d. h., zu jedem (ab jetzt festen) $\varepsilon > 0$ existiert ein $\delta = \delta(\varepsilon)$, so dass für alle $x, y \in [a, b]$ mit $|x - y| < \delta$ gilt:

$$|f(x) - f(y)| < \frac{\varepsilon}{b - a}.$$

Sei nun $m \in \mathbb{N}$ so groß gewählt, dass $(b-a)/m < \delta$ ist. Man beachte, dass m wie auch die folgenden Objekte von ε abhängen. Wir verwenden die Zerlegung $Z = (a, a + \frac{b-a}{m}, a + 2\frac{b-a}{m}, \dots, b)$. Damit ist

$$\begin{aligned} S_Z - s_Z &= \sum_{k=1}^m \frac{b-a}{m} \left[\max_{x \in [a+(k-1)\frac{b-a}{m}, a+k\frac{b-a}{m}]} f(x) - \min_{x \in [a+(k-1)\frac{b-a}{m}, a+k\frac{b-a}{m}]} f(x) \right] \\ &\leq \sum_{k=1}^m \frac{b-a}{m} \cdot \frac{\varepsilon}{b-a} = m \cdot \frac{b-a}{m} \cdot \frac{\varepsilon}{b-a} = \varepsilon. \end{aligned}$$

□

Häufig sind die zu integrierenden Funktionen nicht überall stetig. Einzelne (d. h. endlich viele) Funktionswerte spielen bei der Berechnung eines Integrals keine Rolle.

Satz 2.45 (Weitere hinreichende Bedingungen für Integrierbarkeit)

Sei $f : [a, b] \rightarrow \mathbb{R}$ beschränkt auf $[a, b]$.

- a) Ist f stückweise stetig auf $[a, b]$, d. h., f ist stetig bis auf endlich viele Sprungstellen, so ist f integrierbar auf $[a, b]$.
- b) Ist f monoton auf $[a, b]$, so ist f integrierbar auf $[a, b]$.

Beweis: (Skizze) Die Aussage a) erhält man entweder mit Satz 2.44 oder mit Satz 2.43, indem man das Integral in endlich viele Integrale mit einem stetigen Integranden zerlegt. Die Aussage b) ist etwas schwieriger zu beweisen. Nach Lemma 2.6 auf Seite 286 ist wegen der Monotonie die Menge der Unstetigkeitsstellen von f auf $]a, b[$ höchstens abzählbar unendlich. Man kann zeigen, dass auch diese Menge im Sinne von Satz 2.44 noch klein genug ist. □

Gibt es jedoch zu viele Unstetigkeitsstellen, so kann nicht mehr integriert werden. Das klassische Beispiel dafür ist die Dirichlet-Funktion (siehe Beispiel 2.40 auf Seite 263). Alle Untersummen auf dem Intervall $[0, 1]$ sind für diese Funktion gleich 0, alle Obersummen sind 1.

Ist f integrierbar auf $[a, b]$, so auch auf jedem Teilintervall von $[a, b]$. Insbesondere gilt für $c \in [a, b]$:

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx, \quad \int_a^a f(x) dx = 0. \quad (2.47)$$

Bislang war stets $a \leq b$. Wir lassen auch den umgekehrten Fall zu und definieren

Definition 2.33 (Vertauschte Integrationsgrenzen)

Sei f integrierbar auf $[a, b]$. Dann verstehen wir unter

$$\int_b^a f(x) dx := - \int_a^b f(x) dx.$$

Man beachte, dass diese Definition verträglich ist mit der Additionsregel (2.47), die nun auch ohne die Einschränkung $a \leq c \leq b$ gilt:

$$\int_a^b f(x) dx - \int_c^b f(x) dx = \int_a^b f(x) dx + \int_b^c f(x) dx = \int_a^c f(x) dx.$$

Satz 2.46 (Rückschluss aus Integralwert 0)

Sei f stetig auf $[a, b]$, und es gelte $f(x) \geq 0$ für alle $x \in [a, b]$. Dann folgt aus $\int_a^b f(x) dx = 0$, dass $f(x) = 0$ für alle $x \in [a, b]$.

Beweis: Annahme: Es existiere ein $x_0 \in]a, b[$ mit $f(x_0) > 0$. Da f stetig in x_0 ist, existiert ein $\delta > 0$ (und $\delta < \min\{x_0 - a, b - x_0\}$) mit $|f(x) - f(x_0)| < \varepsilon := \frac{f(x_0)}{2}$ für alle $x \in]x_0 - \delta, x_0 + \delta[$, d. h., auf dieser Umgebung ist $f(x) > \frac{f(x_0)}{2}$. Damit ist

$$\begin{aligned} 0 &= \int_a^b f(x) dx = \int_a^{x_0-\delta} f(x) dx + \int_{x_0-\delta}^{x_0+\delta} f(x) dx + \int_{x_0+\delta}^b f(x) dx \\ &\stackrel{f \geq 0}{\geq} \int_{x_0-\delta}^{x_0+\delta} f(x) dx \geq 2\delta \frac{f(x_0)}{2} = \delta f(x_0) > 0. \end{aligned}$$

Aufgrund dieses Widerspruchs muss $f(x) = 0$ für jedes $x \in]a, b[$ sein. Für $x = a$ bzw. $x = b$ folgt unter Ausnutzung der einseitigen Stetigkeit

$$f(a) = \lim_{x \rightarrow a+} f(x) = \lim_{x \rightarrow a+} 0 = 0 = \lim_{x \rightarrow b-} f(x) = f(b).$$

□

Genauso wie bei der Differenzialrechnung gibt es auch hier einen Mittelwertsatz:

Satz 2.47 (Mittelwertsatz der Integralrechnung)

Sei f stetig auf $[a, b]$, $a < b$. Dann existiert eine (Zwischen-)Stelle $\xi \in [a, b]$ mit

$$f(\xi) = \frac{1}{b-a} \int_a^b f(x) dx.$$

Die Fläche unter dem Funktionsgraphen entspricht also der Fläche eines Rechtecks mit Breite $b - a$ und Höhe $f(\xi)$.

Beweis: Da f stetig auf dem abgeschlossenen Intervall $[a, b]$ ist, ist f integrierbar und nimmt das Minimum $m := \min_{x \in [a, b]} f(x)$ und das Maximum $M := \max_{x \in [a, b]} f(x)$ an Stellen x_m und $x_M \in [a, b]$ an. Damit ist wegen der Monotonie (2.44) des Integrals

$$m(b - a) \leq \int_a^b f(x) dx \leq M(b - a), \quad (2.48)$$

also

$$f(x_m) \leq \frac{\int_a^b f(x) dx}{b - a} \leq f(x_M).$$

Nach dem Zwischenwertsatz wird der Wert $\lambda := \frac{\int_a^b f(x) dx}{b - a}$ von f an einer Stelle $\xi \in [a, b]$ angenommen, die zwischen den Stellen x_m und x_M liegt oder mit einer der Stellen übereinstimmt. $\square$

Im Beweis haben wir das Integral grob mit (2.48) abgeschätzt. Diese naheliegende Ungleichung reicht in Anwendungen gelegentlich bereits aus.

Beispiel 2.86

Für $f(x) = e^x$ auf dem Intervall $[0, 2]$ gilt: $1 \leq e^x \leq e^2$. Hiermit kann der Wert des Integrals grob abgeschätzt werden:

$$1 \cdot 2 \leq \int_0^2 e^x dx \leq 2 \cdot e^2.$$

■

Mit dem Mittelwertsatz lässt sich für eine stetige Funktion f ein Funktionswert als Grenzwert einer Integral-Mittelung über eine kleine Umgebung ausrechnen: Sei f stetig auf $[a, b]$ mit $x_0 \in]a, b[$. Dann ist

$$\lim_{h \rightarrow 0} \frac{1}{h} \int_0^h f(x_0 + x) dx = \lim_{h \rightarrow 0} \frac{1}{h} \int_{x_0}^{x_0+h} f(x) dx = f(x_0). \quad (2.49)$$

Beweis: Sei $\varepsilon > 0$. Da f stetig in x_0 ist, gibt es ein $\delta > 0$ mit $[x_0 - \delta, x_0 + \delta] \subset [a, b]$ und $|f(x_0) - f(x)| < \varepsilon$ für alle $x \in [x_0 - \delta, x_0 + \delta]$. Aus dem Mittelwertsatz der Integralrechnung folgt für $|h| \leq \delta$, dass ein $\xi \in [x_0, x_0 + h]$ oder ein $\xi \in [x_0 + h, x_0]$ existiert (also insbesondere $\xi \in [x_0 - \delta, x_0 + \delta]$) mit

$$\left| \frac{1}{h} \int_0^h f(x_0 + x) dx - f(x_0) \right| = \left| \frac{1}{h} \int_{x_0}^{x_0+h} f(x) dx - f(x_0) \right| = |f(\xi) - f(x_0)| < \varepsilon.$$

Damit ist die Konvergenz mit der Grenzwertdefinition gezeigt. $\square$

2.6.3 Hauptsatz der Differenzial- und Integralrechnung

Um Integrale auszurechnen, kann man Grenzwerte von Ober- und Untersummen berechnen. Das ist aber sehr mühsam. Glücklicherweise gibt es eine elegante Alternative: die Differenzialrechnung. Integration ist quasi eine zur Differenziation inverse Operation. Zuerst hat das schon Newtons Lehrer Isaac Barrow (1630–1677) herausgefunden.

Um den Zusammenhang zwischen Integralen und der Ableitung zu erkennen, betrachten wir eine stetige Funktion f auf $[a, b]$. Für jedes $x \in [a, b]$ existiert das Integral $I(x) := \int_a^x f(t) dt$, so dass wir eine Funktion $I(x)$ der oberen Integrationsgrenze erhalten. Der Wert des Integrals $\int_a^b f(t) dt$ ist damit $I(b)$, und außerdem ist $I(a) = 0$. Kennen wir $I(x)$, so kennen wir den Wert des gesuchten Integrals: $\int_a^b f(t) dt = I(b) = I(b) - I(a)$.

Sei $a \leq x_0 < b$ und $0 < \Delta x \leq b - x_0$ so klein, dass aufgrund der Stetigkeit von f die Funktionswerte von f auf $[x_0, x_0 + \Delta x]$ ungefähr gleich $f(x_0)$ sind. Dann ist

$$I(x_0 + \Delta x) = \int_a^{x_0 + \Delta x} f(t) dt = \int_a^{x_0} f(t) dt + \int_{x_0}^{x_0 + \Delta x} f(t) dt \approx I(x_0) + f(x_0)\Delta x.$$

Wir erhalten eine näherungsweise (rechtsseitige) Ableitung von I an der Stelle x_0 über

$$I'(x_0) \approx \frac{I(x_0 + \Delta x) - I(x_0)}{\Delta x} \approx \frac{I(x_0) + f(x_0)\Delta x - I(x_0)}{\Delta x} = f(x_0).$$

Hier kürzt sich das Δx der Fläche $f(x_0)\Delta x$ mit dem Δx des Differenzenquotienten weg. Die Flächenzunahme an der Stelle x_0 entspricht also anscheinend dem Funktionswert $f(x_0)$. Tatsächlich gilt $I'(x) = f(x)$. Das ist die Aussage des Hauptsatzes/Fundamentalsatzes der Integralrechnung, die wir im Folgenden genauer formulieren und beweisen werden. Damit lässt sich $\int_a^b f(x) dx$ dann so ausrechnen: Suche eine Funktion $F(x)$, deren Ableitung $F'(x)$ gleich $f(x)$ ist. Diese muss zwar nicht mit I identisch sein, aber sie stimmt bis auf eine additive Konstante $c = F(a)$ mit I überein, wie wir unten zeigen werden. Diese Konstante ist aber für den Wert des Integrals nicht wichtig. Dieser ist

$$\int_a^b f(x) dx = I(b) - I(a) = [F(b) - c] - [F(a) - c] = F(b) - F(a).$$

Definition 2.34 (Integralfunktion und Stammfunktion)

Sei f integrierbar auf $[a, b]$ und $c \in [a, b]$, $C \in \mathbb{R}$. Die auf $[a, b]$ definierte Funktion mit der Variable x

$$I(x) := \int_c^x f(t) dt + C$$

heißt eine **Integralfunktion** oder **unbestimmtes Integral** von f .

Ist eine Funktion F auf $[a, b]$ differenzierbar mit $F'(x) = f(x)$, dann (und nur dann) heißt F eine **Stammfunktion** von f .

Bemerkungen:

- Man beachte, dass bei der Integralfunktion insbesondere auch $x \leq c$ erlaubt ist.
- Hier irritiert vielleicht, dass zwei Variablen verwendet werden: x und t . Beim Ausrechnen des Integrals wird x als Konstante angesehen. Die Zahl, die man dabei erhält, hängt natürlich von dieser Konstante x ab, aber durch die Integration kommt die Variable t nicht mehr vor. Damit entsteht eine Funktion mit der Variable x . Diese wird auch als Funktion der oberen Grenze bezeichnet.
- Stammfunktionen erhält man durch eine Umkehrung des Ableitens, quasi durch „Aufleiten“.

Von fundamentaler Bedeutung für das praktische Ausrechnen von Integralen ist nun der bereits angekündigte Hauptsatz, der deswegen auch unter dem Namen **Fundamentalsatz** zu finden ist:

Satz 2.48 (Hauptsatz der Differenzial- und Integralrechnung)

Sei f stetig auf $[a, b]$ und $c \in [a, b]$, dann ist die Integralfunktion $I(x) = \int_c^x f(t) dt + C$ (stetig) differenzierbar auf $[a, b]$, und es gilt für alle $x \in [a, b]$:

$$I'(x) = \frac{dI}{dx}(x) = \frac{d}{dx} \int_c^x f(t) dt = f(x).$$

Beweis: Sei $x \in]a, b[$ beliebig, fest. Für die Randpunkte $x \in \{a, b\}$ funktioniert der Beweis analog mit einseitigen Grenzwerten, die wir hier nicht diskutieren. Wir berechnen $I'(x)$ elementar als Grenzwert des Differenzenquotienten:

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{I(x+h) - I(x)}{h} &= \lim_{h \rightarrow 0} \frac{1}{h} \left(\int_c^{x+h} f(t) dt - \int_c^x f(t) dt \right) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left(\int_x^c f(t) dt + \int_c^{x+h} f(t) dt \right) = \lim_{h \rightarrow 0} \frac{1}{h} \int_x^{x+h} f(t) dt. \end{aligned}$$

An dieser Stelle nutzen wir wie in der Einleitung des Abschnitts die Stetigkeit von f aus: Ist h sehr klein, so sind alle Funktionswerte $f(t)$ auf dem Intervall $[x, x+h]$ ungefähr gleich $f(x)$, der Faktor $\frac{1}{h}$ kürzt sich mit der Länge des Integrationsintervalls weg, und der Grenzwert ist daher $f(x)$. Formal sauber erhalten wir dieses Ergebnis mit der Folgerung (2.49) des Mittelwertsatzes der Integralrechnung. $\square$

Folgerung 2.4 (Integral- und Stammfunktion sind äquivalente Begriffe)

Jede Integralfunktion einer stetigen Funktion ist eine Stammfunktion. Umgekehrt ist aber auch jede Stammfunktion F einer stetigen Funktion eine Integralfunktion I .

Beweis: Nach Satz 2.48 ist die Ableitung der Integralfunktion die Funktion f , die Integralfunktion ist also eine Stammfunktion. Umgekehrt zeigen wir, dass jede Stammfunktion F eine Integralfunktion ist. Dazu erhalten wir ebenfalls über Satz 2.48

$$\frac{d}{dx} \left[F(x) - \int_c^x f(t) dt \right] = f(x) - f(x) = 0,$$

so dass die abzuleitende Funktion nach Folgerung 2.5 a) auf Seite 387 eine Konstante $C \in \mathbb{R}$ sein muss: $F(x) - \int_c^x f(t) dt = C$, also $F(x) = \int_c^x f(t) dt + C$. Die Stammfunktion F ist damit aber genau die Integralfunktion $I(x) := \int_c^x f(t) dt + C$. $\square$

Die Begriffe Stammfunktion und Integralfunktion können für stetiges f also als Synonym benutzt werden.

Eine Stammfunktion F von f wird auch mit $\int f(x) dx$ – also ohne Angabe der Schranken – oder nur unter Angabe der oberen Schranke $\int^x f(t) dt$ als unbestimmtes Integral geschrieben.

Das Berechnen einer Stammfunktion ist eine „Umkehroperator“ zur Differenziation:

- $\frac{d}{dx} \int f(x) dx = f(x)$. Hier kürzt sich das dx weg.

Achtung: Das geht nur dann, wenn die obere Grenze des Integrals die Variable ist, nach der abgeleitet wird! Hat man dagegen eine Funktion f , die von den Variablen x und t abhängig ist, so ist

$$\frac{d}{dx} \int_a^b f(x, t) dt \neq f(x, ?).$$

Wir haben hier auch Schwierigkeiten, die rechte Seite hinzuschreiben, da diese nicht mehr von t abhängig ist. Ohne weitere Überlegungen darf man ein solches Integral auch nicht mit der Ableitung vertauschen.

- $\int \frac{d}{dx} f(x) dx = f(x) - C$ für eine stetig differenzierbare Funktion f . Denn hier ist f eine Stammfunktion für $\frac{d}{dx} f(x)$, die mit einer Integralfunktion $\int_c^x f'(t) dt + C$ übereinstimmt. In diesem Sinne kürzen sich auch in dieser Reihenfolge das Integrieren und Ableiten weg.

Das Berechnen von Integralen lässt sich mit dem Hauptsatz zurückführen auf das Ermitteln von Stammfunktionen. Hierbei hilft dann die Differenzialrechnung. Das ist entscheidend, damit wir nicht bei jedem Integral Grenzwerte wie in (2.42) ausrechnen müssen!

Ist $F(x)$ eine (beliebige) Stammfunktion zu einer stetigen Funktion f , so ist F eine Integralfunktion zu f und hat damit eine Darstellung $F(x) = \int_c^x f(t) dt + C$, $C = F(c)$. Damit erhalten wir die bereits eingangs formulierte wichtige Verbindung zwischen Ableitung und Integration, über die Integrale berechnet werden:

$$\begin{aligned}
 F(b) - F(a) &= \int_c^b f(t) dt + C - \int_c^a f(t) dt - C \\
 &= \int_a^c f(t) dt + \int_c^b f(t) dt = \int_a^b f(t) dt.
 \end{aligned} \tag{2.50}$$

Statt $F(b) - F(a)$ schreiben wir auch kurz $F(x) \Big|_a^b$ oder $[F(x)]_a^b$.

Man beachte, dass dieser Ansatz auch funktioniert, wenn a größer als b ist. In diesem Fall ist

$$\int_a^b f(t) dt = - \int_b^a f(t) dt = -[F(a) - F(b)] = F(b) - F(a).$$

Beispiel 2.87

a) $\sin(x)$ ist eine Stammfunktion von $\cos(x)$. Also ist

$$\int_0^{\pi/2} \cos(x) dx = \sin(x) \Big|_0^{\pi/2} = [\sin(x)]_0^{\pi/2} = \sin(\pi/2) - \sin(0) = 1.$$

b) $\frac{1}{m+1}x^{m+1} + C$ ist eine Stammfunktion zu x^m .

$$\int_0^1 x^m dx = \left[\frac{1}{m+1} x^{m+1} + C \right]_0^1 = \frac{1}{m+1} 1^{m+1} + C - 0 - C = \frac{1}{m+1}. \quad \blacksquare$$

Um ein Integral auszurechnen, muss man also „nur“ eine Stammfunktion ermitteln. Im nächsten Kapitel werden wir dazu einige Techniken betrachten.

Hintergrund: Ein anderer Blick auf die Integration mit Stammfunktion

Wir zeigen, dass sich die Regel (2.50) nicht nur aus dem Hauptsatz der Differenzial und Integralrechnung ergibt, sondern auch direkt aus dem Mittelwertsatz der Differenzialrechnung. Sei dazu f integrierbar auf einem Intervall $[a, b]$ mit einer Stammfunktion F . Wir wollen

$$F(b) - F(a) = \int_a^b f(x) dx$$

zeigen. Dazu erweitern wir die Differenz $F(b) - F(a)$ zu einer Teleskopsumme, auf deren Summanden wir dann den Mittelwertsatz (Satz 2.36 auf Seite 315) anwenden. Es entsteht eine Riemann-Zwischensumme, die wegen des Riemann'schen Integrabilitätskriteriums (siehe Seite 333) gegen das Integral konvergiert. Sei $n \in \mathbb{N}$:

$$\begin{aligned}
 F(b) - F(a) &= F(b) - F\left(b - \frac{b-a}{n}\right) + F\left(b - \frac{b-a}{n}\right) - F\left(b - 2\frac{b-a}{n}\right) \pm \cdots - F(a) \\
 &= \sum_{k=0}^{n-1} F\left(b - k\frac{b-a}{n}\right) - F\left(b - (k+1)\frac{b-a}{n}\right).
 \end{aligned}$$

Auf jeden Summanden wenden wir den Mittelwertsatz an. Es gibt Zwischenstellen ξ_k in den Intervallen $]b - (k+1)\frac{b-a}{n}, b - k\frac{b-a}{n}[$ der Länge $\frac{b-a}{n}$, so dass

$$F(b) - F(a) = \sum_{k=0}^{n-1} F'(\xi_k) \frac{b-a}{n} = \sum_{k=0}^{n-1} f(\xi_k) \frac{b-a}{n}.$$

Die rechte Seite ist eine Riemann-Zwischensumme zur Zerlegung $(a, a + \frac{b-a}{n}, \dots, b)$ des Intervalls $[a, b]$. Da f nach Voraussetzung integrierbar ist, ist das Riemann'sche Integrabilitätskriterium anwendbar, und die rechte Seite strebt für $n \rightarrow \infty$ gegen das Integral. Da die linke Seite unabhängig von n ist, ist damit die Rechenregel bewiesen:

$$F(b) - F(a) = \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} f(\xi_k) \frac{b-a}{n} = \int_a^b f(x) dx.$$

2.6.4 Rechenregeln zur Integration

Das Ausrechnen von Integralen kann deutlich schwieriger sein als die Bestimmung einer Ableitung. Dies liegt daran, dass die Ableitung einer elementaren Funktion wieder eine elementare Funktion ist, während das unbestimmte Integral einer elementaren Funktion keine elementare Funktion sein muss. **Elementare Funktionen** sind die Polynome und gebrochen-rationalen Funktionen, die Funktionen $\sin(x)$ und $\cos(x)$, $\exp(x)$ sowie alle durch Zusammensetzung oder Umkehrung daraus entstehenden Funktionen. Man kann zeigen, dass der als Integralfunktion des stetigen Sinus Cardinalis definierte Integralsinus

$$\text{Si}(x) := \int_0^x \text{sinc}(t) dt = \int_0^x \frac{\sin t}{t} dt$$

keine elementare Funktion ist. Gleiches gilt für $\int_0^x e^{(t^2)} dt$ (Liouville, 1835). Da es kein einfaches allgemeines Verfahren zur Berechnung einer Stammfunktion gibt, werden wir uns mit den folgenden Rechenregeln behelfen. Computer-Algebra-Systeme können aber sehr wohl automatisch Stammfunktionen berechnen, sofern diese sich aus elementaren Funktionen zusammensetzen. Dahinter steckt der komplizierte Risch-Algorithmus (1968), der ausnutzt, dass sich elementare Stammfunktionen aus elementaren Bestandteilen der zu integrierenden Funktion und Logarithmusfunktionen zusammensetzen.

Beispiel 2.88

Bereits mit der Linearität des Integrals und der Stammfunktionen aus Tabelle 2.3 lassen sich einfache Integrale berechnen:

$$\text{a) } \int x^3 - 2x^2 + 5x dx = \int x^3 dx - 2 \int x^2 dx + 5 \int x dx = \frac{1}{4}x^4 - \frac{2}{3}x^3 + \frac{5}{2}x^2 + C.$$

Tab. 2.3: Ausgewählte Stammfunktionen (unbestimmte Integrale)

$\int x^k dx$	$= \frac{x^{k+1}}{k+1} + C, k \neq -1$
$\int \frac{1}{x} dx$	$= \ln x + C$
$\int e^x dx$	$= e^x + C$
$\int \ln(x) dx$	$= x \cdot \ln(x) - x + C$
$\int \cos(x) dx$	$= \sin(x) + C$
$\int \sin(x) dx$	$= -\cos(x) + C$
$\int \cosh(x) dx$	$= \sinh(x) + C$
$\int \sinh(x) dx$	$= \cosh(x) + C$
$\int \frac{1}{1+x^2} dx$	$= \arctan(x) + C$
$\int \frac{1}{1-x^2} dx$	$= \operatorname{artanh}(x) + C = \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right) + C, x < 1$
$\int \frac{1}{1-x^2} dx$	$= \operatorname{arcoth}(x) + C = \frac{1}{2} \ln \left(\frac{x+1}{x-1} \right) + C, x > 1$
$\int \frac{1}{\sqrt{1-x^2}} dx$	$= \arcsin(x) + C$
$\int \frac{1}{\sqrt{1-x^2}} dx$	$= -\arccos(x) + C$
$\int \frac{1}{\sqrt{x^2+1}} dx$	$= \operatorname{arsinh}(x) + C = \ln(x + \sqrt{x^2+1}) + C$
$\int \frac{1}{\sqrt{x^2-1}} dx$	$= \operatorname{arcosh}(x) + C = \ln(x + \sqrt{x^2-1}) + C, x > 1$

b) $\int_0^{\pi/2} (2x + \sin x) dx = 2 \int_0^{\pi/2} x dx + \int_0^{\pi/2} \sin x dx = 2 \left[\frac{1}{2} x^2 \right]_0^{\pi/2} + [-\cos x]_0^{\pi/2} = \frac{\pi^2}{4} + 1.$

■

2.6.4.1 Partielle Integration

Die Produktregel der Differenzialrechnung führt zur **partiellen Integration**.

Satz 2.49 (Partielle Integration)

Seien f und g stetig differenzierbar auf $[a, b]$, dann gilt:

$$\int_a^b f(x) \cdot g'(x) dx = f(x) \cdot g(x) \Big|_a^b - \int_a^b f'(x) \cdot g(x) dx.$$

Beweis: Die Regel folgt sofort aus dem Hauptsatz der Differenzial- und Integralrechnung und der Produktregel der Differenzialrechnung (Satz 2.32):

$$\begin{aligned} f(x)g(x) \Big|_a^b &\stackrel{(2.50)}{=} \int_a^b (f(x)g(x))' dx \text{ (Hauptsatz)} \\ &= \int_a^b f'(x)g(x) + f(x)g'(x) dx \text{ (Produktregel)} \\ &= \int_a^b f'(x)g(x) dx + \int_a^b f(x)g'(x) dx. \end{aligned} \quad \square$$

Sind f und G stetig differenzierbar auf $[a, b]$ und ist $g(x) = G'(x)$, d. h., G ist eine Stammfunktion zu g , so lässt sich die Regel zur partiellen Integration auch schreiben als

$$\int_a^b f(x)g(x) dx = f(x)G(x) \Big|_a^b - \int_a^b f'(x)G(x) dx.$$

Beispiel 2.89

a) $\int x \cdot \sin(x) dx = -x \cdot \cos(x) + \int \cos(x) dx = -x \cdot \cos(x) + \sin(x) + C.$

Hier ist $f(x) = x$, $f'(x) = 1$ und $g'(x) = \sin x$ und $g(x) = -\cos x$.

b) $\int_0^{2\pi} x \cos x dx = x \sin x \Big|_0^{2\pi} - \int_0^{2\pi} 1 \cdot \sin x dx = 0 - [-\cos x]_0^{2\pi} = \cos(2\pi) - \cos 0 = 0.$

c) Beim Integral $\int \ln x dx$ sieht man zunächst nicht, dass partielle Integration weiterhilft. Wenn wir aber $f(x) = \ln x$, $f'(x) = \frac{1}{x}$ und $g'(x) = 1$, $g(x) = x$ wählen, also den Integranden mit 1 multiplizieren, dann ergibt sich

$$\int \ln x dx = \int 1 \cdot \ln x dx = x \cdot \ln(x) - \int x \cdot \frac{1}{x} dx = x \cdot \ln(x) - x + C.$$

d) $\int_1^2 x \cdot \ln x dx = \left[\frac{1}{2} x^2 \cdot \ln x \right]_1^2 - \int_1^2 \frac{1}{2} x dx = \left[\frac{1}{2} x^2 \cdot \ln x \right]_1^2 - \left[\frac{1}{4} x^2 \right]_1^2 = 2 \ln(2) - \frac{3}{4}.$ Hier ist $f(x) = \ln x$, $f'(x) = \frac{1}{x}$ und $g'(x) = x$, $g(x) = \frac{1}{2} x^2$.

e) $\int_0^1 x e^x dx = [x e^x]_0^1 - \int_0^1 e^x dx = e - [e^x]_0^1 = e - e + 1 = 1. \quad \blacksquare$

Beispiel 2.90

Im Umgang mit Schwingungen wird häufig folgendes Integral benötigt, bei dem wir als Trick die trigonometrische Fassung des Satzes von Pythagoras $\sin^2 x + \cos^2 x = 1$ (siehe Seite 118) verwenden:

$$\begin{aligned} \int \sin^2 x \, dx &= -\sin(x) \cdot \cos(x) + \int \cos^2 x \, dx \\ &= -\sin(x) \cdot \cos(x) + \int (1 - \sin^2 x) \, dx = -\sin(x) \cdot \cos(x) + x - \int \sin^2 x \, dx, \end{aligned}$$

und somit folgt

$$\int \sin^2 x \, dx = \frac{1}{2}[x - \sin(x) \cdot \cos(x)] + C. \quad (2.51)$$

Damit haben wir auch

$$\int \cos^2 x \, dx = \int 1 - \sin^2 x \, dx = x - \frac{1}{2}[x - \sin(x) \cdot \cos(x)] + C = \frac{1}{2}[x + \sin(x) \cdot \cos(x)] + C. \quad (2.52)$$

Für die Anwendung der partiellen Integration in Verbindung mit trigonometrischen Funktionen ist es typisch, dass man wieder auf das Ausgangsintegral stößt. Hat es einen von Eins verschiedenen Vorfaktor, so ist man nach Umstellung der Gleichung fertig. Hat es den Vorfaktor Eins, so hat man durch die Rechnung leider nichts erreicht, und man muss einen anderen Lösungsweg suchen. Diese Situation tritt z. B. ein, wenn wir statt der Auflösung $\cos^2 x = 1 - \sin^2 x$ ein weiteres Mal partiell integriert hätten:

$$\begin{aligned} \int \sin^2 x \, dx &= -\sin(x) \cdot \cos(x) + \int \cos^2 x \, dx \\ &= -\sin(x) \cdot \cos(x) + \sin(x) \cdot \cos(x) - \int \sin(x) \cdot (-\sin(x)) \, dx = \int \sin^2 x \, dx. \end{aligned}$$

■

Beispiel 2.91

Der Effektivwert einer Wechselspannung gibt die Größe einer Gleichspannung $U = u_{\text{eff}}$ an, die an einem Ohm'schen Widerstand R die gleiche Wirkleistung $P = UI = \frac{U^2}{R}$ hat wie die Wechselspannung. Dabei ist für eine 2π -periodische Spannung $u(t) = \hat{u} \sin t$ mit Maximalwert $\hat{u} > 0$ die Wirkleistung $P := \frac{1}{2\pi} \int_0^{2\pi} u(t)i(t) \, dt = \frac{1}{2\pi} \int_0^{2\pi} u(t)\frac{u(t)}{R} \, dt$. Damit ergibt sich für den Effektivwert die Gleichung

$$\frac{u_{\text{eff}}^2}{R} = \frac{1}{2\pi} \int_0^{2\pi} u(t) \frac{u(t)}{R} \, dt.$$

Die positive Lösung ist

$$u_{\text{eff}} = \sqrt{\frac{1}{2\pi} \int_0^{2\pi} u(t)^2 \, dt} = \sqrt{\frac{1}{2\pi} \int_0^{2\pi} (\hat{u} \sin t)^2 \, dt} = \hat{u} \cdot \sqrt{\frac{1}{2\pi} \int_0^{2\pi} \sin^2 t \, dt}$$

$$\stackrel{(2.51)}{=} \hat{u} \cdot \sqrt{\frac{1}{4\pi} [x - \sin(x) \cdot \cos(x)]_0^{2\pi}} = \frac{\hat{u}}{\sqrt{2}}.$$

Bei kosinus- bzw. sinusförmigen Wechselspannungen wird häufig der Effektivwert auch in dieser Darstellung definiert. Der Effektivwert der Steckdosenspannung ist $u_{\text{eff}} = 230 \text{ V}$, der Maximalwert ist $\hat{u} = \sqrt{2} \cdot 230 \text{ V} \approx 325 \text{ V}$. ■

2.6.4.2 Integralsubstitution

Analog zur Produktregel führt auch die Kettenregel der Differenzialrechnung (siehe Satz 2.33) zu einer weiteren sehr nützlichen und wichtigen Integrationsregel:

Satz 2.50 (Substitutionsregel)

Seien g stetig differenzierbar auf $[a, b]$ und f stetig auf dem Intervall $g([a, b])$. Dann ist

$$\int_a^b f(g(x))g'(x) dx = \int_{g(a)}^{g(b)} f(t) dt. \quad (2.53)$$

Hier hat man die Substitution $t = g(x)$ vorgenommen. Rein formal kann man sagen: $dt = g'(x) dx$.

Beweis: Zunächst erinnern wir uns, dass die Menge $g([a, b])$ als Konsequenz des Zwischenwertsatzes tatsächlich ein Intervall ist (siehe Satz 2.27 auf Seite 282).

Sei F eine Stammfunktion von f . Nach Voraussetzung ist $F \circ g$ differenzierbar. Mit der Kettenregel (Satz 2.33, Seite 297) gilt:

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

Diese Funktion ist nach Voraussetzung stetig, so dass mit dem Hauptsatz gilt (siehe (2.50)):

$$\int_a^b f(g(x))g'(x) dx = \int_a^b (F(g(x)))' dx = F(g(x)) \Big|_a^b = F(t) \Big|_{g(a)}^{g(b)} = \int_{g(a)}^{g(b)} f(t) dt. \quad \square$$

Die Regel (2.53) kann sowohl von links nach rechts als auch von rechts nach links angewendet werden. Damit kann man auch substituieren, wenn man keine Ableitung einer inneren Funktion sieht. Auch muss g nicht umkehrbar sein. Existiert jedoch die Umkehrfunktion g^{-1} (d. h., die Gleichung $t = g(x)$ kann nach x aufgelöst werden: $x = g^{-1}(t)$), so kann man auch so rechnen ($\alpha := g(a)$, $\beta := g(b)$):

$$\int_{\alpha}^{\beta} f(t) dt = \int_{g^{-1}(\alpha)}^{g^{-1}(\beta)} f(g(x))g'(x) dx, \quad (2.54)$$

wobei wieder formal $t = g(x)$ und $dt = g'(x) dx$ substituiert wurde.

Beispiel 2.92

Mittels Substitution $t = \sin x$, $dt = \cos x dx$ erhalten wir

$$\int_0^{\pi/2} \exp(\sin x) \cos x dx = \int_{\sin(0)}^{\sin(\pi/2)} \exp(t) dt = [e^t]_0^1 = e - 1.$$

■

Beispiel 2.93

Häufig sind Funktionen des Typs $f(c \cdot x)$ mit einer Konstante $c \in \mathbb{R}$ zu integrieren. Hier hilft die Substitution $t = c \cdot x$, $dt = c dx$:

$$\int_a^b f(c \cdot x) dx = \int_{ca}^{cb} f(t) \frac{1}{c} dt = \frac{1}{c} [F(c \cdot x)]_a^b.$$

Hier kann alternativ zur Substitution direkt die Stammfunktion F von f berechnet werden. Dann setzt man die innere Funktion $c \cdot x$ ein. Aufgrund der Kettenregel ist $\frac{d}{dx} F(c \cdot x) = f(c \cdot x) \cdot c$, so dass man noch mit dem Faktor $\frac{1}{c}$ multiplizieren muss, um eine Stammfunktion für $f(c \cdot x)$ zu erhalten. Diese Überlegung entspricht einem Rückwärtslesen der Kettenregel und damit genau dem Beweis der Substitutionsregel. Im Kopf lässt sich also leicht die Stammfunktion $-\frac{1}{2} \cos(2x)$ von $\sin(2x)$ bilden, indem man nur den Sinus hochintegriert zu $-\cos(2x)$ und diese Funktion dann mit der Kettenregel ableitet, um den Korrekturfaktor $\frac{1}{2}$ zu finden. ■

Die formale Anwendung der Substitutionsregel ist unkritisch. Im Ausgangsintegral werde über $[a, b]$ nach x integriert, dies soll durch eine Substitution in eine Integration nach t überführt werden. Hier gibt es zwei Ansätze:

- a) Falls man $t = g(x)$ setzt, ist $dt = g'(x) dx$, und t läuft in den Grenzen von $g(a)$ bis $g(b)$.
- b) Falls man $x = g(t)$ setzt, ist $dx = g'(t) dt$, und t läuft in den Grenzen von $g^{-1}(a)$ bis $g^{-1}(b)$.

Im Fall b) muss die Umkehrfunktion zu g existieren. Ob sie existiert, erkennt man aber ohnehin bei Anwendung der Regel, da man sie berechnen muss.

Kochrezept: Die alte Integrationsvariable sei x , die neue Variable sei $t = g(x)$. Dann gehe man wie folgt vor:

- Ersetze dx durch $dt/g'(x)$.
- Neue Integrationsgrenzen sind $g(a)$ und $g(b)$.
- Ersetze jedes Auftreten von $g(x)$ durch t (das kann auch dann geschehen, wenn $g(x)$ nicht umkehrbar zu $g^{-1}(t)$ ist).

- Falls g umkehrbar ist und die Variable x noch auftritt: Ersetze alle verbleibenden x durch $g^{-1}(t)$. (Dieser Schritt macht den vorangehenden überflüssig, da $g(x)$ ersetzt wird durch $g(g^{-1}(t)) = t$.)
- Falls g nicht umkehrbar ist aber dennoch die Variable x noch auftritt, gelingt die gewählte Substitution nicht.

Die Substitutionsregel gilt auch unter schwächeren Voraussetzungen an die Funktionen f und g . Wir werden sie beispielsweise für eine linear affine Funktion $g(x) := mx + c$ mit $g(a) = \alpha$ und $g(b) = \beta$ und eine Funktion f , die auf dem Intervall zwischen α und β integrierbar (aber nicht unbedingt stetig) ist, anwenden. Direkt mittels der Definition des Integrals kann man nachrechnen, dass gilt ($t = g(x)$, $dt = m dx$):

$$\int_a^b f(g(x)) dx = \frac{1}{m} \int_\alpha^\beta f(t) dt. \quad (2.55)$$

Benutzt man die Substitutionsregel, um eine Stammfunktion (ein unbestimmtes Integral) zu ermitteln, so wird häufig eine Schreibweise ohne Grenzen verwendet. Dann muss man aber am Ende der Rechnung die Substitution wieder rückgängig machen. Beispielsweise kann man so vorgehen:

- $\int x e^{(x^2)} dx = \int_c^x t e^{(t^2)} dt \stackrel{u=t^2}{=} \frac{1}{2} \int_{c^2}^{x^2} e^u du = \frac{1}{2} e^u \Big|_{c^2}^{x^2} = \frac{1}{2} e^{(x^2)} + C,$
- $\int x e^{(x^2)} dx = \int^x t e^{(t^2)} dt \stackrel{u=t^2}{=} \frac{1}{2} \int^{x^2} e^u du = \frac{1}{2} e^u \Big|^{x^2} = \frac{1}{2} e^{(x^2)} + C,$
- $\int x e^{(x^2)} dx \stackrel{u=x^2}{=} \frac{1}{2} \int e^u du = \frac{1}{2} e^u + C \stackrel{x^2=u}{=} \frac{1}{2} e^{(x^2)} + C.$

Beispiel 2.94

- a) $\int \sqrt{1+4x} dx = \int \sqrt{u} \cdot \frac{1}{4} du = \frac{1}{4} \cdot \frac{2}{3} u^{3/2} + C = \frac{1}{6} u^{3/2} + C = \frac{1}{6} \sqrt{1+4x}^3 + C$. Dabei ist $f(u) = \sqrt{u}$, $u = 1 + 4x$, d. h. $du = 4 dx$ und $\int f(u) du = \frac{2}{3} u^{3/2} + C$.
- b) Mit der Substitution $u = -3x + 3$, d. h. $du = -3 dx$ erhalten wir

$$\int e^{-3x+3} dx = -\frac{1}{3} \int e^u du = -\frac{1}{3} e^u + C = -\frac{1}{3} e^{-3x+3} + C.$$

- c) Mit $u = \sin x$ und $du = \cos x dx$ erhalten wir

$$\int \sin^2(x) \cdot \cos(x) dx = \int u^2 du = \frac{1}{3} u^3 + C = \frac{1}{3} \sin^3(x) + C.$$

- d) $\int (8-5x)^7 dx = -\frac{1}{5} \int u^7 du = -\frac{1}{5} \cdot \frac{1}{8} u^8 + C = -\frac{1}{40} (8-5x)^8 + C$. Dabei haben wir $u = 8-5x$ gewählt, so dass $du = -5 dx$.
- e) Mit $u = 1+x^2$, also $du = 2x dx$, erhalten wir

$$\int \frac{x}{\sqrt{1+x^2}} dx = \frac{1}{2} \int \frac{1}{\sqrt{u}} du = \sqrt{u} + C = \sqrt{1+x^2} + C. \quad \blacksquare$$

Ein Spezialfall der Substitutionsregel ist das **logarithmische Integrieren**. Dazu integrieren wir zunächst $\frac{1}{x}$. Für $x > 0$ kennen wir eine Stammfunktion $\ln(x)$. Falls beide Integrationsgrenzen kleiner null sind, also $a, b < 0$, erhalten wir mittels der Substitution $t = -x$, $dt = -dx$ das Integral

$$\int_a^b \frac{1}{x} dx = \int_{-a}^{-b} \frac{1}{t} dt = \int_{|a|}^{|b|} \frac{1}{t} dt = \ln t \Big|_{|a|}^{|b|} = \ln |t| \Big|_a^b.$$

Damit gilt sowohl für $a, b < 0$ als auch für $a, b > 0$:

$$\int_a^b \frac{1}{x} dx = \ln |x| \Big|_a^b.$$

Achtung: Wenn Sie die Stammfunktion von $\frac{1}{x}$ generell mit $\ln|x|$ (statt mit $\ln x$) angeben, machen Sie auch bei einem negativen Integrationsintervall keinen Fehler.

Falls eine stetige Funktion f keine Nullstelle auf $[a, b]$ hat, ist $1/f(x)$ hier stetig und beschränkt. Entweder sind dann sowohl $f(a)$ und $f(b)$ beide positiv oder beide negativ. In diesem Fall ergibt sich mit der Substitution $t = f(x)$, $dt = f'(x) dx$ die Regel für das logarithmische Integrieren:

$$\int_a^b \frac{f'(x)}{f(x)} dx = \int_{f(a)}^{f(b)} \frac{1}{t} dt = \ln |t| \Big|_{f(a)}^{f(b)} = \ln(|f(b)|) - \ln(|f(a)|) = \ln \left(\frac{|f(b)|}{|f(a)|} \right).$$

Man beachte, dass die Einschränkung an $f(a)$ und $f(b)$ erforderlich ist, da $1/t$ im Nullpunkt nicht beschränkt ist und damit in keinem Intervall um Null integriert werden kann.

Beispiel 2.95

$$\int_0^{\pi/4} \tan x dx = - \int_0^{\pi/4} \frac{\sin x}{\cos x} dx = - \ln \left(\frac{|\cos \frac{\pi}{4}|}{|\cos 0|} \right) = - \ln \left(\frac{1}{\sqrt{2}} \right) = \ln(\sqrt{2}).$$

■

- Ein Integral mit Termen der Form $\sqrt{a^2 - x^2}$ mit $a > 0$ lässt sich häufig durch die Substitution $x = a \cdot \cos u$ vereinfachen.
- Ein Integral mit Termen der Form $\sqrt{a^2 + x^2}$ mit $a > 0$ lässt sich mitunter durch die Substitution $x = a \cdot \sinh u$ in den Griff bekommen.
- Bei einem Integral mit Termen der Form $\sqrt{x^2 - a^2}$ mit $a > 0$ führt häufig die Substitution $x = a \cdot \cosh u$ zum Ziel.

Beispiel 2.96

- a) Mit der Substitution $x = a \cdot \sinh u$, $dx = a \cdot \cosh u du$, $u = \operatorname{arsinh} \left(\frac{x}{a} \right)$ erhalten wir ($a > 0$)

$$\int \frac{dx}{\sqrt{a^2 + x^2}} = \int^{\operatorname{arsinh}(x/a)} \frac{a \cdot \cosh u du}{a \sqrt{1 + \sinh^2 u}} = \int^{\operatorname{arsinh}(x/a)} 1 du = \operatorname{arsinh} \left(\frac{x}{a} \right) + C.$$

Dabei haben wir $\sinh^2 x + 1 = \cosh^2 x > 0$ verwendet.

- b) Wir berechnen für $a > 0$ auf dem Intervall $[-a, a]$ eine Stammfunktion von $\sqrt{a^2 - x^2}$. Dazu benutzen wir die Substitution $x = a \cdot \cos u$, $dx = -a \cdot \sin u$, $u = \arccos\left(\frac{x}{a}\right) \in [0, \pi]$ (beachte $\sin u \geq 0$):

$$\begin{aligned} \int \sqrt{a^2 - x^2} dx &= a \int^{\arccos(x/a)} \sqrt{1 - \cos^2 u} \cdot (-a \sin u) du \\ &= -a^2 \int^{\arccos(x/a)} \sin^2 u du \stackrel{(2.51)}{=} -\frac{a^2}{2} [u - \sin(u) \cdot \cos(u)]^{\arccos(x/a)} \\ &= -\frac{a^2}{2} \arccos\left(\frac{x}{a}\right) + \frac{a^2}{2} \left[\sqrt{1 - \cos^2 u} \cdot \cos u \right]^{\arccos(x/a)} \\ &= -\frac{a^2}{2} \arccos\left(\frac{x}{a}\right) + \frac{1}{2} x \sqrt{a^2 - x^2} + C. \end{aligned}$$

Beispiel 2.97 (Fläche des Einheitskreises)

Die Fläche A des halben Einheitskreises ist gleich der Fläche unter dem Graph der Funktion $\sqrt{1 - x^2}$ zwischen $x = -1$ und $x = 1$:

$$A = \int_{-1}^1 \sqrt{1 - x^2} dx.$$

Die Substitution $x = \sin t$, d. h. $dx = \cos t dt$, $t = \arcsin x$, liefert

$$\begin{aligned} A &= \int_{-1}^1 \sqrt{1 - x^2} dx = \int_{-\pi/2}^{\pi/2} \sqrt{1 - \sin^2 t} \cdot \cos t dt = \int_{-\pi/2}^{\pi/2} \cos^2 t dt \\ &\stackrel{(2.52)}{=} \left[\frac{1}{2} (t + \sin(t) \cdot \cos(t)) \right]_{-\pi/2}^{\pi/2} = \frac{\pi}{4} + \frac{\pi}{4} = \frac{\pi}{2}. \end{aligned}$$

Wie zu erwarten war, erhalten wir für die Fläche des Einheitskreises den Wert $2 \cdot A = \pi$.

Beispiel 2.98 (Anwendung der Substitutionsregel in der Physik)

Die kinetische Arbeit ist als Kraft F mal Weg s definiert. Ändert sich die Kraft nur an einzelnen Punkten, 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, siehe Seite 681). 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 $v_b = v(t_b)$

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

Die Substitutionsregel wird häufig kalkülhaft angewendet. Das gelingt, 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“. ■

2.6.4.3 Integration gebrochen-rationaler Funktionen

Seien $p_n(x) = \sum_{k=0}^n a_k x^k$, $q_m(x) := \sum_{k=0}^m b_k x^k$ mit $a_k, b_k \in \mathbb{R}$ und $n, m \in \mathbb{N}$ das Zähler- und das Nennerpolynom einer gebrochen rationalen Funktion. Zu berechnen ist $\int \frac{p_n(x)}{q_m(x)} dx$.

Im Gegensatz zur Anwendung der Substitutionsregel gibt es hier einen klar vorgegebenen Weg, wie diese Stammfunktion bestimmt werden kann. Zunächst muss man mittels Polynomdivision dafür sorgen, dass der Grad des Zählers kleiner als der Grad des Nenners ist, es sich also um eine echt gebrochen-rationale Funktion handelt (vgl. Seite 100). Dabei wird ein Polynom abgespalten, das leicht zu integrieren ist. Wir kümmern uns jetzt um den Rest, bei dem der Grad des Zählerpolynoms p echt kleiner als der Grad m des Nennerpolynoms q ist. Diese Funktion kann man als Summe „einfacherer“ gebrochen-rationaler Funktionen schreiben, als sogenannte **Partialbruchzerlegung**. Hierbei spielen die Nullstellen des Nennerpolynoms q eine entscheidende Rolle. Nach dem Fundamentalsatz der Algebra hat $q(x)$ genau m komplexe Nullstellen unter Berücksichtigung ihrer Vielfachheit. Da wir Polynome mit reellen Koeffizienten

betrachten, treten echt komplexe Nullstellen immer auch konjugiert-komplex auf (siehe Lemma 1.13 auf Seite 148).

Einfache reelle Nullstellen des Nenners

Lemma 2.9 (Partialbruchzerlegung bei einfachen reellen Nullstellen)

Sei $q(x) = \prod_{k=1}^m (x - x_k)$ mit m verschiedenen, einfachen, reellen Nullstellen $x_1, x_2, \dots, x_m$ und $p(x)$ ein Polynom kleineren Grades als $q(x)$. Dann gibt es eine (eindeutige) Darstellung mit reellen Zahlen $A_1, A_2, \dots, A_m$:

$$\frac{p(x)}{q(x)} = \frac{A_1}{x - x_1} + \frac{A_2}{x - x_2} + \dots + \frac{A_m}{x - x_m}. \quad (2.56)$$

Beweis: Multipliziert man beide Seiten mit $q(x)$, so steht nach dem Kürzen sowohl links als auch rechts ein Polynom mit Grad kleiner m :

$$p(x) = A_1 \frac{q(x)}{x - x_1} + A_2 \frac{q(x)}{x - x_2} + \dots + A_m \frac{q(x)}{x - x_m}. \quad (2.57)$$

Setzt man nun in beide Seiten eine Nullstelle x_k ein, so erhält man

$$p(x_k) = A_k \prod_{i=1, i \neq k}^m (x_k - x_i), \text{ also } A_k = \frac{p(x_k)}{\prod_{i=1, i \neq k}^m (x_k - x_i)}.$$

Die Konstanten A_k sind dadurch so festgelegt, dass beide Seiten von (2.57) für die m Werte $x_1, \dots, x_m$ übereinstimmen. Da beide Seiten aber Polynome vom Grad kleiner m sind, sind sie nach Folgerung 1.6 auf Seite 148 gleich. $\square$

Der Beweis liefert direkt das Verfahren, mit dem die Partialbruchzerlegung berechnet wird: Beide Seiten der Gleichung (2.56) werden mit $q(x)$ multipliziert. Dann werden zur Berechnung der Konstanten die Nullstellen eingesetzt. Die Konstanten können so direkt abgelesen werden.

Alternativ kann man die Zahlen $A_1, \dots, A_m$ auch über einen Koeffizientenvergleich bestimmen. Die beiden Seiten von (2.57) stimmen genau dann überein, wenn die Koeffizienten der beiden Seiten gleich sind. Man erhält so ein Gleichungssystem, das eine eindeutige Lösung für $A_1, \dots, A_m$ liefert.

Beispiel 2.99

Wir betrachten die gebrochen-rationale Funktion

$$\frac{p(x)}{q(x)} = \frac{6x^2 - x + 2}{x^3 - x}.$$

Um eine Partialbruchzerlegung durchzuführen, benötigen wir die Linearfaktoren des Nenners $x^3 - x = x(x^2 - 1) = x(x - 1)(x + 1)$. Damit müssen wir die Konstanten A , B und C bestimmen in

$$\frac{6x^2 - x + 2}{x^3 - x} = \frac{A}{x} + \frac{B}{x - 1} + \frac{C}{x + 1}.$$

Multiplikation mit dem Hauptnenner ergibt:

$$6x^2 - x + 2 = A(x - 1)(x + 1) + Bx(x + 1) + Cx(x - 1).$$

Jetzt gibt es mehrere Möglichkeiten zur Bestimmung von A, B, C . Mittels Koeffizientenvergleich erhält man z. B. ein Gleichungssystem. Eleganter ist hier das Einsetzen der Nullstellen, da dabei jeweils nur eine Konstante übrig bleibt:

- Für $x = 0$: $2 = -A$, also $A = -2$.
- Für $x = 1$: $7 = 2B$, also $B = 7/2$.
- Für $x = -1$: $9 = 2C$, also $C = 9/2$.

$$\frac{6x^2 - x + 2}{x^3 - x} = -\frac{2}{x} + \frac{\frac{7}{2}}{x - 1} + \frac{\frac{9}{2}}{x + 1}.$$

Wir integrieren nun diese Funktion:

$$\begin{aligned} \int_2^3 \frac{p(x)}{q(x)} dx &= \int_2^3 -\frac{2}{x} + \frac{7}{2(x - 1)} + \frac{9}{2(x + 1)} dx \\ &= \left[-2 \ln |x| + \frac{7}{2} \ln |x - 1| + \frac{9}{2} \ln |x + 1| \right]_2^3. \quad \blacksquare \end{aligned}$$

Hat man nur einfache reelle Nullstellen des Nenners, so führt die Integration mittels Partialbruchzerlegung zu Integralen des Typs

$$\int \frac{1}{x - x_0} dx = \ln |x - x_0| + C.$$

Beispiel 2.100

$$\begin{aligned} \int_2^3 \frac{4x}{x^2 - 1} dx &= \int_2^3 \frac{2}{x - 1} + \frac{2}{x + 1} dx = 2 [\ln |x - 1| + \ln |x + 1|]_2^3 \\ &= 2 [\ln |(x - 1)(x + 1)|]_2^3 = 2 [\ln(8) - \ln(3)] = 2 \ln \frac{8}{3}. \quad \blacksquare \end{aligned}$$

Dass der Ansatz nicht funktioniert, wenn der Grad des Zählerpolynoms größer oder gleich dem des Nennerpolynoms ist, sieht man an diesem Beispiel: $\frac{x}{x-1} = \frac{A}{x-1}$ kann für keine Konstante A gelten, da nach Multiplikation mit dem Nenner $A = x$ gelten müsste.

Mehrfache reelle Nullstellen des Nenners

Wir erweitern Lemma 2.9:

Lemma 2.10 (Partialbruchzerlegung bei mehrfachen reellen Nullstellen)

Sei $q(x) = \prod_{k=1}^r (x-x_k)^{l_k}$ mit $r \leq m$ und reellen Nullstellen x_1 der Vielfachheit l_1 , x_2 der Vielfachheit l_2 usw. Insbesondere ist damit $l_1 + l_2 + \dots + l_r = m$. Außerdem sei $p(x)$ ein Polynom kleineren Grades als $q(x)$. Dann gibt es eine (eindeutige) Darstellung

$$\frac{p(x)}{q(x)} = \sum_{k=1}^r \left[\sum_{i=1}^{l_k} \frac{A_{k,i}}{(x-x_k)^i} \right].$$

Bei einer l_k -fachen Nullstelle x_k wird also in der Partialbruchzerlegung statt des Terms $\frac{A_k}{x-x_k}$ die Summe

$$\frac{A_{k,1}}{(x-x_k)} + \frac{A_{k,2}}{(x-x_k)^2} + \dots + \frac{A_{k,l_k}}{(x-x_k)^{l_k}}$$

verwendet. Die Konstanten $A_{k,i} \in \mathbb{R}$ kann man wieder mittels Koeffizientenvergleich eindeutig bestimmen. Alternativ dazu kann man auch m verschiedene Werte in die Gleichung einsetzen, um ein Gleichungssystem für die Konstanten zu gewinnen. Im Gegensatz zu den einfachen Nullstellen kann man so aber nicht alle Konstanten direkt ablesen. Man muss etwas rechnen, um das Gleichungssystem zu lösen.

Beispiel 2.101

$$\frac{p(x)}{q(x)} = \frac{x^2 + 9x + 19}{(x+4)^2(x+1)}.$$

Nullstellen des Nenners sind -1 und -4 (doppelt), d. h., wir machen den Ansatz

$$\begin{aligned} \frac{x^2 + 9x + 19}{(x+4)^2(x+1)} &= \frac{A}{x+1} + \frac{B}{x+4} + \frac{C}{(x+4)^2} \\ \implies x^2 + 9x + 19 &= A(x+4)^2 + B(x+4)(x+1) + C(x+1). \end{aligned}$$

Setzt man nun die einfache Nullstelle -1 ein, erhält man $11 = 9A$, $A = \frac{11}{9}$. Setzt man die doppelte Nullstelle -4 ein: $-1 = -3C$, $C = 1/3$. Leider kann man den dritten Koeffizienten B nicht durch Einsetzen einer Nullstelle ablesen. Hier erhalten wir z. B. durch Zusammenfassen der Koeffizienten zu

$$x^2 + 9x + 19 = (A+B)x^2 + (8A+5B+C)x + (16A+4B+C)$$

und mit Koeffizientenvergleich für x^2 die Gleichung $A+B=1$, also $B=1-\frac{11}{9}=-\frac{2}{9}$.

$$\frac{x^2 + 9x + 19}{(x+4)^2(x+1)} = \frac{\frac{11}{9}}{x+1} + \frac{-\frac{2}{9}}{x+4} + \frac{\frac{1}{3}}{(x+4)^2}.$$

Eine entsprechende Gleichung kann man auch ohne Ausmultiplikation mit dem Nenner durch Einsetzen eines weiteren Werts gewinnen, der keine Nullstelle ist. Ist z. B. 0 keine Nullstelle, so erhält man durch Einsetzen die Gleichung, die man anderenfalls bei einem Koeffizientenvergleich für x^0 ablesen kann, also hier $19 = 16A + 4B + C$.

Wir hätten auch ohne Einsetzen von Nullstellen aus dem Koeffizientenvergleich ein komplettes Gleichungssystem gewinnen können:

$$\begin{aligned} A + B &= 1 \\ 8A + 5B + C &= 9 \\ 16A + 4B + C &= 19. \end{aligned}$$

■

Im Beispiel haben wir gesehen, dass man nicht alle Koeffizienten durch Einsetzen der Nullstellen erhält. Allerdings geht es mittels Ableiten. Sei dazu $q(x) = q_0(x)(x - x_k)^{l_k}$. Wir multiplizieren

$$\frac{p(x)}{q(x)} = \frac{A_{k,1}}{(x - x_k)} + \frac{A_{k,2}}{(x - x_k)^2} + \cdots + \frac{A_{k,l_k}}{(x - x_k)^{l_k}} + r(x)$$

mit $(x - x_k)^{l_k}$, leiten sukzessive ab und setzen x_k ein:

$$\begin{aligned} \frac{p(x)}{q_0(x)} &= A_{k,l_k} + A_{k,l_k-1}(x - x_k) + \cdots + A_{k,1}(x - x_k)^{l_k-1} + r(x)(x - x_k)^{l_k} \\ \Rightarrow \frac{p(x_k)}{q_0(x_k)} &= A_{k,l_k}, \\ \frac{d}{dx} \frac{p(x)}{q_0(x)} &= A_{k,l_k-1} + 2A_{k,l_k-2}(x - x_k) + \cdots + (l_k - 1)A_{k,1}(x - x_k)^{l_k-2} \\ &\quad + r'(x)(x - x_k)^{l_k} + l_k r(x)(x - x_k)^{l_k-1} \\ \Rightarrow A_{k,l_k-1} &= \frac{1}{1!} \cdot \frac{d}{dx} \frac{p}{q_0}(x_k), \dots, A_{k,l_k-n} = \frac{1}{n!} \cdot \frac{d^n}{dx^n} \frac{p}{q_0}(x_k). \end{aligned}$$

Damit sind auch bei mehrfachen Nullstellen alle Koeffizienten eindeutig bestimmt, und man kann direkt die Partialbruchzerlegung ohne Lösung eines Gleichungssystems berechnen. Wir wenden dies für das vorangehende Beispiel an und erhalten für $x_0 = -4$:

$$\frac{x^2 + 9x + 19}{x + 1} = C + B(x + 4) + \frac{A}{x + 1}(x + 4)^2$$

und

$$\begin{aligned} C &= \frac{x_0^2 + 9x_0 + 19}{x_0 + 1} = \frac{1}{3}, \\ B &= \frac{(2x_0 + 9)(x_0 + 1) - (x_0^2 + 9x_0 + 19)}{(x_0 + 1)^2} = -\frac{2}{9}. \end{aligned}$$

Den Wert für A erhält man direkt durch Einsetzen der einfachen Nullstelle -1 .

Beispiel 2.102

Wir nutzen nun die gewonnene Partialbruchzerlegung zur Integration

$$\int_0^1 \frac{x^2 + 9x + 19}{(x + 4)^2(x + 1)} dx = \frac{11}{9} \int_0^1 \frac{1}{x + 1} dx - \frac{2}{9} \int_0^1 \frac{1}{x + 4} dx + \frac{1}{3} \int_0^1 \frac{1}{(x + 4)^2} dx$$

$$\begin{aligned}
&= \frac{11}{9} [\ln |x+1|]_0^1 - \frac{2}{9} [\ln |x+4|]_0^1 - \frac{1}{3} \left[\frac{1}{x+4} \right]_0^1 \\
&= \frac{11}{9} \ln(2) - \frac{2}{9} \ln\left(\frac{5}{4}\right) - \frac{1}{3} \left[\frac{1}{5} - \frac{1}{4} \right]. \quad \blacksquare
\end{aligned}$$

Hat man mehrfache reelle Nullstellen des Nenners, so führt die Integration mittels Partialbruchzerlegung zu Integralen des Typs ($n > 1$)

$$\int \frac{1}{(x-x_0)^n} dx = \frac{-1}{(n-1)(x-x_0)^{n-1}} + c, \quad n \geq 2.$$

Echt komplexe Nullstellen des Nenners

Man kann mit komplexen Nullstellen (und auch mit komplexen Koeffizienten) genauso rechnen, wie zuvor für reelle Nullstellen dargestellt. Allerdings kann man diese komplexe Partialbruchzerlegung nicht unmittelbar integrieren. Vor der Integration muss man daher durch konjugiert-komplexe Erweiterung wieder für eine reelle Darstellung sorgen. Wer den Umgang mit komplexen Zahlen scheut, kann aber auch von Anfang an mit echt komplexen Nullstellen rein reell rechnen. Denn für Polynome mit reellen Koeffizienten sind ihre komplexen Nullstellen immer konjugiert-komplexe Zahlenpaare, siehe Lemma 1.13 auf Seite 148. Die zugehörigen Partialbruchsummanden können damit paarweise reell zusammengefasst werden.

Lemma 2.11 (Partialbruchzerlegung bei komplexen Nullstellen)

Hat $q(x)$ unter den sonstigen Voraussetzungen von Lemma 2.10 ein Paar einfacher konjugiert-komplexer Nullstellen $x_k \pm jy_k$, d. h., $q(x)$ ist teilbar durch $(x-x_k-jy_k)(x-x_k+jy_k) = (x-x_k)^2 + y_k^2$, so kann dafür in der Partialbruchzerlegung statt der komplexen Summanden $\frac{C_k}{x-(x_k-jy_k)} + \frac{D_k}{x-(x_k+jy_k)}$ der aus der Summe entstehende reelle Summand $(A_k, B_k \in \mathbb{R})$

$$\frac{A_k x + B_k}{(x-x_k)^2 + y_k^2}$$

verwendet werden.

Sind $x_k \pm jy_k$ l_k -fache Nullstellen, so können dafür in der Partialbruchzerlegung statt der (dann komplexen) Terme aus Lemma 2.10 die reellen Summanden $(A_{k,l_k}, B_{k,l_k}, \dots, A_{k,1}, B_{k,1} \in \mathbb{R})$

$$\frac{A_{k,l_k} x + B_{k,l_k}}{((x-x_k)^2 + y_k^2)^{l_k}} + \frac{A_{k,l_k-1} x + B_{k,l_k-1}}{((x-x_k)^2 + y_k^2)^{l_k-1}} + \dots + \frac{A_{k,1} x + B_{k,1}}{(x-x_k)^2 + y_k^2}$$

verwendet werden.

Beispiel 2.103

Wir ermitteln die Partialbruchzerlegung von $\frac{x^3}{(x^2+1)^2}$ einmal mit einer rein reellen Rechnung und einmal unter Verwendung komplexer Koeffizienten. Der Nenner hat die doppelten Nullstellen j und $-j$:

■ Reelle Rechnung:

$$\begin{aligned}\frac{x^3}{(x^2+1)^2} &= \frac{Ax+B}{(x^2+1)^2} + \frac{Cx+D}{x^2+1} \\ \Rightarrow x^3 &= (A+C)x + Dx^2 + Cx^3 + B + D,\end{aligned}$$

so dass $D = 0$, $B = 0$, $C = 1$ und $A = -1$, also

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

■ Verwendung komplexer Koeffizienten:

$$\begin{aligned}\frac{x^3}{(x^2+1)^2} &= \frac{x^3}{(x-j)^2(x+j)^2} = \frac{A}{x-j} + \frac{B}{(x-j)^2} + \frac{C}{x+j} + \frac{D}{(x+j)^2} \\ \Rightarrow x^3 &= A(x-j)(x+j)^2 + B(x+j)^2 + C(x-j)^2(x+j) + D(x-j)^2 \\ &= (A+C)x^3 + (jA+B-jC+D)x^2 + (A+2jB+C-2jD)x \\ &\quad + jA-B-jC-D\end{aligned}$$

Wir erhalten durch Koeffizientenvergleich das folgende komplexe, lineare Gleichungssystem:

$$\begin{array}{ccccccc} A & & +C & & = & 1 \\ jA & +B & -jC & +D & = & 0 \\ A & +2jB & +C & -2jD & = & 0 \\ jA & -B & -jC & -D & = & 0. \end{array}$$

Dieses System hat die eindeutige Lösung $A = C = \frac{1}{2}$, $B = \frac{j}{4}$, $D = -\frac{j}{4}$:

$$\begin{aligned}\frac{x^3}{(x^2+1)^2} &= \frac{x^3}{(x-j)^2(x+j)^2} = \frac{\frac{1}{2}}{x-j} + \frac{\frac{j}{4}}{(x-j)^2} + \frac{\frac{1}{2}}{x+j} - \frac{\frac{j}{4}}{(x+j)^2} \\ &= \frac{\frac{1}{2}(x+j) + \frac{1}{2}(x-j)}{(x-j)(x+j)} + \frac{\frac{j}{4}(x+j)^2 - \frac{j}{4}(x-j)^2}{(x-j)^2(x+j)^2} \\ &= \frac{x}{x^2+1} + \frac{-x}{(x^2+1)^2}.\end{aligned}$$

Mit dieser Partialbruchzerlegung erhalten wir die Stammfunktion

$$\int \frac{x^3}{(x^2+1)^2} dx = \int \frac{x}{x^2+1} dx + \int \frac{-x}{(x^2+1)^2} dx.$$

Substitution $t = x^2$, $dt = 2x dx$ liefert weiter

$$\begin{aligned}\int \frac{x^3}{(x^2+1)^2} dx &= \frac{1}{2} \int \frac{1}{t+1} dt - \frac{1}{2} \int \frac{1}{(t+1)^2} dt \\ &= \frac{1}{2} \ln|t+1| + \frac{1}{2} \frac{1}{t+1} + C = \frac{1}{2} \ln(x^2+1) + \frac{1}{2} \frac{1}{x^2+1} + C,\end{aligned}$$

wobei wir im letzten Schritt die Substitution rückgängig gemacht haben (was nötig ist, da wir ohne Grenzen gearbeitet haben). ■

Bei komplexen Nullstellen benötigt man zusätzlich die Umkehrfunktionen der trigonometrischen Funktionen als Stammfunktionen (siehe Tabelle 2.3 auf Seite 344).

Achtung: Stammfunktionen sollten erst berechnet werden, wenn die Partialbruchdarstellung reell ist. Beispielsweise haben wir $\ln(x - j)$ nicht definiert. $x - j$ ist eine komplexe Zahl, aber den Logarithmus haben wir nur für positive reelle Zahlen erklärt. Andererseits ist aber $\ln|x - j|$ für die reelle Zahl $|x - j|$ erklärt, jedoch ist

$$\frac{d}{dx} \ln|x - j| = \frac{d}{dx} \ln(\sqrt{x^2 + 1}) = \frac{1}{\sqrt{x^2 + 1}} \frac{2x}{2\sqrt{x^2 + 1}} = \frac{x}{x^2 + 1} = \frac{x}{|x - j|^2} \neq \frac{1}{|x - j|}.$$

2.6.5 Numerische Integration

Viele einfach aussehende Funktionen wie beispielsweise $f(x) = e^{-(x^2)}$ oder $f(x) = \frac{\sin(x)}{x} = \text{sinc}(x)$ lassen sich nicht elementar integrieren, d. h., man findet keine elementare Stammfunktion. Außerdem liegen häufig nur Funktionswerte als Mess- oder Abtastwerte an bestimmten Stellen vor. In diesen Fällen ist man auf numerische Methoden angewiesen, bei denen der Wert des Integrals nur näherungsweise berechnet wird. Dafür sind diese Verfahren leicht als Computeralgorithmen zu implementieren.

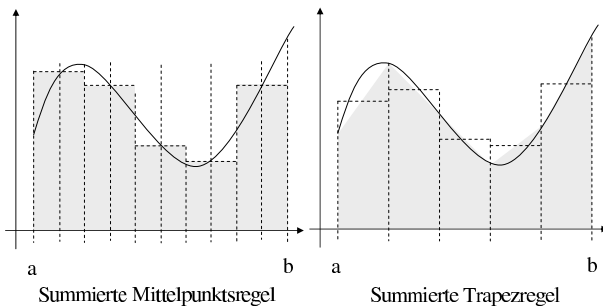


Abb. 2.31: Quadraturformeln

Wir approximieren das exakte Integral einer stetigen Funktion f auf dem Intervall $[a, b]$ durch eine Näherung Q . Ein Algorithmus zur Berechnung der Zahl Q heißt eine **Quadraturformel**. Dabei tritt ein Fehler

$$E := \int_a^b f(x) dx - Q$$

auf, dessen Betrag möglichst klein sein soll.

Eine **interpolatorische Quadraturformel** erhält man, indem man das Intervall $[a, b]$ in n äquidistante Teilintervalle der Länge $h = \frac{b-a}{n}$ mit den **Stützstellen**

$$x_k = a + k \cdot h, \quad k = 0, 1, \dots, n,$$

zerlegt. Dabei stimmen die Randpunkte mit den Intervallgrenzen $a = x_0$ und $b = x_n$ überein. Die Funktionswerte zu den Stützstellen nennen wir $f_k := f(x_k)$, $k = 0, 1, \dots, n$. Zu einer Quadraturformel gelangen wir nun, indem wir f in jedem einzelnen Teilintervall $[x_k, x_{k+1}]$ durch ein **Interpolationspolynom** niedrigen Grades m ersetzen (siehe Seite 92). Das Interpolationspolynom wird auf $[x_k, x_{k+1}]$ exakt integriert, so dass wir einen Näherungswert für $\int_{x_k}^{x_{k+1}} f(x) dx$ bekommen. Die Abweichung des Näherungswerts vom exakten Wert lässt sich dann pro Teilintervall z. B. mit Satz 2.39 auf Seite 321 grob abschätzen.

Die einfachste Quadraturformel entsteht, wenn man die Funktion $f(x)$ in jedem Intervall $[x_k, x_{k+1}]$ durch ein Interpolationspolynom vom Grad null, also durch eine konstante Funktion, z. B. $f_k = f(x_k)$, $k = 0, 1, \dots, n-1$, oder durch $f(x_k + \frac{h}{2})$, wie in Abbildung 2.31 links angedeutet, ersetzt. Durch Summation der Integrale über alle Teilintervalle erhält man die **summierte Rechteckregel**

$$Q_R(h) := \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f_k dx = h \cdot \sum_{k=0}^{n-1} f_k$$

oder die **summierte Mittelpunktsregel**

$$Q_M(h) := \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f\left(x_k + \frac{h}{2}\right) dx = h \cdot \sum_{k=0}^{n-1} f\left(x_k + \frac{h}{2}\right).$$

Die Werte beider Regeln konvergieren für $n \rightarrow \infty$ (also $h \rightarrow 0$) gegen das gesuchte Integral, da es sich um Riemann-Zwischensummen handelt (siehe Satz 2.42 auf Seite 333).

Eine im Allgemeinen genauere Integrationsregel erhält man, indem man die Funktion f im k -ten Teilintervall $[x_k, x_{k+1}]$ linear (also mit einem Polynom vom Grad eins) interpoliert, d. h. die Funktion f durch die Gerade durch die Punkte (x_k, f_k) und (x_{k+1}, f_{k+1}) ersetzt:

$$g(x) = f_k + \frac{f_{k+1} - f_k}{h} \cdot (x - x_k).$$

Diese Gerade wird dann über $[x_k, x_{k+1}]$ exakt integriert:

$$\int_{x_k}^{x_{k+1}} g(x) dx = \frac{x_{k+1} - x_k}{2} \cdot (f_k + f_{k+1}) = \frac{h}{2} \cdot (f_k + f_{k+1}).$$

Summation über alle Teilintervalle ergibt jetzt die **summierte Trapezregel**

$$Q_T(h) := \frac{h}{2} \cdot (f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n),$$

vgl. Abbildung 2.31 rechts.

Da f stetig ist, werden aufgrund des Zwischenwertsatzes (siehe Satz 2.25 auf Seite 278) auf jedem Teilintervall alle Werte zwischen f_k und f_{k+1} angenommen. Insbesondere gibt es also eine Stelle ξ_k in diesem Intervall, an der der Funktionswert genau das arithmetische Mittel $\frac{1}{2} \cdot (f_k + f_{k+1})$ ist. Damit kann auch Q_T als Riemann-Zwischensumme aufgefasst werden, und wir haben Konvergenz gegen das Integral.

Beispiel 2.104

Wir berechnen das Integral $\int_0^\pi \sin(x) dx = 2$ mit der summierten Trapezregel und der Schrittweite $h = 0,2 \cdot \pi$:

$$\begin{aligned} Q_T(0,2) &\approx \frac{1}{2} \cdot 0,2 \cdot \pi (0 + 2 \cdot 0,5878 + 2 \cdot 0,9510 + 2 \cdot 0,9510 + 2 \cdot 0,5879 + 0) \\ &\approx 1,9338. \end{aligned}$$

Mit der Schrittweite $h = 0,1 \cdot \pi$ bekommt man das genauere Ergebnis 1,9835. ■

Ist die Funktion f zweimal differenzierbar auf $[a, b]$, so kann man den Fehler $E_T(h) := |\int_a^b f(t) dt - Q_T(h)|$ nach oben abschätzen gegen $C \cdot h^2$, wobei die Konstante C von der Funktion f und dem Intervall $[a, b]$ abhängt, aber nicht von der Schrittweite h . Denn nach Satz 2.39 können wir den Abstand zwischen Interpolationspolynom vom Grad 1 und Funktion f auf jedem Teilintervall der Länge h gegen $C \cdot h^2$ abschätzen, so dass diese Größenordnung nach Integration über $[a, b]$ erhalten bleibt. Das geht besser:

Bei der **summierten Simpson-Regel** setzen wir voraus, dass die Anzahl n der Teilintervalle eine gerade Zahl ist. Dann können wir Zerlegungsintervalle paarweise zusammenfassen: $[x_k, x_{k+1}] \cup [x_{k+1}, x_{k+2}]$. Auf jedem so zusammengefassten Intervall interpolieren wir f durch ein Polynom zweiten Grades durch die drei Punkte (x_k, f_k) , (x_{k+1}, f_{k+1}) und (x_{k+2}, f_{k+2}) :

$$p(x) = f_k + \frac{f_{k+1} - f_k}{h} \cdot (x - x_k) + \frac{f_{k+2} - 2f_{k+1} + f_k}{2h^2} \cdot (x - x_k) \cdot (x - x_{k+1}).$$

Das Integral über das Näherungspolynom $p(x)$ im zusammengefassten Intervall $[x_k, x_{k+2}]$ lässt sich ohne Tricks aber mit etwas Schreibarbeit ausrechnen zur Simpson-Regel

$$\int_{x_k}^{x_{k+2}} p(x) dx = \frac{1}{3} h \cdot (f_k + 4f_{k+1} + f_{k+2}).$$

Summation über alle zusammengefassten Intervalle ergibt die summierte Simpson-Regel

$$Q_S(h) := \frac{4}{3} h \cdot (f_1 + f_3 + \cdots + f_{n-1}) + \frac{2}{3} h \cdot (f_2 + f_4 + \cdots + f_{n-2}) + \frac{1}{3} h \cdot (f_0 + f_n).$$

Ist f viermal differenzierbar auf $[a, b]$, so kann der Fehler $E_S(h) := |\int_a^b f(t) dt - Q_S(h)|$ der summierten Simpson-Regel sogar gegen $C \cdot h^4$ abgeschätzt werden. Mit Satz 2.39 erhält man direkt allerdings nur $C \cdot h^3$. Zum Beweis der besseren Abschätzung gegen h^4 muss ein zusätzlicher Auslöschungseffekt berücksichtigt werden, hier verweisen wir auf Numerik-Lehrbücher.

Quadraturformeln, die über die exakte Integration von Lagrange-Interpolationspolynomen wie die Trapezregel (Grad 1) und Simpson-Regel (Grad 2) definiert sind, werden in der Literatur als **Newton-Cotes-Formeln** bezeichnet.

Beispiel 2.105

Wir berechnen das Integral $\int_1^3 \frac{1}{x} dx = \ln 3 \approx 1,09861$ mit $n = 2$, d. h. Schrittweite $h = 1$ und Funktionswerte $f_0 = 1$, $f_1 = \frac{1}{2}$, $f_2 = \frac{1}{3}$:

- Summierte Trapezregel:

$$Q_T(1) = \frac{1}{2} \left(1 + 2 \cdot \frac{1}{2} + \frac{1}{3} \right) = \frac{7}{6} = 1,1\bar{6}.$$

- Summierte-Regel (mit nur einem zusammengefassten Intervall):

$$Q_S(1) = \frac{4}{3} \cdot \frac{1}{2} + \frac{1}{3} \left[1 + \frac{1}{3} \right] = \frac{2}{3} + \frac{4}{9} = \frac{10}{9} = 1,1\bar{1}. \quad \blacksquare$$

2.6.6 Uneigentliche Integrale

Das Riemann-Integral ist nur für beschränkte Funktionen f auf beschränkten Intervallen $[a, b]$ erklärt, da man nur unter diesen Voraussetzungen Ober- und Untersummen hinschreiben kann. Häufig (z. B. bei der Fourier- und Laplace-Transformation oder bei Dichten in der Wahrscheinlichkeitsrechnung) sieht man aber Ausdrücke der Form $\int_a^\infty f(x) dx$, $\int_{-\infty}^b f(x) dx$ und $\int_{-\infty}^\infty f(x) dx$ bzw. $\int_a^b f(x) dx$ für eine auf $[a, b]$ unbeschränkte Funktion f . Diese Zahlen sind über Grenzwerte definiert. Man spricht von **uneigentlichen Integralen**.

Definition 2.35 (Uneigentliches Integral)

- Es sei $f(x)$ eine Funktion mit einem uneigentlichen Grenzwert bei b , d. h. $\lim_{x \rightarrow b-} f(x) = \pm\infty$. Weiter sei f auf jedem Intervall $[a, t] \subset [a, b[$ integrierbar. Existiert der einseitige Grenzwert $\lim_{t \rightarrow b-} \int_a^t f(x) dx$ als reelle Zahl, so nennt man f bezüglich $[a, b]$ **uneigentlich integrierbar**, und

$$\int_a^b f(x) dx := \lim_{t \rightarrow b-} \int_a^t f(x) dx$$

heißt **uneigentliches Integral**. Falls der Grenzwert nicht existiert, spricht man von einem **divergenten uneigentlichen Integral**.

Entsprechend verwendet man einen rechtsseitigen Grenzwert, wenn f an der Stelle a einen uneigentlichen Grenzwert $\lim_{x \rightarrow a+} f(x) = \pm\infty$ hat und auf jedem Intervall $[t, b] \subset]a, b]$ integrierbar ist.

Sind Grenzwerte im Inneren des Intervalls gleich $+\infty$ oder $-\infty$, dann ist das Integral so in eine Summe einzelner Integrale aufzutrennen, dass stets nur ein uneigentlicher einseitiger Grenzwert an den Rändern vorliegt.

- Sei der Integrationsbereich $[a, \infty[$ oder $]-\infty, b]$ und die Funktion f auf jedem beschränkten Teilintervall $[c, d]$ integrierbar. Falls die Grenzwerte

$$\int_a^\infty f(x) dx := \lim_{b \rightarrow \infty} \int_a^b f(x) dx \quad \text{bzw.} \quad \int_{-\infty}^b f(x) dx := \lim_{a \rightarrow -\infty} \int_a^b f(x) dx$$

existieren, so spricht man ebenfalls von einem **uneigentlichen Integral** und sagt, dass f **uneigentlich integrierbar** ist. Falls der Grenzwert nicht existiert, so verwendet man wieder den Begriff **divergentes uneigentliches Integral**.
 Ein Integral von $-\infty$ bis ∞ ist an einer beliebigen Stelle in zwei uneigentliche Integrale und damit in die Summe zweier Grenzwerte aufzutrennen.

Achtung: Wie wir in den beiden folgenden Abschnitten sehen werden, gibt es Flächen, die sich ins Unendliche erstrecken, aber dennoch einen endlichen Flächeninhalt besitzen, siehe auch Abbildung 2.32.

Die Begriffe „uneigentliches Integral“ und „unbestimmtes Integral“ klingen sehr ähnlich, sind aber verschieden. Ein unbestimmtes Integral ist eine Stammfunktion. Es ist unbestimmt, da es keine festen Integrationsgrenzen hat. Bei einem uneigentlichen Integral sind die Voraussetzungen der Definition des Integrals nicht erfüllt, und man muss sich mit einem zusätzlichen Grenzwert behelfen.

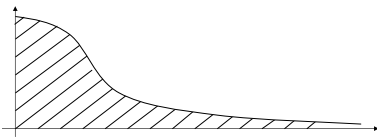


Abb. 2.32: Für $x \rightarrow \infty$ unendlich lange Fläche mit endlichem Inhalt: konvergentes uneigentliches Integral

2.6.6.1 Unbeschränkter Integrand

Wir betrachten zunächst den Fall, dass der Integrand f an einem Rand des Integrationsintervalls einen uneigentlichen Grenzwert besitzt.

Beispiel 2.106

- a) Das Integral $\int_0^1 \frac{1}{\sqrt{1-x^2}} dx$ ist uneigentlich bezüglich der oberen Grenze 1: $\lim_{x \rightarrow 1-} \frac{1}{\sqrt{1-x^2}} = \infty$. Wegen

$$\int_0^t \frac{dx}{\sqrt{1-x^2}} = [\arcsin x]_0^t = \arcsin(t)$$

erhalten wir

$$\lim_{t \rightarrow 1-} \int_0^t \frac{1}{\sqrt{1-x^2}} dx = \lim_{t \rightarrow 1-} \arcsin(t) = \frac{\pi}{2}.$$

- b) Das Integral $\int_0^1 \frac{1}{x} dx$ ist uneigentlich bezüglich der unteren Grenze 0, da $\lim_{x \rightarrow 0+} \frac{1}{x} = \infty$. Zwar existiert

$$\int_t^1 \frac{1}{x} dx = [\ln x]_t^1 = -\ln(t),$$

jedoch ist

$$\lim_{t \rightarrow 0+} -\ln(t) = \infty,$$

und das uneigentliche Integral ist divergent. Anschaulich bedeutet dies, dass der Inhalt der Fläche unter dem Funktionsgraphen unendlich groß ist. ■

2.6.6.2 Unbeschränkter Integrationsbereich

Beispiel 2.107

a) Die Fläche zwischen dem Graphen von e^{-x} und der positiven x -Achse ist:

$$\begin{aligned} \int_0^{\infty} \exp(-x) dx &= \lim_{t \rightarrow \infty} \int_0^t \exp(-x) dx = \lim_{t \rightarrow \infty} [-\exp(-x)]_0^t \\ &= \lim_{t \rightarrow \infty} [-\exp(-t) + 1] = 1. \end{aligned}$$

b) 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 unter a) gezeigt haben. Unter der Annahme, dass $\Gamma(n) = (n-1)!$ ist, müssen wir nun nur noch zeigen, dass $\Gamma(n+1) = n!$ ist. Das gelingt mittels partieller Integration:

$$\begin{aligned} \Gamma(n+1) &= \int_0^{\infty} x^n e^{-x} dx = \lim_{t \rightarrow \infty} \int_0^t x^n e^{-x} dx \\ &= \lim_{t \rightarrow \infty} [-x^n e^{-x}]_0^t + \lim_{t \rightarrow \infty} \int_0^t n x^{n-1} e^{-x} dx \\ &= 0 - 0 + n \int_0^{\infty} x^{n-1} e^{-x} dx = n \cdot \Gamma(n) = n \cdot (n-1)! = n!. \end{aligned}$$

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. ■

Etwas schwieriger ist die Definition des Integrals $\int_{-\infty}^{\infty} f(x) dx$. Laut Definition 2.35 sollen wir hier das Integral an einer beliebigen Stelle $x_0 \in \mathbb{R}$ auftrennen:

$$\int_{-\infty}^{\infty} f(x) dx := \int_{-\infty}^{x_0} f(x) dx + \int_{x_0}^{\infty} f(x) dx.$$

Dabei müssen beide Einzelintegrale existieren, wobei auch Werte $+\infty$ und $-\infty$ zulässig sind. Jedoch dürfen nicht $+\infty$ und $-\infty$ gemeinsam auftreten. Aufgrund der Additivität (2.47), siehe Seite 336, zeigt man, dass diese Definition unabhängig vom konkreten Punkt x_0 ist: Ist z. B. $x_1 > x_0$, so erhalten wir (falls die Integrale für die Grenze x_0 oder alternativ für die Grenze x_1 konvergieren):

$$\begin{aligned} & \lim_{t \rightarrow \infty} \int_{-t}^{x_0} f(x) dx + \lim_{t \rightarrow \infty} \int_{x_0}^t f(x) dx \\ &= \lim_{t \rightarrow \infty} \int_{-t}^{x_0} f(x) dx + \int_{x_0}^{x_1} f(x) dx - \int_{x_0}^{x_1} f(x) dx + \lim_{t \rightarrow \infty} \int_{x_0}^t f(x) dx \\ &= \lim_{t \rightarrow \infty} \left[\int_{-t}^{x_0} f(x) dx + \int_{x_0}^{x_1} f(x) dx \right] + \lim_{t \rightarrow \infty} \left[\int_{x_1}^{x_0} f(x) dx + \int_{x_0}^t f(x) dx \right] \\ &= \lim_{t \rightarrow \infty} \int_{-t}^{x_1} f(x) dx + \lim_{t \rightarrow \infty} \int_{x_1}^t f(x) dx. \end{aligned}$$

Beispiel 2.108

Wir berechnen die rechts und links unbeschränkte Fläche unter dem Funktionsgraphen von $f(x) = \frac{1}{1+x^2}$:

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx &= 2 \lim_{b \rightarrow \infty} \int_0^b \frac{1}{1+x^2} dx = 2 \lim_{b \rightarrow \infty} [\arctan(x)]_0^b \\ &= 2 \lim_{b \rightarrow \infty} \arctan(b) = 2 \cdot \frac{\pi}{2} = \pi. \end{aligned}$$

Achtung: Ein völlig anderes Ergebnis als $\int_{-\infty}^{\infty} f(x) dx$ kann der symmetrischer Grenzwert $\lim_{t \rightarrow \infty} \int_{-t}^t f(x) dx$ liefern, wie das folgende Beispiel zeigt:

Beispiel 2.109

$\int_{-\infty}^{\infty} \sin x dx$ existiert nicht, aber $\int_{-t}^t \sin x dx = 0$ und damit $\lim_{t \rightarrow \infty} \int_{-t}^t \sin x dx = 0$.

Integrale über einen unbegrenzten Integrationsbereich sind eng verwandt mit Zahlenreihen. Wir können das Integral $\int_0^{\infty} \exp(-x) dx$ z. B. über die unendliche Obersumme

$$\sum_{k=0}^{\infty} \frac{1}{n} \exp\left(-\frac{k}{n}\right) = \frac{1}{n} \sum_{k=0}^{\infty} \left[e^{-\frac{1}{n}}\right]^k = \frac{1}{n} \frac{1}{1 - e^{-\frac{1}{n}}}$$

annähern, die eine geometrische Reihe ist (siehe (2.11) auf Seite 237). Dabei haben alle Zerlegungsintervalle von $[0, \infty[$ die Länge $\frac{1}{n}$, und wir verwenden jeweils den größten Funktionswert, der am linken Intervallrand liegt. Nun ist unter Verwendung des Satzes von L'Hospital (Satz 2.40 auf Seite 323 für den Typ $\left[\frac{0}{0}\right]$):

$$\lim_{n \rightarrow \infty} \sum_{k=0}^{\infty} \frac{1}{n} \exp\left(-\frac{k}{n}\right) = \lim_{n \rightarrow \infty} \frac{\frac{1}{n}}{1 - e^{-\frac{1}{n}}} = \lim_{n \rightarrow \infty} \frac{-\frac{1}{n^2}}{-\frac{1}{n^2} e^{-\frac{1}{n}}}$$

$$= \lim_{n \rightarrow \infty} \exp\left(\frac{1}{n}\right) = 1 = \int_0^\infty \exp(-x) dx.$$

Es ist also nicht verwunderlich, dass sich die Konvergenzkriterien für Reihen auf Integrale mit unbeschränktem Integrationsbereich übertragen lassen. Besondere Bedeutung hat dabei das Majoranten-Kriterium:

Satz 2.51 (Vergleichs- oder Majoranten-Kriterium für Integrale)

Sei $f : [a, \infty[\rightarrow \mathbb{R}$, wobei f auf jedem Intervall $[a, b] \subset [a, \infty[$ integrierbar sei. Ist $g : [a, \infty[\rightarrow \mathbb{R}$, so dass das Integral $\int_a^\infty g(x) dx$ existiert, und ist $|f(x)| \leq g(x)$, dann existiert auch $\int_a^\infty f(x) dx$, und es ist

$$\left| \int_a^\infty f(x) dx \right| \leq \int_a^\infty |f(x)| dx \leq \int_a^\infty g(x) dx. \quad (2.58)$$

Dieses Kriterium wird noch wichtig werden, wenn wir (z. B. für die Regelungstechnik) die Laplace-Transformation in Kapitel 6.3 kennenlernen. Diese Transformation basiert auf dem Ausrechnen eines Integrals über $[0, \infty[$, und man muss sicherstellen, dass dieses Integral existiert. Das macht man mit einer Wachstumsbedingung unter Anwendung des Satzes.

Der Satz entspricht genau der Aussage für unendliche Summen in Satz 2.15 auf Seite 246 und wird auch entsprechend über das Cauchy-Kriterium bewiesen:

Beweis: Die Majorantenbedingung sorgt dafür, dass eine Cauchy-Bedingung erfüllt ist. Diese haben wir nur für Folgen definiert, so dass wir die Existenz des Grenzwerts über das Übertragungsprinzip nachrechnen müssen.

Sei $\varepsilon > 0$. Da nach Voraussetzung $\lim_{t \rightarrow \infty} \int_a^t g(x) dx$ existiert, gibt es zu ε eine Stelle T_0 , so dass für alle $u \geq t > T_0$ gilt:

$$\left| \int_t^u f(x) dx \right| \leq \int_t^u |f(x)| dx \leq \int_t^u g(x) dx \leq \int_a^\infty g(x) dx - \int_a^t g(x) dx < \varepsilon. \quad (2.59)$$

Sei $(t_n)_{n=1}^\infty$ mit $t_n > a$ eine Folge mit $\lim_{n \rightarrow \infty} t_n = \infty$. Dann ist $\left(\int_a^{t_n} f(x) dx \right)_{n=1}^\infty$ eine Cauchy-Folge (siehe Definition 2.8 auf Seite 235): Wegen $\lim_{n \rightarrow \infty} t_n = \infty$ gibt es ein $n_0 \in \mathbb{N}$, so dass für $n, m > n_0$: $t_n, t_m > T_0$ und $(t_n < t_m)$:

$$\left| \int_a^{t_m} f(x) dx - \int_a^{t_n} f(x) dx \right| = \left| \int_{t_n}^{t_m} f(x) dx \right| \stackrel{(2.59)}{<} \varepsilon.$$

Als Cauchy-Folge ist sie konvergent (siehe Satz 2.8) gegen einen Grenzwert G , der jetzt aber noch von der konkreten Wahl der Folge abhängen könnte. Sei $(x_n)_{n=1}^\infty$ eine weitere Folge mit $x_n > a$ und $\lim_{n \rightarrow \infty} x_n = \infty$, so dass $\left(\int_a^{x_n} f(x) dx \right)_{n=1}^\infty$ (als Cauchy-Folge) gegen einen Grenzwert H strebt.

Zum vorgegebenen $\varepsilon > 0$ gibt es ein (gemeinsames) $n_0 \in \mathbb{R}$, so dass für $n > n_0$:

$$\left| \int_a^{t_n} f(x) dx - G \right| < \varepsilon, \quad \left| \int_a^{x_n} f(x) dx - H \right| < \varepsilon, \quad t_n, x_n > T_0. \quad (2.60)$$

Damit ist

$$\begin{aligned} |G - H| &= \left| G - \int_a^{t_n} f(x) dx - H + \int_a^{x_n} f(x) dx - \int_{t_n}^{x_n} f(x) dx \right| \\ &\leq \left| G - \int_a^{t_n} f(x) dx \right| + \left| H - \int_a^{x_n} f(x) dx \right| + \left| \int_{t_n}^{x_n} f(x) dx \right| \stackrel{(2.59), (2.60)}{<} 3\varepsilon. \end{aligned}$$

Da ε beliebig klein gewählt werden kann, ist der Grenzwert für jede Folge gleich, und schließlich gilt mit dem Übertragungsprinzip (Satz 2.22, Seite 266) $\lim_{t \rightarrow \infty} \int_a^t f(x) dx = G$. Die Abschätzung (2.58) der Grenzwerte folgt aus (2.44) und Folgerung 2.3 von Seite 269. $\square$

Beispiel 2.110

Wegen $|\sin(x) \exp(-x)| \leq \exp(-x)$ existiert $\int_0^\infty \sin(x) \exp(-x) dx$. $\blacksquare$

Das Majoranten-Kriterium funktioniert auch in entsprechend modifizierter Form für uneigentliche Integrale von unbeschränkten Funktionen auf beschränkten Intervallen.

Beispiel 2.111

Wir zeigen, dass das aus Beispiel 2.107 bekannte Integral $\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$ für jeden festen Parameterwert $x \in]0, \infty[$ existiert. Das Integral liefert den Wert der Gammafunktion an der Stelle x . 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 $t \rightarrow 0+$ 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 $|t^{x-1} e^{-t}| \leq t^{x-1} e^0 = t^{x-1}$ und

$$\lim_{u \rightarrow 0+} \int_u^1 t^{x-1} dt = \frac{1}{x} \lim_{u \rightarrow 0+} [t^x]_{t=u}^{t=1} = \frac{1}{x}$$

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 $|t^{x-1} e^{-t}| \leq 1^{x-1} e^{-t} = e^{-t}$, wobei damit auch hier eine integrierbare Majorante gefunden ist.

- Für $x > 1$ müssen wir uns wegen $\lim_{t \rightarrow \infty} t^{x-1} = \infty$ etwas anstrengen, um eine integrierbare Majorante zu finden:

$$|t^{x-1}e^{-t}| = \underbrace{\frac{t^{x-1}}{\exp\left(\frac{t}{2}\right)}}_{=:g(t)} \exp\left(-\frac{t}{2}\right) \leq C \exp\left(-\frac{t}{2}\right).$$

Dazu müssen wir noch zeigen, dass der Faktor $g(t)$ mit einer Konstante C beschränkt ist. Zunächst gibt es zu $x \in]1, \infty[$ eine Zahl $n \in \mathbb{N}$ mit $n \leq x < n+1$. Mittels n -maliger Anwendung des Satzes von L'Hospital für den Fall $[\infty]$ erhalten wir

$$\lim_{t \rightarrow \infty} g(t) = \lim_{t \rightarrow \infty} \frac{t^{x-1}}{\exp\left(\frac{t}{2}\right)} = \lim_{t \rightarrow \infty} \frac{(x-1)(x-2)(x-3) \cdots (x-n)t^{x-1-n}}{\frac{1}{2^n} \exp\left(\frac{t}{2}\right)} = 0.$$

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 Grenzwerts gibt es eine Stelle t_0 , so dass für $t > t_0$ gilt: $0 \leq g(t) \leq 1$. 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 $\int_1^\infty \exp\left(-\frac{t}{2}\right) dt = 2 \exp\left(-\frac{1}{2}\right) = \frac{2}{\sqrt{e}}$. ■

Wir haben bereits Integrale über unendliche Reihen berechnet. Umgekehrt kann man die Konvergenz unendlicher Reihen mit der Existenz von uneigentlichen Integralen zeigen, die eine Majorante der Reihe sind, z.B. für $\alpha > 1$ so (vgl. Beispiel 2.26 auf Seite 247):

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

Die monoton wachsende Reihe ist daher beschränkt und somit konvergent mit

$$\sum_{k=2}^\infty \frac{1}{k^\alpha} \leq \int_1^\infty \frac{1}{x^\alpha} dx = \frac{1}{\alpha-1}. \quad (2.61)$$

Über einen Vergleich mit Integralen lässt sich auch die Divergenz von Reihen nachweisen, z.B. ist die harmonische Reihe (Fall $\alpha = 1$) nicht beschränkt wegen

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

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**

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n. \quad (2.62)$$

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

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

wobei der zusätzliche Korrekturfaktor für $n \rightarrow \infty$ 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:

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

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

Ein Einsetzen in die Exponentialfunktion liefert

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

und andererseits

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

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

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

2.6.7 Volumen und Flächen

2.6.7.1 Flächenberechnung in der Ebene

Wir haben bereits bei der Einführung des Integrals gesehen, dass sich dessen Wert als Summe bzw. Differenz der Flächen zwischen Funktionsgraph und x -Achse ergibt. Dabei werden die Flächeninhalte oberhalb der x -Achse addiert und die unterhalb der x -Achse subtrahiert (siehe Abbildung 2.30 auf Seite 330).

Insbesondere gilt für integrierbare ungerade Funktionen f , d. h., $f(-x) = -f(x)$, siehe Definition 1.26 auf Seite 81,

$$\int_{-a}^a f(x) dx = 0$$

und für gerade Funktionen f (also $f(x) = f(-x)$)

$$\int_{-a}^a f(x) dt = 2 \int_0^a f(x) dx.$$

Bei einer ungeraden Funktion wie $\sin(x)$ heben sich die Flächen links und rechts vom Nullpunkt auf. Für eine gerade Funktion wie $\cos(x)$ sind die Flächen rechts und links vom Nullpunkt gleich groß.

Zur Bestimmung des Betrags der Gesamtfläche zwischen dem Funktionsgraphen von $f(x)$ und der x -Achse ist das Integral an den Nullstellen von $f(x)$ aufzuspalten. Die Beträge der Teilintegrale sind zu addieren.

Beispiel 2.112

- a) Der Flächeninhalt zwischen der Sinus-Kurve und der x -Achse auf dem Intervall $[0, 2\pi]$ ergibt sich aus den zwei Integralen $I_1 = \int_0^\pi \sin x dx = 2$ und $I_2 = \int_\pi^{2\pi} \sin x dx = -2$ zu $|I_1| + |I_2| = 4$. Man beachte, dass dagegen $\int_0^{2\pi} \sin x dx = 0$ ist.
- b) Der Inhalt der Fläche zwischen x^5 und x -Achse für $-1 \leq x \leq 1$ ist

$$\int_{-1}^1 |x^5| dx = 2 \int_0^1 x^5 dx = 2 \cdot \left[\frac{1}{6} x^6 \right]_0^1 = \frac{2}{6} = \frac{1}{3}.$$

Dagegen ist das Integral der ungeraden Funktion 0. ■

Ganz allgemein ist der Flächeninhalt, den zwei Funktionsgraphen von integrierbaren Funktionen f und g auf einem Intervall $[a, b]$ einschließen, die Zahl

$$\int_a^b |f(x) - g(x)| dx.$$

Da der Betrag über eine Fallunterscheidung definiert ist, benötigt man die Nullstellen der Funktion $f(x) - g(x)$, um den Betrag aufzulösen.

Beispiel 2.113

- a) Wir berechnen den Inhalt der Fläche zwischen den Graphen zu $f(x) = x$ und $g(x) = x^2$ zwischen den Schnittstellen 0 und 1:

$$\int_0^1 |x - x^2| dx = \int_0^1 x - x^2 dx = \left[\frac{1}{2} x^2 - \frac{1}{3} x^3 \right]_0^1 = \frac{1}{6}.$$

- b) Die Graphen zu $f(x) = \sin x$ und $g(x) = \cos x$ schneiden sich im Intervall $[0, 2\pi]$ bei $x_1 = \frac{\pi}{4}$ und $x_2 = \frac{5\pi}{4}$, und der zwischen 0 und 2π eingeschlossene Flächeninhalt ist

$$\begin{aligned}
 & \int_0^{2\pi} |\sin(x) - \cos(x)| dx \\
 &= \int_0^{\pi/4} \cos(x) - \sin(x) dx + \int_{\pi/4}^{5\pi/4} \sin(x) - \cos(x) dx + \int_{5\pi/4}^{2\pi} \cos(x) - \sin(x) dx \\
 &= [\sin(x) + \cos(x)]_0^{\pi/4} + [-\cos(x) - \sin(x)]_{\pi/4}^{5\pi/4} + [\sin(x) + \cos(x)]_{5\pi/4}^{2\pi} \\
 &= \sqrt{2} - 1 + \sqrt{2} + \sqrt{2} + 1 + \sqrt{2} = 4\sqrt{2}.
 \end{aligned}$$

■

2.6.7.2 Volumen eines Rotationskörpers

Lässt man einen Funktionsgraphen um die x -Achse rotieren, so entsteht ein dreidimensionales Objekt – ein Rotationskörper (siehe Abbildung 2.33). Im Alltag kennen wir Rotationskörper als Töpferwaren oder als auf einer Drehbank gedrechselte Werkstücke.

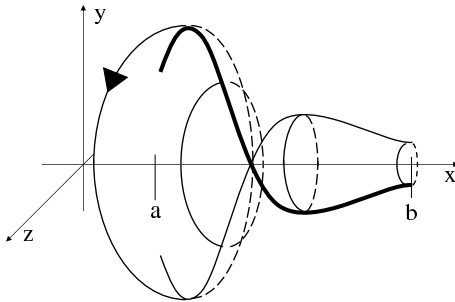


Abb. 2.33: Rotationskörper

In diesem Abschnitt wollen wir das Volumen und im nächsten den Oberflächeninhalt eines Rotationskörpers mittels Integration bestimmen. Dies ist ein Spezialfall der in Kapitel 4.4 beschriebenen Übertragung der Integralrechnung auf Funktionen mit mehreren Variablen, mit der man auch das Volumen komplizierterer Körper berechnen kann.

Wir leiten eine Formel her, mit der das Volumen V eines Rotationskörpers, der durch einen Funktionsgraph auf einem Intervall $[a, b]$ entsteht, berechnet werden kann. Dazu zerlegen wir $[a, b]$ in n Teilintervalle der Breite $\frac{b-a}{n}$ und nähern den Rotationskörper auf jedem Teilintervall durch ein Zylinderstück mit Höhe $\frac{b-a}{n}$ und einem Radius an, der dem Funktionswert am linken Rand des Teilintervalls entspricht. Das Volumen des

Zylinders für das k -te Teilintervall ist $\pi \cdot f\left(a + k \cdot \frac{b-a}{n}\right)^2 \cdot \frac{b-a}{n}$. Damit erhalten wir das gesuchte Volumen als Grenzwert

$$\lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} \pi \cdot f\left(a + k \cdot \frac{b-a}{n}\right)^2 \cdot \frac{b-a}{n} = \pi \int_a^b f(x)^2 dx,$$

sofern die Funktion f^2 integrierbar ist. Denn dann wird der Grenzwert von Riemann-Zwischensummen zu immer feineren Zerlegungen berechnet – und der ist nach dem Riemann'schen Integrabilitätskriterium (siehe Seite 333) gleich dem angegebenen Integral.

Existiert eine Umkehrfunktion f^{-1} von f , so erhält man bei Rotation des Graphen von f um die y -Achse bezüglich des Intervalls $[c, d]$ der y -Achse, das im Wertebereich von f liegt, das Volumen V des Rotationskörpers zu

$$V = \pi \int_c^d f^{-1}(y)^2 dy.$$

Beispiel 2.114

a) Wir berechnen das Volumen von Rotationskörpern zur Funktion $f(x) = x^2 + 3$ für $0 \leq x \leq 1$ (Rotation um x -Achse) bzw. $3 \leq y \leq 4$ (Rotation um y -Achse).

– Bei Drehung um die x -Achse erhalten wir das Volumen:

$$V = \pi \int_0^1 (x^2 + 3)^2 dx = \pi \int_0^1 x^4 + 6x^2 + 9 dx = \pi \left[\frac{1}{5}x^5 + 2x^3 + 9x \right]_0^1 = \frac{56}{5}\pi.$$

– Drehung um die y -Achse mit $x = f^{-1}(y) = \sqrt{y-3}$:

$$V = \pi \int_3^4 \left[\sqrt{y-3} \right]^2 dy = \pi \int_3^4 y - 3 dy = \pi \left[\frac{y^2}{2} - 3y \right]_3^4 = \frac{\pi}{2}. \quad (2.63)$$

b) Eine Kugel mit Radius r entsteht durch Rotation der Funktion $f(x) = \sqrt{r^2 - x^2}$ um die x -Achse auf dem Intervall $[-r, r]$:

$$V = \pi \int_{-r}^r r^2 - x^2 dx = \pi \left[r^2 x - \frac{1}{3}x^3 \right]_{-r}^r = \frac{4}{3}\pi r^3.$$

■

2.6.7.3 Oberfläche eines Rotationskörpers

Auch der Flächeninhalt der Oberfläche eines Rotationskörpers lässt sich über ein „normales“ Integral ausrechnen. Wir benötigen keine mehrdimensionale Integration.

Wir lassen wieder den Funktionsgraphen einer Funktion $f(x)$ im Intervall $[a, b]$ um die x -Achse rotieren und leiten eine Formel für den Inhalt der entstehenden Oberfläche her. Statt das Volumen von Zylinderstücken zu berechnen, verwenden wir nun auf einer Zerlegung des Intervalls $[a, b]$ in n gleich große Teilintervalle Kegelstümpfe und summieren deren angenäherte Mantelflächen (Stirnumfang $2\pi f\left(a + k\frac{b-a}{n}\right)$ des Stumpfs mal Länge der Seite, die sich mittels des Satzes von Pythagoras ergibt):

$$\begin{aligned} & \sum_{k=0}^{n-1} 2\pi f\left(a + k\frac{b-a}{n}\right) \sqrt{\left(\frac{b-a}{n}\right)^2 + \left[f\left(a + (k+1)\frac{b-a}{n}\right) - f\left(a + k\frac{b-a}{n}\right)\right]^2} \\ &= 2\pi \frac{b-a}{n} \sum_{k=0}^{n-1} f\left(a + k\frac{b-a}{n}\right) \sqrt{1 + \left[\frac{f\left(a + (k+1)\frac{b-a}{n}\right) - f\left(a + k\frac{b-a}{n}\right)}{\frac{b-a}{n}}\right]^2}. \end{aligned}$$

Für $n \rightarrow \infty$ entsteht unter der Wurzel aus dem Differenzenquotienten die Ableitung von f . Zusätzlich wird die Summe zum Integral, und wir erhalten die Formel für die Oberfläche O :

$$O = 2\pi \int_a^b f(x) \cdot \sqrt{1 + f'(x)^2} dx.$$

Beispiel 2.115

Wir betrachten wieder die Kugel mit Radius r , die durch Rotation von $y = f(x) = \sqrt{r^2 - x^2}$ für $x \in [-r, r]$ um die x -Achse entsteht. Mit $f'(x) = \frac{-x}{\sqrt{r^2 - x^2}}$ erhalten wir die Größe der Kugeloberfläche

$$2\pi \int_{-r}^r \sqrt{r^2 - x^2} \cdot \sqrt{\frac{r^2}{r^2 - x^2}} dx = 4\pi \cdot r \int_0^r 1 dx = 4\pi \cdot r^2. \quad \blacksquare$$

2.6.8 Lebesgue-Integral*

Es gibt eine für theoretische Zwecke wichtige Erweiterung des Riemann-Integrals: das Lebesgue-Integral. Während Riemann-Integrale in der Ingenieurpraxis völlig ausreichend sind, machen sie in der Mathematik erhebliche Schwierigkeiten. Wir haben gesehen, dass man bei unbeschränkten Funktionen den Integrationsbereich so einschränken muss, dass die Funktion beschränkt ist. Dann ist mindestens ein zusätzlicher Grenzwert hinsichtlich des Integrationsrandes zu berechnen. Ebenso müssen Integrale auf unbeschränktem Integrationsbereich behandelt werden. Bereits eine „einfache“ Funktion wie die Dirichlet-Funktion, die für rationale Argumente den Wert 1 und für irrationale den Funktionswert 0 hat (vgl. Beispiel 2.40 auf Seite 263 und Seite 336), ist auf $[0, 1]$ nicht Riemann-integrierbar, da alle Untersummen gleich null und alle Obersummen gleich eins sind.

Häufig wird die „Größe“ einer Funktion mit Hilfe von Integration gemessen. So lassen sich auch der Abstand zweier Funktionen und die Konvergenz einer Folge von Funktionen gegen eine Grenzfunktion erklären, siehe Kapitel 3.7.3. Das Hauptproblem des Riemann-Integrals ist dann aber, dass ein Vollständigkeitsaxiom fehlt. Die Konvergenz verhält sich anders als bei reellen Zahlen. Eine im geeigneten Sinne konvergente Folge von Riemann-integrierbaren Funktionen kann eine Grenzfunktion besitzen, die nicht Riemann-integrierbar ist. Das sorgt bei Beweisen für massive Probleme und ist damit vergleichbar, dass eine Folge von Brüchen einen irrationalen Grenzwert besitzen kann. Daher benötigt man zumindest für theoretische Zwecke einen „harmloseren“ Integrationsbegriff als das Riemann-Integral, auch wenn dessen Definition zunächst komplizierter erscheint. Dieser ist das **Lebesgue-Integral**.

Bevor wir in den nächsten Abschnitten das Lebesgue-Integral formal einführen, versuchen wir es zunächst einmal anschaulich zu erklären. Die Idee zum Lebesgue-Integral ist überraschend einfach. Beim Riemann-Integral wird die x -Achse in Teilintervalle zerlegt. Zu jedem Zerlegungsintervall berechnet man den Flächeninhalt einer Säule mit der Breite des Intervalls und mit einer Höhe, die z. B. einem Funktionswert an einem Zwischenpunkt entspricht. Statt eines Funktionswerts an einem Zwischenpunkt kann man für eine Untersumme auch die größte untere Schranke (Infimum) der Funktionswerte oder für eine Obersumme die kleinste obere Schranke (Supremum) der Funktionswerte auf diesem Intervall verwenden. Beim Lebesgue-Integral zerlegt man nun nicht die x -, sondern die y -Achse. Zu jedem Zerlegungsintervall sucht man die Menge aller zugehörigen x -Werte (Urbildmenge). Diese kann jetzt allerdings recht schwierig aussehen. In Abbildung 2.34 sind die Urbildmengen eine Vereinigung endlich vieler Intervalle. Wenn man für jede Urbildmenge einen Inhalt (eine Größe, ein Maß) berechnen kann, dann kann man pro Zerlegungsintervall das Produkt eines Funktionswerts mit der Größe der zugehörigen x -Menge bilden und über die Ergebnisse summieren. Wir betrachten das etwas detaillierter. Dazu sei $f : [a, b] \rightarrow \mathbb{R}$ eine nicht-negativwertige Funktion. Hat man eine Funktion, die auch negative Werte besitzt, dann schreibt man diese als Differenz zweier nicht-negativwertiger Funktionen $f = f^+ - f^-$ und integriert diese separat. Statt eines beschränkten Intervalls $[a, b]$ kann man auch unbeschränkte Intervalle oder $\mathbb{R}$ betrachten. Das tun wir in dieser Motivation aber zur Vereinfachung der Darstellung noch nicht. Jetzt zerlegen wir für ein $n \in \mathbb{N}$ die nicht-negative y -Achse in die Intervalle $Y_{k,n} := [(k-1)/n, k/n[, k \in \mathbb{N}$. Zu jedem dieser Intervalle sei $X_{k,n} := \{x \in [a, b] : f(x) \in Y_{k,n}\}$ das Urbild von f . Welches Aussehen $X_{k,n}$ hat, hängt von f ab. Um hier weiterzukommen, müssen wir die Größe von $X_{k,n}$ messen. Dieses „Volumen“ bezeichnen wir mit $v(X_{k,n})$. Dann können wir, sofern der Grenzwert eigentlich oder uneigentlich existiert, das Lebesgue-Integral definieren:

$$\text{L-} \int_a^b f(x) dx := \lim_{n \rightarrow \infty} \sum_{k=1}^{\infty} \frac{k-1}{n} v(X_{k,n}).$$

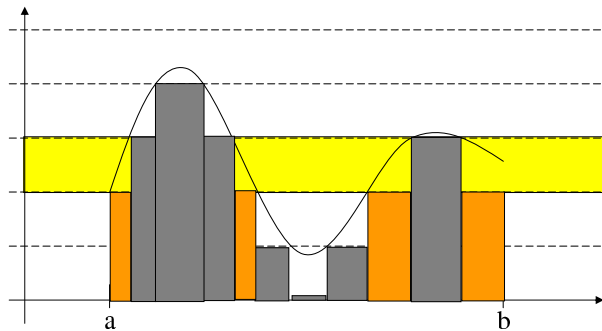


Abb. 2.34: Zu den in den horizontalen Streifen enthaltenen y -Werten werden die zugehörigen x -Werte (Urbilder) ermittelt. Auf diesen Mengen werden dann Säulen gesetzt, die bis an einen Wert aus den horizontalen Streifen (hier der kleinste Wert) heranreichen. Die Summe über alle Säuleninhalte ist ein Näherungswert für das Lebesgue-Integral. Das entsteht als Grenzwert, wenn man die horizontalen Streifen immer feiner wählt. Im Bild sind die vier Säulen zum gelb markierten Zerlegungsintervall der y -Achse orange eingefärbt.

Der Grenzwert $n \rightarrow \infty$ sorgt dafür, dass die Zerlegung der y -Achse immer feiner wird. Zu jeder Zerlegung werden dann mit der unendlichen Summe die Flächen aufaddiert. Dabei wird der kleinste Wert $\frac{k-1}{n}$ des Zerlegungsintervalls wie in Abbildung 2.34 näherungsweise als Funktionswert verwendet. Insbesondere benötigt man bei dieser Definition nicht die Beschränktheit von f . Die verbleibende Schwierigkeit besteht in der Größenmessung von $X_{k,n}$. Diese funktioniert nicht für jede beliebige Funktion, sondern nur für **messbare** Funktionen. Eine messbare Funktion erfüllt die Bedingung, dass alle Urbildmengen $X_{k,n}$ (wie in Abbildung 2.34) Lebesgue-messbar sind. Eine Menge heißt **Lebesgue-messbar**, wenn sie Element einer speziellen Menge von Mengen ist, die **Lebesgue- σ -Algebra** heißt. Der Begriff σ -Algebra stammt aus der Wahrscheinlichkeitsrechnung und wird dort für eine Menge von Ereignissen verwendet. In diesem Buch verwenden wir dafür den Namen Ereignis-Algebra, siehe Seite 998. Die Lebesgue- σ -Algebra hat genau diese Struktur. Wie die Menge genau konstruiert ist, betrachten wir im nächsten Abschnitt. Wichtig sind aber folgende Eigenschaften:

- In dieser Lebesgue- σ -Algebra liegen alle Intervalle I , die damit messbar sind. Jedem Intervall kann der Inhalt (das **Lebesgue-Maß**) $v(I) := \text{„rechter Endpunkt} - \text{linker Endpunkt“}$ zugeordnet werden, außerdem ist $v(\emptyset) = 0$.
- Hat man eine höchstens abzählbar unendliche Vereinigung von disjunkten messbaren Mengen, so ist auch diese Menge messbar, ihr Lebesgue-Maß ist die (ggf. unendliche) Summe der Lebesgue-Maße der einzelnen Mengen. Das entspricht der Additivität eines Wahrscheinlichkeitsmaßes (siehe Definition 7.15 auf Seite 1000).
- Betrachtet man eine messbare Menge $M \subset [a, b]$, so ist ihr Komplement messbar mit $v(\mathcal{C}_{[a,b]} M) = (b - a) - v(M)$.

- Sind A und B messbare Mengen mit $A \subset B$ und $v(A) = v(B)$, dann ist auch jede Menge C mit $A \subset C \subset B$ messbar mit gleichem Inhalt $v(C) = v(A)$. Die Mengen A , B und C unterscheiden sich damit nur durch eine Menge vom Lebesgue-Maß 0. Häufig gelten mathematische Sätze für alle reellen Zahlen bis auf wenige Ausnahmen, die alle in einer (nicht genauer bekannten) Menge vom Lebesgue-Maß null liegen. Dann schreibt man, die Aussage gilt „f. ü.“ (fast überall). In der englischsprachigen Literatur wird entsprechend „a. e.“ (almost everywhere) verwendet.

Tatsächlich ist jede Riemann-integrierbare Funktion Lebesgue-integrierbar. Die Umkehrung gilt allerdings nicht, wie die Dirichlet-Funktion f zeigt. Wir wissen bereits, dass f nicht Riemann-integrierbar ist, aber f ist Lebesgue-integrierbar:

$$\text{L-} \int_0^1 f(x) dx = 0 \cdot v([0, 1] \setminus \mathbb{Q}) + 1 \cdot v([0, 1] \cap \mathbb{Q}) = 0 \cdot 1 + 1 \cdot 0 = 0,$$

denn die rationalen Zahlen im Intervall $[0, 1]$ sind abzählbar, und für jede Summationsreihenfolge gilt

$$\begin{aligned} v([0, 1] \cap \mathbb{Q}) &= \sum_{p \in [0, 1] \cap \mathbb{Q}} v([p, p]) = \sum_{p \in [0, 1] \cap \mathbb{Q}} 0 = 0, \\ v([0, 1] \setminus \mathbb{Q}) &= v(\mathcal{C}_{[0, 1]}(\mathbb{Q} \cap [0, 1])) = 1 - v(\mathbb{Q} \cap [0, 1]) = 1 - 0 = 1. \end{aligned}$$

Wie hier liefern Mengen vom Lebesgue-Maß null keinen Beitrag bei der Berechnung von Integralen. Wenn sich zwei Funktionen nur auf einer Menge vom Maß null unterscheiden, sie also f. ü. gleich sind, dann sind ihre Integrale gleich.

Da die vom Riemann-Integral bekannten Rechenregeln auch für das Lebesgue-Integral gelten, spielt es beim praktischen Rechnen keine Rolle, um welches Integral es sich handelt.

In der Wahrscheinlichkeitsrechnung wird häufig mit Wahrscheinlichkeitsmaßen integriert. Dabei wird $v(X_{k,n})$ über die Wahrscheinlichkeit definiert, dass $X_{k,n}$ eintritt. Das Lebesgue-Maß wird also durch ein Wahrscheinlichkeitsmaß ausgetauscht. Solche Integrale behandelt man in der allgemeinen Maß- und Integrationstheorie, siehe Bauer (1968). Wir gehen in Kapitel 7.2.9 etwas genauer auf die Integration mit Wahrscheinlichkeitsmaßen ein.

2.6.8.1 Messbare Mengen

In diesem Unterkapitel wollen wir exakt definieren, was Lebesgue-Messbarkeit und das Lebesgue-Maß einer Menge von reellen Zahlen bedeutet. Das Folgende kann direkt auf Teilmengen des $\mathbb{R}^n$ übertragen werden (vgl. Kapitel 4.4.5), aber wir begnügen uns mit Teilmengen der reellen Zahlen $\mathbb{R}$. Dadurch werden die Schreibweisen etwas einfacher.

Definition 2.36 (Äußeres Lebesgue-Maß, Lebesgue-Messbarkeit)

Mit dem elementaren **Inhalt** $v(]a, b[) = b - a$ definieren wir das **äußere Lebesgue-Maß** m^* einer Teilmenge $A \subset \mathbb{R}$. Sei dazu

$$U := \left\{ \sum_{k=1}^{\infty} v(I_k) : I_k \text{ sind offene, beschränkte Intervalle mit } A \subset \bigcup_{k=1}^{\infty} I_k \text{ und } \sum_{k=1}^{\infty} v(I_k) < \infty \right\}.$$

Falls $U = \emptyset$ ist, dann setzen wir $m^*(A) := \infty$, anderenfalls existiert das Infimum der nicht-leeren, nach unten mit 0 beschränkten Menge U aufgrund des Vollständigkeitsaxioms der reellen Zahlen (siehe Seite 49), und wir setzen

$$m^*(A) := \inf U.$$

Eine Menge $A \subset \mathbb{R}$ heißt **Lebesgue-messbar** genau dann, wenn

$$m^*(B) = m^*(B \cap A) + m^*(B \setminus A) \text{ für alle } B \subset \mathbb{R} \quad (2.64)$$

gilt. Dabei darf $\infty + \infty = \infty$ gerechnet werden.

Es lässt sich nachrechnen, dass die Menge aller Lebesgue-messbaren Mengen tatsächlich eine σ -Algebra ist (die Lebesgue- σ -Algebra). Wir bezeichnen sie mit $\mathcal{M}$. Insbesondere sind $\emptyset$, $\mathbb{R}$ und alle beschränkten Intervalle messbar.

Das äußere Maß m^* ist für jede Teilmenge der reellen Zahlen erklärt. Wenn wir es auf die Lebesgue-messbaren Mengen mit der Bedingung (2.64) von Constantin Carathéodory einschränken, dann heißt es das **Lebesgue-Maß** und wird mit m bezeichnet. Insbesondere gilt:

$$m(\emptyset) = 0, \quad m(\mathbb{R}) = \infty, \quad m(]a, b[) = m([a, b]) = b - a.$$

Außerdem hat das Lebesgue-Maß die folgenden Eigenschaften:

$$\begin{aligned} m(A) &\leq m(B) \text{ für alle } A, B \in \mathcal{M} \text{ mit } A \subset B, \\ m\left(\bigcup_{k=1}^{\infty} A_k\right) &= \sum_{k=1}^{\infty} m(A_k) \text{ für jede Folge paarweise disjunkter } A_k \in \mathcal{M}, \\ m(B \setminus A) &= m(B) - m(A) \text{ für alle } A, B \in \mathcal{M} \text{ mit } A \subset B \text{ und } m(A) < \infty. \end{aligned} \quad (2.65)$$

Die Einschränkung (2.64) der Mengen ist erforderlich, damit das Lebesgue-Maß σ -additiv wird, das ist die Eigenschaft (2.65). Das äußere Maß auf beliebigen disjunkten Teilmengen von $\mathbb{R}$ erlaubt in (2.65) lediglich kleinergleich und damit σ -Subadditivität.

Alle Mengen mit äußerem Maß null sind Lebesgue-messbar. Weiterhin sind alle abzählbaren Mengen messbar mit Lebesgue-Maß null. Mit der Cantormenge gibt es umgekehrt aber eine überabzählbare Menge, die ebenfalls das Lebesgue-Maß null hat. Die abzählbaren Mengen bilden also eine echte Teilmenge der Mengen mit Lebesgue-Maß null.

2.6.8.2 Messbare Funktionen

Definition 2.37 (Messbare Funktion)

Eine Funktion $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ heißt **Lebesgue-messbar** genau dann, wenn $D \in \mathcal{M}$ ist und für alle $a \in \mathbb{R}$ gilt:

$$f^{-1}([a, \infty[) = \{x \in D : f(x) \geq a\} \in \mathcal{M}.$$

Das Urbild jedes Intervalls $[a, \infty[$ muss also eine messbare Menge sein. Das benötigen wir, wenn wir die y -Achse zerlegen und dann nach Urbildern für die Zerlegungsintervalle suchen. Wir können nun sogar erlauben, dass Funktionen die Funktionswerte „ $\pm\infty$ “ annehmen. In diesem Sinne ist eine Funktion $f : D \subset \mathbb{R} \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ messbar genau dann, wenn ihr Definitionsbereich D messbar ist und $f^{-1}([a, \infty[\cup \{\infty\})$ für jede Zahl $a \in \mathbb{R}$ eine messbare Menge ist. Die Werte „ $\pm\infty$ “ sind beim Rechnen unproblematisch, wenn sie nur auf einer Menge vom Maß null angenommen werden. Wie bereits in der Motivation beschrieben, spielen Mengen vom Maß null für die Integration keine Rolle. Das Integral über eine Menge vom Maß null wird stets zu null definiert, egal ob dort f an einer Stelle gleich „ $\pm\infty$ “ ist.

Wir sagen $f = g$ f. ü. (fast überall) für zwei messbare Funktionen mit gemeinsamem Definitionsbereich D genau dann, wenn die Menge $\{x \in D : f(x) \neq g(x)\}$ das Lebesgue-Maß null hat. Entsprechend sind Schreibweisen wie $f < g$ f. ü. (d. h. $\{x \in D : f(x) \geq g(x)\}$ hat Lebesgue-Maß null) zu verstehen.

2.6.8.3 Definition des Lebesgue-Integrals

Wir haben oben das Lebesgue-Integral durch eine Zerlegung der y -Achse motiviert. Dabei haben wir die Urbilder der Zerlegungsintervalle gesucht und dann eine Funktion integriert, die auf jedem (messbaren) Urbild konstant war. Wir lösen uns jetzt von der Zerlegung der y -Achse und führen das Lebesgue-Integral über solche stückweise konstanten Funktionen ein.

Wir integrieren zunächst einfache Funktionen. Zu einer Menge $A \subset \mathbb{R}$ sei 1_A die **charakteristische Funktion** mit

$$1_A(x) := \begin{cases} 1 & : x \in A \\ 0 & : x \in \mathbb{R} \setminus A. \end{cases}$$

Eine **einfache Funktion** $g : \mathbb{R} \rightarrow \mathbb{R}$ ist eine endliche Linearkombination von charakteristischen Funktionen zu disjunkten messbaren Mengen, die ein endliches Lebesgue-Maß besitzen:

$$g(x) := \sum_{k=1}^n c_k 1_{A_k}(x)$$

für ein $n \in \mathbb{N}$, reelle Konstanten c_k und disjunkte messbare Mengen A_k . Diese Funktionen nehmen nur endlich viele verschiedene Werte an. Die zugehörigen Urbildmengen können aber wie bei der Dirichlet-Funktion, die eine einfache Funktion mit den Parametern $n = 2$, $c_1 = 0$, $c_2 = 1$, $A_1 = [0, 1] \setminus \mathbb{Q}$ und $A_2 = [0, 1] \cap \mathbb{Q}$ ist, sehr kompliziert sein. Sei $A := \bigcup_{k=1}^n A_k$. Dann definiert man für eine einfache Funktion g

$$\mathrm{L}\text{-}\int_A g(x) dx := \sum_{k=1}^n c_k m(A_k).$$

Eine einfache Funktion g kann natürlich unterschiedliche Darstellungen als Linearkombinationen charakteristischer Funktionen besitzen. Der Wert des Integrals ist aber unabhängig von der Wahl der Darstellung. Das folgt aus den Eigenschaften des Lebesgue-Maßes.

Jetzt integrieren wir mit einfachen Funktionen auch kompliziertere Funktionen. Dabei verwenden wir zunächst noch Beschränktheitsbedingungen an f und A und damit ähnliche Einschränkungen wie beim Riemann-Integral. Davon werden wir uns aber in den nächsten Schritten lösen. Für eine messbare beschränkte Funktion f auf einer messbaren Menge A mit endlichem Maß definieren wir

$$\begin{aligned} & \mathrm{L}\text{-}\int_A f(x) dx \\ &:= \sup \left\{ \mathrm{L}\text{-}\int_A g(x) dx : g \text{ ist einfache Funktion mit } g(x) \leq f(x) \text{ für alle } x \in A \right\}. \end{aligned}$$

Das Supremum existiert, da es über eine nicht-leere und mit $m(A) \sup\{f(x) : x \in A\}$ beschränkte Menge gebildet wird.

Jetzt lassen wir die Randbedingungen weg, dass f und A beschränkt sein sollen. Für eine nun nicht-negativwertige messbare Funktion f , die auch Funktionswerte „ ∞ “ besitzen darf, und eine messbare Menge A (mit beliebigem Maß) sei

$$U := \left\{ \mathrm{L}\text{-}\int_{\{x \in A : h(x) \neq 0\}} h(x) dx : h \text{ ist eine messbare, beschränkte Funktion mit} \right.$$

$$h(x) \leq f(x) \text{ für alle } x \in A \text{ und } m(\{x \in A : h(x) \neq 0\}) < \infty \}.$$

Hier wird das bereits erklärte Integral einer beschränkten Funktion über eine Menge mit endlichem Maß benutzt. Damit definieren wir

$$\mathrm{L}\text{-}\int_A f(x) dx := \begin{cases} \sup U, & \text{falls } U \text{ nach oben beschränkt ist.} \\ \infty, & \text{sonst.} \end{cases}$$

Das so definierte Integral kann also durchaus den Wert „ ∞ “ annehmen. Wir sagen, dass eine nicht-negativwertige messbare Funktion f genau dann **Lebesgue-integrierbar** ist, wenn das Integral einen endlichen Wert hat. Insbesondere kann dann der Funktionswert „ ∞ “ nur auf einer Menge vom Maß null angenommen werden (und ist ohne Einfluss auf die Integration auch durch eine echte Zahl ersetzbar).

Zum Abschluss definieren wir das Lebesgue-Integral für allgemeine messbare Funktionen f über messbare Mengen A . Diese lassen sich als Summe von nicht-negativwertigen Funktionen f^+ und f^- schreiben: $f(x) = f^+(x) - f^-(x)$. f ist genau dann **Lebesgue-integrierbar** über A , wenn f^+ und f^- als nicht-negativwertige Funktionen Lebesgue-integrierbar sind. Insbesondere haben ihre Integrale einen endlichen Wert, und damit macht die folgende Definition Sinn:

$$\mathrm{L}\text{-}\int_A f(x) dx := \mathrm{L}\text{-}\int_A f^+(x) dx - \mathrm{L}\text{-}\int_A f^-(x) dx.$$

2.6.8.4 Eigenschaften des Lebesgue-Integrals

Wir stellen in diesem Abschnitt ohne Beweise (hier sei z. B. auf Bauer (1990) verwiesen) die wichtigsten Eigenschaften und Sätze zum Lebesgue-Integral zusammen.

- Das Lebesgue-Integral ist wie das Riemann-Integral linear, d. h., für Lebesgue-integrierbare Funktionen f und g auf einer messbaren Menge A und $c, d \in \mathbb{R}$ gilt:

$$\mathrm{L}\text{-}\int_A c \cdot f(x) + d \cdot g(x) dx = c \cdot \mathrm{L}\text{-}\int_A f(x) dx + d \cdot \mathrm{L}\text{-}\int_A g(x) dx.$$

- Sind $A, B \in \mathcal{M}$ mit $m(A \cap B) = 0$, dann lässt sich der Integrationsbereich so zerlegen:

$$\mathrm{L}\text{-}\int_{A \cup B} f(x) dx = \mathrm{L}\text{-}\int_A f(x) dx + \mathrm{L}\text{-}\int_B f(x) dx.$$

- Sei $(A_k)_{k=1}^\infty$ eine Folge disjunkter messbarer Mengen, und sei f Lebesgue-integrierbar über $A := \bigcup_{k=1}^\infty A_k$. Dann gilt:

$$\mathrm{L}\text{-}\int_A f(x) dx = \sum_{k=1}^\infty \mathrm{L}\text{-}\int_{A_k} f(x) dx. \quad (2.66)$$

- Falls f auf $A \in \mathcal{M}$ messbar ist, dann ist auch $|f|$ auf $A \in \mathcal{M}$ messbar und

$$f \text{ ist auf } A \text{ integrierbar} \iff |f| \text{ ist auf } A \text{ integrierbar}.$$

Man beachte, dass es zu dieser Aussage kein Gegenstück für Riemann-integrierbare Funktionen gibt. So ist die etwas undefinierte messbare Dirichlet-Funktion

$$f(x) := \begin{cases} 1 & : x \in \mathbb{Q} \\ -1 & : x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

auf dem Intervall $[0, 1]$ nicht Riemann-integrierbar, aber $|f(x)| = 1$ sehr wohl.

- Ist f auf A Lebesgue-integrierbar, dann gilt die verallgemeinerte Dreiecksungleichung

$$\left| L- \int_A f(x) dx \right| \leq L- \int_A |f(x)| dx.$$

- Ist f auf der messbaren Menge A Lebesgue-integrierbar und g auf A messbar und gilt $f(x) = g(x)$ f. ü., dann ist auch g Lebesgue-integrierbar und

$$L- \int_A f(x) dx = L- \int_A g(x) dx.$$

- Ist f eine auf $[a, b]$ Riemann-integrierbare Funktion, dann ist f auf der messbaren Menge $[a, b]$ Lebesgue-integrierbar mit

$$L- \int_a^b f(x) dx := L- \int_{[a,b]} f(x) dx = \int_a^b f(x) dx.$$

Das Lebesgue-Integral ist wegen des letzten Punktes eine echte Erweiterung des Riemann-Integrals. Für die in der Ingenieurpraxis häufig auftretenden Riemann-integrierbaren Funktionen benötigt man den Begriff des Lebesgue-Integrals nicht. Er wird aber wichtig, wenn es um Grenzwerte von Funktionenfolgen geht: Aufgrund des Aufbaus des Integrals über die Approximation durch einfache Funktionen von unten ist die folgende Aussage nicht überraschend:

Satz 2.52 (Satz von Beppo-Levi zur monotonen Konvergenz)

Sei $(f_k)_{k=1}^\infty$ eine Folge auf einer messbaren Menge A definierter messbarer, nicht-negativwertiger Funktionen. Die Folge sei monoton wachsend, d. h. $f_k(x) \leq f_{k+1}(x)$ f. ü. für alle $k \in \mathbb{N}$. Außerdem gelte

$$\lim_{k \rightarrow \infty} f_k(x) = f(x) \text{ f. ü.,}$$

d. h., für alle $x \in A$ mit Ausnahmen, die in einer Menge vom Lebesgue-Maß null liegen, konvergiert die zu festgehaltenem x berechnete Zahlenfolge (vgl. Kapitel 2.8.2). Dann ist f messbar, und es gilt:

$$\lim_{k \rightarrow \infty} L- \int_A f_k(x) dx = L- \int_A f(x) dx.$$

Ohne die Einschränkung, dass die Funktionenfolge monoton wachsend ist, kommt der folgende Hauptsatz aus, der dafür eine Lebesgue-integrierbare Majorante verwendet:

Satz 2.53 (Satz von Lebesgue über die majorisierte Konvergenz)

Sei $(f_k)_{k=1}^\infty$ eine Folge auf einer messbaren Menge A definierter Lebesgue-integrierbarer Funktionen mit

$$\lim_{k \rightarrow \infty} f_k(x) = f(x) \text{ f. ü.}$$

Falls eine auf A Lebesgue-integrierbare Funktion g existiert mit $|f_k(x)| \leq g(x)$ f. ü., dann ist f auf A Lebesgue-integrierbar, und es gilt:

$$\lim_{k \rightarrow \infty} L- \int_A f_k(x) dx = L- \int_A f(x) dx \text{ und } \lim_{k \rightarrow \infty} L- \int_A |f_k(x) - f(x)| dx = 0.$$

Mit diesen beiden Sätzen kann man zeigen, dass das in der Motivation des Abschnitts erwähnte Problem der fehlenden Integrierbarkeit einer Grenzfunktion beim Lebesgue-Integral nicht auftritt (siehe Kasten auf Seite 564).

Aufbau und Eigenschaften des Lebesgue-Integrals lassen sich von $\mathbb{R}$ direkt auf den $\mathbb{R}^n$, also auf Funktionen mit n Variablen übertragen, siehe dazu Kapitel 4.4.5.

2.7 Satz von Taylor, Kurvendiskussion und Extremalprobleme

2.7.1 Taylor-Summen

Wir approximieren eine Funktion f durch ein Polynom p_m , das an einer gegebenen Stelle x_0 den gleichen Funktionswert $f(x_0)$ und die gleichen Ableitungswerte $f^{(k)}(x_0)$, $1 \leq k \leq m$, besitzt. Dabei haben wir die Hoffnung, dass dadurch das Polynom das Verhalten von f gut imitiert. Diese Hoffnung wird sich in Kapitel 2.8 bestätigen, aber auch schon im folgenden Satz von Taylor zeigen. Funktionswerte eines Polynoms kann man leicht ausrechnen. Dagegen weiß man bei vielen anderen Funktionen gar nicht, wie man Funktionswerte praktisch berechnen soll. Hier nutzt man dann z. B. die näherungsweise Darstellung durch das zuvor beschriebene Polynom.

Wir verwenden (nur lokal) die Variable m für den Grad des Polynoms p_m , da wir die Ergebnisse später auch für Funktionen mit mehreren Variablen verwenden wollen. Die Anzahl der Variablen (die der Raumdimension des Definitionsbereichs entspricht) wird nämlich üblicherweise mit n bezeichnet.

■ Für $m = 0$ ist $p_0(x) = f(x_0)$ die konstante Funktion.

■ Für $m = 1$ ist

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

die Tangente an f in x_0 . Über den Schnittpunkt der Tangenten mit der x -Achse haben wir bereits beim Newton-Verfahren (siehe Seite 305) näherungsweise die Nullstellen einer differenzierbaren Funktion bestimmt. Bereits diese einfache Näherung ist also praktisch anwendbar.

Allgemein diskutieren wir die **Taylor-Summe** bzw. das **Taylor-Polynom** oder auch die **Taylor-Entwicklung** vom Grad m

$$\begin{aligned} p_m(x) &:= f(x_0) + f'(x_0)(x - x_0) + \frac{f^{(2)}(x_0)}{2}(x - x_0)^2 + \frac{f^{(3)}(x_0)}{6}(x - x_0)^3 + \dots \\ &\quad + \frac{f^{(m)}(x_0)}{m!}(x - x_0)^m = \sum_{k=0}^m \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k. \end{aligned}$$

Wählt man speziell $x_0 = 0$, so spricht man auch von einer **MacLaurin-Entwicklung**.

Wir zeigen, dass an der Stelle x_0 tatsächlich alle Ableitungen bis zur m -ten von f und p_m übereinstimmen, d. h. $f^{(k)}(x_0) = p_m^{(k)}(x_0)$, $0 \leq k \leq m$:

$$\begin{aligned} p_m^{(k)}(x) &= \frac{f^{(k)}(x_0)}{k!} k! + \frac{f^{(k+1)}(x_0)}{(k+1)!} \frac{(k+1)!}{1!} (x - x_0) + \frac{f^{(k+2)}(x_0)}{(k+2)!} \frac{(k+2)!}{2!} (x - x_0)^2 \\ &\quad \dots + \frac{f^{(m)}(x_0)}{m!} \frac{m!}{(m-k)!} (x - x_0)^{m-k}. \end{aligned}$$

Setzen wir $x = x_0$ ein, so bleibt nur $p_m^{(k)}(x_0) = f^{(k)}(x_0)$ stehen. Jedes weitere Summenglied sorgt also dafür, dass die nächsthöhere Ableitung des Polynoms mit der von f an der Stelle x_0 übereinstimmt. Die Abweichung zwischen dem so bestimmten Polynom und der Ausgangsfunktion f lässt sich abschätzen:

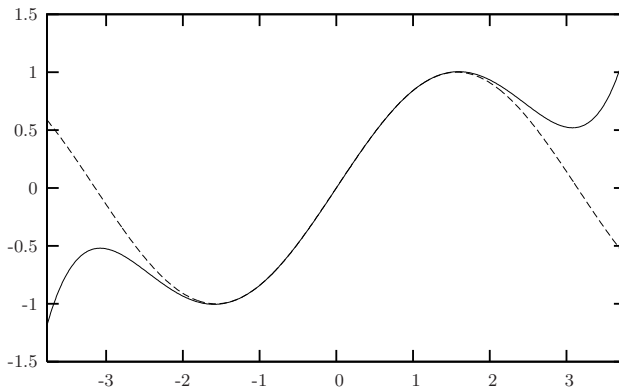


Abb. 2.35: $\sin x$ und zugehörige Taylor-Summe $\sum_{k=0}^2 (-1)^k \frac{x^{2k+1}}{(2k+1)!}$

Satz 2.54 (Satz von Taylor, 1685–1731, hier 1715)

Seien $]a, b[\subset \mathbb{R}$ ein offenes Intervall, $x_0 \in]a, b[$ und f eine $(m+1)$ -mal stetig differenzierbare Funktion auf $]a, b[$, $m \in \mathbb{N}_0$. Für jedes $x \in]a, b[$ gilt:

$$f(x) = \sum_{k=0}^m \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + R_{x_0, m+1, f}(x). \quad (2.67)$$

Dabei hat das Restglied $R_{x_0, m+1, f}(x)$ die Integraldarstellung

$$R_{x_0, m+1, f}(x) = \int_{x_0}^x \frac{(x-t)^m}{m!} f^{(m+1)}(t) dt. \quad (2.68)$$

Alternativ zur Integraldarstellung ist die **Lagrange-Darstellung** des Restglieds nützlich: Zu jedem $x \in]a, b[$ gibt es ein $\xi = \xi(x)$ zwischen x und x_0 , so dass

$$R_{x_0, m+1, f}(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!} (x - x_0)^{m+1}. \quad (2.69)$$

Ist f bereits selbst ein Polynom mit Grad $\leq m$, d. h., der größte auftretende Exponent ist $\leq m$, so ist das Restglied null, und die Taylor-Darstellung ist identisch mit dem Polynom. Man spricht von einer **Projektion**: Polynome vom Grad $\leq m$ werden durch Bilden der Taylor-Summe auf sich selbst abgebildet.

Im Fall $m = 0$ entspricht die Lagrange-Darstellung des Restglieds dem Mittelwertsatz, der damit Spezialfall des Satzes von Taylor ist:

$$f(x) = f(x_0) + R_{x_0, 1, f}(x) = f(x_0) + f'(\xi)(x - x_0), \text{ d. h. } \frac{f(x) - f(x_0)}{x - x_0} = f'(\xi).$$

Beweis: Wir zeigen zuerst (2.67) mit dem Restglied in Integralform (2.68) mittels Hauptsatz und partieller Integration.

- Der Induktionsanfang für $m = 0$ ergibt sich aus dem Hauptsatz (Satz 2.48, Seite 340):

$$f(x) = f(x_0) + \int_{x_0}^x f'(t) dt$$

- Induktionsannahme: (2.67) gelte für ein beliebiges m .
- Induktionsschritt: Wir zeigen, dass die Taylor-Formel auch für $m+1$ gilt. Dazu integrieren wir das Integralrestglied partiell:

$$\begin{aligned} & \frac{f^{(m+1)}(x_0)}{(m+1)!} (x - x_0)^{(m+1)} + R_{x_0, m+2, f}(x) \\ &= \frac{f^{(m+1)}(x_0)}{(m+1)!} (x - x_0)^{(m+1)} + \int_{x_0}^x \frac{(x-t)^{m+1}}{(m+1)!} f^{(m+2)}(t) dt \end{aligned}$$

$$\begin{aligned}
&= \frac{f^{(m+1)}(x_0)}{(m+1)!} (x-x_0)^{(m+1)} + \left[\frac{(x-t)^{m+1}}{(m+1)!} f^{(m+1)}(t) \right]_{t=x_0}^{t=x} \\
&\quad - \int_{x_0}^x -\frac{(x-t)^m}{m!} f^{(m+1)}(t) dt = \int_{x_0}^x \frac{(x-t)^m}{m!} f^{(m+1)}(t) dt = R_{x_0, m+1, f}(x).
\end{aligned}$$

Damit ist aber

$$\sum_{k=0}^{m+1} \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k + R_{x_0, m+2, f}(x) = \sum_{k=0}^m \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k + R_{x_0, m+1, f}(x),$$

und nach Induktionsannahme ist das gleich $f(x)$.

Mit der Induktion ist (2.67) mit der Restglieddarstellung (2.68) bewiesen.

Mit dem Mittelwertsatz der Integralrechnung (Satz 2.47) erhält man direkt aus (2.68) eine Darstellung (das **Cauchy'sche Restglied**)

$$R_{x_0, m+1, f}(x) = \frac{f^{(m+1)}(\xi)}{m!} (x-x_0)(x-\xi)^m$$

für ein ξ im abgeschlossenen Intervall zwischen x_0 und x . Das ist aber leider noch nicht die Lagrange-Darstellung (2.69). Wir beweisen sie separat, indem wir mit

$$R_{x_0, m+1, f}(x) = R(x) := f(x) - \sum_{k=0}^m \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k$$

starten und den verallgemeinerten Mittelwertsatz der Differenzialrechnung (Satz 2.37 von Seite 317) anwenden. Mit f ist auch R mindestens $(m+1)$ -mal stetig differenzierbar auf $]a, b[$. (Beim Einsatz von Satz 2.37 können wir sogar auf die Stetigkeit von $f^{(m+1)}$ verzichten.) Da die Taylor-Summe gerade so konstruiert ist, dass Funktionswert und alle Ableitungen bis zur Ordnung m an der Stelle x_0 mit den entsprechenden Werten von f übereinstimmen, haben wir $R^{(k)}(x_0) = 0$ für $0 \leq k \leq m$. Weiter ist $R^{(m+1)}(x) = f^{(m+1)}(x)$. Damit können wir den verallgemeinerten Mittelwertsatz iteriert anwenden. Für $x \neq x_0$ ist

$$\begin{aligned}
\frac{R(x)}{(x-x_0)^{m+1}} &= \frac{R(x) - \overbrace{R(x_0)}^{=0}}{(x-x_0)^{m+1} - \overbrace{(x_0-x_0)^{m+1}}^{=0}} = \frac{R'(\xi_1)}{(m+1)(\xi_1-x_0)^m} \\
&= \frac{R'(\xi_1) - \overbrace{R'(x_0)}^{=0}}{(m+1)(\xi_1-x_0)^m - \overbrace{(m+1)(x_0-x_0)^m}^{=0}} = \frac{R^{(2)}(\xi_2)}{m(m+1)(\xi_2-x_0)^{m-1}} \\
&= \dots = \frac{R^{(m+1)}(\xi_{m+1})}{(m+1)!} = \frac{f^{(m+1)}(\xi_{m+1})}{(m+1)!}.
\end{aligned}$$

Dabei liegen alle Zwischenstellen ξ_k zwischen x und x_0 . $\xi = \xi_{m+1}$ ist die Zwischenstelle, die im Satz angegeben ist. $\square$

Beispiel 2.116

Wir entwickeln mit dem Satz von Taylor einige Funktionen um $x_0 = 0$:

- a) Für $f(x) = e^x$ gilt $f'(x) = e^x$, $f''(x) = e^x$, ... Da alle Ableitungen an der Stelle $x_0 = 0$ gleich 1 sind, erhalten wir die Taylor-Entwicklung:

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^m}{m!} + R_{0,m+1,e^x}(x) \\ &= \sum_{k=0}^m \frac{x^k}{k!} + R_{0,m+1,e^x}(x), \end{aligned} \quad (2.70)$$

mit einer Stelle ξ zwischen 0 und x und dem Restglied von Lagrange

$$R_{0,m+1,e^x}(x) = \frac{e^\xi}{(m+1)!} x^{m+1}.$$

- b) Die trigonometrischen Funktionen besitzen die Taylor-Entwicklung

$$\begin{aligned} \cos x &= \cos 0 - \frac{\sin 0}{1!} x - \frac{\cos 0}{2!} x^2 + \frac{\sin 0}{3!} x^3 + \frac{\cos 0}{4!} x^4 \pm \cdots + \frac{\frac{d^{m+1} \cos}{dx^{m+1}}(\xi)}{(m+1)!} x^{m+1} \\ &= 1 - \frac{1}{2!} x^2 + \frac{1}{4!} x^4 - \frac{1}{6!} x^6 + \frac{1}{8!} x^8 \pm \cdots + \frac{\frac{d^{m+1} \cos}{dx^{m+1}}(\xi)}{(m+1)!} x^{m+1}, \\ \sin x &= \frac{1}{1!} x - \frac{1}{3!} x^3 + \frac{1}{5!} x^5 - \frac{1}{7!} x^7 \pm \cdots + \frac{\frac{d^{m+1} \sin}{dx^{m+1}}(\xi)}{(m+1)!} x^{m+1}. \end{aligned}$$

Diese Taylor-Entwicklungen sind auch in vielen Anwendungen hilfreich.

- c) Die Restglieddarstellungen ermöglichen Fehlerabschätzungen. Wir können beispielsweise mit der obigen Taylor-Entwicklung der Exponentialfunktion einen Näherungswert für die Zahl e mit vorgegebener Genauigkeit berechnen.

$$e = e^1 \approx \sum_{k=0}^m \frac{1}{k!}.$$

Dabei machen wir einen Fehler, der nach dem Lagrange'schen Restglied die Größe

$$R_{0,m+1,e^x}(x) = \frac{e^\xi}{(m+1)!} \leq \frac{e^1}{(m+1)!} < \frac{3}{(m+1)!}$$

hat. Soll der Fehler kleiner als 10^{-8} sein, so sorgen wir dafür, dass die rechte Seite kleiner als 10^{-8} ist:

$$R_{0,m+1,e^x}(x) < \frac{3}{(m+1)!} < 10^{-8} \implies (m+1)! > 3 \cdot 10^8.$$

Dies ist z. B. für $m \geq 11$ erfüllt, denn $(11+1)! = 479\,001\,600$. Damit können wir e^1 bis auf 8 Dezimalstellen genau mit der Summe $\sum_{k=0}^{11} \frac{1}{k!}$ angeben. ■

Für Potenzreihen (Kapitel 2.8) werden wir diese Summen unendlich fortsetzen und dabei beobachten, dass das Restglied gegen null geht. Damit kann man die Funktionswerte exakt über eine unendliche Summe und näherungsweise über eine endliche Summe ausrechnen.

Eine direkte Anwendung des Satzes von Taylor sind Kurvendiskussionen.

2.7.2 Kurvendiskussion und Extremalprobleme

Bei einer Kurvendiskussion sucht man die charakteristischen Kenngrößen einer Funktion, die einen schnellen Überblick über das Verhalten des Funktionsgraphen geben. Mit ihrer Hilfe kann man den Graphen auch ohne Berechnung weiterer Funktionswerte gut skizzieren. Einen ähnlichen Ansatz werden wir in der beschreibenden Statistik verfolgen, um eine große Datenmenge mittels weniger Kenngrößen zu beschreiben.

Unabhängig von der Skizzierung eines Funktionsgraphen sind Charakteristika wie lokale Maxima und Minima sehr häufig in Anwendungen gesucht.

Zunächst erhalten wir aus dem Mittelwertsatz (Satz 2.36 auf Seite 315) direkt eine Folgerung, die der Anschauung der Ableitung entspricht:

Folgerung 2.5 (Zusammenhang Monotonie und Ableitung)

Sei $f : [a, b] \rightarrow \mathbb{R}$ stetig auf $[a, b]$ und differenzierbar auf $]a, b[$.

- a) Ist $f'(x) = 0$ für alle $x \in]a, b[$, so ist f konstant: $f(x) = c$ für ein $c \in \mathbb{R}$.
- b) Ist $f'(x) \geq 0$ (bzw. $f'(x) > 0$) für alle $x \in]a, b[$, so ist f (streng) monoton steigend auf $[a, b]$.
- c) Ist $f'(x) \leq 0$ (bzw. $f'(x) < 0$) für alle $x \in]a, b[$, so ist f (streng) monoton fallend auf $[a, b]$.

Beweis: Sind $x_1, x_2 \in [a, b]$ mit $x_1 \neq x_2$, so ist $f(x_2) - f(x_1) = (x_2 - x_1)f'(\xi_{x_1, x_2})$.

- a) Hier ist $f(x_2) = f(x_1)$ für alle $x_1, x_2 \in [a, b]$, also ist f konstant.
- b) Für $x_2 > x_1$ ist $f(x_2) - f(x_1) \geq (>) 0$, also $f(x_2) \geq (>) f(x_1)$. Damit ist f (streng) monoton steigend.
- c) analog zu b). □

Beispiel 2.117

Wegen $\frac{d}{dx}e^x = e^x > 0$ erhalten wir nun auch über die Ableitung, dass e^x streng monoton steigend auf $\mathbb{R}$ ist. Als Umkehrfunktion ist nach Lemma 1.7 von Seite 86 auch $\ln x$ streng monoton steigend. Dies sehen wir auch an der Ableitung $\frac{d}{dx} \ln x = \frac{1}{x} > 0$, $x > 0$. ■

Ein sehr wichtiges Charakteristikum einer Funktion sind ihre **lokalen Extrema** (lokale Maxima und Minima). Wir weisen hier nochmals auf den Satz von Fermat 2.34 auf Seite 314 hin:

Hat f in x_0 ein lokales Extremum, dann ist (notwendigerweise) $f'(x_0) = 0$.

Neben Extremstellen sind auch Wendepunkte für Funktionsgraphen charakteristisch. Der Name **Wendepunkt** deutet an, dass der Funktionsgraph hier von einer Rechts- in eine Linkskrümmung oder umgekehrt wechselt:

Definition 2.38 (Krümmung)

Eine zweimal differenzierbare Funktion f heißt auf einem Intervall $]a, b[$ genau dann **rechtsgekrümmt** oder **konkav**, wenn $f''(x) \leq 0$ für alle $x \in]a, b[$. Sie heißt **linksgekrümmt** oder **konvex** genau dann, wenn $f''(x) \geq 0$ für alle $x \in]a, b[$.

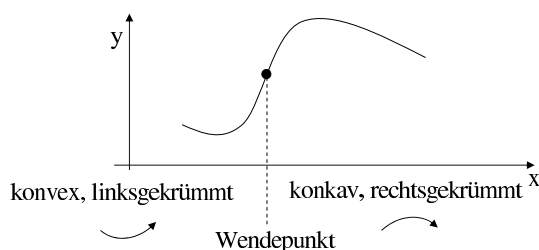


Abb. 2.36: Rechts- und Linkskrümmung

Fährt man entlang des Funktionsgraphen einer rechtsgekrümmten Funktion, so fährt man eine Rechtskurve (siehe Abbildung 2.36). Bei einer linksgekrümmten Funktion fährt man eine Linkskurve. Ist f zweimal stetig differenzierbar, so ist $f''(x_0) = 0$ bei einem Wendepunkt x_0 . Das ist (nach Zwischenwertsatz) notwendig, wenn sich das Vorzeichen der stetigen Funktion f'' in x_0 ändern soll:

Folgerung 2.6 (Notwendige Bedingung für einen Wendepunkt)

Besitzt eine zweimal stetig differenzierbare Funktion f in x_0 einen Wendepunkt, so folgt $f''(x_0) = 0$.

Hinreichende Kriterien für Extremal- und Wendepunkte ergeben sich aus dem Satz von Taylor:

Folgerung 2.7 (Kurvendiskussion)

Sei f m -mal stetig differenzierbar auf $]a, b[$, $m \geq 2$, und $x_0 \in]a, b[$. Außerdem sei $f'(x_0) = \dots = f^{(m-1)}(x_0) = 0$ und $f^{(m)}(x_0) \neq 0$. Dann gilt:

- a) Ist m gerade und $f^{(m)}(x_0) > 0$, so hat f ein striktes lokales Minimum in x_0 (hinreichende Bedingung für lokales Minimum).
 b) Ist m gerade und $f^{(m)}(x_0) < 0$, so hat f ein striktes lokales Maximum in x_0 (hinreichende Bedingung für lokales Maximum).
 c) Ist m ungerade, so hat f kein lokales Extremum in x_0 , sondern einen Wendepunkt (hinreichende Bedingung für einen Wendepunkt).

Ein Wendepunkt mit horizontaler Tangente heißt ein **Sattelpunkt**. Da in der Folgerung $f'(x_0) = 0$ verlangt ist, ist der Wendepunkt aus c) zugleich ein Sattelpunkt.

Beweis: Da alle Ableitungen bis zur $m-1$ -ten in x_0 gleich null sind, ergibt sich mit dem Satz von Taylor:

$$f(x) = f(x_0) + f^{(m)}(\xi(x)) \frac{(x - x_0)^m}{m!}$$

mit $\xi(x)$ zwischen x und x_0 . Für $x \rightarrow x_0$ geht $\xi(x) \rightarrow x_0$. Da $f^{(m)}$ stetig in x_0 ist, strebt $f^{(m)}(\xi(x)) \rightarrow f^{(m)}(x_0) \neq 0$ für $x \rightarrow x_0$. Außerdem hat die stetige Funktion $f^{(m)}$ in einer Umgebung von x_0 keine Nullstelle und verändert das Vorzeichen nicht (siehe Lemma 2.3 auf Seite 274). Für genügend kleines $h > 0$ erhalten wir:

- a) In x_0 liegt ein striktes lokales Minimum, da für gerades m mit $h^m = (-h)^m > 0$ gilt:

$$f(x_0 \pm h) = f(x_0) + \frac{1}{m!} \underbrace{f^{(m)}(\xi(x_0 \pm h))}_{>0} \underbrace{(\pm h)^m}_{>0} > f(x_0).$$

- b) In x_0 liegt ein striktes lokales Maximum, da für gerades m wieder $h^m = (-h)^m > 0$ ist und

$$f(x_0 \pm h) = f(x_0) + \frac{1}{m!} \underbrace{f^{(m)}(\xi(x_0 \pm h))}_{<0} \underbrace{(\pm h)^m}_{>0} < f(x_0).$$

- c) Wir betrachten für ungerades m den Fall $f^{(m)}(\xi(x_0 \pm h)) > 0$ (der Fall $f^{(m)}(\xi(x_0 \pm h)) < 0$ folgt analog):

$$f(x_0 + h) = f(x_0) + \frac{1}{m!} f^{(m)}(\xi(x_0 + h)) h^m > f(x_0),$$

aber

$$f(x_0 - h) = f(x_0) + \frac{1}{m!} f^{(m)}(\xi(x_0 - h)) \underbrace{(-h)^m}_{<0} < f(x_0).$$

Damit hat f kein lokales Extremum in x_0 . f besitzt hier aber einen Wendepunkt, da die Taylor-Entwicklung für die zweite Ableitung liefert:

$$f''(x_0 + h) = \frac{1}{(m-2)!} f^{(2+[m-2])}(\xi(f^{(2)}, x_0 + h)) \cdot h^{m-2} > 0,$$

$$f''(x_0 - h) = \frac{1}{(m-2)!} f^{(2+[m-2])}(\xi(f^{(2)}, x_0 - h)) \cdot (-h)^{m-2} < 0.$$

Da m ungerade ist, ist der letzte Faktor $(-h)^{m-2}$ negativ. □

In der Ingenieurpraxis spielt die Ermittlung von Extremwerten eine bedeutende Rolle. Das folgende Beispiel ist dazu ein Klassiker.

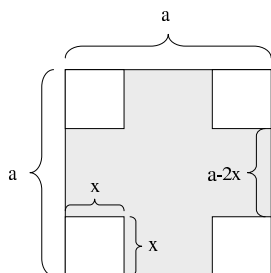


Abb. 2.37: Volumen einer Schachtel

Beispiel 2.118

Aus einem quadratischen Stück Pappe (oder anderen Materialien wie Blech, Stoff usw.) der Seitenlänge a werden an den Ecken Quadrate abgetrennt. Diese haben die Kantenlänge x (siehe Abbildung 2.37). Wir berechnen den Wert für x , für den durch Umknicken der Seitenflächen ein (oben offener) Quader mit maximalem Volumen entsteht. Das Volumen $f(x)$ der Quaders beträgt

$$f(x) = (a - 2x)^2 \cdot x, \quad 0 \leq x \leq \frac{a}{2}.$$

Zum Auffinden des Maximums suchen wir die Nullstellen der ersten Ableitung:

$$f'(x) = (a - 2x)^2 + x \cdot 2(a - 2x) \cdot (-2) = (a - 2x) \cdot (a - 6x) = 0 \quad \Longleftrightarrow \quad x = \frac{a}{2} \vee x = \frac{a}{6}.$$

Mit Hilfe der zweiten Ableitung

$$f''(x) = (a - 6x) \cdot (-2) + (a - 2x) \cdot (-6) = 24x - 8a$$

folgt $f''\left(\frac{a}{2}\right) = \frac{24}{2}a - 8a = 4a > 0$, d. h., bei $x = \frac{a}{2}$ liegt ein lokales Minimum und $f''\left(\frac{a}{6}\right) = \frac{24}{6}a - 8a = -4a < 0$, d. h., bei $x = \frac{a}{6}$ liegt ein lokales Maximum. Für $x = \frac{a}{6}$ wird das Volumen der Schachtel maximal. Es ist

$$f\left(\frac{a}{6}\right) = \left(a - 2 \cdot \frac{a}{6}\right)^2 \cdot \frac{a}{6} = \left(\frac{2}{3}a\right)^2 \cdot \frac{a}{6} = \frac{2}{27}a^3. \quad \blacksquare$$

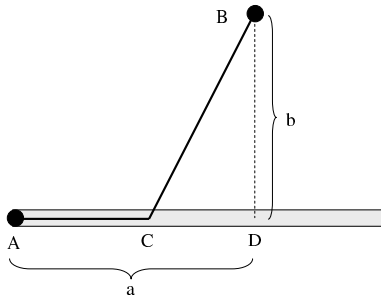


Abb. 2.38: Kommerziell optimale Verlegung einer Leitung

Beispiel 2.119

Vom Campus A aus soll zu einem Studentenwohnheim B eine Glasfaserleitung verlegt werden, siehe Abbildung 2.38, vgl. (Gellrich und Gellrich, 2003, Band 3, S. 218). Für die Lage der beiden Punkte gilt $B - A = (a, b) = (800, 200)$ mit Angaben in Metern. Die erforderlichen Grabungen kosten entlang der Straße 100 [Euro/m] und quer über das angrenzende Grundstück 120 [Euro/m]. An welcher Stelle C muss von der Straße geradlinig abgezweigt werden, wenn die Kosten minimal werden sollen?

Die Gesamtkosten in Euro belaufen sich auf

$$K = 100 \cdot \overline{AC} + 120 \cdot \overline{CB}.$$

Nun soll der Abzweigpunkt C so bestimmt werden, dass K minimal wird. Setzt man $\overline{CD} = x$ und misst man alle Strecken in Meter, so ist ein Minimum von

$$K(x) = 100(a - x) + 120\sqrt{b^2 + x^2}$$

unter der Nebenbedingung $0 \leq x \leq a$ zu bestimmen. Dabei sind a und b durch die Problemstellung gegeben. Es ist

$$K'(x) = -100 + 120 \frac{x}{\sqrt{b^2 + x^2}}$$

und

$$K''(x) = 120 \frac{\sqrt{b^2 + x^2} - x \cdot \frac{x}{\sqrt{b^2 + x^2}}}{b^2 + x^2} = \frac{120 \cdot b^2}{(b^2 + x^2)^{3/2}}.$$

Die Forderung $K'(x) = 0$ liefert für $x \geq 0$

$$\begin{aligned} 120 \frac{x}{\sqrt{b^2 + x^2}} &= 100 \iff 120^2 x^2 = 100^2 (b^2 + x^2) \\ \iff x &= \sqrt{\frac{20\,000^2}{120^2 - 100^2}} \approx 301,5. \end{aligned}$$

Man muss also $800 - 301,5 = 498,5$ Meter von A entfernt von der Straße geradlinig nach B abzweigen, um die Kosten so gering wie möglich zu halten. Dass es sich dabei um ein (lokales) Minimum handelt, sieht man sofort durch Einsetzen in $K''(x)$.

Folgende Vergleichsrechnungen zeigen das Sparpotenzial:

- Der direkte Weg von A nach B kostet $K = 120 \cdot \sqrt{800^2 + 200^2} \approx 98\,954,53$ Euro,
- der Umweg über D (siehe Abbildung 2.38) kostet $K = 800 \cdot 100 + 200 \cdot 120 = 104\,000$ Euro,
- und das berechnete Minimum beträgt $K = 498,5 \cdot 100 + \sqrt{301,5^2 + 200^2} \cdot 120 \approx 93\,266,50$ Euro.

Es macht also Sinn, sich hier Gedanken zu machen, da die Einsparpotenziale durchaus interessant sind. ■

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 $[a, b]$ nach Satz 2.26 auf Seite 281 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:

Satz 2.55 (Globale Extrema)

Sei f auf $\mathbb{R}$ differenzierbar mit existierenden Grenzwerten $g_1 = \lim_{x \rightarrow -\infty} f(x)$ und $g_2 = \lim_{x \rightarrow \infty} f(x)$ (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.

Beweis: 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 $\lim_{x \rightarrow -\infty} f(x) \leq M < M'$ 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 Grenzwerts $\lim_{x \rightarrow \infty} f(x) \leq M < M'$. Der Beweis für das globale Minimum geschieht analog. □

Beispiel 2.120

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 $\lim_{x \rightarrow \pm\infty} f(x) = 0$ 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. ■

Bei einer Kurvendiskussion versucht man, möglichst viele Eigenschaften eines Funktionsgraphen mittels der zuvor hergeleiteten Sätze zu bestimmen. Mit diesen Informationen kann man dann den Graphen leicht skizzieren. Um zusätzlich das Verhalten für $x \rightarrow \pm\infty$ noch genauer zu beschreiben, benötigen wir einen weiteren Begriff:

Definition 2.39 (Asymptote)

Sei $f : D \rightarrow \mathbb{R}$ mit $[a, \infty[\subset D$. Die Gerade $g(x) = mx + b$ heißt genau dann eine **Asymptote** an f für $x \rightarrow \infty$, wenn $\lim_{x \rightarrow \infty} (f(x) - g(x)) = 0$. Die Funktion f nähert sich also der Geraden g im Unendlichen an. Entsprechend definiert man für $f : D \rightarrow \mathbb{R}$, $]-\infty, a] \subset D$, eine Asymptote g an f für $x \rightarrow -\infty$ über $\lim_{x \rightarrow -\infty} (f(x) - g(x)) = 0$.

Falls eine Asymptote für $x \rightarrow \infty$ oder $x \rightarrow -\infty$ existiert, so ist sie eindeutig: Sind $g_1(x) = m_1x + b_1$ und $g_2(x) = m_2x + b_2$ zwei Asymptoten an f für z. B. $x \rightarrow \infty$, so gilt:

$$\lim_{x \rightarrow \infty} [(m_1 - m_2)x + (b_1 - b_2)] = \lim_{x \rightarrow \infty} [g_1(x) - f(x)] - [g_2(x) - f(x)] = 0 - 0 = 0.$$

Da der Grenzwert nicht $\pm\infty$ ist, muss $m_1 = m_2$ sein. Dann ist aber $b_1 - b_2 = 0$, die Geraden sind also identisch. Natürlich können sich die Asymptoten für $x \rightarrow \infty$ und $x \rightarrow -\infty$ unterscheiden.

Beispiel 2.121

a) Das asymptotische Verhalten der gebrochen-rationalen Funktion $f(x) = \frac{3x^2+2x+1}{2x+3}$

kann man mittels Polynomdivision ablesen: $f(x) = \frac{3}{2}x - \frac{5}{4} + \frac{19}{2x+3}$. Damit verhält sich f für sehr große und sehr kleine Werte von x annähernd wie $g(x) = \frac{3}{2}x - \frac{5}{4}$.

Tatsächlich ist $g(x)$ sowohl für $x \rightarrow \infty$ als auch für $x \rightarrow -\infty$ die Asymptote:

$$\lim_{x \rightarrow \pm\infty} \left[\frac{3x^2+2x+1}{2x+3} - \left(\frac{3}{2}x - \frac{5}{4} \right) \right] = \lim_{x \rightarrow \pm\infty} \frac{\frac{19}{4}}{2x+3} = 0.$$

b) Die Funktion $f(x) = x^2$ hat keine Asymptoten, denn $\lim_{x \rightarrow \pm\infty} [x^2 - (mx + b)] = \infty$. Ebenso hat $\exp(x)$ für $x \rightarrow \infty$ keine Asymptote, aber für $x \rightarrow -\infty$ ist die Asymptote $g(x) = 0$. ■

Tab. 2.4: Regeln für die Kurvendiskussion

Charakteristik	Definition/Kriterium
Symmetrie	$f(x) = f(-x)$ Achsensymmetrie (f gerade) $f(x) = -f(-x)$ Nullpunktsymmetrie (f ungerade)
Lokale Extremstellen	notwendig (Satz 2.34): $f'(x_0) = 0$ hinreichend (Folgerung 2.7): n gerade und $f'(x_0) = f''(x_0) = \dots = f^{(n-1)}(x_0) = 0, f^{(n)}(x_0) \neq 0$: $f^{(n)}(x_0) < 0 \implies x_0$ lokales Maximum $f^{(n)}(x_0) > 0 \implies x_0$ lokales Minimum
Monotonie	Folgerung 2.5: f monoton steigend, wenn $f'(x) \geq 0$ f monoton fallend, wenn $f'(x) \leq 0$
Wendepunkte	notwendig (Folgerung 2.6): $f''(x_0) = 0$ hinreichend (Folgerung 2.7): n ungerade und $f'(x_0) = f''(x_0) = \dots = f^{(n-1)}(x_0) = 0, f^{(n)}(x_0) \neq 0$.
Krümmung	Definition 2.38: $f''(x) \geq 0 \implies$ konvex (f' wächst) $f''(x) \leq 0 \implies$ konkav (f' fällt)
Asymptoten	$\lim_{x \rightarrow \pm\infty} (f(x) - (mx + b)) = 0$

Für die Untersuchung einer Funktion kann man sich neben der Analyse des Definitions- und Wertebereichs, der Nullstellen, der Stetigkeit und der Unstetigkeitsstellen sowie dem Grenzwertverhalten an den in der Tabelle 2.4 zusammengestellten Punkten orientieren.

Aus Symmetrien und dem Verhalten im Unendlichen (asymptotisches Verhalten) eines Graphen kann man häufig Maxima, Minima und Wendepunkte allein aus den notwendigen Bedingungen $f'(x) = 0$ und $f''(x) = 0$ ermitteln. Zum Beispiel liegt (gemäß Zwischenwertsatz für stetiges f'') zwischen Maximum (mit $f''(x_M) < 0$) und Minimum (mit $f''(x_m) > 0$) ein Wendepunkt.

Beispiel 2.122

Wir diskutieren die Funktion $f(x) = x^3 - 2x + 1, f : \mathbb{R} \rightarrow \mathbb{R}$.

- Definitions- und Wertebereich sind $\mathbb{R}$, Nullstellen der Funktion: Durch Einsetzen sieht man, dass 1 eine Nullstelle ist. Wir spalten den Linearfaktor $(x - 1)$ mittels Polynomdivision ab: $(x^3 - 2x + 1) : (x - 1) = x^2 + x - 1$. Mit der p-q-Formel erhalten wir die weiteren Nullstellen $-\frac{1}{2} \pm \sqrt{\frac{1}{4} + 1} = \frac{-1 \pm \sqrt{5}}{2}$.

- Symmetrie: Aufgrund der Lage der Nullstellen ist f weder gerade noch ungerade.
- Unstetigkeitsstellen: f ist stetig auf $\mathbb{R}$, es gibt keine Unstetigkeitsstellen.
- Lokale Extrema: Zur Bestimmung der Extremstellen berechnen wir zunächst die Ableitungen: $f'(x) = 3x^2 - 2$, $f^{(2)}(x) = 6x$, $f^{(3)}(x) = 6$, $f^{(k)}(x) = 0$, $k \geq 4$.
Nullstellen der Ableitung: $x^2 - \frac{2}{3} = 0$, also $x = \pm\sqrt{\frac{2}{3}}$.

$$f^{(2)}\left(\sqrt{\frac{2}{3}}\right) = 6 \cdot \sqrt{\frac{2}{3}} > 0, \quad f^{(2)}\left(-\sqrt{\frac{2}{3}}\right) = -6 \cdot \sqrt{\frac{2}{3}} < 0.$$

An der Stelle $-\sqrt{\frac{2}{3}}$ hat f ein lokales Maximum und bei $\sqrt{\frac{2}{3}}$ ein lokales Minimum.

- Monotonie: Zwischen den Nullstellen kann f' als stetige Funktion das Vorzeichen wegen des Zwischenwertsatzes nicht ändern. Ausrechnen jeweils eines Funktionswerts zeigt, dass $f'(x) > 0$ auf $]-\infty, -\sqrt{\frac{2}{3}}[$ und $]\sqrt{\frac{2}{3}}, \infty[$ und f damit dort streng monoton steigend ist. Auf $]-\sqrt{\frac{2}{3}}, \sqrt{\frac{2}{3}}[$ ist $f'(x) < 0$ und f streng monoton fallend.
- Wendepunkte: Die einzige Nullstelle der zweiten Ableitung ist 0. Nur hier kann ein Wendepunkt liegen. Da zwischen dem lokalen Maximum und dem lokalen Minimum ein Wendepunkt liegen muss, liegt er bei 0. Man beachte, dass $f'(0) = -2 \neq 0$. $f'(x_0) = 0$ ist also **keine** notwendige Bedingung für einen Wendepunkt in x_0 (nur für ein lokales Extremum).
- Krümmung: Da $f^{(2)}(x) < 0$, $x < 0$, und $f^{(2)}(x) > 0$, $x > 0$, ist f auf $]-\infty, 0[$ rechtsgekrümmt und auf $]0, \infty[$ linksgekrümmt.
- Grenzwerte im Unendlichen: $\lim_{x \rightarrow \infty} x^3 - 2x + 1 = \lim_{x \rightarrow \infty} x^3 = \infty$, $\lim_{x \rightarrow -\infty} x^3 - 2x + 1 = -\infty$. Damit gibt es insbesondere keine globale Extrema.
- Asymptoten: Wegen $\lim_{x \rightarrow \pm\infty} x^3 - 2x + 1 - mx - b = \pm\infty$ (unabhängig von m und b) existiert keine Gerade als Asymptote für $x \rightarrow \infty$ oder $x \rightarrow -\infty$.
- Funktionsgraph: Mit den Kenngrößen können wir den Funktionsgraphen recht genau skizzieren, siehe Abbildung 2.39.

■

Beispiel 2.123

Wir diskutieren die gebrochen-rationale Funktion $f(x) = \frac{x^3}{x^2-1} = x + \frac{x}{x^2-1}$.

- Definitionsbereich und Nullstellen: Die einzige (dreifache) Nullstelle ist offensichtlich 0. $D(f) = \mathbb{R} \setminus \{\pm 1\}$. Den Wertebereich werden wir später ablesen. Mittels Polynomdivision erhalten wir $f(x) = x + \frac{x}{x^2-1}$.
- Symmetrie: Es gilt $f(x) = -f(-x)$, denn es ist $\frac{x^3}{x^2-1} = \frac{-(-x)^3}{(-x)^2-1}$, d. h., die Funktion ist ungerade bzw. symmetrisch zum Nullpunkt.
- Stetigkeit: Unstetigkeitsstellen liegen bei -1 und 1 vor:

$$\lim_{x \rightarrow -1-} \frac{x^3}{x^2-1} = -\infty, \quad \lim_{x \rightarrow -1+} \frac{x^3}{x^2-1} = +\infty,$$

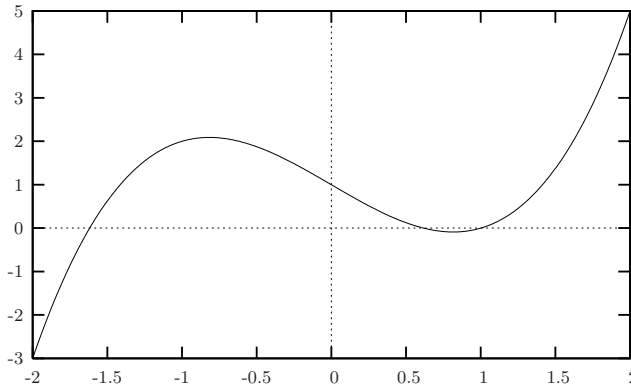


Abb. 2.39: $f(x) = x^3 - 2x + 1$

$$\lim_{x \rightarrow 1^-} \frac{x^3}{x^2 - 1} = -\infty, \quad \lim_{x \rightarrow 1^+} \frac{x^3}{x^2 - 1} = +\infty.$$

Es handelt sich also um Unstetigkeitsstellen zweiter Art (Pole). Da f auf $] -1, 1[$ stetig ist, ist $W(f) = \mathbb{R}$.

- **Lokale Extrema:** Wir berechnen zunächst die Ableitungen:

$$\begin{aligned} f'(x) &= \frac{3x^2(x^2 - 1) - 2x \cdot x^3}{(x^2 - 1)^2} = \frac{x^4 - 3x^2}{(x^2 - 1)^2} \\ f''(x) &= \frac{(4x^3 - 6x)(x^2 - 1)^2 - 2(x^2 - 1)2x(x^4 - 3x^2)}{(x^2 - 1)^4} = \frac{2x^3 + 6x}{(x^2 - 1)^3} \\ &= \frac{2x(x^2 + 3)}{(x^2 - 1)^3}. \end{aligned}$$

Die notwendige Bedingung für lokale Extrema führt zu

$$f'(x) = 0 \iff x^4 - 3x^2 = 0 \iff x^2(x^2 - 3) = 0.$$

Damit sind $x = 0$ und $x = \pm\sqrt{3}$ Kandidaten für lokale Extremstellen. Es ist $f''(0) = 0$, d. h., hier wissen wir zunächst noch nicht viel. Es ist $f''(\sqrt{3}) = \frac{2\sqrt{3}(3+3)}{(3-1)^2} > 0$, d. h., $(\sqrt{3}, \frac{3\sqrt{3}}{2})$ ist lokales Minimum, und aus Symmetriegründen ist $(-\sqrt{3}, \frac{-3\sqrt{3}}{2})$ lokales Maximum.

- **Monotonie:**

$$f'(x) = \frac{x^4 - 3x^2}{(x^2 - 1)^2} \begin{cases} > 0 & \text{für } x < -\sqrt{3}, \\ < 0 & \text{für } -\sqrt{3} < x < 0, x \neq -1, \\ < 0 & \text{für } 0 < x < \sqrt{3}, x \neq 1, \\ > 0 & \text{für } x > \sqrt{3}. \end{cases}$$

f ist also streng monoton steigend auf $] -\infty, \sqrt{3}]$ und auf $[\sqrt{3}, \infty[$. f ist streng monoton fallend auf $[-\sqrt{3}, -1[$, $] -1, 1[$ und $]1, \sqrt{3}]$.

- Wendepunkte: Aufgrund der notwendigen Bedingung $f''(x) = 0 \iff 2x(x^2 + 3) = 0 \iff x = 0$ kann nur bei $x = 0$ ein Wendepunkt vorliegen. Da f'' bei 0 einen Vorzeichenwechsel hat, ist dem auch so.

- Krümmung:

$$f''(x) = \frac{2x(x^2 + 3)}{(x^2 - 1)^3} \begin{cases} < 0 & \text{für } x < -1, \\ > 0 & \text{für } -1 < x < 0, \\ < 0 & \text{für } 0 < x < 1, \\ > 0 & \text{für } x > 1. \end{cases}$$

f ist linksgekrümmt auf $] -1, 0] \cup]1, \infty[$ und rechtsgekrümmt auf $] -\infty, -1[\cup]0, 1[$.

- Grenzwerte im Unendlichen: $\lim_{x \rightarrow \infty} \frac{x^3}{x^2 - 1} = +\infty$, $\lim_{x \rightarrow -\infty} \frac{x^3}{x^2 - 1} = -\infty$. Wieder gibt es keine globale Extrema.
- Asymptoten: Es ist $f(x) \approx x$ für $|x| \rightarrow \infty$:

$$\lim_{x \rightarrow \pm\infty} [f(x) - x] = \lim_{x \rightarrow \pm\infty} \frac{x}{x^2 - 1} = 0.$$

Weiterhin ist $f(x) > x$, für $x \rightarrow \infty$ d.h., der Graph nähert sich der Asymptote $y = x$ von oben. Für $x \rightarrow -\infty$ ist $f(x) < x$, d.h., der Graph nähert sich der Asymptote $y = x$ von unten.

- Funktionsgraph: siehe Abbildung 2.40. ■

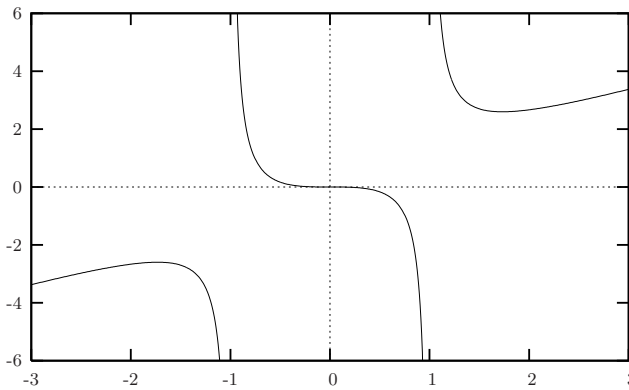


Abb. 2.40: $f(x) = \frac{x^3}{x^2 - 1}$

Beispiel 2.124

Die gedämpfte harmonische Schwingung wird beschrieben durch die Funktion

$$x(t) = e^{-\delta t} \cos(\omega t), \quad t \geq 0,$$

mit $\delta > 0$ und $\omega > 0$. Die Amplitude der Kosinusschwingung wird dabei durch den Faktor $e^{-\delta t}$ gedämpft. Sowohl in der Mechanik als auch in der Elektrotechnik hat man es häufig mit solchen Schwingungen zu tun, so dass wir sie uns später in Kapitel 5.4.3 noch wesentlich genauer ansehen. Wir diskutieren einige Eigenschaften des Graphen von $x(t)$ (siehe Abbildung 2.41):

- Definitionsbereich und Nullstellen: Zunächst gilt $D(f) = [0, \infty[$, $W(f) \subset]-1, 1[$. Die Nullstellen-Bedingung $x(t) = 0$ liefert $\underbrace{e^{-\delta t}}_{>0} \cos(\omega t) = 0 \iff \cos(\omega t) = 0$, d. h.

$$\omega t = (2k + 1)\frac{\pi}{2}, \quad k \in \mathbb{N}_0 \text{ bzw.}$$

$$t = (2k + 1)\frac{\pi}{2\omega}, \quad k \in \mathbb{N}_0.$$

Wegen $|\cos(\omega t)| \leq 1$ verläuft der Graph von $x(t)$ zwischen den Funktionsgraphen von $-e^{-\delta t}$ und $+e^{-\delta t}$. Insbesondere ist $\lim_{t \rightarrow \infty} x(t) = 0$. Der Graph von $x(t)$ berührt die beiden begrenzenden Graphen für $\cos(\omega t) = \pm 1 \iff \omega t = 2\pi k$ (für $\cos(\omega t) = 1$) bzw. $\omega t = 2\pi k + \pi$ (für $\cos(\omega t) = -1$).

- Symmetrie: Der Graph zeigt kein Symmetrieverhalten.
- Stetigkeit: x ist auf $[0, \infty[$ stetig.
- Lokale Extrema: Für die Suche nach Extremstellen leiten wir zunächst ab:

$$x'(t) = -\delta e^{-\delta t} \cos(\omega t) - \omega e^{-\delta t} \sin(\omega t) = -e^{-\delta t} [\delta \cos(\omega t) + \omega \sin(\omega t)].$$

Die notwendige Bedingung führt zu

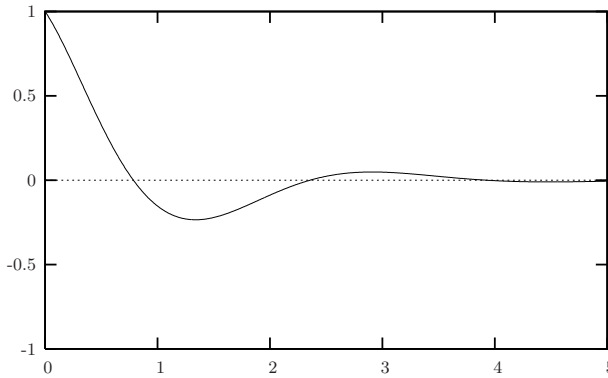
$$\begin{aligned} x'(t) = 0 &\iff \delta \cos(\omega t) + \omega \sin(\omega t) = 0 \\ &\iff \tan(\omega t) = -\frac{\delta}{\omega} \iff \omega t = \arctan\left(-\frac{\delta}{\omega}\right) + k\pi, \quad k \in \mathbb{N}_0, \\ &\iff t = -\frac{1}{\omega} \arctan\left(\frac{\delta}{\omega}\right) + \frac{k\pi}{\omega}, \quad k \in \mathbb{N}_0. \end{aligned}$$

Aufgrund des Schwingungsverhaltens von $\cos(\omega t)$ ist auf jedem Intervall der Länge $\frac{2\pi}{\omega}$ ein lokales Maximum und ein lokales Minimum zu erwarten. Damit liegen hier tatsächlich abwechselnd lokale Maxima und Minima. Jeweils dazwischen liegt ein Wendepunkt. ■

2.8 Potenzreihen

2.8.1 Unendliche Taylor-Summen und Potenzreihen

Mit dem Satz von Taylor haben wir Funktionen in Polynome entwickelt, die bis auf ein Restglied mit den Funktionen übereinstimmen. So können wir z. B. Funktionswerte

Abb. 2.41: $f(t) = e^{-t} \cos(2t)$

von $\exp(x)$ oder $\sin(x)$ bis zu einer vorgegebenen Genauigkeit berechnen. Nun eliminieren wir das Restglied und wollen die Funktionswerte exakt mit einer unendlichen Summe angeben. Das führt zu den Begriffen Taylor-Reihe und Potenzreihe.

Laut Satz von Taylor (Satz 2.54 auf Seite 384 mit $m = n$) kann man eine $(n+1)$ -mal (stetig) differenzierbare Funktion $f :]a, b[\rightarrow \mathbb{R}$ um ein $x_0 \in]a, b[$ entwickeln als

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi),$$

wobei ξ eine von n , f , x und x_0 abhängige Stelle zwischen x und x_0 ist.

Betrachtet man nun den Fall $n \rightarrow \infty$ für eine beliebig oft differenzierbare Funktion f , so erhält man die **Taylor-Reihe**

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k.$$

Der wichtige Spezialfall $x_0 = 0$ wird auch **MacLaurin-Reihe** genannt.

Ob eine Taylor-Reihe an einer Stelle x konvergiert, hängt davon ab, ob die Restglieder $R_{x_0, n+1, f}(x) := \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$ für $n \rightarrow \infty$ gegen null streben (siehe (2.69) auf Seite 384).

Beispiel 2.125 (Wichtige Taylor-Reihen)

Wir greifen die Taylor-Entwicklungen aus Beispiel 2.116 von Seite 386 auf:

- a) Entwickelt um $x_0 = 0$ hat die Funktion $f(x) = e^x$ die Taylor- bzw. MacLaurin-Darstellung

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

mit dem Restglied von Lagrange

$$R_{0, n+1, e^x}(x) = \frac{e^{\xi}}{(n+1)!} x^{n+1}$$

für eine Stelle ξ zwischen 0 und x . Wir zeigen nun $\lim_{n \rightarrow \infty} R_{0,n+1,e^x}(x) = 0$ für alle $x \in \mathbb{R}$. Wegen

$$a_{n+1} := \frac{|x|^{n+1}}{(n+1)!} = \frac{|x|^n}{n!} \cdot \frac{|x|}{n+1} = a_n \cdot \frac{|x|}{n+1}$$

nimmt a_{n+1} monoton ab, sobald $n+1 > |x|$ ist. Ab $n \geq |x|$ ist damit $(a_n)_{n=1}^\infty$ eine monoton fallende Folge nicht-negativer Zahlen, die damit nach dem Monotoniekriterium konvergiert (siehe Seite 224). Mit $a := \lim_{n \rightarrow \infty} a_n$ folgt

$$a = \lim_{n \rightarrow \infty} a_{n+1} = \left(\lim_{n \rightarrow \infty} a_n \right) \cdot \left(\lim_{n \rightarrow \infty} \frac{|x|}{n+1} \right) = a \cdot 0 = 0.$$

Wegen $|\xi| \leq |x|$ folgt $|R_{0,n+1,e^x}(x)| = a_{n+1}e^\xi \leq a_{n+1}e^{|x|}$, so dass das Restglied für $n \rightarrow \infty$ gegen null konvergiert. Damit ist die Konvergenz der **Exponentialreihe**

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

für jeden Wert $x \in \mathbb{R}$ bewiesen. Insbesondere erhalten wir für $x = 1$ die Konvergenz der Reihe $e = \exp(1) = \sum_{k=0}^{\infty} \frac{1}{k!}$, die wir (siehe (1.8) auf Seite 52 und (2.3) auf Seite 226) zur Definition von e vorausgesetzt haben.

b) Wir haben bereits $f(x) = \sin x$ an der Stelle $x_0 = 0$ mit dem Satz von Taylor dargestellt als

$$\sin x = \frac{1}{1!}x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 \pm \cdots + \frac{\frac{d^{n+1}\sin}{dx^{n+1}}(\xi)}{(n+1)!}x^{n+1}.$$

Mit der Folge $(a_n)_{n=1}^\infty$ aus a) ist auch hier

$$\lim_{n \rightarrow \infty} |R_{0,n+1,\sin x}(x)| = \lim_{n \rightarrow \infty} \left| \frac{d^{n+1}\sin}{dx^{n+1}}(\xi) \right| a_{n+1} \leq \lim_{n \rightarrow \infty} a_{n+1} = 0$$

für jedes einzelne $x \in \mathbb{R}$. Damit hat der Sinus eine Darstellung als Taylor-Reihe

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \pm \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}, \quad x \in \mathbb{R}.$$

Analog erhalten wir eine Taylor-Reihe des Kosinus:

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} \pm \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}, \quad x \in \mathbb{R}. \quad \blacksquare$$

Beispiel 2.126 (Taylor-Reihen zu geraden und ungeraden Funktionen)

Entwickelt man eine beliebig oft differenzierbare, gerade Funktion wie den Kosinus um $x_0 = 0$, so kommen in der Reihe nur Terme mit geraden Potenzen vor: Das liegt daran, dass die Ableitungen abwechselnd ungerade und gerade Funktionen sind (siehe Seite 299). Die Ableitungen mit ungradzahliger Ordnung sind ungerade und haben deshalb an der Stelle 0 den Funktionswert 0, so dass die ungradzahligen Summanden verschwinden. Das gilt auch, wenn eine Funktion mit der Eigenschaft $f(x_0 - x) = f(x_0 + x)$ (der Graph von f ist an der Gerade $x = x_0$ gespiegelt) an einer von null verschiedenen Stelle x_0 entwickelt wird. Sofern das Restglied gegen null strebt, ist

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(2k)}(x_0)}{(2k)!} (x - x_0)^{2k}.$$

Entsprechend treten bei einer ungeraden Funktion wie Sinus in der Taylor-Reihe zu $x_0 = 0$ oder allgemeiner bei einer zum Punkt x_0 symmetrischen Funktion ($f(x_0 + x) = -f(x_0 - x)$) in der Taylor-Reihe zu x_0 nur ungerade Exponenten auf:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(2k+1)}(x_0)}{(2k+1)!} (x - x_0)^{2k+1}.$$

■

Jede Taylor-Reihe hat mit den speziellen Koeffizienten $a_k = \frac{f^{(k)}(x_0)}{k!}$ die Gestalt

$$\sum_{k=0}^{\infty} a_k (x - x_0)^k. \quad (2.71)$$

Jetzt lösen wir uns (zunächst) von der konkreten Wahl der Koeffizienten und diskutieren allgemein das Verhalten von Reihen (2.71) mit beliebigen Koeffizienten $a_k \in \mathbb{R}$.

Definition 2.40 (Potenzreihe)

Seien $x_0 \in \mathbb{R}$ und die Folge $(a_k)_{k=0}^{\infty}$ reeller Zahlen gegeben. Die von x abhängende Reihe

$$\sum_{k=0}^{\infty} a_k (x - x_0)^k := \left(\sum_{k=0}^n a_k (x - x_0)^k \right)_{n=0}^{\infty}$$

heißt **Potenzreihe** mit **Entwicklungsmittelpunkt** x_0 und den Koeffizienten a_k .

Mit dem Symbol $\sum_{k=0}^{\infty} a_k (x - x_0)^k$ bezeichnen wir auch die Grenzfunktion, die für die Punkte x erklärt ist, in denen die Potenzreihe konvergiert.

Man beachte, dass eine Potenzreihe für jeden Wert der Variable x eine gewöhnliche Zahlenreihe ist und damit alle Sätze benutzt werden können, die wir für Zahlenreihen kennengelernt haben. Andererseits ist jede Partialsumme $S_n(x) := \sum_{k=0}^n a_k (x - x_0)^k$ eine Funktion, so dass wir eine Potenzreihe auch als Folge von Funktionen $(S_n)_{n=1}^{\infty}$ auffassen können.

Die im Beispiel 2.29 auf Seite 249 beschriebene Z-Transformation ordnet einer Zahlenfolge $(a_k)_{k=0}^{\infty}$ eine Potenzreihe mit Entwicklungsmittelpunkt 0 zu, in die für x die komplexe Zahl $1/z$ eingesetzt wird.

2.8.2 Einschub: Funktionenfolgen*

Dieses Unterkapitel widmet sich der Konvergenz von Folgen, deren Glieder Funktionen sind. Außerdem wird in diesem Zusammenhang die Vertauschung von Grenzwerten diskutiert. Der Inhalt kann ausgelassen werden, ohne dass dadurch das weitere Verständnis deutlich eingeschränkt wird. Andererseits wird der Begriff der gleichmäßigen Konvergenz eingeführt, mit dem tiefere Einblicke in Potenz- und Fourier-Reihen möglich werden. Das Nichtvorhandensein von gleichmäßiger Konvergenz führt z. B. zu Artefakten in Bildern, die im JPEG-Format gespeichert werden. Ähnliche Effekte treten in der Signalverarbeitung auf.

Definition 2.41 (Funktionenfolge)

Ist $(f_n)_{n=1}^{\infty}$ eine Folge, deren Glieder $f_n = f_n(x)$ reell-(oder komplex-)wertige Funktionen mit gemeinsamen Definitionsbereich $D \subset \mathbb{R}$ sind, so heißt $(f_n)_{n=1}^{\infty}$ eine **Funktionenfolge** auf D . Für jedes $x_0 \in D$ ist $(f_n(x_0))_{n=1}^{\infty}$ eine Zahlenfolge.

Die Partialsummen einer Potenzreihe bilden also eine Funktionenfolge. Wie bei den Potenzreihen wird sie für jeden konkreten Zahlenwert für x zu einer Zahlenfolge.

Definition 2.42 (Punktweise Konvergenz)

Eine Funktionenfolge $(f_n)_{n=1}^{\infty}$ auf D heißt genau dann **konvergent** im Punkt $x_0 \in D$, wenn die Zahlenfolge $(f_n(x_0))_{n=1}^{\infty}$ konvergent ist. Ist $(f_n(x_0))_{n=1}^{\infty}$ konvergent für jeden einzelnen Punkt $x_0 \in D$, so heißt die Funktionenfolge **punktweise konvergent** auf D . Dann heißt eine Funktion f mit Definitionsbereich D und $f(x) := \lim_{n \rightarrow \infty} f_n(x)$ für alle $x \in D$ die **Grenzfunktion** von $(f_n)_{n=1}^{\infty}$.

Wir hoffen, dass die Grenzfunktion der Partialsummen einer Potenzreihe genau die Ausgangsfunktion ist, zu der die Potenzreihe über Auswertung der Ableitungen berechnet wurde. Dies sehen wir uns im nächsten Abschnitt genauer an.

Häufig möchte man aus Eigenschaften der Funktionen der Funktionenfolge auf Eigenschaften der Grenzfunktion schließen. Hier reicht in der Regel die punktweise Konvergenz aber nicht aus.

Beispiel 2.127 (Klassische Gegenbeispiele)

- a) Die Funktionenfolge $(f_n)_{n=1}^\infty$ mit $f_n(x) := x^n$ ist auf $[0, 1]$ punktweise konvergent gegen

$$f(x) = \begin{cases} 0, & x \in [0, 1[, \\ 1, & x = 1. \end{cases} \quad (2.72)$$

Obwohl alle f_n stetig auf $[0, 1]$ sind, gilt dies für die Grenzfunktion nicht. Anders ausgedrückt: Die beiden folgenden Grenzwerte lassen sich nicht vertauschen:

$$0 = \lim_{x \rightarrow 1-} \left[\lim_{n \rightarrow \infty} f_n(x) \right] \neq \lim_{n \rightarrow \infty} \left[\lim_{x \rightarrow 1-} f_n(x) \right] = 1.$$

- b) Zur Vertauschbarkeit von Grenzwert der Funktionenfolge und Ableitung betrachten wir die Folge $(f_n)_{n=1}^\infty$,

$$f_n(x) = \frac{\sin(nx)}{\sqrt{n}}, \quad (2.73)$$

auf $\mathbb{R}$. Diese konvergiert auf $\mathbb{R}$ punktweise gegen $f(x) = 0$, denn für jedes $x \in \mathbb{R}$ ist $|f_n(x)| \leq 1/\sqrt{n} \rightarrow 0$, $n \rightarrow \infty$. Die Folge der Ableitungen $f'_n(x) = \frac{n \cos(nx)}{\sqrt{n}} = \sqrt{n} \cos(nx)$ ist jedoch nicht punktweise konvergent auf $\mathbb{R}$. Beispielsweise gilt für $x_0 = 0$: $\lim_{n \rightarrow \infty} f'_n(0) = \lim_{n \rightarrow \infty} \sqrt{n} = \infty$. Damit ist

$$\lim_{n \rightarrow \infty} \left[\frac{d}{dx} f_n(x) \right] \neq \frac{d}{dx} \left[\lim_{n \rightarrow \infty} f_n(x) \right] = 0.$$

- c) Mittels zweimaliger Anwendung des Satzes von L'Hospital (siehe Satz 2.41 auf Seite 325) können wir die Grenzfunktion der Funktionenfolge $(f_n)_{n=1}^\infty$ mit

$$f_n(x) := n^2 x(1-x)^n$$

auf $[0, 1]$ bestimmen. Für $x \in]0, 1[$ ist

$$\begin{aligned} \lim_{y \rightarrow \infty} y^2 x(1-x)^y &= x \lim_{y \rightarrow \infty} \frac{y^2}{(1-x)^{-y}} = x \lim_{y \rightarrow \infty} \frac{y^2}{\exp(y \ln(\frac{1}{1-x}))} \\ &= 2x \lim_{y \rightarrow \infty} \frac{y}{\ln(\frac{1}{1-x}) \exp(y \ln(\frac{1}{1-x}))} = \frac{2x}{(\ln(\frac{1}{1-x}))^2} \lim_{y \rightarrow \infty} \frac{1}{\exp(y \ln(\frac{1}{1-x}))} = 0. \end{aligned}$$

Da $f_n(0) = f_n(1) = 0$ ist, ist die Grenzfunktion $f(x) = 0$ für $x \in [0, 1]$ und $\int_0^1 f(x) dx = 0$. Andererseits erhalten wir jedoch mittels partieller Integration

$$\begin{aligned} \int_0^1 f_n(x) dx &= n^2 \left(\left[-x \frac{(1-x)^{n+1}}{n+1} \right]_0^1 + \frac{1}{n+1} \int_0^1 (1-x)^{n+1} dx \right) \\ &= \frac{n^2}{n+1} \int_0^1 u^{n+1} du = \frac{n^2}{n+1} \left[\frac{1}{n+2} u^{n+2} \right]_0^1 = \frac{n^2}{(n+1)(n+2)}. \end{aligned}$$

Damit ist aber

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx = \lim_{n \rightarrow \infty} \frac{n^2}{n^2 + 3n + 2} = 1 \neq 0.$$

Die Integrale der Folgenglieder konvergieren nicht gegen das Integral der Grenzfunktion:

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx \neq \int_0^1 \left[\lim_{n \rightarrow \infty} f_n(x) \right] dx. \quad \blacksquare$$

In der Ingenieurmathematik werden Grenzwerte in der Regel ohne Überprüfung der Voraussetzungen vertauscht. Wie wir an den vorangehenden Beispielen sehen, kann das aber ins Auge gehen. Man benötigt eine stärkere Voraussetzung als die punktweise Konvergenz:

Definition 2.43 (Gleichmäßige Konvergenz)

Eine Funktionenfolge $(f_n)_{n=1}^\infty$ auf D heißt genau dann **gleichmäßig konvergent** auf D gegen eine Grenzfunktion f , wenn zu jedem (noch so kleinen) $\varepsilon > 0$ ein $n_0 = n_0(\varepsilon)$ (unabhängig von $x \in D$) existiert, so dass $|f_n(x) - f(x)| < \varepsilon$ für alle $n > n_0$ und $x \in D$ gilt.

Offensichtlich folgt aus der gleichmäßigen die punktweise Konvergenz, bei der n_0 abhängig von x gewählt werden darf.

Gleichmäßige Konvergenz ist dazu äquivalent, dass man unabhängig von x eine positive Nullfolge $(a_n)_{n=1}^\infty$, $\lim_{n \rightarrow \infty} a_n = 0$, angeben kann mit $|f_n(x) - f(x)| < a_n$. Es gibt also unabhängig von x eine gewisse Mindestkonvergenzgeschwindigkeit a_n , mit der sich die Funktionenfolge der Grenzfunktion nähert. Bei punktweiser Konvergenz muss das nicht so sein, hier kann die Konvergenzgeschwindigkeit in Abhängigkeit von x beliebig langsam werden. Ein entsprechendes Beispiel sehen wir für Fourier-Reihen in Kapitel 6.1.6 (Gibbs-Phänomen). Dort konvergiert eine Fourier-Reihe punktweise, aber nicht gleichmäßig.

Um die gleichmäßige Konvergenz zu illustrieren, betrachten wir eine Läufergruppe mit m Joggern $x_1, \dots, x_m$, die gemeinsam einen Marathon bestreitet. Für die Strecke gibt es kein Zeitlimit. Nach n Sekunden haben die Jogger $f_n(x_1), \dots, f_n(x_m)$ Meter zurückgelegt. Zu jedem $n \in \mathbb{N}$ gibt die Funktion f_n also an, wie weit die Jogger gekommen sind. Wenn keiner vorzeitig aufgibt, gilt $\lim_{n \rightarrow \infty} f_n(x_k) = 42,195$ für $1 \leq k \leq m$. Als Grenzwert erhält man also für jeden Läufer die Gesamtdistanz von 42,195 Kilometern. Betrachtet man als Zwischenmarke die Halbmarathondistanz, so gibt es für jeden Läufer x_k ein $n_{0,k}$, so dass er nach $n_{0,k}$ Sekunden diese Marke überquert hat, d. h. $|42\,195 - f_n(x_k)| < 21\,097,5$ für $n > n_{0,k}$. Bilden wir nun das Maximum n_0 über die m Werte $n_{0,k}$, $1 \leq k \leq m$, so sind nach n_0 Sekunden alle Läufer näher am Ziel als die Halbmarathondistanz. Wählt man einen anderen Abstand ε zum Ziel, so sind nach einer Zeit $n_0(\varepsilon)$ alle Läufer näher als ε am Ziel. Die auf dem Definitionsbereich

$\{x_1, x_2, \dots, x_m\}$ erklärte Funktionenfolge $(f_n)_{n=1}^\infty$ konvergiert gleichmäßig gegen die konstante Grenzfunktion $f(x) = 42,195$.

Jetzt vergrößern wir in einem Gedankenexperiment die Läufergruppe zu einer unendlichen Menge $\{x_k : k \in \mathbb{N}\}$. Dann kann man nicht mehr wie zuvor das Maximum der Zeiten für eine Zwischenmarke bilden. Zwar passiert jeder Läufer nach einer endlichen Zeit $n_{0,k}$ eine Marke, aber wenn beispielsweise $n_{0,k} = 3600 + k$ gilt, so passiert der erste Läufer x_1 die Marke nach gut einer Stunde, aber für $k \rightarrow \infty$ gehen auch die Zeiten gegen Unendlich, so dass man keinen Zeitpunkt finden kann, nach dem alle Läufer die Zwischenmarke überquert haben. Auf diesem unendlich großen Definitionsbereich ist die Funktionenfolge $(f_n)_{n=1}^\infty$ nun nicht mehr gleichmäßig konvergent.

Beispiel 2.128

Die Folge $(f_n)_{n=1}^\infty$ mit $f_n(x) := \sin(x) + \frac{\sin(nx)}{n}$ ist gleichmäßig konvergent gegen $f(x) = \sin x$:

$$|f_n(x) - f(x)| = \frac{|\sin(nx)|}{n} \leq \frac{1}{n} \rightarrow 0, n \rightarrow \infty,$$

unabhängig von x . ■

Hat man eine Funktionenfolge $(g_n)_{n=1}^\infty$, so erhält man durch Bildung von Partialsummen eine **Funktionenreihe** $(f_n)_{n=1}^\infty$ mit $f_n(x) := \sum_{k=1}^n g_k(x)$. Sowohl die Funktionenreihe als auch ihr möglicher Grenzwert werden wie bei Zahlenreihen und Potenzreihen mit $\sum_{k=1}^\infty g_k(x)$ bezeichnet. Eine Potenzreihe ist eine Funktionenreihe, die über die Funktionenfolge $(a_k(x - x_0)^k)_{k=0}^\infty$ gebildet wird. So wie Reihen per Definition nichts anderes als Folgen sind, sind auch Funktionenreihen nichts anderes als Funktionenfolgen. Um die gleichmäßige Konvergenz einer Funktionenreihe nachzuweisen, lässt sich das Majorantenkriterium für Reihen (siehe Seite 246) übertragen:

Lemma 2.12 (Majorantenkriterium für gleichmäßige Konvergenz)

Sei $(f_n)_{n=1}^\infty = (\sum_{k=1}^n g_k)_{n=1}^\infty$ auf D eine Funktionenreihe mit $|g_k(x)| \leq a_k$ für alle $x \in D$. Falls nun die Zahlenreihe (Majorante) $\sum_{k=1}^\infty a_k$ (absolut) konvergent ist, so ist $(f_n)_{n=1}^\infty$ gleichmäßig konvergent auf D .

Beweis: Für jedes $x \in D$ hat die Zahlenreihe $\sum_{k=1}^n g_k(x)$ die angegebene konvergente Majorante und konvergiert damit nach dem Majorantenkriterium für Zahlenreihen (Satz 2.15) gegen einen Grenzwert $f(x)$. Damit ist eine Grenzfunktion $f(x)$ für jedes $x \in D$ erklärt, gegen die die Funktionenreihe punktweise konvergiert.

Wegen der Konvergenz der Majorante gibt es zu jedem $\varepsilon > 0$ eine Stelle $n_0 \in \mathbb{N}$, so dass für $n > n_0$ gilt:

$$\left| \sum_{k=1}^\infty a_k - \sum_{k=1}^n a_k \right| = \sum_{k=n+1}^\infty a_k < \varepsilon.$$

Damit erhalten wir die gleichmäßige Konvergenz gegen die Grenzfunktion f , denn unabhängig von $x \in D$ ist für $n > n_0$:

$$\begin{aligned} \left| \sum_{k=1}^n g_k(x) - f(x) \right| &= \left| \sum_{k=1}^n g_k(x) - \sum_{k=1}^{\infty} g_k(x) \right| = \left| \sum_{k=n+1}^{\infty} g_k(x) \right| \\ &\leq \sum_{k=n+1}^{\infty} |g_k(x)| \leq \sum_{k=n+1}^{\infty} a_k < \varepsilon. \end{aligned} \quad \square$$

Beispiel 2.129

Eine **Weierstraß-Funktion** (siehe Abbildung 2.42) ist definiert als

$$f(x) := \lim_{n \rightarrow \infty} \sum_{k=1}^n \left(\frac{2}{3}\right)^k \sin(2^k x). \quad (2.74)$$

Diese Funktionenreihe ist gleichmäßig konvergent gegen die Grenzfunktion f , da die Majorante $\sum_{k=1}^{\infty} \left(\frac{2}{3}\right)^k \stackrel{(2.12)}{=} \frac{\frac{2}{3}}{1-\frac{2}{3}} = 2$ als geometrische Reihe konvergiert. ■

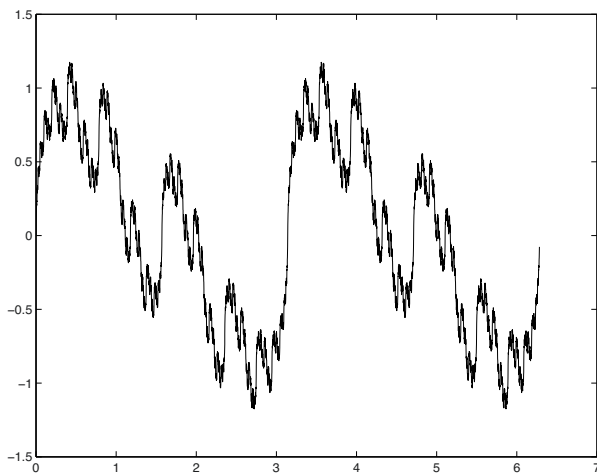


Abb. 2.42: Weierstraß-Funktion $\sum_{k=1}^{\infty} \left(\frac{2}{3}\right)^k \sin(2^k x)$

Jetzt zeigen wir, dass sich bei gleichmäßiger Konvergenz wichtige Eigenschaften der Folgenglieder auf die Grenzfunktion übertragen und man tatsächlich Grenzwerte vertauschen kann.

Satz 2.56 (Stetigkeit der Grenzfunktion)

Sei $(f_n)_{n=1}^{\infty}$ eine gleichmäßig konvergente Funktionenfolge auf D , die gegen f konvergiert. Ist jedes f_n in einem Punkt $x_0 \in D$ stetig, so ist auch die Grenzfunktion f stetig in x_0 .

Ist z. B. x_0 innerer Punkt von D , so dürfen als Konsequenz die folgenden Grenzwerte vertauscht werden:

$$\lim_{x \rightarrow x_0} \underbrace{\left[\lim_{n \rightarrow \infty} f_n(x) \right]}_{=f(x)} = \lim_{n \rightarrow \infty} \underbrace{\left[\lim_{x \rightarrow x_0} f_n(x) \right]}_{=f_n(x_0)} = \lim_{n \rightarrow \infty} f_n(x_0) = f(x_0).$$

Die Weierstraß-Funktion (2.74) aus dem vorangehenden Beispiel ist als Grenzfunktion einer gleichmäßig konvergenten Reihe stetiger Funktionen stetig.

Beweis: Wir zeigen die Stetigkeit der Grenzfunktion f , indem wir von f zu einem Glied f_{n_0+1} der Funktionenfolge übergehen, das aufgrund der gleichmäßigen Konvergenz nah genug bei f liegt. Dann können wir die Stetigkeit von f_{n_0+1} ausnutzen:

Sei $\varepsilon > 0$. Da die Funktionenfolge gleichmäßig konvergent ist, gibt es zu $\varepsilon/3$ ein $n_0 \in \mathbb{N}$, so dass für $n > n_0$ und alle $x \in D$ gilt: $|f_n(x) - f(x)| < \varepsilon/3$. Insbesondere gilt dies speziell für $n = n_0 + 1$. Nach Voraussetzung ist die Funktion f_{n_0+1} stetig in x_0 , d. h., zu $\varepsilon/3$ existiert ein $\delta > 0$, so dass für jedes $x \in D$ mit $|x - x_0| < \delta$ gilt: $|f_{n_0+1}(x_0) - f_{n_0+1}(x)| < \varepsilon/3$. Wir setzen nun für alle $x \in D$ mit $|x - x_0| < \delta$ die beiden Abschätzungen mittels Dreiecksungleichung zusammen:

$$\begin{aligned} |f(x) - f(x_0)| &= |f(x) - f_{n_0+1}(x) - f(x_0) + f_{n_0+1}(x_0) + f_{n_0+1}(x) - f_{n_0+1}(x_0)| \\ &\leq |f(x) - f_{n_0+1}(x)| + |f(x_0) - f_{n_0+1}(x_0)| + |f_{n_0+1}(x) - f_{n_0+1}(x_0)| \leq 3 \cdot \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Damit ist f stetig auf D . □

Die Folge $(f_n)_{n=1}^\infty$ mit stetigen Gliedern $f_n(x) := x^n$ aus dem obigen Beispiel konvergiert gegen eine unstetige Grenzfunktion (2.72) und kann daher nicht gleichmäßig konvergent sein. Mit der Folge $x_n = 1 - \frac{1}{n}$ erhalten wir unter Verwendung des Grenzwerts (2.40) von Seite 327:

$$\lim_{n \rightarrow \infty} |f_n(x_n) - f(x_n)| = \lim_{n \rightarrow \infty} x_n^n = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n \stackrel{(2.40)}{=} e^{-1}.$$

Es gibt also keine Nullfolge $(a_n)_{n=1}^\infty$ mit $|f_n(x) - f(x)| < a_n$ unabhängig von x .

Satz 2.57 (Differenzierbarkeit der Grenzfunktion)

Sei $(f_n)_{n=1}^\infty$ eine auf $[a, b]$ gleichmäßig gegen f konvergente Funktionenfolge stetig differenzierbarer Funktionen. Außerdem sei die Folge $(f'_n)_{n=1}^\infty$ ebenfalls auf $[a, b]$ gleichmäßig konvergent gegen eine Funktion g . Dann ist f stetig differenzierbar mit $f' = g$. Für alle $x \in [a, b]$ gilt also:

$$\frac{d}{dx} \lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} \frac{d}{dx} f_n(x).$$

Beweis: Wir zeigen $f' = g$ mit der Definition der Ableitung. Sei dazu $\varepsilon > 0$ beliebig und $x_0 \in [a, b]$ fest. Wir müssen zeigen, dass es ein $\delta = \delta(\varepsilon, x_0)$ gibt, so dass

$$\left| \frac{f(x_0 + h) - f(x_0)}{h} - g(x_0) \right| \leq \varepsilon \quad (2.75)$$

für jedes $h \neq 0$ mit $|h| < \delta$ und $x_0 + h \in [a, b]$ gilt.

Zunächst wählen wir das δ mittels der Stetigkeit von g : Nach Voraussetzung sind die Funktionen f'_n stetig, so dass auch die Grenzfunktion g auf $[a, b]$ nach Satz 2.56 stetig ist. Damit existiert ein $\delta > 0$, so dass

$$|g(x) - g(x_0)| < \varepsilon/4 \quad (2.76)$$

für alle $x \in [a, b]$ mit $|x - x_0| < \delta$ ist.

Wir zeigen jetzt, dass (2.75) nun für jedes $h \neq 0$ mit $x_0 + h \in [a, b]$ und $|h|$ kleiner als das soeben gewählte δ gilt. Dazu nutzen wir die Definitionen aller Voraussetzungen aus und schachteln sie mittels Dreiecksungleichung zusammen, so wie wir es auch im vorangehenden Beweis getan haben. Wegen der gleichmäßigen Konvergenz der Folge $(f_n)_{n=1}^\infty$ existiert ein $n_1 = n_1(\varepsilon, h) \in \mathbb{N}$, so dass für $n > n_1$ und alle $x \in [a, b]$:

$$|f_n(x) - f(x)| < |h| \cdot \frac{\varepsilon}{4}. \quad (2.77)$$

Zu verschiedenen Werten für h erhält man eventuell auch verschiedene n_1 . Das ist aber kein Problem.

Wegen der gleichmäßigen Konvergenz der Folge $(f'_n)_{n=1}^\infty$ existiert ein $n_2 = n_2(\varepsilon) \in \mathbb{N}$, so dass für $n > n_2$ und alle $x \in [a, b]$:

$$|f'_n(x) - g(x)| < \varepsilon/4. \quad (2.78)$$

Sei $n > n_0 = n_0(\varepsilon, h) := \max\{n_1(\varepsilon, h), n_2(\varepsilon)\}$. Wegen des Mittelwertsatzes (siehe Seite 315) existiert ein ξ_n zwischen x_0 und $x_0 + h$ (das von x_0 , h und f_n abhängt) mit

$$\frac{f_n(x_0 + h) - f_n(x_0)}{h} = f'_n(\xi_n). \quad (2.79)$$

Damit haben wir alle Voraussetzungen ausgenutzt und setzen sie nun zusammen, um (2.75) zu erhalten:

$$\begin{aligned} & \left| \frac{f(x_0 + h) - f(x_0)}{h} - g(x_0) \right| \\ &= \left| \frac{f(x_0 + h) - f_n(x_0 + h) + f_n(x_0 + h) - f_n(x_0) + f_n(x_0) - f(x_0)}{h} \right. \\ & \quad \left. - g(x_0) + f'_n(\xi_n) - f'_n(\xi_n) - g(\xi_n) + g(\xi_n) \right| \\ &\leq \left| \frac{f(x_0 + h) - f_n(x_0 + h)}{h} \right| + \left| \frac{f_n(x_0 + h) - f_n(x_0)}{h} - f'_n(\xi_n) \right| \end{aligned}$$

$$\begin{aligned}
& + \left| \frac{f_n(x_0) - f(x_0)}{h} \right| + |f'_n(\xi_n) - g(\xi_n)| + |g(\xi_n) - g(x_0)| \\
& \stackrel{(2.77)}{\leq} \frac{\varepsilon}{4} + \left| \frac{f_n(x_0 + h) - f_n(x_0)}{h} - f'_n(\xi_n) \right| + \frac{\varepsilon}{4} + |f'_n(\xi_n) - g(\xi_n)| + |g(\xi_n) - g(x_0)| \\
& \stackrel{(2.79)}{=} \frac{\varepsilon}{2} + |f'_n(\xi_n) - g(\xi_n)| + |g(\xi_n) - g(x_0)| \stackrel{(2.78)}{\leq} \frac{3}{4}\varepsilon + |g(\xi_n) - g(x_0)| \stackrel{(2.76)}{\leq} \varepsilon. \quad \square
\end{aligned}$$

Die Voraussetzungen des Satzes können abgeschwächt werden. Statt der gleichmäßigen Konvergenz von $(f_n)_{n=1}^\infty$ ist nur die Konvergenz in einem Punkt $x_0 \in [a, b]$ erforderlich. Benötigt man nicht die Stetigkeit von f' , so genügt die Differenzierbarkeit der f_n , die stetige Differenzierbarkeit muss nicht vorausgesetzt werden. Bei abgeschwächten Voraussetzungen gestaltet sich der Beweis entsprechend aufwändiger.

Man muss aber weiterhin eine Bedingung wie die der gleichmäßigen Konvergenz an die Folge der Ableitungen stellen. Im Beispiel (2.73) war eine Vertauschung von Folgengrenzwert und Ableitung nicht möglich. Die Folge $f_n(x) = \frac{\sin(nx)}{\sqrt{n}}$ konvergiert zwar punktweise und sogar gleichmäßig gegen $f(x) = 0$, aber die Folge der Ableitungen ist divergent.

Beispiel 2.130

Die Folge $(f_n)_{n=1}^\infty$ mit $f_n(x) := x^2 + \frac{\sin(nx)}{n^2}$ ist gleichmäßig konvergent gegen $f(x) = x^2$:

$$|f_n(x) - f(x)| = \frac{|\sin(nx)|}{n^2} \leq \frac{1}{n^2} \rightarrow 0, \quad n \rightarrow \infty,$$

unabhängig von x . Die Folge der Ableitungen konvergiert gleichmäßig gegen $g(x) := 2x$:

$$|f'_n(x) - g(x)| = \left| 2x + \frac{\cos(nx)}{n} - 2x \right| = \frac{|\cos(nx)|}{n} \leq \frac{1}{n} \rightarrow 0, \quad n \rightarrow \infty.$$

Tatsächlich ist $\frac{d}{dx} [\lim_{n \rightarrow \infty} f_n(x)] = f'(x) = 2x = g(x) = \lim_{n \rightarrow \infty} f'_n(x)$. ■

Es ist bekannt, dass die stetige Weierstraß-Funktion (2.74) an keiner einzigen Stelle differenzierbar ist. Damit kann die Reihe der Ableitungen $\sum_{k=1}^\infty \left(\frac{2}{3}\right)^k 2^k \cos(2^k x)$ nicht gleichmäßig konvergent sein. Für $x = 0$ haben wir tatsächlich sogar die divergente Zahlenreihe $\sum_{k=1}^\infty \left(\frac{4}{3}\right)^k$.

Ähnlich zur Ableitung verhält sich die Integration:

Satz 2.58 (Integrierbarkeit der Grenzfunktion)

Sei $(f_n)_{n=1}^\infty$ eine gleichmäßig konvergente Funktionenfolge auf $[a, b]$, die gegen f konvergiert. Falls jedes f_n integrierbar auf $[a, b]$ ist, so ist auch f integrierbar auf $[a, b]$, und es gilt:

$$\lim_{n \rightarrow \infty} \int_a^b f_n(x) dx = \int_a^b \left[\lim_{n \rightarrow \infty} f_n(x) \right] dx = \int_a^b f(x) dx.$$

Beweis: Wir beweisen zunächst die Integrierbarkeit der Grenzfunktion f , indem wir zeigen, dass das Ober- mit dem Unterintegral übereinstimmt. Das schaffen wir, indem wir mittels der gleichmäßigen Konvergenz in den Unter- und Obersummen f durch ein f_n ersetzen.

Sei $\varepsilon > 0$ beliebig und im Folgenden fest. Da $(f_n)_{n=1}^\infty$ gleichmäßig gegen f konvergiert, existiert zu diesem $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$, so dass für alle $n > n_0$ und $x \in [a, b]$ gilt:

$$|f(x) - f_n(x)| < \varepsilon. \quad (2.80)$$

Da ein solches f_n als integrierbare Funktion beschränkt ist, gilt dies wegen (2.80) auch für f . Wir können also Ober- und Untersummen von f betrachten. Sei $Z = (x_0, x_1, \dots, x_m)$ eine Zerlegung von $[a, b]$ und $\mathcal{Z}$ die Menge aller Zerlegungen von $[a, b]$. Für die Obersumme $S_Z(f)$ von f zur Zerlegung Z gilt für $n > n_0$:

$$\begin{aligned} S_Z(f) &= \sum_{k=1}^m \left[\sup_{x \in [x_{k-1}, x_k]} f(x) \right] (x_k - x_{k-1}) \\ &= \sum_{k=1}^m \left[\sup_{x \in [x_{k-1}, x_k]} (f(x) - f_n(x) + f_n(x)) \right] (x_k - x_{k-1}) \\ &\leq \sum_{k=1}^m \left[\sup_{x \in [x_{k-1}, x_k]} (|f(x) - f_n(x)| + f_n(x)) \right] (x_k - x_{k-1}) \\ &\stackrel{(2.80)}{\leq} \sum_{k=1}^m \left[\varepsilon + \sup_{x \in [x_{k-1}, x_k]} f_n(x) \right] (x_k - x_{k-1}) \\ &= (b-a)\varepsilon + \sum_{k=1}^m \left[\sup_{x \in [x_{k-1}, x_k]} f_n(x) \right] (x_k - x_{k-1}) = (b-a)\varepsilon + S_Z(f_n). \end{aligned}$$

Analog gilt für die Untersumme $s_Z(f)$ von f zur Zerlegung Z :

$$\begin{aligned} s_Z(f) &= \sum_{k=1}^m \left[\inf_{x \in [x_{k-1}, x_k]} f(x) \right] (x_k - x_{k-1}) \\ &\geq \sum_{k=1}^m \left[\inf_{x \in [x_{k-1}, x_k]} (f_n(x) - |f_n(x) - f(x)|) \right] (x_k - x_{k-1}) \\ &\stackrel{(2.80)}{\geq} -(b-a)\varepsilon + s_Z(f_n). \end{aligned}$$

Damit erhalten wir für das Ober- und Unterintegral:

$$\bar{I}(f) = \inf_{Z \in \mathcal{Z}} S_Z(f) \leq (b-a)\varepsilon + \inf_{Z \in \mathcal{Z}} S_Z(f_n) = (b-a)\varepsilon + \bar{I}(f_n), \quad (2.81)$$

$$\underline{I}(f) = \sup_{Z \in \mathcal{Z}} s_Z(f) \geq -(b-a)\varepsilon + \sup_{Z \in \mathcal{Z}} s_Z(f_n) = -(b-a)\varepsilon + \underline{I}(f_n). \quad (2.82)$$

Also ist

$$|\bar{I}(f) - \underline{I}(f)| = \bar{I}(f) - \underline{I}(f) \leq 2(b-a)\varepsilon + \bar{I}(f_n) - \underline{I}(f_n) = 2(b-a)\varepsilon,$$

da f_n integrierbar ist, so dass $\bar{I}(f_n) = \underline{I}(f_n)$. Da zu Beginn $\varepsilon > 0$ beliebig gewählt wurde, gilt $\bar{I}(f) = \underline{I}(f)$, d. h., auch f ist integrierbar.

Jetzt müssen wir nur noch nachrechnen, dass das Integral der Grenzfunktion mit dem Grenzwert der Integrale der f_n übereinstimmt:

$$\begin{aligned} \int_a^b f(x) dx - \int_a^b f_n(x) dx &= \bar{I}(f) - \bar{I}(f_n) \stackrel{(2.81)}{\leq} (b-a)\varepsilon, \\ \int_a^b f_n(x) dx - \int_a^b f(x) dx &= \underline{I}(f_n) - \underline{I}(f) \stackrel{(2.82)}{\leq} (b-a)\varepsilon. \end{aligned}$$

Es ist also für $n > n_0$:

$$\left| \int_a^b f(x) dx - \int_a^b f_n(x) dx \right| \leq (b-a)\varepsilon$$

und damit $\lim_{n \rightarrow \infty} \int_a^b f_n(x) dx = \int_a^b f(x) dx$. □

Bei gleichmäßiger Konvergenz kann also die Integration mit der Bestimmung des Grenzwerts vertauscht werden. Der Satz von Arzela (1847–1912), Osgood (1864–1943) und Lebesgue (1875–1941) besagt sogar darüber hinaus, dass man die beiden Grenzwerte auch ohne gleichmäßige Konvergenz bei nur punktwiser Konvergenz vertauschen darf. Allerdings müssen dazu alle Funktionen der Folge mit der gleichen Konstante beschränkt sein, und die Grenzfunktion muss integrierbar sein.

Im zuvor betrachteten Beispiel $f_n(x) := n^2 x(1-x)^n$ haben wir punktweise Konvergenz auf $[0, 1]$ gegen $f(x) = 0$. Aber eine Vertauschung des Grenzwerts mit dem Integral war nicht möglich, obwohl jedes einzelne f_n und auch f integrierbar ist. Folglich kann weder gleichmäßige Konvergenz vorliegen noch können alle Funktionen der Folge mit der gleichen Konstante beschränkt sein. Dass dem so ist, sieht man, wenn $x_n = \frac{1}{n+1}$ in f_n eingesetzt wird und man wieder die Darstellung (2.8) von e verwendet (siehe Seite 228):

$$|f_n(x_n) - f(x_n)| = f_n(x_n) = \frac{n^2}{n+1} \left(\frac{n}{n+1} \right)^n = \underbrace{\frac{n^2}{n+1}}_{\rightarrow \infty} \underbrace{\left(1 + \frac{1}{n} \right)^{-n}}_{\rightarrow e^{-1}} \rightarrow \infty, \quad n \rightarrow \infty.$$

2.8.3 Konvergenz von Potenzreihen

Satz 2.59 (Konvergenzradius)

Jede Potenzreihe $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ erfüllt genau eine der drei folgenden Bedingungen:

- a) Die Potenzreihe konvergiert nur in x_0 , d. h. nur auf $[x_0 - \rho, x_0 + \rho]$ mit $\rho = 0$.
- b) Die Potenzreihe konvergiert (absolut) auf $]x_0 - \rho, x_0 + \rho[$ und divergiert für jedes $x \notin [x_0 - \rho, x_0 + \rho]$ für ein $0 < \rho < \infty$.

c) Die Potenzreihe konvergiert (absolut) auf $\mathbb{R}$, d. h. auf „ $]x_0 - \rho, x_0 + \rho[$ für $\rho = \infty$ “.

Der damit eindeutig bestimmte Wert ρ heißt **Konvergenzradius** der Potenzreihe.

Jetzt wird verständlich, warum x_0 Entwicklungsmittelpunkt heißt, x_0 liegt in der Mitte des Konvergenzintervalls. Bei b) wird keine Aussage über das Konvergenzverhalten in $x_0 - \rho$ und $x_0 + \rho$ gemacht. Tatsächlich kann hier abhängig von der Reihe Konvergenz oder Divergenz in jeder denkbaren Kombination vorliegen.

Der Beweis des Satzes geschieht über einen Vergleich mit der konvergenten geometrischen Reihe unter Ausnutzung der Vollständigkeit der reellen Zahlen. Dazu benötigen wir den folgenden Hilfssatz:

Lemma 2.13 (Punktweise absolute Konvergenz)

Für eine Potenzreihe $\sum_{k=0}^{\infty} a_k(x - x_0)^k$ gilt: Falls die Potenzreihe für einen Punkt $x_1 \neq x_0$ konvergiert, so konvergiert die Reihe absolut für jedes $x \in \mathbb{R}$ mit $|x - x_0| < |x_1 - x_0|$. Das heißt, für diese Punkte x ist $\sum_{k=0}^{\infty} |a_k(x - x_0)^k|$ konvergent (und damit insbesondere auch $\sum_{k=0}^{\infty} a_k(x - x_0)^k$, siehe Satz 2.13 auf Seite 243).

Beweis: Da $\sum_{k=0}^{\infty} a_k(x_1 - x_0)^k$ konvergiert, ist notwendigerweise $\lim_{k \rightarrow \infty} a_k(x_1 - x_0)^k = 0$. Die Folge $(a_k(x_1 - x_0)^k)_{k=0}^{\infty}$ ist damit eine beschränkte Folge: $|a_k(x_1 - x_0)^k| \leq M$ für ein $M \in \mathbb{R}$. Sei x ein Punkt, in dem die Funktionenreihe nach Aussage des Lemmas konvergieren soll, d. h. $|x - x_0| < |x_1 - x_0|$. Dann ist $q := |x - x_0|/|x_1 - x_0| < 1$ und

$$|a_k(x - x_0)^k| = \frac{|a_k(x_1 - x_0)^k| |x - x_0|^k}{|x_1 - x_0|^k} \leq M \left| \frac{x - x_0}{x_1 - x_0} \right|^k \leq M q^k. \quad (2.83)$$

Da die geometrische Reihe $\sum_{k=0}^{\infty} q^k$ nach (2.11) auf Seite 237 für $q < 1$ konvergent ist, folgt die absolute Konvergenz im Punkt x mit dem Vergleichskriterium für Zahlenfolgen (Seite 246). $\square$

Bemerkung: Wegen Lemma 2.12 folgt für $0 < p < 1$ aus diesem Beweis sogar die gleichmäßige Konvergenz (siehe Definition 2.43) der Potenzreihe auf $[x_0 - p \cdot |x_1 - x_0|, x_0 + p \cdot |x_1 - x_0|]$, da wir für alle Werte x aus diesem Intervall die Abschätzung (2.83) mit $|x - x_0| \leq p|x_1 - x_0|$ modifizieren können zu $|a_k(x - x_0)^k| \leq M p^k$.

Beweis: (zu Satz 2.59) Sei

$$\rho := \sup \left\{ |x - x_0| : \text{Es gibt ein } x \in \mathbb{R}, \text{ so dass } \sum_{k=0}^{\infty} a_k(x - x_0)^k \text{ konvergent ist.} \right\},$$

sofern das Supremum existiert. ρ ist dann der kleinste Wert, so dass alle Punkte x , in denen die Reihe konvergiert, in $[x_0 - \rho, x_0 + \rho]$ liegen. Da die Potenzreihe für $x = x_0$ konvergent ist, ist 0 in der betrachteten Menge enthalten, die damit nicht-leer ist. Ist sie nach oben beschränkt, folgt mit dem Vollständigkeitsaxiom der reellen Zahlen (siehe Seite 49) die Existenz des Supremums ρ . Anderenfalls setzen wir $\rho = \infty$. Damit ist ρ wohldefiniert. In Abhängigkeit von ρ erhalten wir genau einen der drei folgenden Fälle:

- a) Ist $\rho = 0$, so konvergiert die Potenzreihe nur in $x = x_0$.
- b) Ist $0 < \rho < \infty$, so existiert zu jedem $0 < \delta < \rho$ ein $x_1 \in [x_0 - \rho, x_0 + \rho]$ mit $|x_1 - x_0| > \rho - \delta$, und $\sum_{k=0}^{\infty} a_k(x_1 - x_0)^k$ konvergiert (sonst wäre ρ keine kleinste obere Schranke, da auch $\rho - \delta$ eine obere Schranke wäre). Nach Lemma 2.13 konvergiert die Potenzreihe damit aber auf $]x_0 - \rho + \delta, x_0 + \rho - \delta[$ absolut. Da $\delta > 0$ beliebig gewählt werden kann, ist die Potenzreihe absolut konvergent auf $]x_0 - \rho, x_0 + \rho[$. Nach Konstruktion des Supremums divergiert sie für jedes x außerhalb von $[x_0 - \rho, x_0 + \rho]$.
- c) Ist $\rho = \infty$, so folgt mit einem ähnlichen Argument wie zu b) die Konvergenz der Potenzreihe auf $\mathbb{R}$. □

Auf jedem abgeschlossenen Intervall $[a, b] \subset]x_0 - \rho, x_0 + \rho[$ konvergiert eine Potenzreihe sogar gleichmäßig. Das ergibt sich aus der Bemerkung zur gleichmäßigen Konvergenz zu Lemma 2.13.

Eine Anwendung des Wurzel- bzw. Quotientenkriteriums für Zahlenreihen liefert den folgenden Satz, mit dem der Konvergenzradius ausgerechnet werden kann. Diese beiden Kriterien eignen sich deshalb besonders gut, da die Summanden Faktoren $(x - x_0)^k$ aufweisen, so dass bei Anwendung der beiden Kriterien nur die Konstante $|x - x_0|$ übrig bleibt. Hat man es zudem wie in der Taylor-Entwicklung mit Fakultäten zu tun, so kürzen diese sich in der Regel bei Anwendung des Quotientenkriteriums weg. Aber auch das Wurzelkriterium lässt sich dann in Verbindung mit der Stirling'schen Formel (2.62) von Seite 369 einsetzen.

Satz 2.60 (Formel von Cauchy-Hadamard, Hadamard (1865–1963))

Sei $\sum_{k=0}^{\infty} a_k(x - x_0)^k$ eine Potenzreihe mit Konvergenzradius ρ . Weiter existiere mindestens einer der beiden Grenzwerte

$$r_1 := \lim_{k \rightarrow \infty} |a_k|^{1/k},$$

oder

$$r_2 := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|,$$

wobei auch $r_1 = \infty$ oder $r_2 = \infty$ zugelassen sei. In diesem Sinne sei $r = r_1$ oder $r = r_2$ ein existierender Grenzwert. Dann gilt:

- a) Ist $r = \infty$, so ist $\rho = 0$.

- b) Ist $0 < r < \infty$, so ist $\rho = \frac{1}{r}$.
 c) Ist $r = 0$, so ist $\rho = \infty$.

Wir beweisen hier nur die Aussage für den ersten Grenzwert $r = r_1 = \lim_{k \rightarrow \infty} |a_k|^{1/k}$ mit dem Wurzelkriterium. Analog zeigt man den Konvergenzradius für den zweiten Grenzwert $r = r_2 = \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|$ mit dem Quotientenkriterium, siehe Aufgabe 2.45.

Beweis: a) Falls $r = \infty$ ist, gilt für $x \neq x_0$:

$$\lim_{k \rightarrow \infty} \left| a_k (x - x_0)^k \right|^{1/k} = |x - x_0| \underbrace{\lim_{k \rightarrow \infty} |a_k|^{1/k}}_{= r = \infty} = \infty.$$

Damit ist aber $(a_k (x - x_0)^k)_{k=0}^\infty$ keine Nullfolge, so dass die Potenzreihe für jedes $x \neq x_0$ divergent ist. Sie ist nur für $x = x_0$ offensichtlich konvergent. Daher ist $\rho = 0$.

b) Sei $0 < r < \infty$ und $x \in \mathbb{R}$ mit $|x - x_0| < \frac{1}{r}$. Dann ist

$$\lim_{k \rightarrow \infty} \left| a_k (x - x_0)^k \right|^{1/k} = r \cdot |x - x_0| < 1.$$

Damit ist die Potenzreihe nach dem Wurzelkriterium (Satz 2.18 auf Seite 253) für Reihen an der Stelle x (absolut) konvergent. Da für $|x - x_0| > \frac{1}{r}$ gilt:

$$\lim_{k \rightarrow \infty} \left| a_k (x - x_0)^k \right|^{1/k} = r \cdot |x - x_0| > 1,$$

ist die Potenzreihe für dieses x aufgrund des Wurzelkriteriums divergent.

c) Ist $r = 0$, dann gilt für alle $x \in \mathbb{R}$:

$$\lim_{k \rightarrow \infty} \left| a_k (x - x_0)^k \right|^{1/k} = r \cdot |x - x_0| = 0,$$

so dass die Potenzreihe gemäß des Wurzelkriteriums für jedes x konvergiert. $\square$

Beispiel 2.131

- a) In Beispiel 2.125 auf Seite 399 haben wir bereits nachgerechnet, dass die Taylor-Reihen der Exponential-, der Sinus- und der Kosinus-Funktion für jeden Wert $x \in \mathbb{R}$ konvergieren. Damit ist ihr Konvergenzradius jeweils $\rho = \infty$. Beispielsweise hat die Exponentialreihe $\sum_{k=0}^\infty \frac{x^k}{k!}$ die Koeffizienten $a_k = \frac{1}{k!}$, $a_{k+1} = \frac{1}{(k+1)!}$, und wir erhalten $r := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k \rightarrow \infty} \frac{k!}{(k+1)!} = \lim_{k \rightarrow \infty} \frac{1}{k+1} = 0$. Also folgt auch mit Satz 2.60, dass $\rho = \infty$ ist.
- b) Für die geometrische Reihe $\sum_{k=0}^\infty x^k = 1 + x + x^2 + \dots$ gilt $a_k = 1$ für alle $k \in \mathbb{N}_0$, und wir erhalten sowohl

$$\lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k \rightarrow \infty} 1 = 1 \text{ als auch } \lim_{k \rightarrow \infty} \sqrt[k]{|a_k|} = \lim_{k \rightarrow \infty} \sqrt[k]{1} = 1$$

und damit als Kehrwert den Konvergenzradius $\rho = 1$. Das stimmt mit unserem Wissen über die geometrische Reihe überein (vgl. Seite 237).

$x = 1$: $\sum_{k=0}^{\infty} 1 = \infty$ ist bestimmt divergent.

$x = -1$: Die Partialsummen haben die Folgehäufungspunkte 0 und 1 und sind damit divergent.

- c) Für die Potenzreihe $\sum_{k=1}^{\infty} \frac{x^k}{2^k \cdot k} = \frac{x}{2} + \frac{x^2}{4 \cdot 2} + \dots$ gilt $a_k = \frac{1}{2^k \cdot k}$ für alle k , so dass

$$r := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k \rightarrow \infty} \left| \frac{2^k \cdot k}{2^{k+1} \cdot (k+1)} \right| = \frac{1}{2} \lim_{k \rightarrow \infty} \frac{k}{k+1} = \frac{1}{2}.$$

Damit ist $\rho = 2$, und die Reihe konvergiert für $x \in]-2, 2[$ absolut und divergiert für $|x| > 2$.

Für $x = 2$ ist $\sum_{k=1}^{\infty} \frac{x^k}{2^k \cdot k} = \sum_{k=1}^{\infty} \frac{1}{k}$ divergent (harmonische Reihe).

Für $x = -2$ ist $\sum_{k=1}^{\infty} \frac{x^k}{2^k \cdot k} = \sum_{k=1}^{\infty} \frac{(-1)^k}{k}$ konvergent (alternierende harmonische Reihe).

Die Reihe konvergiert also genau für $x \in [-2, 2[$.

- d) Für die Reihe $\sum_{k=1}^{\infty} (-1)^{k-1} \cdot \frac{x^k}{k} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$ ist $a_k = (-1)^{k-1} \cdot \frac{1}{k}$, $a_{k+1} = (-1)^k \cdot \frac{1}{k+1}$, d. h. $\frac{a_{k+1}}{a_k} = -\frac{k}{k+1}$ und weiter $r := \lim_{k \rightarrow \infty} \frac{k}{k+1} = 1$.

Die Reihe konvergiert für $|x| < 1$ absolut und divergiert für $|x| > 1$. In den Randpunkten gilt:

$x = 1$: $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$ ist konvergent nach dem Leibniz-Kriterium;

$x = -1$: $-1 - \frac{1}{2} - \frac{1}{3} - \frac{1}{4} - \dots$ ist als negative harmonische Reihe divergent.

- e) Für die Reihe $\sum_{k=1}^{\infty} (kx)^k = x + (2x)^2 + (3x)^3 + \dots = x + 4x^2 + 27x^3 + \dots$ mit $a_k = k^k$ folgt $\sqrt[k]{|a_k|} = k$ und weiter $r := \lim_{k \rightarrow \infty} k = \infty$. Wegen $\rho = 0$ ist die Reihe divergent für alle $x \neq x_0 = 0$.

- f) Zur Folge der Fibonacci-Zahlen $a_0 = 0$, $a_1 = 1$, $a_k = a_{k-2} + a_{k-1}$ für $k \geq 2$ von Seite 231 bilden wir eine Potenzreihe $\sum_{k=0}^{\infty} a_k x^k$. Da der Grenzwert $\lim_{k \rightarrow \infty} \frac{a_{k+1}}{a_k} = \Phi = \frac{1+\sqrt{5}}{2}$ 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. Wir werden diese Formel auf Seite 551 mit Mitteln der Linearen Algebra beweisen, für den Potenzreihenansatz verweisen wir auf (Heuser, 2009, S. 378) in Verbindung mit Aufgabe 2.48 d). ■

2.8.4 Differenziation und Integration von Potenzreihen

Im nächsten Abschnitt werden wir uns intensiver mit dem Zusammenhang zwischen Potenzreihen und Taylor-Reihen beschäftigen. Dazu werden wir die Taylor-Reihe der Grenzfunktion einer Potenzreihe berechnen. Wir benötigen dazu alle Ableitungen dieser Grenzfunktion. Hier sehen wir, wie man diese berechnen kann:

Satz 2.61 (Gliederweise Ableitung)

Ist $f(x)$ die Grenzfunktion einer Potenzreihe $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ mit Konvergenzradius $\rho > 0$, so ist f auf $]x_0 - \rho, x_0 + \rho[$ beliebig oft differenzierbar. Die Ableitungen sind als Potenzreihen mit gleichem Konvergenzradius ρ entwickelbar. Es darf gliedweise differenziert werden, z. B.

$$f'(x) = \sum_{k=1}^{\infty} k a_k (x - x_0)^{k-1}.$$

Beweis: (Skizze) Sofern man den Konvergenzradius von $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ mit dem Grenzwert aus der Formel von Cauchy-Hadamard (Satz 2.60) ermitteln kann, erhält man so auch den Konvergenzradius der gliedweise differenzierten Potenzreihe. Ist z. B. $0 < \rho < \infty$, so gilt wie zuvor mit dem Satz von L'Hospital (Seite 325):

$$\lim_{k \rightarrow \infty} |k a_k|^{\frac{1}{k}} = \lim_{k \rightarrow \infty} \exp\left(\frac{\ln k}{k}\right) \lim_{k \rightarrow \infty} |a_k|^{\frac{1}{k}} = \exp\left(\lim_{k \rightarrow \infty} \frac{\frac{1}{k}}{1}\right) \frac{1}{\rho} = \frac{1}{\rho}.$$

Damit sind beide Reihen gleichmäßig konvergente Funktionenfolgen auf jedem Intervall $[a, b]$ innerhalb des gemeinsamen Konvergenzbereiches. Die Voraussetzungen von Satz 2.57 sind erfüllt, so dass einmal gliedweise differenziert werden darf. Nun kann man das Argument sukzessive wiederholen, um auch die Aussage für höhere Ableitungen zu erhalten. $\square$

Beispiel 2.132

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 Differenzieren nun auch die Ableitungen:

$$\frac{d}{dx} e^x = \frac{d}{dx} \sum_{k=0}^{\infty} \frac{x^k}{k!} = \sum_{k=0}^{\infty} \frac{d}{dx} \frac{x^k}{k!} = \sum_{k=1}^{\infty} \frac{k x^{k-1}}{k!} = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x. \quad (2.84) \quad \blacksquare$$

Beispiel 2.133

Der Sinus Cardinalis $\text{sinc}(x)$ ist die an der Stelle $x = 0$ stetig mit dem Wert 1 ergänzte Funktion $\frac{\sin(x)}{x}$ (siehe Seite 271). Mit der Potenzreihe des Sinus erhalten wir für $x \neq 0$ die Darstellung

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

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

Da die Reihe insbesondere für den Entwicklungsmittelpunkt 0 den Wert 1 hat, erhalten wir über die Potenzreihe den Grenzwert $\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1$ ohne die geometrischen Überlegungen von Seite 262. ■

Nachdem wir gesehen haben, dass Potenzreihen gliedweise differenzierbar sind, stellt sich nun die Frage nach gliedweiser Integration. Diese lässt sich aber aufgrund der gleichmäßigen Konvergenz der Potenzreihen sofort mit Satz 2.58 beantworten: Innerhalb des Konvergenzbereichs einer Potenzreihe darf man gliedweise integrieren.

Beispiel 2.134

$$\begin{aligned} \int_0^1 \exp(x) dx &= \int_0^1 \sum_{k=0}^{\infty} \frac{x^k}{k!} dx = \sum_{k=0}^{\infty} \int_0^1 \frac{x^k}{k!} dx = \sum_{k=0}^{\infty} \left[\frac{x^{k+1}}{(k+1)!} \right]_0^1 \\ &= \sum_{k=0}^{\infty} \frac{1}{(k+1)!} = \sum_{k=1}^{\infty} \frac{1}{k!} = -\frac{1}{0!} + \sum_{k=0}^{\infty} \frac{1}{k!} = -1 + e. \end{aligned}$$

2.8.5 Der Zusammenhang zwischen Potenzreihen und Taylor-Reihen

Wir sind zu Potenzreihen gelangt, indem wir allgemeine Koeffizienten für Taylor-Reihen betrachtet haben (siehe Seite 401). Insbesondere ist damit jede Taylor-Reihe eine Potenzreihe. Unklar ist an dieser Stelle noch, ob auch jede Potenzreihe $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ eine Taylor-Reihe ist, d. h., ob sich ihre Koeffizienten darstellen lassen als $a_k = \frac{f^{(k)}(x_0)}{k!}$ mit den Ableitungen einer an einer Stelle x_0 beliebig oft differenzierbaren Funktion f , z. B. der Grenzfunktion der Reihe. Dazu betrachten wir eine auf $]x_0 - \rho, x_0 + \rho[$ konvergente Potenzreihe mit Konvergenzradius $\rho > 0$ und Grenzfunktion f . Durch Einsetzen von x_0 in die Potenzreihe erhält man sofort $a_0 = f(x_0) = \frac{f^{(0)}(x_0)}{0!}$. Nach Satz 2.61 ist die Grenzfunktion f beliebig oft differenzierbar auf $]x_0 - \rho, x_0 + \rho[$ mit $f'(x) = \sum_{k=1}^{\infty} k a_k (x-x_0)^{k-1}$. Damit ist $f'(x_0) = a_1$. Gliedweises Ableiten der Potenzreihe von f' führt zu $f''(x_0) = 2! \cdot a_2$. Entsprechend erhält man durch fortgesetztes Ableiten $f^{(k)}(x_0) = k! \cdot a_k$. Damit haben wir gezeigt:

Jede Potenzreihe mit Konvergenzradius $\rho > 0$ ist die Taylor-Reihe ihrer Grenzfunktion.

Eine Potenzreihe konvergiert auf $]x_0 - \rho, x_0 + \rho[$ gegen genau eine Grenzfunktion, und ihre Koeffizienten ergeben sich aus den Ableitungen dieser Grenzfunktion an der Stelle x_0 . Wenn man umgekehrt mit einer Funktion f beginnt und dazu an der Stelle x_0 die Taylor-Reihe berechnet, dann konvergiert diese Potenzreihe innerhalb ihres Konvergenzintervalls gegen eine Grenzfunktion g . Insbesondere wissen wir jetzt, dass

alle Ableitungen von f und g an der Stelle x_0 übereinstimmen. Heißt das aber auch, dass f und g im Konvergenzintervall gleich sind? Im Allgemeinen ist das leider nicht der Fall. Es ist nicht schwierig nachzurechnen, dass die Funktion

$$f(x) := \begin{cases} \exp(-\frac{1}{x}), & x > 0, \\ 0, & x \leq 0, \end{cases}$$

trotz der stückweisen Definition auf $\mathbb{R}$ beliebig oft differenzierbar ist. Außerdem sind alle Ableitungen an der Stelle $x_0 = 0$ gleich null: $f^{(k)}(x_0) = 0$, $k \in \mathbb{N}_0$. Damit sind aber alle Koeffizienten der zugehörigen Taylor-Reihe gleich null, und die Grenzfunktion dieser trivialen Potenzreihe ist die Nullfunktion mit Konvergenzradius $\rho = \infty$, die aber offensichtlich nicht mit f übereinstimmt.

Wir können zwar mit der Formel von Cauchy-Hadamard (Satz 2.60) ausrechnen, wo eine Taylor-Reihe konvergiert, wir wissen dann aber nicht, ob sie tatsächlich gegen die Funktion konvergiert, aus deren Ableitungen die Taylor-Reihe aufgebaut wurde. Wenn wir das zeigen sollen, dann müssen wir das mittels des Verhaltens der Taylor-Restglieder überprüfen.

Wir fassen die Ergebnisse über Taylor-Reihen zusammen:

- Da eine Taylor-Reihe eine Potenzreihe ist, kann ihr Konvergenzradius mit der Formel von Cauchy-Hadamard bestimmt werden. Allerdings kann der Konvergenzradius null sein.
- Falls die Taylor-Reihe von f konvergiert, muss sie nicht notwendigerweise gegen f konvergieren.
- Die Taylor-Reihe konvergiert genau dann gegen $f(x)$, wenn das Restglied der Taylor-Entwicklung $R_{x_0, n+1, f}(x)$ für $n \rightarrow \infty$ gegen null geht. In diesem Fall stimmen Taylor-Reihe (als Grenzfunktion) und Ausgangsfunktion der Taylor-Reihe (Funktion, mit deren Ableitungen die Taylor-Reihe berechnet wird) überein.

2.8.6 Die komplexe Exponentialfunktion

Über die Entwicklung der Exponentialfunktion als Potenzreihe wird die Bezeichnung $e^{j\varphi} := \cos(\varphi) + j \sin(\varphi)$ in der Euler'schen Gleichung (Definition 1.41 auf Seite 139) verständlich. Dazu setzen wir in die Exponentialreihe die komplexe Zahl $j\varphi$ ein. Das geht, da hier lediglich Potenzen von $j\varphi$ gebildet werden. Die zunächst nur für reelle Zahlen definierte Exponentialfunktion kann so auch für imaginäre Zahlen einen Sinn bekommen:

$$\begin{aligned} \exp(j\varphi) &= \sum_{k=0}^{\infty} \frac{(j\varphi)^k}{k!} = \sum_{k=0}^{\infty} \left[\frac{(j\varphi)^{2k}}{(2k)!} + \frac{(j\varphi)^{2k+1}}{(2k+1)!} \right] = \sum_{k=0}^{\infty} \frac{(j\varphi)^{2k}}{(2k)!} + \sum_{k=0}^{\infty} \frac{(j\varphi)^{2k+1}}{(2k+1)!} \\ &= \sum_{k=0}^{\infty} (-1)^k \frac{\varphi^{2k}}{(2k)!} + j \sum_{k=0}^{\infty} (-1)^k \frac{\varphi^{2k+1}}{(2k+1)!} = \cos(\varphi) + j \sin(\varphi). \end{aligned}$$

Hier haben wir zuerst Summanden zu geraden und ungeraden Indizes zusammengefasst, also in der unendlichen Summe Klammern gesetzt. Das ist erlaubt, da die Potenzreihen absolut konvergent sind. Die Aufspaltung in zwei Summen ist dann erlaubt, da die Grenzwerte der beiden einzelnen Summen als Potenzreihen von Kosinus und Sinus existieren.

Nachdem wir zunächst $e^{j\varphi} := \cos(\varphi) + j \sin(\varphi)$ nur als abkürzende Schreibweise verwendet haben, sehen wir nun, dass sich dahinter tatsächlich eine Erweiterung der reellen Exponentialfunktion für imaginäre Zahlen verbirgt. Entsprechend kann man sich überlegen, dass

$$\exp(z) := e^{\operatorname{Re}(z)} e^{j \operatorname{Im}(z)}$$

eine Erweiterung der reellen Exponentialfunktion auf alle komplexen Zahlen ist, die selbst für alle $z \in \mathbb{C}$ eine Potenzreihendarstellung

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

besitzt und für die die bereits bekannten Rechenregeln der reellen Exponentialfunktion auch für komplexes z gelten. Damit haben wir nun nachträglich die Definition 1.42 auf Seite 140 gerechtfertigt.

Für Potenzreihen einer komplexen Variable $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ kann man einen Konvergenzradius ρ ebenfalls mit der Formel von Cauchy-Hadamard berechnen (siehe Seite 413). Die Potenzreihe konvergiert innerhalb der Kreisscheibe $\{z \in \mathbb{C} : |z - z_0| < \rho\}$. Daher wird ρ als Radius bezeichnet.

Beispiel 2.135

Die komplexe geometrische Reihe $\sum_{k=0}^{\infty} z^k = 1 + z + z^2 + \dots$ mit $a_k = 1$, $k \in \mathbb{N}_0$, besitzt als Konvergenzradius den Kehrwert des Grenzwerts (vgl. Beispiel 2.131)

$$r := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k \rightarrow \infty} 1 = 1.$$

Somit konvergiert die Reihe für alle z innerhalb des Einheitskreises $|z - 0| < 1$. Sie ist divergent für alle $z \in \mathbb{C}$ mit $|z| > 1$. Über die Konvergenz auf dem Einheitskreis $|z| = 1$ macht die Formel von Cauchy-Hadamard keine Aussage. ■

Die Ableitung von Funktionen einer komplexen Variable wird über einen komplexen Differenzenquotienten

$$f'(z_0) := \lim_{z \rightarrow z_0} \frac{f(z_0) - f(z)}{z_0 - z} \quad (z \in \mathbb{C})$$

definiert. Innerhalb der Kreisscheibe ist die Grenzfunktion der Potenzreihe beliebig oft komplex differenzierbar. Sie heißt auf dieser Menge **analytisch** oder **holomorph**. Mit holomorphen Funktionen beschäftigt sich die Funktionentheorie. Wir werden die komplexe Ableitung aus Sicht der Differenzialrechnung für Funktionen mit mehreren Variablen auf Seite 609 noch kurz aus einer anderen Perspektive beleuchten.

2.9 Aufgaben

Aufgabe 2.1

Untersuchen Sie die Folgen auf Monotonie und Beschränktheit und geben Sie – falls vorhanden – das Infimum und Supremum der Menge der Folgenglieder (der Wertemenge) an:

a) $a_n = 1 + n + \ln(n)$, b) $b_n = \frac{3n}{n+1}$, c) $c_n = \frac{1}{n^2+1}$.

Aufgabe 2.2

Bestimmen Sie Kandidaten für die Grenzwerte der Folgen, und bestimmen Sie zu jedem $\varepsilon > 0$ jeweils eine Stelle n_0 , ab der alle Glieder in einer ε -Umgebung des jeweiligen Grenzwerts liegen:

a) $a_n = 10^{-n}$, b) $b_n = \sqrt[n]{4}$, c) $c_n = \frac{n^2}{n^2+2}$.

Aufgabe 2.3

Man zeige mit der ε - n_0 -Definition des Folngengrenzwerts $\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}+2+\sin(n)} = 0$.

Hinweis: Man schätze in einem ersten Schritt geeignet durch Weglassen ab.

Aufgabe 2.4

Zeigen Sie, dass die Folgen den jeweils angegebenen Grenzwert besitzen:

a) $a_n = 2 + (-1)^n \frac{1}{n^2}$, $\lim_{n \rightarrow \infty} a_n = 2$, b) $b_n = \frac{4-4n+n^2}{2+3n-n^2}$, $\lim_{n \rightarrow \infty} b_n = -1$,
 c) $c_n = \sqrt{n+1} - \sqrt{n}$, $\lim_{n \rightarrow \infty} c_n = 0$, d) $d_n = \frac{1+2+\dots+n}{n^2+n+1}$, $\lim_{n \rightarrow \infty} d_n = \frac{1}{2}$.

Bei a) verwende man die Definition des Folngengrenzwerts, bei b) – d) verwende man bekannte Sätze über Folngengrenzwerte. Bei c) hilft die dritte Binomische Formel.

Aufgabe 2.5

Berechnen Sie die Folngengrenzwerte:

a) $\lim_{n \rightarrow \infty} \frac{3n^3+4n^2+7}{6n^3+2n+1}$, b) $\lim_{n \rightarrow \infty} \frac{3n^3+4n+7}{6n^2+2n+1}$, c) $\lim_{n \rightarrow \infty} \frac{3n^3+4n^2+7}{6n^4+2n^2+n}$.

Aufgabe 2.6

Die durchschnittliche Inflationsrate während der Existenz der D-Mark (1948–2002) betrug ca. 2,7 %. Wir gehen für die Rechnung von einer konstanten Inflationsrate aus. Welchen Wert hatte Die D-Mark 2002 im Vergleich zu 1948, und wie groß war die Halbwertszeit der D-Mark?

Aufgabe 2.7

Man prüfe mit dem Quotientenkriterium, ob $\sum_{k=0}^{\infty} \frac{1}{[(k+1)!]^2}$ und $\sum_{k=0}^{\infty} \frac{k^2}{\pi^k}$ konvergieren.

Aufgabe 2.8

Man prüfe die alternierende Reihe $\sum_{k=1}^{\infty} (-1)^k e^{-k}$ auf Konvergenz.

Aufgabe 2.9

Untersuchen Sie die folgenden Reihen auf Konvergenz:

$$\text{a) } \sum_{k=1}^{\infty} \frac{5k}{5^k}, \quad \text{b) } \sum_{k=1}^{\infty} \frac{3^{2k}}{(2k)!}, \quad \text{c) } \sum_{k=1}^{\infty} k \cdot \left(\frac{1}{3}\right)^{k-1}, \quad \text{d) } \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k \cdot 5^{2k-1}}.$$

Aufgabe 2.10

Zeigen Sie, dass die folgenden (Funktionen-) Reihen (punktweise) für jedes $x \in \mathbb{R}$ konvergieren:

$$\text{a) } \sum_{k=1}^{\infty} \frac{1}{k^2 + x^4}, \quad \text{b) } \sum_{k=1}^{\infty} \frac{\sin(kx^2)}{k^{3/2}} \text{ für } x \in \mathbb{R}.$$

Aufgabe 2.11

Man berechne den Grenzwert von $\sum_{k=1}^{\infty} \frac{1}{4k^2 - 1}$.

Hinweis: In den Partialsummen addieren sich viele Terme zu 0 (Teleskopsumme).

Aufgabe 2.12

Gegeben ist die Funktion $f(x) = \begin{cases} \frac{|9-3x|}{3-x} & \text{für } x \neq 3, \\ 0 & \text{für } x = 3. \end{cases}$

Welche der folgenden Grenzwerte existieren, und welchen Wert haben sie?

a) $\lim_{x \rightarrow 0} f(x)$, **b)** $\lim_{x \rightarrow 3-} f(x)$, **c)** $\lim_{x \rightarrow 3+} f(x)$, **d)** $\lim_{x \rightarrow 3} f(x)$, **e)** $\lim_{x \rightarrow \infty} f(x)$.

Aufgabe 2.13

Für welche der folgenden Funktionen $f(x)$ existiert an der Stelle x_0 der rechtsseitige bzw. linksseitige Grenzwert? Ist $f(x)$ an der Stelle x_0 stetig bzw., lässt sich $f(x)$ an der Stelle x_0 stetig ergänzen?

$$\text{a) } f(x) = \frac{x}{|x|}, x_0 = 0, \quad \text{b) } f(x) = |x - 1|, x_0 = 1, \quad \text{c) } f(x) = \sqrt{1 - x}, x_0 = 1,$$

$$\text{d) } f(x) = \begin{cases} x^2 - 2 & \text{für } x < 0, \\ \ln\left(\frac{1+x}{e^2}\right) & \text{für } x > 0, \end{cases} \quad x_0 = 0.$$

Aufgabe 2.14

Zeigen Sie mit der ε - δ -Definition der Stetigkeit:

a) Die Funktion $f(x) = 4x - 4$ ist stetig für alle $x \in \mathbb{R}$.

b) Die Funktion $f(x) = \begin{cases} 2x^2 \cdot \sin\left(\frac{1}{x}\right), & x \neq 0, \\ 0, & x = 0, \end{cases}$ ist stetig in $x_0 = 0$.

c) Die Funktion $\sigma(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases}$ ist unstetig in $x_0 = 0$.

Aufgabe 2.15

Man berechne die Ableitungen der folgenden Funktionen:

$$f_1(x) = \frac{x^2 + 3x + 10}{x + 1}, \quad f_2(x) = \sin(x)[x^2 + 2x + 1], \quad f_3(x) = e^{1+\sin x}, \quad f_4(x) = x^{2 \cos x}.$$

Aufgabe 2.16

Differenzieren Sie folgende Funktionen $f(x)$ einmal nach x :

$$\begin{aligned} \text{a)} \quad & f(x) = (2 - 2x^2)^9, \quad \text{b)} \quad f(x) = (1 + x^2)(1 + x)^3, \quad \text{c)} \quad f(x) = 3\sqrt[4]{(1 - x)^7}, \\ \text{d)} \quad & f(x) = \frac{1 - x^2}{\sqrt{x}}, \quad \text{e)} \quad f(x) = \frac{1 - 2x}{1 + x}, \quad \text{f)} \quad f(x) = \frac{x^3}{(1 - 2x)^4}, \\ \text{g)} \quad & f(x) = \sqrt{1 + \sqrt{x}}, \quad \text{h)} \quad f(x) = \tan(1 - 3x^3), \quad \text{i)} \quad f(x) = e^{-4x^2}, \\ \text{j)} \quad & f(x) = (e^{3x})^5, \quad \text{k)} \quad f(x) = 3^{(3x)^5}, \quad \text{l)} \quad f(x) = \cos(x \ln x). \end{aligned}$$

Aufgabe 2.17

Untersuchen Sie, ob folgende Funktionen $f(x)$ an der Stelle x_0 differenzierbar sind:

$$\text{a)} \quad f(x) = |x + 1|, \quad x_0 = \pm 1, \quad \text{b)} \quad f(x) = \begin{cases} |x - 1| & \text{für } x \geq 0, \\ 1 + \cos(x + \frac{\pi}{2}) & \text{für } x < 0, \end{cases} \quad x_0 = 0.$$

Aufgabe 2.18

Differenzieren Sie folgende Funktionen nach x :

$$\text{a)} \quad \ln(xe^{-\cos(2x)}), \quad \text{b)} \quad \ln\left(\frac{\sin^2(x)}{\sqrt{x}}\right), \quad \text{c)} \quad x^{\cos(1-x)}, \quad \text{d)} \quad x^2 \arctan(\sqrt{x}), \quad \text{e)} \quad \frac{\sqrt{1-x}}{\sqrt{1+x}}.$$

Aufgabe 2.19

Bestimmen Sie die ersten Ableitungen der folgenden Funktionen für konstante Parameter $a, \delta, \omega \in \mathbb{R}$:

$$\begin{aligned} \text{a)} \quad & f(t) = a \sin(\omega t + \varphi), \quad \text{b)} \quad f(t) = a e^{-\delta t} \cos(\omega t + \varphi), \quad \text{c)} \quad f(\varphi) = \cos\left(\frac{1}{\varphi}\right), \\ \text{d)} \quad & f(\varphi) = a \sqrt{1 + \cos \varphi}. \end{aligned}$$

Aufgabe 2.20

Berechnen Sie die Ableitung der Funktion $f(x) = |e^x - 1|$, $x \in [-1, 1]$.

Aufgabe 2.21

Zeigen Sie, dass die Funktion $f(x) = 2x - \arctan x$, $x \in \mathbb{R}$, monoton steigend ist. Berechnen Sie an der Stelle $y_0 = 2 - \frac{\pi}{4}$ die Ableitung der Umkehrfunktion.

Aufgabe 2.22

- a) Stellen Sie die Gleichung der Tangente an den Funktionsgraphen von $f(x) = \frac{2}{\arctan x}$ im Punkt $(1, f(1))$ auf.
- b) Unter welchem Winkel α schneidet der Graph der Funktion $y = \frac{\ln(x^2)}{2x}$ die x -Achse?

Aufgabe 2.23

Zeigen Sie mittels Vollständiger Induktion die folgende Verallgemeinerung der Produktregel: Die Funktionen $f_1, \dots, f_n$ seien an der Stelle x_0 differenzierbar. Dann gilt:

$$\left[\prod_{k=1}^n f_k \right]'(x_0) = \sum_{k=1}^n f'_k(x_0) \cdot \prod_{\substack{i=1 \\ i \neq k}}^n f_i(x_0).$$

Aufgabe 2.24

Nach dem Mittelwertsatz gibt es im Intervall $[-1, 1]$ Stellen ξ , an denen die Funktion $f(x) = 2x^3 - 3x^2 - x + 5$ eine Tangente mit der gleichen Steigung wie die Gerade durch die Punkte $(-1, f(-1))$, $(1, f(1))$ hat. Bestimmen Sie alle diese Stellen.

Aufgabe 2.25

Berechnen Sie die folgenden Grenzwerte mit den L'Hospital'schen Regeln:

$$\begin{array}{lll} \text{a)} \lim_{x \rightarrow 0} \frac{\sin(4x)}{x}, & \text{b)} \lim_{x \rightarrow 0} \frac{1 - \cosh x}{2x^2}, & \text{c)} \lim_{x \rightarrow -2} \frac{x^9 + 2^9}{x^7 + 2^7}, \\ \text{d)} \lim_{x \rightarrow \frac{3\pi}{2}} \frac{\tan(3x)}{\tan(5x)}, & \text{e)} \lim_{x \rightarrow 0} \frac{1}{x} \operatorname{artanh}(x), & \text{f)} \lim_{x \rightarrow \infty} x \cdot \left(\cos\left(\frac{1}{x}\right) - 1 \right), \\ \text{g)} \lim_{x \rightarrow -\infty} x^2 \cdot e^x. \end{array}$$

Aufgabe 2.26

Mit dem Satz von L'Hospital berechne man a) $\lim_{x \rightarrow 0} \frac{e^x - 1}{\sin x}$, b) $\lim_{x \rightarrow \infty} \left(\frac{x+1}{x} \right)^{(x^2)}$.

Aufgabe 2.27

Man berechne das bestimmte Integral $\int_0^1 x^2 dx$ mit Riemann'schen Zwischensummen.

Hinweis: $\sum_{k=0}^n k^2 = \frac{n(n+1)(2n+1)}{6}.$

Aufgabe 2.28

Berechnen Sie mittels partieller Integration $\int_0^{\frac{\pi}{2}} x \cdot \cos(2x) dx$.

Aufgabe 2.29

Berechnen Sie mittels Substitution $t = x^2$ das Integral $\int_0^1 \exp(x^2) \cdot 2x dx$.

Aufgabe 2.30

Lösen Sie die folgenden Integrale mit einer geeigneten Substitution:

$$\begin{array}{lll} \text{a)} \int_0^2 \frac{-3}{3x+1} dx, & \text{b)} \int \frac{3}{e^{3x-1}} dx, & \text{c)} \int \frac{\sin x}{\cos^5 x} dx, \\ \text{d)} \int \frac{2e^x}{e^{2x} + 1} dx, & \text{e)} \int 3x^2(x^3 + 1)^3 dx. \end{array}$$

Aufgabe 2.31

Lösen Sie die folgenden Aufgaben mit Hilfe partieller Integration:

a) $\int x \cdot e^{-3x} dx$, b) $\int_1^e x^{10} \cdot \ln x dx$, c) $\int_0^\pi x^2 \cdot \cos x dx$, d) $\int_0^{\pi/2} x \cdot \sin(x) \cos(x) dx$.

Aufgabe 2.32

Lösen Sie die folgenden Integrale:

a) $\int_0^1 2x \cdot (x^2 + 1)^{100} dx$, b) $\int x \cdot (x + 1)^{1000} dx$, c) $\int \frac{1}{(x + 1)^{4711}} dx$,
d) $\int \frac{x - 2}{\sqrt{1 - x^2}} dx$, e) $\int \frac{1}{\sqrt{4 - x^2}} dx$.

Aufgabe 2.33

Berechnen Sie Partialbruchzerlegungen von

a) $\frac{3x - 8}{x^2 - 6x + 8}$, b) $\frac{x + 2}{(x - 2)^3}$, c) $\frac{x - 1}{x^2 + x - 6}$, d) $\frac{2x + 1}{(x - 2)^2}$.

Aufgabe 2.34

Berechnen Sie jeweils mittels Partialbruchzerlegung

$$\int_0^2 \frac{8x}{x^2 + 4x + 4} dx, \quad \int_2^3 \frac{2x + 1}{x^2 + x - 2} dx.$$

Aufgabe 2.35

Lösen Sie die folgenden Aufgaben mit Hilfe einer Partialbruchzerlegung:

a) $\int \frac{(2x)^3 + 8}{x^2} dx$, b) $\int \frac{2x^2 - 2^2}{(x + 2)^2(x - 1)} dx$, c) $\int \frac{x + 1}{(x - 1)(x + 2)^2} dx$,
d) $\int \frac{x^4}{x^2 + 3} dx$, e) $\int \frac{x + 4}{x^2 - 4x + 8} dx$, f) $\int \frac{16x - 4}{x^2 - 6x + 5} dx$,
g) $\int \frac{x^2 + 15x + 8}{x^3 - 3x^2 - 9x - 5} dx$.

Aufgabe 2.36

Gegeben sei die Funktion $f(x) = \frac{4}{(x+2)^2}$, $x \geq 0$.

- a) Legt man im Punkt $(0, f(0))$ eine Tangente an den Funktionsgraphen von $f(x)$, so schneidet diese die x -Achse an einer Stelle x_0 . Bestimmen Sie x_0 .
b) Der Funktionsgraph schließt zwischen $x = 0$ und $x = u$ mit der x -Achse ein endliches Flächenstück ein. Berechnen Sie den Inhalt $F(u)$ dieser Fläche, also eine Stammfunktion von f .
c) Welchen Wert hat das uneigentliche Integral $\int_0^\infty \frac{4}{(x+2)^2} dx = \lim_{u \rightarrow \infty} F(u)$?

Aufgabe 2.37

$f(x) = 4x^4 - 16x^2 + 16$ schließt mit der x -Achse eine endliche Fläche ein. Berechnen Sie ihre Größe.

Aufgabe 2.38

Berechnen Sie die uneigentlichen Integrale, sofern sie konvergieren:

a) $\int_{\frac{1}{2}}^{\infty} e^{-2x} dx$, b) $\int_0^1 \ln(x) + 2 dx$, c) $\int_0^{\infty} \frac{2x}{1+x^2} dx$, d) $\int_0^1 \frac{x}{\sqrt{|x-1|}} dx$.

Hinweis: Bei b) verwende man die Regel von L'Hospital.

Aufgabe 2.39

Entwickeln Sie die Funktion $y = f(x)$ an der Stelle x_0 in ein Taylor-Polynom vom Grad n :

a) $y = f(x) = \cos x$, $x_0 = \frac{\pi}{2}$, $n = 3$, b) $y = f(x) = x^5$, $x_0 = 2$, $n = 6$,

c) $y = f(x) = \sqrt{1+x}$, $x_0 = 0$, $n = 3$, d) $y = f(x) = \frac{1}{\sqrt{1+x}}$, $x_0 = 0$, $n = 2$.

Aufgabe 2.40

Führen Sie für die Funktion $y = f(x) = \frac{2x^3 - 8x}{x^2 + 1}$ eine Kurvendiskussion durch. Untersuchen Sie dazu den Definitionsbereich, Symmetrien, Unstetigkeitsstellen, asymptotisches Verhalten, Extrema, Wendepunkte und Schnittpunkte mit den Koordinatenachsen.

Aufgabe 2.41

Untersuchen Sie die Funktionen $f(x) = \sin(\ln x)$ und $f(x) = \frac{\cos(\ln x)}{x}$:

- Geben Sie jeweils den maximalen Definitionsbereich, Nullstellen und die Unstetigkeitsstellen an.
- Existieren $\lim_{x \rightarrow \infty} f(x)$ und $\lim_{x \rightarrow 0+} f(x)$? Begründen Sie Ihre Antwort.
- Berechnen Sie f' und f'' .
- Bestimmen Sie mit den zuvor berechneten Ableitungen die lokalen Minima von f .

Aufgabe 2.42

Untersuchen Sie die Funktion $f(x) = \exp(\sin^2(x) + 2 \sin(x) + 1) - 1$:

- Geben Sie den maximalen Definitionsbereich, Wertebereich, Nullstellen und die Unstetigkeitsstellen an.
- Begründen Sie, ob $\lim_{x \rightarrow \infty} f(x)$ und $\lim_{x \rightarrow -\infty} f(x)$ existieren.
- Berechnen Sie f' und f'' .
- Bestimmen Sie mit den zuvor berechneten Ableitungen mindestens eine Stelle, an der ein lokales Maximum von f liegt.

Aufgabe 2.43

Untersuchen Sie die beiden reellen Funktionen

$$f(x) = \ln \left(\frac{x-1}{\exp(x)} \right) \quad \text{und} \quad f(x) = \exp \left(\frac{x-1}{x^2+3} \right).$$

- a) Geben Sie jeweils den maximalen Definitionsbereich, die Nullstellen und die Unstetigkeitsstellen an. **Hinweis:** $e^x > x$.
- b) Berechnen Sie $\lim_{x \rightarrow \infty} f(x)$ und $\lim_{x \rightarrow 1+} f(x)$ bzw. $\lim_{x \rightarrow -\infty} f(x)$.
- c) Berechnen Sie f' und f'' .
- d) Bestimmen Sie mit den zuvor berechneten Ableitungen lokale Extrema der Funktionen. Wie lautet jeweils der Wertebereich?

Aufgabe 2.44

Welchen Durchmesser und welche Höhe hat ein Zylinder mit einem vorgegebenen Volumen V von genau einem Liter, wenn seine Oberfläche minimal sein soll?

Aufgabe 2.45

Beweisen Sie mittels des Quotientenkriteriums den folgenden Satz:

Sei $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ eine Potenzreihe mit Konvergenzradius ρ , so dass der Grenzwert $r := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|$ existiert. Dann gilt:

- a) Ist $r = \infty$, so ist $\rho = 0$. b) Ist $0 < r < \infty$, so ist $\rho = \frac{1}{r}$. c) Ist $r = 0$, so ist $\rho = \infty$.

Aufgabe 2.46

Bestimmen Sie den Konvergenzbereich der Potenzreihen (bei c) nur den Konvergenzradius)

$$\text{a) } \sum_{k=1}^{\infty} k x^k, \quad \text{b) } \sum_{k=1}^{\infty} \frac{(x-4)^k}{k^k}, \quad \text{c) } \sum_{k=0}^{\infty} \frac{k^k}{k!} x^k.$$

Aufgabe 2.47

Berechnen Sie den Konvergenzradius ρ der folgenden Potenzreihen um den Entwicklungsmittelpunkt 0:

$$\text{a) } \sum_{k=1}^{\infty} x^k = x + x^2 + x^3 + \dots, \quad \text{b) } \sum_{k=1}^{\infty} \frac{x^k}{k^2} = x + \frac{1}{4}x^2 + \frac{1}{9}x^3 + \dots, \quad \text{c) } \sum_{k=1}^{\infty} \frac{x^k}{k^k}.$$

Aufgabe 2.48

Man entwickle die folgenden Funktionen in eine Potenzreihe um 0 und bestimme den Konvergenzradius:

$$\text{a) } x^4 + 3x^2 + 3x, \quad \text{b) } \ln(x+1), \quad \text{c) } \sinh x = \frac{1}{2}(e^x - e^{-x}), \quad \text{d) } \frac{1}{x-x_0} \text{ für } x_0 \neq 0.$$

Aufgabe 2.49

Man untersuche die Potenzreihen um den Punkt 0 aus Aufgabe 2.47 und Aufgabe 2.48, für die der Konvergenzradius $\rho < \infty$ ist, auf Konvergenz in den Punkten $-\rho$ und ρ .

Aufgabe 2.50

Ist eine Funktion f in eine Potenzreihe um x_0 mit Konvergenzradius $\rho > 0$ entwickelbar, so ist f nach Satz 2.61 auf $]x_0 - \rho, x_0 + \rho[$ beliebig oft differenzierbar, und man erhält die ebenfalls auf $]x_0 - \rho, x_0 + \rho[$ konvergente Potenzreihe der Ableitung f' durch gliedweises Differenzieren. Berechnen Sie damit eine Potenzreihe für $\frac{1}{x+1}$.



3 Lineare Algebra

Übersicht

3.1	Vektoren in der Ebene und im Raum	427
3.2	Analytische Geometrie	452
3.3	Vektorräume	466
3.4	Lineare Abbildungen	500
3.5	Lösungstheorie linearer Gleichungssysteme	521
3.6	Eigenwerte und Eigenvektoren	537
3.7	Normierte Vektorräume: Lineare Algebra trifft Analysis*	557
3.8	Aufgaben	585

3.1 Vektoren in der Ebene und im Raum

Viele physikalische Größen sind durch Angabe eines reellen Werts bestimmt, wie z. B. die Masse, die Temperatur oder die Leistung. Man bezeichnet sie als **Skalare**. Feldstärken und Kräfte wirken in eine Richtung, und auch Geschwindigkeiten werden nicht allein durch ihren Betrag, sondern erst durch Angabe der Richtung eindeutig.

3.1.1 Vektoren: Grundbegriffe und elementare Rechenregeln

Definition 3.1 (Vektor)

Ein **Vektor** $\vec{a}$ ist eine Größe, die durch einen **Betrag** $|\vec{a}|$ und eine **Richtung** charakterisiert ist.

In der Ebene oder im Raum werden Vektoren durch Pfeile veranschaulicht. Ein Vektor $\vec{a}$ lässt sich durch einen Anfangspunkt A und einen Endpunkt B festlegen. Man schreibt $\vec{a} = \overrightarrow{AB}$. Dieser Pfeil zeigt von A nach B und hat damit eine Richtung. Den Begriff

Richtung verwenden wir anschaulich ohne weitere Definition. Den Abstand von A und B bezeichnet man als Betrag $|\vec{a}| \geq 0$. Der Betrag ist also die Länge des Pfeils.

Zwei Vektoren $\vec{a}$ und $\vec{b}$ sind **gleich**, $\vec{a} = \vec{b}$, wenn sie in Betrag und, falls der Betrag ungleich null ist, in der Richtung übereinstimmen. Mit dieser Festlegung kann man den Anfangspunkt eines Vektors beliebig wählen. Durch den Vektor ist der Endpunkt festgelegt. In Kapitel 3.3 werden wir den Begriff des Vektors noch wesentlich weiter und unabhängig von der Anschauung axiomatisch genau fassen. Wir werden alle mathematischen Objekte Vektor nennen, für die Rechenregeln gelten, die wir für die Pfeile identifizieren. Dazu gehören insbesondere auch die Matrizen, die wir bei linearen Gleichungssystemen verwendet haben, aber auch Funktionen.

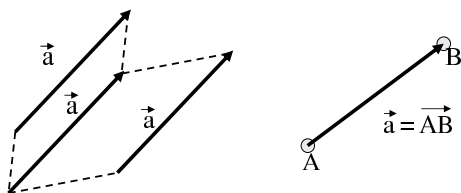


Abb. 3.1: Vektoren des $\mathbb{R}^2$ als Pfeile

Von besonderer Bedeutung sind die folgenden Vektoren.

- Ein Vektor $\vec{a}$ mit Betrag $|\vec{a}| = 0$ heißt **Nullvektor** $\vec{0}$. Dem Nullvektor wird keine Richtung zugeordnet.
- Ein Vektor $\vec{e}$ mit Betrag $|\vec{e}| = 1$ heißt **Einheitsvektor**.

Vektoren im Sinne von Zeigern haben wir schon bei Zeigerdiagrammen in Kapitel 1.5.9.5 kennengelernt. Mit ihnen konnten wir Wechselströme ausrechnen.

3.1.1.1 Vektorarithmetik

Um mit Vektoren zu rechnen, müssen wir zunächst die Rechenoperationen für Vektoren erklären. Diese sind für uns nicht neu, wir haben sie schon beim Rechnen mit Zeigerdiagrammen und Matrizen verwendet. Sie bilden das Verhalten physikalischer Vektoren wie Kräfte oder Feldstärken ab.

Definition 3.2 (Multiplikation mit einem Skalar, kollineare Vektoren)

Es seien $\vec{a}$ und $\vec{b}$ zwei Vektoren. Dann gilt:

- a) Der Vektor $\lambda \cdot \vec{a}$, $\lambda \in \mathbb{R}$, hat den $|\lambda|$ -fachen Betrag (Länge) von $\vec{a}$. Für $\lambda > 0$ zeigt $\lambda \cdot \vec{a}$ in dieselbe Richtung wie $\vec{a}$, für $\lambda < 0$ zeigt $\lambda \cdot \vec{a}$ in die zu $\vec{a}$ entgegengesetzte Richtung.
- b) Haben $\vec{a}$ und $\vec{b}$ gleiche oder entgegengesetzte Richtung oder ist ein Vektor $\vec{0}$, d. h., gilt $\vec{a} = \lambda \cdot \vec{b}$ oder $\vec{b} = \lambda \cdot \vec{a}$ mit $\lambda \in \mathbb{R}$, so heißen $\vec{a}$ und $\vec{b}$ zueinander **kollinear**.

Die Summe und die Differenz von zwei Vektoren können geometrisch festgelegt werden (vgl. Abbildung 3.2):

Definition 3.3 (Vektoraddition und Vektorsubtraktion)

- a) Für Vektoren $\vec{a}$ und $\vec{b}$ ist die **Summe** $\vec{a} + \vec{b}$ definiert als der Vektor, der der Diagonalen des Parallelogramms entspricht, das durch $\vec{a}$ und $\vec{b}$ aufgespannt wird. Um auch die Richtung des Vektors festzulegen, verwendet man als Anfangspunkt von $\vec{b}$ den Endpunkt von $\vec{a}$. Dann ist $\vec{a} + \vec{b}$ der Vektor mit dem Anfangspunkt von $\vec{a}$ und dem Endpunkt von $\vec{b}$.
- b) Die **Differenz** $\vec{a} - \vec{b}$ ist gleich der Summe der Vektoren $\vec{a}$ und $-\vec{b} := (-1) \cdot \vec{b}$.

Für die Vektoraddition und die Multiplikation mit einem Skalar können wir folgende Rechenregeln formulieren:

Satz 3.1 (Vektorarithmetik)

Für Vektoren $\vec{a}, \vec{b}, \vec{c}$ in der Ebene oder im Raum gelten:

- a) Regeln zur Vektoraddition:
- i) **Kommutativgesetz:** $\vec{a} + \vec{b} = \vec{b} + \vec{a}$.
 - ii) **Assoziativgesetz:** $\vec{a} + (\vec{b} + \vec{c}) = (\vec{a} + \vec{b}) + \vec{c}$.
 - iii) Es gibt ein **neutrales Element** $\vec{0}$, das für jeden Vektor $\vec{a}$ das eindeutige Element ist, für das $\vec{a} + \vec{0} = \vec{a}$ gilt.
 - iv) Zu jedem $\vec{a}$ existiert bezüglich der Addition genau ein **inverses Element** $-\vec{a}$, so dass $\vec{a} + (-\vec{a}) = \vec{0}$. Wir benutzen statt $\vec{b} + (-\vec{a})$ die Kurzschreibweise $\vec{b} - \vec{a}$.
 - v) Es gilt die **Dreiecksungleichung** $|\vec{a} + \vec{b}| \leq |\vec{a}| + |\vec{b}|$.
- b) Regeln zur Multiplikation mit einem Skalar:
- i) Zu jedem $\vec{a}$ und jedem $\lambda \in \mathbb{R}$ ist ein Produkt $\vec{b} = \lambda \cdot \vec{a} = \vec{a} \cdot \lambda$ erklärt.
 - ii) Das Produkt mit einem Skalar ist **assoziativ**: $(\lambda\mu) \cdot \vec{a} = \lambda \cdot (\mu \cdot \vec{a})$.
 - iii) $1 \cdot \vec{a} = \vec{a}$.
- c) Die Vektoraddition und Multiplikation mit einem Skalar erfüllen die **Distributivgesetze**

$$\lambda(\vec{a} + \vec{b}) = \lambda\vec{a} + \lambda\vec{b}, \quad (\lambda + \mu)\vec{a} = \lambda\vec{a} + \mu\vec{a}.$$

Das erste Distributivgesetz aus Satz 3.1 c) ist für Pfeile in Abbildung 3.4 dargestellt. Daraus können wir direkt für $r > 1$ die **Strahlensätze** ablesen: Zwei parallele Geraden schneiden zwei Strahlen mit gleichem Ursprung. Für die Streckenverhältnisse gilt mit den Bezeichnungen aus Abbildung 3.4:

- a) Erster Strahlensatz:

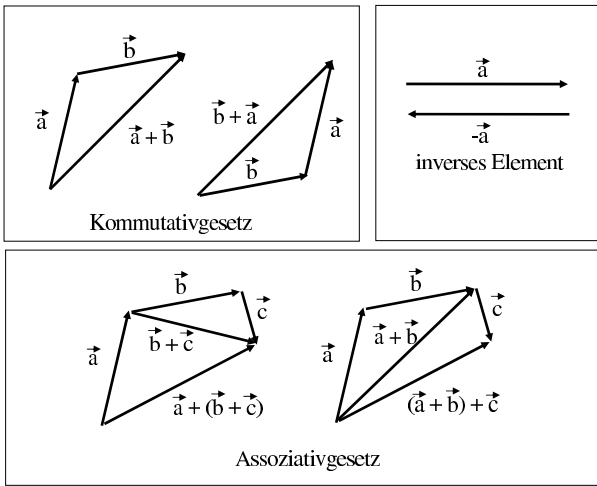


Abb. 3.2: Regeln der Vektorarithmetik

$$\frac{\overline{OC}}{\overline{OA}} = \frac{\overline{OD}}{\overline{OB}},$$

d. h., das Verhältnis der Abschnittslängen ist auf beiden Strahlen identisch. Es gilt nämlich

$$\frac{\overline{OC}}{\overline{OA}} = \frac{r|\vec{a}|}{|\vec{a}|} = r = \frac{r|\vec{a} + \vec{b}|}{|\vec{a} + \vec{b}|} = \frac{\overline{OD}}{\overline{OB}}.$$

b) Zweiter Strahlensatz:

$$\frac{\overline{CD}}{\overline{AB}} = \frac{\overline{OC}}{\overline{OA}},$$

d. h., das Verhältnis der Parallelenabschnitte entspricht dem Verhältnis der Abschnittslängen eines Strahls. Der Satz gilt, da

$$\frac{\overline{CD}}{\overline{AB}} = \frac{r|\vec{b}|}{|\vec{b}|} = r = \frac{r|\vec{a}|}{|\vec{a}|} = \frac{\overline{OC}}{\overline{OA}}.$$

3.1.1.2 Koordinaten und Komponenten von Vektoren in der Ebene und im Raum

Wir haben zwar Rechenregeln für Vektoren in Form von Pfeilen kennengelernt, aber damit kann man nicht gut rechnen. So wie wir Wechselstromzeiger durch komplexe Zahlen ersetzt haben, die ein Zahlenpaar aus Real- und Imaginärteil sind, stellen wir

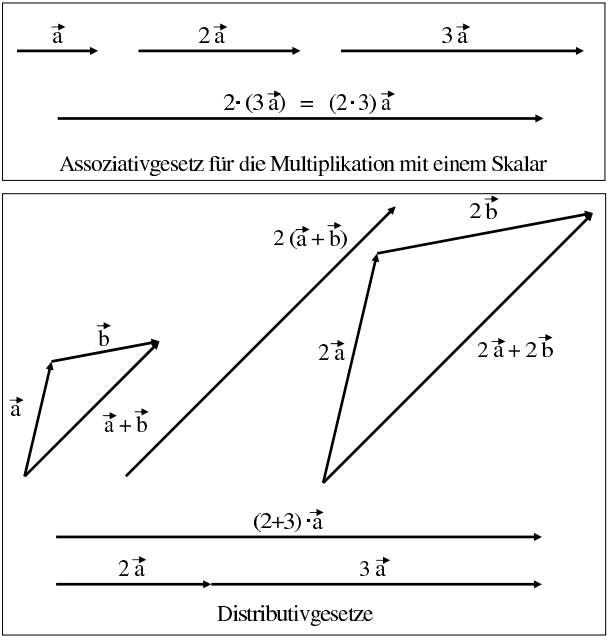


Abb. 3.3: Regeln der skalaren Multiplikation

hier die Zeiger der Ebene ebenfalls über Zahlenpaare und die Zeiger im dreidimensionalen Raum über Tupel mit drei Elementen dar. Dazu benötigen wir ein rechtwinkliges Koordinatensystem mit x -, y - und ggf. z -Achse. Bei drei Achsen verwendet man ein **Rechtssystem**, das über die „Rechte-Hand-Regel“ festgelegt ist: Zeigt der Daumen der rechten Hand in Richtung x und der Zeigefinger in Richtung y , so zeigt der gespreizte Mittelfinger in Richtung z . Diese Festlegung passt zum Vektorprodukt, das wir später einführen.

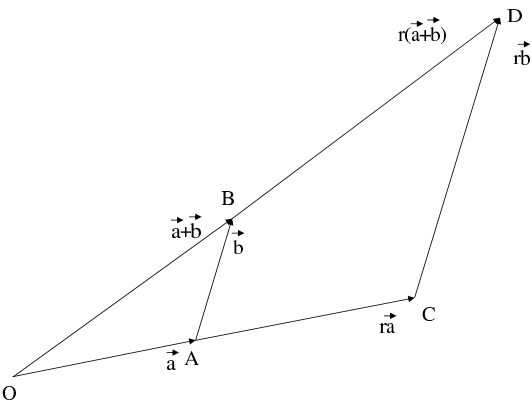


Abb. 3.4: Distributivgesetz als Strahlensatz

Jeden Punkt A der Ebene oder B des dreidimensionalen Raums können wir mit dem Koordinatensystem über eine x -, eine y - und ggf. eine z -**Koordinate** ausdrücken: $A = (x, y)$ bzw. $B = (x, y, z)$. Dabei können wir entweder eine Darstellung als Zeilen- oder alternativ als Spaltenmatrix wählen. Die Ebene entspricht somit der Menge

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y) : x \in \mathbb{R}, y \in \mathbb{R}\},$$

und der dreidimensionale Raum ist mit dem Kreuzprodukt darstellbar als $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$. Der Punkt $O = (0, 0)$ bzw. $O = (0, 0, 0)$ heißt der **Nullpunkt** oder **Ursprung**.

Jeden Punkt A identifizieren wir mit einem **Ortsvektor** $\vec{a} := \overrightarrow{OA}$. Das war zuvor in der Ebene der Zeiger zu einer komplexen Zahl. Jeder Pfeil ist als Ortsvektor und damit als Punkt darstellbar, da es nur auf seine Länge und Richtung ankommt. Der Vektor $\vec{a}$ von O zum Punkt $A = (x, y)$ der Ebene entspricht also einem Zahlenpaar

$$\vec{a} := \overrightarrow{OA} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2,$$

wobei wir x und y auch die **Koordinaten** von $\vec{a}$ nennen. Entsprechend sind Vektoren im dreidimensionalen Raum über drei Koordinaten als Elemente von $\mathbb{R}^3$ darstellbar.

Für Vektoren in Koordinatendarstellung verwenden wir bevorzugt die Spaltenschreibweise, benutzen aber auch die platzsparendere Zeilenschreibweise und transponieren ggf. den Zeilenvektor (siehe Definition 1.52 auf Seite 171). Auf das Transponieren verzichten wir bisweilen aber auch, da das Format nur dann wichtig wird, wenn Vektoren mit Matrizen von links oder rechts multipliziert werden sollen.

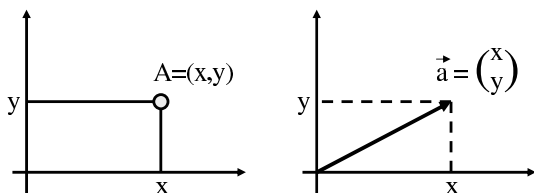


Abb. 3.5: Punkt und zugehöriger Ortsvektor im $\mathbb{R}^2$

Zwei Ortsvektoren sind genau dann gleich, wenn sie dieselben Koordinaten besitzen. Denn nur dann stimmen Länge und Richtung überein. Die Vektoraddition aus Definition 3.1 heißt, dass die Koordinaten des Summenvektors die Summen der einzelnen Koordinaten sind. In der Ebene gilt z. B. für $\vec{a} = (a_1, a_2)^\top$ und $\vec{b} = (b_1, b_2)^\top$ (vgl. Abbildung 3.6)

$$\vec{a} + \vec{b} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \end{pmatrix}.$$

Entsprechend erhält man mit $\lambda \in \mathbb{R}$ die Koordinaten des skalaren Vielfachen $\lambda \vec{a}$, indem man die Koordinaten von $\vec{a}$ mit λ multipliziert:

$$\lambda \cdot \vec{a} = \begin{pmatrix} \lambda \cdot a_1 \\ \lambda \cdot a_2 \end{pmatrix}.$$

Das ist konsistent mit den Rechenoperationen, die wir für Matrizen eingeführt haben.

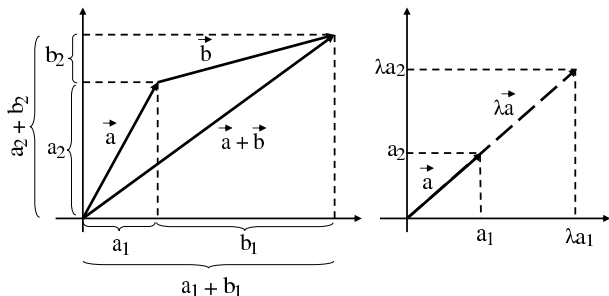


Abb. 3.6: Summe und skalares Vielfaches von Vektoren in Koordinatenschreibweise

Bis jetzt haben wir die koordinatenweise Addition und die skalare Multiplikation für Ortsvektoren betrachtet. Jeder Pfeil ist als Ortsvektor darstellbar, der im Nullpunkt O beginnt. Aber wir verbinden mit Pfeilen die Vorstellung, dass wir sie an einem beliebigen Punkt ansetzen können. Ein **freier Vektor** $\vec{v}$ wird durch Parallelverschiebung auf einen Ortsvektor zurückgeführt und besitzt dieselben Koordinaten wie der Ortsvektor.

Der freie Vektor $\vec{v} = \overrightarrow{AB}$ kann als Differenz zweier Ortsvektoren $\vec{v} = \overrightarrow{OB} - \overrightarrow{OA}$ geschrieben werden. In der Ebene erhalten wir so die Koordinaten

$$\vec{v} = \overrightarrow{OB} - \overrightarrow{OA} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} - \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} b_1 - a_1 \\ b_2 - a_2 \end{pmatrix}.$$

Ein Pfeil (Vektor) hat eine Darstellung als genau ein Ortsvektor oder als unendlich viele freie Vektoren. In der Mathematik fasst man alle diese Darstellungen zu einer sogenannten Äquivalenzklasse zusammen. Man rechnet mit Äquivalenzklassen, indem man mit einem beliebigen Repräsentanten der Klasse arbeitet.

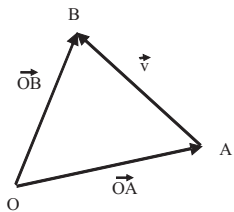


Abb. 3.7: Differenz von Vektoren

Die **Länge** oder der **Betrag** des Vektors $\vec{a} = (a_1, a_2)^\top$ in der Ebene ist wie bei einer komplexen Zahl (siehe Seite 136) mit dem Satz von Pythagoras gegeben durch

$$|\vec{a}| = \sqrt{a_1^2 + a_2^2}.$$

Die Länge der Strecke zwischen den Punkten A und B ist gleich dem Betrag des Vektors $\vec{v} = \overrightarrow{AB}$:

$$|\vec{v}| = |\overrightarrow{OB} - \overrightarrow{OA}| = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2}.$$

Völlig analog ergibt sich die Länge (bzw. der Betrag) eines Vektors $\vec{a} = (a_1, a_2, a_3)^\top$ über $|\vec{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}$.

Eine besondere Rolle kommt Vektoren in Richtung der Koordinatenachsen zu, die die Länge eins haben. Dies sind die **Standard-Einheitsvektoren** $\vec{e}_1, \vec{e}_2$ der Ebene bzw. $\vec{e}_1, \vec{e}_2, \vec{e}_3$ des dreidimensionalen Raums. Ihre Koordinatendarstellung lautet

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{bzw.} \quad \vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Ein beliebiger Vektor $\vec{a} \in \mathbb{R}^2$ (oder $\vec{a} \in \mathbb{R}^3$) kann als Summe aus Vielfachen dieser Standard-Einheitsvektoren geschrieben werden (vgl. Abbildung 3.8):

$$\vec{a} = a_1 \cdot \vec{e}_1 + a_2 \cdot \vec{e}_2 \quad (\vec{a} = a_1 \cdot \vec{e}_1 + a_2 \cdot \vec{e}_2 + a_3 \cdot \vec{e}_3).$$

In dieser Darstellung bezeichnet man die Vektoren $a_1 \vec{e}_1$ und $a_2 \vec{e}_2$ als **Komponenten**, und die Skalare a_1 und a_2 sind die Koordinaten von $\vec{a}$.

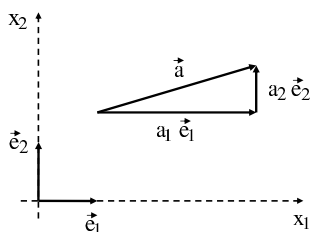


Abb. 3.8: Darstellung eines Vektors als Summe von Vielfachen der Standard-Einheitsvektoren

Beispiel 3.1

- a) Für die Vektoren $\vec{a} = (1, 2)^\top$ und $\vec{b} = (2, -5)^\top$ sind die Summe $\vec{a} + \vec{b}$ und die Differenz $\vec{a} - \vec{b}$ gegeben durch

$$\begin{aligned} \vec{a} + \vec{b} &= \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ -5 \end{pmatrix} = \begin{pmatrix} 1+2 \\ 2+(-5) \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \end{pmatrix}, \\ \vec{a} - \vec{b} &= \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ -5 \end{pmatrix} = \begin{pmatrix} -1 \\ 7 \end{pmatrix}. \end{aligned}$$

Die Beträge lauten $|\vec{a}| = \sqrt{1^2 + 2^2} = \sqrt{5}$ und $|\vec{b}| = \sqrt{2^2 + (-5)^2} = \sqrt{29}$.

- b) Für $\vec{a} = (1, 4)^\top$ und $\lambda = 3$ bzw. $\mu = -2$ erhält man

$$\lambda \cdot \vec{a} = 3 \cdot \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 3 \\ 12 \end{pmatrix}, \quad \mu \cdot \vec{a} = -2 \cdot \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} -2 \\ -8 \end{pmatrix}.$$

c) Für den Vektor $\vec{a} = 3\vec{e}_1 - 2\vec{e}_2$ lauten die Koordinaten $a_1 = 3$ und $a_2 = -2$.

Die Komponenten sind $\vec{a}_1 = 3\vec{e}_1$ und $\vec{a}_2 = -2\vec{e}_2$. Es gilt $|\vec{a}| = \sqrt{3^2 + (-2)^2} = \sqrt{13}$.

d) Für die Vektoren $\vec{a}$ und $\vec{b}$ in $\mathbb{R}^3$ mit den Koordinaten $a_1 = 1, a_2 = 1, a_3 = -1$ und $b_1 = 1, b_2 = 2, b_3 = 3$ hat der Vektor $\vec{c} = 2\vec{a} - \vec{b}$ die Koordinaten

$$\vec{c} = 2 \cdot \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}$$

bzw. in Komponentenform: $\vec{c} = \vec{e}_1 - 5\vec{e}_3$. Die Länge von $\vec{c}$ ist

$$|\vec{c}| = \sqrt{1^2 + 0^2 + (-5)^2} = \sqrt{26}.$$

■

3.1.2 Skalarprodukt und Orthogonalität

Das Skalarprodukt lässt sich durch den Begriff der mechanischen Arbeit aus der Physik motivieren: Ein Wagen bewegt sich längs des geraden Wegs $\vec{s}$ aufgrund der konstanten Zugkraft $\vec{F}$. Die Kraft $\vec{F}$ und der Weg $\vec{s}$ schließen den Winkel α ein (siehe Abbildung 3.9). Die verrichtete **Arbeit**, die geleistet wird, wenn der Wagen um $\vec{s}$ verschoben

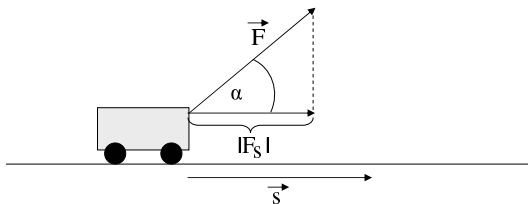


Abb. 3.9: Zur Berechnung der Arbeit als Skalarprodukt aus Kraft $\vec{F}$ und Weg $\vec{s}$

wird, berechnet sich als Produkt aus dem Anteil der Kraft in Richtung des Weges $F_s = |\vec{F}| \cos \alpha$ und dem zurückgelegten Weg $|\vec{s}|$:

$$W = |\vec{F}| \cos(\alpha) \cdot |\vec{s}| = |\vec{F}| \cdot |\vec{s}| \cdot \cos \alpha.$$

Diese Art der Multiplikation heißt Skalarprodukt der beiden Vektoren $\vec{F}$ und $\vec{s}$.

3.1.2.1 Definition des Skalarprodukts

Definition 3.4 (Skalarprodukt)

Es seien $\vec{a}$ und $\vec{b}$ Vektoren in der Ebene $\mathbb{R}^2$ oder im Raum $\mathbb{R}^3$, die den Winkel φ einschließen. Das **Skalarprodukt** (**inneres Produkt**) $\vec{a} \cdot \vec{b}$ ist definiert durch

$$\vec{a} \cdot \vec{b} = |\vec{a}| \cdot |\vec{b}| \cdot \cos \varphi. \quad (3.1)$$

In Worten: Das Skalarprodukt von $\vec{a}$ und $\vec{b}$ ist gleich dem Produkt aus den Beträgen beider Vektoren und dem Kosinus des eingeschlossenen Winkels φ . Ausgehend von $\vec{a}$ misst man einen Winkel im Gegenuhrzeigersinn (im mathematisch positiven Sinn) zum Vektor $\vec{b}$. Allerdings spielt hier wegen der Achsensymmetrie des Kosinus die Richtung keine Rolle.

Das Produkt $|\vec{b}| \cdot \cos \varphi$ entspricht der (vorzeichenbehafteten) Länge der **Projektion** des Vektors $\vec{b}$ in die Richtung von $\vec{a}$, siehe Abbildung 3.10.

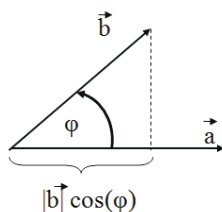


Abb. 3.10: Zur Definition des Skalarprodukts

Achtung: Das Skalarprodukt ist ein Produkt zwischen Vektoren und liefert als Ergebnis eine reelle Zahl. Dagegen ist das Ergebnis des Produkts eines Vektors mit einem Skalar (skalares Produkt) wieder ein Vektor.

Bereits das Vorzeichen des Skalarprodukts sagt viel über die Lage zweier Vektoren $\vec{a}, \vec{b} \neq \vec{0}$ aus:

- Für $0 < \varphi < \frac{\pi}{2}$ und $\frac{3}{2}\pi < \varphi < 2\pi$ gilt $\vec{a} \cdot \vec{b} > 0$, da $\cos \varphi > 0$ ist.
- Für $\frac{\pi}{2} < \varphi < \frac{3}{2}\pi$ gilt $\vec{a} \cdot \vec{b} < 0$, da $\cos \varphi < 0$ ist.
- Für $\varphi = 0$ haben $\vec{a}$ und $\vec{b}$ die gleiche Richtung, es gilt $\vec{a} \cdot \vec{b} = |\vec{a}| \cdot |\vec{b}|$. Speziell folgt

$$\vec{a} \cdot \vec{a} = |\vec{a}|^2 \quad \text{bzw.} \quad |\vec{a}| = \sqrt{\vec{a} \cdot \vec{a}}.$$

- Für $\varphi = \pi$ haben die Vektoren $\vec{a}$ und $\vec{b}$ entgegengesetzte Richtung, es gilt $\vec{a} \cdot \vec{b} = -|\vec{a}| \cdot |\vec{b}|$.
- Für $\varphi = \frac{\pi}{2}$ und $\varphi = \frac{3}{2}\pi$ stehen $\vec{a}$ und $\vec{b}$ aufeinander senkrecht, und es gilt $\vec{a} \cdot \vec{b} = 0$. Man nennt $\vec{a}$ und $\vec{b}$ dann **orthogonal** (vgl. Kapitel 3.1.2.4), und man schreibt $\vec{a} \perp \vec{b}$.

Achtung: Beim Skalarprodukt darf nicht wie beim Produkt von reellen Zahlen gekürzt werden: Man kann aus $\vec{a} \cdot \vec{b}_1 = \vec{a} \cdot \vec{b}_2$ nicht schließen, dass $\vec{b}_1 = \vec{b}_2$ ist. Lediglich die Anteile von $\vec{b}_1$ und von $\vec{b}_2$ in Richtung des Vektors $\vec{a}$ stimmen überein. Die Gleichung $\vec{a} \cdot \vec{b} = p$ lässt sich also bei gegebenem $\vec{a}$ und p nicht eindeutig nach $\vec{b}$ auflösen.

3.1.2.2 Rechenregeln, Koordinatenform und Winkelberechnung

Satz 3.2 (Rechenregeln des Skalarprodukts)

Es seien $\vec{a}, \vec{b}$ und $\vec{c}$ Vektoren in der Ebene $\mathbb{R}^2$ oder im Raum $\mathbb{R}^3$ und $r \in \mathbb{R}$. Dann gelten für das Skalarprodukt folgende Rechenregeln:

- a) **Kommutativgesetz:** $\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{a}$,
b) **Distributivgesetz:** $\vec{a} \cdot (\vec{b} + \vec{c}) = \vec{a} \cdot \vec{b} + \vec{a} \cdot \vec{c}$,
c) $\vec{0} \cdot \vec{a} = 0$ (hier ist es unerheblich, dass der Winkel φ nicht eindeutig ist),
d) **Cauchy-Schwarz'sche Ungleichung:** $|\vec{a} \cdot \vec{b}| \leq |\vec{a}| \cdot |\vec{b}|$,
e) **Homogenität:** $(r \cdot \vec{a}) \cdot \vec{b} = r \cdot (\vec{a} \cdot \vec{b})$.

Beweis: a) Das Kommutativgesetz folgt direkt aus der Definition, da Kosinus eine gerade Funktion ist und damit die Orientierung des eingeschlossenen Winkels unerheblich ist.

b) Das Distributivgesetz ist in Abbildung 3.11 veranschaulicht: Die Projektion von $\vec{b} + \vec{c}$ auf $\vec{a}$ ist gleich der Summe der Projektionen von $\vec{b}$ und $\vec{c}$ auf $\vec{a}$. Da das Skalarprodukt jeweils durch Multiplikation mit $\vec{a}$ entsteht, folgt die Gleichheit.

c) Dies folgt direkt aus der Definition und $|\vec{0}| = 0$.

d) Die Cauchy-Schwarz'sche Ungleichung folgt aus der Definition des Skalarprodukts und der Abschätzung $|\cos \varphi| \leq 1$:

$$|\vec{a} \cdot \vec{b}| = |\vec{a}| \cdot |\vec{b}| \cdot |\cos \varphi| \leq |\vec{a}| \cdot |\vec{b}|.$$

e) Für $r = 0$ ist nichts zu zeigen. Sei α der Winkel zwischen $\vec{a}$ und $\vec{b}$. Ist $r > 0$, so ist α auch der Winkel zwischen $r \cdot \vec{a}$ und $\vec{b}$ und

$$|(r \cdot \vec{a}) \cdot \vec{b}| = |r \cdot \vec{a}| \cdot |\vec{b}| \cdot \cos \alpha = r \cdot |\vec{a}| \cdot |\vec{b}| \cdot \cos \alpha = r \cdot (\vec{a} \cdot \vec{b}).$$

Ist $r < 0$, so ist der Winkel zwischen $r \cdot \vec{a}$ und $\vec{b}$ gleich $\alpha + \pi$, und es ergibt sich ebenfalls

$$|(r \cdot \vec{a}) \cdot \vec{b}| = |r| \cdot |\vec{a}| \cdot |\vec{b}| \cdot \cos(\alpha + \pi) = -|r| \cdot |\vec{a}| \cdot |\vec{b}| \cdot \cos \alpha = r \cdot (\vec{a} \cdot \vec{b}). \quad \square$$

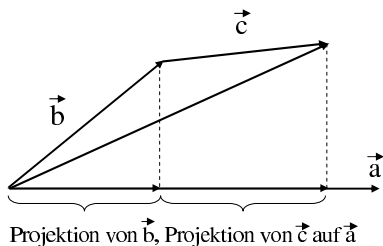


Abb. 3.11: Zum Beweis von Satz 3.2: Die Summe der Projektionen von $\vec{b}$ und $\vec{c}$ auf $\vec{a}$ ist gleich der Projektion von $\vec{b} + \vec{c}$ auf $\vec{a}$.

In $\mathbb{R}^2$ gelten die Beziehungen

$$\vec{e}_i \cdot \vec{e}_i = 1, \quad \vec{e}_i \cdot \vec{e}_k = 0 \quad \text{für } k \neq i, \quad i, k \in \{1, 2\}.$$

Mit dem Distributivgesetz erhält man unter Verwendung der Komponentendarstellung für das Skalarprodukt

$$\vec{a} \cdot \vec{b} = (a_1 \cdot \vec{e}_1 + a_2 \cdot \vec{e}_2) \cdot (b_1 \cdot \vec{e}_1 + b_2 \cdot \vec{e}_2)$$

$$\begin{aligned}
&= a_1 b_1 (\vec{e}_1 \cdot \vec{e}_1) + a_1 b_2 (\vec{e}_1 \cdot \vec{e}_2) + a_2 b_1 (\vec{e}_2 \cdot \vec{e}_1) + a_2 b_2 (\vec{e}_2 \cdot \vec{e}_2) \\
&= a_1 b_1 \cdot 1 + a_1 b_2 \cdot 0 + a_2 b_1 \cdot 0 + a_2 b_2 \cdot 1 = a_1 b_1 + a_2 b_2.
\end{aligned}$$

Damit haben wir gezeigt:

Satz 3.3 (Koordinatenform des Skalarprodukts in $\mathbb{R}^2$ und $\mathbb{R}^3$)

- *Das Skalarprodukt der Vektoren $\vec{a}$ und $\vec{b}$ in $\mathbb{R}^2$ wird berechnet, indem man $\vec{a}$ und $\vec{b}$ koordinatenweise multipliziert und die Produkte addiert:*

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2. \quad (3.2)$$

Für den Betrag von $\vec{a}$ gilt

$$|\vec{a}| = \sqrt{\vec{a} \cdot \vec{a}} = \sqrt{a_1^2 + a_2^2}. \quad (3.3)$$

- *Entsprechend gilt im Raum $\mathbb{R}^3$: Das Skalarprodukt der Vektoren $\vec{a}$ und $\vec{b}$ wird berechnet, indem man $\vec{a}$ und $\vec{b}$ koordinatenweise multipliziert und die Produkte addiert:*

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + a_3 b_3. \quad (3.4)$$

Für den Betrag von $\vec{a}$ gilt

$$|\vec{a}| = \sqrt{\vec{a} \cdot \vec{a}} = \sqrt{a_1^2 + a_2^2 + a_3^2}. \quad (3.5)$$

Mit Hilfe des Skalarprodukts kann man den **Winkel** zwischen zwei Vektoren $\vec{a}$ und $\vec{b}$ berechnen. Umstellung von (3.1) liefert

$$\cos \varphi = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} \quad \text{bzw.} \quad \varphi = \arccos \left(\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} \right).$$

Beispiel 3.2

a) Für die Vektoren

$$\vec{a} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}, \quad \vec{c} = \frac{1}{4} \cdot \begin{pmatrix} -4 \\ -4 \\ 3 \end{pmatrix}$$

berechnet man

$$\vec{a} \cdot \vec{b} = 1 \cdot 1 + 2 \cdot (-2) + 4 \cdot 2 = 5, \quad \vec{a} \cdot \vec{c} = \frac{1}{4} \cdot (1 \cdot (-4) + 2 \cdot (-4) + 4 \cdot 3) = 0.$$

Das erste Skalarprodukt ist positiv, d. h., die Vektoren $\vec{a}$ und $\vec{b}$ bilden einen spitzen Winkel. Die Vektoren $\vec{a}$ und $\vec{c}$ sind orthogonal.

b) Für den eingeschlossenen Winkel φ zwischen den Vektoren

$$\vec{a} = 4\vec{e}_1 - 3\vec{e}_3, \quad \vec{b} = \vec{e}_1 - 2\vec{e}_2 + 2\vec{e}_3$$

erhalten wir

$$\varphi = \arccos \left(\frac{4 \cdot 1 + 0 \cdot (-2) + (-3) \cdot 2}{\sqrt{4^2 + 0^2 + (-3)^2} \cdot \sqrt{1^2 + (-2)^2 + 2^2}} \right) = \arccos \left(-\frac{2}{15} \right) \approx 1,7045...$$

bzw. $\varphi \approx 97,66^\circ$. ■

3.1.2.3 Anwendungen des Skalarprodukts in der Geometrie

Mit dem Skalarprodukt lassen sich elementare Aussagen aus der Geometrie ableiten und beweisen.

Beispiel 3.3

Der Kosinus-Satz (siehe Seite 118) besagt, dass für ein Dreieck mit Seitenlängen a, b und c sowie dem Winkel γ zwischen den Seiten zu a und b als Verallgemeinerung des Satzes von Pythagoras gilt:

$$c^2 = a^2 + b^2 - 2ab \cos \gamma.$$

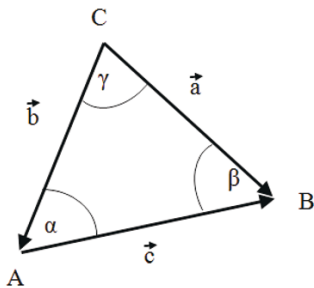


Abb. 3.12: Vektorieller Beweis des Kosinus-Satzes

Fassen wir die Seiten des Dreiecks als Vektoren $\vec{a}$, $\vec{b}$ und $\vec{c} = \vec{a} - \vec{b}$ auf (siehe Abbildung 3.12), so gestaltet sich nun der Beweis mit dem Skalarprodukt so:

$$|\vec{c}|^2 = (\vec{a} - \vec{b}) \cdot (\vec{a} - \vec{b}) = |\vec{a}|^2 - 2\vec{a} \cdot \vec{b} + |\vec{b}|^2 = |\vec{a}|^2 - 2|\vec{a}||\vec{b}|\cos(\gamma) + |\vec{b}|^2. \quad \blacksquare$$

Satz 3.4 (Satz von Thales (ca. 624–546 v. Chr.))

Verbindet man die Endpunkte der Grundlinie eines Halbkreises mit einem Punkt auf dem Halbkreis zu einem Dreieck, so ist dieses rechtwinklig. Diese Konstruktion wird als **Thaleskreis** bezeichnet.

Beweis: Mit den Vektoren aus Abbildung 3.13 gilt

$$\vec{a} = \vec{u} + \vec{v}, \quad \vec{b} = \vec{u} - \vec{v}, \quad |\vec{u}| = |\vec{v}| \quad (\text{Kreisradien}) \text{ und}$$

$$\vec{a} \cdot \vec{b} = (\vec{u} + \vec{v}) \cdot (\vec{u} - \vec{v}) = |\vec{u}|^2 - |\vec{v}|^2 = 0.$$

Wegen $\vec{a} \cdot \vec{b} = 0 \iff \vec{a} \perp \vec{b}$ folgt daraus die Behauptung des Satzes von Thales. $\square$

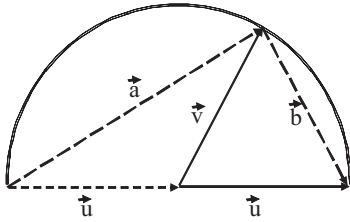


Abb. 3.13: Zum Beweis des Satzes von Thales

Beispiel 3.4 (Ausrichtung von Linien um eine Mittellinie)

In Katasteramtsdaten sind beispielsweise Mauern, Hecken und Gleise als einzelner Streckenzug dargestellt, der die Mitte beschreibt. Dabei sind Stützpunkte angegeben, die durch gerade Linien verbunden werden. Wenn wir die Objekte in einem virtuellen Stadtmodell zeichnen wollen, dann müssen wir um die vorhandene Mittellinie zwei Randlinien legen. Wie erhalten wir an den Stellen, an denen Knicke sind, die Stützpunkte für die Randlinien? Dazu betrachten wir einen Abschnitt, der vom Punkt (x_0, y_0) über den Punkt (x_1, y_1) zum Punkt (x_2, y_2) verläuft, siehe Abbildung 3.14. Zwischen (x_0, y_0) und (x_1, y_1) beschreibe der Einheitsvektor

$$\vec{a} := (y_0 - y_1, x_1 - x_0) / \sqrt{(y_0 - y_1)^2 + (x_1 - x_0)^2},$$

der senkrecht zur Mittellinie steht, den Abstand von der Mitte zum Rand. Entsprechend beschreibt der Einheitsvektor

$$\vec{b} := (y_1 - y_2, x_2 - x_1) / \sqrt{(y_1 - y_2)^2 + (x_2 - x_1)^2},$$

den Abstand zwischen Mitte und Rand im Sektor von (x_1, y_1) bis (x_2, y_2) . Daraus resultiert an der Knickstelle (x_1, y_1) die Richtung $\vec{a} + \vec{b}$ zum zu berechnenden Stützpunkt der Randlinie. Entlang dieses Vektors sei r der Abstand von der Mitte zum Rand. Damit ist $r = 1 / \cos(\alpha)$, wobei α der Winkel zwischen dem Einheitsvektor $\vec{a}$ und $\vec{a} + \vec{b}$ bzw. zwischen $\vec{a} + \vec{b}$ und dem Einheitsvektor $\vec{b}$ ist. Über das Standardskalarprodukt erhalten wir $\cos(\alpha)$:

$$\cos(\alpha) = \frac{\vec{a} \cdot (\vec{a} + \vec{b})}{|\vec{a} + \vec{b}|} = \frac{(\vec{a} + \vec{b}) \cdot \vec{b}}{|\vec{a} + \vec{b}|}.$$

Der Vektor von (x_1, y_1) zum Knick des Randes ist damit das r -fache des zur Länge 1 normierten Vektors $\vec{a} + \vec{b}$:

$$r \cdot \frac{\vec{a} + \vec{b}}{|\vec{a} + \vec{b}|} = \frac{|\vec{a} + \vec{b}|}{\vec{a} \cdot (\vec{a} + \vec{b})} \frac{\vec{a} + \vec{b}}{|\vec{a} + \vec{b}|} = \frac{1}{\vec{a} \cdot (\vec{a} + \vec{b})} (\vec{a} + \vec{b}).$$

■

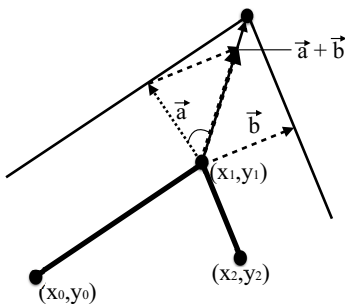


Abb. 3.14: Ausrichtung paralleler Linien: Jedes Gleis besteht aus zwölf Linien (vier Kanten des Bahndamms, vier Kanten pro Schiene), die parallel zur Oberleitung verlaufen.

3.1.2.4 Orthogonale Projektion und Lot

In der Physik, vor allem in der Mechanik, benötigt man oft die Zerlegung eines Vektors $\vec{b}$ in einen „Anteil“ $\vec{u}$, der zu einem gegebenen Vektor $\vec{a}$ parallel ist (d. h., $\vec{u}$ ist ein skalares Vielfaches von $\vec{a}$), und einen dazu senkrechten Anteil $\vec{v}$ gemäß

$$\vec{b} = \vec{u} + \vec{v}.$$

Ist beispielsweise $\vec{b}$ eine Kraft, so ist der Anteil dieser Kraft in Richtung von $\vec{a}$ der Vektor $\vec{u}$. Der Vektor $\vec{u}$ heißt **orthogonale Projektion** oder kurz **Projektion** von $\vec{b}$ auf $\vec{a}$, der Vektor $\vec{v} = \vec{b} - \vec{u}$ wird als **Lot** von $\vec{b}$ auf $\vec{a}$ bezeichnet. Beide Vektoren sind tatsächlich eindeutig festgelegt:

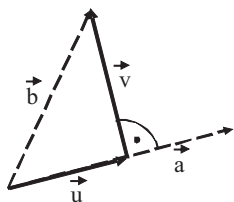


Abb. 3.15: Zerlegung des Vektors $\vec{b}$ in Projektion $\vec{u}$ und Lot $\vec{v}$ auf den Vektor $\vec{a}$

Satz 3.5 (Projektion und Lot)

Es seien $\vec{b}$ und $\vec{a} \neq \vec{0}$ Vektoren in der Ebene $\mathbb{R}^2$ oder im Raum $\mathbb{R}^3$. Der Vektor $\vec{b}$ lässt sich eindeutig als Summe eines zu $\vec{a}$ parallelen Anteils $\vec{u}$ und eines zu $\vec{a}$ senkrechten Anteils $\vec{v}$ schreiben: $\vec{b} = \vec{u} + \vec{v}$. Dabei sind $\vec{u}$ und $\vec{v}$ eindeutig festgelegt über:

$$\vec{u} = \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2} \vec{a} \quad \text{und} \quad \vec{v} = \vec{b} - \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2} \vec{a}.$$

Beweis: Offensichtlich gilt für die angegebenen Vektoren $\vec{u}$ und $\vec{v}$, dass $\vec{b} = \vec{u} + \vec{v}$. Als Vielfaches von $\vec{a}$ ist $\vec{u}$ parallel zu $\vec{a}$. Dass $\vec{v}$ senkrecht zu $\vec{a}$ steht, rechnen wir mit dem Skalarprodukt nach:

$$\vec{a} \cdot \vec{v} = \vec{a} \cdot \vec{b} - \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2} \vec{a} \cdot \vec{a} = \vec{a} \cdot \vec{b} - \vec{b} \cdot \vec{a} = 0.$$

Damit hat die angegebene Zerlegung die gewünschten Eigenschaften. Es bleibt die Eindeutigkeit zu zeigen. Ist $\vec{u}$ parallel zu $\vec{a}$, so handelt es sich um ein skalares Vielfaches $\vec{u} = \lambda \vec{a}$. Da in Summe $\vec{b}$ entsteht, ist $\vec{v} = \vec{b} - \lambda \vec{a}$. Da $\vec{v}$ senkrecht zu $\vec{a}$ steht, gilt:

$$0 = \vec{a} \cdot (\vec{b} - \lambda \vec{a}) = \vec{a} \cdot \vec{b} - \lambda |\vec{a}|^2,$$

so dass wie angegeben $\lambda = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}|^2}$ sein muss. □

Beispiel 3.5

Für die Vektoren $\vec{b} = (1, 1, 2)^\top$ und $\vec{a} = (2, 1, 0)^\top$ bestimmen wir die Zerlegung von $\vec{b}$ in Projektion und Lot bezüglich $\vec{a}$.

Wir berechnen $\vec{b} \cdot \vec{a} = 1 \cdot 2 + 1 \cdot 1 + 2 \cdot 0 = 3$ und $|\vec{a}|^2 = \vec{a} \cdot \vec{a} = 2^2 + 1^2 + 0^2 = 5$. Weiter folgt die Zerlegung $\vec{b} = \vec{u} + \vec{v}$ mit

$$\vec{u} = \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2} \vec{a} = \frac{3}{5} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{v} = \vec{b} - \vec{u} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} - \frac{3}{5} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{5} \begin{pmatrix} -1 \\ 2 \\ 10 \end{pmatrix}. \quad \blacksquare$$

3.1.3 Vektorprodukt und Spatprodukt

Neben dem Skalarprodukt werden zwei weitere Produkte von Vektoren für die Anwendungen benötigt.

3.1.3.1 Vektorprodukt

Wie beim Skalarprodukt gibt es auch für das Vektorprodukt eine Motivation aus der Mechanik. Wir betrachten eine Masse in einem Punkt P , der gemeinsam mit einem weiteren Punkt O in einer Ebene liegt. Senkrecht zur Ebene verläuft durch O eine Drehachse (siehe Abbildung 3.16). Die Masse ist an der Drehachse befestigt und kann nur in der vorgegebenen Ebene um O gedreht werden. Auf die Masse im Punkt P wirkt die Kraft $\vec{F}$, die über einen Vektor in der gewählten Ebene dargestellt werden kann. Sie hat also in diesem Beispiel keine Komponente parallel zum Verlauf der Drehachse. Die Masse wird durch die Kraft in eine Drehbewegung versetzt, man sagt: Die Kraft $\vec{F}$ erzeugt ein **Drehmoment** M um die Drehachse.

- Gilt nun $\vec{r} := \overrightarrow{OP} \perp \vec{F}$, so ist $M = |\vec{F}| \cdot |\vec{r}|$, d. h., es gilt das Hebelgesetz: „Kraft mal Kraftarm“.
- Zeigt $\vec{F}$ in Richtung von $\vec{r}$, dann wird kein Moment erzeugt, d. h., es ist $M = 0$.

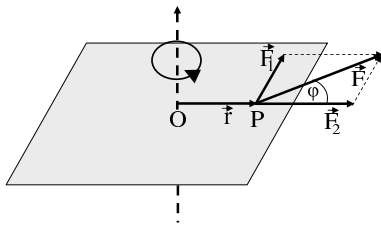


Abb. 3.16: Die Kraft $\vec{F}$ erzeugt ein Drehmoment, das einem Vektor in Richtung der Drehachse nach oben entspricht.

Im allgemeinen Fall schließen die Vektoren $\vec{r}$ und $\vec{F}$ einen Winkel φ ein. Hier zerlegt man $\vec{F}$ in das Lot $\vec{F}_1$ senkrecht zu $\overrightarrow{OP}$ und die Projektion $\vec{F}_2$ parallel zu $\overrightarrow{OP}$. Dann liefert nur $\vec{F}_1$ mit $|\vec{F}_1| = |\vec{F}| \sin(\varphi)$ einen Beitrag zum Moment:

$$M = |\vec{F}_1| \cdot |\vec{r}| = |\vec{F}| \sin(\varphi) \cdot |\vec{r}|.$$

Das Drehmoment M sagt noch nichts über die Richtung der Drehung aus. Möchten wir sowohl das Drehmoment als auch den Drehsinn in einer Größe angeben, so können wir einen **Drehmomentvektor** $\vec{M}$ einführen, dessen Betrag M ist und der parallel zur Drehachse liegt, also senkrecht zu der Ebene, in der sich $\vec{r}$ und $\vec{F}$ befinden. Die Orientierung von $\vec{M}$ wird so gewählt, dass $\vec{r}$, $\vec{F}$, $\vec{M}$ ein Rechtssystem bilden, siehe Seite 431, d. h., zeigt der Daumen in Richtung von $\vec{r}$ und der Zeigefinger in Richtung von $\vec{F}$, dann zeigt der gespreizte Mittelfinger in Richtung von $\vec{M}$.

Die Rechenoperation, die im Beispiel den Vektoren $\vec{r}$ und $\vec{F}$ den Vektor $\vec{M}$ zuordnet, nennt man Vektorprodukt, Kreuzprodukt oder auch äußeres Produkt. Es hat vielfältige Anwendungen, z. B. kann man damit auch die Richtung der Lorentzkraft berechnen, die auf eine bewegte Ladung in einem Magnetfeld wirkt.

Definition 3.5 (Vektorprodukt im $\mathbb{R}^3$)

Es seien $\vec{a}, \vec{b}$ Vektoren im Raum $\mathbb{R}^3$, die einen Winkel $\varphi \in [0, 2\pi[$ einschließen. Das **Vektorprodukt (äußere Produkt)**

$$\vec{c} = \vec{a} \times \vec{b}$$

ist der eindeutige **Vektor** mit folgenden Eigenschaften:

a) Der Betrag ist

$$|\vec{c}| = |\vec{a}| \cdot |\vec{b}| \cdot |\sin \varphi| \quad (3.6)$$

und entspricht damit dem Flächeninhalt des von $\vec{a}$ und $\vec{b}$ aufgespannten Parallelogramms,

b) $\vec{c}$ ist orthogonal zu $\vec{a}$ und $\vec{b}$,

c) $\vec{a}, \vec{b}$ und $\vec{c}$ bilden ein Rechtssystem.

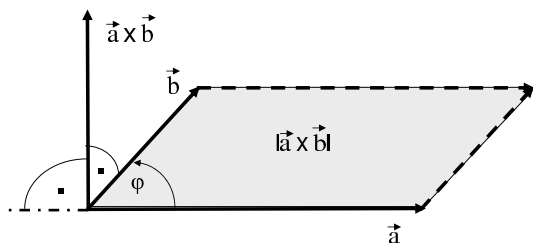


Abb. 3.17: Definition des Vektorprodukts im $\mathbb{R}^3$

Abbildung 3.17 veranschaulicht den Inhalt von Definition 3.5. Man beachte die Interpretation des Betrags von $\vec{a} \times \vec{b}$ als Maßzahl des Flächeninhalts des von $\vec{a}$ und $\vec{b}$ aufgespannten Parallelogramms.

Für das Vektorprodukt gelten folgende Rechenregeln:

Satz 3.6 (Rechenregeln des Vektorprodukts)

Für die Vektoren $\vec{a}, \vec{b}, \vec{c} \in \mathbb{R}^3$ gelten die folgenden Rechenregeln:

a) **Anti-Kommutativgesetz:** $\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$,

b) **Distributivgesetz:** $\vec{a} \times (\vec{b} + \vec{c}) = \vec{a} \times \vec{b} + \vec{a} \times \vec{c}$, $(\vec{a} + \vec{b}) \times \vec{c} = \vec{a} \times \vec{c} + \vec{b} \times \vec{c}$,

c) $\vec{a} \times \vec{a} = \vec{a} \times \vec{0} = \vec{0}$ (bei $\vec{0}$ spielt wieder die fehlende Eindeutigkeit von φ keine Rolle),

d) $\vec{a} \times \vec{b} = \vec{0} \iff \vec{a}$ und $\vec{b}$ sind parallel, oder einer der beiden Vektoren ist der Nullvektor,

e) $\vec{a} \times (\lambda \cdot \vec{b}) = \lambda \cdot \vec{a} \times \vec{b}$, $\lambda \in \mathbb{R}$.

- Beweis: (Skizze) a)** Vertauscht man Daumen und Zeigefinger, so zeigt der Mittelfinger in die entgegengesetzte Richtung wie zuvor.
- b)** Die Distributivgesetze sind schwieriger nachzuweisen als beim Skalarprodukt und werden hier nicht ausgeführt.
- c)** Der Winkel zwischen $\vec{a}$ und $\vec{a}$ ist $\varphi = 0$, so dass nach Definition $\vec{a} \times \vec{a} = \vec{0}$. Wegen $|\vec{0}| = 0$ folgt auch $\vec{a} \times \vec{0} = \vec{0}$ mit der Definition.
- d)** $\vec{a} \times \vec{b}$ ist genau dann $\vec{0}$, wenn die Länge eines Vektors null ist oder der Sinus des Winkels null ergibt.
- e)** Unter Berücksichtigung, dass sich durch Multiplikation mit einer negativen Zahl die Richtung von $\vec{b}$ ändert, folgt dies unmittelbar aus der Definition. $\square$

Damit erhalten wir für die Vektorprodukte der Standard-Einheitsvektoren:

$$\vec{e}_i \times \vec{e}_i = \vec{0}, \quad \vec{e}_1 \times \vec{e}_2 = -\vec{e}_2 \times \vec{e}_1 = \vec{e}_3, \quad \vec{e}_2 \times \vec{e}_3 = -\vec{e}_3 \times \vec{e}_2 = \vec{e}_1, \quad \vec{e}_3 \times \vec{e}_1 = -\vec{e}_1 \times \vec{e}_3 = \vec{e}_2.$$

Unter Verwendung der Komponentendarstellung der Vektoren $\vec{a} \times \vec{b}$ ergibt sich mit den Rechenregeln eine praktisch viel einfachere Möglichkeit zur Berechnung des Vektorprodukts:

$$\begin{aligned} \vec{a} \times \vec{b} &= (a_1 \cdot \vec{e}_1 + a_2 \cdot \vec{e}_2 + a_3 \cdot \vec{e}_3) \times (b_1 \cdot \vec{e}_1 + b_2 \cdot \vec{e}_2 + b_3 \cdot \vec{e}_3) \\ &= a_1 b_1 \cdot \vec{e}_1 \times \vec{e}_1 + a_1 b_2 \cdot \vec{e}_1 \times \vec{e}_2 + \cdots + a_3 b_3 \cdot \vec{e}_3 \times \vec{e}_3 \\ &= a_1 b_2 \cdot \vec{e}_3 - a_1 b_3 \cdot \vec{e}_2 - a_2 b_1 \cdot \vec{e}_3 + a_2 b_3 \cdot \vec{e}_1 + a_3 b_1 \cdot \vec{e}_2 - a_3 b_2 \cdot \vec{e}_1 \\ &= (a_2 b_3 - a_3 b_2) \cdot \vec{e}_1 + (a_3 b_1 - a_1 b_3) \cdot \vec{e}_2 + (a_1 b_2 - a_2 b_1) \cdot \vec{e}_3. \end{aligned}$$

Satz 3.7 (Koordinatenform des Vektorprodukts)

Für zwei Vektoren $\vec{a}, \vec{b} \in \mathbb{R}^3$ ist

$$\vec{a} \times \vec{b} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix}.$$

Dieses Ergebnis lässt sich über das formale Ausrechnen von Determinanten merken:

$$\vec{a} \times \vec{b} = \begin{vmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \begin{pmatrix} \begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix} \\ \begin{vmatrix} a_3 & b_3 \\ a_1 & b_1 \end{vmatrix} \\ \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \end{pmatrix} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix}.$$

Beispiel 3.6

Für die Vektoren $\vec{a} = 2\vec{e}_1 - \vec{e}_2 + 3\vec{e}_3$ und $\vec{b} = -\vec{e}_2 + 4\vec{e}_3$ erhalten wir

$$\begin{aligned}\vec{c} &:= \vec{a} \times \vec{b} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix} \times \begin{pmatrix} 0 \\ -1 \\ 4 \end{pmatrix} = \begin{pmatrix} (-1) \cdot 4 - 3 \cdot (-1) \\ 3 \cdot 0 - 2 \cdot 4 \\ 2 \cdot (-1) - (-1) \cdot 0 \end{pmatrix} = \begin{pmatrix} -1 \\ -8 \\ -2 \end{pmatrix} \\ &= -\vec{e}_1 - 8\vec{e}_2 - 2\vec{e}_3.\end{aligned}$$

■

Hintergrund: Vektorprodukt über die Cramer'sche Regel erklärt

Zu zwei linear unabhängigen Vektoren (x_1, y_1, z_1) und (x_2, y_2, z_2) ist ein dritter Vektor (x_3, y_3, z_3) gesucht, der senkrecht zu den beiden gegebenen steht. Die Skalarprodukte mit diesem Vektor müssen also null sein:

$$\begin{aligned}x_1x_3 + y_1y_3 + z_1z_3 &= 0 \\ \wedge \quad x_2x_3 + y_2y_3 + z_2z_3 &= 0.\end{aligned}$$

Damit haben wir bereits zwei Gleichungen für die drei Unbekannten x_3, y_3 und z_3 . Als dritte Gleichung verlangen wir, dass für ein $a \neq 0$

$$\begin{aligned}\det \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix} &= a \\ \Leftrightarrow \det \begin{bmatrix} y_1 & z_1 \\ y_2 & z_2 \end{bmatrix} \cdot x_3 - \det \begin{bmatrix} x_1 & z_1 \\ x_2 & z_2 \end{bmatrix} \cdot y_3 + \det \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix} \cdot z_3 &= a\end{aligned}$$

ist. Das von den drei Zeilenvektoren aufgespannte Parallelepipiped soll somit das Volumen $|a|$ haben. Zudem wird die Richtung (das Vorzeichen) des zu den beiden anderen senkrecht stehenden Vektors (x_3, y_3, z_3) festgelegt. Insgesamt haben wir nun ein lineares Gleichungssystem mit drei Gleichungen. Dieses hat die Matrix

$$\mathbf{A} := \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \det \begin{bmatrix} y_1 & z_1 \\ y_2 & z_2 \end{bmatrix} & -\det \begin{bmatrix} x_1 & z_1 \\ x_2 & z_2 \end{bmatrix} & \det \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix} \end{bmatrix}$$

und die Inhomogenität $(0, 0, a)^\top$. Wählen wir speziell $a := \det(\mathbf{A})$, dann erhalten wir die Lösung mit der Cramer'schen-Regel (auf das Nachrechnen, dass die Determinante tatsächlich ungleich null ist, verzichten wir):

$$\begin{aligned}
 x_3 &= \frac{\begin{vmatrix} 0 & y_1 & z_1 \\ 0 & y_2 & z_2 \\ \det(\mathbf{A}) & - \begin{vmatrix} x_1 & z_1 \\ x_2 & z_2 \end{vmatrix} & \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} \end{vmatrix}}{\det(\mathbf{A})} = \frac{\begin{vmatrix} y_1 & z_1 \\ y_2 & z_2 \end{vmatrix}}{\det(\mathbf{A})} = y_1 z_2 - z_1 y_2, \\
 y_3 &= \frac{\begin{vmatrix} x_1 & 0 & z_1 \\ x_2 & 0 & z_2 \\ \begin{vmatrix} y_1 & z_1 \\ y_2 & z_2 \end{vmatrix} & \det(\mathbf{A}) & \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} \end{vmatrix}}{\det(\mathbf{A})} = - \frac{\begin{vmatrix} x_1 & z_1 \\ x_2 & z_2 \end{vmatrix}}{\det(\mathbf{A})} = z_1 x_2 - x_1 z_2, \\
 z_3 &= \frac{\begin{vmatrix} x_1 & y_1 & 0 \\ x_2 & y_2 & 0 \\ \begin{vmatrix} y_1 & z_1 \\ y_2 & z_2 \end{vmatrix} & - \begin{vmatrix} x_1 & z_1 \\ x_2 & z_2 \end{vmatrix} & \det(\mathbf{A}) \end{vmatrix}}{\det(\mathbf{A})} = \frac{\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix}}{\det(\mathbf{A})} = x_1 y_2 - x_2 y_1.
 \end{aligned}$$

Dies ist genau der Vektor, den man über das Vektorprodukt von (x_1, y_1, z_1) und (x_2, y_2, z_2) erhält. Jetzt sehen wir uns noch an, was die Wahl $a = \det(\mathbf{A})$ bedeutet. Wenn wir diese Determinante nach der letzten Zeile entwickeln, dann erhalten wir $\det(\mathbf{A}) = x_3^2 + y_3^2 + z_3^2 = |(x_3, y_3, z_3)|^2$ als Volumen des von (x_1, y_1, z_1) , (x_2, y_2, z_2) und (x_3, y_3, z_3) aufgespannten Parallelepipeds. Aufgrund der konstruierten Orthogonalität hat das von (x_1, y_1, z_1) , (x_2, y_2, z_2) aufgespannte Parallelogramm den Flächeninhalt $|x_3, y_3, z_3|^2 / |(x_3, y_3, z_3)| = |(x_3, y_3, z_3)|$, wie in der Definition des Vektorprodukts gefordert.

Beispiel 3.7 (Orientierung eines Dreiecks)

Beim Zeichnen von Objekten mit dem Computer ist es wichtig zu wissen, welche Flächen sichtbar und welche unsichtbar sind. Soll eine Fläche nur gezeichnet werden, wenn man sie von der einen, nicht aber von der anderen Seite sieht, so kann der Umlaufsinn der Eckpunkte genutzt werden. Dieser kehrt sich bei Betrachtung von der anderen Seite um. Nur einseitig sichtbare Flächen liegen z. B. bei Oberflächen geschlossener Körper vor. Ist der Umlaufsinn mathematisch positiv, dann wird gezeichnet, sonst ist die Fläche vom Betrachter abgewandt und kann weggelassen werden. Wir betrachten in der Bildebene das Dreieck mit den Eckpunkten (x_0, y_0) , (x_1, y_1) und (x_2, y_2) . Durch die Reihenfolge der Eckpunkte ist der Umlaufsinn vorgegeben. Die beiden Vektoren

$$\vec{a} := (x_0 - x_1, y_0 - y_1, 0), \quad \vec{b} := (x_2 - x_1, y_2 - y_1, 0),$$

zeigen ausgehend von $(x_1, y_1, 0)$ nach $(x_0, y_0, 0)$ und $(x_2, y_2, 0)$. Gemäß der Rechte-Hand-Regel erhalten wir über das Vektorprodukt $\vec{a} \times \vec{b} \in \mathbb{R}^3$ einen Vektor, der den Umlaufsinn des Dreiecks anzeigt:

- Ist die dritte Komponente von $\vec{a} \times \vec{b}$ größer null, so wird das Dreieck im Uhrzeigersinn, d. h. im mathematisch negativen Sinn, umlaufen.
- Ist die dritte Komponente kleiner null, so wird das Dreieck entgegen des Uhrzeigersinns, d. h. im mathematisch positiven Sinn, umlaufen.

■

3.1.3.2 Spatprodukt

Drei Vektoren $\vec{a}$, $\vec{b}$ und $\vec{c}$, die nicht in einer Ebene liegen, spannen einen **Spat** oder ein **Parallelepiped** auf (siehe Abbildung 3.18).

Definition 3.6 (Spatprodukt)

Für die Vektoren $\vec{a}, \vec{b}, \vec{c} \in \mathbb{R}^3$ heißt die Zahl

$$[\vec{a}, \vec{b}, \vec{c}] := (\vec{a} \times \vec{b}) \cdot \vec{c} \quad (3.7)$$

das **Spatprodukt**.

Der Betrag des Spatprodukts ist das Volumen des von den drei Vektoren aufgespannten Spats. Denn $|\vec{a} \times \vec{b}|$ gibt den Inhalt der Grundfläche des durch $\vec{a}$ und $\vec{b}$ aufgespannten Spats an. Die Länge der Projektion von $\vec{c}$ auf $\vec{a} \times \vec{b}$ ist die Höhe

$$h = \left| \frac{\vec{c} \cdot (\vec{a} \times \vec{b})}{|\vec{a} \times \vec{b}|^2} (\vec{a} \times \vec{b}) \right| \quad (3.8)$$

des Spats. Das Volumen ist somit

$$|\vec{a} \times \vec{b}| \cdot \frac{|\vec{c} \cdot (\vec{a} \times \vec{b})|}{|\vec{a} \times \vec{b}|^2} |\vec{a} \times \vec{b}| = |\vec{c} \cdot (\vec{a} \times \vec{b})| = |(\vec{a} \times \vec{b}) \cdot \vec{c}|.$$

Bilden $\vec{a}, \vec{b}, \vec{c}$ ein Rechtssystem, so ist das Spatprodukt nicht-negativ, und auf den Betrag kann verzichtet werden.

Rechnen wir das Spatprodukt

$$[\vec{a}, \vec{b}, \vec{c}] = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{matrix} a_1 b_2 c_3 + a_2 b_3 c_1 + a_3 b_1 c_2 \\ -a_3 b_2 c_1 - a_1 b_3 c_2 - a_2 b_1 c_3 \end{matrix}$$

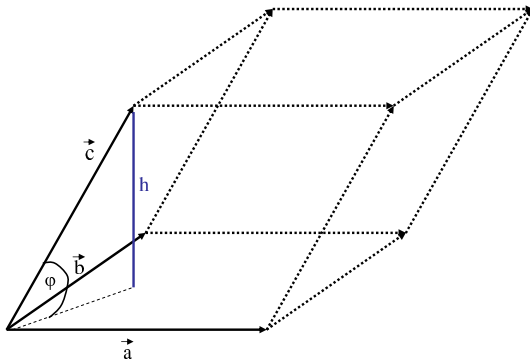


Abb. 3.18: Durch die Vektoren $\vec{a}$, $\vec{b}$ und $\vec{c}$ aufgespanntes Parallellepiped (Spat)

in dieser allgemeinen Form aus, so erkennen wir die Sarrus'sche Regel wieder. Das Spatprodukt lässt sich als Determinante einer (3×3) -Matrix berechnen:

$$[\vec{a}, \vec{b}, \vec{c}] = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}.$$

Diese Darstellung eignet sich als Merkmittel für die Anwendung.

Satz 3.8 (Rechenregeln des Spatprodukts)

Für Vektoren $\vec{a}, \vec{b}$ und $\vec{c}$ im Raum $\mathbb{R}^3$ und das Spatprodukt gelten folgende Rechenregeln:

- a) $[\vec{a}, \vec{b}, \vec{c}] = [\vec{c}, \vec{a}, \vec{b}] = -[\vec{a}, \vec{c}, \vec{b}] = -[\vec{c}, \vec{b}, \vec{a}] = [\vec{b}, \vec{c}, \vec{a}],$
- b) $[\vec{a}, \vec{b}, \vec{c}] = 0 \iff \vec{a}, \vec{b} \text{ und } \vec{c} \text{ liegen in einer Ebene, wobei weder } \vec{a} \text{ noch } \vec{b} \text{ noch } \vec{c} \text{ der Nullvektor sein muss (aber sein kann).}$

Beweis: a) Eine Determinante ändert ihr Vorzeichen, wenn man zwei Zeilen oder Spalten vertauscht (siehe Folgerung 1.57 d) auf Seite 182). Damit folgen die Gleichungen aus der obigen Determinanten-Darstellung des Spatprodukts.

b) Das Volumen des von den Vektoren aufgespannten Spats (und damit das Spatprodukt) ist genau dann null, wenn die Vektoren in einer Ebene liegen.

□

Beispiel 3.8

Mit dem Spatprodukt kann einfach nachgeprüft werden, ob drei Vektoren in einer Ebene liegen:

a) Die drei Vektoren $\vec{a}_1 = (3, 2, -1)^\top$, $\vec{a}_2 = (-3, -4, 2)^\top$ und $\vec{a}_3 = (12, 6, -3)^\top$ liegen

in einer Ebene, da
$$\begin{vmatrix} 3 & 2 & -1 \\ -3 & -4 & 2 \\ 12 & 6 & -3 \end{vmatrix} = 0.$$

b) $\vec{b}_1 = (1, 1, 4)^\top$, $\vec{b}_2 = (1, -2, 1)^\top$ und $\vec{b}_3 = (3, -3, 4)^\top$ liegen nicht in einer Ebene,

da
$$[\vec{b}_1, \vec{b}_2, \vec{b}_3] = \begin{vmatrix} 1 & 1 & 4 \\ 1 & -2 & 1 \\ 3 & -3 & 4 \end{vmatrix} = 6 \neq 0.$$

■

3.1.4 Anwendungen des Skalar-, Vektor- und Spatprodukts

Beispiel 3.9 (Sinus-Satz)

Auf Seite 119 haben wir den Sinus-Satz bewiesen: In einem Dreieck mit Seitenlängen a , b und c sowie den Seiten gegenüberliegenden Winkeln α , β und γ gilt

$$\frac{a}{\sin \alpha} = \frac{b}{\sin \beta} = \frac{c}{\sin \gamma}.$$

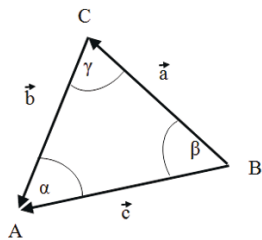


Abb. 3.19: Vektorieller Beweis des Sinussatzes

Nun wiederholen wir den Beweis unter Verwendung des Vektorprodukts. Dazu fassen wir die Seiten als Vektoren auf, wobei $\vec{c} = \vec{a} + \vec{b}$ sei. Das Vektorprodukt mit $\vec{c}$ von rechts ergibt (beachte $\vec{c} \times \vec{c} = \vec{0}$)

$$\vec{0} = (\vec{a} \times \vec{c}) + (\vec{b} \times \vec{c}) \iff \vec{a} \times \vec{c} = \vec{c} \times \vec{b}.$$

Bei gleichen Vektoren müssen die Beträge gleich sein (beachte $0 < \alpha, \beta < \pi$, $\sin \alpha > 0$, $\sin \beta > 0$):

$$|\vec{a}||\vec{c}| \sin \beta = |\vec{b}||\vec{c}| \sin \alpha.$$

Damit ergibt sich eine Aussage des Sinussatzes:

$$\frac{|\vec{a}|}{\sin \alpha} = \frac{|\vec{b}|}{\sin \beta}.$$

Die zweite Gleichung folgt analog.

Beispiel 3.10 (Dreiecksberechnung)

Die Punkte $A = (1, 0, 1)$, $B = (2, 1, 0)$ und $C = (1, 1, 0)$ sind die Ecken eines räumlichen Dreiecks. Den Punkten in $\mathbb{R}^3$ entsprechen die Ortsvektoren

$$\vec{OA} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{OB} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{OC} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Die Seitenvektoren und Seitenlängen des Dreiecks sind

$$\begin{aligned} \vec{a} &= \vec{OC} - \vec{OB} = (-1, 0, 0)^\top \quad \text{mit} \quad |\vec{a}| = 1, \\ \vec{b} &= \vec{OA} - \vec{OC} = (0, -1, 1)^\top \quad \text{mit} \quad |\vec{b}| = \sqrt{2}, \\ \vec{c} &= \vec{OB} - \vec{OA} = (1, 1, -1)^\top \quad \text{mit} \quad |\vec{c}| = \sqrt{3}. \end{aligned}$$

Zur Kontrolle prüfen wir, dass die Summe der Vektoren den Nullvektor ergibt: $\vec{a} + \vec{b} + \vec{c} = (0, 0, 0)^\top$. Für die Winkel zwischen den Schenkeln des Dreiecks (siehe Abbildung 3.20) erhalten wir über das Skalarprodukt

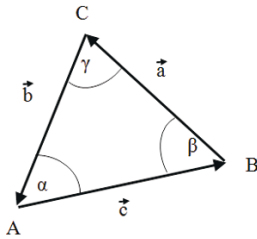


Abb. 3.20: Zur Dreiecksberechnung

$$\begin{aligned} (-\vec{a}) \cdot \vec{b} &= |\vec{a}||\vec{b}| \cos \gamma \implies 0 = \sqrt{2} \cos \gamma \implies \gamma = \frac{\pi}{2} \\ (-\vec{b}) \cdot \vec{c} &= |\vec{b}||\vec{c}| \cos \alpha \implies 2 = \sqrt{2}\sqrt{3} \cos \alpha \implies \alpha = 0,6155... \\ (-\vec{c}) \cdot \vec{a} &= |\vec{c}||\vec{a}| \cos \beta \implies 1 = \sqrt{3} \cos \beta \implies \beta = 0,9553... \end{aligned}$$

Zur Kontrolle rechnen wir $\alpha + \beta + \gamma = \pi$.

Die Fläche des Dreiecks mit den Eckpunkten A, B, C ist gleich der halben Fläche des Parallelogramms, das von $\vec{a}, \vec{b}$ aufgespannt wird. Wir überprüfen dies gleich dreifach:

$$\begin{aligned} \vec{a} \times \vec{b} &= (0, 1, 1)^\top, \quad |\vec{a} \times \vec{b}| = \sqrt{2}, \\ \vec{c} \times \vec{b} &= (0, -1, -1)^\top, \quad |\vec{c} \times \vec{b}| = \sqrt{2}, \\ \vec{a} \times \vec{c} &= (0, -1, -1)^\top, \quad |\vec{a} \times \vec{c}| = \sqrt{2}. \end{aligned}$$

Die Fläche des Dreiecks beträgt also $\frac{\sqrt{2}}{2}$ Flächeneinheiten.

Beispiel 3.11 (Tetraedervolumen)

Es seien $\vec{a}, \vec{b}, \vec{c} \in \mathbb{R}^3$ drei Vektoren, die nicht in einer Ebene liegen. Dann spannen $\vec{a}, \vec{b}, \vec{c}$ ein **Tetraeder** auf (siehe Abbildung 3.21). Die Grundfläche G ist gleich der halben Fläche des Parallelogramms, das von $\vec{a}$ und $\vec{b}$ aufgespannt wird. Die Höhe h entspricht der Höhe des Spats über der Ebene von $\vec{a}$ und $\vec{b}$. Das Volumen des Tetraeders berechnet sich mit der Formel

$$V = \frac{1}{3} G \cdot h.$$

Mit $G = \frac{1}{2} |\vec{a} \times \vec{b}|$ und h nach (3.8) erhalten wir

$$V = \frac{1}{3} \cdot \frac{1}{2} |\vec{a} \times \vec{b}| \cdot \left| \frac{\vec{c} \cdot (\vec{a} \times \vec{b})}{|\vec{a} \times \vec{b}|^2} (\vec{a} \times \vec{b}) \right| = \frac{1}{6} |(\vec{a} \times \vec{b}) \cdot \vec{c}| = \frac{1}{6} |[\vec{a}, \vec{b}, \vec{c}]|.$$

Für $\vec{a} = (-1, 0, 0)^\top$, $\vec{b} = (0, -1, 1)^\top$ und $\vec{c} = (1, 2, 3)^\top$ erhalten wir das Volumen des Tetraeders

$$V = \frac{1}{6} |[\vec{a}, \vec{b}, \vec{c}]| = \frac{1}{6} \left| \begin{vmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 1 & 2 & 3 \end{vmatrix} \right| = \frac{5}{6}.$$

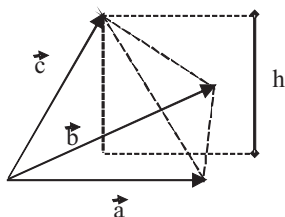


Abb. 3.21: Volumen eines Tetraeders

3.2 Analytische Geometrie

Ein dreidimensionales CAD-Modell lässt sich auf einem Bildschirm, durch den ein Beobachter das Objekt betrachtet, perspektivisch darstellen. Wo muss man in der Bildelebene einen Punkt des Objektes einzeichnen? Liegen im dreidimensionalen Raum das Auge des Betrachters und das Objekt auf unterschiedlichen Seiten der Bildelebene, dann kann man durch jeden für den Betrachter sichtbaren Punkt des Objekts und durch das Auge jeweils eine Gerade legen, die die Bildelebene schneidet. Der Schnittpunkt ist die korrekte perspektivische Darstellung des Punktes auf dem Bildschirm.

Die **Analytische Geometrie** ist ein Teilgebiet der Geometrie, in dem solche geometrischen Zusammenhänge mittels Vektoren ausgedrückt werden. Einige Ergebnisse haben wir zuvor schon betrachtet. Nun sehen wir uns systematisch Geraden und Ebenen an.

3.2.1 Geraden in der Ebene und im Raum

3.2.1.1 Darstellungsformen von Geraden in $\mathbb{R}^2$ und $\mathbb{R}^3$

Aus der Schule kennen Sie die Gerade in der Ebene $\mathbb{R}^2$ als Funktion $y = f(x) = mx + b$ mit Steigung m und $f(0) = b$. In Kapitel 1.5.6.1 wurde dieser Funktionstyp bereits diskutiert. Für die x -Achse als Gerade gilt beispielsweise $y = f(x) = 0$. Die y -Achse ist auch eine Gerade in der Ebene, lässt sich aber nicht in dieser Form beschreiben. Man benötigt daher andere Darstellungen für Geraden.

Lemma 3.1 (Koordinatengleichung der Gerade in $\mathbb{R}^2$)

Die Elemente $\vec{x} = (x_1, x_2)^\top$ der Lösungsmenge der linearen Gleichung

$$a_1x_1 + a_2x_2 = c \quad (3.9)$$

mit $(a_1, a_2) \neq (0, 0)$ und $c \in \mathbb{R}$ bilden eine Gerade g in der Ebene $\mathbb{R}^2$. Man nennt (3.9) die **Koordinatengleichung** oder kurz die **Gleichung** von g .

Bei dieser Gleichung mit zwei Variablen x_1 und x_2 kann eine der beiden Variablen frei gewählt werden. Die andere ist dann wegen $(a_1, a_2) \neq (0, 0)$ eindeutig bestimmt. Jeder Punkt $\vec{x} = (x_1, x_2)^\top \in g$ erfüllt die Gleichung und umgekehrt entspricht jede Lösung von (3.9) einem Punkt $\vec{x} = (x_1, x_2)^\top \in g$. Die y -Achse wird nun durch die Gleichung $x_1 = 0$ beschrieben.

Ist eine Gerade über eine Koordinatengleichung gegeben, so hat man keine unmittelbare Vorstellung von ihrem Verlauf. Das ist bei der Punkt-Richtungsform anders. Eine Gerade g ist eindeutig festgelegt durch einen beliebigen Punkt $\vec{x}_0$ der Gerade (wir identifizieren hier Punkte mit ihren Ortsvektoren, d. h., wir schreiben sie als Spaltenvektoren) und einen **Richtungsvektor** $\vec{v} \neq \vec{0}$. Man erreicht jeden Punkt $\vec{x}$ von g , indem man zu $\vec{x}_0$ ein Vielfaches des Richtungsvektors $\lambda \vec{v}$ addiert (vgl. Abbildung 3.22).

Lemma 3.2 (Punkt-Richtungsform der Gerade)

Es seien $\vec{x}_0$ und $\vec{v}$ Vektoren in der Ebene $\mathbb{R}^2$ oder im Raum $\mathbb{R}^3$. Dann sind die Vektoren

$$\vec{x} = \vec{x}_0 + \lambda \vec{v}, \quad \lambda \in \mathbb{R} \quad (3.10)$$

Ortsvektoren der Punkte einer Geraden g in der Ebene oder im Raum. Der Punkt $\vec{x}_0$ heißt **Aufpunkt** und der Vektor $\vec{v} \neq \vec{0}$ heißt **Richtungsvektor** von g . Man nennt (3.10) die **Punkt-Richtungsform** oder **Parameterform** von g .

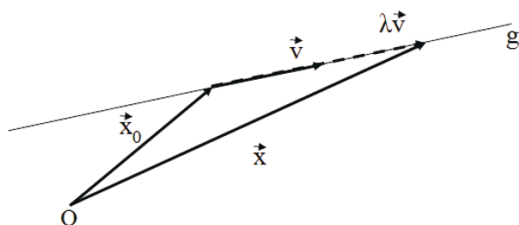


Abb. 3.22: Punkt-Richtungsform einer Geraden

In der Ebene betrachten wir den Zusammenhang zwischen (3.9) und (3.10) etwas genauer: Ein Vektor $\vec{n} \in \mathbb{R}^2$ mit $\vec{n} \cdot \vec{v} = 0$ (d. h. $\vec{n} \perp \vec{v}$) wird als **Normalenvektor** von g bzw. von $\vec{v}$ bezeichnet. Zu $\vec{v} = (v_1, v_2)^\top$ stehen beispielsweise $\vec{n} = (-v_2, v_1)^\top$ oder $\vec{n} = (v_2, -v_1)^\top$ senkrecht. Bilden wir mit $\vec{x} = \vec{x}_0 + \lambda \vec{v}$ und dem Normalenvektor $\vec{n} = (n_1, n_1)^\top$ das Skalarprodukt, so folgt

$$\vec{n} \cdot \vec{x} = \vec{n} \cdot \vec{x}_0 + \lambda \underbrace{\vec{n} \cdot \vec{v}}_{=0} \iff n_1 x_1 + n_2 x_2 = \vec{n} \cdot \vec{x}_0,$$

und wir erhalten eine Koordinatengleichung von g . Da jede Gerade durch zwei verschiedene Punkte festgelegt ist und wir damit eine Punkt-Richtungsform angeben können, haben wir jetzt gezeigt, dass es zu jeder Gerade in $\mathbb{R}^2$ auch eine Koordinatengleichung gibt.

Geht man umgekehrt von der Koordinatengleichung (3.9) für g aus, so erhält man (wie im folgenden Beispiel) die Punkt-Richtungsform, indem man in der unterbestimmten linearen Gleichung (3.9) eine Variable (d. h. eine Koordinate x_1 oder x_2) frei wählt und die jeweils andere damit ausdrückt.

Beispiel 3.12

- a) Die Gerade $g \subset \mathbb{R}^2$ sei gegeben durch die Gleichung $2x_1 + 3x_2 = 1$. Wir bestimmen eine Punkt-Richtungsform von g , indem wir in der Gleichung $x_2 := t$ setzen und nach x_1 auflösen: $x_1 = \frac{1}{2}(1 - 3t)$. In Vektorschreibweise folgt

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} + t \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix}, \quad t \in \mathbb{R}.$$

- b) Die Gerade $g \subset \mathbb{R}^2$ sei in Punkt-Richtungsform gegeben durch

$$g: \quad \vec{x} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad \lambda \in \mathbb{R}.$$

Wir berechnen eine Koordinatengleichung von g . Wegen $\vec{v} = (2, -1)^\top$ ist $\vec{n} = (1, 2)^\top$ ein Normalenvektor von g . Bildung des Skalarprodukts mit $\vec{n}$ liefert

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \underbrace{\lambda \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \end{pmatrix}}_{=0} \iff x_1 + 2x_2 = 5. \quad \blacksquare$$

Eine Gerade g in $\mathbb{R}^3$ kann in Punkt-Richtungsform (3.10) mit dem Aufpunkt $\vec{x}_0 \in \mathbb{R}^3$ und dem Richtungsvektor $\vec{v} \in \mathbb{R}^3$ dargestellt werden. Die Darstellung von g in Koordinatenform benötigt in Erweiterung zu (3.9) zwei lineare Gleichungen für $\vec{x} = (x_1, x_2, x_3)^\top \in g$:

$$\begin{aligned} a_1x_1 + a_2x_2 + a_3x_3 &= c_1, \\ b_1x_1 + b_2x_2 + b_3x_3 &= c_2 \end{aligned} \quad (3.11)$$

mit $c_1, c_2 \in \mathbb{R}$, wobei die Vektoren (a_1, a_2, a_3) und (b_1, b_2, b_3) nicht kollinear sind. Um von der Koordinatenform (3.11) auf eine Punkt-Richtungsform umzurechnen, löst man das unterbestimmte lineare Gleichungssystem.

Umgekehrt erhält man aus der Punkt-Richtungsform die Koordinatenform, indem man ausnutzt, dass der Richtungsvektor $\vec{v}$ senkrecht zu $\vec{a} = (a_1, a_2, a_3)^\top$ und $\vec{b} = (b_1, b_2, b_3)^\top$ stehen muss. Das sieht man, indem man die Punkt-Richtungsform in (3.11) einsetzt, denn beispielsweise kann

$$\vec{a} \cdot (\vec{x}_0 + \lambda \vec{v}) = c_1 \iff \lambda \cdot \vec{a} \cdot \vec{v} = c_1 - \vec{a} \cdot \vec{x}_0$$

nur für alle $\lambda \in \mathbb{R}$ gelten, wenn $\vec{a} \cdot \vec{v} = 0$ ist.

Wählt man einen zu $\vec{v}$ orthogonalen Vektor $\vec{a} \neq (0, 0, 0)$, so erhält man die rechte Seite c_1 in (3.11) durch Einsetzen des Aufpunkts $\vec{x}_0$ in die linke Seite der Gleichung. Die zweite Gleichung erhält man entsprechend über einen weiteren orthogonalen Vektor $\vec{b}$, der nicht kollinear zu $\vec{a}$ ist.

Beispiel 3.13

a) Die Gerade $g \subset \mathbb{R}^3$ sei gegeben durch die beiden Gleichungen

$$\begin{aligned} 2x_1 + 2x_2 + x_3 &= 1 \\ x_1 - x_2 + x_3 &= 0. \end{aligned}$$

Wir bestimmen eine Punkt-Richtungsform von g , indem wir das lineare Gleichungssystem lösen:

$$\left[\begin{array}{ccc|c} 2 & 2 & 1 & 1 \\ 1 & -1 & 1 & 0 \end{array} \right] \implies \left[\begin{array}{ccc|c} 1 & -1 & 1 & 0 \\ 0 & 4 & -1 & 1 \end{array} \right] \implies \left[\begin{array}{ccc|c} 1 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 1 & -\frac{1}{4} & \frac{1}{4} \end{array} \right].$$

Setzen wir $x_3 = t$, so erhalten wir $x_2 = \frac{1}{4}(1+t)$ und $x_1 = \frac{1}{4} - \frac{3}{4}t$, also

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ 0 \end{pmatrix} + t \begin{pmatrix} -\frac{3}{4} \\ \frac{1}{4} \\ 1 \end{pmatrix}.$$

b) Die Gerade $g \subset \mathbb{R}^3$ sei in Punkt-Richtungsform gegeben durch

$$g: \quad \vec{x} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \quad \lambda \in \mathbb{R}.$$

Wir berechnen Koordinatengleichungen von g . Aus der Forderung $\vec{n} \cdot \vec{v} = 0$ bzw. $n_1 - n_2 + n_3 = 0$ bestimmen wir die beiden nicht kollinearen Normalenvektoren $\vec{n}_1 = (1, 1, 0)^\top$ und $\vec{n}_2 = (0, 1, 1)^\top$. Wir bilden das Skalarprodukt der Punkt-Richtungsform mit $\vec{n}_1$ und $\vec{n}_2$ und erhalten die Gleichungen $\vec{n}_1 \cdot \vec{x} = \vec{n}_1 \cdot \vec{x}_0$ und $\vec{n}_2 \cdot \vec{x} = \vec{n}_2 \cdot \vec{x}_0$. Im Beispiel: $x_1 + x_2 = 3$ und $x_2 + x_3 = 2$. ■

3.2.1.2 Typische Aufgabenstellungen für Geraden in $\mathbb{R}^2$ und $\mathbb{R}^3$

Wir beginnen mit der Konstruktion einer Geraden durch zwei vorgegebene Punkte $\vec{x}_0, \vec{x}_1$. Wählen wir $\vec{x}_0$ als Aufpunkt, so ergibt sich der Richtungsvektor als die Differenz der beiden Ortsvektoren $\vec{v} = \vec{x}_1 - \vec{x}_0$. Wir erhalten

$$\vec{x} = \vec{x}_0 + \lambda \cdot (\vec{x}_1 - \vec{x}_0).$$

Für $\lambda = 0$ folgt $\vec{x} = \vec{x}_0$, für $\lambda = 1$ folgt $\vec{x} = \vec{x}_1$. Um festzustellen, ob ein gegebener Punkt $\vec{x}_2$ auf g liegt, untersucht man, ob ein $\lambda \in \mathbb{R}$ existiert mit $\vec{x}_2 = \vec{x}_0 + \lambda \cdot \vec{v}$.

Beispiel 3.14

Wir legen die Gerade g durch die Punkte $\vec{x}_0 = (1, 2, 3)^\top$ und $\vec{x}_1 = (1, 0, 2)^\top$. Als Aufpunkt wählen wir $\vec{x}_0 = (1, 2, 3)^\top$ und der Richtungsvektor ist $\vec{v} = \vec{x}_1 - \vec{x}_0 = (0, -2, -1)^\top$. Die Punkt-Richtungsform lautet

$$\vec{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \lambda \cdot \begin{pmatrix} 0 \\ -2 \\ -1 \end{pmatrix}$$

bzw. in Koordinaten $x_1 = 1$, $x_2 = 2 - 2\lambda$, $x_3 = 3 - \lambda$. Der Punkt $\vec{x}_3 = (1, 6, 5)^\top$ liegt auf g , denn es gilt $\begin{pmatrix} 1 \\ 6 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ -2 \\ -1 \end{pmatrix}$. Der Punkt $\vec{x}_4 = (2, 0, 3)^\top$ liegt nicht auf g , denn $\begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \lambda \begin{pmatrix} 0 \\ -2 \\ -1 \end{pmatrix}$ besitzt keine Lösung $\lambda \in \mathbb{R}$. ■

Als zweite Anwendung betrachten wir die Lage zweier Geraden in $\mathbb{R}^3$ an drei Beispielen.

Beispiel 3.15

a) Die Geraden

$$g_1: \quad \vec{x} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad g_2: \quad \vec{x} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

haben keinen Schnittpunkt. Das lineare Gleichungssystem in λ, μ für $g_1 \cap g_2$:

$$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \iff \lambda \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} - \mu \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}$$

bzw.

$$\left[\begin{array}{cc|c} 2 & -1 & 0 \\ 1 & -1 & 2 \\ 0 & -1 & -1 \end{array} \right] \implies \left[\begin{array}{cc|c} 2 & -1 & 0 \\ 1 & -1 & 2 \\ 0 & 1 & 1 \end{array} \right] \implies \left[\begin{array}{cc|c} 1 & 0 & -2 \\ 1 & 0 & 3 \\ 0 & 1 & 1 \end{array} \right]$$

besitzt keine Lösung. Die erste und die zweite Gleichung widersprechen sich.

b) Die Geraden

$$g_1: \quad \vec{x} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad g_2: \quad \vec{x} = \begin{pmatrix} 7 \\ 2 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

haben genau einen Schnittpunkt. Das lineare Gleichungssystem in λ, μ für $g_1 \cap g_2$:

$$\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 7 \\ 2 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \iff \lambda \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} - \mu \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \\ -2 \end{pmatrix}$$

bzw.

$$\left[\begin{array}{cc|c} 2 & -1 & 6 \\ 1 & -1 & 2 \\ 0 & -1 & -2 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 2 & 0 & 8 \\ 1 & 0 & 4 \\ 0 & 1 & 2 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 1 & 0 & 4 \\ 1 & 0 & 4 \\ 0 & 1 & 2 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 1 & 0 & 4 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{array} \right]$$

besitzt die Lösung $\lambda = 4$ und $\mu = 2$. Der Schnittpunkt ist $\begin{pmatrix} 9 \\ 4 \\ 2 \end{pmatrix}$.

c) Die Geraden

$$g_1 : \vec{x} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \quad \text{und} \quad g_2 : \vec{x} = \begin{pmatrix} 7 \\ -3 \\ 5 \end{pmatrix} + \lambda \begin{pmatrix} 6 \\ -3 \\ 3 \end{pmatrix}$$

sind identisch. Das lineare Gleichungssystem in λ, μ für $g_1 \cap g_2$:

$$\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 7 \\ -3 \\ 5 \end{pmatrix} + \mu \begin{pmatrix} 6 \\ -3 \\ 3 \end{pmatrix}$$

$$\Leftrightarrow \lambda \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} - \mu \begin{pmatrix} 6 \\ -3 \\ 3 \end{pmatrix} = \begin{pmatrix} 6 \\ -3 \\ 3 \end{pmatrix}$$

bzw.

$$\left[\begin{array}{cc|c} 2 & -6 & 6 \\ -1 & 3 & -3 \\ 1 & -3 & 3 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 1 & -3 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right]$$

besitzt die Lösung $\lambda - 3\mu = 3$ bzw. $\lambda = 3 + 3\mu$, $\mu = \frac{\lambda}{3} - 1$. Jeder Punkt, der über die eine Punkt-Richtungsform erreicht werden kann, wird damit auch über die jeweils andere erreicht. Also sind g_1 und g_2 identisch. ■

Beispiel 3.15 zeigt exemplarisch, dass die Schnittmenge zweier Geraden genau der Lösungsmenge eines linearen Gleichungssystems der Koordinaten entspricht. Für dieses Gleichungssystem gibt es folgende Lösungsmöglichkeiten (vgl. Kapitel 1.7.3.2):

- Es existiert **keine** Lösung. Dann haben g_1 und g_2 keinen Schnittpunkt. Die beiden Geraden sind entweder parallel oder windschief (nur in $\mathbb{R}^3$). Im Falle der Parallelität von g_1 und g_2 ist $\vec{v}_1$ ein Vielfaches von $\vec{v}_2$, und es folgt $\vec{v}_1 \times \vec{v}_2 = \vec{0}$.
- Es existiert **genau eine** Lösung $\lambda, \mu \in \mathbb{R}$. Dann haben g_1 und g_2 genau einen Schnittpunkt.

■ Es existieren **unendlich viele** Lösungen $\lambda, \mu \in \mathbb{R}$. Dann sind g_1 und g_2 identisch.

Eine weitere wichtige Aufgabe ist die Berechnung des Abstands eines Punkts $\vec{x}_1$ zu der Geraden g . Wir betrachten einen Punkt $\vec{x}_1 \in \mathbb{R}^2$ und die Gerade $g: \vec{x} = \vec{x}_0 + \lambda \vec{v}$, $\lambda \in \mathbb{R}$, mit Aufpunkt $\vec{x}_0 \in \mathbb{R}^2$ und Richtung $\vec{v} \in \mathbb{R}^2$. Um den Abstand von $\vec{x}_1$ zu g zu berechnen, fällt man das Lot von $\vec{x}_1$ auf g . Der Lotfußpunkt $\vec{x}_L \in g$ hat die Darstellung $\vec{x}_L = \vec{x}_0 + \lambda_L \vec{v}$ mit $\lambda_L \in \mathbb{R}$. Für den Lotvektor hat man

$$\vec{l} = \vec{x}_L - \vec{x}_1$$

entsprechend der Abbildung 3.23. Die Forderung $\vec{l} \perp \vec{v}$ bzw. $\vec{l} \cdot \vec{v} = 0$ liefert durch Bildung des Skalarprodukts mit $\vec{v}$ die Bedingung

$$\underbrace{\vec{v} \cdot \vec{l}}_{=0} = \vec{v} \cdot (\vec{x}_0 + \lambda_L \vec{v} - \vec{x}_1) \iff 0 = \vec{v} \cdot (\vec{x}_0 - \vec{x}_1) + \lambda_L |\vec{v}|^2 \iff \lambda_L = \frac{\vec{v} \cdot (\vec{x}_1 - \vec{x}_0)}{|\vec{v}|^2}.$$

Mit λ_L erhält man $\vec{x}_L$ und daraus $\vec{l} = \vec{x}_L - \vec{x}_1$. Der Abstand von $\vec{x}_1$ zu g ist gerade die Länge des Lotvektors $\vec{l}$.

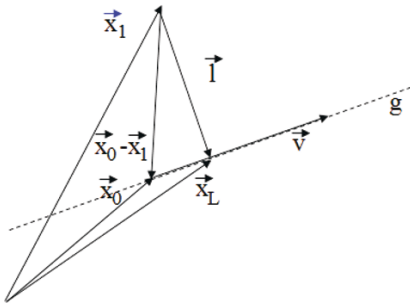


Abb. 3.23: Lot von einem Punkt auf eine Gerade

Beispiel 3.16

Wir berechnen den Abstand von $\vec{x}_1 = (1, 0, 1)^\top$ zur Geraden $g: \vec{x} = (1, 2, 3)^\top + \lambda(0, -2, 1)^\top$. Hier ist der Aufpunkt $\vec{x}_0 = (1, 2, 3)^\top$, der Richtungsvektor $\vec{v} = (0, -2, 1)^\top$ mit $|\vec{v}|^2 = 5$, und wir erhalten

$$\lambda_L = \frac{\vec{v} \cdot (\vec{x}_1 - \vec{x}_0)}{|\vec{v}|^2} = \frac{1}{5} \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -2 \\ -2 \end{pmatrix} = \frac{2}{5}$$

und weiter

$$\vec{x}_L = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \frac{2}{5} \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{6}{5} \\ \frac{17}{5} \end{pmatrix}, \quad \vec{l} = \begin{pmatrix} 1 \\ \frac{6}{5} \\ \frac{17}{5} \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{6}{5} \\ \frac{12}{5} \end{pmatrix},$$

sowie $|\vec{l}| = \sqrt{\frac{36}{25} + \frac{144}{25}} = \frac{\sqrt{180}}{5}$. Der Abstand von $\vec{x}_1$ und g beträgt $\frac{\sqrt{180}}{5}$ Längeneinheiten. ■

3.2.2 Ebenen im Raum

3.2.2.1 Darstellungsformen von Ebenen im Raum

Eine Ebene E in $\mathbb{R}^3$ ist eindeutig bestimmt durch einen Punkt $\vec{x}_0$ der Ebene und zwei nicht kollineare Richtungsvektoren $\vec{v}$ und $\vec{w}$ (siehe Definition 3.2 auf Seite 428), d. h., $\vec{v}$ lässt sich nicht als Vielfaches von $\vec{w}$ schreiben und umgekehrt. Analog zu Lemma 3.2 gilt:

Lemma 3.3 (Punkt-Richtungsform der Ebene)

Die Ortsvektoren der Ebene $E \subset \mathbb{R}^3$, die durch einen Punkt $\vec{x}_0$ geht und durch die nicht kollinearen **Richtungsvektoren** $\vec{v}$ und $\vec{w}$ aufgespannt wird (d. h. $\vec{v} \times \vec{w} \neq \vec{0}$), werden beschrieben durch die Gleichung

$$\vec{x} = \vec{x}_0 + \lambda \cdot \vec{v} + \mu \cdot \vec{w}, \quad \lambda, \mu \in \mathbb{R}. \quad (3.12)$$

Man nennt diese Darstellung **Parameterform** der Ebenengleichung oder **Punkt-Richtungsform** von E . Man bezeichnet $\vec{x}_0$ als **Aufpunkt** der Parameterform.

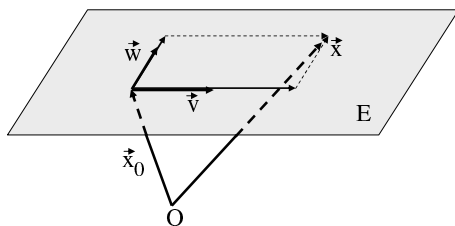


Abb. 3.24: Punkt-Richtungsform einer Ebene

Beispiel 3.17

- a) Gegeben sind die drei Punkte $\vec{x}_0 = (1, 1, 0)^\top$, $\vec{x}_1 = (2, 0, 2)^\top$ und $\vec{x}_2 = (4, 2, 1)^\top$. Die Ebene E durch $\{\vec{x}_0, \vec{x}_1, \vec{x}_2\}$ erhält man, indem ein Punkt als Aufpunkt gewählt wird. Die Richtungsvektoren erhält man als die Differenzen der beiden anderen Punkte zum Aufpunkt. Wählen wir $\vec{x}_0 = (1, 1, 0)^\top$ als Aufpunkt, so folgt mit $\vec{v}_1 = \vec{x}_1 - \vec{x}_0 = (1, -1, 2)^\top$ und $\vec{w}_1 = \vec{x}_2 - \vec{x}_0 = (3, 1, 1)^\top$ die Punkt-Richtungsform

$$E: \quad \vec{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} + \mu \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}, \quad \lambda, \mu \in \mathbb{R}.$$

Wählt man hingegen $\vec{x}_1 = (2, 0, 2)^\top$ als Aufpunkt, erhält man mit $\vec{v}_2 = \vec{x}_0 - \vec{x}_1$, $\vec{w}_2 = \vec{x}_2 - \vec{x}_1$ die Punkt-Richtungsform

$$E: \quad \vec{x} = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix} + \mu \begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix}, \quad \lambda, \mu \in \mathbb{R}.$$

Beide Punkt-Richtungsformen beschreiben dieselbe Ebene E .

- b) Der Punkt $\vec{x}_3 = (8, 2, 4)^\top$ liegt in E , denn das lineare Gleichungssystem $\vec{x}_3 = \vec{x}_0 + \lambda \vec{v}_1 + \mu \vec{w}_1$ bzw. $\lambda \vec{v}_1 + \mu \vec{w}_1 = \vec{x}_3 - \vec{x}_0$ ist lösbar:

$$\left[\begin{array}{cc|c} 1 & 3 & 7 \\ -1 & 1 & 1 \\ 2 & 1 & 4 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 0 & 4 & 8 \\ 1 & -1 & -1 \\ 0 & 3 & 6 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 1 & -1 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{array} \right].$$

Es gilt $\vec{x}_3 = \vec{x}_0 + \vec{v}_1 + 2\vec{w}_1$. ■

Alternativ zur Punkt-Richtungsform lässt sich eine Ebene durch eine **Koordinatengleichung** beschreiben.

Lemma 3.4 (Koordinatengleichung der Ebene in $\mathbb{R}^3$)

In $\mathbb{R}^3$ bilden die Lösungsvektoren $\vec{x} = (x_1, x_2, x_3)^\top$ der linearen Gleichung

$$a_1x_1 + a_2x_2 + a_3x_3 = c \tag{3.13}$$

mit $(a_1, a_2, a_3) \neq (0, 0, 0)$ und $c \in \mathbb{R}$ die Ortsvektoren einer Ebene E . Man nennt (3.13) die **Koordinatengleichung** oder kurz die **Gleichung** von E .

Da hier eine Gleichung mit drei Unbekannten verwendet wird, können zwei der Variablen x_1, x_2, x_3 beliebig gewählt werden und bestimmen den Wert einer dritten Variable. Wir erhalten also zwei Freiheitsgrade, die eine Ebene ausmachen.

Der Zusammenhang zwischen der Punkt-Richtungsform (3.12) und der Koordinatengleichung (3.13) lässt sich analog zu den Betrachtungen für Geraden mit einem Normalenvektor $\vec{n}$ von E darstellen. Ausgehend von der Punkt-Richtungsform $\vec{x} = \vec{x}_0 + \lambda \vec{v} + \mu \vec{w}$ erhalten wir zunächst mit

$$\vec{n} := \vec{v} \times \vec{w}$$

einen Normalenvektor von E , d. h. einen Vektor, der senkrecht auf beiden Richtungsvektoren steht. Die Bildung des Skalarprodukts von $\vec{n} = (n_1, n_2, n_3)^\top$ mit einem Vektor $\vec{x} = (x_1, x_2, x_3)^\top$ der Ebene E ergibt

$$\vec{n} \cdot \vec{x} = \vec{n} \cdot \vec{x}_0 + \underbrace{\lambda \vec{n} \cdot \vec{v}}_{=0} + \underbrace{\mu \vec{n} \cdot \vec{w}}_{=0} \iff \vec{n} \cdot \vec{x} = \vec{n} \cdot \vec{x}_0,$$

oder ausgeschrieben

$$n_1x_1 + n_2x_2 + n_3x_3 = c, \quad c = \vec{n} \cdot \vec{x}_0.$$

Da man für jede Ebene aus drei Punkten eine Punkt-Richtungsform konstruieren kann, hat nun insbesondere jede Ebene auch eine Darstellung als Koordinatengleichung.

Lemma 3.5 (Koordinatenform der Ebene in $\mathbb{R}^3$)

Jeder Punkt bzw. Ortsvektor $\vec{x}$ der Ebene $E \subset \mathbb{R}^3$, die durch den Punkt $\vec{x}_0$ geht und den Normalenvektor $\vec{n}$ besitzt, erfüllt die Gleichung

$$\vec{n} \cdot \vec{x} = c \quad \text{mit } c := \vec{n} \cdot \vec{x}_0 \quad \text{bzw.} \quad \vec{n} \cdot (\vec{x} - \vec{x}_0) = 0. \quad (3.14)$$

Umgekehrt ist jede Lösung der Gleichung ein Punkt der Ebene. Man nennt diese Darstellung **Koordinatenform** der Ebenengleichung.

Ist E in Koordinatenform gegeben, so erhält man eine Darstellung in Punkt-Richtungsform über drei Vektoren als Lösungen der Koordinatengleichung, die nicht auf einer Geraden liegen. Alternativ kann man zum Normalenvektor zwei orthogonale Vektoren $\vec{v}, \vec{w}$ bestimmen. Zusammen mit einem Punkt $\vec{x}_0$ der Ebene als Lösung der Koordinatengleichung hat man dann die Daten für die Punkt-Richtungsform.

Beispiel 3.18

a) Für die Ebene

$$E: \quad \vec{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} + \mu \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$$

erhalten wir den Normalenvektor

$$\vec{n} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} \times \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -4 \\ 5 \\ 7 \end{pmatrix}.$$

Bildung des Skalarprodukts von $\vec{n}$ mit der Punkt-Richtungsform von E ergibt die Koordinatengleichung

$$\begin{pmatrix} -4 \\ 5 \\ 7 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -4 \\ 5 \\ 7 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \iff -4x_1 + 5x_2 + 7x_3 = 1.$$

b) Ist E durch eine Koordinatengleichung gegeben, z. B.

$$E: \quad -2x_1 + 3x_2 + 4x_3 = 1,$$

so kann man daraus den Normalenvektor $\vec{n} = (-2, 3, 4)^\top$ ablesen. Durch „zielgerichtetes Raten“ findet man leicht die beiden nicht kollinearen und zu $\vec{n}$ senkrechten Vektoren $\vec{v} = (2, 0, 1)^\top$ und $\vec{w} = (3, 2, 0)^\top$. Jetzt benötigen wir noch einen Punkt $\vec{x}_0 \in E$, den wir ebenfalls leicht aus der Gleichung ablesen: Wegen $-2 \cdot 1 + 3 \cdot 1 + 4 \cdot 0 = 1$ liegt $\vec{x}_0 = (1, 1, 0)^\top$ in der Ebene. Insgesamt ergibt sich:

$$E: \quad \vec{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + \mu \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix}, \quad \lambda, \mu \in \mathbb{R}. \quad \blacksquare$$

Wählt man als Normalenvektor $\vec{n}_0$ einer Ebene E speziell einen Vektor der Länge 1, etwa indem man einen beliebigen Normalenvektor $\vec{n}$ zu $\vec{n}_0 := \frac{\vec{n}}{|\vec{n}|}$ normiert, so ist die Konstante c in der Ebenengleichung $\vec{n}_0 \cdot \vec{x} = c$ der **Abstand von E zum Nullpunkt**. Man sieht dies wie folgt: Der Ortsvektor des Punkts $\vec{d} \in E$ mit kürzestem Abstand zu 0 ist parallel zu $\vec{n}_0$, d. h. $\vec{d} = d\vec{n}_0$ mit $d \in \mathbb{R}$. Setzen wir $\vec{x} = \vec{d}$ in die Ebenengleichung ein, so ergibt sich

$$\vec{n}_0 \cdot \vec{d} = \vec{n}_0 \cdot \vec{x}_0 \iff d\vec{n}_0 \cdot \vec{n}_0 = \vec{n}_0 \cdot \vec{x}_0 \iff d = \vec{n}_0 \cdot \vec{x}_0.$$

Wir wählen nun die Richtung $\vec{n}_0$ so, dass $d = \vec{n}_0 \cdot \vec{x}_0 \geq 0$ ist und erhalten:

Satz 3.9 (Hesse'sche Normalform)

Jede Ebene $E \subset \mathbb{R}^3$ besitzt eine Darstellung über die **Hesse'sche Normalform**

$$\vec{n}_0 \cdot \vec{x} = d \text{ mit } d = \vec{n}_0 \cdot \vec{x}_0.$$

Ein Punkt $\vec{x}$ liegt genau dann in E , wenn $\vec{x}$ die Gleichung erfüllt. Dabei ist $\vec{x}_0$ ein beliebiger Punkt der Ebene, und der **Normaleneinheitsvektor** $\vec{n}_0$ von E ist ein Normalenvektor der Länge 1, für den $d = \vec{n}_0 \cdot \vec{x}_0 \geq 0$ ist. Die Zahl $d = \vec{n}_0 \cdot \vec{x}_0$ ist der Abstand der Ebene E zum Nullpunkt.

Beispiel 3.19

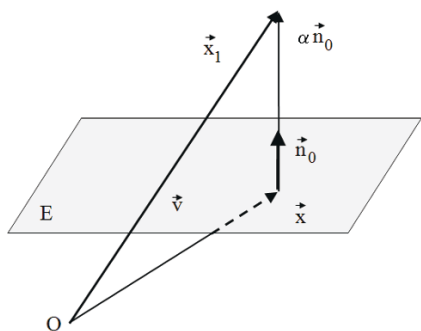
Für die Ebene

$$E: \quad 2x_1 + x_2 - x_3 = 4$$

erhalten wir einen Normalenvektor $\vec{n} = (2, 1, -1)^\top$. Mit $|\vec{n}| = \sqrt{6}$ erhalten wir die Hesse'sche Normalform

$$E: \quad \frac{2}{\sqrt{6}}x_1 + \frac{1}{\sqrt{6}}x_2 - \frac{1}{\sqrt{6}}x_3 = \frac{4}{\sqrt{6}}.$$

Der Abstand von E zum Ursprung beträgt $d = \frac{4}{\sqrt{6}}$. ■

Abb. 3.25: Abstand eines Punkts zur Ebene E

Wir können jetzt den Abstand einer Ebene zum Nullpunkt ablesen. Zur Berechnung des Abstands α eines beliebigen Punkts $\vec{x}_1$ zu der Ebene E in der Hesse'schen Normalform $\vec{n}_0 \cdot \vec{x} = d$ verwenden wir den Ansatz $\vec{x}_1 = \vec{x} + \alpha \vec{n}_0$ mit einem Vektor $\vec{x} \in E$ (siehe Abbildung 3.25). Wir bilden das Skalarprodukt

$$\vec{n}_0 \cdot \vec{x}_1 = \vec{n}_0 \cdot \vec{x} + \underbrace{\alpha \vec{n}_0 \cdot \vec{n}_0}_{=1} \iff \vec{n}_0 \cdot \vec{x}_1 = d + \alpha$$

und erhalten $\alpha = \vec{n}_0 \cdot \vec{x}_1 - d$. Der Abstand von $\vec{x}_1$ zu E beträgt $|\alpha| = |\vec{n}_0 \cdot \vec{x}_1 - d|$.

Beispiel 3.20

Wir berechnen den Abstand des Punkts $\vec{x}_1 = (2, 2, 3)^\top$ zur Ebene in Punkt-Richtungsform

$$E: \vec{x} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

Zunächst gewinnen wir die Hesse'sche Normalform von E . Ein Normalenvektor $\vec{n}$ lautet

$$\vec{n} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \quad \text{bzw.} \quad \vec{n}_0 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Die Hesse'sche Normalform von E ergibt sich zu

$$E: \vec{n}_0 \cdot \vec{x} = \vec{n}_0 \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \iff \frac{1}{\sqrt{3}} (x_1 - x_2 + x_3) = \frac{2}{\sqrt{3}}.$$

Mit dem Ansatz $\vec{x}_1 = \vec{x} + \alpha \vec{n}_0$ bestimmen wir α so, dass $\vec{x} \in E$ gilt; $|\alpha|$ ist dann der Abstand von $\vec{x}_1$ zu E . Also:

$$\vec{n}_0 \cdot \vec{x}_1 = \vec{n}_0 \cdot \vec{x} + \alpha \iff \alpha = \vec{n}_0 \cdot \vec{x}_1 - \frac{2}{\sqrt{3}}$$

bzw.

$$\alpha = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix} - \frac{2}{\sqrt{3}} = \frac{3}{\sqrt{3}} - \frac{2}{\sqrt{3}} = \frac{1}{\sqrt{3}}.$$

Der Abstand von $\vec{x}_1$ zu E beträgt $\frac{1}{\sqrt{3}}$. ■

3.2.2.2 Typische Aufgabenstellungen für Ebenen in $\mathbb{R}^3$

Wir betrachten abschließend noch zwei Problemstellungen der Analytischen Geometrie, wie sie in den Anwendungen (und in Klausuren) häufig vorkommen.

Beim Schnitt von zwei Ebenen E_1 und E_2 in $\mathbb{R}^3$ können folgende Situationen auftreten:

- $E_1 \cap E_2 = \emptyset$, d. h., E_1 und E_2 sind parallel,
- $E_1 \cap E_2 = g$, d. h., E_1 und E_2 schneiden sich in einer Geraden g ,
- $E_1 = E_2$, d. h., E_1 und E_2 sind gleich.

Liegen die Ebenen E_1 und E_2 als Gleichung (Koordinatenform oder Hesse'sche Normalform) vor, so löst man ein lineares Gleichungssystem mit zwei Gleichungen in drei Variablen. Ansonsten bietet es sich an, die Ebenen in Koordinatenform umzurechnen.

Beispiel 3.21

Wir bilden die Schnittmenge der Ebenen

$$E_1: 2x_1 + x_2 - x_3 = 1 \quad \text{und} \quad E_2: x_1 + x_2 + x_3 = 3.$$

Die Normalenvektoren lesen wir direkt aus den Gleichungen ab: $\vec{n}_1 = (2, 1, -1)^\top$ und $\vec{n}_2 = (1, 1, 1)^\top$. Mit $\vec{n}_1$ und $\vec{n}_2$ kann der Schnittwinkel von E_1 und E_2 als Winkel zwischen den Normalenvektoren berechnet werden:

$$\cos \varphi = \frac{\vec{n}_1 \cdot \vec{n}_2}{|\vec{n}_1| |\vec{n}_2|} = \frac{2}{\sqrt{6}\sqrt{3}} \implies \varphi = 1,08\dots \approx 61,87^\circ.$$

Damit wissen wir, dass die Ebenen nicht parallel verlaufen oder identisch sind. Sie schneiden sich also in einer Geraden, die wir über ein Gleichungssystem berechnen:

$$\left[\begin{array}{ccc|c} 2 & 1 & -1 & 1 \\ 1 & 1 & 1 & 3 \end{array} \right] \implies \left[\begin{array}{ccc|c} 1 & 0 & -2 & -2 \\ 1 & 1 & 1 & 3 \end{array} \right] \implies \left[\begin{array}{ccc|c} 1 & 0 & -2 & -2 \\ 0 & 1 & 3 & 5 \end{array} \right].$$

Wählt man $x_3 = t$, so folgt $x_2 = 5 - 3t$ und $x_1 = -2 + 2t$ bzw. die Punkt-Richtungsform der Schnittgeraden

$$g: \vec{x} = \begin{pmatrix} -2 \\ 5 \\ 0 \end{pmatrix} + t \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}, \quad t \in \mathbb{R}. \quad \text{■}$$

Die Schnittmenge einer Geraden g mit einer Ebene E in $\mathbb{R}^3$ erhält man über ein lineares Gleichungssystem mit drei Gleichungen (zwei für die Gerade und eine Gleichung für die Ebene) und drei Unbekannten. Entsprechend dem Verhalten von linearen Gleichungssystemen (vgl. Kapitel 1.7.3.2) können hier die folgenden Fälle auftreten:

- $E \cap g = \emptyset$, d. h., g verläuft parallel zu E ,
- $E \cap g = \vec{x}_S$, d. h., E und g schneiden sich in genau einem Punkt,
- $E \cap g = g$, d. h., g verläuft in E , $g \subset E$.

Beispiel 3.22

Wir berechnen die Schnittmenge der Geraden g mit der Ebene E in Punkt-Richtungsform

$$g: \vec{x} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \text{und} \quad E: \vec{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \lambda \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

Aus $\vec{x} \in g$ und $\vec{x} \in E$ erhalten wir das lineare Gleichungssystem

$$\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \lambda \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

Wir rechnen in erweiterter Matrixschreibweise weiter:

$$\left[\begin{array}{ccc|c} 1 & 0 & -2 & 0 \\ 0 & -1 & -1 & 0 \\ 1 & 0 & -1 & 3 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 3 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & 6 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & 3 \end{array} \right].$$

Also folgt $t = 6$, $\lambda = -3$ und $\mu = 3$. Der Schnittpunkt lautet $\begin{pmatrix} 7 \\ 2 \\ 6 \end{pmatrix}$. ■

3.3 Vektorräume

Nachdem wir uns mit Vektoren im Anschauungsraum beschäftigt haben, übertragen wir nun die dabei entdeckten Strukturen in allgemeinere Situationen. Wir haben eine Menge von Objekten wie z. B. Pfeile vorgefunden, für die eine (Vektor-)Addition erklärt ist. Außerdem konnten wir die Objekte mit Zahlen (Skalaren) multiplizieren. Für diese Skalarmultiplikation und die Vektoraddition gelten gewisse Rechenregeln. Damit ist für die Menge der Objekte eine Struktur vorgegeben. In der Mathematik nennt man eine Menge mit einer Struktur einen **Raum**. Bei der von uns betrachteten Struktur spricht man von einem **Vektorraum**.

Sie fragen sich zu Recht, warum es nötig ist, die vorangehenden Überlegungen zu verallgemeinern. Wir werden schnell sehen, dass man in vielen Zusammenhängen auf Vektorräume trifft – und dann hilft die allgemeine Theorie, so dass man die benötigten Aussagen nicht jeweils neu herleiten muss. Konkret bildet zum Beispiel die Lösungsmenge eines homogenen linearen Gleichungssystems einen Vektorraum. Die Eigenschaften, die wir für allgemeine Vektorräume finden, helfen uns, bei einer technischen Aufgabenstellung die Gleichungen richtig aufzustellen, damit wir über das Gleichungssystem eine eindeutige Lösung erhalten. Wir vertiefen diesen Aspekt in Kapitel 3.5 am Beispiel der Berechnung von Strömen und Spannungen in elektrischen Netzwerken. Auch bei Differenzialgleichungen werden wir auf Eigenschaften von Vektorräumen zurückgreifen. Die Vektoren werden dann Funktionen sein. Nahezu alle mathematischen Objekte lassen sich im Kontext von Vektorräumen verstehen, so bilden sowohl die Polynome als auch die $(m \times n)$ -Matrizen Vektorräume.

3.3.1 Definition des Vektorraums

Wir beginnen mit der abstrakten Definition des Vektorraums, indem wir die Vektoren des Anschauungsraums durch Elemente einer beliebigen Menge ersetzen. Die als Axiome formulierten Rechenregeln können wir aus dem Anschauungsraum übernehmen.

3.3.1.1 Vektorraumaxiome

Vielleicht betrachten Sie beim Lesen der folgenden Definition gleichzeitig das sich anschließende Beispiel 3.23.

Definition 3.7 (Vektorraum)

Eine Menge V bildet einen Vektorraum $(V, +; \mathbb{K}, \cdot)$ über $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$, falls in V eine Vektoraddition „+“ und eine Multiplikation „ $\cdot$ “ mit Skalaren aus $\mathbb{K}$ erklärt ist, so dass die folgenden Axiome gelten:

a) Regeln zur Vektoraddition: Seien $\vec{a}, \vec{b}, \vec{c} \in V$.

- i) Zu je zwei Elementen $\vec{a}, \vec{b} \in V$ ist eine eindeutige Summe $\vec{c} = \vec{a} + \vec{b} \in V$ erklärt.
- ii) Diese Addition ist **kommutativ**: $\vec{a} + \vec{b} = \vec{b} + \vec{a}$.
- iii) Die Addition ist **assoziativ**: $\vec{a} + (\vec{b} + \vec{c}) = (\vec{a} + \vec{b}) + \vec{c}$.
- iv) Es gibt ein **neutrales Element** $\vec{0} \in V$, das für jedes $\vec{a} \in V$ das eindeutige Element ist, für das $\vec{a} + \vec{0} = \vec{a}$ gilt.
- v) Zu jedem $\vec{a} \in V$ existiert bezüglich der Addition genau ein **inverses Element** $-\vec{a} \in V$, so dass $\vec{a} + (-\vec{a}) = \vec{0}$. Wir benutzen statt $\vec{b} + (-\vec{a})$ die Kurzschreibweise $\vec{b} - \vec{a}$.

b) Regeln zur Multiplikation mit einem Skalar: Seien $\lambda, \mu \in \mathbb{K}$ und $\vec{a}, \vec{b} \in V$.

- i) Zu jedem $\vec{a} \in V$ und jedem $\lambda \in \mathbb{K}$ ist ein Produkt $\vec{b} = \lambda \cdot \vec{a} = \vec{a} \cdot \lambda \in V$ eindeutig erklärt.
- ii) Das Produkt mit einem Skalar ist **assoziativ**: $(\lambda\mu) \cdot \vec{a} = \lambda \cdot (\mu \cdot \vec{a})$.
- iii) $1 \cdot \vec{a} = \vec{a}$, wobei $1 \in \mathbb{K}$.
- iv) Addition in $\mathbb{K}$, Vektoraddition und Multiplikation mit einem Skalar erfüllen die **Distributivgesetze**

$$\lambda \cdot (\vec{a} + \vec{b}) = \lambda \cdot \vec{a} + \lambda \cdot \vec{b}, \quad (\lambda + \mu) \cdot \vec{a} = \lambda \cdot \vec{a} + \mu \cdot \vec{a}.$$

Die Elemente von V heißen **Vektoren**.

Statt über $\mathbb{R}$ oder $\mathbb{C}$ kann man Vektorräume über beliebige Körper $\mathbb{K}$ definieren, siehe Definition 1.18 auf Seite 45.

Achtung:

- $-\vec{a}$ ist (zunächst) nicht gleichbedeutend mit $(-1) \cdot \vec{a}$ für $-1 \in \mathbb{K}$. $-\vec{a}$ ist (zunächst) nur ein Symbol für das inverse Element. Wir werden im Anschluss zeigen, dass aufgrund der Axiome $-\vec{a}$ berechnet werden kann über $(-1) \cdot \vec{a}$. Aber das ist eine Aussage, die bewiesen werden muss.
- Die Multiplikation mit einem Skalar wird als **skalare Multiplikation** oder **Skalarmultiplikation** bezeichnet. Das Ergebnis ist ein Vektor. In Kapitel 3.1.2 haben wir das **Skalarprodukt** eingeführt. Mit dem Skalarprodukt werden zwei Vektoren auf einen Skalar abgebildet, über den man den Winkel zwischen den Vektoren berechnen kann. Obwohl beide Multiplikationen ähnlich bezeichnet werden, handelt es sich um völlig unterschiedliche Begriffe!
- Da $\vec{0} \in V$ gefordert ist, kann die Menge V nicht leer sein. Der einfachste Vektorraum besteht nur aus diesem **Nullvektor** und heißt **Nullvektorraum**.

Wenn aus dem Zusammenhang klar ist, welcher Körper verwendet wird, schreiben wir statt $(V, +; \mathbb{K}, \cdot)$ auch kurz nur V , wobei wir die Doppelbezeichnung von V als Menge und als Vektorraum in Kauf nehmen.

Beispiel 3.23

Wir haben bereits mehrere Vektorräume kennengelernt:

- a) $(\mathbb{R}, +; \mathbb{R}, \cdot)$ ist ein Vektorraum. Dabei ist die Vektoraddition die „normale“ Addition und die Multiplikation mit einem Skalar die „normale“ Multiplikation in $\mathbb{R}$. Analog ist $(\mathbb{C}, +; \mathbb{C}, \cdot)$ ein Vektorraum.
- b) Da die Multiplikation einer reellen Zahl mit einer komplexen Zahl eine komplexe Zahl ist, ist auch $(\mathbb{C}, +; \mathbb{R}, \cdot)$ ein Vektorraum.
- c) $(\mathbb{R}, +; \mathbb{C}, \cdot)$ ist jedoch kein Vektorraum, da die Multiplikation mit dem Skalar j aus der Menge $V = \mathbb{R}$ hinausführt.

- d) In Kapitel 3.1 haben wir die Vektorräume $(\mathbb{R}^2, +; \mathbb{R}, \cdot)$ und $(\mathbb{R}^3, +; \mathbb{R}, \cdot)$ als Menge von Pfeilen mit einer geometrisch erklärten Addition und skalaren Multiplikation erklärt. Die Axiome des Vektorraums können dafür geometrisch bewiesen werden, siehe Satz 3.1 und die Abbildungen 3.2 und 3.3.

Man beachte, dass auch komplexe Zahlen $x + jy$ als Punkte (x, y) und damit als Ortsvektoren dargestellt werden können (siehe Kapitel 1.6.3). Dabei haben wir auch die komplexe Addition bereits mittels Pfeilen visualisiert, so dass der Vektorraum der Pfeile der Ebene gleichbedeutend ist mit dem Vektorraum $(\mathbb{C}, +; \mathbb{R}, \cdot)$.

- e) Im vorangehenden Punkt haben wir die Vektoren der Ebene und des Raums als Spaltenmatrizen ihrer Koordinaten dargestellt. Nun sind $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$ mit

$$\mathbb{R}^n = \left\{ \vec{a} = (a_1, \dots, a_n)^\top : a_1, \dots, a_n \in \mathbb{R} \right\}$$

und $(\mathbb{C}^n, +; \mathbb{C}, \cdot)$ mit der koordinatenweisen Addition und koordinatenweisen skalaren Multiplikation

$$\vec{a} + \vec{b} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ \vdots \\ a_n + b_n \end{pmatrix}, \quad \lambda \cdot \vec{a} = \begin{pmatrix} \lambda \cdot a_1 \\ \vdots \\ \lambda \cdot a_n \end{pmatrix}, \quad \lambda \in \mathbb{R} \text{ bzw. } \mathbb{C},$$

nicht nur für $n = 2$ oder $n = 3$ Standardbeispiele für Vektorräume. Wir werden diese Räume allgemein für jedes $n \in \mathbb{N}$ verwenden.

Wir fassen die Vektoren dieser Räume (also die n -Tupel) alternativ als Zeilen- oder Spaltenvektoren auf. Das ergibt sich jeweils aus dem Zusammenhang. Relevant wird die Schreibweise nur bei der Multiplikation einer Matrix mit einem Vektor. Spaltenvektoren werden von rechts, Zeilenvektoren von links gegen eine Matrix multipliziert.

- f) $(\mathbb{C}^{m \times n}, +; \mathbb{C}, \cdot)$ ist ein Vektorraum mit der gleichen Struktur wie $(\mathbb{C}^{m \cdot n}, +; \mathbb{C}, \cdot)$.
g) Die Menge der Funktionen von $D \subset \mathbb{R} \rightarrow \mathbb{R}$ ist ebenfalls ein reeller Vektorraum mit den Verknüpfungen (vgl. Kapitel 1.5.1):

$$(f + g)(x) := f(x) + g(x), \quad (c \cdot f)(x) := cf(x). \quad (3.15)$$

Man beachte, dass auf der jeweils linken Seite mit Funktionen gerechnet wird, während auf der rechten Seite mit Funktionswerten gearbeitet wird.

- h) Die Menge der Polynome $\mathbb{P}_n$ vom Grad $\leq n$ mit reellen Koeffizienten ist ein Vektorraum über $\mathbb{K} = \mathbb{R}$. Insbesondere führen die Addition zweier Polynome und die Multiplikation mit einem Skalar nicht aus dieser Menge hinaus, der Grad bleibt $\leq n$.
i) Der vielleicht für die Anwendungen wichtigste Vektorraum wird durch die Menge der Lösungen eines homogenen linearen Gleichungssystems gegeben. Mit ihm beschäftigen wir uns intensiv in Kapitel 3.5. ■

Auf den Vektorraumaxiomen baut die gesamte Theorie der Linearen Algebra auf. Als Beispiel zeigen wir mit den Axiomen einige anscheinend offensichtliche Rechenregeln. Allerdings sind sie nicht Teil der Vektorraumdefinition und müssen daher bewiesen werden.

Lemma 3.6 (Folgerungen aus den Vektorraumaxiomen)

- a) *Jedes Vielfache des Nullvektors ist wieder der Nullvektor: $\lambda \cdot \vec{0} = \vec{0}$ für jedes $\lambda \in \mathbb{K}$.*
 b) *Multipliziert man den Skalar Null mit einem beliebigen Vektor $\vec{a} \in V$, so entsteht der Nullvektor:*

$$0 \cdot \vec{a} = \vec{0}. \quad (3.16)$$

- c) *Man erhält das zu $\vec{a} \in V$ bezüglich der Vektoraddition inverse Element $-\vec{a}$, indem man $\vec{a}$ mit dem Skalar -1 multipliziert: $-\vec{a} = (-1) \cdot \vec{a}$.*

Beweis: a) Da

$$\lambda \cdot \vec{0} \stackrel{\text{Def. 3.7a)iv)}}{=} \lambda \cdot (\vec{0} + \vec{0}) \stackrel{\text{Def. 3.7b)iv)}}{=} \lambda \cdot \vec{0} + \lambda \cdot \vec{0}$$

und $\lambda \cdot \vec{0} = \lambda \cdot \vec{0} + \vec{b}$ wegen Definition 3.7 a)iv) nur für $\vec{b} = \vec{0}$ gilt, muss $\lambda \cdot \vec{0} = \vec{0}$ sein.

b) Der Beweis dieser Regel ist analog: Da

$$0 \cdot \vec{a} = (0 + 0) \cdot \vec{a} \stackrel{\text{Def. 3.7b)iv)}}{=} 0 \cdot \vec{a} + 0 \cdot \vec{a}$$

und $0 \cdot \vec{a} = 0 \cdot \vec{a} + \vec{b}$ wegen Definition 3.7 a)iv) nur für $\vec{b} = \vec{0}$ gilt, muss $0 \cdot \vec{a} = \vec{0}$ sein.

c) Für jeden Vektor $\vec{a}$ ist

$$\vec{a} + (-1) \cdot \vec{a} \stackrel{\text{Def. 3.7b)iii)}}{=} 1 \cdot \vec{a} + (-1) \cdot \vec{a} \stackrel{\text{Def. 3.7b)iv)}}{=} (1 - 1) \cdot \vec{a} = 0 \cdot \vec{a} \stackrel{(3.16)}{=} \vec{0}.$$

Damit ist aber nach Definition 3.7 a)v) genau $(-1) \cdot \vec{a}$ das inverse Element zu $\vec{a}$.

□

3.3.1.2 Linearkombination, Erzeugendensysteme und lineare Hülle

Jeden Pfeil in der Ebene, also jeden Vektor $(x, y)^\top$ aus $(\mathbb{R}^2, +; \mathbb{R}, \cdot)$ kann man schreiben als Summe von gestreckten oder gestauchten Pfeilen in Richtung der beiden Koordinatenachsen und damit als eine „Linearkombination“ von nur zwei festgelegten Pfeilen $(1, 0)^\top$ und $(0, 1)^\top$: $(x, y)^\top = x \cdot (1, 0)^\top + y \cdot (0, 1)^\top$.

Definition 3.8 (Linearkombination)

Sind $\vec{a}_1, \dots, \vec{a}_n \in V$, so nennt man

$$x_1 \vec{a}_1 + x_2 \vec{a}_2 + \dots + x_n \vec{a}_n \in V$$

mit $x_1, x_2, \dots, x_n \in \mathbb{K}$ eine **Linearkombination** der Vektoren $\vec{a}_1, \dots, \vec{a}_n$.

Beispiel 3.24

a) In $\mathbb{R}^2$ kombinieren wir den Vektor $\vec{x} = (1, 2)^\top$ aus den Vektoren $\vec{a} = (0, 1)^\top$ und $\vec{b} = (7, 1)^\top$ linear. Wir suchen dazu nach $\lambda, \mu \in \mathbb{R}$ mit

$$\vec{x} = \lambda \cdot \vec{a} + \mu \cdot \vec{b} \iff \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \lambda \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \mu \cdot \begin{pmatrix} 7 \\ 1 \end{pmatrix} \iff \begin{array}{l} 1 = 7\mu \\ 2 = \lambda + \mu. \end{array}$$

Damit ist $\mu = \frac{1}{7}$ und $\lambda = \frac{13}{7}$. Also gilt $\vec{x} = \frac{13}{7} \cdot \vec{a} + \frac{1}{7} \cdot \vec{b}$.

b) Mit den n -dimensionalen Standard-Einheitsvektoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \quad \vec{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}, \quad (3.17)$$

kann jeder Vektor $\vec{x} \in \mathbb{R}^n$ als Linearkombination

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \cdot \vec{e}_1 + x_2 \cdot \vec{e}_2 + \dots + x_n \cdot \vec{e}_n$$

dargestellt werden. Dies ist die Summe der Komponenten, siehe Seite 434.

c) Der Vektor $\vec{x} = (0, 0, 1)^\top$ ist nicht durch $\vec{a} = (1, 0, 0)^\top$ und $\vec{b} = (0, 1, 0)^\top$ linear kombinierbar, da das folgende Gleichungssystem keine Lösung besitzt:

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \lambda \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \mu \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \iff \begin{array}{l} 0 = \lambda, \\ 0 = \mu, \\ 1 = 0. \end{array} \quad \blacksquare$$

Im Beispiel haben wir gesehen, wie sich bestimmte Vektoren aus gegebenen Vektoren linear kombinieren lassen. Wie sieht nun die Menge aller Vektoren aus, die sich als Linearkombination von einigen vorgegebenen Vektoren schreiben lassen?

Lemma 3.7 (Lineare Hülle)

Hat man einen Vektorraum $(V, +; \mathbb{K}, \cdot)$ und Vektoren $\vec{a}_1, \dots, \vec{a}_n \in V$, so ist die Menge

$$\{\vec{x} = \lambda_1 \vec{a}_1 + \dots + \lambda_n \vec{a}_n : \lambda_k \in \mathbb{K}\}$$

zusammen mit der Addition in V und Multiplikation mit Skalaren aus $\mathbb{K}$ selbst bereits wieder ein Vektorraum. Er heißt die **lineare Hülle** der Vektoren $\vec{a}_1, \dots, \vec{a}_n$. Sie wird von diesen Vektoren **aufgespannt** oder **erzeugt**.

Das rechnet man sofort nach, da die Vektoraddition und die Multiplikation mit einem Skalar nicht aus der von $\vec{a}_1, \dots, \vec{a}_n \in V$ aufgespannten Menge führt. Die Vektorraumaxiome sind somit erfüllt, da sie für V gelten. Dieses Argument formulieren wir wegen seiner praktischen Bedeutung im Anschluß explizit als Satz 3.10.

Definition 3.9 (Erzeugendensystem)

Eine (endliche) Menge $\{\vec{a}_1, \dots, \vec{a}_n\} \subset V$ heißt genau dann ein **Erzeugendensystem** für V , wenn jeder Vektor $\vec{b} \in V$ als Linearkombination der Vektoren $\vec{a}_1, \dots, \vec{a}_n$ geschrieben werden kann, falls also V die lineare Hülle von $\vec{a}_1, \dots, \vec{a}_n$ ist.

Beispiel 3.25

- a) Nach Beispiel 3.24 b) bilden die drei Vektoren $(1, 0, 0)^\top$, $(0, 1, 0)^\top$ und $(0, 0, 1)^\top$ ein Erzeugendensystem des $\mathbb{R}^3$.
 b) Ein Erzeugendensystem für $(\mathbb{C}^3, +; \mathbb{C}, \cdot)$ ist beispielsweise

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

Denn jeder Vektor $\vec{x} \in \mathbb{C}^3$ lässt sich damit als komplexe Linearkombination schreiben:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + 0 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Offensichtlich ist der letzte Vektor gar nicht erforderlich, die ersten drei bilden bereits ein Erzeugendensystem. Das gegebene Erzeugendensystem ist also nicht **minimal**. ■

3.3.1.3 Unterräume

Die lineare Hülle von Vektoren $\{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n\}$ liefert einen Vektorraum, der im Ausgangsvektorraum V enthalten ist. Man spricht daher von einem **Untervektorraum** oder kurz einem **Unterraum**:

Definition 3.10 (Unterraum)

Seien $(V, +; \mathbb{K}, \cdot)$ und $(U, +; \mathbb{K}, \cdot)$ Vektorräume mit $U \subset V$ und mit gleicher Vektoraddition und skalarer Multiplikation, so heißt $(U, +; \mathbb{K}, \cdot)$ ein **Unterraum** des Vektorraums $(V, +; \mathbb{K}, \cdot)$.

Unterräume werden später z. B. beim Lösen linearer Gleichungssysteme benötigt. Um zu prüfen, dass eine Teilmenge U eines Vektorraums V einen Unterraum bildet, muss man wie zuvor bei der linearen Hülle nur zeigen, dass die Addition und skalare Multiplikation nicht aus der Menge U hinausführen. Alle Rechenregeln, die für V gelten, gelten automatisch auch für U .

Satz 3.10 (Unterraum-Kriterium)

Es sei V ein Vektorraum. Eine Teilmenge $U \subset V$ ist genau dann ein Unterraum von V , wenn gilt:

- Für alle $\vec{a}, \vec{b} \in U$ ist $\vec{a} + \vec{b} \in U$, und
- für alle $\vec{a} \in U$ und alle $\lambda \in \mathbb{K}$ folgt $\lambda \cdot \vec{a} \in U$.

Da ein Unterraum ein Vektorraum ist, muss er zwangsläufig den Nullvektor enthalten. Jeder Vektorraum V besitzt mindestens als Unterräume den Raum V selbst und den Nullvektorraum $\{\vec{0}\}$.

Für die geometrische Interpretation von Vektoren des $\mathbb{R}^n$ unterscheiden wir jetzt nicht mehr explizit zwischen Punkten und Ortsvektoren, d. h., Punkte sind bei Vektorraum Betrachtungen als Ortsvektoren zu verstehen.

Beispiel 3.26

- a) Eine durch den Ursprung verlaufende Gerade U in $\mathbb{R}^2$ ist ein Unterraum von $\mathbb{R}^2$. Es ist geometrisch offensichtlich, dass Summen und skalare Vielfache von Vektoren aus U wieder auf der Geraden U liegen. Somit ist das Unterraum-Kriterium für U erfüllt.
- b) Mit der gleichen Begründung ist eine durch den Ursprung $\vec{0}$ verlaufende Ebene in $\mathbb{R}^3$ ein Unterraum.
- c) Geraden und Ebenen, die in $\mathbb{R}^3$ nicht durch den Ursprung verlaufen, sind keine Unterräume, da sie den Nullvektor nicht enthalten.
- d) Für $V = \mathbb{R}^3$ ist die Menge $U = \{\vec{x} = (x_1, x_2, x_3)^\top : x_3 = 0\}$ ein Unterraum von V , denn mit $\vec{a}, \vec{b} \in U$ und $\lambda \in \mathbb{R}$ folgt

$$\begin{pmatrix} a_1 \\ a_2 \\ 0 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ 0 \end{pmatrix} \in U, \quad \lambda \cdot \begin{pmatrix} a_1 \\ a_2 \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda \cdot a_1 \\ \lambda \cdot a_2 \\ 0 \end{pmatrix} \in U.$$

Dieser Unterraum ist zwar formal nicht der $\mathbb{R}^2$, hat aber die exakt gleiche Struktur.

- e) Der Raum der Polynome $\mathbb{P}_n$ vom Grad $\leq n$ ist ein Unterraum des Vektorraums aller Funktionen: Die Summe zweier Polynome aus $\mathbb{P}_n$ ist wieder ein Polynom aus $\mathbb{P}_n$, und das skalare Vielfache eines Polynoms aus $\mathbb{P}_n$ ist wieder aus $\mathbb{P}_n$. ■

3.3.2 Lineare Unabhängigkeit, Basis und Dimension

3.3.2.1 Lineare Unabhängigkeit und lineare Abhängigkeit

In Beispiel 3.25 haben wir ein Erzeugendensystem vorgefunden, das nicht minimal war. Für die Suche nach minimalen Erzeugendensystemen benötigen wir:

Definition 3.11 (Lineare Unabhängigkeit und Abhängigkeit)

Sei $(V, +; \mathbb{K}, \cdot)$ ein Vektorraum. Die Menge von Vektoren $\{\vec{a}_1, \dots, \vec{a}_n\} \subset V$ heißt **linear unabhängig** genau dann, wenn die Gleichung

$$x_1 \vec{a}_1 + x_2 \vec{a}_2 + \dots + x_n \vec{a}_n = \vec{0}, \quad \vec{x} \in \mathbb{K}^n$$

nur die Lösung $\vec{x} = (x_1, x_2, \dots, x_n)^\top = \vec{0}$ besitzt. Anderenfalls heißt die Menge **linear abhängig**. Statt von der linearen (Un-)Abhängigkeit einer Menge sprechen wir auch von der linearen (Un-)Abhängigkeit der Vektoren der Menge.

Lineare Unabhängigkeit von Vektoren bedeutet, dass man den Nullvektor nur als **triviale Linearkombination** dieser Vektoren schreiben kann. Mit anderen Worten besagt die Definition:

$\{\vec{a}_1, \dots, \vec{a}_n\}$ ist linear unabhängig

$$\iff \left[\sum_{k=1}^n x_k \vec{a}_k = \vec{0} \implies x_1 = \dots = x_n = 0 \right].$$

$\{\vec{a}_1, \dots, \vec{a}_n\}$ ist linear abhängig

$$\iff \text{Es existieren } x_1, x_2, \dots, x_n \in \mathbb{K} \text{ mit } \sum_{k=1}^n x_k \vec{a}_k = \vec{0} \text{ und } (x_1, x_2, \dots, x_n) \neq \vec{0}.$$

Beispiel 3.27

Die Vektoren

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

sind linear unabhängig in $(\mathbb{R}^3, +; \mathbb{R}, \cdot)$, $(\mathbb{C}^3, +; \mathbb{R}, \cdot)$ und $(\mathbb{C}^3, +; \mathbb{C}, \cdot)$, da

$$x_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \iff \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

nur die Lösung $x_1 = x_2 = x_3 = 0$ besitzt. ■

Lemma 3.8 (Charakterisierung der linearen Unabhängigkeit)

Eine äquivalente Formulierung der linearen Unabhängigkeit einer Menge aus $n > 1$ Vektoren ist:

$$\text{„Keiner der Vektoren ist als Linearkombination} \quad (3.18) \\ \text{der anderen } n - 1 \text{ Vektoren darstellbar.“}$$

Beweis: a) Wir zeigen mittels Widerspruch, dass aus der linearen Unabhängigkeit die Aussage (3.18) folgt. Wir nehmen an, dass ein Vektor durch die anderen Vektoren darstellbar ist, also (3.18) nicht gilt. o. B. d. A. sei $\vec{a}_1$ durch die anderen darstellbar:

$$\vec{a}_1 = x_2 \vec{a}_2 + \cdots + x_n \vec{a}_n \implies \vec{a}_1 - x_2 \vec{a}_2 - \cdots - x_n \vec{a}_n = \vec{0},$$

d. h., die Vektoren sind im Widerspruch zur Voraussetzung linear abhängig. Die Annahme ist falsch, und stattdessen gilt (3.18).

b) Umgekehrt folgt aus (3.18) die lineare Unabhängigkeit: Wären die Vektoren nicht unabhängig, so gäbe es eine nicht-triviale Linearkombination des Nullvektors, z. B. $x_1 \neq 0$ und $x_1 \vec{a}_1 + x_2 \vec{a}_2 + \cdots + x_n \vec{a}_n = \vec{0}$. Dann ist aber im Widerspruch zu (3.18):

$$\vec{a}_1 = -\frac{x_2}{x_1} \vec{a}_2 - \cdots - \frac{x_n}{x_1} \vec{a}_n.$$

Folglich müssen die Vektoren linear unabhängig sein. $\square$

Linear unabhängige Vektoren sind eigenständig und können nicht durch die anderen ersetzt werden.

Beispiel 3.28

- a) Der Vektor $\vec{0}$ ist linear abhängig, da $c \cdot \vec{0} = \vec{0}$ für jedes $c \in \mathbb{K}$. Betrachtet man n Vektoren, von denen einer $\vec{0}$ ist, so sind diese linear abhängig. Ist ein einzelner Vektor linear abhängig, so muss es sich um $\vec{0}$ handeln.
- b) Zwei Vektoren $\vec{a}, \vec{b}$ sind linear abhängig, wenn $\lambda_1 \cdot \vec{a} + \lambda_2 \cdot \vec{b} = \vec{0}$ mit $\lambda_1 \neq 0$ oder $\lambda_2 \neq 0$. Dann gilt

$$\vec{a} = -\frac{\lambda_2}{\lambda_1} \cdot \vec{b} \quad \text{oder} \quad \vec{b} = -\frac{\lambda_1}{\lambda_2} \cdot \vec{a},$$

d. h., $\vec{a}$ und $\vec{b}$ sind kollinear.

- c) Die Vektoren $\vec{a}, \vec{b}, \vec{c} \in \mathbb{R}^3$ sind linear abhängig, wenn sie in einer Ebene (oder auf einer Gerade) liegen. Dann nennt man $\vec{a}, \vec{b}, \vec{c}$ auch **komplanar**.
- d) Die drei Vektoren in $\mathbb{R}^2$

$$\vec{a}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \vec{a}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \vec{a}_3 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

sind linear abhängig, denn der Nullvektor kann als nicht-triviale Linearkombination aus den drei Vektoren dargestellt werden:

$$\vec{a}_1 + (-3) \cdot \vec{a}_2 + \vec{a}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} -3 \\ -3 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Alternativ erkennt man die lineare Abhängigkeit auch daran, dass ein Vektor als Linearkombination der übrigen geschrieben werden kann:

$$\vec{a}_2 = \frac{1}{3} \cdot \vec{a}_1 + \frac{1}{3} \cdot \vec{a}_3.$$

e) Die drei Vektoren in $\mathbb{R}^3$

$$\vec{a}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{a}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{a}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

sind linear unabhängig, denn der Nullvektor kann nur trivial als Linearkombination aus den drei Vektoren dargestellt werden. Aus

$$\vec{0} = \lambda_1 \cdot \vec{a}_1 + \lambda_2 \cdot \vec{a}_2 + \lambda_3 \cdot \vec{a}_3 = \begin{pmatrix} \lambda_1 \\ \lambda_1 + \lambda_2 + \lambda_3 \\ \lambda_3 \end{pmatrix}$$

folgt $\lambda_1 = 0$, $\lambda_3 = 0$ und somit auch $\lambda_2 = 0$.

f) Die Vektoren

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

sind linear abhängig in $(\mathbb{R}^3, +; \mathbb{R}, \cdot)$, da

$$\vec{0} = 1 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - 1 \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Insbesondere sehen wir, dass der Vektor $(1, 1, 1)^\top$ als Summe (und damit als Linearkombination) der Vektoren $(1, 1, 0)^\top$ und $(0, 0, 1)^\top$ darstellbar ist.

Es gibt aber noch eine weitere wichtige Beobachtung: Je zwei der Vektoren sind linear unabhängig. Aus der paarweisen linearen Unabhängigkeit folgt also **nicht** die lineare Unabhängigkeit aller Vektoren.

g) Wie zuvor schon für den $\mathbb{R}^3$ gesehen, sind in $\mathbb{R}^n$ die Vektoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \vec{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \quad (3.19)$$

linear unabhängig.

h) Man beachte, dass zwei Funktionen $f, g : D \subset \mathbb{R} \rightarrow \mathbb{R}$ gleich sind, wenn alle Funktionswerte übereinstimmen. Damit sind $f_1, f_2, \dots, f_k$ linear unabhängig genau dann, wenn aus

$$0 = c_1 f_1(x) + c_2 f_2(x) + \dots + c_k f_k(x) \text{ für alle } x \in D$$

folgt, dass $c_1 = c_2 = \dots = c_k = 0$.

Die Monome $1, x, x^2, \dots, x^n$ sind linear unabhängig: Aus

$$0 = c_0 + c_1 x + c_2 x^2 + \dots + c_n x^n$$

folgt durch Einsetzen von $x = 0$ sofort $c_0 = 0$. Ableiten beider Seiten ergibt

$$0 = c_1 + 2c_2 x + \dots + n c_n x^{n-1}.$$

Wir setzen wieder $x = 0$ ein und erhalten $c_1 = 0$. Dies kann man fortsetzen, so dass $c_0 = c_1 = \dots = c_n = 0$ gilt. Die Monome sind also linear unabhängig. Alternativ können wir das auch direkt mit einem Koeffizientenvergleich des Polynoms 0 mit dem Polynom $c_0 + c_1 x + c_2 x^2 + \dots + c_n x^n$ sehen. ■

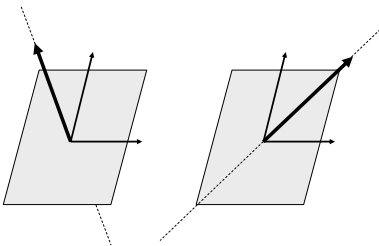


Abb. 3.26: Zwei linear unabhängige Vektoren spannen eine Ebene auf. Drei Vektoren sind linear unabhängig (links) bzw. linear abhängig (rechts).

3.3.2.2 Basis und Dimension

Mit Hilfe eines Erzeugendensystems lassen sich alle Vektoren eines Vektorraums V als Linearkombination schreiben. Die Frage nach einem (minimalen) Erzeugendensystem mit einer möglichst kleinen Anzahl von Vektoren führt auf den Begriff der Basis.

Definition 3.12 (Basis)

Es sei V ein Vektorraum. Die Menge der Vektoren $B := \{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n\} \subset V$ heißt genau dann eine **Basis** von V , wenn sie linear unabhängig ist und ein Erzeugendensystem von V bildet.

Satz 3.11 (Basisdarstellung)

Es sei $B = \{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n\}$ eine Basis von V . Dann kann jeder Vektor $\vec{x} \in V$ als **Linearkombination**

$$\vec{x} = \lambda_1 \cdot \vec{a}_1 + \lambda_2 \cdot \vec{a}_2 + \dots + \lambda_n \cdot \vec{a}_n = \sum_{k=1}^n \lambda_k \cdot \vec{a}_k$$

geschrieben werden. Diese Darstellung ist (bezüglich der Faktoren $\lambda_1, \dots, \lambda_n$) **eindeutig**.

Beweis: Da B insbesondere ein Erzeugendensystem ist, lässt sich $\vec{x}$ als Linearkombination schreiben. Die Eindeutigkeit der Darstellung sieht man wie folgt ein. Angenommen, es gibt zwei Darstellungen von $\vec{x}$, $\vec{x} = \lambda_1 \vec{a}_1 + \dots + \lambda_n \vec{a}_n$ und $\vec{x} = \mu_1 \vec{a}_1 + \dots + \mu_n \vec{a}_n$. Dann folgt

$$\vec{0} = \vec{x} - \vec{x} = (\lambda_1 - \mu_1) \vec{a}_1 + \dots + (\lambda_n - \mu_n) \vec{a}_n,$$

und aus der linearen Unabhängigkeit von $\{\vec{a}_1, \dots, \vec{a}_n\}$ folgt $\lambda_1 = \mu_1, \dots, \lambda_n = \mu_n$. $\square$

Beispiel 3.29

In $\mathbb{R}^2$ bilden die Vektoren

$$\vec{a}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{a}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

eine Basis. $\vec{a}_1$ und $\vec{a}_2$ sind linear unabhängig:

$$\lambda_1 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_2 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} \lambda_1 + \lambda_2 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

d. h., es folgt $\lambda_2 = 0$ und $\lambda_1 = 0$. Außerdem bilden sie ein Erzeugendensystem: Jeder Vektor $\vec{x} = (x_1, x_2)^\top$ kann als Linearkombination von $\vec{a}_1$ und $\vec{a}_2$ geschrieben werden:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1 - x_2) \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad \blacksquare$$

Beispiel 3.30

a) Eine Basis von $(\mathbb{R}^3, +; \mathbb{R}, \cdot)$ bzw. von $(\mathbb{C}^3, +; \mathbb{C}, \cdot)$ ist

$$B := \left\{ \vec{e}_1 := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \vec{e}_2 := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \vec{e}_3 := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

b) Allgemeiner ist $\{\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n\}$ wie in (3.19) die **Standardbasis** von $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$ bzw. von $(\mathbb{C}^n, +; \mathbb{C}, \cdot)$.

c) Eine andere Basis von $(\mathbb{C}^3, +; \mathbb{C}, \cdot)$ ist

$$\left\{ \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ j \end{pmatrix} \right\}.$$

d) Diese Mengen sind aber kein Erzeugendensystem und damit keine Basis von $(\mathbb{C}^3, +; \mathbb{R}, \cdot)$. Dieser Vektorraum hat z. B. die Basis

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} j \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ j \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ j \end{pmatrix} \right\}.$$

e) Der reelle Vektorraum der Polynome $\mathbb{P}_n$ von Grad $\leq n$ hat eine Basis $\{1, x, x^2, x^3, \dots, x^n\}$. Die lineare Unabhängigkeit der Monome haben wir bereits auf Seite 477 gezeigt. Zudem ist klar, dass sie ein Erzeugendensystem bilden. ■

Lemma 3.9 (Existenz einer Basis)

Hat man in einem Vektorraum $(V, +; \mathbb{K}, \cdot)$ mit $V \neq \{\vec{0}\}$ ein Erzeugendensystem mit m Vektoren, so gibt es eine Basis mit $n \leq m$ Vektoren.

Beweis: Ist im Erzeugendensystem der Nullvektor vorhanden, so kann er wegen $V \neq \{\vec{0}\}$ weggelassen werden. Falls das resultierende Erzeugendensystem dann noch linear abhängig ist, so hat es mindestens zwei Vektoren, und man kann einen Vektor $\vec{a}$ als Linearkombination der anderen schreiben. Lässt man $\vec{a}$ weg, hat man immer noch ein Erzeugendensystem. Solange linear abhängige Erzeugendensysteme entstehen, wiederholt man diesen Schritt. Das danach verbleibende Erzeugendensystem (mit mindestens einem Vektor) ist linear unabhängig und damit eine Basis. □

Satz 3.12 (Austauschsatz)

Sei $\{\vec{a}_1, \dots, \vec{a}_n\}$ eine Basis in $(V, +; \mathbb{K}, \cdot)$. Weiter sei $\vec{b} = c_1 \vec{a}_1 + c_2 \vec{a}_2 + \dots + c_n \vec{a}_n$ eine Linearkombination mit $c_1 \neq 0$. Dann ist auch $\{\vec{b}, \vec{a}_2, \dots, \vec{a}_n\}$ eine Basis.

Beweis: Zum Nachweis der Basis-Eigenschaft müssen wir zeigen, dass die Vektoren $\{\vec{b}, \vec{a}_2, \dots, \vec{a}_n\}$ ein linear unabhängiges Erzeugendensystem bilden.

a) Wir zeigen zunächst die lineare Unabhängigkeit. Aus der Gleichung

$$\begin{aligned}\vec{0} &= d_1\vec{b} + d_2\vec{a}_2 + \dots + d_n\vec{a}_n = d_1[c_1\vec{a}_1 + \dots + c_n\vec{a}_n] + d_2\vec{a}_2 + \dots + d_n\vec{a}_n \\ &= (d_1c_1)\vec{a}_1 + (d_1c_2 + d_2)\vec{a}_2 + \dots + (d_1c_n + d_n)\vec{a}_n\end{aligned}$$

folgt wegen der linearen Unabhängigkeit von $\vec{a}_1, \dots, \vec{a}_n$ das Gleichungssystem

$$d_1c_1 = 0 \wedge d_1c_2 + d_2 = 0 \wedge \dots \wedge d_1c_n + d_n = 0.$$

Da $c_1 \neq 0$ gilt, ist $d_1 = 0$. Damit sind aber auch $d_2 = d_3 = \dots = d_n = 0$, und die lineare Unabhängigkeit ist bewiesen.

b) Wir müssen noch zeigen, dass es sich um ein Erzeugendensystem handelt. Jeder Vektor $\vec{a} \in V$ hat nach Voraussetzung eine Darstellung $\vec{a} = d_1\vec{a}_1 + d_2\vec{a}_2 + \dots + d_n\vec{a}_n$. Damit kann er aber wegen

$$\vec{a}_1 = \frac{1}{c_1}[\vec{b} - c_2\vec{a}_2 - \dots - c_n\vec{a}_n]$$

auch so geschrieben werden:

$$\begin{aligned}\vec{a} &= d_1 \left[\frac{1}{c_1}[\vec{b} - c_2\vec{a}_2 - \dots - c_n\vec{a}_n] \right] + d_2\vec{a}_2 + \dots + d_n\vec{a}_n \\ &= \frac{d_1}{c_1}\vec{b} + \left[d_2 - \frac{d_1c_2}{c_1} \right] \vec{a}_2 + \dots + \left[d_n - \frac{d_1c_n}{c_1} \right] \vec{a}_n.\end{aligned}$$

Somit handelt es sich auch bei $\{\vec{b}, \vec{a}_2, \dots, \vec{a}_n\}$ um ein Erzeugendensystem. $\square$

Satz 3.13 (Anzahl der Vektoren einer Basis)

Der Vektorraum V habe eine Basis mit n Vektoren. Dann haben alle Basen dieses Vektorraums genau n Vektoren.

Beweis: Seien $\{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n\}$ und $\{\vec{b}_1, \vec{b}_2, \dots, \vec{b}_m\}$ Basen.

- Fall $m > n$: $\vec{b}_1$ ist darstellbar als eine Linearkombination $\sum_{k=1}^n c_k \vec{a}_k$. Da $\vec{b}_1 \neq \vec{0}$ ist, muss mindestens ein $c_k \neq 0$ sein. Wir sortieren o. B. d. A. die $\vec{a}_k$ so um, dass $c_1 \neq 0$ ist. Nach Austauschsatz ist damit auch $\{\vec{b}_1, \vec{a}_2, \dots, \vec{a}_n\}$ eine Basis. Jetzt ist $\vec{b}_2 = d_1\vec{b}_1 + \sum_{k=2}^n d_k \vec{a}_k$. Da $\vec{b}_1$ und $\vec{b}_2$ linear unabhängig sind, muss ein d_k , $k \geq 2$, ungleich 0 sein. Wir sortieren die Vektoren so um, dass $d_2 \neq 0$. Nach dem Austauschsatz ist damit auch $\{\vec{b}_1, \vec{b}_2, \dots, \vec{a}_n\}$ eine Basis. Dieses Argument lässt sich nun fortsetzen, so dass $\{\vec{b}_1, \vec{b}_2, \dots, \vec{b}_n\}$ eine Basis ist. Damit kann aber $\vec{b}_{n+1}$ als Linearkombination der Vektoren $\vec{b}_1, \dots, \vec{b}_n$ geschrieben werden – im Widerspruch zur linearen Unabhängigkeit.

- Fall $m < n$: Analog erhält man einen Widerspruch mit vertauschten Basen.

Damit bleibt nur $n = m$. □

Wir haben damit die vielleicht wichtigste charakteristische Größe eines Vektorraums entdeckt und geben ihr einen Namen:

Definition 3.13 (Dimension)

*Man sagt, dass ein Vektorraum mit einer Basis aus n Vektoren die **Dimension** n besitzt, Bezeichnung: $\dim V = n$.*

Beispiel 3.31

- a) $(\mathbb{R}^3, +; \mathbb{R}, \cdot)$ hat die Dimension 3. Das passt zur umgangssprachlichen Formulierung „dreidimensionaler Raum“ und zur bisherigen informellen Verwendung des Begriffs „Dimension“.
- b) $(\mathbb{C}^3, +; \mathbb{C}, \cdot)$ hat ebenfalls die Dimension 3,
- c) Dagegen hat $(\mathbb{C}^3, +; \mathbb{R}, \cdot)$ die Dimension 6,
- d) $(\mathbb{C}^{m \times n}, +; \mathbb{C}, \cdot)$ hat die Dimension $m \cdot n$. ■

Folgerung 3.1 (Bedeutung der Dimension)

Sei V ein Vektorraum der Dimension n , d. h. V besitzt eine Basis aus n Vektoren.

- a) *Weniger als n Vektoren können kein Erzeugendensystem bilden.*
- b) *Es sind nie mehr als n Vektoren linear unabhängig.*

Beweis: a) Hätte man ein Erzeugendensystem mit weniger als n Vektoren, dann gäbe es nach Lemma 3.9 auch eine Basis mit weniger als n Vektoren. Das wäre aber ein Widerspruch dazu, dass bei der Dimension n alle Basen n Vektoren haben (Satz 3.13).

b) Hätte man $m > n$ linear unabhängige Vektoren, so gibt es zwei Fälle:

- Sie sind ein Erzeugendensystem. Das wäre wieder ein Widerspruch zu Satz 3.13.
- Sie sind kein Erzeugendensystem und damit keine Basis des Vektorraums. Sei $\{\vec{a}_1, \dots, \vec{a}_n\}$ die vorhandene Basis. Dann lässt sich mindestens ein Vektor $\vec{a}_k$ nicht als Linearkombination der m linear unabhängigen Vektoren schreiben. Damit hat man dann aber schon $m + 1$ linear unabhängige Vektoren. Man kann nun so lange weitere Elemente der Basis ergänzen, bis man schließlich ein linear unabhängiges Erzeugendensystem (eine Basis) hat. Das hat aber mehr als n Elemente im Widerspruch zum Satz 3.13. □

Mittels der Folgerung können wir direkt zeigen, dass nicht alle Vektorräume eine endliche Dimension haben:

Lemma 3.10 (Unendlich-dimensionaler Vektorraum)

Der Vektorraum aller Polynome auf $\mathbb{R}$ hat keine endliche Dimension.

Beweis: Für jedes $n \in \mathbb{N}$ sind die Polynome $1, x, x^2, \dots, x^n$ linear unabhängig, so dass wegen Folgerung 3.1 b) die Dimension des Vektorraums der Polynome $\geq n$ sein muss. $\square$

Satz 3.14 (Basissatz)

Es sei V ein Vektorraum und $\dim(V) = n$.

- a) *Ist die Menge der Vektoren $\{\vec{a}_1, \dots, \vec{a}_n\}$ linear unabhängig, so bildet sie eine Basis.*
- b) *Ist die Menge der Vektoren $\{\vec{a}_1, \dots, \vec{a}_n\}$ ein Erzeugendensystem, so bildet sie eine Basis.*

Beweis: a) Sei $\vec{x} \in V$ beliebig. Wegen $\dim(V) = n$ sind $n+1$ Vektoren nach Folgerung 3.1 stets linear abhängig. Insbesondere gilt dies für $\{\vec{x}, \vec{a}_1, \dots, \vec{a}_n\}$. Also existiert eine Linearkombination

$$\lambda_1 \cdot \vec{a}_1 + \lambda_2 \cdot \vec{a}_2 + \dots + \lambda_n \cdot \vec{a}_n + \lambda_{n+1} \cdot \vec{x} = \vec{0}$$

mit mindestens einem $\lambda_k \neq 0$, $1 \leq k \leq n+1$. Nun ist $\lambda_{n+1} = 0$ nicht möglich, denn sonst wäre $\lambda_1 \cdot \vec{a}_1 + \dots + \lambda_n \cdot \vec{a}_n = \vec{0}$ mit einem $\lambda_k \neq 0$, $1 \leq k \leq n$, entgegen der Annahme, dass $\{\vec{a}_1, \dots, \vec{a}_n\}$ linear unabhängig ist. Also folgt

$$\vec{x} = -\frac{\lambda_1}{\lambda_{n+1}} \cdot \vec{a}_1 - \frac{\lambda_2}{\lambda_{n+1}} \cdot \vec{a}_2 - \dots - \frac{\lambda_n}{\lambda_{n+1}} \cdot \vec{a}_n,$$

und $\vec{x}$ ist als Linearkombination aus $\{\vec{a}_1, \dots, \vec{a}_n\}$ darstellbar. Damit ist $\{\vec{a}_1, \dots, \vec{a}_n\}$ ein Erzeugendensystem, das wegen der linearen Unabhängigkeit eine Basis ist.

- b) Wären die Vektoren nicht linear unabhängig, so könnte man einen als Linearkombination der anderen schreiben. Lässt man ihn weg, so hat man immer noch ein Erzeugendensystem. Das hat aber weniger als n Elemente im Widerspruch zu Folgerung 3.1. $\square$

Schreibt man die Vektoren einer Basis des $\mathbb{R}^n$ (oder des $\mathbb{C}^n$) als Spalten (oder Zeilen) in eine $(n \times n)$ -Matrix, so ist diese invertierbar, denn jeder Standard-Einheitsvektor kann aus den Spalten (Zeilen) kombiniert werden, so dass das Gauß-Verfahren zur Berechnung der inversen Matrix zu einer Lösung führt (vgl. Seite 170). Insbesondere ist damit die Determinante der Matrix ungleich null (vgl. Seite 189). Hat man umgekehrt eine Matrix $\mathbf{A}$ mit von null verschiedener Determinante, so sind die Spalten (Zeilen) eine Basis des $\mathbb{R}^n$. Denn da die Matrix invertierbar ist, kann der Nullvektor nur als triviale Linearkombination der Spalten (Zeilen) dargestellt werden ($\mathbf{A} \cdot \vec{x} = \vec{0} \iff \vec{x} =$

$\mathbf{A}^{-1} \cdot \vec{0} = \vec{0}$), die Spalten (Zeilen) sind linear unabhängig und bilden nach Satz 3.14 eine Basis. Über die Berechnung einer Determinante kann man also die Basiseigenschaft prüfen:

Satz 3.15 (Basis und Determinante)

Die Menge $\{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n\} \subset \mathbb{R}^n$ ist genau dann eine Basis des Vektorraums $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$, wenn die $(n \times n)$ -Matrix $\mathbf{A}$, deren Spalten (oder alternativ Zeilen) genau die Vektoren $\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n$ sind, eine Determinante ungleich null hat.

3.3.3 Skalarprodukt und Norm

Der Begriff des Vektorraums abstrahiert vom Anschauungsraum. Statt Mengen von Pfeilen kann man nun Mengen von beliebigen Objekten betrachten, sofern für diese eine Addition und eine Multiplikation mit Skalaren erklärt ist, die den Vektorraumaxiomen genügt. Die Axiome stellen sicher, dass sich diese Verknüpfungen wie die entsprechenden Verknüpfungen für Pfeile verhalten.

Jetzt übertragen wir die Begriffe „Länge“ und „Winkel“ auf beliebige Elemente eines Vektorraums. Dazu müssen wir auch das Skalarprodukt allgemeiner fassen.

3.3.3.1 Euklid'scher Raum und Skalarprodukt

Definition 3.14 (Euklid'scher Raum, Skalarprodukt)

Ein **Euklid'scher Raum** ist ein reeller Vektorraum $(V, +; \mathbb{R}, \cdot)$ mit einem **Skalarprodukt**, d. h. einer Abbildung

$$„ \cdot “ : V \times V \mapsto \mathbb{R},$$

die für beliebige Vektoren $\vec{a}, \vec{b}, \vec{c} \in V$ und Skalare $\lambda \in \mathbb{R}$ die folgenden Regeln (Axiome) erfüllt:

- a) **Positive Definitheit:** Das Skalarprodukt eines Vektors mit sich selbst ist nicht-negativ: $\vec{a} \cdot \vec{a} \geq 0$. Aus $\vec{a} \cdot \vec{a} = 0$ folgt $\vec{a} = \vec{0}$.
- b) **Symmetrie:** Es gilt das Kommutativgesetz $\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{a}$.
- c) **Homogenität:** $(\lambda \vec{a}) \cdot \vec{b} = \lambda(\vec{a} \cdot \vec{b})$.
- d) **Additivität (Distributivgesetz):** $(\vec{a} + \vec{b}) \cdot \vec{c} = \vec{a} \cdot \vec{c} + \vec{b} \cdot \vec{c}$.

Wie zuvor sind Skalarprodukt und Skalarmultiplikation aus der Definition des Vektorraums zwei unterschiedliche Verknüpfungen, die nur einen ähnlichen Namen besitzen. Daher wird häufig (aber nicht in diesem Buch) das Skalarprodukt mit spitzen Klammern geschrieben: $\langle \vec{a}, \vec{b} \rangle := \vec{a} \cdot \vec{b}$.

Die Homogenität und die Additivität führen dazu, dass ein Skalarprodukt in beiden Argumenten linear ist, also z. B.

$$(\lambda \vec{a} + \mu \vec{b}) \cdot \vec{c} = \lambda(\vec{a} \cdot \vec{c}) + \mu(\vec{b} \cdot \vec{c}).$$

Man beachte aber, dass $(\lambda \vec{a}) \cdot (\lambda \vec{b}) = \lambda^2(\vec{a} \cdot \vec{b})$ ist.

Sofort aus der Definition des allgemeinen Skalarprodukts folgt:

$$\vec{a} \cdot \vec{0} = \vec{0} \cdot \vec{a} = (0 \vec{0}) \cdot \vec{a} = 0 \quad (\vec{0} \cdot \vec{a}) = 0, \quad (3.20)$$

d. h., die Regel zur positiven Definitheit in der Definition kann als Äquivalenz geschrieben werden.

Beispiel 3.32

a) Der Anschauungsraum $\mathbb{R}^3$ ist mit dem Skalarprodukt (siehe Definition 3.4 auf Seite 435 und Satz 3.3 auf Seite 438)

$$\vec{a} \cdot \vec{b} = |\vec{a}| \cdot |\vec{b}| \cdot \cos \varphi = a_1 \cdot b_1 + a_2 \cdot b_2 + a_3 \cdot b_3$$

ein Euklid'scher Raum. Dabei ist φ der Winkel zwischen $\vec{a}$ und $\vec{b}$.

b) Das Skalarprodukt unter a) lässt sich direkt auf den Raum $\mathbb{R}^n$ mit allgemeiner Dimension $n \in \mathbb{N}$ übertragen. Er wird mit dem **Standardskalarprodukt**

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

ebenfalls zu einem Euklid'schen Raum. Für die Definition benötigt man keinen Winkelbegriff. Die Axiome sind leicht zu verifizieren: Es ist

$$\begin{aligned} \vec{a} \cdot \vec{a} &= \sum_{k=1}^n a_k^2 \geq 0 \text{ und } > 0 \text{ genau dann, wenn mindestens ein } a_k \neq 0, \\ \vec{a} \cdot \vec{b} &= \sum_{k=1}^n a_k \cdot b_k = \sum_{k=1}^n b_k \cdot a_k = \vec{b} \cdot \vec{a}, \\ (\lambda \vec{a}) \cdot \vec{b} &= \sum_{k=1}^n \lambda a_k \cdot b_k = \lambda \sum_{k=1}^n a_k \cdot b_k = \lambda(\vec{a} \cdot \vec{b}), \\ (\vec{a} + \vec{b}) \cdot \vec{c} &= \sum_{k=1}^n (a_k + b_k) \cdot c_k = \sum_{k=1}^n a_k \cdot c_k + \sum_{k=1}^n b_k \cdot c_k = \vec{a} \cdot \vec{c} + \vec{b} \cdot \vec{c}. \end{aligned}$$

Damit haben die Vektoren $\vec{a} = (1, 2, 3, 4) \in \mathbb{R}^4$ und $\vec{b} = (4, 3, 2, 1) \in \mathbb{R}^4$ das Standardskalarprodukt

$$\vec{a} \cdot \vec{b} = \sum_{k=1}^4 a_k \cdot b_k = 4 + 6 + 6 + 4 = 20,$$

und für die Standardskalarprodukte der Standard-Einheitsvektoren aus $\mathbb{R}^n$ gilt

$$\vec{e}_i \cdot \vec{e}_k = \begin{cases} 1, & \text{falls } i = k, \\ 0, & \text{falls } i \neq k, \end{cases} \quad i, k = 1, \dots, n.$$

- c) Die auf dem Intervall $[-1, 1]$ stetigen Funktionen bilden einen Unterraum des Vektorraums aller reellen Funktionen auf diesem Intervall. Im Vektorraum $C[-1, 1]$ der stetigen Funktionen ist durch

$$f \bullet g := \int_{-1}^1 f(x) \cdot g(x) dx \quad (3.21)$$

ein Skalarprodukt erklärt. Die Eigenschaften prüft man leicht nach, so ist $f \bullet f = \int_{-1}^1 f^2(x) dx \geq 0$ und $f \bullet f = 0 \iff f(x) = 0$ für alle $x \in [-1, 1]$ wegen der Stetigkeit von f . Damit ist $\bullet$ positiv definit. Die Symmetrie ist offensichtlich, und die Additivität und Homogenität folgen aus den Eigenschaften des bestimmten Integrals. ■

Hintergrund: Skalarprodukt für einen komplexen Vektorraum

Definition 3.15 (Skalarprodukt)

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

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

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:

- a) **Positive Definitheit:** Das Skalarprodukt eines Vektors mit sich selbst ist eine nicht-negative reelle Zahl: $\vec{a} \cdot \vec{a} \geq 0$. Aus $\vec{a} \cdot \vec{a} = 0$ folgt $\vec{a} = \vec{0}$.
b) 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 $\vec{a} \cdot \vec{a} = \vec{a} \cdot \vec{a}$ ist insbesondere das Skalarprodukt eines Vektors mit sich selbst eine reelle Zahl. Das müssten wir bei der positiven Definitheit damit gar nicht fordern.

- c) **Multiplikation mit einem Skalar:** $(\lambda \vec{a}) \cdot \vec{b} = \bar{\lambda}(\vec{a} \cdot \vec{b})$, 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.
d) **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 $\vec{a} = (a_1, a_2, \dots, a_n)$ und $\vec{b} = (b_1, b_2, \dots, b_n) \in \mathbb{C}^n$ ist definiert als

$$\vec{a} \cdot \vec{b} := \sum_{k=1}^n \overline{a_k} \cdot b_k. \quad (3.22)$$

3.3.3.2 Betrag, Norm und Abstand

Mit dem Skalarprodukt lässt sich nach dem Vorbild des Anschauungsraums und konsistent zu den zuvor verwendeten Bezeichnungen ein Längen- und Abstandsbegriff in allgemeinen Vektorräumen einführen.

Definition 3.16 (Betrag, Norm und Abstand)

Es sei V ein Euklid'scher Raum und $\vec{a}, \vec{b} \in V$.

- $|\vec{a}| := \sqrt{\vec{a} \cdot \vec{a}}$ ist der **Betrag** oder die **Norm** von $\vec{a}$.
- Ein Vektor $\vec{a}$ mit Norm 1 (also $|\vec{a}| = 1$) heißt **Einheitsvektor**.
- $d(\vec{a}, \vec{b}) := |\vec{b} - \vec{a}|$ ist der **Abstand** von $\vec{a}$ und $\vec{b}$.

In einem komplexen Vektorraum mit einem komplexen Skalarprodukt werden diese Begriffe genauso definiert.

Beispiel 3.33

a) Für Vektoren in $\mathbb{R}^n$ ergibt das Standardskalarprodukt den Betrag

$$|\vec{a}| = \sqrt{\vec{a} \cdot \vec{a}} = \sqrt{\sum_{k=1}^n a_k^2}, \quad (3.23)$$

der auch **Euklid'sche Norm** heißt. Für $n = 2$ und $n = 3$ ist das die bekannte Definition von Seite 434, die mit dem Satz von Pythagoras übereinstimmt.

Insbesondere sind die Standard-Einheitsvektoren $\vec{e}_1, \vec{e}_2$ und $\vec{e}_3$ tatsächlich Einheitsvektoren in Richtung der Koordinatenachsen.

b) Der Vektor $\vec{a} \in \mathbb{R}^4$ mit Koordinaten $\vec{a} = (2, -3, \sqrt{2}, 1)$ hat den Betrag

$$|\vec{a}| = \sqrt{2^2 + (-3)^2 + \sqrt{2}^2 + 1^2} = 4.$$

Der Vektor $\vec{e}_a = \frac{1}{4}\vec{a}$ ist der Einheitsvektor, der in die gleiche Richtung wie $\vec{a}$ zeigt (Richtungseinheit von $\vec{a}$).

c) Die beiden Vektoren $\vec{a} = (1, 2, 3, 4)$ und $\vec{b} = (4, 3, 2, 1)$ in $\mathbb{R}^4$ haben den Abstand

$$d(\vec{b}, \vec{a}) = |\vec{a} - \vec{b}| = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + (a_4 - b_4)^2}$$

$$= \sqrt{(-3)^2 + (-1)^2 + 1^2 + 3^2} = \sqrt{20}. \quad \blacksquare$$

Die Begriffe „Länge“ und „Abstand“ sind in $\mathbb{R}^n$ geometrisch nachvollziehbar, in einem Funktionenraum kann man den Abstand dagegen nutzen, um zu beschreiben, wie gut eine Funktion f durch eine Funktion g angenähert wird (siehe beispielsweise die Berechnung eines Fourier-Koeffizienten auf Seite 620).

Beispiel 3.34

In $C[-1, 1]$ mit dem Skalarprodukt $\bullet$ aus (3.21) betrachten wir $f(x) = 1$ und $g(x) = x$. Dann ist

$$\begin{aligned} [d(f, g)]^2 &= (f - g) \bullet (f - g) = \int_{-1}^1 (f(x) - g(x))^2 dx \\ &= \int_{-1}^1 (1 - x)^2 dx = \left[-\frac{1}{3}(1 - x)^3 \right]_{-1}^1 = \frac{8}{3}. \end{aligned}$$

Die Funktionen haben also den Abstand $\sqrt{\frac{8}{3}}$. $\blacksquare$

Satz 3.16 (Eigenschaften des Skalarprodukts und des Betrags)

Für Vektoren $\vec{x}$ und $\vec{y}$ eines Euklid'schen Raums V gilt

- a) $|\vec{x}| = 0 \iff \vec{x} = \vec{0}$,
- b) $|\lambda \cdot \vec{x}| = |\lambda| \cdot |\vec{x}|$ für alle $\lambda \in \mathbb{R}$,
- c) **Cauchy-Schwarz'sche Ungleichung:** $|\vec{x} \cdot \vec{y}| \leq |\vec{x}| \cdot |\vec{y}|$,
- d) **Dreiecksungleichung:** $|\vec{x} + \vec{y}| \leq |\vec{x}| + |\vec{y}|$.

Beweis: a) Dies ist nichts anderes als (3.20) zusammen mit der positiven Definitheit des Skalarprodukts, siehe Definition 3.14 a).

b) $|\lambda \cdot \vec{x}| = \sqrt{(\lambda \cdot \vec{x}) \cdot (\lambda \cdot \vec{x})} = \sqrt{\lambda^2 (\vec{x} \cdot \vec{x})} = |\lambda| \cdot |\vec{x}|$.

c) Wir zeigen die Ungleichung für $\vec{y} \neq \vec{0}$, d. h. $|\vec{y}| > 0$. Für $\vec{y} = \vec{0}$ ist die Ungleichung $0 \leq 0$ offensichtlich erfüllt (vgl. a)). Für jedes $\lambda \in \mathbb{R}$ gilt:

$$0 \leq (\vec{x} - \lambda \vec{y}) \cdot (\vec{x} - \lambda \vec{y}) = \vec{x} \cdot \vec{x} - \lambda(\vec{y} \cdot \vec{x}) - \lambda(\vec{x} \cdot \vec{y}) + \lambda^2(\vec{y} \cdot \vec{y}) = |\vec{x}|^2 - 2\lambda(\vec{x} \cdot \vec{y}) + \lambda^2|\vec{y}|^2.$$

Wählen wir nun speziell $\lambda := \frac{\vec{x} \cdot \vec{y}}{|\vec{y}|^2}$, so erhalten wir

$$0 \leq |\vec{x}|^2 - 2 \frac{(\vec{x} \cdot \vec{y})^2}{|\vec{y}|^2} + \frac{(\vec{x} \cdot \vec{y})^2}{|\vec{y}|^2} = |\vec{x}|^2 - \frac{(\vec{x} \cdot \vec{y})^2}{|\vec{y}|^2}.$$

Damit ist

$$|\vec{x} \cdot \vec{y}|^2 \leq |\vec{x}|^2 |\vec{y}|^2,$$

und die Ungleichung folgt durch Ziehen der Wurzel auf beiden Seiten, da die Beträge nicht-negativ sind.

d) Mit der Definition des Skalarprodukts und der Cauchy-Schwarz'schen Ungleichung folgt:

$$\begin{aligned} |\vec{x} + \vec{y}|^2 &= (\vec{x} + \vec{y}) \cdot (\vec{x} + \vec{y}) = \vec{x} \cdot \vec{x} + 2\vec{x} \cdot \vec{y} + \vec{y} \cdot \vec{y} = |\vec{x}|^2 + 2\vec{x} \cdot \vec{y} + |\vec{y}|^2 \\ &\leq |\vec{x}|^2 + 2|\vec{x}||\vec{y}| + |\vec{y}|^2 = (|\vec{x}| + |\vec{y}|)^2, \end{aligned}$$

und Ziehen der Wurzel ergibt $|\vec{x} + \vec{y}| \leq |\vec{x}| + |\vec{y}|$. □

Der Satz gilt auch für komplexe Vektorräume mit einem komplexen Skalarprodukt. Lediglich im Beweis zu b) und c) muss man die komplexe Konjugation beachten.

3.3.4 Orthogonalität, Orthogonal- und Orthonormalsysteme

Rechte Winkel sind in Wohnungen praktisch, da man gut Möbel platzieren kann. Auch beim Rechnen mit Vektoren erleichtern sie die Arbeit erheblich. Besonders angenehm sind Basen, deren Vektoren orthogonal (senkrecht) zueinander stehen. Noch schöner wird es, wenn auch die Längen der Basisvektoren zu eins normiert sind. Man spricht dann von einer orthonormalen Basis.

3.3.4.1 Winkel und Orthogonalität

Mit Hilfe des Skalarprodukts lässt sich in einem Euklid'schen Raum ein (abstrakter) Winkelbegriff einführen. Aus der Cauchy-Schwarz'schen Ungleichung $|\vec{a} \cdot \vec{b}| \leq |\vec{a}| \cdot |\vec{b}|$ folgt für alle $\vec{a}, \vec{b} \in V$

$$-1 \leq \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|} \leq 1.$$

In Anlehnung an das Skalarprodukt aus Definition 3.4, das wir elementargeometrisch für Vektoren in der Ebene und im Raum eingeführt haben, wird die Zahl $\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$ als Kosinus eines Winkel $\varphi \in [0, \pi]$ interpretiert.

Definition 3.17 (Winkel und Orthogonalität)

Es sei $(V, +; \mathbb{R}, \cdot)$ ein Euklid'scher Raum mit Skalarprodukt „ $\cdot$ “.

- Für $\vec{a}, \vec{b} \neq \vec{0}$ heißt die Zahl $\varphi = \varphi(\vec{a}, \vec{b}) \in [0, \pi]$ mit

$$\cos \varphi = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$$

der **Winkel** zwischen $\vec{a}$ und $\vec{b}$.

- Zwei Vektoren $\vec{a}, \vec{b} \in V$ heißen **orthogonal** ($\vec{a} \perp \vec{b}$) genau dann, wenn $\vec{a} \cdot \vec{b} = 0$. In diesem Fall ist $\varphi(\vec{a}, \vec{b}) = \frac{\pi}{2}$.
- Entsprechend sind ein Vektor $\vec{a} \in V$ und eine Menge $U \subset V$ **orthogonal** ($\vec{a} \perp U$) genau dann, wenn $\vec{a} \cdot \vec{b} = 0$ für alle $\vec{b} \in U$ gilt.

Die hier definierte Orthogonalität entspricht genau der Definition auf Seite 436 für das Standardskalarprodukt in $\mathbb{R}^2$ oder $\mathbb{R}^3$. Jetzt können wir aber auch von orthogonalen Vektoren in höheren Dimensionen und abstrakten Räumen sprechen. Man beachte, dass bei Verwendung eines komplexen Skalarprodukts in einem komplexen Vektorraum die Zahl $\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$ komplex ist. Damit können wir den Arkuskosinus in der Regel nicht berechnen. Ist der Wert aber null, so nennen wir auch in diesem Fall die Vektoren orthogonal.

Beispiel 3.35

Die Vektoren $\vec{a} = (3, -1, 1, 5)$ und $\vec{b} = (-2, 2, 3, 1)$ sind orthogonal, denn es ist

$$\vec{a} \cdot \vec{b} = 3 \cdot (-2) + (-1) \cdot 2 + 1 \cdot 3 + 5 \cdot 1 = 0.$$

Weiter gilt $|\vec{a}| = \sqrt{3^2 + (-1)^2 + 1^2 + 5^2} = 6$ und $|\vec{b}| = \sqrt{(-2)^2 + 2^2 + 3^2 + 1^2} = \sqrt{18}$. ■

Beispiel 3.36 (Suchmaschinen)

Suchmaschinen im Internet nutzen das Standardskalarprodukt in $\mathbb{R}^n$, um die Ähnlichkeit einer Anfrage mit einer Internetseite zu bewerten. Dazu zählt man alle relevanten Wörter der Anfrage und schreibt die Anzahlen in 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. Dies ist der 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. ■

Beispiel 3.37

In $C[-1, 1]$ mit dem Skalarprodukt $\bullet$ aus (3.21) betrachten wir wieder $f(x) = 1$ und $g(x) = x$. Dann ist

$$f \bullet g = \int_{-1}^1 1 \cdot x \, dx = \left[\frac{1}{2} x^2 \right]_{-1}^1 = 0.$$

Die Funktionen sind orthogonal in diesem Vektorraum. Das heißt aber **nicht**, dass die Graphen im Schnittpunkt senkrecht zueinander stehen. ■

Sind $\vec{a}$ und $\vec{b}$ zueinander orthogonale Vektoren, so gilt wegen $\vec{a} \cdot \vec{b} = 0$:

$$|\vec{a} + \vec{b}|^2 = (\vec{a} + \vec{b}) \cdot (\vec{a} + \vec{b}) = |\vec{a}|^2 + 2\vec{a} \cdot \vec{b} + |\vec{b}|^2 = |\vec{a}|^2 + |\vec{b}|^2.$$

Damit haben wir gezeigt:

Satz 3.17 (Pythagoras in allgemeinen Euklid'schen Räumen)

Es seien $\vec{a}$ und $\vec{b}$ zueinander orthogonale Vektoren in einem Euklid'schen Raum. Dann gilt

$$|\vec{a} + \vec{b}|^2 = |\vec{a}|^2 + |\vec{b}|^2.$$

Insbesondere haben wir damit den Satz von Pythagoras in $\mathbb{R}^n$ für beliebiges n zur Verfügung.

3.3.4.2 Orthogonal- und Orthonormalsysteme

Die Vektoren der Standardbasis des $\mathbb{R}^3$, nämlich $\vec{e}_1 = (1, 0, 0)$, $\vec{e}_2 = (0, 1, 0)$, $\vec{e}_3 = (0, 0, 1)$, stehen paarweise senkrecht aufeinander und bilden das Grundgerüst für ein kartesisches Koordinatensystem. Dieser Begriff wird nun allgemeiner gefasst:

Definition 3.18 (Orthogonalsystem und Orthonormalsystem)

- Sind die Vektoren einer Menge $U \subset V$ paarweise orthogonal und ungleich $\vec{0}$, so bezeichnet man U als **Orthogonalsystem**.
- Ein Orthogonalsystem U aus lauter Einheitsvektoren (d.h. Vektoren mit Betrag eins) heißt **Orthonormalsystem**.
- Ein Orthonormalsystem U , das zugleich Basis von V ist, heißt **Orthonormalbasis**.

Für ein Orthonormalsystem $U = \{\vec{c}_1, \dots, \vec{c}_m\}$ gilt

$$\vec{c}_i \cdot \vec{c}_k = \delta_{i,k} := \begin{cases} 1, & \text{falls } i = k, \\ 0, & \text{falls } i \neq k, \end{cases} \quad i, k = 1, \dots, m. \quad (3.24)$$

Das **Kronecker-Delta** $\delta_{i,k}$ wird häufig als abkürzende Schreibweise für eine Fallunterscheidung verwendet, bei der bei nicht übereinstimmenden Indizes der Wert Null und sonst die Zahl Eins verlangt wird.

Beispiel 3.38

- a) Die Vektoren $\vec{a}$ und $\vec{b}$ aus Beispiel 3.35 bilden ein Orthogonalsystem U in $\mathbb{R}^4$ bezüglich des Standardskalarprodukts. Da die Beträge nicht 1 sind, ist U kein Orthonormalsystem. Ein Orthonormalsystem erhält man, indem man die beiden Vektoren normiert:

$$\vec{e}_a = \frac{1}{6} \cdot \vec{a}, \quad \vec{e}_b = \frac{1}{\sqrt{18}} \cdot \vec{b}.$$

Es handelt sich nicht um eine Orthonormalbasis von $\mathbb{R}^4$, da eine Basis vier Vektoren enthält.

- b) Die Standardbasis $U_S = \{\vec{e}_1, \dots, \vec{e}_n\} \subset \mathbb{R}^n$ ist eine Orthonormalbasis von $\mathbb{R}^n$ für das Standardskalarprodukt. Jede Teilmenge von U_S ist ein Orthonormalsystem in $\mathbb{R}^n$. ■

3.3.4.3 Euklid'sche Räume endlicher Dimension

Wir betrachten im Folgenden Euklid'sche Räume V endlicher Dimension n . Vektorräume wie die Polynome beliebigen Grades schließen wir damit aus. Mit einer Orthonormalbasis gestaltet sich dann vieles relativ einfach. Beispielsweise erhalten wir die Komponenten eines Vektors $\vec{a} \in V$ bezüglich einer Orthonormalbasis unmittelbar als das Skalarprodukt von $\vec{a}$ mit den Basisvektoren. Daneben lässt sich die Berechnung des Skalarprodukts auf das Standardskalarprodukt in $\mathbb{R}^n$ zurückführen, die Räume verhalten sich wie der $\mathbb{R}^n$.

Satz 3.18 (Basisdarstellung, Skalarprodukt und Satz von Pythagoras)

Es sei $U = \{\vec{c}_1, \dots, \vec{c}_n\}$ eine Orthonormalbasis in V , d. h. $\vec{c}_i \cdot \vec{c}_k = \delta_{i,k}$, $i, k = 1, \dots, n$.

- a) Für $\vec{a} \in V$ gilt die Komponentendarstellung

$$\vec{a} = \sum_{i=1}^n a_i \cdot \vec{c}_i \quad \text{mit} \quad a_k = \vec{a} \cdot \vec{c}_k, \quad k = 1, \dots, n.$$

Analog zur Standardbasis der Ebene oder des dreidimensionalen Anschauungsraums nennt man die Summanden $a_i \cdot \vec{c}_i$ die **Komponenten** von $\vec{a}$ zur Basis U , die Faktoren a_i heißen **Koordinaten**.

- b) Das Skalarprodukt von zwei Vektoren $\vec{a}$ und $\vec{b}$ in V kann über das Standardskalarprodukt der Koordinaten in $\mathbb{R}^n$ berechnet werden:

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^n a_i \cdot b_i.$$

- c) Der Betrag eines Vektors $\vec{a} \in V$ ist gleich dem Betrag der Koordinaten im Raum $\mathbb{R}^n$ mit dem Standardskalarprodukt:

$$|\vec{a}|^2 = \sum_{i=1}^n a_i^2.$$

Beweis: a) Da U insbesondere ein Erzeugendensystem ist, hat $\vec{a}$ eine Darstellung als Linearkombination $\vec{a} = \sum_{i=1}^n a_i \cdot \vec{c}_i$, wobei wir noch die Koeffizienten a_i ausrechnen müssen. Das Skalarprodukt von $\vec{a}$ mit $\vec{c}_k$ lautet $\vec{a} \cdot \vec{c}_k = \sum_{i=1}^n a_i \vec{c}_i \cdot \vec{c}_k$, und mit (3.24) folgt $\vec{a} \cdot \vec{c}_k = a_k$.

b) Für das Skalarprodukt von $\vec{a} = \sum_{i=1}^n a_i \cdot \vec{c}_i$ mit $\vec{b} = \sum_{k=1}^n b_k \cdot \vec{c}_k$ erhält man

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^n a_i \cdot \left(\vec{c}_i \cdot \sum_{k=1}^n b_k \cdot \vec{c}_k \right) = \sum_{i=1}^n a_i \cdot \sum_{k=1}^n b_k \underbrace{\vec{c}_i \cdot \vec{c}_k}_{=\delta_{i,k}} = \sum_{i=1}^n a_i \cdot b_i.$$

c) erhält man aus b), wenn man $\vec{b} = \vec{a}$ setzt. $\square$

Satz 3.19 (Lineare Unabhängigkeit und Orthonormalsystem)

*Bildet die Menge $\{\vec{a}_1, \dots, \vec{a}_n\}$ ein **Orthonormalsystem**, dann ist $\{\vec{a}_1, \dots, \vec{a}_n\}$ linear unabhängig.*

In einem Euklid'schen Raum V der Dimension n hat ein Orthonormalsystem U also höchstens n Vektoren.

Beweis: Angenommen, es ist $\vec{0} = \sum_{i=1}^n \lambda_i \vec{a}_i$. Dann folgt durch Bildung des Skalarprodukts mit $\vec{a}_k$:

$$0 = \vec{0} \cdot \vec{a}_k = \sum_{i=1}^n \lambda_i \underbrace{\vec{a}_i \cdot \vec{a}_k}_{\delta_{i,k}} = \lambda_k, \quad k = 1, \dots, n.$$

Damit kann $\vec{0}$ nur trivial linear kombiniert werden, und die Menge $\{\vec{a}_1, \dots, \vec{a}_n\}$ ist linear unabhängig. $\square$

Ein Orthonormalsystem U ist genau dann eine Basis von V , wenn es n Vektoren enthält.

Beispiel 3.39

Gegeben sei das Orthonormalsystem U in $\mathbb{R}^2$ aus den Vektoren

$$\vec{c}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \vec{c}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Wir bestimmen die Komponenten des Vektors $\vec{a} = (3, 4)^\top$ bezüglich U . Für die Koordinaten erhält man

$$\vec{a} \cdot \vec{c}_1 = \frac{1}{\sqrt{5}}(3 \cdot 1 + 4 \cdot 2) = \frac{11}{\sqrt{5}}, \quad \vec{a} \cdot \vec{c}_2 = \frac{1}{\sqrt{5}}(3 \cdot 2 + 4 \cdot (-1)) = \frac{2}{\sqrt{5}}$$

und damit die Komponentendarstellung $\vec{a} = \frac{11}{\sqrt{5}} \cdot \vec{c}_1 + \frac{2}{\sqrt{5}} \cdot \vec{c}_2$. $\blacksquare$

3.3.4.4 Gram-Schmidt'sches Orthonormierungsverfahren

In einem Euklid'schen Raum V kann aus jeder linear unabhängigen Menge $U = \{\vec{a}_1, \dots, \vec{a}_m\}$ ein Orthonormalsystem machen, das den gleichen Unterraum wie die Vektoren aus U erzeugt. Bei einem endlich-dimensionalen Raum lässt sich so insbesondere eine Orthonormalbasis gewinnen. Das **Gram-Schmidt'sche Orthonormierungsverfahren** liefert dieses Orthonormalsystem $U_o = \{\vec{c}_1, \dots, \vec{c}_m\}$ schrittweise:

- Da die Vektoren aus U linear unabhängig sind, sind sie vom Nullvektor verschieden und haben einen Betrag größer null. Damit können wir den ersten Vektor $\vec{a}_1$ zu einem Einheitsvektor normieren:

$$\vec{c}_1 := \frac{\vec{a}_1}{|\vec{a}_1|}.$$

- Der zweite Vektor $\vec{a}_2$ lässt sich als Summe eines Vektors, der senkrecht zum ersten Vektor steht (Lot), und eines Vektors, der parallel zum ersten Vektor liegt (Projektion), schreiben. Das geschieht genau so, wie es in Kapitel 3.1.2.4 für den Anschauungsraum beschrieben ist. Den senkrechten Anteil erhalten wir über

$$\vec{d}_2 := \vec{a}_2 - (\vec{a}_2 \cdot \vec{c}_1) \vec{c}_1,$$

denn $\vec{d}_2 \cdot \vec{c}_1 = \vec{a}_2 \cdot \vec{c}_1 - (\vec{a}_2 \cdot \vec{c}_1) \underbrace{\vec{c}_1 \cdot \vec{c}_1}_{=1} = 0$. Hier benötigt man, dass $\vec{c}_1$ normiert

ist. Da das auch in den folgenden Schritten wichtig ist und $\vec{d}_2$ im Allgemeinen noch nicht die Länge eins hat, wird auch dieser Vektor normiert:

$$\vec{c}_2 := \frac{\vec{d}_2}{|\vec{d}_2|}.$$

- Für die nächsten Schritte verallgemeinern wir die Vorgehensweise des vorherigen Schritts. Angenommen, $\vec{c}_1, \dots, \vec{c}_l$ liegen bereits vor, dann wird der nächste Vektor so konstruiert, dass er orthogonal zu allen diesen Vektoren steht:

$$\vec{d}_{l+1} := \vec{a}_{l+1} - \sum_{k=1}^l (\vec{a}_{l+1} \cdot \vec{c}_k) \vec{c}_k.$$

Wie für $\vec{d}_2$ rechnet man nach, dass $\vec{d}_{l+1}$ tatsächlich senkrecht zu den zuvor konstruierten Vektoren steht und muss nun nur noch normieren:

$$\vec{c}_{l+1} := \frac{\vec{d}_{l+1}}{|\vec{d}_{l+1}|}, \quad l = 1, \dots, m-1.$$

Beispiel 3.40

- a) Als erstes einfaches Beispiel für das Gram-Schmidt'sche Verfahren betrachten wir die Vektoren $\vec{a}_1 = (1, 0)^\top$ und $\vec{a}_2 = (1, 1)^\top$.

Im ersten Schritt ist lediglich $\vec{a}_1$ zu normieren:

$$\vec{c}_1 = \frac{\vec{a}_1}{|\vec{a}_1|} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Im zweiten Schritt betrachten wir

$$\vec{d}_2 := \vec{a}_2 - (\vec{a}_2 \cdot \vec{c}_1) \vec{c}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - (1 \cdot 1 + 1 \cdot 0) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Wegen $|\vec{d}_2| = 1$ erhalten wir $\vec{c}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

b) Wir wollen das System der linear unabhängigen Vektoren in $\mathbb{R}^4$

$$\vec{a}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{a}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{a}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

nach dem Verfahren von Gram-Schmidt orthonormieren. Zunächst wird der erste Vektor $\vec{d}_1 := \vec{a}_1$ normiert:

$$\vec{c}_1 = \frac{\vec{d}_1}{|\vec{d}_1|} = \frac{1}{\sqrt{2}}(1, 0, 1, 0)^\top.$$

Ein dazu orthogonaler Vektor ist

$$\vec{d}_2 = \vec{a}_2 - (\vec{a}_2 \cdot \vec{c}_1) \vec{c}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \\ -1 \\ 0 \end{pmatrix}.$$

Der entsprechende Einheitsvektor lautet

$$\vec{c}_2 = \frac{\vec{d}_2}{|\vec{d}_2|} = \frac{1}{\sqrt{6}}(1, 2, -1, 0)^\top.$$

Der folgende Vektor ist zu $\vec{c}_1$ und zu $\vec{c}_2$ orthogonal:

$$\begin{aligned} \vec{d}_3 &= \vec{a}_3 - (\vec{a}_3 \cdot \vec{c}_1) \vec{c}_1 - (\vec{a}_3 \cdot \vec{c}_2) \vec{c}_2 \\ &= \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} - \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} - (-1) \cdot \frac{1}{\sqrt{6}} \cdot \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 2 \\ -1 \\ 0 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} -1 \\ 1 \\ 1 \\ 3 \end{pmatrix}. \end{aligned}$$

Die Normierung ergibt

$$\vec{c}_3 = \frac{\vec{d}_3}{|\vec{d}_3|} = \frac{1}{\sqrt{12}}(-1, 1, 1, 3)^\top.$$

Die Menge $U_o = \{\vec{c}_1, \vec{c}_2, \vec{c}_3\}$ ist das gesuchte Orthonormalsystem. ■

3.3.4.5 Orthogonale Projektion

Die Zerlegung eines Vektors $\vec{b}$ in $\mathbb{R}^3$ als Summe $\vec{b} = \vec{u} + \vec{v}$ mit einem Vektor $\vec{u}$, der parallel zu einem gegebenen Vektor $\vec{a}$ liegt (Projektion auf $\vec{a}$), und einem Vektor $\vec{v}$, der senkrecht zu $\vec{a}$ steht (Lot), ist nach Satz 3.5 für das Standardskalarprodukt in $\mathbb{R}^3$ gegeben durch

$$\vec{u} = \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2} \vec{a} \quad \text{und} \quad \vec{v} = \vec{b} - \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2} \vec{a}.$$

Diesen Sachverhalt haben wir bei der Herleitung des Gram-Schmidt'schen Verfahrens bereits für allgemeine Euklid'sche Räume ausgenutzt. Außerdem haben wir dort ab dem dritten Schritt Vektoren berechnet, die senkrecht auf mehreren Vektoren stehen. Das lässt sich nun verallgemeinern, indem wir Vektoren auf Unterräume projizieren (vgl. Abbildung 3.27).

Definition 3.19 (Orthogonale Projektion)

Es sei V ein Euklid'scher Raum, und U sei ein endlich-dimensionaler Unterraum von V . Für den Vektor $\vec{a} \in V$ definieren wir die **orthogonale Projektion** von $\vec{a}$ auf U als den Vektor $\vec{p}(\vec{a}|U)$ mit

$$|\vec{p}(\vec{a}|U) - \vec{a}| \leq |\vec{q} - \vec{a}| \quad \text{für alle } \vec{q} \in U.$$

Diese Definition ist so noch nicht in Ordnung, denn es ist zunächst nicht klar, ob es tatsächlich ein eindeutiges Minimum $\vec{p}(\vec{a}|U)$ gibt. Um dies zu finden, benutzen wir eine orthogonale Basis $\{\vec{c}_1, \dots, \vec{c}_m\}$ des Unterraums U , die man immer aus einer beliebigen Basis von U mittels des Verfahrens von Gram-Schmidt berechnen kann.

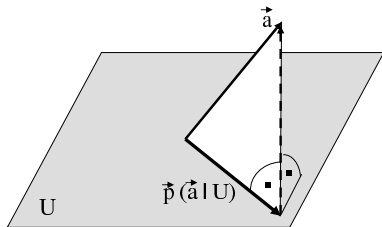


Abb. 3.27: Orthogonale Projektion

Satz 3.20 (Projektionsdarstellung)

Es sei V ein Euklid'scher Raum mit einem Unterraum U , der von einem Orthonormalsystem $\vec{c}_1, \dots, \vec{c}_m$ erzeugt wird. Die orthogonale Projektion eines Vektors $\vec{a} \in V$ auf den Unterraum U ist gegeben durch

$$\vec{p}(\vec{a}|U) = \sum_{i=1}^m (\vec{a} \cdot \vec{c}_i) \vec{c}_i.$$

Der **Lotvektor** $\vec{l} := \vec{a} - \vec{p}(\vec{a}|U)$ steht senkrecht auf allen Elementen von U (d. h. $\vec{l} \perp U$), daher der Name „orthogonale Projektion“.

In jedem Schritt des Gram-Schmidt-Verfahrens wird eine Projektion auf den Unterraum, der von den bereits berechneten Vektoren aufgespannt wird, durchgeführt.

Beweis: Im Satz ist der Kandidat für das in der Definition geforderte Minimum angegeben:

$$\vec{p} := \vec{p}(\vec{a}|U) = \sum_{i=1}^m (\vec{a} \cdot \vec{c}_i) \vec{c}_i \in U.$$

Analog zu den Überlegungen beim Gram-Schmidt'schen Orthogonalisierungsverfahren steht $\vec{l} = \vec{a} - \vec{p}$ senkrecht auf jedem Element von U : Wegen $\vec{p} \cdot \vec{c}_k = \sum_{i=1}^m (\vec{a} \cdot \vec{c}_i) (\vec{c}_i \cdot \vec{c}_k) = \vec{a} \cdot \vec{c}_k$ folgt $(\vec{a} - \vec{p}) \cdot \vec{c}_k = 0$ für $k = 1, \dots, m$. Weiter gilt $(\vec{a} - \vec{p}) \cdot \vec{q} = 0$ für einen beliebigen Vektor $\vec{q} \in U$, denn $\vec{q}$ lässt sich als Linearkombination aus $\{\vec{c}_1, \dots, \vec{c}_m\}$ darstellen.

Jetzt zeigen wir, dass $\vec{p}$ der eindeutige Vektor aus U ist, der einen minimalen Abstand zu $\vec{a}$ hat (so dass die vorangehende Definition sinnvoll ist). Dazu sei $\vec{q} \in U$ beliebig gewählt:

$$\begin{aligned} |\vec{a} - \vec{q}|^2 &= |(\vec{a} - \vec{p}) + (\vec{p} - \vec{q})|^2 = ((\vec{a} - \vec{p}) + (\vec{p} - \vec{q})) \cdot ((\vec{a} - \vec{p}) + (\vec{p} - \vec{q})) \\ &= |\vec{a} - \vec{p}|^2 + |\vec{p} - \vec{q}|^2 + 2(\vec{a} - \vec{p}) \cdot (\vec{p} - \vec{q}), \end{aligned}$$

wobei der letzte Summand wegen $\vec{p} - \vec{q} \in U$ und $\vec{a} - \vec{p} \perp U$ verschwindet. Damit ist

$$|\vec{a} - \vec{p}|^2 + \underbrace{|\vec{p} - \vec{q}|^2}_{\geq 0} = |\vec{a} - \vec{q}|^2,$$

und wir erhalten $|\vec{a} - \vec{p}| \leq |\vec{a} - \vec{q}|$, $\vec{p}$ ist minimal. Da für $\vec{p} \neq \vec{q}$ sogar $|\vec{a} - \vec{p}| < |\vec{a} - \vec{q}|$ gilt, haben wir auch die Eindeutigkeit des Minimums bei $\vec{p}$ gezeigt. $\square$

Beispiel 3.41

Wir betrachten die Vektoren

$$\vec{a} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad \vec{c}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \vec{c}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Die Komponenten

$$\vec{a}_1 = (\vec{a} \cdot \vec{c}_1) \vec{c}_1 = \frac{11}{5} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \vec{a}_2 = (\vec{a} \cdot \vec{c}_2) \vec{c}_2 = \frac{2}{5} \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

von $\vec{a}$ bezüglich der Orthonormalbasis $\{\vec{c}_1, \vec{c}_2\}$ sind die orthogonalen Projektionen von $\vec{a}$ auf die Richtungen $\vec{c}_1$ und $\vec{c}_2$ bzw. auf die von ihnen erzeugten Teilräume. ■

Als ein weiteres Beispiel zeigt Abbildung 3.28 die Projektion von farbigen Laserscanpunkten auf Fassaden eines 3-D-Stadtmodells des Landes NRW. Im Stadtmodell liegt jede Fassade in einer Ebene U . Da in der Realität aber Fassaden nie ganz eben sind, werden zur Texturierung auch Laserscanpunkte $\vec{a}$ verwendet, die maximal einen Meter vor oder hinter der Fassade liegen. Diese werden orthogonal als Vektor $\vec{p}(\vec{a}|U)$ in die Ebene projiziert, so dass die so gefundene Stelle eingefärbt werden kann.



Abb. 3.28: Wandtexturen, die durch Orthogonalprojektion von Laserscandaten erzeugt wurden.

3.3.4.6 Orthogonale Matrizen

In Kapitel 1.7.5 haben wir orthogonale Matrizen $\mathbf{A}$ über die Eigenschaft $\mathbf{A}^{-1} = \mathbf{A}^\top$ eingeführt. Damals konnten wir aber die Bezeichnung noch nicht verstehen. Jetzt haben wir das Standardskalarprodukt in $\mathbb{R}^n$ zur Verfügung und wissen, dass Vektoren genau dann senkrecht zueinander stehen, wenn ihr Skalarprodukt null ist.

Folgerung 3.2 (Eigenschaften einer orthogonalen Matrix)

Ist $\mathbf{A} \in \mathbb{R}^{n \times n}$ eine orthogonale Matrix, d. h. $\mathbf{A} \cdot \mathbf{A}^\top = \mathbf{A}^\top \cdot \mathbf{A} = \mathbf{E}$ (siehe Definition 1.55 auf Seite 175), so gilt bezüglich des Standardskalarprodukts in $\mathbb{R}^n$:

- Die Spaltenvektoren von $\mathbf{A}$ stehen senkrecht zueinander.
- Die Zeilenvektoren von $\mathbf{A}$ stehen senkrecht zueinander.
- Bei Multiplikation mit $\mathbf{A}$ bleibt der Betrag eines Vektors erhalten.
- Bei der Multiplikation mit orthogonalen Matrizen bleibt der Kosinus des Winkels zwischen Vektoren erhalten.

Verwendet man das komplexe Standardskalarprodukt (3.22) von Seite 486, so gelten die Punkte a), b) und c) nicht nur für reelle orthogonale Matrizen, sondern auch für komplexe unitäre Matrizen $\mathbf{A} \in \mathbb{C}^{n \times n}$ (siehe Kasten auf Seite 176). Dabei stehen komplexe Spalten- oder Zeilenvektoren paarweise senkrecht zueinander, wenn ihr komplexes Standardskalarprodukt null ist.

Beweis: a) Bildet man das Skalarprodukt zweier verschiedener Spalten $i \neq k$, so ergibt sich der Eintrag der i -ten Zeile und k -ten Spalte von $\mathbf{A}^\top \cdot \mathbf{A} = \mathbf{E}$, also 0.

b) Analog zu a) folgt die Orthogonalität der Zeilen aus $\mathbf{A} \cdot \mathbf{A}^\top = \mathbf{E}$.

c) Für $\vec{a} \in \mathbb{R}^n$ ist

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

d) Seien $\vec{a}, \vec{b} \in \mathbb{R}^n$ mit $\vec{a} \neq \vec{0} \neq \vec{b}$. Für das Skalarprodukt der Bildvektoren gilt:

$$(\mathbf{A} \cdot \vec{a}) \cdot (\mathbf{A} \cdot \vec{b}) = \vec{a}^\top \cdot \mathbf{A}^\top \cdot \mathbf{A} \cdot \vec{b} = \vec{a}^\top \cdot \mathbf{A}^{-1} \cdot \mathbf{A} \cdot \vec{b} = \vec{a}^\top \cdot \vec{b} = \vec{a} \cdot \vec{b}.$$

Für den Kosinus des Winkels φ zwischen den Vektoren $\vec{a}$ und $\vec{b}$ folgt mit c)

$$\cos \varphi = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} = \frac{(\mathbf{A}\vec{a}) \cdot (\mathbf{A}\vec{b})}{|\mathbf{A}\vec{a}| |\mathbf{A}\vec{b}|},$$

d.h., der Kosinus des Winkels zwischen den Vektoren $\vec{a}$, $\vec{b}$ und der Kosinus des Winkels zwischen $\mathbf{A}\vec{a}$ und $\mathbf{A}\vec{b}$ sind gleich. $\square$

Orthogonale Matrizen sind sogar genau die Matrizen, für die c) oder d) gilt:

Lemma 3.11 (Charakterisierung orthogonaler Matrizen)

Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$. Äquivalent sind:

- a) $\mathbf{A}$ ist orthogonal.
- b) $(\mathbf{A} \cdot \vec{x}) \cdot (\mathbf{A} \cdot \vec{y}) = \vec{x} \cdot \vec{y}$ für alle $\vec{x}, \vec{y} \in \mathbb{R}^n$.
- c) $|\mathbf{A} \cdot \vec{x}| = |\vec{x}|$ für alle $\vec{x} \in \mathbb{R}^n$.

Orthogonale Matrizen sind also genau die Matrizen, die bei der Multiplikation längenerhaltend sind.

Beweis: Nach Beweis der Folgerung 3.2 d) folgt b) aus a). Indem wir $\vec{y} = \vec{x}$ setzen, folgt aus b) die Aussage c). Das Lemma ist also bewiesen, wenn wir c) $\implies$ a) zeigen können. Seien dazu $\vec{a}_1, \dots, \vec{a}_n$ die n Zeilenvektoren der Matrix $\mathbf{A}$. Nach c) ist

$$|\vec{a}_k|^2 = |\mathbf{A} \cdot \vec{a}_k^\top|^2 = \sum_{i=1}^n [\vec{a}_i \cdot \vec{a}_k]^2 = |\vec{a}_k|^2 + \sum_{i=1, i \neq k}^n [\vec{a}_i \cdot \vec{a}_k]^2.$$

Damit muss jeder (nicht-negative) Summand der verbleibenden Summe null sein, die Zeilenvektoren von $\mathbf{A}$ stehen also senkrecht zueinander. $\mathbf{A}^\top \cdot \mathbf{A} = \mathbf{D} = \mathbf{D}^\top = \mathbf{A} \cdot \mathbf{A}^\top$ ist damit eine Diagonalmatrix. Wir müssen noch zeigen, dass die Hauptdiagonalelemente $d_k = 1$ sind. Dazu multiplizieren wir mit dem Standard-Einheitsvektor $\vec{e}_k$, der eine 1 an der k -ten Stelle hat:

$$d_k = \vec{e}_k^\top \mathbf{D} \vec{e}_k = \vec{e}_k^\top \mathbf{A}^\top \cdot \mathbf{A} \vec{e}_k = (\mathbf{A} \vec{e}_k) \cdot (\mathbf{A} \vec{e}_k) = |\mathbf{A} \cdot \vec{e}_k|^2 \stackrel{c)}{=} |\vec{e}_k|^2 = 1. \quad \square$$

Beispiel 3.42

Die Spaltenvektoren der Matrix $\mathbf{A}$ aus (1.48) stehen senkrecht aufeinander, da die Matrix orthogonal ist (siehe Seite 173). $\mathbf{A}$ beschreibt eine Drehung. Dreht man zwei Spaltenvektoren $\vec{a}$ und $\vec{b}$ um den gleichen Winkel durch Multiplikation mit $\mathbf{A}$, so bleibt der Winkel zwischen diesen Vektoren gleich. Allgemeiner beschreiben orthogonale Matrizen Drehungen oder Spiegelungen. $\blacksquare$

Folgerung 3.3 (Orthogonale Matrix als orthonormale Basis)

- a) Die Spalten und ebenso die Zeilen einer orthogonalen Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ bilden eine Orthonormalbasis des $\mathbb{R}^n$.
- b) Hat man umgekehrt eine Orthonormalbasis der $\mathbb{R}^n$, so erhält man eine orthogonale Matrix $\mathbf{A}$, indem man die Basisvektoren als Zeilen oder Spalten von $\mathbf{A}$ verwendet.

Beweis: a) Nach ihrer Definition ist die Matrix $\mathbf{A}$ invertierbar, die n Spalten sind damit linear unabhängig und bilden eine Basis des $\mathbb{R}^n$. Außerdem stehen die Spalten senkrecht zueinander (siehe Folgerung 3.2). Der Betrag der k -ten Spalte ist die Wurzel aus dem Eintrag an der Stelle (k, k) der Matrix $\mathbf{A}^\top \cdot \mathbf{A} = \mathbf{E}$ und ist damit 1.

b) Seien die Spalten von $\mathbf{A}$ eine orthonormale Basis. Dann ist $\mathbf{A}^\top \cdot \mathbf{A} = \mathbf{E}$, so dass $\mathbf{A}^\top = \mathbf{A}^{-1}$ und $\mathbf{A}$ orthogonal ist. $\square$

Zusammenfassend sind damit äquivalent:

- $\mathbf{A}$ ist orthogonal.
- $(\mathbf{A} \cdot \vec{x}) \cdot (\mathbf{A} \cdot \vec{y}) = \vec{x} \cdot \vec{y}$ für alle $\vec{x}, \vec{y} \in \mathbb{R}^n$.
- $|\mathbf{A} \cdot \vec{x}| = |\vec{x}|$ für alle $\vec{x} \in \mathbb{R}^n$.
- $\mathbf{A}^{-1} = \mathbf{A}^\top$.
- $\mathbf{A}^\top \cdot \mathbf{A} = \mathbf{E}$.
- $\mathbf{A} \cdot \mathbf{A}^\top = \mathbf{E}$.
- Die Spalten von $\mathbf{A}$ bilden eine Orthonormalbasis des $\mathbb{R}^n$.
- Die Zeilen von $\mathbf{A}$ bilden eine Orthonormalbasis des $\mathbb{R}^n$.

3.4 Lineare Abbildungen

Die **lineare Abbildung** ist in der Linearen Algebra das, was in der Analysis eine stetige Funktion ist. Das liegt daran, dass beide Arten von Abbildungen die jeweils untersuchten Eigenschaften erhalten. Stetige Abbildungen bilden Intervalle $[a, b]$ des Definitionsbereichs auf Intervalle $[c, d]$ des Wertebereichs ab (siehe Seite 282) und zeigen ein entsprechendes Verhalten für die in der Analysis wichtigen kleinen Umgebungen. Lineare Abbildungen erhalten die lineare Struktur, d. h., es spielt keine Rolle, ob man Vektoren zuerst addiert und mit Skalaren multipliziert und dann abbildet, oder ob man sie zuerst abbildet und dann verknüpft.

Geometrische Transformationen wie Drehungen, Spiegelungen und Streckungen werden über lineare Abbildungen dargestellt. So wie wir Drehungen bereits über ein Matrixprodukt ausgedrückt haben, lassen sich für endlich-dimensionale Vektorräume alle linearen Abbildungen über Matrizen schreiben. Hier erwartet uns nichts Neues, wir betrachten lediglich den Matrix-Begriff aus einer anderen Perspektive. Aber auch die Differenziation von Funktionen und die Bildung der Stammfunktion sind lineare Abbildungen. Diese lassen sich für unendlich-dimensionale Vektorräume von Funktionen dann aber nicht mehr mit Matrizen darstellen.

3.4.1 Lineare Abbildungen und Matrizen

Eine **lineare Funktion** $f : \mathbb{R} \rightarrow \mathbb{R}$ mit $f(x) = a \cdot x$ beschreibt eine Gerade durch den Koordinatenursprung mit der Steigung $a \in \mathbb{R}$. Für $x_1, x_2 \in \mathbb{R}$ gilt

$$\begin{aligned} f(x_1 + x_2) &= a \cdot (x_1 + x_2) = a \cdot x_1 + a \cdot x_2 = f(x_1) + f(x_2), \\ f(\lambda x_1) &= a \cdot (\lambda \cdot x_1) = \lambda \cdot (a \cdot x_1) = \lambda \cdot f(x_1), \quad \lambda \in \mathbb{R}. \end{aligned}$$

Diese beiden Eigenschaften nennt man **Linearität**. Abbildungen mit diesen Eigenschaften heißen **linear**. Hier beschäftigen wir uns mit linearen Abbildungen zwischen Vektorräumen:

Definition 3.20 (Lineare Abbildung)

*Es seien V und W Vektorräume über $\mathbb{R}$ (oder $\mathbb{C}$). Eine Abbildung $L : V \rightarrow W$, $\vec{x} \mapsto L(\vec{x})$ heißt **linear** genau dann, wenn für alle $\vec{x}, \vec{y} \in V$ und $\lambda \in \mathbb{R}$ (bzw. $\mathbb{C}$) gilt:*

- **Additivität:** $L(\vec{x} + \vec{y}) = L(\vec{x}) + L(\vec{y})$,
- **Homogenität:** $L(\lambda \cdot \vec{x}) = \lambda \cdot L(\vec{x})$.

Auf der linken Seite der Gleichungen zur Additivität und Homogenität wird die Addition und Skalarmultiplikation des Vektorraums V verwendet, während auf der rechten Seite die entsprechenden Verknüpfungen in W gemeint sind.

Im Folgenden formulieren wir für reelle Vektorräume, überall kann aber ohne Probleme $\mathbb{R}$ durch $\mathbb{C}$ ersetzt werden.

Lemma 3.12 (Eigenschaften linearer Abbildungen)

Gegeben sei eine lineare Abbildung $L: V \rightarrow W$:

a) Der Nullvektor $\vec{0}$ in V wird auf den Nullvektor $\vec{0}$ in W abgebildet:

$$L(\vec{0}) = L(0 \cdot \vec{x}) = 0 \cdot L(\vec{x}) = \vec{0}.$$

Umgekehrt kann es aber auch Vektoren $\vec{x} \neq \vec{0}$ mit $L(\vec{x}) = \vec{0}$ geben.

b) Das Bild einer Linearkombination von Vektoren ist gleich der Linearkombination der Bildvektoren:

$$L(\lambda_1 \cdot \vec{x}_1 + \lambda_2 \vec{x}_2) = \lambda_1 \cdot L(\vec{x}_1) + \lambda_2 \cdot L(\vec{x}_2), \quad \vec{x}_1, \vec{x}_2 \in V, \lambda_1, \lambda_2 \in \mathbb{R}.$$

Beispiel 3.43

a) $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $\vec{x} \mapsto L(\vec{x}) := \vec{x}$ heißt die **identische Abbildung**. Sie ist offensichtlich linear und entspricht der Multiplikation mit der Einheitsmatrix.

b) $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$, $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} \cos(\varphi) \cdot x_1 - \sin(\varphi) \cdot x_2 \\ \sin(\varphi) \cdot x_1 + \cos(\varphi) \cdot x_2 \end{pmatrix}.$

Die Abbildung ordnet jedem Vektor in $\mathbb{R}^2$ den um den Winkel φ gedrehten Vektor zu. Diese Drehung haben wir bereits auf Seite 160 über die (orthogonale) Matrix

$$\mathbf{A} = \mathbf{A}(\varphi) = \begin{bmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{bmatrix}$$

als Matrix-Vektorprodukt dargestellt: $L(\vec{x}) = \mathbf{A}(\varphi) \cdot \vec{x}$ für alle $\vec{x} \in \mathbb{R}^2$.

c) $L: \mathbb{R}^3 \rightarrow \mathbb{R}^3$, $\vec{x} \mapsto L(\vec{x}) := -\vec{x}$.

Dies ist eine Spiegelung im Ursprung. Man erhält sie auch durch Multiplikation mit der (orthogonalen) Matrix $(-1) \cdot \mathbf{E}$, wobei $\mathbf{E}$ die Einheitsmatrix ist.

d) $P: \mathbb{R}^3 \rightarrow \mathbb{R}^2$, $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$

Dies ist eine Projektion des dreidimensionalen Raums auf die Ebene. Sie kann ebenfalls mit einer Matrix beschrieben werden: $P(\vec{x}) = \mathbf{A} \cdot \vec{x}$ für alle $\vec{x} \in \mathbb{R}^3$ mit

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

e) $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^{\frac{1}{3}} \\ x_2^2 \end{pmatrix}$

ist eine Abbildung, die nicht linear ist. Zum Beispiel ist

$$T\left(2 \cdot \begin{pmatrix} 4 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \neq 2T\left(\begin{pmatrix} 4 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 2^{\frac{5}{3}} \\ 2 \end{pmatrix}.$$

- f) $L: \mathbb{R}^n \rightarrow \mathbb{R}^n, \vec{x} \mapsto L(\vec{x}) := \vec{0}$ heißt die **Nullabbildung** und ist ebenfalls linear. Man kann sie auch durch Multiplikation mit einer Matrix, deren Elemente alle 0 sind, beschreiben.
- g) Es sei $\mathbb{P}$ der (unendlich-dimensionale) Vektorraum der Polynome. Dann wertet die Abbildung $T: \mathbb{P} \rightarrow \mathbb{R}, p \mapsto p(1)$ das Polynom p an der Stelle 1 aus. Jedem Polynom $p \in \mathbb{P}$ wird so sein Funktionswert an der Stelle $x = 1$ zugeordnet. T ist linear. Lineare Abbildungen, die in die reellen (oder komplexen) Zahlen abbilden, nennt man **lineare Funktionale**. Man beachte, dass $(\mathbb{R}, +; \mathbb{R}, \cdot)$ und $(\mathbb{C}, +; \mathbb{C}, \cdot)$ Vektorräume sind.
- h) Die Menge der stetig differenzierbaren reellwertigen Funktionen auf einem Intervall $[a, b]$ bildet bezüglich der üblichen Addition von Funktionen und Multiplikation mit reellen Zahlen einen (unendlich-dimensionalen) Vektorraum V , den wir mit $V = C^1[a, b]$ bezeichnen. Ebenso bilden die auf $[a, b]$ stetigen Funktionen einen Vektorraum $W = C[a, b]$. Die Ableitung $\frac{d}{dx}$ ist eine lineare Abbildung von $V \rightarrow W$, denn es gilt

$$\begin{aligned} \frac{d}{dx}(f(x) + g(x)) &= \frac{d}{dx}f(x) + \frac{d}{dx}g(x), \\ \frac{d}{dx}(\lambda f(x)) &= \lambda \cdot \frac{d}{dx}f(x), \quad \lambda \in \mathbb{R}. \end{aligned} \quad \blacksquare$$

Wie das Beispiel zeigt, lassen sich lineare Abbildungen auf endlich-dimensionalen Vektorräumen durch Matrizen übersichtlich darstellen. Mit der Definition können wir unmittelbar nachrechnen:

Lemma 3.13 (Matrix-Vektor-Multiplikation als lineare Abbildung)

Sei $\mathbf{A}$ eine reelle $(m \times n)$ -Matrix. Die Abbildung $L: \mathbb{R}^n \rightarrow \mathbb{R}^m, \vec{x} \mapsto L(\vec{x}) := \mathbf{A} \cdot \vec{x}$ ist linear. $\vec{y} = \mathbf{A} \cdot \vec{x} \in \mathbb{R}^m$ ist das Bild bzw. der Funktionswert von $\vec{x}$ unter der Abbildung L .

Beispiel 3.44

a) Wir betrachten die Matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 \\ -1 & -6 & 0 \\ 0 & 5 & 2 \end{bmatrix}.$$

Die Matrix $\mathbf{A}$ beschreibt eine lineare Abbildung $\vec{y} = \mathbf{A} \cdot \vec{x}$ des $\mathbb{R}^3$ in sich. In Koordinatenform lautet die Abbildung

$$\begin{aligned} y_1 &= 3x_1 + 2x_3 \\ y_2 &= -x_1 - 6x_2 \\ y_3 &= 5x_2 + 2x_3. \end{aligned}$$

Der Vektor $(1, 2, 3)^\top$ wird dabei in den Vektor $(9, -13, 16)^\top$ überführt.

b) Die Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -4 & 5 & 6 \end{bmatrix}$$

definiert eine lineare Abbildung des Raums $\mathbb{R}^3$ in die Ebene $\mathbb{R}^2$. Speziell wird $(1, -1, 1)^\top$ auf $(2, -3)^\top$ abgebildet.

c) Eine Gerade in $\mathbb{R}^2$ oder $\mathbb{R}^3$ hat eine Punkt-Richtungsdarstellung $\vec{x} + \lambda \cdot \vec{v}$, $\lambda \in \mathbb{R}$. Als Bild der Geraden unter einer linearen Abbildung L entsteht die Menge

$$L(\vec{x} + \lambda \cdot \vec{v}) = L(\vec{x}) + \lambda \cdot L(\vec{v}), \quad \lambda \in \mathbb{R}$$

d. h., das Bild ist bei $L(\vec{v}) \neq \vec{0}$ wieder eine Gerade und im Fall $L(\vec{v}) = \vec{0}$ ein einzelner Punkt. ■



Abb. 3.29: Die Reihenfolge bei der Matrixmultiplikation spielt eine Rolle. Links ist die Startszene (Campus Süd der Hochschule Niederrhein) abgebildet, daneben ist das Resultat einer Drehung in der x - y -Ebene um 16 Grad mit anschließender Drehung in der x - z -Ebene um 44 Grad abgebildet. Zum rechten Bild führt die umgekehrte Reihenfolge der Drehungen.

Beispiel 3.45 (Vektorgrafik und homogene Koordinaten)

Wenn wir uns in einem Computerspiel oder in einer Simulation durch eine 3-D-Szene bewegen, dann müssen die Objekte je nach Position verschoben und je nach Blickwinkel gedreht werden. Dies geschieht durch die Multiplikation mit Matrizen. In Beispiel 3.43 haben wir bereits gesehen, wie Punkte in der Ebene durch Multiplikation mit einer Matrix gedreht werden können. Eine Verschiebung lässt sich so aber nicht darstellen.

Hier wird ein Trick angewendet: Eine weitere Koordinate, die stets 1 ist, wird eingeführt. Punkte im $\mathbb{R}^3$ werden dann mittels sogenannter **homogener Koordinaten** in $\mathbb{R}^4$ beschrieben. Der Punkt $(x, y, z) \in \mathbb{R}^3$ wird also als $(x, y, z, 1) \in \mathbb{R}^4$ geschrieben, so dass er mit 4×4 -Matrizen multipliziert werden kann. Eine Drehung um den Winkel φ in der x - y -Ebene lässt sich nun durch Multiplikation mit der Matrix

$$\mathbf{A} = \mathbf{A}(\varphi) = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) & 0 & 0 \\ \sin(\varphi) & \cos(\varphi) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

beschreiben. Entsprechend lassen sich Drehungen in anderen Ebenen formulieren. Eine Addition des Vektors $\vec{d} = (d_1, d_2, d_3)$ (Verschiebung im $\mathbb{R}^3$) erhalten wir durch Multiplikation mit

$$\mathbf{T} = \mathbf{T}(\vec{d}) = \begin{bmatrix} 0 & 0 & 0 & d_1 \\ 0 & 0 & 0 & d_2 \\ 0 & 0 & 0 & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Um diese Verschiebung auszudrücken, benötigen wir die Eins in der vierten Komponente der Vektoren.

Die Hintereinanderausführung mehrerer Drehungen und Verschiebungen entspricht der Multiplikation der zugehörigen Matrizen. Dass diese nicht kommutativ ist, sieht man in Abbildung 3.29, bei der einerseits zunächst in der x - y -Ebene und dann in der x - z -Ebene gedreht und andererseits in umgekehrter Reihenfolge gedreht wird. ■

Wir haben in Lemma 3.13 gesehen, dass die Multiplikation einer Matrix $\mathbf{A}$ mit einem Vektor $\vec{x}$ eine lineare Abbildung ist. Es stellt sich Frage, ob umgekehrt alle linearen Abbildungen $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ durch eine Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ mit $L(\vec{x}) = \mathbf{A} \cdot \vec{x}$ darstellbar sind. Das legen einige der vorangehenden Beispiele nahe.

Der folgende Satz erklärt sich direkt mit Lemma 3.12: Wenn wir mit L ein Element $\vec{x} \in V$ nach W abbilden, dann lässt sich $L(\vec{x})$ als Linearkombination von Bildern der Basisvektoren aus V darstellen.

Satz 3.21 (Darstellung als Matrix)

Zu jeder linearen Abbildung $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\vec{x} \mapsto L(\vec{x})$ gibt es genau eine $(m \times n)$ -Matrix $\mathbf{A}$ mit

$$L(\vec{x}) = \mathbf{A} \cdot \vec{x}, \quad \vec{x} \in \mathbb{R}^n.$$

Beweis: Mit den Standard-Einheitsvektoren $\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n$ in $\mathbb{R}^n$ lässt sich ein beliebiger Vektor $\vec{x} = (x_1, \dots, x_n)^\top$ als Linearkombination schreiben:

$$\vec{x} = x_1 \cdot \vec{e}_1 + x_2 \cdot \vec{e}_2 + \dots + x_n \cdot \vec{e}_n.$$

Das Bild $\vec{y}$ von $\vec{x}$ unter der linearen Abbildung L ist

$$\begin{aligned} \vec{y} &= L(\vec{x}) = x_1 \cdot L(\vec{e}_1) + x_2 \cdot L(\vec{e}_2) + \dots + x_n \cdot L(\vec{e}_n) \\ &= \underbrace{[L(\vec{e}_1), L(\vec{e}_2), \dots, L(\vec{e}_n)]}_{=: \mathbf{A} \in \mathbb{R}^{m \times n}} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{A} \cdot \vec{x}. \end{aligned} \quad \square$$

Die k -te Spalte der Abbildungsmatrix $\mathbf{A}$ ist das Bild des k -ten Standard-Einheitsvektors von $\mathbb{R}^n$.

Beispiel 3.46

a) Wir betrachten den $\mathbb{R}^2$ und $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ mit der linearen Abbildung

$$L\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} x + y \\ 3x - 2y \end{pmatrix}.$$

Über die Bilder der Standard-Einheitsvektoren

$$L(\vec{e}_1) = L\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \quad L(\vec{e}_2) = L\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

erhalten wir die Matrixdarstellung

$$L(\vec{x}) = \begin{bmatrix} 1 & 1 \\ 3 & -2 \end{bmatrix} \cdot \vec{x}.$$

b) Ebenso erhalten wir für die lineare Abbildung $L : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, die über

$$L\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} 3x_2 \\ x_1 - x_2 \\ x_1 + x_2 \end{pmatrix}$$

definiert ist, mit

$$L(\vec{e}_1) = L\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad \text{und} \quad L(\vec{e}_2) = L\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

die Matrixdarstellung

$$L(\vec{x}) = \begin{bmatrix} 0 & 3 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \cdot \vec{x}.$$

Beispielsweise wird $\vec{x} = (1, 2)^\top$ auf $L(\vec{x}) = (6, -1, 3)^\top$ abgebildet. ■

Allgemein können wir jede lineare Abbildung von einem endlich-dimensionalen Vektorraum in einen anderen endlich-dimensionalen Vektorraum über eine Matrix beschreiben, da man die Räume durch Übergang zu Koordinaten bezüglich einer Basis in Räume $\mathbb{R}^n$ und $\mathbb{R}^m$ überführen kann (siehe Kapitel 3.1.1.2 und Seite 491). Bei unendlich-dimensionalen Vektorräumen wie dem der Polynome beliebigen Grades ist die Darstellung über eine Matrix (mit endlich vielen Zeilen und Spalten) nicht möglich.

3.4.2 Summe, skalares Vielfaches und Verkettung linearer Abbildungen

Wie wir eben gesehen haben, hängen lineare Abbildungen und Matrizen eng zusammen. Die Summe, das skalare Vielfache und die Verkettung linearer Abbildungen lassen sich bei endlich-dimensionalen Vektorräumen über entsprechende Matrix-Operationen berechnen.

Für die Summe zweier linearer Abbildungen und das skalare Vielfache gilt:

Satz 3.22 (Summe und skalares Vielfaches von linearen Abbildungen)

Es seien $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ und $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ lineare Abbildungen mit zugehörigen $(m \times n)$ -Matrizen $\mathbf{A}$ und $\mathbf{B}$.

- a) Die Summe $L + S : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\vec{x} \mapsto L(\vec{x}) + S(\vec{x})$ ist eine lineare Abbildung, und die zugehörige $(m \times n)$ -Matrix lautet $\mathbf{A} + \mathbf{B}$.
- b) Mit $\lambda \in \mathbb{R}$ ist das skalare Vielfache $\lambda \cdot L : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\vec{x} \mapsto \lambda \cdot L(\vec{x})$ eine lineare Abbildung mit zugehöriger $(m \times n)$ -Matrix $\lambda \cdot \mathbf{A}$.

Beweis: Mit der Linearität von L und S folgt

$$\begin{aligned} (L + S)(\vec{x} + \vec{y}) &\stackrel{\text{Definition}}{=}^{L+S} L(\vec{x} + \vec{y}) + S(\vec{x} + \vec{y}) \\ &\stackrel{\text{Linearität}}{=} L(\vec{x}) + L(\vec{y}) + S(\vec{x}) + S(\vec{y}) \stackrel{\text{Definition}}{=}^{L+S} (L + S)(\vec{x}) + (L + S)(\vec{y}), \end{aligned}$$

und mit $\lambda \in \mathbb{R}$ folgt

$$\begin{aligned} (L + S)(\lambda \cdot \vec{x}) &\stackrel{\text{Definition}}{=}^{L+S} L(\lambda \cdot \vec{x}) + S(\lambda \cdot \vec{x}) \stackrel{\text{Linearität}}{=} \lambda \cdot (L(\vec{x}) + S(\vec{x})) \\ &\stackrel{\text{Definition}}{=}^{L+S} \lambda \cdot (L + S)(\vec{x}). \end{aligned}$$

Für die Abbildungsmatrix von $L + S$ gilt: $(L + S)(\vec{x}) = L(\vec{x}) + S(\vec{x}) = \mathbf{A} \cdot \vec{x} + \mathbf{B} \cdot \vec{x} = (\mathbf{A} + \mathbf{B}) \cdot \vec{x}$. Entsprechend erhält man für $\lambda \cdot L$ die Matrix $\lambda \mathbf{A}$. $\square$

Die Menge der linearen Abbildungen von $\mathbb{R}^n$ nach $\mathbb{R}^m$ ist eine bezüglich Addition und skalarer Multiplikation abgeschlossene Teilmenge des Vektorraums aller Abbildungen von $\mathbb{R}^n$ nach $\mathbb{R}^m$ und damit nach dem Unterraum-Kriterium (Satz 3.10 auf Seite 473) selbst ein Vektorraum.

Dem Matrixprodukt entspricht die Verkettung der zugehörigen Abbildungen:

Satz 3.23 (Verkettung linearer Abbildungen)

Es sei $L : \mathbb{R}^l \rightarrow \mathbb{R}^n$ eine lineare Abbildung mit $(n \times l)$ -Matrix $\mathbf{A}$ und $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ebenfalls linear mit $(m \times n)$ -Matrix $\mathbf{B}$. Dann ist die Verkettung oder Verschachtelung $S \circ L : \mathbb{R}^l \rightarrow \mathbb{R}^m$, $\vec{x} \mapsto S(L(\vec{x}))$, eine lineare Abbildung, und die zugehörige $(m \times l)$ -Matrix lautet $\mathbf{C} = \mathbf{B} \cdot \mathbf{A}$.

Beweis: Mit der Linearität von L und S folgt

$$\begin{aligned} (S \circ L)(\vec{x} + \vec{y}) &= S(L(\vec{x} + \vec{y})) = S(L(\vec{x}) + L(\vec{y})) \\ &= S(L(\vec{x})) + S(L(\vec{y})) = (S \circ L)(\vec{x}) + (S \circ L)(\vec{y}) \end{aligned}$$

und mit $\lambda \in \mathbb{R}$ entsprechend

$$(S \circ L)(\lambda \cdot \vec{x}) = S(L(\lambda \cdot \vec{x})) = S(\lambda \cdot L(\vec{x})) = \lambda \cdot S(L(\vec{x})) = \lambda \cdot (S \circ L)(\vec{x}).$$

Für die Abbildungsmatrix von $S \circ L$ gilt: $(S \circ L)(\vec{x}) = S(\mathbf{A} \cdot \vec{x}) = \mathbf{B} \cdot \mathbf{A} \cdot \vec{x}$. $\square$

Beispiel 3.47

Gegeben seien die linearen Abbildungen $S : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, $\vec{y} \mapsto S(\vec{y}) = \mathbf{B} \cdot \vec{y}$ und $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, $\vec{x} \mapsto L(\vec{x}) = \mathbf{A} \cdot \vec{x}$ mit Abbildungsmatrizen

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad \text{und} \quad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Wir berechnen die Abbildungsvorschrift der Verkettung $S \circ L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$:

$$\begin{aligned} \mathbf{A} \cdot \vec{x} &= \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + 2x_2 \\ 3x_1 + 4x_2 \end{pmatrix} \\ \mathbf{B} \cdot (\mathbf{A} \cdot \vec{x}) &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_1 + 2x_2 \\ 3x_1 + 4x_2 \end{pmatrix} = \begin{pmatrix} 3x_1 + 4x_2 \\ x_1 + 2x_2 \end{pmatrix} = \mathbf{C} \cdot \vec{x}, \end{aligned}$$

wobei wir die Matrix $\mathbf{C}$ durch Einsetzen der Standard-Einheitsvektoren erhalten:

$$\mathbf{B} \cdot (\mathbf{A} \cdot \vec{e}_1) = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad \mathbf{B} \cdot (\mathbf{A} \cdot \vec{e}_2) = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \quad \text{also} \quad \mathbf{C} = \begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix}.$$

Die Matrix $\mathbf{C}$ erhält man auch direkt durch $\mathbf{C} = \mathbf{B} \cdot \mathbf{A}$. $\blacksquare$

Beispiel 3.48

Obwohl die Ableitung eine lineare Operation ist, besitzt sie keine Matrixdarstellung. Da die zugrunde liegenden Räume der stetig differenzierbaren Funktionen bzw. der stetigen Funktionen nicht endlich-dimensional sind, ist dies auch nicht zu erwarten.

Wir betrachten nun den Vektorraum der Polynome vom Grad höchstens n

$$\mathbb{P}_n = \{p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 : a_n, \dots, a_0 \in \mathbb{R}\}.$$

Es gilt $\dim(\mathbb{P}_n) = n + 1$. Jedes Polynom $p \in \mathbb{P}_n$ kann durch den Vektor seiner Koeffizienten $\vec{p} = (a_n, a_{n-1}, \dots, a_1, a_0)^\top \in \mathbb{R}^{n+1}$ dargestellt werden. Es ist

$$p'(x) = na_n x^{n-1} + (n-1)a_{n-1} x^{n-2} + \cdots + 2a_2 x + a_1 \iff \vec{p}' = \begin{pmatrix} na_n \\ \vdots \\ 2a_2 \\ a_1 \end{pmatrix} \in \mathbb{R}^n.$$

Die Ableitung ist eine lineare Abbildung von $\mathbb{P}_n \rightarrow \mathbb{P}_{n-1}$ bzw. bezüglich der Koeffizienten eine lineare Abbildung von $\mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ und besitzt die Matrixdarstellung

$$\begin{pmatrix} na_n \\ \vdots \\ 2a_2 \\ a_1 \end{pmatrix} = \begin{bmatrix} n & 0 & & 0 \\ 0 & n-1 & & 0 \\ & 0 & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \cdot \begin{pmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_1 \\ a_0 \end{pmatrix}$$

Die Matrixdarstellung ist möglich, da $\mathbb{P}_n$ und $\mathbb{P}_{n-1}$ endlich-dimensionale Vektorräume sind. ■

3.4.3 Kern und Bild einer linearen Abbildung, Dimensionssatz**Definition 3.21 (Kern einer linearen Abbildung)**

Es sei $L : V \rightarrow W$ eine lineare Abbildung. Die Menge

$$\text{Kern}(L) := \{\vec{x} \in V : L(\vec{x}) = \vec{0}\}$$

heißt der **Kern** von L . Die Menge

$$\text{Bild}(L) := \{L(\vec{x}) \in W : \vec{x} \in V\}$$

ist das **Bild** von L .

Das Bild einer linearen Abbildung L ist ihre Wertemenge, wie wir sie aus Definition 1.4 kennen. Neu ist lediglich der Begriff Kern. Der Kern ist die Menge aller der Vektoren, die durch L auf den Nullvektor abgebildet werden. Ist L über eine Matrix $\mathbf{A}$ definiert, also $L(\vec{x}) = \mathbf{A} \cdot \vec{x}$, so ist der Kern genau die Lösungsmenge des homogenen linearen Gleichungssystems $\mathbf{A} \cdot \vec{x} = \vec{0}$. Allgemeiner gilt:

Satz 3.24 (Kern und Bild sind Vektorräume)

Der Kern der linearen Abbildung $L : V \rightarrow W$ ist ein Unterraum von V , und das Bild von L ist ein Unterraum von W .

Beweis: Seien $\vec{x}, \vec{y} \in \text{Kern}(L)$, d. h. $L(\vec{x}) = L(\vec{y}) = \vec{0}$. Dann folgt mit der Linearität von L : $\vec{0} = L(\vec{x}) + L(\vec{y}) = L(\vec{x} + \vec{y})$, d. h. $\vec{x} + \vec{y} \in \text{Kern}(L)$. Mit $\lambda \in \mathbb{R}$ folgt: $L(\lambda\vec{x}) = \lambda \cdot L(\vec{x}) = \lambda \cdot \vec{0} = \vec{0}$, d. h. $\lambda\vec{x} \in \text{Kern}(L)$. Damit ist aber der Kern bereits ein Unterraum (siehe Satz 3.10 auf Seite 473). Entsprechend zeigt man die Unterräumeigenschaft von $\text{Bild}(L)$. $\square$

Beispiel 3.49

a) Für die lineare Abbildung $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ mit

$$L\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} x + y \\ 3x - 2y \end{pmatrix}, \text{ d. h. } L(\vec{x}) = \mathbf{A} \cdot \vec{x} \text{ mit } \mathbf{A} = \begin{bmatrix} 1 & 1 \\ 3 & -2 \end{bmatrix},$$

sind $\text{Kern}(L)$ und $\text{Bild}(L)$ zu bestimmen. Für den Kern suchen wir die Vektoren $\vec{x} \in \mathbb{R}^2$ mit $L(\vec{x}) = \vec{0}$, d. h., wir suchen nach Lösungen des homogenen linearen Gleichungssystems $\mathbf{A} \cdot \vec{x} = \vec{0}$:

$$\left[\begin{array}{cc|c} 1 & 1 & 0 \\ 3 & -2 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \end{array} \right] \iff \vec{x} = \vec{0}.$$

Das System besitzt nur die Lösung $\vec{x} = \vec{0}$, und es ist $\text{Kern}(L) = \{\vec{0}\}$. Da $L(\vec{e}_1) = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ und $L(\vec{e}_2) = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ linear unabhängig sind, ist $\text{Bild}(L) = \mathbb{R}^2$.

b) Die Abbildung $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ sei erklärt durch

$$L\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} x + y \\ -x - y \end{pmatrix}, \text{ also } L(\vec{x}) = \mathbf{A} \cdot \vec{x} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \cdot \vec{x}.$$

Den Kern von L erhalten wir als Lösung des homogenen Gleichungssystems

$$\left[\begin{array}{cc|c} 1 & 1 & 0 \\ -1 & -1 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array} \right] \iff x + y = 0,$$

also als die Gerade $y = -x$. Diese geht durch den Ursprung, da es sich beim $\text{Kern}(L)$ um einen Unterraum des $\mathbb{R}^2$ handelt.

Die Koordinaten der Vektoren des Bildes unterscheiden sich offenbar nur durch ihr Vorzeichen, sie sind also von der Gestalt $(z, -z)^\top$ und befinden sich somit ebenfalls auf einer Geraden durch den Ursprung. ■

Wir betrachten den Zusammenhang zwischen der Dimension des Kerns und des Bildes der linearen Abbildung $L : V \rightarrow W$. Das Bild entspricht ja gerade dem Unterraum in W , den man durch Anwendung von L auf alle Vektoren aus V erhält. Beim Abbilden mittels L geht aber einiges „verloren“: Alle Vektoren, die im Kern von L liegen, werden auf den Nullvektor in W abgebildet. Sie leisten keinen Beitrag zur Dimension des Bildes.

Satz 3.25 (Dimensionssatz)

Es sei $L : V \rightarrow W$ eine lineare Abbildung und $\dim V$ sei endlich. Dann gilt

$$\dim V = \dim \text{Bild}(L) + \dim \text{Kern}(L).$$

Die lineare Abbildung im zuletzt betrachteten Beispiel bildet den $\mathbb{R}^2$ in den $\mathbb{R}^2$ ab. Die Dimension von $\text{Kern}(L)$ und $\text{Bild}(L)$ ist jeweils eins, was die Aussage des bestätigt.

Bemerkungen:

- Die Dimension des Kerns einer linearen Abbildung L wird auch als **Defekt** von L bezeichnet.
- Die Dimension des Bildes einer linearen Abbildung L wird als **Rang** von L bezeichnet.
- Der Dimensionssatz lautet damit:

$$\dim V = \text{Defekt}(L) + \text{Rang}(L).$$

Nach Satz 3.21 besitzt jede lineare Abbildung $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine Abbildungsmatrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ mit $L(\vec{x}) = \mathbf{A} \cdot \vec{x}$. Entsprechend werden die Begriffe Kern und Bild auch auf Matrizen bezogen, und in dieser Form werden wir später den Dimensionssatz beweisen. Allerdings haben sich bei Matrizen andere Begriffe eingebürgert, die wir kurz vorstellen.

Definition 3.22 (Nullraum, Spaltenraum und Zeilenraum einer Matrix)

- Der Kern der linearen Abbildung $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\vec{x} \mapsto \mathbf{A} \cdot \vec{x}$ mit $\mathbf{A} \in \mathbb{R}^{m \times n}$ heißt **Nullraum** von $\mathbf{A}$, kurz $N(\mathbf{A})$.
- Das Bild der linearen Abbildung $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\vec{x} \mapsto \mathbf{A} \cdot \vec{x}$ mit $\mathbf{A} \in \mathbb{R}^{m \times n}$ ist die lineare Hülle der Spaltenvektoren $\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n$ von $\mathbf{A}$. Dieser Raum ist der **Spaltenraum** von $\mathbf{A}$, kurz $S(\mathbf{A})$.

■ Als **Zeilenraum** von $\mathbf{A} \in \mathbb{R}^{m \times n}$ bezeichnet man die lineare Hülle der m Zeilenvektoren $\vec{z}_1, \vec{z}_2, \dots, \vec{z}_m$ von $\mathbf{A}$, kurz $Z(\mathbf{A})$.

Beispiel 3.50

a) Wir bestimmen den Nullraum der Matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$. Das zugehörige homogene lineare System $\mathbf{A} \cdot \vec{x} = \vec{0}$ lautet

$$\left[\begin{array}{cc|c} 1 & 2 & 0 \\ 4 & 8 & 0 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 0 & 0 \end{array} \right].$$

Setzen wir $x_2 = t$, so folgt $x_1 = -2t$, $t \in \mathbb{R}$, d. h. wir erhalten den eindimensionalen Nullraum $N(\mathbf{A}) = \{(-2t, t) \in \mathbb{R}^2 : t \in \mathbb{R}\}$. Der ebenfalls eindimensionale Spaltenraum wird vom Vektor $(1, 4)^\top$ erzeugt.

b) Den Nullraum der Matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$ erhalten wir durch Lösung von $\mathbf{A} \cdot \vec{x} = \vec{0}$.

$$\left[\begin{array}{cc|c} 1 & 2 & 0 \\ 3 & 2 & 0 \end{array} \right] \Rightarrow \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 2 & 0 & 0 \end{array} \right],$$

d. h. $x_1 = 0$ und $x_2 = 0$. Also ist $N(\mathbf{A}) = \{\vec{0}\}$. Der Spaltenraum ist der $\mathbb{R}^2$.

c) Wir beschreiben den Spalten- und Zeilenraum der Matrix $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$. Der

zweidimensionale Spaltenraum von $\mathbf{A}$ wird von $\vec{a}_1 = (1, 1, 0)^\top$ und $\vec{a}_2 = (0, 1, 1)^\top$ aufgespannt und liegt in $\mathbb{R}^3$. Er beschreibt damit eine Ebene durch den Ursprung. Der Zeilenraum von $\mathbf{A}$ ist ein Unterraum des $\mathbb{R}^2$ und wird von $\vec{z}_1 = (1, 0)$, $\vec{z}_2 = (1, 1)$, und $\vec{z}_3 = (0, 1)$ erzeugt. Da $\vec{z}_1, \vec{z}_3$ linear unabhängig sind, ist $Z(\mathbf{A}) = \mathbb{R}^2$. ■

Mit den Begriffen Nullraum und Spaltenraum liest sich der Dimensionssatz (Satz 3.25) für Matrizen so:

Satz 3.26 (Dimensionssatz für Matrizen)

Für die Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ (oder $\mathbb{C}^{m \times n}$) gilt

$$n = \dim S(\mathbf{A}) + \dim N(\mathbf{A}).$$

Bevor wir den Satz in dieser Form beweisen, setzen wir uns intensiver mit dem Null-, dem Zeilen- und dem Spaltenraum auseinander.

Lemma 3.14 (Elementare Zeilenumformungen)

Elementare Zeilenumformungen (wie beim Gauß-Verfahren) ändern die Lösungsmenge eines linearen Systems $\mathbf{A} \cdot \vec{x} = \vec{b}$ nicht. Folglich lassen sie den Nullraum $N(\mathbf{A})$ und den Zeilenraum $Z(\mathbf{A})$ unverändert. Anders formuliert: Entsteht die Matrix $\mathbf{B}$ aus der Matrix $\mathbf{A}$ durch elementare Zeilenumformungen, so gilt $Z(\mathbf{A}) = Z(\mathbf{B})$ und $N(\mathbf{A}) = N(\mathbf{B})$.

Man könnte nun vermuten, dass eine analoge Aussage auch für Spaltenräume gilt.

Beispiel 3.51

Die Spalten der Matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$ sind linear abhängig. Der Spaltenraum $S(\mathbf{A})$ wird z. B. durch $\vec{a}_1 = (1, 4)^\top$ aufgespannt. Die Matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$$

entsteht aus $\mathbf{A}$ durch elementare Zeilenumformung, ihre Spalten sind ebenfalls linear abhängig, aber es ist $S(\mathbf{A}) \neq S(\mathbf{B})$. Elementare Zeilenumformungen können also den Spaltenraum einer Matrix ändern. Der Spaltenraum bleibt unter elementaren Spaltenumformungen, die wir zuvor bei der Berechnung von Determinanten eingesetzt haben, unverändert. Bei diesen Umformungen ändert sich jedoch der Zeilenraum. ■

Satz 3.27 (Spaltenvektoren von Matrizen)

Die Matrix $\mathbf{B} = [\vec{b}_1, \dots, \vec{b}_n]$ gehe aus der Matrix $\mathbf{A} = [\vec{a}_1, \dots, \vec{a}_n]$ durch elementare Zeilenumformungen hervor. Die Spaltenvektoren von $\mathbf{A}$ sind genau dann linear unabhängig, wenn die Spaltenvektoren von $\mathbf{B}$ linear unabhängig sind. Die Vektoren $\vec{a}_1, \dots, \vec{a}_n$ bilden genau dann eine Basis von $S(\mathbf{A})$, wenn die Vektoren $\vec{b}_1, \dots, \vec{b}_n$ eine Basis von $S(\mathbf{B})$ bilden.

Beweis: Die Matrizen $\mathbf{A}$ und $\mathbf{B}$ haben denselben Nullraum, d. h., die Lösungsmengen von $\mathbf{A} \cdot \vec{x} = \vec{0}$ und $\mathbf{B} \cdot \vec{x} = \vec{0}$ sind gleich. Wir können die Gleichungen in der Form

$$x_1 \cdot \vec{a}_1 + \dots + x_n \cdot \vec{a}_n = \vec{0}$$

bzw.

$$x_1 \cdot \vec{b}_1 + \dots + x_n \cdot \vec{b}_n = \vec{0}$$

schreiben. Die erste Gleichung besitzt genau dann eine nicht-triviale Lösung, wenn die zweite Gleichung eine nicht-triviale Lösung hat. Damit sind die Vektoren $\vec{a}_1, \dots, \vec{a}_n$ genau dann linear (un-)abhängig, wenn $\vec{b}_1, \dots, \vec{b}_n$ linear (un-)abhängig sind. □

Diese Aussage gilt auch, wenn man nur einige ausgewählte Spaltenvektoren betrachtet. Zeilenumformungen ändern an ihrer linearen (Un-)Abhängigkeit nichts.

Mit Gauß'schen Zeilenumformungen kann man jede Matrix in eine Gestalt bringen, bei der in jeder Zeile der erste von null verschiedene Koeffizient auf eins normiert ist und in der Spalte, in der dieser Koeffizient steht, keine weiteren von null verschiedenen Einträge stehen. In dieser Darstellung, die wir **normierte Zeilenstufenform** nennen, kann man beispielsweise die Lösungen eines linearen Gleichungssystems ablesen (vgl. Beispiel 3.53).

Mittels Zeilenumformungen sind wir auch in der Lage, zu einer gegebenen Menge von Vektoren eine Basis zu konstruieren. Angenommen, die Menge hat m Vektoren aus $\mathbb{R}^n$, dann schreiben wir diese in eine $(m \times n)$ -Matrix, transformieren diese Matrix mit Hilfe elementarer Zeilenumformungen in die normierte Zeilenstufenform und lesen die Basisvektoren des Zeilenraums ab.

Beispiel 3.52

Wir bestimmen eine Basis der linearen Hülle der drei Vektoren $\vec{v}_1 = (1, 1)$, $\vec{v}_2 = (2, 0)$ und $\vec{v}_3 = (2, 1)$. Die Basis der linearen Hülle besteht aus höchstens zwei Vektoren. $\vec{v}_1, \vec{v}_2, \vec{v}_3$ spannen den Zeilenraum der Matrix

$$\begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 2 & 1 \end{bmatrix}$$

auf. Die normierte Zeilenstufenform der Matrix ergibt sich über

$$\begin{bmatrix} 1 & 1 \\ 0 & -2 \\ 0 & -1 \end{bmatrix} \quad \text{und} \quad \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{zu} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Da die Zeilenvektoren $(1, 0)$ und $(0, 1)$ vom Nullvektor verschieden sind, bilden sie eine Basis des Zeilenraums und damit eine Basis der linearen Hülle der Vektoren $\vec{v}_1, \vec{v}_2, \vec{v}_3$. ■

Mit der Zeilenstufenform kann man neben linear unabhängigen Zeilenvektoren auch direkt linear unabhängige Spaltenvektoren ablesen und abzählen:

Satz 3.28 (Basis des Spaltenraums)

Es sei $\mathbf{B} \in \mathbb{R}^{m \times n}$ eine Matrix in normierter Zeilenstufenform. Eine Basis des Zeilenraums von $\mathbf{B}$ erhält man aus denjenigen Zeilenvektoren von $\mathbf{B}$, die ungleich dem Nullvektor sind. Zu dieser Basis des Zeilenraums bilden wir ein System von Spaltenvektoren: Wir nehmen die Spalte k als Vektor genau dann auf, wenn es in der Basis des Zeilenraums einen Vektor gibt, dessen erster von null verschiedener Eintrag an

der Stelle k vorkommt. Diese Spaltenvektoren bilden (wie im folgenden Beispiel) eine Basis des Spaltenraums von $\mathbf{B}$. Insbesondere sieht man nun direkt, dass Zeilen- und Spaltenraum einer Matrix in Zeilenstufenform die gleiche Dimension haben.

Beispiel 3.53

Wir bestimmen eine Basis des Zeilen- und Spaltenraums der (4×5) -Matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Die Matrix $\mathbf{B}$ hat normierte Zeilenstufenform. Die Vektoren $\vec{z}_1 = (1, 0, 0, 0, 1)$, $\vec{z}_2 = (0, 1, 0, 0, 2)$ und $\vec{z}_3 = (0, 0, 0, 1, 3)$ bilden eine Basis von $Z(\mathbf{B})$ in $\mathbb{R}^5$. Die zu den führenden Einsen der Zeilen ausgewählten Vektoren $\vec{b}_1 = (1, 0, 0, 0)^\top$, $\vec{b}_2 = (0, 1, 0, 0)^\top$ und $\vec{b}_3 = (0, 0, 1, 0)^\top$ bilden eine Basis von $S(\mathbf{B})$ in $\mathbb{R}^4$. ■

Es sei $\mathbf{A}$ eine reelle $(m \times n)$ -Matrix, und $\mathbf{B}$ sei eine normierte Zeilenstufenmatrix von $\mathbf{A}$, die über Gauß-Umformungen aus $\mathbf{A}$ entstanden ist. Wegen Lemma 3.14 ist $Z(\mathbf{A}) = Z(\mathbf{B})$ und insbesondere $\dim Z(\mathbf{A}) = \dim Z(\mathbf{B})$. Soeben haben wir gesehen, dass $\dim Z(\mathbf{B}) = \dim S(\mathbf{B})$. Betrachten wir die Vektoren einer Basis des Spaltenraums von $\mathbf{B}$, so sind nach Satz 3.27 (und der anschließenden Bemerkung, da in der Basis nicht alle Spaltenvektoren enthalten sein müssen) auch die zugehörigen Spalten in $\mathbf{A}$ linear unabhängig (und keine weiteren), d. h. $\dim S(\mathbf{B}) = \dim S(\mathbf{A})$. Insgesamt haben wir gezeigt: Der Zeilenraum und der Spaltenraum einer Matrix haben die gleiche Dimension:

$$\dim Z(\mathbf{A}) = \dim S(\mathbf{A}).$$

Die Idee der Herleitung dieser Aussage ist ganz einfach: Man formt eine Matrix mit Zeilenumformungen so lange um, bis man sieht, dass es eine Basis des Zeilen- und eine Basis des Spaltenraums mit gleich vielen Elementen gibt. Die Umformungen ändern an den Dimensionen der Zeilen- und Spaltenräume nichts.

Definition 3.23 (Rang einer Matrix)

Da Zeilenraum und Spaltenraum einer Matrix die gleiche Dimension haben, spricht man vom **Rang** der Matrix $\mathbf{A}$:

$$\dim S(\mathbf{A}) = \dim Z(\mathbf{A}) =: \text{Rang}(\mathbf{A}).$$

Beispiel 3.54

a) Die Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

hat vier Zeilen und fünf Spalten, d. h. $\mathbf{A} \in \mathbb{R}^{4 \times 5}$. Es ist

$$\text{Rang}(\mathbf{A}) = \dim S(\mathbf{A}) = \dim Z(\mathbf{A}) = 3.$$

Nach (dem noch zu beweisenden) Satz 3.26 besitzt der Nullraum von $\mathbf{A}$ die Dimension

$$\dim N(\mathbf{A}) = 5 - \dim S(\mathbf{A}) = 5 - 3 = 2.$$

Dies leuchtet auch ein: Die Anzahl der führenden Einsen ist 3, und die Zahl der freien Variablen bei Verwendung der Matrix in einem Gleichungssystem ist 2.

b) Die Matrix

$$\mathbf{A} = [4 \ 6 \ 9]$$

hat eine Zeile und drei Spalten, d. h. $\mathbf{A} \in \mathbb{R}^{1 \times 3}$. Es ist

$$\text{Rang}(\mathbf{A}) = \dim S(\mathbf{A}) = \dim Z(\mathbf{A}) = 1.$$

Der Zeilenraum ist eine Gerade durch den Ursprung in $\mathbb{R}^3$. Der Nullraum ist die Ebene $4x_1 + 6x_2 + 9x_3 = 0$, hat also Dimension 2, $\dim N(\mathbf{A}) = 3 - 1 = 2$. ■

Wir tragen noch den Beweis des Dimensionssatzes für Matrizen nach, der zugleich den Dimensionssatz für lineare Abbildungen zeigt:

Beweis: (Satz 3.26) Nach Lemma 3.14 ändern elementare Zeilenumformungen weder den Nullraum noch den Zeilenraum. Mit ihnen gelangt man zu einer Matrix in normierter Zeilenstufenform. An dieser kann man ablesen, dass die Dimension des Zeilenraums plus die Dimension des Nullraums n ergibt. Damit hat man für die Ausgangsmatrix

$$n = \dim Z(A) + \dim N(A) = \text{Rang}(A) + \dim N(A) = \dim S(A) + \dim N(A). \quad \square$$

3.4.4 Umkehrabbildung und inverse Matrix

Wir haben mit der Definition 1.7 auf Seite 9 für Abbildungen die Begriffe „injektiv“ (jedes Element des Bildes wird nur einmal angenommen), „surjektiv“ (jedes Element der Zielmenge wird angenommen) und „bijektiv“ (injektiv und surjektiv) erklärt. Diese Begriffe sind insbesondere auf lineare Abbildungen anwendbar. Eine bijektive lineare Abbildung $L : V \rightarrow W$ hat demnach eine Umkehrabbildung, für die gilt:

$$L^{-1}(L(\vec{v})) = \vec{v}, \quad \vec{v} \in V \quad \text{und} \quad L(L^{-1}(\vec{w})) = \vec{w}, \quad \vec{w} \in W.$$

Beispiel 3.55

- a) $L : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, $(x_1, x_2) \mapsto (x_1, x_2, 0)$ ist injektiv, aber nicht surjektiv. Jeder Bildvektor $(x_1, x_2, 0)$ in $\mathbb{R}^3$ hat genau ein Urbild, aber es ist $(0, 0, 1) \notin \text{Bild}(L)$.
- b) $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, $(x_1, x_2, x_3) \mapsto (x_1, x_2)$ ist surjektiv, aber nicht injektiv. Jeder Vektor (x_1, x_2) liegt in $\text{Bild}(L)$, aber $L((1, 1, 0)) = L((1, 1, 1)) = (1, 1)$.
- c) Die Drehung um den Winkel φ , definiert durch (1.48) auf Seite 160

$$L : \mathbb{R}^2 \rightarrow \mathbb{R}^2, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} \cos(\varphi) \cdot x_1 - \sin(\varphi) \cdot x_2 \\ \sin(\varphi) \cdot x_1 + \cos(\varphi) \cdot x_2 \end{pmatrix}$$

ist bijektiv und hat eine anschauliche Umkehrabbildung: Man dreht um den Winkel $-\varphi$:

$$L^{-1} : \mathbb{R}^2 \rightarrow \mathbb{R}^2, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} \cos(\varphi) \cdot x_1 + \sin(\varphi) \cdot x_2 \\ -\sin(\varphi) \cdot x_1 + \cos(\varphi) \cdot x_2 \end{pmatrix}. \quad \blacksquare$$

Im Rest dieses Abschnitts beschränken wir uns auf lineare Abbildungen $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$, die beschrieben werden durch $(n \times n)$ -Matrizen: $L(\vec{x}) = \mathbf{A} \cdot \vec{x}$ für alle $\vec{x} \in \mathbb{R}^n$.

Die Bedingung, dass L injektiv ist, also

$$L(\vec{x}_1) = \mathbf{A} \cdot \vec{x}_1 = L(\vec{x}_2) = \mathbf{A} \cdot \vec{x}_2 \implies \vec{x}_1 = \vec{x}_2,$$

führt auf die Bedingung

$$\mathbf{A} \cdot \vec{x}_1 - \mathbf{A} \cdot \vec{x}_2 = \mathbf{A} \cdot (\vec{x}_1 - \vec{x}_2) = \vec{0} \implies \vec{x}_1 - \vec{x}_2 = \vec{0}.$$

Es gilt also

Satz 3.29 (Matrix einer injektiven Abbildung)

Sei $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ mit $L(\vec{x}) = \mathbf{A} \cdot \vec{x}$ eine lineare Abbildung mit der dazugehörigen $(n \times n)$ -Matrix $\mathbf{A}$. Dann ist L injektiv genau dann, wenn die Matrix $\mathbf{A}$ regulär (d. h. invertierbar) ist.

Ist eine lineare Abbildung $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ injektiv, so hat das Bild die Dimension n , und sie ist automatisch auch surjektiv. L ist also bijektiv und damit invertierbar.

Wir betrachten nun eine bijektive lineare Abbildung $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ mit zugehöriger Matrix $\mathbf{A}$. Wird der Vektor $\vec{x}$ auf $\vec{b}$ abgebildet, so gilt $\mathbf{A} \cdot \vec{x} = \vec{b}$. Für die Beschreibung der Umkehrabbildung L^{-1} stellt man sich die Aufgabe, eine Matrix $\mathbf{B}$ zu finden, die den Vektor $\vec{b}$ auf den Vektor $\vec{x}$ abbildet:

$$\mathbf{B} \cdot \vec{b} = \mathbf{B} \cdot (\mathbf{A} \cdot \vec{x}) = \vec{x}.$$

Die Matrix $\mathbf{B}$ ist genau durch die inverse Matrix zu $\mathbf{A}$ gegeben: $\mathbf{B} = \mathbf{A}^{-1}$.

Satz 3.30 (Invertierbare lineare Abbildung)

Es sei $\mathbf{A}$ eine $(n \times n)$ -Matrix und L die lineare Abbildung $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $\vec{x} \mapsto L(\vec{x}) = \mathbf{A} \cdot \vec{x}$. Dann sind folgende Aussagen äquivalent:

- $\mathbf{A}$ ist regulär, d. h. invertierbar mit inverser Matrix $\mathbf{A}^{-1}$.
- $\det \mathbf{A} \neq 0$.
- Das homogene lineare Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{0}$ hat nur die triviale Lösung $\vec{x} = \vec{0}$.
- Für jedes $\vec{b} \in \mathbb{R}^n$ hat das inhomogene lineare Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ genau eine Lösung.
- Die lineare Abbildung L ist bijektiv.
- Der Kern von L besteht nur aus dem Nullvektor, $\dim \text{Kern}(L) = 0$.
- Die Spalten von $\mathbf{A}$ sind linear unabhängig, $\dim S(\mathbf{A}) = n$.
- Die Zeilen von $\mathbf{A}$ sind linear unabhängig, $\dim Z(\mathbf{A}) = n$.
- Der Rang der Matrix $\mathbf{A}$ ist n .

Trifft eine der Aussagen zu, so ist die zu $\mathbf{A}$ inverse Matrix $\mathbf{A}^{-1}$ die Matrix der Umkehrabbildung L^{-1} , d. h., es gilt

$$L^{-1}(\vec{x}) = \mathbf{A}^{-1} \cdot \vec{x}, \quad \vec{x} \in \mathbb{R}^n.$$

Beispiel 3.56

Betrachten wir die lineare Abbildung $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ mit

$$L\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} x+y \\ 3x+2y \end{pmatrix}, \text{ also } L(\vec{x}) = \mathbf{A} \cdot \vec{x} \text{ mit } \mathbf{A} = \begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix}.$$

Wegen $\text{Kern}(L) = \{\vec{0}\}$ ist L umkehrbar. Die Abbildungsmatrix von L^{-1} ist die Inverse der Matrix $\mathbf{A}$:

$$\left[\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 3 & 2 & 0 & 1 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 0 & -1 & -3 & 1 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & 1 & 3 & -1 \end{array} \right],$$

$$\text{also } L^{-1}(\vec{x}) = \mathbf{A}^{-1} \cdot \vec{x} \text{ mit } \mathbf{A}^{-1} = \begin{bmatrix} -2 & 1 \\ 3 & -1 \end{bmatrix}. \quad \blacksquare$$

3.4.5 Koordinaten- und Basistransformationen*

Ein Vektor $\vec{x} \in \mathbb{R}^n$ besitzt bezüglich einer Basis des $\mathbb{R}^n$ eine eindeutige Darstellung als Linearkombination aus den Basisvektoren. Bei vielen Anwendungen arbeitet man mit mehreren Basen und man muss die Darstellungen von $\vec{x}$ bezüglich einer Basis in die Darstellung bezüglich einer anderen Basis umrechnen.

Wir betrachten im Vektorraum $\mathbb{R}^n$ die Basis $A := \{\vec{a}_1, \dots, \vec{a}_n\}$ und die Basis $B := \{\vec{b}_1, \dots, \vec{b}_n\}$. Der Vektor $\vec{x} \in \mathbb{R}^n$ besitze die Darstellungen als Linearkombination

$$\vec{x} = \sum_{k=1}^n x_{A,k} \vec{a}_k \quad \text{und} \quad \vec{x} = \sum_{k=1}^n x_{B,k} \vec{b}_k.$$

In A hat $\vec{x}$ die Koordinaten $\vec{x}_A = (x_{A,1}, \dots, x_{A,n})$, in B gilt entsprechend $\vec{x}_B = (x_{B,1}, \dots, x_{B,n})$. Wir definieren die **Basismatrix** $\mathbf{A} := [\vec{a}_1, \dots, \vec{a}_n]$ als die Matrix mit den Basisvektoren $\vec{a}_k$ als Spalten und entsprechend $\mathbf{B} := [\vec{b}_1, \dots, \vec{b}_n]$. Dann folgt

$$\vec{x} = \mathbf{A} \cdot \vec{x}_A \quad \text{und} \quad \vec{x} = \mathbf{B} \cdot \vec{x}_B.$$

Den Übergang von A zu B nennt man **Basistransformation**. Die dabei stattfindende Koordinatenänderung heißt **Koordinatentransformation**. Bei einem Basiswechsel kann man den neuen Koordinatenvektor $\vec{x}_B$ aus dem alten Koordinatenvektor $\vec{x}_A$ berechnen.

Satz 3.31 (Koordinatentransformation)

Es sei A eine Basis mit Basismatrix $\mathbf{A} = [\vec{a}_1, \dots, \vec{a}_n]$ und B eine Basis mit Basismatrix $\mathbf{B} = [\vec{b}_1, \dots, \vec{b}_n]$. Der Vektor $\vec{x} \in \mathbb{R}^n$ besitze bezüglich A die Koordinaten $\vec{x}_A$ und bezüglich B die Koordinaten $\vec{x}_B$. Dann gilt

$$\begin{aligned} \vec{x}_B &= \mathbf{T} \cdot \vec{x}_A \quad \text{mit} \quad \mathbf{T} := \mathbf{B}^{-1} \cdot \mathbf{A}, \\ \vec{x}_A &= \mathbf{T}^{-1} \cdot \vec{x}_B \quad \text{mit} \quad \mathbf{T}^{-1} := \mathbf{A}^{-1} \cdot \mathbf{B}. \end{aligned} \tag{3.25}$$

Beweis: Die Matrizen $\mathbf{A}$ und $\mathbf{B}$ sind wegen der Basiseigenschaft der Spalten invertierbar. Aus $\vec{x} = \mathbf{A} \cdot \vec{x}_A = \mathbf{B} \cdot \vec{x}_B$ folgt (3.25) durch Multiplikation mit einer inversen Basismatrix. $\square$

Die Matrix $\mathbf{T} = \mathbf{T}(\mathbf{A}, \mathbf{B})$ heißt **Transformationsmatrix**. Die Koordinatentransformation ist eine umkehrbare lineare Abbildung.

Die Spaltenvektoren $\vec{a}_k$ von $\mathbf{A}$ haben in ihrer Basis A die Koordinatendarstellung $(\vec{a}_k)_A = \vec{e}_k$. Daher sind nach (3.25) die Spalten von $\mathbf{T}$ ihre Koordinatenvektoren in der Basis B . Damit können wir mittels $\mathbf{T}$ auch die Basen ineinander umrechnen. Mit der Transformationsmatrix $\mathbf{T} = \mathbf{B}^{-1} \cdot \mathbf{A}$ ist $\mathbf{A} = \mathbf{B} \cdot \mathbf{T}$: Die k -te Spalte von $\mathbf{A}$ (d.h. der Basisvektor $\vec{a}_k$) ist die Linearkombination der Spalten von $\mathbf{B}$ (d.h. der Basisvektoren $\vec{b}_i$) mit den skalaren Faktoren $t_{i,k}$:

$$\vec{a}_k = \sum_{i=1}^n t_{i,k} \cdot \vec{b}_i, \quad \mathbf{T} = [t_{i,k}]_{i,k=1,\dots,n}.$$

Umgekehrt folgt aus $\mathbf{B} = \mathbf{A} \cdot \mathbf{T}^{-1}$: Die k -te Spalte von $\mathbf{B}$ (d. h. der Basisvektor $\vec{b}_k$) ist die Linearkombination der Spalten von $\mathbf{A}$ (d. h. der Basisvektoren $\vec{a}_i$) mit den skalaren Faktoren $t_{i,k}^{(-1)}$ (wobei mit $t_{i,k}^{(-1)}$ nicht der Kehrwert $1/t_{i,k}$ gemeint ist):

$$\vec{b}_k = \sum_{i=1}^n t_{i,k}^{(-1)} \cdot \vec{a}_i, \quad \mathbf{T}^{-1} = [t_{i,k}^{(-1)}]_{i,k=1,\dots,n}.$$

Handelt es sich bei A und B um **Orthonormalbasen**, so gilt $\mathbf{A}^{-1} = \mathbf{A}^\top$ bzw. $\mathbf{B}^{-1} = \mathbf{B}^\top$, und für die Transformationsmatrizen $\mathbf{T}$ und $\mathbf{T}^{-1}$ folgt ebenfalls

$$\mathbf{T} = \mathbf{B}^\top \cdot \mathbf{A} \quad \text{und} \quad \mathbf{T}^{-1} = \mathbf{A}^\top \cdot \mathbf{B} = (\mathbf{B}^\top \cdot \mathbf{A})^\top = \mathbf{T}^\top.$$

Wir betrachten nun die Auswirkung eines Basiswechsels auf die Matrix einer linearen Abbildung.

Satz 3.32 (Transformation der Matrix einer linearen Abbildung)

Es sei $\mathbf{L}$ eine lineare Abbildung von $\mathbb{R}^n \rightarrow \mathbb{R}^n$. Bezüglich der Basis A besitze $\mathbf{L}$ die Abbildungsmatrix $\mathbf{K}$, d. h. es gelte

$$(\mathbf{L}(\vec{x}))_A = \mathbf{K} \cdot \vec{x}_A.$$

Bezüglich der Basis B besitze $\mathbf{L}$ die Abbildungsmatrix $\mathbf{M}$:

$$(\mathbf{L}(\vec{x}))_B = \mathbf{M} \cdot \vec{x}_B.$$

Dann gilt mit der Transformationsmatrix $\mathbf{T}(A, B)$:

$$\mathbf{K} = \mathbf{T}^{-1} \cdot \mathbf{M} \cdot \mathbf{T}. \quad (3.26)$$

Beweis: Nach (3.25) ist $\vec{x}_B = \mathbf{T} \cdot \vec{x}_A$ und $(\mathbf{L}(\vec{x}))_B = \mathbf{T} \cdot (\mathbf{L}(\vec{x}))_A$. Damit erhalten wir

$$\begin{aligned} (\mathbf{L}(\vec{x}))_A &= \mathbf{T}^{-1} \cdot (\mathbf{L}(\vec{x}))_B = \mathbf{T}^{-1} \cdot \mathbf{M} \cdot \vec{x}_B \\ &= \mathbf{T}^{-1} \cdot \mathbf{M} \cdot \mathbf{T} \cdot \vec{x}_A. \end{aligned}$$

Andererseits ist $(\mathbf{L}(\vec{x}))_A = \mathbf{K} \cdot \vec{x}_A$. Da beides für alle Vektoren $\vec{x}_A \in \mathbb{R}^n$ gilt, stimmen die Matrizen überein:

$$\mathbf{K} = \mathbf{T}^{-1} \cdot \mathbf{M} \cdot \mathbf{T}. \quad \square$$

Matrizen, die die gleiche lineare Abbildung bezüglich unterschiedlicher Basen beschreiben, bekommen einen Namen:

Definition 3.24 (Ähnlichkeit von Matrizen)

Zwei $(n \times n)$ -Matrizen $\mathbf{K}$ und $\mathbf{M}$ heißen genau dann zueinander **ähnlich**, wenn es eine invertierbare $(n \times n)$ -Matrix $\mathbf{T}$ gibt mit

$$\mathbf{K} = \mathbf{T}^{-1} \cdot \mathbf{M} \cdot \mathbf{T}.$$

Wir werden später in den Kapiteln 3.6.2 und 3.6.3 noch sehen, dass eine komplexe $(n \times n)$ -Matrix immer ähnlich zu einer Diagonalmatrix oder einer Tridiagonalmatrix mit Einträgen auf nur zwei Diagonalen ist.

Beispiel 3.57

- a) Die Vektoren $\vec{a}_1 = (1, 1)^\top$ und $\vec{a}_2 = (0, 2)^\top$ bilden eine Basis des $\mathbb{R}^2$. Wir berechnen die Transformationsmatrizen bezüglich $A = \{\vec{a}_1, \vec{a}_2\}$ und der Standardbasis $B = \{\vec{e}_1, \vec{e}_2\}$ in $\mathbb{R}^2$. Für einen beliebigen Vektor $\vec{c}$ mit Koordinaten c_1 und c_2 bezüglich der Standardbasis B berechnen wir die Koordinaten $(x_1, x_2)^\top := \vec{c}_A$:

$$\vec{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = x_1 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + x_2 \cdot \begin{pmatrix} 0 \\ 2 \end{pmatrix} \iff \begin{array}{lcl} c_1 & = & x_1 \\ c_2 & = & x_1 + 2x_2 \end{array}.$$

In Matrixschreibweise:

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Durch Linksmultiplikation mit der inversen Matrix folgt

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

- b) Wir betrachten in $\mathbb{R}^2$ die Basen

$$A = \left\{ \vec{a}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \vec{a}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\} \quad \text{und} \quad B = \left\{ \vec{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}.$$

Die lineare Abbildung $\mathbf{L} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ sei bezüglich A erklärt durch

$$\mathbf{L}(\vec{a}_1) = 2\vec{a}_1, \quad \mathbf{L}(\vec{a}_2) = \vec{a}_1 - \vec{a}_2.$$

Die Abbildungsmatrix $\mathbf{K}$ von $\mathbf{L}$ bezüglich A lautet:

$$\mathbf{K} = [(\mathbf{L}(\vec{a}_1))_A, (\mathbf{L}(\vec{a}_2))_A] = \left[\begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right] = \begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix}.$$

Nun bestimmen wir die Abbildungsmatrix $\mathbf{M}$ von $\mathbf{L}$ bezüglich B . Die Transformationsmatrix lautet

$$\mathbf{T} = \mathbf{B}^{-1} \cdot \mathbf{A} = \mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}.$$

Die dazu inverse Matrix erhalten wir über

$$\left[\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 2 & -1 & 0 & 1 \end{array} \right] \Rightarrow \left[\begin{array}{cc|cc} 3 & 0 & 1 & 1 \\ 0 & -3 & -2 & 1 \end{array} \right] \Rightarrow \left[\begin{array}{cc|cc} 1 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & \frac{2}{3} & -\frac{1}{3} \end{array} \right].$$

Wegen $\mathbf{K} = \mathbf{T}^{-1} \cdot \mathbf{M} \cdot \mathbf{T} \Leftrightarrow \mathbf{M} = \mathbf{T} \cdot \mathbf{K} \cdot \mathbf{T}^{-1}$ ist die gesuchte Abbildungsmatrix

$$\mathbf{M} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{2}{3} \\ \frac{10}{3} & \frac{1}{3} \end{bmatrix}. \quad \blacksquare$$

3.5 Lösungstheorie linearer Gleichungssysteme

Nach dem einführenden Kapitel 1.7.3.2 gleichen Namens haben wir nun das nötige Rüstzeug, uns noch einmal intensiver mit linearen Gleichungssystemen zu beschäftigen und diese insbesondere im Kontext der Vektorräume zu betrachten. Anschließend wenden wir die Theorie für die Berechnung von Strömen und Spannungen in einem Netzwerk an.

Hat man ein Gleichungssystem mit einer $(n \times n)$ -Matrix $\mathbf{A}$, dann ist es genau dann für jede Inhomogenität eindeutig lösbar, wenn die Spalten der Matrix eine Basis des $\mathbb{C}^n$ bilden, so dass jede Inhomogenität eindeutig linear aus den Spaltenvektoren kombiniert werden kann. Insbesondere ist nach Satz 3.15 dazu äquivalent, dass $\det \mathbf{A} \neq 0$ ist und es eine inverse Matrix $\mathbf{A}^{-1}$ gibt. Diese Bedingung haben wir bereits aus anderer Sicht in Kapitel 1.8.2 kennengelernt.

Beispiel 3.58 (Kreisbögen, definiert über drei Punkte auf dem Bogen)

Bei der Vermessung von Gebäuden werden Kreisbögen über den Anfangs- und Endpunkt sowie einen weiteren Punkt auf dem Bogen gespeichert. Wir wollen zeigen: Liegen drei Punkte (x_1, y_1) , (x_2, y_2) und (x_3, y_3) im $\mathbb{R}^2$ nicht auf einer Geraden, so gibt es genau einen Kreis, auf dem diese drei Punkte liegen.

Dazu konstruieren wir den Mittelpunkt (x_0, y_0) des Kreises, auf dem die Punkte liegen. Der Trick dabei ist, dass alle Punkte auf einem Kreis den gleichen Abstand zum Mittelpunkt haben, es muss also für die quadrierten Abstände gelten:

$$\begin{aligned} (x_1 - x_0)^2 + (y_1 - y_0)^2 &= (x_2 - x_0)^2 + (y_2 - y_0)^2 \\ \wedge \quad (x_1 - x_0)^2 + (y_1 - y_0)^2 &= (x_3 - x_0)^2 + (y_3 - y_0)^2. \end{aligned}$$

Wenn wir ausmultiplizieren, erhalten wir daraus ein lineares Gleichungssystem für die Unbekannten x_0 und y_0 :

$$\begin{aligned} (-2x_1 + 2x_2)x_0 + (-2y_1 + 2y_2)y_0 &= -x_1^2 - y_1^2 + x_2^2 + y_2^2 \\ \wedge \quad (-2x_1 + 2x_3)x_0 + (-2y_1 + 2y_3)y_0 &= -x_1^2 - y_1^2 + x_3^2 + y_3^2. \end{aligned}$$

Das Gleichungssystem ist genau dann eindeutig lösbar, wenn die Determinante der zugehörigen Matrix ungleich null ist:

$$\begin{aligned} D &:= \det \begin{bmatrix} (-2x_1 + 2x_2) & (-2y_1 + 2y_2) \\ (-2x_1 + 2x_3) & (-2y_1 + 2y_3) \end{bmatrix} \\ &= (-2x_1 + 2x_2) \cdot (-2y_1 + 2y_3) - (-2y_1 + 2y_2) \cdot (-2x_1 + 2x_3). \end{aligned}$$

Die Determinante ist genau dann ungleich null, wenn die Zeilenvektoren linear unabhängig sind. Diese lauten $2 \cdot [(x_2, y_2) - (x_1, y_1)]$ und $2 \cdot [(x_3, y_3) - (x_1, y_1)]$, so dass Lineare Unabhängigkeit genau dann vorliegt, wenn die gegebenen Punkte nicht auf einer Geraden liegen. Mit der Cramer'schen Regel (siehe Seite 197) lassen sich in diesem Fall die Koordinaten des Kreismittelpunktes berechnen:

$$\begin{aligned} x_0 &= [(-x_1^2 - y_1^2 + x_2^2 + y_2^2)(-2y_1 + 2y_3) - (-x_1^2 - y_1^2 + x_3^2 + y_3^2)(-2y_1 + 2y_2)]/D, \\ y_0 &= [(-x_1^2 - y_1^2 + x_3^2 + y_3^2)(-2x_1 + 2x_2) - (-x_1^2 - y_1^2 + x_2^2 + y_2^2)(-2x_1 + 2x_3)]/D. \end{aligned}$$

Um den Kreisbogen von (x_1, y_1) bis (x_3, y_3) durch einen Polygonzug (eine Folge von geraden Strecken) anzunähern, können wir rekursiv Zwischenpunkte bilden, bis diese näher als eine vorgegebene Entfernung ε zusammen liegen oder bis der (über das Skalarprodukt berechnete) Winkel zwischen dem Mittelpunkt und benachbarten Zwischenpunkten kleiner als δ ist. Dies geschieht durch den Aufruf

ZEICHNEBOGEN($(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3)$)

für das in einem halb-formalen Pseude-Code geschriebene Programm

```
procedure BESTIMMEREKURSIVPUNKTE( $(x_0, y_0), (x_1, y_1), (x_2, y_2)$ )
  if  $|(x_1, y_1) - (x_2, y_2)| < \varepsilon$  then return
  if  $\frac{|(x_1, y_1) - (x_0, y_0)| \cdot |(x_2, y_2) - (x_0, y_0)|}{|(x_1, y_1) - (x_0, y_0)| |(x_2, y_2) - (x_0, y_0)|} > \cos(\delta)$  then return
   $\vec{m} := [(x_1, y_1) - (x_0, y_0) + (x_2, y_2) - (x_0, y_0)]$ 
  BESTIMMEREKURSIVPUNKTE( $(x_0, y_0), (x_1, y_1), (x_0, y_0) + \frac{|(x_1, y_1) - (x_0, y_0)|}{|\vec{m}|} \vec{m}$ )
  Zeichne eine Strecke zum Punkt  $(x_0, y_0) + \frac{|(x_1, y_1) - (x_0, y_0)|}{|\vec{m}|} \vec{m}$ .
  BESTIMMEREKURSIVPUNKTE( $(x_0, y_0), (x_0, y_0) + \frac{|(x_1, y_1) - (x_0, y_0)|}{|\vec{m}|} \vec{m}, (x_2, y_2)$ )
return

procedure ZEICHNEBOGEN( $(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3)$ )
  Starte im Punkt  $(x_1, y_1)$ .
  BESTIMMEREKURSIVPUNKTE( $(x_0, y_0), (x_1, y_1), (x_2, y_2)$ )
  Zeichne eine Strecke zum Punkt  $(x_2, y_2)$ .
```

BESTIMMEREKURSIVPUNKTE($((x_0, y_0), (x_2, y_2), (x_3, y_3))$)

Zeichne eine Strecke zum Punkt (x_3, y_3) .

return

Die Tanks des Krefelder Hafens im rechten Bild von Abbildung 3.30 wurden mit diesem Algorithmus gezeichnet. ■

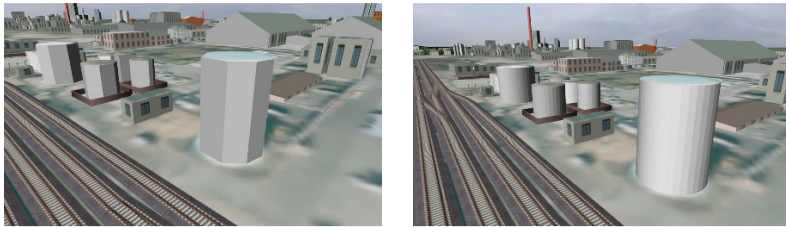


Abb. 3.30: Kreisbögen: links nur mit den drei definierenden Punkten durch je zwei Linien gezeichnet und rechts mit weiteren Zwischenpunkten versehen.

Nun beschäftigen wir uns aber auch mit dem Fall, dass $\mathbf{A}$ nicht invertierbar ist oder dass die Anzahl der Gleichungen nicht mit der Anzahl der Variablen übereinstimmt.

3.5.1 Lösungsraum eines linearen Gleichungssystems

Die Lösungsmenge eines **homogenen Gleichungssystems** $\mathbf{A} \cdot \vec{x} = \vec{0}$ mit $\mathbf{A} \in \mathbb{C}^{n \times n}$ ist ein Unterraum des $\mathbb{C}^n$. Um das zu sehen, müssen wir nach Satz 3.10 lediglich zeigen, dass auch Summen und skalare Vielfache von Lösungen wieder Lösungen sind:

Lemma 3.15 (Struktur homogener Lösungen)

Addiert man zwei homogene Lösungen $\vec{x}$ und $\vec{y}$ eines linearen Gleichungssystems (also $\mathbf{A} \cdot \vec{x} = \vec{0}$ und $\mathbf{A} \cdot \vec{y} = \vec{0}$), so erhält man wieder eine homogene Lösung:

$$\mathbf{A} \cdot (\vec{x} + \vec{y}) = \mathbf{A} \cdot \vec{x} + \mathbf{A} \cdot \vec{y} = \vec{0} + \vec{0} = \vec{0}.$$

Multipliziert man eine homogene Lösung $\vec{x}$ mit einer Zahl c , so ist das Ergebnis wieder eine homogene Lösung: $\mathbf{A} \cdot (c \cdot \vec{x}) = c \cdot (\mathbf{A} \cdot \vec{x}) = c \cdot \vec{0} = \vec{0}$.

Der Unterraum der homogenen Lösungen ist gleich dem Nullraum der Matrix $\mathbf{A}$ des Gleichungssystems und gleich dem Kern der zugehörigen linearen Abbildung.

Die Menge der Lösungen eines **inhomogenen Gleichungssystems** $\mathbf{A} \cdot \vec{x} = \vec{b} \neq \vec{0}$ ist **kein** Vektorraum. Addiert man zwei Lösungen $\vec{x}$ und $\vec{y}$, so ist bereits $\vec{x} + \vec{y}$ keine inhomogene Lösung mehr und damit nicht in der Menge enthalten: $\mathbf{A} \cdot (\vec{x} + \vec{y}) =$

$\mathbf{A} \cdot \vec{x} + \mathbf{A} \cdot \vec{y} = \vec{b} + \vec{b} \neq \vec{b}$. Interessant ist dennoch die Betrachtung der Differenz zweier inhomogener Lösungen:

Lemma 3.16 (Struktur inhomogener Lösungen)

Man erhält alle Lösungen eines inhomogenen linearen Gleichungssystems $\mathbf{A} \cdot \vec{x} = \vec{b}$, indem man eine inhomogene Lösung berechnet und dann Lösungen des homogenen Gleichungssystems $\mathbf{A} \cdot \vec{x} = \vec{0}$ hinzuaddiert.

Beweis: Die Differenz zweier inhomogener Lösungen $\vec{x}$ und $\vec{y}$ ist eine homogene Lösung:

$$\mathbf{A} \cdot (\vec{x} - \vec{y}) = \mathbf{A} \cdot \vec{x} - \mathbf{A} \cdot \vec{y} = \vec{b} - \vec{b} = \vec{0}.$$

Kennt man also alle homogenen Lösungen (Lösungen von $\mathbf{A} \cdot \vec{x} = \vec{0}$), so benötigt man nur eine inhomogene Lösung, um alle inhomogenen Lösungen zu erhalten. $\square$

Beispiel 3.59

Betrachten wir das homogene (unterbestimmte) reelle Gleichungssystem (vgl. (1.52) auf Seite 167)

$$\left[\begin{array}{ccccc|c} 1 & 0 & 0 & 1 & 3 & 0 \\ 0 & 1 & 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] \quad \text{bzw.} \quad \begin{array}{lcl} x_1 & = & -x_4 - 3x_5 \\ \wedge \quad x_2 & = & -x_4 - 2x_5 \\ \wedge \quad x_3 & = & -2x_4 - 1x_5, \end{array}$$

so ist die Lösungsmenge

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \in \mathbb{R}^5 : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = x_4 \begin{pmatrix} -1 \\ -1 \\ -2 \\ 1 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} -3 \\ -2 \\ -1 \\ 0 \\ 1 \end{pmatrix}, x_4, x_5 \in \mathbb{R} \right\}$$

ein Vektorraum mit Erzeugendensystem $\{(-1, -1, -2, 1, 0)^\top, (-3, -2, -1, 0, 1)^\top\}$.

Die beiden Vektoren des Erzeugendensystems ergeben sich als Lösungen, wenn man eine der beiden freien Variablen x_4 oder x_5 zu 1 und die jeweils andere zu 0 wählt. Dadurch erhält man linear unabhängige Vektoren, also eine Basis. Die Menge der homogenen Lösungen ist also hier ein Vektorraum der Dimension zwei. Die Dimension stimmt mit der Anzahl der freien Variablen überein. $\blacksquare$

Lemma 3.17 (Lösbarkeit eines linearen Gleichungssystems)

Das lineare Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ ist genau dann lösbar, wenn $\vec{b} \in S(\mathbf{A})$, d. h. wenn die Inhomogenität $\vec{b}$ im Spaltenraum von $\mathbf{A}$ liegt.

Beispiel 3.60

- a) Hat man ein Gleichungssystem, dessen Matrix die $(n \times n)$ -Einheitsmatrix $\mathbf{E}$ ist, also $\mathbf{E}\vec{x} = \vec{b}$, so kann man für jede Inhomogenität $\vec{b}$ die eindeutige Lösung $\vec{x} = \vec{b}$ direkt ablesen. Das passt auch zum vorangehenden Lemma, denn der Spaltenraum der Matrix ist der Vektorraum $\mathbb{R}^n$ oder $\mathbb{C}^n$.
- b) Das lineare System

$$\begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & -3 \\ 9 & 6 & -5 \end{bmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \\ 5 \end{pmatrix}$$

ist lösbar, denn die rechte Seite lässt sich als Linearkombination der Spalten von $\mathbf{A}$ schreiben:

$$1 \cdot \begin{pmatrix} 2 \\ 6 \\ 9 \end{pmatrix} + 1 \cdot \begin{pmatrix} 1 \\ 4 \\ 6 \end{pmatrix} + 2 \cdot \begin{pmatrix} 2 \\ -3 \\ -5 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \\ 5 \end{pmatrix}.$$

Eine Lösung des Gleichungssystems kann man direkt an den Faktoren der Linearkombination ablesen: $x_1 = 1$, $x_2 = 1$ und $x_3 = 2$.

- c) Der Spaltenraum $S(\mathbf{A})$ der Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 4 \end{bmatrix}$$

ist der gesamte Vektorraum $\mathbb{C}^2$, da wir bereits mit der ersten Spalte $(1, 0)^\top$ und der dritten Spalte $(3, 4)^\top$ jeden Vektor aus $\mathbb{C}^2$ linear kombinieren können. Damit lässt sich ein zugehöriges Gleichungssystem für jede Inhomogenität lösen. Die Lösung ist aber nicht eindeutig. Beispielsweise erhalten wir für $\vec{b} = (1, 2)^\top$ die Lösungen $\vec{x} = (-\frac{1}{2}, 0, \frac{1}{2})^\top + \lambda(-2, 1, 0)^\top$, $\lambda \in \mathbb{C}$. ■

Beispiel 3.61

Wir betrachten ein weiteres homogenes Gleichungssystem mit einer $(m \times n)$ -Matrix für $m = 4$ und $n = 5$:

$$\begin{bmatrix} 2 & 1 & 3 & 2 & 1 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \iff$$

$$\begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} x_1 + \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} x_2 + \begin{pmatrix} 3 \\ 1 \\ 0 \\ 0 \end{pmatrix} x_3 + \begin{pmatrix} 2 \\ 3 \\ 1 \\ 0 \end{pmatrix} x_4 + \begin{pmatrix} 1 \\ 1 \\ 3 \\ 0 \end{pmatrix} x_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Eine Variable kann frei gewählt werden, wenn der zugehörige Spaltenvektor der Matrix des Gleichungssystems sich als Linearkombination der anderen Spalten schreiben lässt. Im Beispiel kann die letzte Spalte, die zur Variable x_5 gehört, aus den anderen linear kombiniert werden:

$$\begin{pmatrix} 1 \\ 1 \\ 3 \\ 0 \end{pmatrix} = 3 \begin{pmatrix} 2 \\ 3 \\ 1 \\ 0 \end{pmatrix} - 8 \begin{pmatrix} 3 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 19 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Hat man sich für eine Variable entschieden, betrachtet man die Matrix ohne die entsprechende Spalte und prüft, ob sich eine weitere Spalte als Linearkombination der restlichen schreiben lässt. Ist das so, so kann man auch die dazugehörige Variable frei wählen. Im Beispiel kann nun z. B. die erste Spalte als zweimal die zweite geschrieben werden. Dies lässt sich fortsetzen, bis nur noch eine Menge von k linear unabhängigen Spalten übrig bleibt. Dann ist keine der verbleibenden Spalten mehr eine Linearkombination von anderen (siehe (3.8)). Dafür lässt sich aber jede der weggelassenen Spalten als Linearkombination der verbleibenden schreiben. Im Beispiel sind keine weiteren Variablen frei wählbar, und die Spalten

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 1 \\ 0 \end{pmatrix}$$

sind linear unabhängig. Dies ist eine Basis eines Unterraums von $\mathbb{C}^4$ mit Dimension 3. ■

Hat man ein Gleichungssystem mit n Variablen, so ist dieses über eine $(m \times n)$ -Matrix $\mathbf{A}$ darstellbar. Die Dimension des Raums der homogenen Lösungen ist die Dimension des Nullraums $\dim N(\mathbf{A})$. Nach dem Dimensionssatz für Matrizen (Satz 3.26 auf Seite 511) gilt $n = \dim S(\mathbf{A}) + \dim N(\mathbf{A})$, d. h., die Dimension des Spaltenraums, die gleich dem Rang der Matrix ist, plus die Dimension des Raums homogener Lösungen ist gleich der Anzahl der Variablen:

Satz 3.33 (Dimensionssatz für Gleichungssysteme)

Die Dimension des Vektorraums der Lösungen eines homogenen linearen Gleichungssystems $\mathbf{A} \cdot \vec{x} = \vec{0}$ mit n Variablen (also $\mathbf{A} \in \mathbb{C}^{m \times n}$) ist $n - \text{Rang}(\mathbf{A})$.

Ist man also an einer eindeutigen Lösung interessiert, so muss ausschließlich $\vec{0}$ eine homogene Lösung sein, d. h. $n - \text{Rang}(\mathbf{A}) = n - \dim S(\mathbf{A}) = 0$. Der Spaltenraum kann aber nur dann die Dimension n haben, wenn die Matrix mindestens n Zeilen hat, d. h., wenn mindestens n Gleichungen existieren. Für n Unbekannte braucht man n Gleichungen.

Beispiel 3.62 (Schnittgerade zweier Ebenen)

Wir suchen die Schnittgerade zwischen zwei Ebenen im $\mathbb{R}^3$, die über die Gleichungen $\vec{n}_1 \cdot \vec{x} = d_1$ und $\vec{n}_2 \cdot \vec{x} = d_2$ gegeben sind. Dabei sind $\vec{n}_1 = (n_{1,1}, n_{1,2}, n_{1,3})$ und $\vec{n}_2 = (n_{2,1}, n_{2,2}, n_{2,3})$ Normalenvektoren der Ebenen. Wählt man die Normalenvektoren als Normaleneinheitsvektoren mit Länge eins und $d_1 > 0$, $d_2 > 0$, so sind die beiden Gleichungen Hesse'sche Normalformen (siehe Seite 463) und d_1 bzw. d_2 ist jeweils der Abstand der Ebene zum Nullpunkt. Zeigen die Normalenvektoren in unterschiedliche Richtungen, d. h., sind sie linear unabhängig, dann bilden alle Lösungen $\vec{x} = (x_1, x_2, x_3)$ beider Gleichungen eine Gerade. Eine Anwendung ist beispielsweise die Rekonstruktion von Dächern aus Laserscandaten (3-D-Punktwolke). Mit Methoden der Bildverarbeitung (z. B. Hough-Transformation) können hier Ebenen erkannt und die Hesse'schen Normalformen aufgestellt werden. Für die Berechnung der Ränder der Dachflächen werden dann die Schnittgeraden der Ebenen benötigt, vgl. Abbildung 3.31.

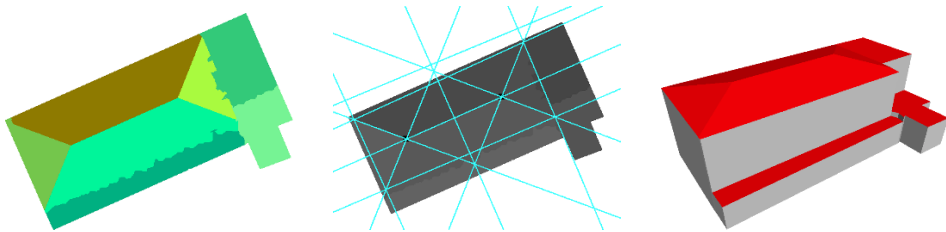


Abb. 3.31: Links sind die Punkte markiert, die Dachflächenebenen eines Gebäudes zugeordnet werden können. Daneben sind die Schnittgeraden und Geraden durch Stufen zwischen den Ebenen eingezeichnet. Daraus wird das rechts abgebildete 3-D-Modell berechnet.

Zur Bestimmung der Schnittgeraden müssen wir das unterbestimmte lineare Gleichungssystem

$$\begin{bmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ n_{2,1} & n_{2,2} & n_{2,3} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$$

lösen. Obwohl die Matrix nicht quadratisch ist, können wir die Lösung mit der Cramer'schen Regel (siehe Seite 197) berechnen. Die beiden Zeilen der Matrix sind laut Voraussetzung linear unabhängig, so dass der Rang der Matrix 2 ist. Wenn wir eine Gerade als Lösungsmenge erwarten, dann muss das laut Dimensionssatz auch genau so sein: Die Lösungsmenge ist ein Vektorraum der Dimension eins, also muss der Rang $3 - 1 = 2$ sein. Damit gibt es aber auch zwei linear unabhängige Spalten. Wir können nun drei Fälle unterscheiden: Die Spalten 1 und 2 sind linear unabhängig, 2 und 3 sind unabhängig oder 1 und 3 sind unabhängig. Wir betrachten jetzt nur den ersten Fall, die beiden anderen löst man völlig analog. Bei linear unabhängigen Spalten 1 und 2 wählen wir einmal die Variable $x_3 = 0$ und lösen dann das eindeutig lösbares Gleichungssystem

$$\mathbf{A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \quad \mathbf{A} := \begin{bmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{bmatrix},$$

für x_1 und x_2 mit der Cramer'schen Regel:

$$\begin{aligned} x_1 &= \det \begin{bmatrix} d_1 & n_{1,2} \\ d_2 & n_{2,2} \end{bmatrix} / \det \mathbf{A} = \frac{d_1 n_{2,2} - d_2 n_{1,2}}{n_{1,1} n_{2,2} - n_{2,1} n_{1,2}} \\ x_2 &= \det \begin{bmatrix} n_{1,1} & d_1 \\ n_{2,1} & d_2 \end{bmatrix} / \det \mathbf{A} = \frac{n_{1,1} d_2 - n_{2,1} d_1}{n_{1,1} n_{2,2} - n_{2,1} n_{1,2}}. \end{aligned}$$

Damit haben wir einen Punkt gefunden, der auf der gesuchten Gerade liegt. Mit einem weiteren Punkt ist die Gerade aber eindeutig bestimmt. Diesen erhalten wir, indem wir z. B. $x_3 = 1$ wählen und das resultierende Gleichungssystem

$$\mathbf{A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} d_1 - n_{1,3} \\ d_2 - n_{2,3} \end{pmatrix}$$

lösen. Die Cramer'sche Regel führt zu

$$\begin{aligned} x_1 &= \det \begin{bmatrix} d_1 - n_{1,3} & n_{1,2} \\ d_2 - n_{2,3} & n_{2,2} \end{bmatrix} / \det \mathbf{A} = \frac{(d_1 - n_{1,3}) n_{2,2} - (d_2 - n_{2,3}) n_{1,2}}{n_{1,1} n_{2,2} - n_{2,1} n_{1,2}} \\ x_2 &= \det \begin{bmatrix} n_{1,1} & d_1 - n_{1,3} \\ n_{2,1} & d_2 - n_{2,3} \end{bmatrix} / \det \mathbf{A} = \frac{n_{1,1} (d_2 - n_{2,3}) - n_{2,1} (d_1 - n_{1,3})}{n_{1,1} n_{2,2} - n_{2,1} n_{1,2}}. \end{aligned}$$

■

3.5.2 Berechnung von linearen elektrischen Netzwerken*

Dies ist ein Vertiefungskapitel, das sich damit beschäftigt, wie die Lineare Algebra (und Graphentheorie) bei der Berechnung von elektrischen Netzwerken mittels des Maschenstromverfahrens hilft. Insbesondere zeigt sich die Bedeutung der linearen Unabhängigkeit bei der Konstruktion ausreichend vieler Gleichungen.

3.5.2.1 Elektrische Netzwerke und Graphen

Im Einführungsbeispiel (1.45) zu linearen Gleichungssystemen auf Seite 153 (siehe Abbildung 1.59) haben wir die Ströme in einem elektrischen Netzwerk berechnet und dabei die Kirchhoff'schen Regeln verwendet. Auf diese Art können beliebige (lineare) Gleich- oder Wechselstromschaltungen analysiert werden. Um für große Netzwerke einen strukturierten Ansatz zu finden, der auch mit einem Computerprogramm implementiert werden kann, wird die Schaltung zunächst in einen Graphen überführt, aus dem dann das Gleichungssystem gewonnen wird.

In (1.45) nennen wir die Spannungen U_1 und U_2 um in U_4 und U_5 , damit wir Spannungen und Ströme in den Netzwerkzweigen konsistent durchnummerieren können. Das umbenannte Netzwerk ist in Abbildung 3.32 (links) dargestellt.

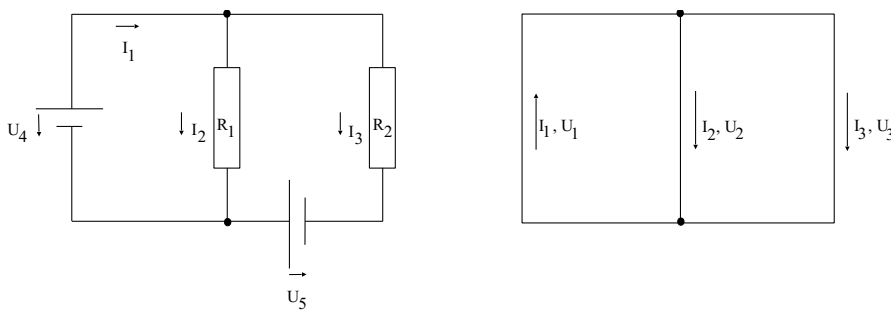


Abb. 3.32: Beispielnetzwerk mit zugehörigem Graphen

Definition 3.25 (Graph)

Ein (ungerichteter) Graph $G = (V, E)$ ist eine Struktur, die aus einer Menge V von **Knoten** (vertices) und einer Menge E von **Kanten** (edges) besteht, wobei jeder Kante $e \in E$ genau zwei Knoten v_1 und $v_2 \in V$ zugeordnet sind. Man sagt, die Kante e verbindet die Knoten v_1 und v_2 . Da wir ungerichtete Graphen betrachten, verbindet e sowohl v_1 mit v_2 als auch v_2 mit v_1 . In der Elektrotechnik wird statt von Kanten von **Zweigen** gesprochen. Ein **Teilgraph** von V ist ein Graph, dessen Knotenmenge eine Teilmenge von V und dessen Kantenmenge eine Teilmenge von E ist, so dass nur Knoten des Teilgraphen verbunden werden.

Wir werden im Folgenden nur endliche Graphen betrachten, d. h. Graphen mit einer endlichen Anzahl von Knoten und Zweigen. Üblicherweise werden Graphen mittels geometrischer Diagramme visualisiert. Dabei werden die Knoten durch Punkte und die Zweige durch Linien dargestellt. Eine kompakte Einführung in Graphen und Algorithmen für Graphen ist z. B. (Goebbels und Rethmann, 2014, Kap. 2).

Aus einem elektrischen Netzwerk erhält man einen Graphen, indem man die Knotenpunkte der Leiter als Knoten des Graphen interpretiert und die Verbindung zweier Knotenpunkte durch Zweipolelemente wie Widerstände, Kondensatoren, Spulen und Spannungsquellen als Zweige auffasst (siehe Abbildung 3.32). Kurzschlusszweige sind nicht erlaubt. Ideale Stromquellen werden nicht über Zweige dargestellt. Sie speisen einen Strom an einem Knoten ein, der an einem anderen wieder abfließt.

Definition 3.26 (Graphenbegriffe)

Sei $G = (V, E)$ ein (ungerichteter) Graph.

- Zwei Knoten $u, v \in V$ heißen genau dann **benachbart**, wenn sie durch einen Zweig verbunden sind.
- Eine **Zweigefolge** von $u_0 := u$ nach $u_n := v$ ist eine endliche Folge von Zweigen $e_1, e_2, \dots, e_n \in E$, so dass e_1 den Knoten u_0 mit einem Knoten $u_1 \in V$ verbindet, e_2 den Knoten u_1 mit einem Knoten $u_2 \in V$ verbindet usw. Schließlich möge e_n den Knoten u_{n-1} mit $u_n = v$ verbinden. Diese Zweigefolge schreiben wir kurz mittels der zugehörigen benachbarten Knoten $u_0, u_1, \dots, u_n$. Die Zweigefolge hat die **Länge** $n > 0$. u und v heißen **Endknoten** der Zweigefolge.
- Ist $u = v$, so heißt die Zweigefolge **geschlossen**, sonst heißt sie **offen**.
- Ein **Weg** ist eine offene Zweigefolge, bei der alle Knoten verschieden sind.
- Ein **Kreis** ist eine geschlossene Zweigefolge, bei der alle Knoten mit Ausnahme des Anfangs- oder des Endknotens voneinander verschieden sind.
- Ein Graph heißt genau dann **zusammenhängend**, wenn zu je zwei Knoten $u \neq v$, $u, v \in V$ ein Weg existiert, der u mit v verbindet.
- Ein (ungerichteter) zusammenhängender Graph ohne Kreise heißt ein **Baum** (vgl. Abbildung 3.34).

Achtung: In der Literatur gibt es unterschiedliche Definitionen für Wege (und Pfade). Hier ist genau auf die Definition des jeweiligen Autors zu achten.

Ein elektrisches Netzwerk wird zu einem zusammenhängenden Graphen. Eine **Masche** in einem elektrischen Netzwerk entspricht genau einem Kreis im zugehörigen Graphen. Wir verwenden daher den Begriff Masche als Synonym für Kreis.

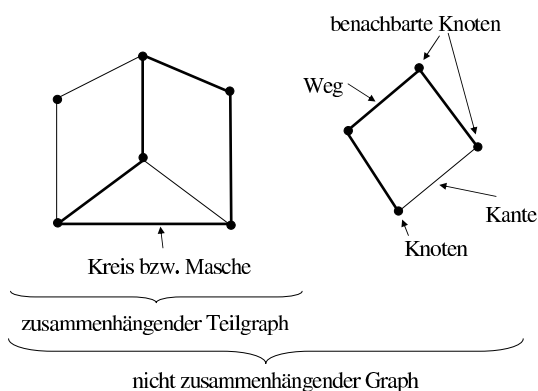


Abb. 3.33: Darstellung eines Graphen, Graph-Begriffe

Lemma 3.18 (Erhalt des Zusammenhangs)

Erzeugt man einen Graphen, indem man aus einer Masche eines zusammenhängenden Graphen einen Zweig entfernt, so ist auch der so reduzierte Graph zusammenhängend.

Beweis: Der entfernte Zweig möge die Knoten u und v verbinden. Da er Teil einer Masche ist, gibt es einen Weg W zwischen u und v und damit auch zwischen v und u , der den Zweig nicht verwendet. Jeder Weg, der im Ausgangsgraphen zwei Knoten verbindet, wird zu einem Weg im reduzierten Graphen, indem man die Zweige zwischen u und v durch die Zweigefolge von W ersetzt und dann solche Zweige eliminiert, die dazu führen, dass ein Knoten mehrfach durchlaufen wird. Da der Ausgangsgraph zusammenhängend ist, gilt dies damit auch für den reduzierten Graphen. $\square$

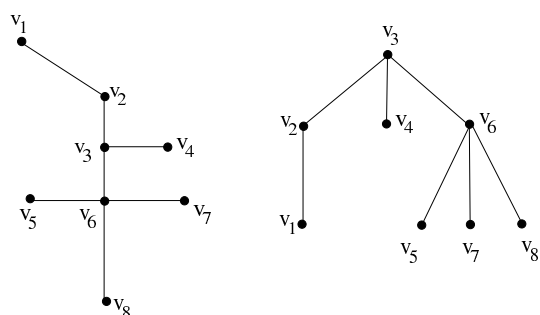


Abb. 3.34: Zwei Darstellungen eines Baums

Man kann die Knoten eines Baums so anordnen, dass der Graph tatsächlich wie ein Baum mit einer **Wurzel** (hier ist nicht die Rechenoperation gemeint!) und **Blättern** aussieht (siehe Abbildung 3.34). Jeder Knoten kann als Wurzel gewählt werden. Im Gegensatz zu einem realen Baum wird üblicherweise die Wurzel ganz oben eingezeichnet. Eine Reihe tiefer werden die Knoten angeordnet, die direkt mit der Wurzel verbunden

sind. In der nächsten Reihe die Knoten, die mit Knoten der vorangehenden Reihe verbunden sind. Da es keine Maschen gibt, können diese nur mit genau einem Knoten der Vorgängerreihe verbunden sein. Gäbe es zwei Zweige zu Knoten der Vorgängerreihe, gäbe es eine Masche über einen gemeinsamen Vorgänger dieser Knoten. Auch können keine Knoten innerhalb einer Reihe über eine Kante verbunden sein, da es sonst eine Masche gäbe. Setzt man diese Konstruktion fort, erreicht man schließlich alle Knoten, da der Graph zusammenhängend ist.

Gibt es zu einem Knoten keinen Folgeknoten in der nächsten Reihe, heißt er ein **Blatt**. In Abbildung 3.34 sind v_1, v_4, v_5, v_7 und v_8 Blätter. In einer anderen Darstellung des Baumes könnte eines dieser Blätter aber auch die Wurzel sein, die in der Elektrotechnik als **Masse** bezeichnet wird. Die Masse kann damit frei gewählt werden.

Das entscheidende Charakteristikum dieser Baumdarstellung ist, dass jeder Knoten mit Ausnahme der Wurzel genau einen Vorgänger hat.

Lemma 3.19 (Anzahl der Zweige eines Baums)

Ein Baum mit n Knoten hat genau $n - 1$ Zweige.

Beweis: Wir zählen die Zweige in der zuvor beschriebenen Baumdarstellung: Jeder der $n - 1$ Knoten ohne die Wurzel hat genau einen Zweig zum Vorgängerknoten. Dies sind genau $n - 1$ Zweige. $\square$

Lemma 3.20 (Spannender Baum)

Zu jedem zusammenhängenden Graphen mit n Knoten gibt es einen Baum mit n Knoten, der Teilgraph ist.

Beweis: Wie bei der obigen Darstellung des Baums beginnt man mit einem Knoten und konstruiert zeilenweise den Baum. Gibt es zu einem Knoten mehr als einen Zweig zur vorangehenden Reihe, so lässt man dabei einfach alle bis auf einen Zweig weg. Ebenso lässt man alle Zweige weg, die Knoten innerhalb einer Reihe miteinander verbinden. $\square$

Einen solchen Teilgraphen nennen wir **maximalen Baum** oder **Gerüst** des Ausgangsgraphen. Die Zweige des Ausgangsgraphen, die nicht zu diesem Baum gehören, heißen **Verbindungszweige** bezüglich dieses Baums.

3.5.2.2 Maschengleichungen

Zunächst wird der Begriff „Lineare Unabhängigkeit“ in einem neuen Zusammenhang außerhalb von Vektorräumen definiert. Wir werden später sehen, dass sich daraus linear unabhängige Zeilenvektoren der Matrix eines Gleichungssystems ergeben.

Definition 3.27 (Lineare Unabhängigkeit von Maschen)

Eine Menge von Maschen (Kreisen) eines endlichen Graphen heißt genau dann **linear unabhängig**, wenn die Maschen in eine Reihenfolge gebracht werden können, so dass jede Masche mindestens einen Zweig enthält, der in den vorangehenden Maschen nicht enthalten ist.

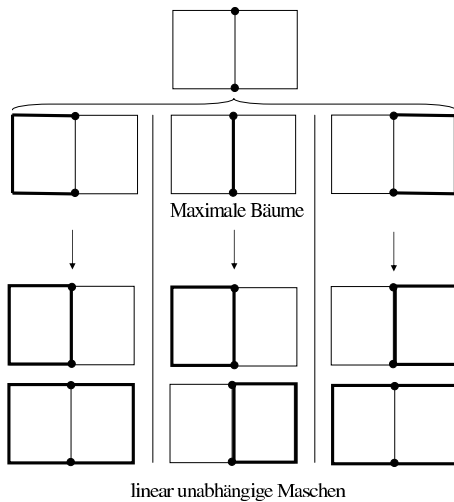


Abb. 3.35: Maximale Bäume und daraus gewonnene linear unabhängige Maschen

Satz 3.34 (Anzahl unabhängiger Maschen)

Ein zusammenhängender Graph mit n Knoten und m Zweigen besitzt $m - n + 1$ linear unabhängige Maschen, aber nicht mehr.

- Beweis:** a) Wir konstruieren zunächst $m - n + 1$ linear unabhängige Maschen: Der Graph besitzt einen maximalen Baum mit $n - 1$ Zweigen und keine Maschen. Jede Masche des Ausgangsgraphen muss daher einen Zweig besitzen, der nicht zum Baum gehört, also einen Verbindungszweig. Umgekehrt ist jeder der $m - n + 1$ Verbindungszweige Teil einer Masche, da die Knoten sowohl durch den Zweig als auch durch den Baum verbunden sind. Wir erhalten $m - n + 1$ linear unabhängige Maschen, indem wir jeweils genau einen Verbindungszweig und Zweige des Baums auswählen. Der Verbindungszweig ist nur in dieser Masche und sorgt damit für die lineare Unabhängigkeit. Es gibt also $m - n + 1$ linear unabhängige Maschen.
- b) Jetzt zeigen wir, dass es nicht mehr als $m - n + 1$ linear unabhängige Maschen geben kann. Wir nehmen an, es gäbe $k > m - n + 1$ linear unabhängige Maschen. Die Letzte besitzt wegen der linearen Unabhängigkeit einen Zweig, der in den anderen nicht vorkommt. Entfernt man diesen Zweig, bleibt der verbleibende Graph zusammenhängend (Lemma 3.18). Der Restgraph hat außerdem $k - 1$ linear unabhängige

Maschen, da in diesen der entfernte Zweig nicht vorkommt. Jetzt entfernt man iteriert aus jeder Masche einen Zweig, bis man schließlich einen zusammenhängenden Graphen mit $m - k < m - (m - n + 1) = n - 1$ Zweigen hat. Da er n Knoten hat, müsste er aber $n - 1$ Zweige haben, um zusammenhängend zu sein. Wegen des Widerspruches kann es nicht mehr als $m - n + 1$ linear unabhängige Maschen geben. $\square$

Im Teil a) des Beweises haben wir ein Verfahren kennengelernt, um eine maximale Menge linear unabhängiger Maschen zu finden: Man wählt einen maximalen Baum aus. Jeder Verbindungsweig zu diesem Baum liefert dann zusammen mit Zweigen des Baums eine Masche, so dass die Menge aller dieser Maschen linear unabhängig ist. In Abbildung 3.35 sind die linear unabhängigen Maschen zu allen maximalen Bäumen des Beispiels angegeben.

Wendet man nun auf ein elektrisches Netzwerk die Maschenregel auf linear unabhängige Maschen an, so erhält man bei n Knoten und m Zweigen $m - n + 1$ homogene lineare Gleichungen für m Spannungen zwischen den Knoten. In Matrixschreibweise ergeben die Gleichungen linear unabhängige Zeilenvektoren, da man keine als Linearkombination der anderen schreiben kann: Da jeder Verbindungsweig in genau einer Masche vorkommt, kommt jede Spannung eines Verbindungszeigs als Variable nur in genau einer Gleichung vor. Die Matrix des zugehörigen Gleichungssystems hat den Rang $m - n + 1$. Dies ist der Grund, warum man die Maschen als linear unabhängig bezeichnet.

Im Beispiel erhält man so für die beiden linear unabhängigen Maschen des mittleren maximalen Baums in Abbildung 3.35 die beiden Gleichungen

$$\begin{array}{rcl} U_1 & +U_2 & = 0 \\ -U_2 & +U_3 & = 0. \end{array}$$

3.5.2.3 Knotengleichungen

Wendet man die Knotenregel auf alle Knoten an, erhält man n Gleichungen für die m Ströme in den Zweigen. An jedem Knoten ist die Summe der Ströme gleich null. Dabei werden zum Knoten fließende Ströme positiv und wegfließende negativ gewertet. Hinzu kommen die ebenfalls orientierten Ströme von Stromquellen. Da jeder Zweig genau zwei Knoten verbindet, wird jeder Zweigstrom in einem Knoten positiv und in einem zweiten negativ addiert. Ebenso tritt jeder Strom einer Stromquelle bei einem Knoten positiv und bei einem negativ auf. Summiert man alle Gleichungen, so ergibt sich $0 = 0$. Damit führen höchstens $n - 1$ dieser Gleichungen zu linear unabhängigen Zeilen in der Matrix des Gleichungssystems.

Wählt man $n - 1$ beliebige Knoten aus, so sind die Zeilenvektoren der zugehörigen Gleichungen linear unabhängig. Das sieht man, indem man einen (existierenden) maxi-

malen Baum auswählt und als Wurzel den nicht ausgewählten Knoten nimmt. Zeichnet man den Baum in Baumstruktur mit dieser Wurzel an der Spitze und durchläuft alle Knoten von Ebene zu Ebene von den Blättern zur Wurzel, dann kommt mit jeder Knotengleichung ein Strom zu dem Zweig neu zum Gleichungssystem dazu, der den Knoten mit dem übergeordneten Knoten der nächsten Ebene verbindet. Alle Knoten mit Ausnahme der Wurzel liefern also sukzessive mindestens eine neue Stromvariable für das Gleichungssystem.

Das Gleichungssystem ist bei fehlenden Stromquellen homogen, anderenfalls treten die Ströme der Stromquellen als Inhomogenität auf. Im Beispiel findet man eine Gleichung für die Ströme:

$$I_1 - I_2 - I_3 = 0.$$

Insgesamt hat man so ein Gleichungssystem mit $m - n + 1 + n - 1 = m$ linear unabhängigen Zeilen (Gleichungen) für die gesuchten m Ströme und m Spannungen. Im Beispiel:

$$\begin{array}{rcl} U_1 + U_2 & & = 0 \\ -U_2 + U_3 & & = 0 \\ I_1 - I_2 - I_3 & = & 0. \end{array}$$

3.5.2.4 Gleichungen zwischen Spannungen und Strömen

Bislang wurden die Zweipole, die die Zweige bilden, nicht berücksichtigt. Für jeden Zweig ist die Stromstärke an allen Bauteilen gleich, die Spannungen addieren sich. Zudem können an Widerständen, Kondensatoren und Spulen Spannungen und Ströme mittels der Widerstände bzw. Impedanzen ineinander umgerechnet werden. Als Inhomogenitäten kommen die Spannungen der Spannungsquellen hinzu.

Für jeden der k Zweige mit mindestens einem komplexen Widerstand ($k \leq m$) kommt so eine (evtl. inhomogene) Gleichung zwischen der Spannung der beiden Knoten des Zweigs und dem Strom des Zweigs hinzu. Die zuvor aus den Kirchhoff'schen Regeln gewonnenen m Gleichungen enthalten entweder nur Ströme oder nur Spannungen. Jede der hinzukommenden Gleichungen betrifft einen anderen Strom bzw. eine andere Spannung. Damit erhalten wir ein Gleichungssystem mit Rang $m + k$. Gibt es keinen Zweig, der nur aus einer idealen Spannungsquelle besteht, so sind alle Ströme und Spannungen eindeutig über das Gleichungssystem mit Rang $2m$ bestimmt, das nun z. B. mittels des Gauß'schen Eliminationsverfahrens gelöst werden kann.

Besitzen Zweige lediglich eine ideale Spannungsquelle und keine weiteren Zweipole, so ist die Situation nicht so eindeutig. Hier ist die Spannung zwischen den Knoten des Zweigs direkt durch eine Inhomogenität bestimmt. Parallelschaltung von zwei Spannungsquellen kann daher dazu führen, dass sich der Rang durch diese Gleichung nicht erhöht, so dass das (inhomogene) Gleichungssystem ggf. sogar unlösbar wird. Bei realen Spannungsquellen tritt dieses Problem nicht auf. Sie können als ideale Spannungsquelle

mit in Reihe geschaltetem Widerstand aufgefasst werden, so dass das Gleichungssystem den vollen Rang $2m$ hat (s. o.). Im Beispiel kommen so die drei Gleichungen

$$\begin{array}{rcl} U_1 & & = -U_4 \\ U_2 & -R_1 I_2 & = 0 \\ U_3 & -R_2 I_3 & = -U_5 \end{array}$$

hinzu und führen zum vollständigen Gleichungssystem

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -R_1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -R_2 \end{bmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -U_4 \\ 0 \\ -U_5 \end{pmatrix},$$

dessen eindeutige Lösung man nach einigen Gauß-Umformungen ablesen kann:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} -U_4 \\ U_4 \\ U_4 \\ \frac{U_4}{R_1} + \frac{U_4+U_5}{R_2} \\ \frac{U_4}{R_1} \\ \frac{U_4+U_5}{R_2} \end{pmatrix}.$$

Insgesamt haben wir eine Variante des Maschenstromverfahrens zur Berechnung aller Ströme und Spannungen beschrieben:

- a) Überführe das elektrische Netzwerk in einen Graphen.
- b) Wähle einen maximalen Baum aus.
- c) Stelle die Maschengleichungen über die Verbindungszweige des Graphen auf.
- d) Stelle die Knotengleichungen auf.
- e) Komplettiere das Gleichungssystem über die Beziehungen zwischen Strömen und Spannungen.
- f) Löse das Gleichungssystem z.B. mittels des Gauß-Verfahrens oder der Cramer'schen Regel.

Als weiteren Lösungsalgorithmus, der auf den Kirchhoff'schen Gesetzen beruht, verwendet man das Knotenpotenzialverfahren. Dieses wird hauptsächlich bei der numerischen Berechnung von Netzwerken mit Simulatoren eingesetzt, da man keinen maximalen Baum finden muss und häufig an Potenzialen und nicht an Strömen interessiert

ist. Beim Knotenpotenzialverfahren berechnet man die Potenziale $\varphi_1, \dots, \varphi_n$ der n Netzwerknoten. Genauer kann man ein Potenzial beliebig festlegen (Erde) und muss nur $n - 1$ Potenziale als Spannungsdifferenz zur Erde berechnen. Dies ist genau die Anzahl der linear unabhängigen Zeilen der Knotengleichungen. Die Spannung zwischen zwei Knoten v_i und v_k eines Zweigs ist die Potentialdifferenz $\varphi_i - \varphi_k$ der Knoten. Der Strom des Zweigs in Richtung von v_i nach v_k ist das Produkt dieser Differenz $\varphi_i - \varphi_k$ mit dem komplexen Widerstand des Zweigs. Eingesetzt für die Ströme in den $n - 1$ Knotengleichungen (Spannungsquellen werden zuvor durch Stromquellen ersetzt) erhält man $n - 1$ Gleichungen für die gesuchten $n - 1$ Potenziale.

3.6 Eigenwerte und Eigenvektoren

In der Elektrotechnik, bei mechanischen Schwingungen, bei chemischen Reaktionen sowie bei Aufgabenstellungen aus den Wirtschaftswissenschaften sucht man oft nach Gleichgewichtszuständen. Durch die Modellierung des Problems ist dabei eine Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ gegeben, und man sucht einen Vektor $\vec{d} \in \mathbb{R}^n$, den die Matrix bis auf ein skalares Vielfaches nicht ändert, d. h., die beiden Vektoren $\vec{d}$ und $\mathbf{A} \cdot \vec{d}$ sollen in $\mathbb{R}^n$ parallel liegen. In diesem Kapitel sehen wir, wie man diese Vektoren findet.

Beispiel 3.63

Über Verkehrszählungen wurde ermittelt, dass 80 % der Pendler, die mit öffentlichen Verkehrsmitteln ihre Arbeitsstätte erreichen, auch im nächsten Jahr wieder mit öffentlichen Verkehrsmitteln fahren werden. 20 % wollen allerdings auf das Auto umsteigen. Von den Autofahren wollen auch im nächsten Jahr 60 % dem Auto treu bleiben, dagegen wollen 40 % auf öffentliche Verkehrsmittel wechseln. Ist a_n die Zahl der Autofahrer im Jahr n , und steht o_n für die Nutzerzahl des öffentlichen Nahverkehrs, dann kann dieses Verhalten als Matrixmultiplikation

$$\begin{pmatrix} a_{n+1} \\ o_{n+1} \end{pmatrix} = \underbrace{\begin{bmatrix} 0,6 & 0,2 \\ 0,4 & 0,8 \end{bmatrix}}_{\mathbf{A}:=} \begin{pmatrix} a_n \\ o_n \end{pmatrix} \quad (3.27)$$

dargestellt werden. Verkehrsplaner suchen nach einem **stationären Zustand**, für den sich die Nutzerzahlen nicht verändern, also nach einem Vektor, der durch Multiplikation mit der Matrix $\mathbf{A}$ auf sich selbst abgebildet wird. Dies ist ein Fixpunkt der so gegebenen Abbildung (siehe Definition 2.28 auf Seite 318). Als lineares Gleichungssystem geschrieben, wird nach nicht-trivialen Lösungen von $\mathbf{A} \cdot \vec{d} = \vec{d}$ gesucht. $\mathbf{A} \cdot \vec{d}$ und $\vec{d}$ sollen hier also nicht nur parallel, sondern sogar gleich sein. ■

3.6.1 Eigenwerte und Eigenvektoren

Definition 3.28 (Eigenvektor, Eigenwert)

Ein Vektor $\vec{d} \neq \vec{0}$ aus $\mathbb{R}^n$ (oder $\mathbb{C}^n$) heißt **Eigenvektor** zum **Eigenwert** $s \in \mathbb{C}$ der $(n \times n)$ -Matrix **A** genau dann, wenn

$$\mathbf{A} \cdot \vec{d} = s \vec{d}.$$

Im Eingangsbeispiel sind also Eigenvektoren zu einem Eigenwert 1 gesucht. Eigenwerte und ihre Eigenvektoren können viel über die Struktur einer Matrix preisgeben (das sehen wir uns später z. B. mit Satz 3.39 an). Die Matrix bildet einen Eigenvektor auf ein skalares Vielfaches seiner selbst ab. Dies ist eine Art von Resonanz. Bei linearen Differenzialgleichungssystemen, die z. B. bei der Berechnung von Schaltungen auftreten, erhält man Lösungen über Eigenwerte und Eigenvektoren (siehe Kapitel 5.3). Auch werden Eigenvektoren in der Statistik eingesetzt (siehe Abbildung 7.6 auf Seite 991).

Beispiel 3.64

a) Wir zeigen, dass $\vec{d} = (1, 1)^\top$ ein Eigenvektor der Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -3 & 5 \end{bmatrix}$$

zum Eigenwert $s = 2$ ist. Es gilt

$$\mathbf{A} \cdot \vec{d} = \begin{bmatrix} 1 & 1 \\ -3 & 5 \end{bmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} = 2\vec{d}.$$

b) In Abbildung 3.36 sehen wir Eigenvektoren, die bei der Bildmanipulation auftreten. Eine Spiegelung an der y -Achse entspricht einer Multiplikation der Vektoren (bzw. Punkte) mit der Matrix $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ (oben rechts). $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$ bewirkt eine zusätzliche Spiegelung an der x -Achse (unten links). Schließlich ist unten rechts auch noch eine Drehung (Multiplikation mit (1.48)) dargestellt.

c) Hier ist $(j, 1)^\top$ ein Eigenvektor zum Eigenwert $1 - j$:

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{pmatrix} j \\ 1 \end{pmatrix} = (1 - j) \begin{pmatrix} j \\ 1 \end{pmatrix}.$$

■

Wir haben in der Definition den Nullvektor als Eigenvektor ausgeschlossen, da so einige der folgenden Aussagen leichter zu formulieren sind. Denn jede Zahl wäre Eigenwert zum Nullvektor, während sonst zu einem Eigenvektor der Eigenwert eindeutig bestimmt ist. Dagegen ist der Eigenwert 0 erlaubt.

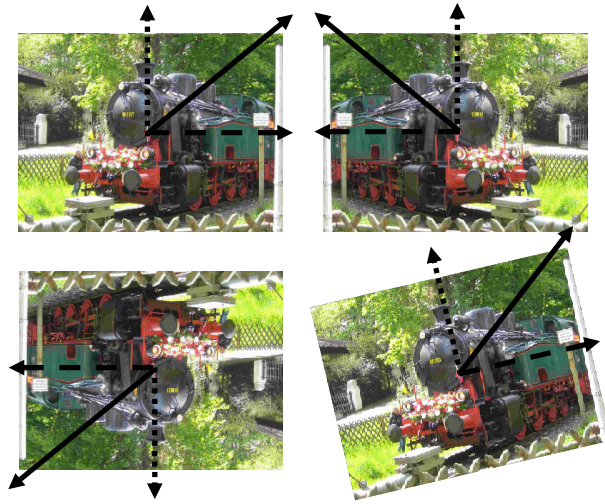


Abb. 3.36: Eigenvektoren zu Spiegelungen und Drehungen: Spiegelt man das Ausgangsbild (links oben) an der y -Achse, so ist der Vektor $(0,1)^\top$ ein Eigenvektor zum Eigenwert $s = 1$, der Vektor $(1,0)^\top$ ein Eigenvektor zum Eigenwert $s = -1$. Der Vektor $(1,1)^\top$ ist kein Eigenvektor (rechts oben). Spiegelt man an der x - und an der y -Achse (unten links), so sind alle Vektoren Eigenvektoren zum Eigenwert $s = -1$. Dreht man wie rechts unten abgebildet um den Mittelpunkt um einen Winkel, der kein Vielfaches von π ist, so ist kein Vektor ein Eigenvektor.

Ein Eigenvektor $\vec{d}$ zum Eigenwert 1 erfüllt $\mathbf{A} \cdot \vec{d} = \vec{d}$ und ist damit analog zu Definition 2.28 ein Fixpunkt der Abbildung $\vec{f}: \mathbb{C}^n \rightarrow \mathbb{C}^n, \vec{d} \mapsto \mathbf{A} \cdot \vec{d}$ (vgl. Abbildung 3.36). Wie beim Pendlerproblem sucht man bei Fixpunktverfahren einen stationären Zustand als Lösung, so z. B. bei der Nullstellenberechnung mit dem Newton-Verfahren auf Seite 306 oder bei der iterativen Lösung linearer Gleichungssysteme (siehe Kasten auf Seite 179). Wie berechnet man nun die Eigenwerte einer Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$? Die Gleichung $\mathbf{A} \cdot \vec{d} = s \vec{d}$ ist äquivalent zu $\mathbf{A} \cdot \vec{d} = s \mathbf{E} \cdot \vec{d}$ bzw.

$$(\mathbf{A} - s \mathbf{E}) \cdot \vec{d} = \vec{0}.$$

Damit dieses homogene lineare Gleichungssystem nicht-triviale Lösungen besitzt, muss die Determinante der Koeffizientenmatrix $\mathbf{A} - s \mathbf{E}$ verschwinden,

$$\det(\mathbf{A} - s \mathbf{E}) = 0.$$

Definition 3.29 (Charakteristisches Polynom)

Sei $\mathbf{A}$ eine $(n \times n)$ -Matrix und $\mathbf{E}$ die $(n \times n)$ -Einheitsmatrix. $p(s) = \det(\mathbf{A} - s \mathbf{E})$ ist ein Polynom der Variable $s \in \mathbb{C}$. Es wird **charakteristisches Polynom** der Matrix

$\mathbf{A}$ genannt. Der Grad von p ist gleich der Anzahl n der Zeilen und der Spalten von $\mathbf{A}$. Die Gleichung

$$p(s) = \det(\mathbf{A} - s\mathbf{E}) = 0 \quad (3.28)$$

heißt **charakteristische Gleichung**.

Die Nullstellen des charakteristischen Polynoms $p(s)$ sind die Eigenwerte von $\mathbf{A}$:

Lemma 3.21 (Berechnung von Eigenwerten)

Die Eigenwerte s sind genau die Lösungen der charakteristischen Gleichung

$$\det(\mathbf{A} - s\mathbf{E}) = \begin{vmatrix} a_{1,1} - s & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} - s & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} - s \end{vmatrix} = 0.$$

Beweis: Die Zahl s ist ein Eigenwert genau dann, wenn $(\mathbf{A} - s\mathbf{E}) \cdot \vec{d} = \vec{0}$ für ein $\vec{d} \neq \vec{0}$. Das ist äquivalent mit der Existenz einer nicht-trivialen Linearkombination der Spalten von $\mathbf{A} - s\mathbf{E}$, die $\vec{0}$ ergibt, was wiederum äquivalent zur linearen Abhängigkeit der Spalten ist. Wegen Folgerung 1.7 ist diese äquivalent mit $\det(\mathbf{A} - s\mathbf{E}) = 0$. $\square$

Beispiel 3.65

Die Matrix $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ besitzt das charakteristische Polynom

$$\det(\mathbf{A} - s\mathbf{E}) = \begin{vmatrix} 1 - s & 1 \\ -1 & 1 - s \end{vmatrix} = (1 - s)^2 + 1 = s^2 - 2s + 2$$

mit den Nullstellen $s_{1,2} = 1 \pm j$. ■

Da die charakteristische Gleichung $p(s) = 0$ nach dem Fundamentalsatz der Algebra (siehe Satz 1.24 auf Seite 147) höchstens n verschiedene (komplexe) Nullstellen hat, besitzt eine $(n \times n)$ -Matrix höchstens n Eigenwerte.

Eigenvektoren $\vec{d}$ zum Eigenwert s erhält man wegen $(\mathbf{A} - s\mathbf{E}) \cdot \vec{d} = \vec{0}$ als nicht-triviale Lösungen $\vec{d} \neq \vec{0}$ dieses Gleichungssystems. Der Nullvektor ist zwar auch immer Lösung eines homogenen linearen Gleichungssystems, ist aber wie bereits betont als Eigenvektor laut Definition explizit ausgeschlossen. Die Menge der Eigenvektoren zum Eigenwert s zuzüglich $\vec{0}$ ist als Lösungsmenge von $(\mathbf{A} - s\mathbf{E}) \cdot \vec{d} = \vec{0}$ ein Unterraum von $\mathbb{R}^n$ bzw. $\mathbb{C}^n$. Man nennt diesen Unterraum den **Eigenraum** zu s .

Um Eigenwertaufgaben zu lösen kann, daher wie folgt vorgegangen werden:

- a) Bestimme alle Nullstellen des charakteristischen Polynoms $\det(\mathbf{A} - s\mathbf{E})$. Diese sind die Eigenwerte der Matrix $\mathbf{A}$.
- b) Für jeden Eigenwert s müssen alle nicht-trivialen Lösungen des homogenen linearen Gleichungssystems $(\mathbf{A} - s\mathbf{E}) \cdot \vec{d} = \vec{0}$ gefunden werden. Diese sind dann die Eigenvektoren von $\mathbf{A}$ zum Eigenwert s .

Beispiel 3.66

- a) Wir bestimmen die Eigenwerte und Eigenvektoren von $\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}$. Die charakteristische Gleichung lautet

$$\det(\mathbf{A} - s\mathbf{E}) = \begin{vmatrix} 1-s & 3 \\ 4 & 2-s \end{vmatrix} = (1-s)(2-s) - 12 = s^2 - 3s - 10 = 0.$$

Dies ist eine quadratische Gleichung für s mit den Lösungen $s_1 = -2$ und $s_2 = 5$. Wir berechnen nun die Eigenvektoren zu den Eigenwerten.

Für $s_1 = -2$ erhält man das homogene lineare Gleichungssystem

$$\begin{bmatrix} 1+2 & 3 \\ 4 & 2+2 \end{bmatrix} \vec{d} = \vec{0} \iff \begin{bmatrix} 3 & 3 & | & 0 \\ 4 & 4 & | & 0 \end{bmatrix} \iff \begin{bmatrix} 1 & 1 & | & 0 \\ 0 & 0 & | & 0 \end{bmatrix},$$

d. h., ein Eigenvektor zum Eigenwert $s_1 = -2$ ist $\vec{d}_1 = (1, -1)^\top$. Auch ein beliebiges skalares Vielfaches von $\vec{d}_1$ ist Eigenvektor von $\mathbf{A}$ zu s_1 . Man wählt $\vec{d}_1$ so einfach wie möglich.

Für $s_2 = 5$ erhält man das homogene lineare Gleichungssystem

$$\begin{bmatrix} 1-5 & 3 \\ 4 & 2-5 \end{bmatrix} \vec{d} = \vec{0} \iff \begin{bmatrix} -4 & 3 & | & 0 \\ 4 & -3 & | & 0 \end{bmatrix} \iff \begin{bmatrix} -4 & 3 & | & 0 \\ 0 & 0 & | & 0 \end{bmatrix},$$

d. h., ein Eigenvektor zum Eigenwert $s_2 = 5$ ist $\vec{d}_2 = (3, 4)^\top$.

- b) Für das Pendlerproblem (3.27), mit dem wir das Kapitel begonnen haben, wird nach Eigenwerten der Matrix

$$\mathbf{A} = \begin{bmatrix} 0,6 & 0,2 \\ 0,4 & 0,8 \end{bmatrix}$$

gesucht. Die Eigenwerte sind Lösungen der charakteristischen Gleichung

$$\begin{aligned} p(s) = \det(\mathbf{A} - s\mathbf{E}) &= \begin{vmatrix} 0,6-s & 0,2 \\ 0,4 & 0,8-s \end{vmatrix} = s^2 - 1,4s + 0,4 \\ &= (s-1)(s-0,4) = 0, \end{aligned}$$

und als Nullstellen von $p(s)$ erhält man $s = 1$ und $s = 0,4$. Wir haben damit Glück (vgl. Aufgabe 3.22), dass $s = 1$ tatsächlich Eigenwert der Matrix ist und wir so einen Eigenvektor bestimmen können. Dazu lösen wir das homogene lineare System

$$\left[\begin{array}{cc|c} -0,4 & 0,2 & 0 \\ 0,4 & -0,2 & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} -0,4 & 0,2 & 0 \\ 0 & 0 & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} 2 & -1 & 0 \\ 0 & 0 & 0 \end{array} \right].$$

Wir erhalten einen Eigenvektor $\vec{d}_1 = (1, 2)^\top$. Alle Eigenvektoren zum Eigenwert 1 haben die Gestalt $r \cdot \vec{d}_1$, $r \in \mathbb{R} \setminus \{0\}$.

Der Eigenwert $s = 0,4$ wird für die Lösung des Pendlerproblems nicht benötigt. Dennoch bestimmen wir zur Übung auch einen Eigenvektor zu $s = 0,4$:

$$\left[\begin{array}{cc|c} 0,2 & 0,2 & 0 \\ 0,4 & 0,4 & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array} \right].$$

Wir lesen einen Eigenvektor $\vec{d}_2 = (1, -1)^\top$ ab. Alle Eigenvektoren zum Eigenwert 0,4 sind $r \cdot \vec{d}_2$, $r \in \mathbb{R} \setminus \{0\}$.

Der Eigenvektor $\vec{d}_1 = (1, 2)^\top$ bedeutet für das Pendlerproblem, dass im stationären Zustand $\frac{1}{3}$ der Pendler mit dem Auto und $\frac{2}{3}$ mit öffentlichen Verkehrsmitteln fahren.

- c) Wir bestimmen die Eigenwerte der (orthogonalen) Matrix, die einen Punkt im $\mathbb{R}^2$ um den Winkel φ um den Ursprung dreht:

$$\begin{aligned} \det \begin{bmatrix} \cos(\varphi) - s & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) - s \end{bmatrix} &= [\cos(\varphi) - s]^2 + \sin^2(\varphi) \\ &= \underbrace{\cos^2(\varphi) + \sin^2(\varphi)}_{=1} - 2s \cos(\varphi) + s^2. \end{aligned}$$

Also sind die Eigenwerte $s = \cos(\varphi) \pm \sqrt{\cos^2(\varphi) - 1}$. Einen reellen Eigenwert gibt es nur für $\varphi = k\pi$, $k \in \mathbb{Z}$, also wenn man um das Vielfache von 180 Grad dreht. Dieser ist dann entweder 1 (falls $\varphi = 2k\pi$) oder -1 (falls $\varphi = (2k+1)\pi$). Jeder

Vektor $\vec{d} \in \mathbb{R}^2$ mit $\vec{d} \neq \vec{0}$ ist dann Eigenvektor, da $\mathbf{A} - s\mathbf{E} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. ■

Die Eigenvektoren zum selben Eigenwert bilden zusammen mit dem Nullvektor den Eigenraum genannten Unterraum. Eigenräume zu verschiedenen Eigenwerten $s \neq t$ können nur den Nullvektor gemeinsam haben: Aus $\mathbf{A} \cdot \vec{d} = s\vec{d} = t\vec{d}$ folgt $(s-t)\vec{d} = \mathbf{A} \cdot \vec{d} - \mathbf{A} \cdot \vec{d} = \vec{0}$ und somit $\vec{d} = \vec{0}$. Darüber hinaus gilt sogar:

Lemma 3.22 (Lineare Unabhängigkeit von Eigenvektoren)

Eigenvektoren zu unterschiedlichen Eigenwerten sind linear unabhängig.

Dieses Lemma gilt nur deshalb, weil wir in der Definition den Nullvektor als Eigenvektor ausgeschlossen haben:

Beweis: Seien $\vec{d}_1, \dots, \vec{d}_n$ Eigenvektoren zu unterschiedlichen Eigenwerten. Annahme: Diese Vektoren sind linear abhängig. Wir führen diese Annahme zum Widerspruch. Lineare Abhängigkeit bedeutet, dass es Skalare $r_1, \dots, r_n$ gibt, die nicht alle gleich 0 sind, so dass

$$r_1 \vec{d}_1 + r_2 \vec{d}_2 + \dots + r_n \vec{d}_n = \vec{0};$$

o. B. d. A. sei $r_1 \neq 0$, dann ist

$$\vec{d}_1 = -\frac{r_2}{r_1} \vec{d}_2 - \dots - \frac{r_n}{r_1} \vec{d}_n. \quad (3.29)$$

Da $\vec{d}_1$ als Eigenvektor ungleich $\vec{0}$ ist, sind nicht alle $\frac{r_k}{r_1} = 0$. Wir wenden nun $\mathbf{A}$ auf die Gleichung an und erhalten

$$\begin{aligned} \mathbf{A} \cdot \vec{d}_1 &= -\frac{r_2}{r_1} \mathbf{A} \cdot \vec{d}_2 - \dots - \frac{r_n}{r_1} \mathbf{A} \cdot \vec{d}_n \\ \implies s_1 \vec{d}_1 &= -\frac{r_2 s_2}{r_1} \vec{d}_2 - \dots - \frac{r_n s_n}{r_1} \vec{d}_n. \end{aligned}$$

■ Falls $s_1 \neq 0$ ist, folgt weiter:

$$\vec{d}_1 = -\frac{r_2 s_2}{r_1 s_1} \vec{d}_2 - \dots - \frac{r_n s_n}{r_1 s_1} \vec{d}_n.$$

Subtrahieren wir nun von (3.29) diese Gleichung, so erhalten wir

$$\vec{0} = \frac{r_2}{r_1} \left(\frac{s_2}{s_1} - 1 \right) \vec{d}_2 + \dots + \frac{r_n}{r_1} \left(\frac{s_n}{s_1} - 1 \right) \vec{d}_n.$$

Da die Eigenwerte unterschiedlich sind, sind die Faktoren $\left(\frac{s_k}{s_1} - 1 \right)$ ungleich 0.

■ Ist $s_1 = 0$, so müssen alle anderen Eigenwerte $s_2, \dots, s_n$ von 0 verschieden sein (unterschiedliche Eigenwerte), und es ist

$$\vec{0} = -\frac{r_2 s_2}{r_1} \vec{d}_2 - \dots - \frac{r_n s_n}{r_1} \vec{d}_n.$$

Oben haben wir bereits gesehen, dass nicht alle $\frac{r_k}{r_1} = 0$ sind. Damit sind aber in beiden Fällen die Vektoren $\vec{d}_2, \dots, \vec{d}_n$ linear abhängig. Der Beweis lässt sich nun sukzessive fortsetzen, bis man schließlich die lineare Abhängigkeit von $\vec{d}_n$ erhält, die aber im Widerspruch zu $\vec{d}_n \neq \vec{0}$ als Eigenvektor steht. Also ist die Annahme falsch, und die Vektoren sind linear unabhängig. $\square$

Satz 3.35 (Eigenschaften von Eigenwerten)

Es sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$. Dann gilt:

a) $\mathbf{A}$ ist genau dann invertierbar, falls $s = 0$ kein Eigenwert von $\mathbf{A}$ ist.

- b) Ist s Eigenwert von $\mathbf{A}$ mit einem zugehörigen Eigenvektor $\vec{d}$, so ist für $k \in \mathbb{N}$ die Zahl s^k Eigenwert von $\mathbf{A}^k = \underbrace{\mathbf{A} \cdot \mathbf{A} \cdot \dots \cdot \mathbf{A}}_{k\text{-mal}}$ mit dem gleichen Eigenvektor $\vec{d}$.
- c) Sei $\mathbf{A}$ invertierbar und $\vec{d}$ ein Eigenvektor zum Eigenwert s der Matrix $\mathbf{A}$. Dann ist $\vec{d}$ ebenfalls ein Eigenvektor zum Eigenwert $\frac{1}{s}$ der Inversen $\mathbf{A}^{-1}$.

Beweis: a) $s = 0$ ist Eigenwert genau dann, wenn $\det \mathbf{A} = \det(\mathbf{A} - 0 \cdot \mathbf{E}) = 0$, also genau dann, wenn $\mathbf{A}$ nicht invertierbar ist.

b) $\mathbf{A}^2 \cdot \vec{d} = \mathbf{A} \cdot (\mathbf{A} \cdot \vec{d}) = \mathbf{A} \cdot (s\vec{d}) = s(\mathbf{A} \cdot \vec{d}) = s^2 \vec{d}$ usw.

c) Nach a) ist $s \neq 0$ und

$$\mathbf{A}\vec{d} = s\vec{d} \iff \mathbf{A}^{-1}\mathbf{A}\vec{d} = s\mathbf{A}^{-1}\vec{d} \iff \vec{d} = s\mathbf{A}^{-1}\vec{d} \iff \mathbf{A}^{-1}\vec{d} = \frac{1}{s}\vec{d}. \quad \square$$

Setzt man in das charakteristische Polynom $p(s)$ einer $(n \times n)$ -Matrix $\mathbf{A}$ die Matrix 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 $\mathbf{A}$ wie in Satz 3.35 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 die Multiplikation von $p(\mathbf{A})$ mit jedem Eigenvektor den Nullvektor ergibt. Sei dazu $\vec{d}$ ein Eigenvektor zum Eigenwert s . Wegen

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

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

Eigenwerte werden uns später beim Lösen von Differenzialgleichungssystemen helfen. Dafür benötigen wir auch die folgende Aussage:

Lemma 3.23 (Konjugiert komplexe Eigenwerte)

Ist $s \in \mathbb{C}$ ein komplexer Eigenwert der reellen Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, so ist auch $\bar{s}$ ein Eigenwert. Ist $\vec{d} = \vec{d}_1 + j\vec{d}_2$, $\vec{d}_1, \vec{d}_2 \in \mathbb{R}^n$ ein Eigenvektor zu s , so ist $\bar{\vec{d}} := \vec{d}_1 - j\vec{d}_2$ ein Eigenvektor zu $\bar{s}$.

Beweis: Komplexe Nullstellen des reellen charakteristischen Polynoms der reellen Matrix $\mathbf{A}$ treten immer in Paaren von konjugiert-komplexen Zahlen auf (siehe Seite 148). Damit ist mit s auch $\bar{s}$ ein Eigenwert. Das sehen wir aber auch, indem wir alle Zahlen der Gleichung $(\mathbf{A} - s\mathbf{E})\vec{d} = \vec{0}$ konjugieren:

$$\overline{(\mathbf{A} - s\mathbf{E})\vec{d} = \vec{0}} \iff (\mathbf{A} - \bar{s}\mathbf{E})\bar{\vec{d}} = \vec{0},$$

da die Einträge der Matrix $\mathbf{A}$ reell sind, also $\overline{\mathbf{A}} = \mathbf{A}$. Insbesondere ist damit $\vec{d}$ ein Eigenvektor zum Eigenwert $\bar{s}$. $\square$

Besonders angenehm verhalten sich Eigenwerte symmetrischer, reeller Matrizen:

Satz 3.36 (Eigenwerte symmetrischer Matrizen)

Es sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix, d. h. $\mathbf{A} = \mathbf{A}^\top$.

- a) Alle Eigenwerte von $\mathbf{A}$ sind reell.
- b) Alle reellen Eigenvektoren zu **verschiedenen** (reellen) Eigenwerten sind orthogonal bezüglich des (reellen) Standardskalarprodukts.

Beweis: a) Ist $s \in \mathbb{C}$ ein Eigenwert mit Eigenvektor $\vec{d} = (d_1, d_2, \dots, d_n)^\top \in \mathbb{C}^n$, so ist nach Lemma 3.23 $\vec{d}$ ein Eigenvektor zum Eigenwert $\bar{s}$. Damit betrachten wir das Produkt des Zeilenvektors $s\vec{d}^\top$ mit dem Spaltenvektor $\vec{d}$:

$$s\vec{d}^\top \cdot \vec{d} = (\mathbf{A}\vec{d})^\top \cdot \vec{d} = \vec{d}^\top \mathbf{A}^\top \cdot \vec{d} \stackrel{\mathbf{A}^\top = \mathbf{A}}{=} \vec{d}^\top \mathbf{A} \vec{d} = \vec{d}^\top \cdot \bar{s}\vec{d} = \bar{s}\vec{d}^\top \cdot \vec{d}.$$

Wegen $\vec{d} \neq \vec{0}$ ist $\vec{d}^\top \cdot \vec{d} = \sum_{k=1}^n |d_k|^2 \neq 0$, so dass $s = \bar{s}$ gilt. Das bedeutet aber $s \in \mathbb{R}$.

- b) Die Orthogonalität von Eigenvektoren $\vec{d}_1$ und $\vec{d}_2 \in \mathbb{R}^n$ zu verschiedenen Eigenwerten $s_1 \neq s_2$ folgt direkt aus $\mathbf{A} = \mathbf{A}^\top$:

$$\begin{aligned} s_1(\vec{d}_1 \cdot \vec{d}_2) &= s_1(\vec{d}_1^\top \cdot \vec{d}_2) = (s_1\vec{d}_1)^\top \cdot \vec{d}_2 = (\mathbf{A}\vec{d}_1)^\top \cdot \vec{d}_2 = \vec{d}_1^\top \mathbf{A}^\top \vec{d}_2 \\ &= \vec{d}_1^\top \mathbf{A} \vec{d}_2 = \vec{d}_1^\top s_2 \vec{d}_2 = s_2(\vec{d}_1 \cdot \vec{d}_2). \end{aligned}$$

Damit ist $(s_1 - s_2)(\vec{d}_1 \cdot \vec{d}_2) = 0$, und da die Eigenwerte verschieden sind, stehen die Vektoren senkrecht zueinander. $\square$

Hintergrund: Eigenwerte selbstadjungierter Matrizen

Analog zu symmetrischen reellen Matrizen verhalten sich selbstadjungierte komplexe Matrizen. Der zuvor gezeigte Satz ist ein Spezialfall des folgenden:

Satz 3.37 (Eigenwerte selbstadjungierter Matrizen)

Es sei $\mathbf{A} \in \mathbb{C}^{n \times n}$ eine selbstadjungierte (hermitesche) Matrix, d. h. $\mathbf{A} = \mathbf{A}^* := \overline{\mathbf{A}^\top}$.

- a) Alle Eigenwerte von $\mathbf{A}$ sind reell.
- b) Alle Eigenvektoren zu **verschiedenen** (reellen) Eigenwerten sind orthogonal bezüglich des komplexen Standardskalarprodukts (3.22) auf Seite 486, d. h., ihr komplexes Standardskalarprodukt ist null.

Beweis: Der Beweis verläuft wie zuvor:

- a) 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

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

muss daher $s = \bar{s}$ und damit s reell sein.

- b) 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 ($s_1 = \bar{s}_1$):

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

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

Linear unabhängige Eigenvektoren zum gleichen Eigenwert sind in der Regel nicht orthogonal zueinander. Sie können mit dem Verfahren von Gram-Schmidt (siehe Seite 493) orthogonalisiert werden.

Beispiel 3.67

- a) In Beispiel 3.65 haben wir die konjugiert komplexen Eigenwerte der Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$s_1 = 1 + j$ und $s_2 = 1 - j$ bereits berechnet. Als Eigenvektoren sind nun zu den Eigenwerten nicht-triviale Lösungen des homogenen Systems $(\mathbf{A} - s\mathbf{E}) \cdot \vec{d} = \vec{0}$, d. h.

$$\begin{bmatrix} 1-s & 1 \\ -1 & 1-s \end{bmatrix} \cdot \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

zu bestimmen. Für $s = 1 + j$ ergibt sich

$$\left[\begin{array}{cc|c} -j & 1 & 0 \\ -1 & -j & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} 1 & j & 0 \\ -1 & -j & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} 1 & j & 0 \\ 0 & 0 & 0 \end{array} \right].$$

Wählen wir $d_2 = 1$, so ergibt sich $d_1 = -j$, und wir erhalten einen Eigenvektor $\vec{d}_1 = (-j, 1)^\top$. Einen Eigenvektor $\vec{d}_2$ zu $s = 1 - j$ können wir ohne Rechnung durch komplexe Konjugation bestimmen: $\vec{d}_2 = \overline{\vec{d}_1} = (j, 1)^\top$ ist Eigenvektor zu $s = 1 - j$.

b) Wir bestimmen die Eigenwerte und Eigenvektoren der Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ -1 & 1 & 2 \end{bmatrix}. \quad (3.30)$$

Die charakteristische Gleichung von $\mathbf{A}$ lautet

$$\det(\mathbf{A} - s\mathbf{E}) = (2-s)(1-s)(2-s) = (1-s)(2-s)^2 = 0.$$

Also hat die Matrix $\mathbf{A}$ die Eigenwerte 1 und 2 und besitzt somit zwei Eigenräume. Eigenwerte zu $s = 2$ erhalten wir über

$$\begin{bmatrix} -1 & 1 & 0 \\ 0 & 0 & 0 \\ -1 & 1 & 0 \end{bmatrix} \cdot \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Die allgemeine Lösung ist durch $d_1 = t$, $d_2 = t$, $d_3 = r$ mit reellen Parametern t, r gegeben. Damit sind die Eigenvektoren zum Eigenwert $s = 2$ die von Null verschiedenen Vektoren der Form

$$\vec{d} = \begin{pmatrix} t \\ t \\ r \end{pmatrix} = t \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + r \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Da die Vektoren $\vec{d}_1 = (1, 1, 0)^\top$ und $\vec{d}_2 = (0, 0, 1)^\top$ linear unabhängig sind, bilden sie eine Basis des Eigenraums von $\mathbf{A}$ zu $s = 2$. Für $s = 1$ ergibt sich aus

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \cdot \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

die allgemeine Lösung $d_1 = t$, $d_2 = 0$, $d_3 = t$ mit $t \in \mathbb{R}$. Die Eigenvektoren zum Eigenwert $s = 1$ sind die von Null verschiedenen Vektoren der Form $\vec{d} = (t, 0, t)^\top = t(1, 0, 1)^\top$, so dass $\vec{d}_3 = (1, 0, 1)^\top$ den Eigenraum zu $s = 1$ aufspannt. ■

Falls man ohne exakte Berechnung nur abschätzen möchte, welche Zahlen als Eigenwerte in Frage kommen, dann hilft die folgende Aussage.

Satz 3.38 (Kreissatz von Gerschgorin)

Es sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$. Dann liegt jeder Eigenwert s von $\mathbf{A}$ in mindestens einem der Kreise $K_i \subset \mathbb{C}$, $1 \leq i \leq n$:

$$K_i = \left\{ z \in \mathbb{C} : |z - a_{i,i}| \leq r_i := \sum_{k=1, k \neq i}^n |a_{i,k}| \right\}.$$

Beweis: Sei $\vec{d}$ ein Eigenvektor zum Eigenwert $s \in \mathbb{C}$ mit dominierender Komponente d_i , d. h. $|d_i| \geq |d_k|$ für alle $k \in \{1, \dots, n\}$. Als Eigenvektor ist $\vec{d}$ nicht der Nullvektor, so dass $|d_i| > 0$ ist. Mit diesem Index i rechnen wir jetzt:

$$\begin{aligned} \mathbf{A} \cdot \vec{d} = s\vec{d} &\implies a_{i,1}d_1 + a_{i,2}d_2 + \dots + a_{i,n}d_n = s \cdot d_i \\ &\implies (s - a_{i,i}) \cdot d_i = \sum_{k=1, k \neq i}^n a_{i,k} d_k \\ &\implies |s - a_{i,i}| \leq \sum_{k=1, k \neq i}^n |a_{i,k}| \cdot \underbrace{\left| \frac{d_k}{d_i} \right|}_{\leq 1} \leq \sum_{k=1, k \neq i}^n |a_{i,k}| = r_i. \end{aligned}$$

Damit ist $s \in K_i$. □

3.6.2 Diagonalisierung von Matrizen *

Wir interessieren uns im letzten Abschnitt der Linearen Algebra dafür, ob zu einer gegebenen Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$ eine invertierbare Matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ (bzw. $\mathbb{C}^{n \times n}$) existiert, so dass $\mathbf{D} := \mathbf{X}^{-1} \cdot \mathbf{A} \cdot \mathbf{X}$ eine Diagonalmatrix mit den Diagonalelementen $d_1, \dots, d_n$ ist. Die Transformationsmatrix $\mathbf{X}$ entspricht der Matrix $\mathbf{T}$, mit der in Kapitel 3.4.5 Basiswechsel für die Matrizen linearer Abbildungen beschrieben werden. $\mathbf{A}$ ist dann ähnlich zur Diagonalmatrix $\mathbf{D}$, und Vieles wird einfacher:

$$\mathbf{A}^m = (\mathbf{X} \cdot \mathbf{D} \cdot \mathbf{X}^{-1})^m = \mathbf{X} \cdot \mathbf{D} \cdot (\mathbf{X}^{-1} \cdot \mathbf{X}) \cdot \mathbf{D} \cdot \dots \cdot (\mathbf{X}^{-1} \cdot \mathbf{X}) \cdot \mathbf{D} \cdot \mathbf{X}^{-1} = \mathbf{X} \cdot \mathbf{D}^m \cdot \mathbf{X}^{-1}.$$

Dabei ist $\mathbf{D}^m$ ebenfalls eine Diagonalmatrix mit Diagonalelementen $d_1^m, \dots, d_n^m$. Die Potenzierung wird also sehr einfach. Potenzen von Matrizen findet man z. B. bei Iterationsverfahren, bei denen man $\mathbf{A}$ stets auf das nächste Zwischenergebnis anwendet. Dazu sehen wir uns später noch einmal das Pendlerproblem an.

Hat man ein Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$, so ist dieses gelöst, wenn man eine Diagonalisierung hat: $\mathbf{X} \cdot \mathbf{D} \cdot \mathbf{X}^{-1} \cdot \vec{x} = \vec{b} \iff \vec{x} = \mathbf{X} \cdot \mathbf{D}^{-1} \cdot \mathbf{X}^{-1} \cdot \vec{b}$, wobei $\mathbf{D}^{-1}$ eine Diagonalmatrix mit Diagonalelementen $\frac{1}{d_1}, \dots, \frac{1}{d_n}$ ist.

Zudem wird uns die Diagonalisierung bei der Lösung linearer Differenzialgleichungssysteme helfen (siehe Kapitel 5.3). Es lohnt daher, die Existenz von $\mathbf{X}$ zu untersuchen.

Definition 3.30 (Diagonalisierbarkeit)

Eine Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$ heißt genau dann **diagonalisierbar**, wenn es eine Matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$ gibt, so dass

$$\mathbf{X}^{-1} \cdot \mathbf{A} \cdot \mathbf{X}$$

eine Diagonalmatrix ist.

Sei $\mathbf{A}$ diagonalisierbar mit einer entsprechenden Matrix $\mathbf{X}$, und seien $\vec{x}_1, \dots, \vec{x}_n$ die Spalten von $\mathbf{X}$. Dann folgt aus $\mathbf{A} \cdot \mathbf{X} = \mathbf{X} \cdot \mathbf{D}$:

$$\mathbf{A} \cdot \vec{x}_k = \vec{x}_k d_k.$$

Die Spalten von $\mathbf{X}$ sind also Eigenvektoren von $\mathbf{A}$. Da $\mathbf{X}$ invertierbar ist, müssen die Spalten von $\mathbf{X}$ linear unabhängig sein: Es gibt n linear unabhängige Eigenvektoren.

Hat man umgekehrt n linear unabhängige Eigenvektoren zu (nicht notwendigerweise verschiedenen) Eigenwerten $d_1, \dots, d_n$, so kann man diese in eine Matrix $\mathbf{X}$ als Spalten schreiben. $\mathbf{X}$ ist invertierbar, und $\mathbf{X}^{-1} \cdot \mathbf{A} \cdot \mathbf{X}$ ist eine Diagonalmatrix $\mathbf{D}$ mit den Eigenwerten als Hauptdiagonalelemente: $\mathbf{X}^{-1} \cdot \mathbf{A} \cdot \mathbf{X} = \mathbf{X}^{-1} \cdot \mathbf{X} \cdot \mathbf{D} = \mathbf{D}$. Damit haben wir bewiesen:

Satz 3.39 (Diagonalisierbarkeit einer Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$)

Eine Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ oder $\mathbb{C}^{n \times n}$ ist genau dann diagonalisierbar, wenn sie n linear unabhängige Eigenvektoren hat.

Beispiel 3.68

a) Wir untersuchen die Matrix des Pendlerproblems aus (3.27) mit

$$\mathbf{A} = \begin{bmatrix} 0,6 & 0,2 \\ 0,4 & 0,8 \end{bmatrix}$$

auf Diagonalisierbarkeit. $\mathbf{A}$ besitzt die Eigenwerte $s_1 = 1$ und $s_2 = 0,4$ mit den Eigenvektoren $\vec{x}_1 = (1, 2)^\top$ und $\vec{x}_2 = (1, -1)^\top$. Eine Eigenvektormatrix ist gegeben durch

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}, \text{ und die Inverse ist } \mathbf{X}^{-1} = \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}.$$

Es gilt

$$\mathbf{X}^{-1} \cdot \mathbf{A} \cdot \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 0,4 \end{bmatrix}.$$

Als Anwendung der Diagonalisierung berechnen wir die Pendlerverteilung im Jahr n ausgehend von den Startwerten a_0 und $o_0 = 1 - a_0$. Mit $\mathbf{A}^n = \mathbf{X} \cdot \mathbf{D}^n \cdot \mathbf{X}^{-1}$ erhalten wir:

$$\begin{aligned} \begin{pmatrix} a_n \\ o_n \end{pmatrix} &= \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0,4^n \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{pmatrix} a_0 \\ 1 - a_0 \end{pmatrix} \\ &= \frac{1}{3} \begin{bmatrix} 1 & 0,4^n \\ 2 & -(0,4^n) \end{bmatrix} \begin{pmatrix} 1 \\ 3a_0 - 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 + 0,4^n[3a_0 - 1] \\ 2 - 0,4^n[3a_0 - 1] \end{pmatrix}. \end{aligned}$$

Damit erhalten wir unabhängig von der Startsituation $\lim_{n \rightarrow \infty} a_n = \frac{1}{3}$ und $\lim_{n \rightarrow \infty} o_n = \frac{2}{3}$. Der Grenzwert bedeutet, dass irgendwann nahezu $\frac{2}{3}$ der Pendler mit öffentlichen Verkehrsmitteln fahren werden. Das kann bei der Planung berücksichtigt werden.

b) Wir bestimmen eine Eigenvektormatrix $\mathbf{X}$, die die Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ -1 & 1 & 2 \end{bmatrix}$$

aus (3.30) diagonalisiert. Zu den Eigenwerten $s_1 = 1$ und $s_2 = 2$ haben wir bereits die Eigenvektoren $\vec{x}_1 = (1, 0, 1)^\top$ (zu $s_1 = 1$) und $\vec{x}_2 = (1, 1, 0)^\top$ sowie $\vec{x}_3 = (0, 0, 1)^\top$ (zu $s_2 = 2$) gefunden. Die beiden Vektoren zu $s_2 = 2$ sind so ausgewählt, dass sie linear unabhängig sind.

Da es drei linear unabhängige Eigenvektoren gibt, ist die Matrix $\mathbf{A}$ diagonalisierbar mit

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad \mathbf{X}^{-1} \cdot \mathbf{A} \cdot \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

c) Die Matrix $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ mit $\det(\mathbf{A} - s\mathbf{E}) = (1 - s)^2$ ist nicht diagonalisierbar.

Denn man findet zum doppelten Eigenwert $s = 1$ nur die Eigenvektoren $(c, 0)^\top$, $0 \neq c \in \mathbb{C}$, und hat damit nicht $n = 2$ linear unabhängige Eigenvektoren. ■

Es gibt keine festgelegte Reihenfolge für die Spaltenvektoren der Eigenvektormatrix $\mathbf{X}$. Daher führt eine Vertauschung der Spalten von $\mathbf{X}$ zu einer entsprechenden Vertauschung der Hauptdiagonalelemente von $\mathbf{D} = \mathbf{X}^{-1} \cdot \mathbf{A} \cdot \mathbf{X}$.

Da nach Lemma 3.22 Eigenvektoren zu unterschiedlichen Eigenwerten linear unabhängig sind, erhalten wir mit Satz 3.39 die folgende hinreichende Bedingung für Diagonalisierbarkeit:

Satz 3.40 (Kriterium für Diagonalisierbarkeit)

Besitzt eine Matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ n paarweise verschiedene Eigenwerte, so ist $\mathbf{A}$ diagonalisierbar.

Es gibt aber auch diagonalisierbare Matrizen wie (3.30), die weniger als n verschiedene Eigenwerte besitzen. Diese haben dann k linear unabhängige Eigenvektoren zu k -fachen Eigenwerten.

Beispiel 3.69

Zum Abschluss kommen wir noch einmal auf den goldenen Schnitt zurück. Wir können die Fibonacci-Folge (siehe Seite 231) $(a_n)_{n=1}^\infty$ mit $a_0 := 0$, $a_1 := 1$, $a_n := a_{n-1} + a_{n-2}$ für $n \geq 2$ auch so angeben:

$$\begin{pmatrix} a_n \\ a_{n-1} \end{pmatrix} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}}_{=: \mathbf{A}} \begin{pmatrix} a_{n-1} \\ a_{n-2} \end{pmatrix}.$$

Da $\det \begin{bmatrix} 1-s & 1 \\ 1 & -s \end{bmatrix} = s^2 - s - 1$ die Nullstellen $s = \frac{1}{2} \pm \frac{\sqrt{5}}{2}$ besitzt, hat die Matrix $\mathbf{A}$ diese beiden Eigenwerte. Dabei ist $\frac{1}{2} + \frac{\sqrt{5}}{2} = \Phi$ der goldene Schnitt, und für den zweiten Eigenwert gilt:

$$\frac{1}{2} - \frac{\sqrt{5}}{2} = \frac{\left(\frac{1}{2} - \frac{\sqrt{5}}{2}\right) \left(\frac{1}{2} + \frac{\sqrt{5}}{2}\right)}{\frac{1}{2} + \frac{\sqrt{5}}{2}} = -\frac{1}{\Phi}.$$

Zum Eigenwert Φ erhält man als Lösung des entsprechenden homogenen Gleichungssystems einen Eigenvektor $(\Phi, 1)^\top$, und zum Eigenwert $-\frac{1}{\Phi}$ erhält man analog den Eigenvektor $(-\frac{1}{\Phi}, 1)^\top$. Die Matrix $\mathbf{A}$ ist also diagonalisierbar mit

$$\mathbf{X} = \begin{bmatrix} \Phi & -\frac{1}{\Phi} \\ 1 & 1 \end{bmatrix} \quad \text{und} \quad \mathbf{X}^{-1} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & \frac{1}{\Phi} \\ -1 & \Phi \end{bmatrix}.$$

Damit erhalten wir:

$$\begin{aligned} \begin{pmatrix} a_n \\ a_{n-1} \end{pmatrix} &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^{n-1} \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} = \left[\mathbf{X} \begin{bmatrix} \Phi & 0 \\ 0 & -\frac{1}{\Phi} \end{bmatrix} \mathbf{X}^{-1} \right]^{n-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \mathbf{X} \begin{bmatrix} \Phi^{n-1} & 0 \\ 0 & [-\frac{1}{\Phi}]^{n-1} \end{bmatrix} \mathbf{X}^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{bmatrix} \Phi & -\frac{1}{\Phi} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \Phi^{n-1} & 0 \\ 0 & [-\frac{1}{\Phi}]^{n-1} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{\Phi} \\ -1 & \Phi \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{bmatrix} \Phi & -\frac{1}{\Phi} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \Phi^{n-1} & 0 \\ 0 & [-\frac{1}{\Phi}]^{n-1} \end{bmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{bmatrix} \Phi & -\frac{1}{\Phi} \\ 1 & 1 \end{bmatrix} \begin{pmatrix} \Phi^{n-1} \\ -[-\frac{1}{\Phi}]^{n-1} \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} \Phi^n - [-\frac{1}{\Phi}]^n \\ \Phi^{n-1} - [-\frac{1}{\Phi}]^{n-1} \end{pmatrix}. \end{aligned}$$

Dies ist die **Binet-Formel** für die Fibonacci-Folge. Damit ist aber

$$\lim_{n \rightarrow \infty} \frac{a_n}{a_{n-1}} = \lim_{n \rightarrow \infty} \frac{\Phi^n - [-\frac{1}{\Phi}]^n}{\Phi^{n-1} - [-\frac{1}{\Phi}]^{n-1}} = \Phi, \quad (3.31)$$

denn wegen $0 < \frac{1}{\Phi} < 1$ ist $\lim_{n \rightarrow \infty} \left[-\frac{1}{\Phi}\right]^n = \lim_{n \rightarrow \infty} \left[-\frac{1}{\Phi}\right]^{n-1} = 0$. Damit strebt die Folge der Quotienten $\frac{a_n}{a_{n-1}}$ tatsächlich gegen den goldenen Schnitt Φ . ■

Symmetrische Matrizen lassen sich besonders schön diagonalisieren. Durch eine sukzessive Konstruktion kann man zeigen, dass sich jede symmetrische reelle Matrix $\mathbf{A}$ zu einer Diagonalmatrix mit den reellen Eigenwerten auf der Hauptdiagonalen diagonalisieren lässt. Insbesondere gibt es einen vollen Satz von n zueinander orthogonalen Eigenvektoren (vgl. Satz 3.36) und damit eine Transformationsmatrix $\mathbf{X}$, deren Spalten eine Orthonormalbasis bilden und für die somit $\mathbf{X}^{-1} = \mathbf{X}^\top$ gilt.

Satz 3.41 (Diagonalisierbarkeit symmetrischer Matrizen)

Sei $\mathbf{A}$ eine reelle Matrix. Dann sind äquivalent:

- a) $\mathbf{A}$ ist symmetrisch.
- b) Dann existiert eine orthogonale Matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n \times n}$ mit $\mathbf{D} = \mathbf{X}^{-1} \mathbf{A} \mathbf{X} = \mathbf{X}^\top \mathbf{A} \mathbf{X}$.

Gilt b), so hat $\mathbf{A}$ die Darstellung $\mathbf{A} = \mathbf{X} \mathbf{D} \mathbf{X}^\top$. Damit ist $\mathbf{A}^\top = (\mathbf{X} \mathbf{D} \mathbf{X}^\top)^\top = (\mathbf{X}^\top)^\top \mathbf{D}^\top \mathbf{X}^\top = \mathbf{X} \mathbf{D} \mathbf{X}^\top = \mathbf{A}$, d. h., $\mathbf{A}$ ist symmetrisch.

Entsprechend (vgl. Kasten auf Seite 545) sind für jede komplexe Matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ äquivalent:

- a) $\mathbf{A}$ ist selbstadjungiert.
- b) Es existiert eine unitäre Matrix $\mathbf{X} \in \mathbb{C}^{n \times n}$ 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 $\mathbf{A}^* \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A}^*$.

3.6.3 Hauptvektoren und Jordan-Normalform*

Selbst wenn man eine Matrix verwendet, die nicht genügend linear unabhängige Eigenvektoren hat, um diagonalisierbar zu sein, so kann man sie unter Zuhilfenahme von sogenannten Hauptvektoren zumindest in eine komplexe Matrix mit nur zwei besetzten Diagonalen bringen. Damit lassen sich Differenzialgleichungssysteme lösen (vgl. Kapitel 5.3.5.2).

Definition 3.31 (Hauptvektoren)

Seien $\mathbf{A} \in \mathbb{C}^{n \times n}$ und $s \in \mathbb{C}$ ein Eigenwert von $\mathbf{A}$. Ein Vektor $\vec{d} \in \mathbb{C}^n$ heißt genau dann **Hauptvektor** von $\mathbf{A}$ zu s , wenn es ein $k \in \mathbb{N}$ gibt mit $(\mathbf{A} - s\mathbf{E})^k \vec{d} = \vec{0}$.

Der Hauptvektor $\vec{d}$ heißt genau dann von **k -ter Stufe**, wenn (mit $(\mathbf{A} - s\mathbf{E})^0 = \mathbf{E}$)

$$(\mathbf{A} - s\mathbf{E})^k \vec{d} = \vec{0}, \text{ aber } (\mathbf{A} - s\mathbf{E})^{k-1} \vec{d} \neq \vec{0}.$$

Eigenvektoren sind Hauptvektoren der Stufe 1, da $(\mathbf{A} - s\mathbf{E})^1 \vec{d} = (\mathbf{A} - s\mathbf{E}) \vec{d} = \vec{0}$ und $(\mathbf{A} - s\mathbf{E})^0 \vec{d} = \mathbf{E} \vec{d} = \vec{d}$, wobei $\vec{d} \neq \vec{0}$ nach Definition ist.

Ist $\vec{d}$ ein Hauptvektor der Stufe $k > 1$, so ist $(\mathbf{A} - s\mathbf{E}) \vec{d}$ ein Hauptvektor der Stufe $k - 1$. So kann man durch fortgesetzte Multiplikation mit $(\mathbf{A} - s\mathbf{E})$ eine **Kette von k Hauptvektoren** erzeugen.

Lemma 3.24 (Lineare Unabhängigkeit von Hauptvektoren)

Die k Hauptvektoren $\vec{d}_l := (\mathbf{A} - s\mathbf{E})^{k-l} \vec{d}$, $1 \leq l \leq k$ einer Kette zu einem Hauptvektor $\vec{d}_k = \vec{d}$ der Stufe k sind (komplex) linear unabhängig.

Beweis: Falls die Vektoren nicht linear unabhängig sind, dann müsste es mindestens einen Koeffizienten $c_i \neq 0$ geben, so dass $c_1 \vec{d}_1 + c_2 \vec{d}_2 + \dots + c_k \vec{d}_k = \vec{0}$ ist. In dieser Gleichung sei l der größte Index mit $c_l \neq 0$. Dann lässt sich der Hauptvektor $\vec{d}_l$ der Stufe l als Linearkombination von Hauptvektoren kleinerer Stufen schreiben. Dann wäre aber $(\mathbf{A} - s\mathbf{E})^{l-1} \vec{d}_l = \vec{0}$ im Widerspruch zur Definition eines Hauptvektors der Stufe l . Folglich kann der Nullvektor nur als triviale Linearkombination geschrieben werden, die Vektoren sind linear unabhängig. $\square$

Lemma 3.25 (Charakterisierung von Hauptvektoren)

Seien $\mathbf{A} \in \mathbb{C}^{n \times n}$ und $s \in \mathbb{C}$ ein Eigenwert von $\mathbf{A}$. Hauptvektoren zum Eigenwert s sind genau die von $\vec{0}$ verschiedenen Lösungen $\vec{d}$ der Gleichung

$$(\mathbf{A} - s\mathbf{E})^n \vec{d} = \vec{0}.$$

Beweis: Man sieht sofort, dass jeder Hauptvektor zu einer Stufe $k \leq n$ Lösung der Gleichung ist und umgekehrt. Hauptvektoren einer Stufe $k > n$ kann es aber wegen der linearen Unabhängigkeit der Vektoren einer Kette mit k Vektoren (Lemma 3.24) im n -dimensionalen Raum $\mathbb{C}^n$ nicht geben. $\square$

Ist s ein Eigenwert von $\mathbf{A}$ der Vielfachheit k , dann existieren genau k komplex linear unabhängige Hauptvektoren von $\mathbf{A}$ zu s (die den **Hauptraum** der Dimension k zum Eigenwert s erzeugen und in Ketten organisiert sind). Gibt es darunter mehr als einen Eigenvektor und damit mehr als eine Kette, so kann die volle Stufe k bei

Hauptvektoren nicht erreicht werden. Wie Eigenvektoren sind (abgesehen von $\vec{0}$) auch Hauptvektoren zu unterschiedlichen Eigenwerten linear unabhängig, insgesamt findet man stets einen vollen Satz von n linear unabhängigen Hauptvektoren. Auf den (hinsichtlich der Dimension und Struktur der Haupträume) aufwändigen Beweis verzichten wir hier, vgl. (Arens et al., 2012, Kapitel 18.8) und die dort zitierte Literatur.

Beispiel 3.70

a) Die Matrix $\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ hat das charakteristische Polynom $p(s) = s^2$ und da-

mit den doppelten Eigenwert $s = 0$. $(1, 0)^\top$ und $(0, 1)^\top$ sind linear unabhängige Eigenvektoren. Es gibt hier keinen Hauptvektor $\vec{d}$ der Stufe 2, da die Bedingung $(\mathbf{A} - s\mathbf{E})^1 \vec{d} \neq \vec{0}$ nicht erfüllbar ist. Es gibt dafür aber zwei linear unabhängige Hauptvektoren der Stufe 1, nämlich z.B. die beiden gefundenen Eigenvektoren. Diese bilden Ketten der Länge 1.

b) Die Matrix $\mathbf{A}$ hat den dreifachen Eigenwert 2 und den fünffachen Eigenwert 3:

$$\mathbf{A} := \begin{bmatrix} \boxed{2} & \boxed{0} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \boxed{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \boxed{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \boxed{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \boxed{3} \end{bmatrix}.$$

Sei $\vec{e}_k \in \mathbb{C}^8$ der Einheitsvektor, der in der k -ten Komponente eine Eins und sonst nur Nullen aufweist. Zum Eigenwert 2 gibt es einen Hauptvektor $\vec{e}_1$ der Stufe 2 und zwei Hauptvektoren $\vec{e}_2$ und $\vec{e}_3$ der Stufe 1 (also Eigenvektoren), die linear unabhängig sind. Zu 3 gibt es eine „Kette“, bestehend aus je einem Hauptvektor $\vec{e}_4$ der Stufe 3, einem Hauptvektor $\vec{e}_5$ der Stufe 2 und einem Eigenvektor (Stufe 1) $\vec{e}_6$, sowie eine Kette, bestehend aus je einem Hauptvektor $\vec{e}_7$ der Stufe 2 und einem Eigenvektor $\vec{e}_8$, so dass alle Haupt- bzw. Eigenvektoren linear unabhängig sind. ■

Die Hauptvektoren zu einem Eigenwert der Vielfachheit k können so ermittelt werden:

a) Berechne eine Lösung $\vec{d}_k$ von $(\mathbf{A} - s\mathbf{E})^k \vec{d}_k = \vec{0}$, für die $(\mathbf{A} - s\mathbf{E})^{k-1} \vec{d}_k \neq \vec{0}$. Falls es keine Lösung gibt, versuche linear unabhängige Lösungen von $(\mathbf{A} - s\mathbf{E})^{k-1} \vec{d}_k = \vec{0}$ mit $(\mathbf{A} - s\mathbf{E})^{k-2} \vec{d}_k \neq \vec{0}$ zu bestimmen. Gibt es auch hier keine Lösungen, dann mache mit der Aufgabe $(\mathbf{A} - s\mathbf{E})^{k-2} \vec{d}_k = \vec{0}$ mit $(\mathbf{A} - s\mathbf{E})^{k-3} \vec{d}_k \neq \vec{0}$ weiter usw.

- b) Hat man so Hauptvektoren $\vec{d}_r$ einer Stufe r gefunden, erhält man mittels $\vec{d}_{r-1} = (\mathbf{A} - s\mathbf{E})\vec{d}_r$, $\vec{d}_{r-2} = (\mathbf{A} - s\mathbf{E})\vec{d}_{r-1}$ usw., zu jedem dieser Hauptvektoren $r-1$ weitere. Falls damit die gesuchte Anzahl k noch nicht erreicht wird, muss mit dem ersten Schritt weitergemacht werden. Dabei werden neue Hauptvektoren, die zu bereits gefundenen linear abhängig sind, nicht berücksichtigt.

Bei der Diagonalisierung haben wir eine Matrix $\mathbf{X}$ erstellt, deren Spalten linear unabhängige Eigenvektoren sind. Jetzt benutzen wir eine Transformationsmatrix $\mathbf{X}$, deren Spalten Hauptvektoren sind. Gibt es zum Eigenwert s eine Kette von Hauptvektoren, beginnend mit $\vec{d}_r$ der Stufe r , und sind $\vec{d}_{r-1} = (\mathbf{A} - s\mathbf{E})\vec{d}_r$, $\vec{d}_{r-2} = (\mathbf{A} - s\mathbf{E})\vec{d}_{r-1}, \dots, \vec{d}_1 = (\mathbf{A} - s\mathbf{E})\vec{d}_2$ zugehörige Hauptvektoren kleinerer Stufen, so schreiben wir sie beginnend mit $\vec{d}_r$ in Spalten $i, i+1, \dots, i+r-1$ der Matrix $\mathbf{X}$. Entsprechend tragen wir alle gefundenen linear unabhängigen Ketten und Eigenvektoren als Spalten in $\mathbf{X}$ ein. Damit entsteht eine invertierbare Matrix $\mathbf{X} \in \mathbb{C}^{n \times n}$.

Die Matrix $\mathbf{A}\mathbf{X}$ hat die i -te Spalte $\mathbf{A}\vec{d}_r = s\vec{d}_r + \vec{d}_{r-1}$, die nächste lautet $\mathbf{A}\vec{d}_{r-1} = s\vec{d}_{r-1} + \vec{d}_{r-2}$ usw. Da $\vec{d}_1$ ein Eigenvektor ist, steht schließlich in der $i+r-1$ -ten Spalte $\mathbf{A}\vec{d}_1 = s\vec{d}_1$. Damit können wir $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{T}$ schreiben, wobei $\mathbf{T}$ eine Matrix mit maximal zwei besetzten Diagonalen ist. Die Kette von Hauptvektoren, die in der Spalte i von $\mathbf{X}$ beginnt, führt dabei zu einem Kästchen in $\mathbf{T}$, dessen linke obere Ecke an der Position (i, i) liegt. Für eine Kettenlänge $r = 3$ sieht es so aus:

$$\begin{array}{|ccc|} \hline s & 0 & 0 \\ 1 & s & 0 \\ 0 & 1 & s \\ \hline \end{array}.$$

Zu jeder Kette von Hauptvektoren erhält man ein entsprechendes Kästchen um die Hauptdiagonale in $\mathbf{T}$, wobei der Eigenwert auf der Hauptdiagonalen steht. Darunter kennzeichnen Einsen die zusammengehörende Kette. Hat man mehrere Ketten, so sieht $\mathbf{T}$ beispielsweise wie die Matrix $\mathbf{A}$ in Beispiel 3.70 b) aus, die aus Kästchen zu vier Ketten besteht. Die Matrix $\mathbf{T} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ heißt eine **Jordan-Normalform** von $\mathbf{A}$.

Wenn man die Hauptvektoren einer Kette in umgekehrter Reihenfolge (beginnend mit dem Eigenvektor) in $\mathbf{X}$ einträgt, so stehen die Einsen oberhalb der Hauptdiagonalen. Auch dann spricht man von einer Jordan-Normalform.

Jede quadratische Matrix ist also zu einer Matrix in Jordan-Normalform ähnlich (siehe Definition 3.24 auf Seite 520). Damit kann man jede lineare Abbildung von $\mathbb{C}^n$ nach $\mathbb{C}^n$ über eine Matrix mit höchstens zwei besetzten Diagonalen ausdrücken, was beim Rechnen eine große Erleichterung sein kann.

Beispiel 3.71

$\mathbf{A}$ in Beispiel 3.70 b) ist bereits in Jordan-Normalform mit $\mathbf{T} = \mathbf{A}$ und $\mathbf{X} = \mathbf{E}$. Mehr Arbeit macht

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 4 & 2 & 1 & -2 & 0 \\ -2 & -1 & 1 & 4 & 0 \\ 0 & 0 & 1 & 2 & 3 \end{bmatrix}.$$

Das charakteristische Polynom $\det(\mathbf{A} - s\mathbf{E}) = (2 - s)^2(3 - s)[(1 - s)(4 - s) + 2] = (2 - s)^3(3 - s)^2$ liefert den dreifachen Eigenwert 2 und den doppelten Eigenwert 3.

Wir berechnen eine Kette von Hauptvektoren zum Eigenwert 2:

$$(\mathbf{A} - 2\mathbf{E})^2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 4 & 2 & -1 & -2 & 0 \\ -2 & -1 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 & 1 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & -1 & -2 & 0 \\ -1 & 0 & 1 & 2 & 0 \\ 0 & 0 & 2 & 4 & 1 \end{bmatrix},$$

$$(\mathbf{A} - 2\mathbf{E})^3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 3 & 6 & 1 \end{bmatrix}.$$

Da beispielsweise $(\mathbf{A} - 2\mathbf{E})^3 \cdot (1, 0, 0, 0, 0)^\top = \vec{0}$, aber $(\mathbf{A} - 2\mathbf{E})^2 \cdot (1, 0, 0, 0, 0)^\top = (0, 0, 2, -1, 0)^\top \neq \vec{0}$ ist, haben wir einen Hauptvektor der Stufe 3 gefunden, den wir in die erste Spalte der Transformationsmatrix $\mathbf{X}$ schreiben. Eine andere Wahl und damit ein anderes $\mathbf{X}$ wäre auch möglich. Die beiden weiteren Hauptvektoren der zugehörigen Kette sind $(\mathbf{A} - 2\mathbf{E}) \cdot (1, 0, 0, 0, 0)^\top = (0, 1, 4, -2, 0)^\top$ und $(\mathbf{A} - 2\mathbf{E}) \cdot (0, 1, 4, -2, 0)^\top = (0, 0, 2, -1, 0)^\top$. Diese schreiben wir in die Spalten zwei und drei von $\mathbf{X}$.

Entsprechend verfahren wir nun mit dem doppelten Eigenwert 3:

$$(\mathbf{A} - 3\mathbf{E})^2 = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 4 & 2 & -2 & -2 & 0 \\ -2 & -1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 \end{bmatrix}^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ -6 & 4 & 2 & 2 & 0 \\ 4 & 2 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Da $(\mathbf{A} - 3\mathbf{E})^2 \cdot (0, 0, 1, -1, 0)^\top = \vec{0}$, aber $(\mathbf{A} - 3\mathbf{E}) \cdot (0, 0, 1, -1, 0)^\top = (0, 0, 0, 0, -1)^\top \neq \vec{0}$ ist, haben wir einen Hauptvektor der Stufe 2 gefunden, den wir in die vierte Spalte der Transformationsmatrix $\mathbf{X}$ schreiben. Dabei haben wir direkt auch den Haupt- bzw. Eigenvektor $(0, 0, 0, 0, -1)^\top$ gewonnen, der $\mathbf{X}$ komplettiert. Insgesamt erhalten wir $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{T}$ mit

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 4 & 2 & 1 & 0 \\ 0 & -2 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix} \quad \text{und} \quad \mathbf{T} = \begin{bmatrix} \boxed{2} & 0 & 0 & 0 & 0 \\ 1 & \boxed{2} & 0 & 0 & 0 \\ 0 & 1 & \boxed{2} & 0 & 0 \\ 0 & 0 & 0 & \boxed{3} & 0 \\ 0 & 0 & 0 & 1 & \boxed{3} \end{bmatrix}.$$

■

3.7 Normierte Vektorräume: Lineare Algebra trifft Analysis*

In diesem Kapitel wollen wir die Lineare Algebra mit der Analysis verbinden. Der Betrag einer reellen Zahl wird dabei durch eine Norm ersetzt. Die Norm legt einen Abstandsbegriff für Vektoren (und damit eine Topologie, siehe Kasten auf Seite 276) fest, so dass man über die Stetigkeit von Abbildungen zwischen Vektorräumen sprechen kann. Wir betrachten aber keine beliebigen Abbildungen, sondern lineare Abbildungen. Das Teilgebiet der Mathematik, das sich mit der Stetigkeit linearer Abbildungen beschäftigt, ist die **Funktionalanalysis**. In Kapitel 4 werden wir die Analysis auch für nicht notwendigerweise lineare Abbildungen auf dem $\mathbb{R}^n$ betreiben. Dabei handelt es sich aber um eine zwar wichtige, aber geradlinige Erweiterung des Kapitels 2. Dagegen bietet die Funktionalanalysis tatsächlich neue interessante Einsichten.

3.7.1 Norm

Wir haben bereits eine Euklid'sche Norm über ein Skalarprodukt in Kapitel 3.3.3.2 definiert. Im Folgenden setzen wir aber kein Skalarprodukt voraus, die Euklid'sche Norm ist Spezialfall des folgenden Begriffs.

Definition 3.32 (Norm)

Sei V ein reeller oder komplexer Vektorraum, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. Eine Abbildung $\|\cdot\| = \|\cdot\|_V : V \rightarrow [0, \infty[$ heißt eine **Norm** auf V , falls für alle $\vec{x}, \vec{y} \in V$ und $r \in \mathbb{K}$ gilt:

■ Aus $\|\vec{x}\| = 0$ folgt $\vec{x} = \vec{0}$,

- $\|r\vec{x}\| = |r|\|\vec{x}\|$ (**absolute Homogenität**),
- $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ (**Dreiecksungleichung**).

$(V, \|\cdot\|)$ heißt ein **normierter Vektorraum** oder ein **linear normierter Raum**.

Beispiel 3.72 (Normierte Vektorräume)

- Offensichtlich ist der Betrag eine Norm für die Vektorräume $(\mathbb{R}, +; \mathbb{R}, \cdot)$ und $(\mathbb{C}, +; \mathbb{C}, \cdot)$.
- Auf dem Vektorraum $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$ sind für jedes $p \in [1, \infty[$ die Abbildungen $\|\cdot\|_p : \mathbb{R}^n \rightarrow [0, \infty[$ mit

$$\|(x_1, x_2, \dots, x_n)\|_p := (|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{1/p}$$

Normen. Für $p = 2$ erhält man den Betrag für Vektoren des $\mathbb{R}^n$, der gemäß dem Satz von Pythagoras den geometrischen Abstand zum Nullvektor angibt und sich darüber hinaus auch aus dem Standardskalarprodukt ergibt.

Wir werden auch die Wahl $p = \infty$ zulassen:

$$\|(x_1, x_2, \dots, x_n)\|_\infty := \max\{|x_1|, |x_2|, \dots, |x_n|\}.$$

- Hat man einen endlich-dimensionalen Vektorraum, dann kann man jeden Vektor eindeutig als Linearkombination von endlich vielen Basisvektoren schreiben. Dem Vektor sind so eindeutig die Koeffizienten dieser Linearkombination zugeordnet. Eine Norm erhält man, indem man eine der zuvor diskutierten Normen des Vektorraums $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$ auf die Koeffizienten anwendet.
- Wir erhalten einen normierten Vektorraum l^p von Folgen $(x_k)_{k=1}^\infty$ mit der Norm

$$\|(x_k)_{k=1}^\infty\|_p := \|(x_k)_{k=1}^\infty\|_{l^p} := \left(\sum_{k=1}^\infty |x_k|^p \right)^{1/p}$$

für $1 \leq p < \infty$, indem wir nur solche Folgen in den Vektorraum nehmen, für die die unendliche Summe eine endliche Zahl ergibt. Dass es sich hier um eine Norm handelt, muss nachgerechnet werden. Darauf verzichten wir hier aber. Insbesondere ist der Nachweis der Dreiecksungleichung nicht trivial (und nutzt die Hölder-Ungleichung). Für $p = \infty$ setzen wir

$$\|(x_k)_{k=1}^\infty\|_\infty := \sup\{|x_k| : k \in \mathbb{N}\}$$

und definieren l^∞ als Menge aller beschränkten Folgen. Die Gegenstücke der Folgenräume l^p sind für Funktionen die L^p -Räume, die wir weiter unten besprechen, siehe Seite 562.

- Der Vektorraum der auf dem Intervall $[a, b]$ stetigen Funktionen wird mit $C[a, b]$ bezeichnet. Mittels

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

wird dieser normiert. Da $|f|$ stetig ist, nimmt $|f|$ auf dem Intervall $[a, b]$ ein Maximum an (siehe Satz 2.26 auf Seite 281), so dass die Norm wohldefiniert ist. Die Eigenschaften einer Norm ergeben sich sofort. ■

Vielfach sind Normen wie in Kapitel 3.3.3.2 über Skalarprodukte definiert. Wir werden später dazu unterschiedliche Skalarprodukte verwenden. Diese lassen sich nicht unterscheiden, wenn wir sie mit dem Multiplikationszeichen schreiben. Daher verwenden wir lokal in diesem Kapitel 3.7 eine Notation mittels Klammern, bei der wir über einen Index dem Skalarprodukt einen Namen geben können. Ist der Vektorraum X über ein Skalarprodukt normiert, dann bezeichnen wir es mit $(\cdot, \cdot)_X$.

Die Eigenschaften eines Skalarprodukts $(\cdot, \cdot)_X$ führen dazu, dass $\|\cdot\|_X$ mit $\|\vec{x}\|_X := \sqrt{(\vec{x}, \vec{x})_X}$ die Eigenschaften einer Norm erfüllt. Wir erinnern dazu an die Definition 3.14 des Skalarprodukts für Vektorräume über dem Körper $\mathbb{R}$ auf Seite 483 und an die Definition 3.15 im Kasten auf Seite 485 für Vektorräume über $\mathbb{C}$.

Die normierten Vektorräume sind übrigens Topologische Räume (siehe Kasten auf Seite 276), bei denen die Topologie durch die Norm erzeugt wird. Zu $\delta > 0$ lässt sich beispielsweise über $\{\vec{x} \in V : \|\vec{x} - \vec{x}_0\|_V < \delta\}$ eine offene Umgebung des Vektors $\vec{x}_0$ definieren. Dabei können wir über die Norm den Abstand eines Vektors $\vec{x}$ zu einem Vektor $\vec{x}_0$ angeben. Mit den offenen Umgebungen lässt sich dann eine Topologie analog wie für die reellen Zahlen festlegen.

Definition 3.33 (Äquivalente Normen)

Zwei Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ auf einem Vektorraum V heißen **äquivalent** genau dann, wenn es reelle Konstanten $c, C > 0$ gibt mit

$$c\|\vec{x}\|_a \leq \|\vec{x}\|_b \leq C\|\vec{x}\|_a \text{ für alle } \vec{x} \in V.$$

Offensichtlich gilt auch

$$\|\vec{x}\|_b \leq C\|\vec{x}\|_a \leq \frac{C}{c}\|\vec{x}\|_b \text{ für alle } \vec{x} \in V.$$

Äquivalente Normen erzeugen die gleiche Topologie. Hinsichtlich der Begriffe Stetigkeit und Konvergenz wird es keine Unterschiede geben, so dass die Normen austauschbar sind.

Lemma 3.26 (Normen auf einem endlich-dimensionalen Vektorraum)

Sind $\|\cdot\|_a$ und $\|\cdot\|_b$ Normen auf dem gleichen endlich-dimensionalen Vektorraum V , dann sind sie äquivalent.

Wir verzichten auf einen Beweis, der über ein Stetigkeitsargument geführt werden kann.

3.7.2 Banach- und Hilbert-Räume

Eine Folge $(\vec{x}_k)_{k=1}^\infty$ von Vektoren eines normierten Vektorraums $(V, \|\cdot\|_V)$ heißt (völlig analog zu Folgen reeller oder komplexer Zahlen) eine **Cauchy-Folge** genau dann, wenn es zu jedem $\varepsilon > 0$ eine Stelle $n_0 \in \mathbb{N}$ gibt, so dass für alle $n, m > n_0$ gilt:

$$\|\vec{x}_n - \vec{x}_m\|_V < \varepsilon.$$

Eine Folge $(\vec{x}_k)_{k=1}^\infty$ heißt (in der Norm) **konvergent** gegen einen (eindeutigen) Grenzwert $\vec{x}$ genau dann, wenn

$$\lim_{n \rightarrow \infty} \|\vec{x}_n - \vec{x}\|_V = 0.$$

In Kapitel 2.8.2 besprechen wir die gleichmäßige Konvergenz einer Folge von Funktionen. Dabei handelt es sich ebenfalls um eine Konvergenz in einer Norm, nämlich der Supremum-Norm $\|f\|_{B[a,b]} := \sup_{x \in [a,b]} |f(x)|$, die für beschränkte Funktionen erklärt ist und für stetige Funktionen mit $\|f\|_{C[a,b]}$ übereinstimmt.

Ist V endlich-dimensional, dann folgt mit Lemma 3.26 aus der Konvergenz einer Folge in einer Norm automatisch auch die Konvergenz in allen anderen (dann äquivalenten) Normen auf V . Entsprechend ist dann eine Cauchy-Folge in einer Norm auch eine Cauchy-Folge in allen anderen Normen.

Wir wissen, dass jede Cauchy-Folge reeller Zahlen konvergent gegen eine reelle Zahl ist. Dahinter steckt die Vollständigkeit der reellen Zahlen. Diese Eigenschaft möchte man auch für normierte Vektorräume besitzen. Sie ist aber nicht generell gegeben. Vielmehr sind vollständige Räume ausgezeichnet:

Definition 3.34 (Banach-Raum)

*Ist in einem normierten Vektorraum V jede Cauchy-Folge (in der Norm) konvergent gegen ein Element des Vektorraums, dann heißt dieser **vollständig**. Ein vollständiger normierter Vektorraum wird als **Banach-Raum** bezeichnet. Ist die Norm durch ein Skalarprodukt auf V definiert, dann nennt man den Banach-Raum einen **Hilbert-Raum**.*

Banach-Räume sind in vielen Anwendungen der Ingenieurmathematik zumindest für die Theorie sehr wichtig, um die Existenz und Eindeutigkeit von Lösungen (z. B. im Zusammenhang mit Differenzialgleichungen) zu beweisen.

Beispiel 3.73 (Banach-Raum der reellen Vektoren)

Der mit $\|\cdot\|_p$, $1 \leq p \leq \infty$ normierte Vektorraum $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$ ist ein Banach-Raum. Hat man eine Cauchy-Folge $(\vec{x}_k)_{k=1}^\infty$ von Vektoren, so ist auch jede Komponentenfolge eine Cauchy-Folge und damit wegen der Vollständigkeit der reellen Zahlen konvergent gegen einen Komponentengrenzwert in $\mathbb{R}$. Der aus diesen Grenzwerten zusammengesetzte Vektor ist dann Grenzwert von $(\vec{x}_k)_{k=1}^\infty$ (vgl. mit (4.1) für $\|\cdot\|_2$ auf Seite 593). So folgt auch, dass $(\mathbb{C}^n, +; \mathbb{C}, \cdot)$ ein Banach-Raum ist.

Aufgrund dieser Argumentation, aber auch da auf dem endlich-dimensionalen Vektorraum $\mathbb{R}^n$ oder $\mathbb{C}^n$ alle Normen äquivalent sind, bedeutet Konvergenz in einer $\|\cdot\|_p$ -Norm auch automatisch Konvergenz in allen anderen Normen.

Das Kapitel 4 beschäftigt sich im Wesentlichen mit Funktionen von $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$ in $(\mathbb{R}^m, +; \mathbb{R}, \cdot)$, wobei beide Vektorräume mit der $\|\cdot\|_2$ -Norm ausgestattet und damit Banach-Räume sind. ■

Beispiel 3.74 (Banach-Raum der stetigen Funktionen)

Wir zeigen, dass der Vektorraum $C[a, b]$ der reellen, auf $[a, b]$ stetigen Funktionen zusammen mit der Maximum-Norm $\|\cdot\|_{C[a, b]}$ ein Banach-Raum ist. Sei dazu $(f_k)_{k=1}^\infty$ eine Cauchy-Folge stetiger Funktionen, d. h., zu jedem $\varepsilon > 0$ existiere ein $n_0(\varepsilon) \in \mathbb{N}$, so dass für $m, n > n_0(\varepsilon)$ gilt:

$$\|f_n - f_m\|_{C[a, b]} := \max\{|f_n(x) - f_m(x)| : x \in [a, b]\} \leq \varepsilon.$$

Wir müssen zeigen, dass diese Cauchy-Folge gegen eine stetige Funktion f in der Maximum-Norm konvergiert. Zunächst beschaffen wir uns einen Kandidaten für diese Grenzfunktion. Für jedes $x \in [a, b]$ ist die Zahlenfolge $(f_k(x))_{k=1}^\infty$ eine Cauchy-Folge, die damit gegen einen reellen Grenzwert $f(x)$ konvergiert. Über diese punktweise Konvergenz ist damit die Grenzfunktion erklärt.

Aus der Cauchy-Bedingung lässt sich für $\varepsilon = 2^{-n}$ eine streng monoton wachsende Folge natürlicher Zahlen $(k_n)_{n=1}^\infty$ über $k_n > n_0(2^{-n})$ finden mit

$$\|f_{k_{n+1}} - f_{k_n}\|_{C[a, b]} \leq 2^{-n}.$$

Die Folge $(f_{k_n})_{n=1}^\infty$ konvergiert natürlich auch punktweise gegen f . Damit erhalten wir für f und ein $n \in \mathbb{N}$ die Darstellung:

$$\begin{aligned} f(x) &= \lim_{l \rightarrow \infty} f_{k_l}(x) = \lim_{l \rightarrow \infty} \left[f_{k_n}(x) + \sum_{i=n}^{l-1} [f_{k_{i+1}}(x) - f_{k_i}(x)] \right] \\ &= f_{k_n}(x) + \sum_{i=n}^{\infty} [f_{k_{i+1}}(x) - f_{k_i}(x)]. \end{aligned}$$

Da wir noch nicht wissen, ob f stetig ist und damit ein Maximum annimmt, verwenden wir zunächst $\|f\|_{B[a, b]} := \sup\{|f(x)| : x \in [a, b]\}$. Damit gilt unter Verwendung der geometrischen Summe (1.10) von Seite 60 und der geometrischen Reihe (2.11), siehe Seite 237:

$$\begin{aligned} \|f - f_{k_n}\|_{B[a, b]} &= \left\| \sum_{i=n}^{\infty} [f_{k_{i+1}} - f_{k_i}] \right\|_{B[a, b]} \leq \sum_{i=n}^{\infty} \|f_{k_{i+1}} - f_{k_i}\|_{C[a, b]} \\ &\leq \sum_{i=n}^{\infty} 2^{-i} = \frac{1}{1 - \frac{1}{2}} - \frac{1 - 2^{-n}}{1 - \frac{1}{2}} = 2^{-n+1}. \end{aligned}$$

Somit konvergiert die Folge stetiger Funktionen $(f_{k_n})_{n=1}^\infty$ gleichmäßig gegen die Grenzfunktion f , die nach Satz 2.56 auf Seite 406 stetig und damit im Raum enthalten ist. Jetzt können wir statt $\|\cdot\|_{B[a,b]}$ auch $\|\cdot\|_{C[a,b]}$ schreiben, und unsere Teilfolge konvergiert in der Norm gegen f . Aus der Konvergenz der Teilfolge ergibt sich wie im Beweis zu Satz 2.8 (Seite 235, Teil b) die Konvergenz der gesamten Folge. Dabei ist lediglich die Zahlenfolge durch die Funktionenfolge und der Betrag durch die Norm zu ersetzen. ■

3.7.3 L^p -Räume

Wir haben zuvor in Kapitel 2.6.8 das Lebesgue-Integral mit der Begründung eingeführt, dass man diesen Integrationsbegriff benötigt, damit gewisse Funktionenräume Banach-Räume werden. Das sind die L^p -Räume, die vielleicht wichtigsten Banach-Räume, die in der angewandten Mathematik benutzt werden.

Wir beginnen mit einem Vektorraum, der aus den auf dem Intervall $[a, b]$ Riemann-integrierbaren, reellen Funktionen besteht. Diesen Vektorraum können wir (analog zum Vektorraum der Folgen) mit unterschiedlichen Normen versehen. Wichtig sind die folgenden **L^p -Normen**:

$$\|f\|_p := \|f\|_{L^p[a,b]} := \sqrt[p]{\int_a^b |f(x)|^p dx},$$

wobei $1 \leq p < \infty$ ein fester Parameter ist. Wenn wir die Eigenschaften der Normen überprüfen wollen, dann stellen wir fest, dass alle bis auf die erste erfüllt sind (die Dreiecksungleichung ergibt sich aus der weiter unten angegebenen Minkowski-Ungleichung). Aus $\|f\|_p = 0$ muss ohne Zusatzwissen nicht $f = 0$ folgen, da z.B. einzelne Funktionswerte keinen Einfluss auf den Wert des Integrals haben. Daher fasst man solche Funktionen f und g zu einem Element des Raums zusammen, für die $\|f - g\|_p = 0$ ist. Wir haben damit einen Vektorraum, dessen Elemente Äquivalenzklassen von Funktionen sind, die an gewissen Stellen unterschiedliche Funktionswerte haben. Die Stellen dürfen lediglich für das Ergebnis der Integration keine Rolle spielen. Wenn wir im Folgenden eine Funktion hinschreiben, dann meinen wir damit eigentlich eine ganze Klasse von Funktionen. Auf diesem Vektorraum sind die L^p -Normen tatsächlich Normen. Insbesondere für $p = 2$ ergibt sich die Norm dann aus dem Skalarprodukt $(f, g)_2 := \int_a^b f(x) \cdot g(x) dx$, das wir bereits als Beispiel (3.21) auf Seite 485 unter der Zusatzeigenschaft der Stetigkeit der Funktionen benutzt haben. Die Bedingung $(f, f)_2 = 0 \implies f = 0$ ist dagegen jetzt durch die Klasseneinteilung erfüllt.

Hat man eine Cauchy-Folge von Riemann-integrierbaren Funktionen, dann kann es durchaus sein, dass die Kandidaten-Funktion für den Grenzwert nicht Riemann-

integrierbar ist und damit nicht im Funktionenraum liegt, so dass er nicht vollständig ist. Wir betrachten dazu als Beispiel die Funktionen

$$f_n(x) := \begin{cases} \left[\frac{1}{x}\right]^{\frac{1}{4}}, & \text{falls } x \in \left[\frac{1}{n}, 1\right]. \\ 0, & \text{sonst.} \end{cases}$$

Jede dieser Funktionen ist Riemann-integrierbar und hat eine endliche Norm. Wir zeigen, dass es sich um eine Cauchy-Folge bezüglich der Norm $\|\cdot\|_2$ handelt: Für $m \geq n > 4/\varepsilon^2$ gilt:

$$\|f_n - f_m\|_2^2 = \int_{1/m}^{1/n} \left[\frac{1}{x}\right]^{\frac{2}{4}} dx = \int_{1/m}^{1/n} \frac{1}{\sqrt{x}} dx = [2\sqrt{x}]_{1/m}^{1/n} = 2\sqrt{\frac{1}{n}} - 2\sqrt{\frac{1}{m}} \leq 2\sqrt{\frac{1}{n}} < \varepsilon.$$

Man findet hier aber keine Riemann-integrierbare Grenzfunktion, da $f(x) = \left[\frac{1}{x}\right]^{1/4}$ auf $]0, \infty]$ nicht beschränkt ist.

Mit dem Lebesgue-Integral (das insbesondere auch für unbeschränkte Funktionen erklärt ist und für Funktionen, die sogar den Funktionswert „ ∞ “ auf Mengen vom Maß null annehmen, sinnvoll ist) lassen sich dagegen die Funktionenräume so definieren, dass sie vollständig sind. Der Vektorraum $L^p[a, b]$ mit der Norm $\|\cdot\|_p$, $1 \leq p < \infty$, besteht aus den messbaren Funktionen f , für die (im Sinne des Lebesgue-Integrals) $\|f\|_p < \infty$ ist. Statt für Funktionen auf dem Intervall $[a, b]$ können wir L^p -Räume ohne Probleme direkt auch für Funktionen über beliebigen messbaren Mengen A (auch unbeschränkten Mengen und Mengen mit unendlichem Maß) erklären und verwenden hier $[a, b]$ nur für eine einfachere Darstellung. Wie oben bereits beschrieben, sind die Elemente des Raums streng genommen keine einzelnen Funktionen, sondern Äquivalenzklassen von Funktionen, die hinsichtlich der Norm gleich sind, d. h., die f. ü. gleich sind. Man kann dann mit einem beliebigen Repräsentanten der Äquivalenzklasse rechnen. Aus reiner Bequemlichkeit zur Vermeidung von Fallunterscheidungen lässt man $\pm\infty$ als „Funktionswerte“ zu. Man beachte aber, dass aus der Lebesgue-Integrierbarkeit folgt, dass f höchstens auf einer Menge vom Maß null die „Werte“ $\pm\infty$ annehmen kann, so dass diese beim Rechnen keine Rolle spielen. Die Funktion f lässt sich durch eine gleichwertige reellwertige Funktion $\tilde{f}$ ersetzen mit $f = \tilde{f}$ f. ü.

Bislang haben wir $p = \infty$ ausgeschlossen. Aber auch für diese Parameterwahl entsteht ein wichtiger Banach-Raum – der Raum der wesentlich beschränkten Funktionen. Eine (messbare) Funktion $f \in L^\infty[a, b]$ muss nicht beschränkt sein, aber sie muss auf einer Menge $[a, b] \setminus N$ beschränkt sein, wobei N eine Menge vom Lebesgue-Maß null ist. Die Norm (das **wesentliche Supremum**) ist dann so definiert:

$$\|f\|_\infty := \inf \left\{ \sup_{x \in [a, b] \setminus N} |f(x)| : N \subset [a, b] \text{ ist eine messbare Menge mit } v(N) = 0 \right\}.$$

Man beachte, dass die Definition

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

nicht sinnvoll ist, wenn Funktionen gleich sein dürfen, obwohl sie sich auf einer Menge vom Maß null unterscheiden. Dann könnten gleiche Funktionen eine unterschiedliche Norm haben. Daher verwendet man $\|f\|_{B[a,b]}$ nur in Verbindung mit beschränkten Funktionen, bei denen jeder Funktionswert wichtig ist.

Die L^p -Normen sind tatsächlich Normen. Schwierig ist nur das Nachrechnen der Dreiecksungleichung, die hier **Minkowski-Ungleichung** heißt:

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Für $p = 1$ und $p = \infty$ ergibt sich die Minkowski-Ungleichung direkt aus der Dreiecksungleichung für Beträge. Für $1 < p < \infty$ folgt sie (nicht-trivial) aus der **Hölder-Ungleichung**

$$\|f \cdot g\|_1 \leq \|f\|_r \|g\|_s \quad (3.32)$$

für $\frac{1}{r} + \frac{1}{s} = 1$, $r, s \geq 1$, wobei $r = \infty$ oder $s = \infty$ zugelassen ist mit Kehrwert 0.

Die $L^p[a, b]$ -Räume sind ineinander enthalten, es gilt $L^p[a, b] \subset L^q[a, b]$ für $1 \leq q \leq p \leq \infty$. Wir wenden zum Beweis die Hölder-Ungleichung für $q < p$ mit $r = \frac{p}{q}$, $s = 1 / \left(1 - \frac{q}{p}\right)$ und $g(x) = 1$ an:

$$\|f\|_q^q = \|f^q \cdot 1\|_1 \leq \|f^q\|_{\frac{p}{q}} \|1\|_{1/(1-\frac{q}{p})} = \|f\|_p^q (b-a)^{1-\frac{q}{p}}.$$

Gilt also $\|f\|_p < \infty$, so ist auch $\|f\|_q < \infty$. Hier ist wichtig, dass $v([a, b]) = b - a$ endlich ist. Betrachtet man die L^p -Räume auf einem unbeschränkten Intervall (z. B. $\mathbb{R}$) oder einer messbaren Menge mit unendlichem Lebesgue-Maß, dann liegen die Räume nicht ineinander.

Hintergrund: Vollständigkeit der L^p -Räume

Wir haben oben gesehen, dass das Riemann-Integral nicht zu einem vollständigen Funktionenraum führt. Wir überlegen uns nun, dass die L^p -Räume tatsächlich vollständig sind (wobei im Folgenden $[a, b]$ durch eine beliebige messbare Menge ausgetauscht werden darf, selbst wenn sie unendliches Lebesgue-Maß hat). Dabei beschränken wir uns auf den Fall $1 \leq p < \infty$. Der folgende technische Beweis basiert im Wesentlichen auf den Eigenschaften des Lebesgue-Integrals, die in den Sätzen 2.52 und 2.53 (Seite 381) formuliert sind. Wir müssen zeigen, dass jede Cauchy-Folge konvergiert. Sei dazu $(f_k)_{k=1}^\infty$ eine Cauchy-Folge in L^p . Nach Definition der Cauchy-Folge gibt es zu $\varepsilon = 2^{-n}$ eine Stelle $n_0 = n_0(2^{-n})$, so dass für alle $m, i > n_0$ gilt: $\|f_m - f_i\|_p \leq 2^{-n}$. Damit existiert analog zu Beispiel 3.74 eine Teilfolge $(k_n)_{n=1}^\infty$ der natürlichen Zahlen mit $\|f_{k_n} - f_{k_{n+1}}\|_p \leq 2^{-n}$. Es genügt zu zeigen, dass diese Teilfolge in L^p konvergiert. Daraus folgt wieder wie im Beweis zu Satz 2.8 (Seite 235, Teil b) die Konvergenz der gesamten Folge in L^p . Also zeigen wir die Konvergenz der Teilfolge in L^p :

Wir definieren ähnlich wie in Beispiel 3.74 eine Folge von Hilfsfunktionen

$$h_n(x) := \sum_{i=1}^{n-1} |f_{k_{i+1}}(x) - f_{k_i}(x)|.$$

Durch die Summation nicht-negativer Zahlen ist diese Folge (wie auch die Folge $|h_n(x)|^p$) für jeden Wert x monoton wachsend (wie es im Satz 2.52 gefordert ist). Sie konvergiert damit gegen eine Grenzfunktion h , die aber an Stellen x Funktionswerte ∞ annimmt, wenn die Folge für x unbeschränkt ist.

Mit der Dreiecksungleichung und der Formel (1.10) für die geometrische Summe (Seite 60) erhalten wir

$$\|h_n\|_p \leq \sum_{i=1}^{n-1} \|f_{k_{i+1}} - f_{k_i}\|_p \leq \sum_{i=1}^{n-1} 2^{-i} = \frac{1 - \frac{1}{2^n}}{1 - \frac{1}{2}} - 1 = 1 - \frac{1}{2^{n-1}} < 1.$$

Nach dem Satz von Beppo Levi (siehe Satz 2.52) ist die Grenzfunktion h bzw. $|h|^p$ messbar, und es gilt

$$\lim_{n \rightarrow \infty} \underbrace{L_- \int_a^b |h_n(x)|^p dx}_{= \|h_n\|_p^p < 1} = L_- \int_a^b |h(x)|^p dx.$$

Also ist $L_- \int_a^b |h(x)|^p dx \leq 1$, so dass $h \in L^p[a, b]$ ist. Das wiederum bedeutet, dass h nur auf einer Menge vom Maß null den Funktionswert „ ∞ “ annehmen kann.

Jetzt sind wir nicht an der Funktion h , sondern am Grenzwert unserer Teilfolge interessiert. Da aus der absoluten Konvergenz einer Zahlenfolge ihre Konvergenz folgt, konvergiert die Folge

$$f_{k_n}(x) = f_{k_1}(x) + \sum_{i=1}^{n-1} (f_{k_{i+1}}(x) - f_{k_i}(x))$$

punktweise für alle Stellen x , an denen $h(x)$ nicht den Wert „ ∞ “ hat, also f.ü., gegen eine Grenzfunktion $f(x)$. Mit dem Wissen über h zeigen wir, dass diese Grenzfunktion zu $L^p[a, b]$ gehört. Dafür benötigen wir eine Ungleichung ($1 \leq p < \infty$)

$$\left(\frac{x+y}{2}\right)^p \leq \frac{1}{2}x^p + \frac{1}{2}y^p \text{ für } x, y \geq 0, \quad (3.33)$$

die verständlich wird, da für $t > 0$ gilt: $(t^p)'' = p(p-1)t^{p-2} \geq 0$. Der Graph von t^p ist also linksgekrümmt. In der Mitte $\frac{x+y}{2}$ zwischen den Stellen x und y muss daher der Funktionswert kleiner oder gleich dem Mittelwert der Funktionswerte bei x und bei y sein. Damit gilt (3.33), und wir können die Ungleichung nutzen. Wegen

$$|f_{k_n}|^p \leq [|f_{k_1}| + h_n]^p \stackrel{(3.33)}{\leq} 2^{p-1}(|f_{k_1}|^p + h_n^p) \leq 2^{p-1}(|f_{k_1}|^p + h^p)$$

gibt es eine integrierbare Majorante. Hier benötigen wir die zuvor bewiesene Zugehörigkeit von h zu $L^p[a, b]$. Nach dem Satz über die majorisierte Konvergenz (siehe Satz

2.53) ist daher die Grenzfunktion $|f|^p = \lim_{n \rightarrow \infty} |f_{k_n}|^p$ Lebesgue-integrierbar, d. h. $f \in L^p[a, b]$.

Damit sind wir fast fertig. Wir wissen, dass unsere Teilfolge punktweise f. ü. gegen f konvergiert und f im Funktionenraum liegt. Wir müssen nur noch zeigen, dass die Teilfolge auch in der L^p -Norm gegen f strebt, dass also f der Grenzwert im Sinne des Funktionenraums ist. Jetzt sind mit f_{k_n} und f auch $f_{k_n} - f \in L^p[a, b]$, d. h., die Funktionen $g_n := |f_{k_n} - f|^p = |f_{k_n} - [f_{k_n} + \sum_{i=n}^{\infty} (f_{k_{i+1}} - f_{k_i})]|^p$ sind Lebesgue-integrierbar. Außerdem sind sie beschränkt mit der Majorante

$$|g_n| = \left| \sum_{i=n}^{\infty} (f_{k_{i+1}} - f_{k_i}) \right|^p \leq \left[\sum_{i=n}^{\infty} |f_{k_{i+1}} - f_{k_i}| \right]^p \leq \left[\sum_{i=1}^{\infty} |f_{k_{i+1}} - f_{k_i}| \right]^p = h(x)^p,$$

von der wir ja wissen, dass sie Lebesgue-integrierbar ist. Außerdem konvergiert die Folge g_n punktweise f. ü. gegen null. Damit können wir den Satz über die majorisierte Konvergenz (Satz 2.53) ein zweites Mal anwenden und erhalten

$$\lim_{n \rightarrow \infty} \int_a^b |f_{k_n}(x) - f(x)|^p dx = \lim_{n \rightarrow \infty} \int_a^b g_n(x) dx = \int_a^b 0 dx = 0.$$

Damit ist gezeigt, dass die Teilfolge tatsächlich in der L^p -Norm gegen eine in L^p enthaltene Grenzfunktion konvergiert.

Als Gegenstücke der L^p -Funktionenräume sind auch die l^p -Folgenräume vollständig.

In Kapitel 4.4.5.2 werden L^p -Räume für Funktionen mit mehreren Variablen eingeführt.

3.7.4 Stetige Abbildungen zwischen normierten Vektorräumen

3.7.4.1 Stetigkeit und Operatornormen

Völlig konsistent und analog zur Stetigkeit reeller Funktionen sowie zum Stetigkeitsbegriff der Topologie (siehe Kasten auf Seite 276) heißt eine Abbildung $S : (V, \|\cdot\|_V) \rightarrow (W, \|\cdot\|_W)$ **stetig an einer Stelle** $\vec{x}_0$ genau dann, wenn zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass für alle $\vec{x} \in V$ mit $\|\vec{x}_0 - \vec{x}\|_V < \delta$ gilt: $\|S(\vec{x}_0) - S(\vec{x})\|_W < \varepsilon$. Ist sie für jedes $\vec{x}_0 \in V$ stetig, so heißt S **stetig auf V** .

Viele Eigenschaften stetiger reeller Funktionen übertragen sich daher auf stetige Abbildungen zwischen normierten Vektorräumen. Wenn es dabei um Grenzwerte geht, müssen die beteiligten Räume allerdings Banach-Räume sein.

Insbesondere gilt der Banach'sche Fixpunktsatz (siehe Satz 2.38 auf Seite 318) in der folgenden Fassung, die ohne die Einschränkung auf reelle Funktionen viel anwendungsfreundlicher ist. So können numerische Verfahren oft als Abbildungen eines normierten Vektorraums in sich selbst aufgefasst werden.

Satz 3.42 (Banach'scher Fixpunktsatz für normierte Vektorräume)

Sei $(V, \|\cdot\|_V)$ ein Banach-Raum und $0 \leq \lambda < 1$. Jede Abbildung $S : (V, \|\cdot\|_V) \rightarrow (V, \|\cdot\|_V)$, die eine Kontraktion im Sinne von

$$\|S(\vec{x}) - S(\vec{y})\|_V < \lambda \|\vec{x} - \vec{y}\|_V \text{ für alle } \vec{x}, \vec{y} \in V$$

ist, besitzt genau einen **Fixpunkt** $\vec{x}_0$, d. h. $S(\vec{x}_0) = \vec{x}_0$, und dieser ist der Grenzwert jeder Folge $(\vec{x}_n)_{n=1}^\infty$ mit beliebig wählbarem $\vec{x}_1 \in V$ und $\vec{x}_{n+1} := S(\vec{x}_n)$. Es gelten die **A-priori-Abschätzung**

$$\|\vec{x}_0 - \vec{x}_n\|_V \leq \frac{\lambda^{n-1}}{1-\lambda} \cdot \|\vec{x}_2 - \vec{x}_1\|_V \quad (3.34)$$

und die **A-posteriori-Abschätzung**

$$\|\vec{x}_0 - \vec{x}_n\|_V \leq \frac{\lambda}{1-\lambda} \cdot \|\vec{x}_n - \vec{x}_{n-1}\|_V.$$

Man beachte, dass aus der Bedingung der Kontraktion sofort die Stetigkeit an jeder Stelle in V folgt. Der Beweis des Satzes lässt sich direkt aus dem zu Satz 2.38 und den Herleitungen der Fehlerabschätzungen (2.36) und (2.37) übertragen.

Bislang haben wir noch nicht die Linearität von Abbildungen vorausgesetzt. Liegt dieser wichtige Spezialfall vor, so wird der Begriff der Stetigkeit ganz besonders einfach.

Definition 3.35 (Operatornorm)

Eine lineare Abbildung (ein linearer Operator) $L : (V, \|\cdot\|_V) \rightarrow (W, \|\cdot\|_W)$, $V \neq \{\vec{0}\}$, heißt **beschränkt**, falls die **Operatornorm**

$$\|L\|_{[V,W]} := \sup_{\vec{0} \neq \vec{x} \in V} \frac{\|L\vec{x}\|_W}{\|\vec{x}\|_V}$$

endlich ist.

Da $\frac{\|L\vec{x}\|_W}{\|\vec{x}\|_V} = \left\| L \frac{\vec{x}}{\|\vec{x}\|_V} \right\|_W$ gilt, kann man sich in der Definition auf Vektoren $\vec{x}$ beschränken, die zu eins normiert sind:

$$\|L\|_{[V,W]} = \sup_{\vec{x} \in V \text{ mit } \|\vec{x}\|_V=1} \|L\vec{x}\|_W. \quad (3.35)$$

Für $\vec{x} \neq \vec{0}$ gilt

$$\|L\vec{x}\|_W = \frac{\|L\vec{x}\|_W}{\|\vec{x}\|_V} \|\vec{x}\|_V \leq \|L\|_{[V,W]} \|\vec{x}\|_V.$$

Ist $\vec{x} = \vec{0}$, so gilt die Abschätzung ebenfalls:

$$\|L\vec{x}\|_W = \|\vec{0}\|_W = 0 = \|L\|_{[V,W]} \cdot 0 = \|L\|_{[V,W]} \|\vec{x}\|_V.$$

Insgesamt haben wir damit die ganz wichtige Normabschätzung

$$\|L\vec{x}\|_W \leq \|L\|_{[V,W]} \|\vec{x}\|_V \quad (3.36)$$

gezeigt. Damit erhält man auch sofort eine Normabschätzung für die Verkettung (Hintereinanderausführung) $L \circ S = L(S(\cdot))$ zweier linearer Abbildungen $L : U \rightarrow W$ und $S : V \rightarrow U$. Da für alle $\vec{x} \in V$

$$\|(L \circ S)(\vec{x})\|_W = \|L(S(\vec{x}))\|_W \leq \|L\|_{[U,W]} \|S(\vec{x})\|_U \leq \|L\|_{[U,W]} \|S\|_{[V,U]} \|\vec{x}\|_V$$

ist, folgt

$$\|L \circ S\|_{[V,W]} \leq \|L\|_{[U,W]} \|S\|_{[V,U]}. \quad (3.37)$$

Die linke Seite kann aber echt kleiner als die rechte sein, im Allgemeinen gilt nicht die Gleichheit.

Ist V endlich-dimensional, dann ist die Operatornorm grundsätzlich endlich. Um das zu sehen, sei $\{\vec{x}_1, \dots, \vec{x}_n\}$ eine Basis des Raums. Dann gilt für jeden Vektor $\vec{x} = r_1\vec{x}_1 + \dots + r_n\vec{x}_n \in V$:

$$\begin{aligned} \|L\vec{x}\|_W &= \|L(r_1\vec{x}_1 + \dots + r_n\vec{x}_n)\|_W = \|r_1L(\vec{x}_1) + \dots + r_nL(\vec{x}_n)\|_W \\ &\leq \sum_{k=1}^n |r_k| \|L(\vec{x}_k)\|_W \leq \max\{\|L(\vec{x}_1)\|_W, \dots, \|L(\vec{x}_n)\|_W\} \sum_{k=1}^n |r_k|. \end{aligned}$$

Nun definiert $\sum_{k=1}^n |r_k|$ eine Norm für $\vec{x}$ auf V , wie man leicht nachrechnen kann. Da V endlich-dimensional ist, ist diese Norm laut Lemma 3.26 äquivalent zu $\|\cdot\|_V$:

$$\|L\vec{x}\|_W \leq C_1 \sum_{k=1}^n |r_k| \leq C_2 \|\vec{x}\|_V.$$

Das aber ist die Beschränktheit von L .

Lemma 3.27 (Normierter Vektorraum der linearen Abbildungen)

Seien V und W normierte Vektorräume. Die Menge der beschränkten linearen Abbildungen von V nach W bildet mit der Operatornorm selbst einen normierten Vektorraum.

Die Aussage des Lemmas lässt sich unmittelbar anhand der Definition überprüfen.

Aus der Beschränktheit $\|L\|_{[V,W]} \leq C$ einer linearen Abbildung $L : V \rightarrow W$ folgt ihre Stetigkeit auf V . Denn mit (3.36) folgt sofort

$$\|L\vec{x} - L\vec{y}\|_W = \|L(\vec{x} - \vec{y})\|_W \leq C \|\vec{x} - \vec{y}\|_V$$

für alle $\vec{x}, \vec{y} \in V$. Das ist nicht nur die Stetigkeit von L an jeder Stelle $\vec{x}$, sondern sogar die gleichmäßige Stetigkeit: Zu $\varepsilon > 0$ existiert ein $\delta := \varepsilon/C$ unabhängig von $\vec{x}$, so dass für alle $\vec{y}$ mit $\|\vec{x} - \vec{y}\|_V < \delta$ gilt: $\|L\vec{x} - L\vec{y}\|_W < \varepsilon$.

Umgekehrt ist es nicht schwer indirekt zu beweisen, dass jede stetige lineare Abbildung automatisch beschränkt ist. Wir nehmen dazu an, dass die Abbildung L nicht beschränkt ist. Damit gibt es für jede Zahl $n \in \mathbb{N}$ einen Vektor $\vec{0} \neq \vec{x}_n \in V$, so dass

$$\|L(\vec{x}_n)\|_W \geq n\|\vec{x}_n\|_V.$$

Die Folge $\vec{y}_n := \frac{1}{n\|\vec{x}_n\|_V} \vec{x}_n$ konvergiert in V gegen den Nullvektor, da

$$\lim_{n \rightarrow \infty} \|\vec{y}_n - \vec{0}\|_V = \lim_{n \rightarrow \infty} \left\| \frac{1}{n\|\vec{x}_n\|_V} \vec{x}_n \right\|_V = \lim_{n \rightarrow \infty} \frac{1}{n} \frac{\|\vec{x}_n\|_V}{\|\vec{x}_n\|_V} = 0.$$

Da L stetig ist, muss auch

$$0 = \lim_{n \rightarrow \infty} \|L(\vec{y}_n) - L(\vec{0})\|_W = \lim_{n \rightarrow \infty} \|L(\vec{y}_n)\|_W = \lim_{n \rightarrow \infty} \frac{1}{n\|\vec{x}_n\|_V} \|L(\vec{x}_n)\|_W$$

gelten. Im Widerspruch dazu ist aber die rechte Seite nach Wahl der Vektoren $\vec{x}_n$ größer oder gleich eins. Wir haben bewiesen:

Lemma 3.28 (Stetigkeit linearer Abbildungen)

Bei linearen Abbildungen zwischen normierten Vektorräumen sind die Begriffe Stetigkeit auf dem Vektorraum, gleichmäßige Stetigkeit und Beschränktheit identisch.

In der Numerik und in den Ingenieurwissenschaften ist daneben der Begriff der Stabilität üblich. Wenn eine lineare Abbildung als stabil bezeichnet wird, ist damit die Stetigkeit/Beschränktheit gemeint.

Beispiel 3.75 (Ableitungsoperator)

Die Menge der auf einem Intervall $[a, b]$ stetig differenzierbaren Funktionen ist ein Vektorraum $C^{(1)}[a, b]$, der mit

$$\|f\|_{C^{(1)}[a, b]} := \max\{\|f\|_{C[a, b]}, \|f'\|_{C[a, b]}\}$$

normiert werden kann. Dabei entsteht ein Banach-Raum. Um das nachzurechnen, benötigt man Satz 2.57 auf Seite 407. Eine besonders wichtige beschränkte lineare Abbildung $L : C^{(1)}[a, b] \rightarrow C[a, b]$ ist die Ableitung:

$$L(f) := f'.$$

Die Linearität ist durch die Linearität der Ableitung gegeben. Die Beschränktheit folgt sofort aus

$$\|L(f)\|_{C[a, b]} = \|f'\|_{C[a, b]} \leq \|f\|_{C^{(1)}[a, b]},$$

so dass die Operatornorm kleinergleich eins ist. Für $f(x) = \exp(x)$ ist außerdem

$$\|L(\exp(x))\|_{C[a, b]} = \|\exp(x)\|_{C[a, b]} = \|\exp(x)\|_{C^{(1)}[a, b]},$$

so dass die Operatornorm gleich eins ist: $\|L\|_{[C^{(1)}[a, b], C[a, b]]} = 1$. ■

Die Namensgebung Differenzialoperator für $\frac{d}{dx}$ ist dadurch entstanden, dass es sich tatsächlich um eine beschränkte lineare Abbildung (einen beschränkten linearen Operator) handelt. Die linken Seiten von Differenzialgleichungen, die wir in Kapitel 5 betrachten, sind Linearkombinationen von Differenzialoperatoren und können so ebenfalls häufig als beschränkte lineare Abbildungen verstanden werden. Allerdings sind die hier gewählten Supremum-Normen in der Praxis vielfach zu einschränkend, da die gesuchten Lösungsfunktionen nicht genügend glatt sind. Deshalb wird mittels partieller Integration ein schwächerer Ableitungsbegriff für Funktionen aus den L^p -Räumen eingeführt. Indem man die l^p -Norm des Vektors der L^p -Normen der Funktion und ihrer sogenannten schwachen Ableitungen bis zu einer vorgegebenen Ordnung verwendet, entsteht ein Banach-Raum, der als **Sobolev-Raum** bezeichnet wird (siehe Kapitel 3.7.6). Dieser spielt beim Lösen (partieller) Differenzialgleichungen eine wichtige Rolle.

Namensgeber der Funktionalanalysis sind die Funktionalen. Eine lineare Abbildung eines Vektorraums in die reellen oder komplexen Zahlen $\mathbb{K}$ heißt ein **lineares Funktional**. Ist V ein normierter Vektorraum und fasst man $\mathbb{K}$ ebenfalls als normierten Vektorraum (mit dem Betrag als Norm) auf, so können wir den Vektorraum der beschränkten linearen Funktionalen auf V betrachten, der gleich dem Vektorraum der beschränkten linearen Abbildungen von V nach $\mathbb{K}$ ist. Dieser Vektorraum heißt der **Dualraum** zu V und wird mit V^* bezeichnet.

Mit Funktionalen hatten wir auch schon zu tun, nur haben wir sie nicht so genannt.

Beispiel 3.76 (Integral als Funktional)

Auf dem normierten Vektorraum $(C[a, b], \|\cdot\|_{C[a, b]})$ der auf dem Intervall $[a, b]$ stetigen Funktionen ist beispielsweise die Abbildung $F(f) := \int_a^b f(x) dx$, $F : C[a, b] \rightarrow \mathbb{R}$ ein beschränktes lineares Funktional. Die Linearität folgt sofort aus der des Integrals. Das Funktional ist beschränkt, da $|F(f)| \leq (b - a)\|f\|_{C[a, b]}$. Damit ist insbesondere die Operatornorm

$$\|F\|_{C[a, b]^*} := \|F\|_{[C[a, b], \mathbb{R}]} = \sup_{0 \neq f \in C[a, b]} \frac{|F(f)|}{\|f\|_{C[a, b]}} = \sup_{0 \neq f \in C[a, b]} \frac{\left| \int_a^b f(x) dx \right|}{\sup_{x \in [a, b]} |f(x)|}$$

kleiner oder gleich $b - a$. Sie ist sogar genau gleich $b - a$, da die Funktion $f(x) = 1$ in $C[a, b]$ enthalten ist. ■

Wichtige Beispiele für lineare Funktionalen sind die Distributionen, also verallgemeinerte Funktionen, siehe Kasten ab Seite 898. Die Delta-Distribution (der Dirac-Stoß, die Impulsfunktion) wird z.B. in der Regelungstechnik eingesetzt, um das Verhalten linearer zeitinvarianter Systeme vollständig zu ermitteln.

3.7.4.2 Matrix-Normen

Die Frage, ob eine lineare Abbildung beschränkt ist, wird bei unendlich-dimensionalen Vektorräumen wie Funktionenräumen wichtig. Der tatsächliche Wert einer Operatornorm wird aber häufig auch bei linearen Abbildungen zwischen endlich-dimensionalen Räumen benötigt. Diese Abbildungen können als Matrixmultiplikation dargestellt werden. Eine **Matrix-Norm** ist definiert als Operatornorm der durch die Matrix beschriebenen linearen Abbildung. Sei beispielsweise $\mathbf{A} \in \mathbb{R}^{m \times n}$ eine Matrix, die eine lineare Abbildung von $\mathbb{R}^n$ in $\mathbb{R}^m$ beschreibt, oder $\mathbf{A} \in \mathbb{C}^{m \times n}$ eine Matrix, die eine lineare Abbildung von $\mathbb{C}^n$ in $\mathbb{C}^m$ festlegt. Wir staten sowohl $\mathbb{R}^n$ bzw. $\mathbb{C}^n$ als auch $\mathbb{R}^m$ bzw. $\mathbb{C}^m$ mit der $\|\cdot\|_p$ -Norm aus. Dann ist

$$\|\mathbf{A}\|_p := \sup_{\vec{0} \neq \vec{x} \in \mathbb{R}^n \text{ bzw. } \mathbb{C}^n} \frac{\|\mathbf{A}\vec{x}\|_p}{\|\vec{x}\|_p}.$$

Generell sind Matrix-Normen quadratischer Matrizen $\mathbf{A} \in \mathbb{C}^{n \times n}$ mindestens so groß wie der betragsmäßig größte Eigenwert s der Matrix $\mathbf{A}$. Ist $\vec{d} \neq \vec{0}$ ein Eigenvektor zu s , dann folgt aus

$$|s| \|\vec{d}\|_p = \|s\vec{d}\|_p = \|\mathbf{A}\vec{d}\|_p \leq \|\mathbf{A}\|_p \|\vec{d}\|_p$$

direkt $\|\mathbf{A}\|_p \geq |s|$. Der betragsmäßig größte Eigenwert heißt der **Spektralradius** der Matrix $\mathbf{A}$. Er kann mit dem Satz von Gerschgorin (Seite 547) abgeschätzt werden.

Von besonderer Bedeutung sind die Normen zu $p = 1$, $p = 2$ und $p = \infty$. Die $\|\cdot\|_2$ -Vektornorm misst die Euklid'sche Länge des Vektors und wird daher oft verwendet, z. B. in Kapitel 4 als Betrag. Die darüber definierte Matrix-Norm $\|\mathbf{A}\|_2$ einer $m \times n$ -Matrix $\mathbf{A}$ heißt **Spektralnorm** von $\mathbf{A}$, denn sie kann, bei reellwertigem $\mathbf{A}$, als Wurzel des Spektralradius der quadratischen symmetrischen Matrix $\mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{n \times n}$ berechnet werden. Diese Matrix hat nur reelle nicht-negative Eigenwerte, so dass der Spektralradius gleich dem größten Eigenwert von $\mathbf{A}^\top \mathbf{A}$ ist. Diese Darstellung der Matrix-2-Norm werden wir im Beispiel 4.20 auf Seite 619 mittels mehrdimensionaler Extremwertrechnung herleiten. Für Matrizen mit komplexen Einträgen muss $\mathbf{A}^\top$ durch die Adjungierte $\mathbf{A}^*$ ersetzt werden, dann ist $\|\mathbf{A}\|_2$ die Wurzel des größten Eigenwertbetrags von $\mathbf{A}^* \mathbf{A}$.

Lemma 3.29 (Matrix-Normen zu $p = 1$ und $p = \infty$)

Für eine Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ oder $\mathbb{C}^{m \times n}$ ist

$$\|\mathbf{A}\|_1 = \max_{1 \leq i \leq n} \sum_{k=1}^m |a_{k,i}|, \quad \|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{k=1}^n |a_{i,k}|.$$

Bei der Matrix-1-Norm wird das Maximum über die absoluten Spaltensummen (**Spaltensummennorm**) der Matrix gebildet, bei der Matrix- ∞ -Norm wird dagegen das Maximum über die absoluten Zeilensummen (**Zeilensummennorm**) berechnet.

Beweis: a) Wir bestimmen $\|\mathbf{A}\|_1$: Dazu verwenden wir die Standard-Einheitsvektoren $\vec{e}_i \in \mathbb{R}^n \subset \mathbb{C}^n$, die in der i -ten Komponente eine 1 und sonst nur Null-Komponenten besitzen. Dann ist $\|\vec{e}_i\|_1 = 1$ und

$$\|\mathbf{A}\|_1 \geq \|\mathbf{A}\vec{e}_i\|_1 = \sum_{k=1}^m |a_{k,i}|$$

für jedes $1 \leq i \leq n$, so dass $\|\mathbf{A}\|_1 \geq \max_{1 \leq i \leq n} \sum_{k=1}^m |a_{k,i}|$ folgt. Für die umgekehrte Abschätzung sei $\vec{y} \in \mathbb{R}^n$ bzw. $\mathbb{C}^n$ ein beliebiger Vektor:

$$\begin{aligned} \|\mathbf{A}\vec{y}\|_1 &= \sum_{k=1}^m \left| \sum_{i=1}^n a_{k,i} y_i \right| \leq \sum_{k=1}^m \sum_{i=1}^n |a_{k,i}| \cdot |y_i| = \sum_{i=1}^n \left[\sum_{k=1}^m |a_{k,i}| \right] \cdot |y_i| \\ &\leq \sum_{i=1}^n \left[\max_{1 \leq k \leq m} \sum_{i=1}^n |a_{k,i}| \right] \cdot |y_i| = \left[\max_{1 \leq i \leq n} \sum_{k=1}^m |a_{k,i}| \right] \cdot \sum_{i=1}^n |y_i| \\ &= \left[\max_{1 \leq i \leq n} \sum_{k=1}^m |a_{k,i}| \right] \|\vec{y}\|_1. \end{aligned}$$

Damit folgt aber $\|\mathbf{A}\|_1 \leq \max_{1 \leq i \leq n} \sum_{k=1}^m |a_{k,i}|$, und zusammen haben wir die Darstellung der Matrix-1-Norm bewiesen.

b) Wir bestimmen $\|\mathbf{A}\|_\infty$: Sind in der i -ten Zeile von $\mathbf{A}$ alle Komponenten gleich null, so gilt offensichtlich $\|\mathbf{A}\|_\infty \geq \sum_{k=1}^n |a_{i,k}| = 0$. Anderenfalls sei bei einer reellen Matrix $\mathbf{A}$

$$\vec{y}_i := (\text{sign}(a_{i,1}), \text{sign}(a_{i,2}), \dots, \text{sign}(a_{i,n})).$$

In Verallgemeinerung dieser Definition setzen wir bei einer komplexen Matrix $\mathbf{A}$

$$\vec{y}_i := (e^{-j\varphi_{i,1}}, e^{-j\varphi_{i,2}}, \dots, e^{-j\varphi_{i,n}}),$$

wobei $\varphi_{i,k}$ der Winkel der komplexen Zahl $a_{i,k}$ in Polarform ist. Damit ist $a_{i,k} e^{-j\varphi_{i,k}} = |a_{i,k}| e^{j\varphi_{i,k}} e^{-j\varphi_{i,k}} = |a_{i,k}|$. Insbesondere gilt für reelles $a_{i,k}$, dass $\text{sign}(a_{i,k}) = e^{-j\varphi_{i,k}}$. Damit ist $\|\vec{y}_i\|_\infty = 1$, und es gilt ebenfalls

$$\|\mathbf{A}\|_\infty \geq \frac{\|\mathbf{A}\vec{y}_i\|_\infty}{\|\vec{y}_i\|_\infty} = \|\mathbf{A}\vec{y}_i\|_\infty \geq \left| \sum_{k=1}^n a_{i,k} e^{-j\varphi_{i,k}} \right| = \sum_{k=1}^n |a_{i,k}|.$$

So können wir für jedes $1 \leq i \leq m$ abschätzen und erhalten $\|\mathbf{A}\|_\infty \geq \max_{1 \leq i \leq m} \sum_{k=1}^n |a_{i,k}|$. Für die umgekehrte Abschätzung benutzen wir (3.35). Sei dazu $\vec{y}$ ein beliebiger Vektor mit $\|\vec{y}\|_\infty = 1$. Insbesondere sind alle Komponenten des Vektors betragsmäßig höchstens eins. Der Betrag der i -ten Komponente des Vektors $\mathbf{A}\vec{y}$ lässt sich damit so abschätzen:

$$\left| \sum_{k=1}^n a_{i,k} y_k \right| \leq \sum_{k=1}^n |a_{i,k}| \cdot |y_k| \leq \sum_{k=1}^n |a_{i,k}| \cdot \|\vec{y}\|_\infty = \sum_{k=1}^n |a_{i,k}|,$$

so dass auch $\|\mathbf{A}\|_\infty \leq \max_{1 \leq i \leq m} \sum_{k=1}^n |a_{i,k}|$ gilt.

□

3.7.4.3 Kondition, Stabilität und Konsistenz

Matrix-Normen werden in der numerischen Mathematik z. B. benutzt, um auszudrücken, wie genau die Lösung eines (eindeutig lösbaren) linearen Gleichungssystems oder eines anderen Problems berechnet werden kann. Gegeben sei beispielsweise ein lineares Gleichungssystem $\mathbf{A}\vec{x} = \vec{b}$, $\mathbf{A} \in \mathbb{R}^{n \times n}$ mit $\det \mathbf{A} \neq 0$ und $\vec{b} \in \mathbb{R}^n$. Der exakte Algorithmus (die Abbildung) f berechnet $\vec{x} = f(\vec{b}) := \mathbf{A}^{-1}\vec{b}$. Bei der numerischen Lösung des Problems mit dem Computer werden die gegebenen Daten aber nicht nur mit endlich vielen Bits gespeichert und damit gerundet, sondern in praktischen Anwendungen kommen auch noch Messfehler hinzu. Also hat man es beispielsweise mit einem Vektor $\tilde{\vec{b}}$ statt $\vec{b}$ zu tun. Außerdem werden Probleme oft nur näherungsweise und unter Verwendung der Computer-Arithmetik mit einer Abbildung $\tilde{f}$ und nicht exakt mit f gelöst. Damit erhält man $\tilde{\vec{x}} := \tilde{f}(\tilde{\vec{b}})$ statt $f(\vec{b})$. Diesen Fehler kann man u. a. auf die folgenden beiden Weisen abschätzen:

$$\|\tilde{f}(\tilde{\vec{b}}) - f(\vec{b})\|_p = \|\tilde{f}(\tilde{\vec{b}}) - \tilde{f}(\vec{b}) + \tilde{f}(\vec{b}) - f(\vec{b})\|_p \leq \underbrace{\|\tilde{f}(\tilde{\vec{b}}) - \tilde{f}(\vec{b})\|_p}_{\text{Stabilität}} + \underbrace{\|\tilde{f}(\vec{b}) - f(\vec{b})\|_p}_{\text{Konsistenz}},$$

$$\|\tilde{f}(\tilde{\vec{b}}) - f(\vec{b})\|_p = \|\tilde{f}(\tilde{\vec{b}}) - f(\tilde{\vec{b}}) + f(\tilde{\vec{b}}) - f(\vec{b})\|_p \leq \underbrace{\|\tilde{f}(\tilde{\vec{b}}) - f(\tilde{\vec{b}})\|_p}_{\text{Konsistenz}} + \underbrace{\|f(\tilde{\vec{b}}) - f(\vec{b})\|_p}_{\text{Kondition}}.$$

Die **Konsistenz** ist der Fehler, den der Algorithmus bei exakter Eingabe macht. Die **Stabilität** ist gleichbedeutend mit der Stetigkeit des Algorithmus $\tilde{f}$: Die Abweichung $\|\tilde{f}(\tilde{\vec{b}}) - \tilde{f}(\vec{b})\|_p$ soll klein sein, wenn $\|\tilde{\vec{b}} - \vec{b}\|_p$ klein genug ist. Mittels der Konsistenz können wir die Stabilität auf die Kondition zurückführen:

$$\underbrace{\|\tilde{f}(\tilde{\vec{b}}) - \tilde{f}(\vec{b})\|_p}_{\text{Stabilität}} \leq \underbrace{\|\tilde{f}(\tilde{\vec{b}}) - f(\tilde{\vec{b}})\|_p}_{\text{Konsistenz}} + \underbrace{\|f(\tilde{\vec{b}}) - f(\vec{b})\|_p}_{\text{Kondition}} + \underbrace{\|f(\vec{b}) - \tilde{f}(\vec{b})\|_p}_{\text{Konsistenz}}.$$

Die **Kondition** entspricht der Stetigkeit und damit der Gutmütigkeit der exakten Berechnungsabbildung f und ist eine Eigenschaft des gegebenen Problems, nicht aber des Lösungsalgorithmus. Wie die vorangehende Abschätzung zeigt, beeinflusst sie aber die Stabilität eines Lösungsalgorithmus. Das gewünschte stetige Verhalten des gegebenen Problems können wir so für das Gleichungssystem etwas genauer beschreiben und quantisieren:

$$\|f(\vec{b}) - f(\tilde{\vec{b}})\|_p = \|\mathbf{A}^{-1}\vec{b} - \mathbf{A}^{-1}\tilde{\vec{b}}\|_p = \|\mathbf{A}^{-1}(\vec{b} - \tilde{\vec{b}})\|_p \leq \|\mathbf{A}^{-1}\|_p \|\vec{b} - \tilde{\vec{b}}\|_p.$$

Die nur durch das Gleichungssystem vorgegebene Konstante $\|\mathbf{A}^{-1}\|_p$ wird auch als **absolute Kondition** des Gleichungssystems (hinsichtlich der Norm) bezeichnet. Üblicher ist aber die Verwendung der **relativen Kondition**, bei der die Größen der Inhomogenität und der Lösung $\vec{x} = f(\vec{b}) \neq \vec{0}$ berücksichtigt werden. Unter Verwendung von

$$\|\vec{b}\|_p = \|\mathbf{A}\vec{x}\|_p \leq \|\mathbf{A}\|_p \|\vec{x}\|_p \iff \frac{1}{\|\vec{x}\|_p} \leq \frac{\|\mathbf{A}\|_p}{\|\vec{b}\|_p}$$

ergibt sich

$$\frac{\|\vec{x} - f(\vec{b})\|_p}{\|\vec{x}\|_p} = \frac{\|\mathbf{A}^{-1}\vec{b} - \mathbf{A}^{-1}\tilde{\vec{b}}\|_p}{\|\vec{x}\|_p} \leq \|\mathbf{A}^{-1}\|_p \|\vec{b} - \tilde{\vec{b}}\|_p \frac{1}{\|\vec{x}\|_p} \leq \|\mathbf{A}\|_p \|\mathbf{A}^{-1}\|_p \frac{\|\vec{b} - \tilde{\vec{b}}\|_p}{\|\vec{b}\|_p}.$$

Die Zahl $\|\mathbf{A}\|_p \|\mathbf{A}^{-1}\|_p$ heißt (hinsichtlich der Norm) die **Kondition** der Matrix $\mathbf{A}$. Wählen wir $p = 2$, so ist nach Beispiel 4.20 auf Seite 619 $\|\mathbf{A}\|_2$ gleich der Wurzel aus dem größten Eigenwert $s_{\max}$ von $\mathbf{A}^\top \mathbf{A}$ (und $\mathbf{A}^\top \mathbf{A}$ hat nur reelle, nicht-negative Eigenwerte). Entsprechend ist $\|\mathbf{A}^{-1}\|_2$ die Wurzel des größten Eigenwerts von $(\mathbf{A}^{-1})^\top \mathbf{A}^{-1}$. Wegen Satz 1.29 auf Seite 172 ist

$$(\mathbf{A}^{-1})^\top \mathbf{A}^{-1} = (\mathbf{A}^\top)^{-1} \mathbf{A}^{-1} = (\mathbf{A} \mathbf{A}^\top)^{-1}.$$

Falls nun zusätzlich $\mathbf{A} \mathbf{A}^\top = \mathbf{A}^\top \mathbf{A}$ gilt (z. B. falls $\mathbf{A}$ symmetrisch ist), so sind nach Satz 3.35 a) und c) auf Seite 543 die Eigenwerte von $(\mathbf{A}^{-1})^\top \mathbf{A}^{-1}$ gleich den Kehrwerten der Eigenwerte von $\mathbf{A}^\top \mathbf{A}$. Damit ist $\|\mathbf{A}^{-1}\|_2$ gleich der Wurzel aus dem Kehrwert des kleinsten Eigenwerts $s_{\min} > 0$ von $\mathbf{A}^\top \mathbf{A}$,

$$\|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 = \sqrt{\frac{s_{\max}}{s_{\min}}}.$$

Generell wird der Begriff Kondition in der Numerik für ein Verfahren mit Eingabe und daraus berechneter Ausgabe so verwendet:

$$\text{Kondition} := \frac{\text{relative \AA{nderung der Ausgabe}}}{\text{relative \AA{nderung der Eingabe}}.$$

Die Kondition eines Problems lässt sich durch **Vorkonditionierung** beeinflussen. So kann beispielsweise ein Gleichungssystem $\mathbf{A}\vec{x} = \vec{b}$ durch Multiplikation mit einer invertierbaren Matrix $\mathbf{C}$ in ein neues Gleichungssystem

$$\mathbf{C} \mathbf{A} \vec{x} = \mathbf{C} \vec{b}$$

überführt werden. Betrachten wir auf der rechten Seite beliebige Eingabevektoren $\vec{c}$, so berechnet sich die Kondition des Systems $\mathbf{C} \mathbf{A} \vec{x} = \vec{c}$ über die Matrix $\mathbf{C} \mathbf{A}$, die hoffentlich eine bessere (gleich kleinere) Kondition als $\mathbf{A}$ hat. Eine Strategie besteht darin, $\mathbf{C}$ mit geringem Aufwand so zu bestimmen, dass sich die Matrix „ähnlich wie $\mathbf{A}^{-1}$ “ verhält. Bei der Jacobi-Vorkonditionierung wählt man z. B. $\mathbf{C}$ als Diagonalmatrix, deren Diagonalelemente genau die Kehrwerte der Diagonalelemente von $\mathbf{A}$ sind.

Eine zweite Technik zur Vorkonditionierung eines Gleichungssystems $\mathbf{A}\vec{x} = \vec{b}$ basiert auf der Substitution des Variablenvektors durch $\vec{x} = \mathbf{C}\vec{y}$ mit einer ebenfalls invertierbaren Matrix $\mathbf{C}$. Dann wird das Gleichungssystem $\mathbf{A}\mathbf{C}\vec{y} = \vec{b}$ mit der neuen Matrix $\mathbf{A}\mathbf{C}$ gelöst, so dass man im Nachgang aus der Lösung $\vec{y}$ auch $\vec{x}$ erhält.

Das Verfahren der konjugierten Gradienten (Beispiel 4.23 auf Seite 623) setzt eine symmetrische Matrix voraus. Hier ist es bei der Vorkonditionierung hilfreich, wenn man beide Vorkonditionierungsansätze kombiniert und neben der Substitution $\vec{x} = \mathbf{C}\vec{y}$ die gesamte Gleichung mit $\mathbf{C}^\top$ multipliziert. Dann ist

$$\mathbf{C}^\top \mathbf{A} \mathbf{C} \vec{y} = \mathbf{C}^\top \vec{b}$$

zu lösen. So erhält man zu einer symmetrischen Matrix $\mathbf{A}$ auch eine symmetrische Matrix $\mathbf{C}^\top \mathbf{A} \mathbf{C} = (\mathbf{C}^\top \mathbf{A} \mathbf{C})^\top$.

3.7.4.4 Fixpunktverfahren für lineare Gleichungssysteme

Unter Verwendung einer Matrix-Norm können wir uns nun die Konvergenz des Jacobi- und des Gauß-Seidel-Verfahrens zur iterativen Lösung von Gleichungssystemen noch einmal genauer ansehen. Sie berechnen die Lösung $\vec{x} \in \mathbb{C}^n$ des Gleichungssystems $\mathbf{A}\vec{x} = \vec{b}$ über die Fixpunktiteration

$$\vec{x}^{(k+1)} = S(\vec{x}^{(k)}) := \mathbf{B} \cdot \vec{x}^{(k)} + \vec{c},$$

siehe (1.56) auf Seite 181. Wenn wir zeigen können, dass $S : (\mathbb{C}^n, \|\cdot\|_p) \rightarrow (\mathbb{C}^n, \|\cdot\|_p)$ eine Kontraktion ist, dann folgt die Konvergenz der Verfahren mit dem Banach'schen Fixpunktsatz (Satz 3.42). Wegen

$$\|S(\vec{x}) - S(\vec{y})\|_p = \|\mathbf{B}(\vec{x} - \vec{y})\|_p \leq \|\mathbf{B}\|_p \|\vec{x} - \vec{y}\|_p$$

liegt Konvergenz vor, falls $\|\mathbf{B}\|_p < 1$ ist. Das erkennt man auch an der Fehlerdarstellung (1.58) auf Seite 181, die zur folgenden Fehlerabschätzung führt:

$$\begin{aligned} \|\vec{x}^{(k+1)} - \vec{x}\|_p &= \|\mathbf{B}^{(k+1)}(\vec{x}^{(0)} - \vec{x})\|_p \leq \|\mathbf{B}^{(k+1)}\|_p \cdot \|\vec{x}^{(0)} - \vec{x}\|_p \\ &\stackrel{(3.37)}{\leq} \|\mathbf{B}\|_p^{k+1} \cdot \|\vec{x}^{(0)} - \vec{x}\|_p. \end{aligned}$$

Da man aber den Fixpunkt $\vec{x}$ zunächst nicht kennt, ist die rechte Seite dieser Abschätzung in der Praxis nicht brauchbar. Hier ist die A-priori-Abschätzung (3.34) besser geeignet:

$$\|\vec{x}^{(k+1)} - \vec{x}\|_p \leq \frac{\|\mathbf{B}\|_p^k}{1 - \|\mathbf{B}\|_p} \cdot \|\vec{x}^{(2)} - \vec{x}^{(1)}\|_p.$$

Man benötigt also die tatsächliche Größe der Matrix-Norm. Wir betrachten dazu die etwas einfachere Matrix $\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ des Jacobi-Verfahrens mit ihren Komponenten (siehe Seite 181, $a_{i,i} \neq 0$)

$$b_{i,k} = \begin{cases} -\frac{a_{i,k}}{a_{i,i}}, & i \neq k, \\ 0, & i = k. \end{cases}$$

Um zu einer besonders einfachen Bedingung an $\mathbf{B}$ zu gelangen, verwenden wir die Maximum-Norm $p = \infty$. Man beachte, dass aus einer Konvergenz der Folge $(\vec{x}^{(k)})_{k=1}^{\infty}$ in der $p = \infty$ -Norm automatisch auch die Konvergenz in jeder anderen Norm auf dem endlich-dimensionalen Vektorraum $\mathbb{C}^n$ folgt (vgl. Lemma 3.26). Wegen Lemma 3.29 ist

$$\|\mathbf{B}\|_{\infty} = \max_{1 \leq i \leq n} \sum_{k=1}^n |b_{i,k}| = \max_{1 \leq i \leq n} \frac{1}{|a_{i,i}|} \sum_{k=1, k \neq i}^n |a_{i,k}|.$$

$\|\mathbf{B}\|_{\infty}$ ist kleiner als eins, falls die quadratische Matrix $\mathbf{A}$ **streng diagonaldominant** ist, d. h., wenn für alle $i \in \{1, \dots, n\}$ gilt:

$$|a_{i,i}| > \sum_{k=1, k \neq i}^n |a_{i,k}|.$$

Streng diagonaldominante Matrizen sind invertierbar nach Satz 3.35a) (Seite 543), denn aufgrund des Kreissatzes von Gerschgorin (Satz 3.38 auf Seite 547) besitzen sie keinen Eigenwert null. Wir wissen, dass der Betrag jedes Eigenwerts von $\mathbf{B}$ kleiner $\|\mathbf{B}\|_{\infty}$ ist (Seite 571). Ist $\mathbf{A}$ streng diagonaldominant, dann ist $\|\mathbf{B}\|_{\infty} < 1$, und der Spektralradius der Matrix $\mathbf{B}$ ist kleiner Eins. Das folgt übrigens auch direkt mit dem Kreissatz aus der strengen Diagonaldominanz.

Mit etwas mehr Aufwand als beim Jacobi-Verfahren lässt sich auch für das Gauß-Seidel-Verfahren zeigen, dass es z. B. für strikt diagonaldominante Matrizen $\mathbf{A}$ konvergiert, vgl. (Meister, 2011, S. 76).

Neben dem Jacobi- und Gauß-Seidel-Verfahren kann man leicht weitere Fixpunktiterationen zum Lösen eines Gleichungssystems $\mathbf{A}\vec{x} = \vec{b}$ aufstellen. Dazu benötigt man eine invertierbare Matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$. Mit der Einheitsmatrix $\mathbf{E} \in \mathbb{R}^{n \times n}$ ist

$$\vec{0} = -\mathbf{A}\vec{x} + \vec{b} \iff \mathbf{C}\vec{x} = \mathbf{C}\vec{x} - \mathbf{A}\vec{x} + \vec{b} \iff \vec{x} = (\mathbf{E} - \mathbf{C}^{-1}\mathbf{A})\vec{x} + \mathbf{C}^{-1}\vec{b}.$$

Aus dieser Fixpunktgleichung erhalten wir die Iterationsvorschrift

$$\vec{x}^{(k+1)} = (\mathbf{E} - \mathbf{C}^{-1}\mathbf{A})\vec{x}^{(k)} + \mathbf{C}^{-1}\vec{b}.$$

Konvergenz liegt vor, falls eine Matrix-Norm von $\mathbf{E} - \mathbf{C}^{-1}\mathbf{A}$ kleiner eins ist. Für die spezielle Wahl $\mathbf{C} = \mathbf{E}$ ergibt sich das **Richardson-Verfahren**

$$\vec{x}^{(k+1)} = (\mathbf{E} - \mathbf{A})\vec{x}^{(k)} + \vec{b}.$$

Jacobi- bzw. Gauß-Seidel-Verfahren entstehen aus dem Richardson-Verfahren durch Vorkonditionierung (siehe Seite 574), indem beide Seiten des Gleichungssystems mit den Matrizen $\mathbf{D}^{-1}$ bzw. $(\mathbf{D} + \mathbf{L})^{-1}$ (in den Bezeichnungen von Seite 181) multipliziert werden.

3.7.5 Einige zentrale Sätze der Funktionalanalysis

Wir beginnen mit einem Satz von Stefan Banach und Hugo Steinhaus bzw. Hans Hahn über Operatornormen, der im englischsprachigen Raum als „uniform boundedness principle“ bekannt ist:

Satz 3.43 (Prinzip gleichgradiger Beschränktheit)

Sei X ein Banach-Raum (der nicht nur aus $\vec{0}$ besteht) und Y ein normierter Vektorraum. Weiter sei $\mathcal{L}$ eine Menge beschränkter (d. h. stetiger) linearer Abbildungen von X nach Y . Falls für jeden einzelnen Vektor $\vec{x} \in X$ die Zahlenmenge der Normen der Bilder $\{\|L(\vec{x})\|_Y : L \in \mathcal{L}\}$ beschränkt ist (punktweise Beschränktheit), dann ist auch die Menge der Operatornormen $\{\|L\|_{[X,Y]} : L \in \mathcal{L}\}$ beschränkt.

Man hat also zunächst nur für jedes einzelne $\vec{x} \in X$ eine Schranke. Daraus wird dann eine gemeinsame Schranke für alle Operatornormen kondensiert.

Der Satz eignet sich sehr gut zur Konstruktion von Gegenbeispielen, wenn man ihn indirekt (rückwärts) anwendet: Weiß man, dass die Normen einer Menge von Operatoren nicht beschränkt sind, dann muss es ein Gegenbeispiel $\vec{x} \in X$ geben, für das $\{\|L(\vec{x})\|_Y : L \in \mathcal{L}\}$ nicht beschränkt ist. So werden wir im Kasten auf Seite 842 zeigen, dass nicht jede stetige periodische Funktion als Fourier-Reihe (Reihe von Sinus- und Kosinus-Funktionen) geschrieben werden kann.

Hintergrund: Beweis des Prinzips gleichgradiger Beschränktheit

Es gibt einen nicht-konstruktiven Beweis mit dem Baire'schen Kategoriensatz, siehe z. B. (Heuser, 2006, S. 246 f.). Dabei wird nur die Existenz eines Gegenbeispiels $\vec{x}$ bei nicht beschränkten Operatornormen bewiesen, das Beispiel $\vec{x}$ wird aber nicht angegeben. Neben diesem nicht-konstruktiven Beweis gibt es aber auch eine konstruktive Variante, bei der das Gegenbeispiel $\vec{x}$ mit dem Prinzip des gleitenden Höckers über eine Reihe konstruiert wird. Beim Beweis über einen gleitenden Höcker, wie er ursprünglich von Banach und Hahn geführt wurde, nimmt man an, dass die Operatornormen unbeschränkt sind, dass es also eine Folge $(L_n)_{n=1}^\infty$ von Elementen aus $\mathcal{L}$ gibt, deren Operatornormen gegen unendlich streben. Aus dieser Folge konstruiert man ein Gegenbeispiel als Reihe $\vec{x} = \sum_{k=1}^\infty \vec{x}_{n_k} \in X$, für das $(L_n(\vec{x}))_{n=1}^\infty$ unbeschränkt ist. Dabei wählt man eine Teilfolge $(L_{n_k})_{k=1}^\infty$ und Vektoren $\vec{x}_{n_k}$ so aus, dass $L_{n_k}(\vec{x}_{n_k})$ für $k \rightarrow \infty$ gegen unendlich strebt und L_{n_k} , angewendet auf die anderen Summanden $\vec{x}_{n_i}$, $i \neq k$, keinen signifikanten Beitrag liefert. In diesem Sinne ist für L_{n_k} der Summand $\vec{x}_{n_k}$ der Höcker; hier passiert etwas, während sich die anderen Summanden für L_{n_k} gutmütig verhalten.

In Sokal (2011) finden Sie einen vereinfachten konstruktiven Beweis, bei dem eine Cauchy-Folge aufgebaut wird, die gegen das Gegenbeispiel konvergiert. Dabei dient quasi der gesamte Reihenbeginn als Höcker, wenn man die Cauchy-Folge als Reihe auffasst.

Dieser elementare Beweis wird hier in etwas modifizierter Form wiedergegeben. Wir beginnen mit einer Abschätzung der Operatornorm:

Lemma 3.30 (Abschätzung der Operatornorm)

Sei $L : X \rightarrow Y$ eine beschränkte lineare Abbildung vom normierten Vektorraum $X \neq \{\vec{0}\}$ in den normierten Vektorraum Y . Für jedes $\vec{x} \in X$ und jedes $R > 0$ gilt:

$$\sup_{\vec{u} \in X: \|\vec{u}\|_X = R} \|L(\vec{x} + \vec{u})\|_Y \geq R \cdot \|L\|_{[X,Y]}. \quad (3.38)$$

Beweis: Sei $\vec{v} \in X$ mit $\|\vec{v}\|_X = R$. Dann ist auch $\|\vec{v}\|_X = R$, und es gilt unter Verwendung der Dreiecksungleichung der Norm:

$$\begin{aligned} \sup_{\vec{u} \in X: \|\vec{u}\|_X = R} \|L(\vec{x} + \vec{u})\|_Y &\geq \max\{\|L(\vec{x} + \vec{v})\|_Y, \|L(\vec{x} - \vec{v})\|_Y\} \\ &\geq \frac{1}{2} [\|L(\vec{x} + \vec{v})\|_Y + \|L(\vec{x} - \vec{v})\|_Y] \geq \frac{1}{2} [\|L(\vec{x} + \vec{v}) - L(\vec{x} - \vec{v})\|_Y] \\ &= \frac{1}{2} [\|L(\vec{x} + \vec{v} - \vec{x} + \vec{v})\|_Y] = \|L(\vec{v})\|_Y. \end{aligned} \quad (3.39)$$

Nun liefert der Übergang zum Supremum über alle $\vec{v} \in X$ mit $\|\vec{v}\|_X = R$ die Aussage (3.38), da zu $\vec{0} \neq \vec{u} \in X$ der Vektor $\vec{v} = R \frac{\vec{u}}{\|\vec{u}\|_X}$ die Norm R hat:

$$\|L\|_{[X,Y]} = \sup_{\vec{0} \neq \vec{u} \in X} \frac{\|L(\vec{u})\|_Y}{\|\vec{u}\|_X} = \sup_{\|\vec{v}\|_X = R} \frac{\|L(\vec{v})\|_Y}{\|\vec{v}\|_X} = \frac{1}{R} \sup_{\|\vec{v}\|_X = R} \|L(\vec{v})\|_Y, \quad (3.40)$$

also

$$\sup_{\vec{u} \in X: \|\vec{u}\|_X = R} \|L(\vec{x} + \vec{u})\|_Y \stackrel{(3.39)}{\geq} \sup_{\|\vec{v}\|_X = R} \|L(\vec{v})\|_Y \stackrel{(3.40)}{=} R \cdot \|L\|_{[X,Y]}.$$

□

Damit zeigen wir das Prinzip gleichgradiger Beschränktheit mittels Widerspruch.

Beweis: Falls die Menge der Operatornormen unbeschränkt ist, dann können wir eine Folge von Abbildungen wählen mit

$$\|L_n\|_{[X,Y]} \geq 4^n. \quad (3.41)$$

Darauf wenden wir das Lemma iterativ beginnend mit $\vec{x}_0 := \vec{0}$ an, indem wir zu einem $\vec{x}_n$ einen nächsten Vektor $\vec{x}_{n+1} := \vec{u}_n + \vec{x}_n$ so wählen, dass

$$\|\vec{u}_n\|_X = 3^{-n} =: R_n \quad (3.42)$$

und wir mindestens $\frac{3}{4}$ des Supremums in (3.38) erhalten:

$$\|L_{n+1}(\vec{x}_{n+1})\|_Y = \|L_{n+1}(\vec{x}_n + \vec{u}_n)\|_Y \geq \frac{3}{4} R_n \|L_{n+1}\|_{[X,Y]} = \frac{3}{4} 3^{-n} \|L_{n+1}\|_{[X,Y]}. \quad (3.43)$$

Die so konstruierte Folge ist eine Cauchy-Folge, da für $n > m$ unter Verwendung der geometrischen Summe (1.10) von Seite 60 gilt:

$$\begin{aligned} \|\vec{x}_n - \vec{x}_m\|_X &= \left\| \sum_{k=m}^{n-1} \vec{u}_k \right\|_X \leq \sum_{k=m}^{n-1} \|\vec{u}_k\|_X \\ &\stackrel{(3.42)}{=} \sum_{k=m}^{n-1} 3^{-k} < \sum_{k=m}^{\infty} 3^{-k} = \frac{1}{1 - \frac{1}{3}} - \frac{1 - \frac{1}{3^m}}{1 - \frac{1}{3}} \rightarrow 0, \quad m \rightarrow \infty. \end{aligned}$$

Denn so lässt sich zu jedem $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$ wählen, so dass für alle $n > m > n_0$ gilt: $\|\vec{x}_n - \vec{x}_m\|_X < \varepsilon$. Da X ein Banach-Raum ist, konvergiert die Cauchy-Folge in X gegen einen Grenzwert

$$\vec{x} = \vec{x}_0 + \lim_{l \rightarrow \infty} \sum_{k=0}^l \vec{u}_k = \vec{x}_0 + \sum_{k=0}^{\infty} \vec{u}_k = \sum_{k=0}^{\infty} \vec{u}_k = \vec{x}_n + \sum_{k=n}^{\infty} \vec{u}_k.$$

Für das Gegenbeispiel $\vec{x}$ ist die punktweise Beschränktheit verletzt:

$$\begin{aligned} \|L_n(\vec{x})\|_Y &= \left\| L_n \left(\vec{x}_n + \sum_{k=n}^{\infty} \vec{u}_k \right) \right\|_Y \geq \underbrace{\|L_n(\vec{x}_n)\|_Y}_{\text{Höcker}} - \underbrace{\sum_{k=n}^{\infty} \|L_n(\vec{u}_k)\|_Y}_{\text{kleiner Rest}} \\ &\stackrel{(3.43)}{\geq} \frac{3}{4} 3^{-n+1} \|L_n\|_{[X,Y]} - \sum_{k=n}^{\infty} \|L_n\|_{[X,Y]} \|\vec{u}_k\|_X \stackrel{(3.42)}{=} \|L_n\|_{[X,Y]} \left[\frac{3}{4} 3^{-n+1} - \sum_{k=n}^{\infty} 3^{-k} \right] \\ &= \|L_n\|_{[X,Y]} \left[\frac{3}{4} 3^{-n+1} - \frac{3}{2} + \frac{3}{2} \left(1 - \frac{1}{3^n} \right) \right] = \|L_n\|_{[X,Y]} \left[\frac{3}{4} 3^{-n+1} - \frac{3}{2} 3^{-n} \right] \\ &= \|L_n\|_{[X,Y]} \frac{3}{4} 3^{-n} \stackrel{(3.41)}{\geq} 4^n \frac{3}{4} 3^{-n} = \left(\frac{4}{3} \right)^{n-1} \rightarrow \infty, \quad n \rightarrow \infty, \end{aligned}$$

im Widerspruch zur Voraussetzung der punktweisen Beschränktheit. ■

Das Prinzip der gleichgradigen Beschränktheit hilft aber nicht nur bei solch negativen Aussagen. Es gibt beispielsweise in der digitalen Signalverarbeitung bzw. Regelungstechnik unterschiedlich erscheinende Stabilitätsbegriffe, die aber so unterschiedlich gar nicht sind, wie wir mit Satz 3.43 zeigen.

Ein lineares zeitinvariantes Abtastsystem überführt eine Eingangsfolge $(x_k)_{k=0}^{\infty}$ in eine Ausgangsfolge $(y_k)_{k=0}^{\infty}$ als Lösung einer sogenannten Differenzengleichung

$$y_k + a_{n-1}y_{k-1} + \cdots + a_0y_{k-n} = b_nx_k + b_{n-1}x_{k-1} + \cdots + b_0x_{k-n}, \quad (3.44)$$

wobei $x_k = y_k := 0$ für $k < 0$.

Tatsächlich sieht man in dieser Gleichung keine Differenzen. Wenn man aber Ableitungen durch Differenzen annähert, dann bekommt man Linearkombinationen von Abtastwerten, wie sie auf beiden Seiten der Gleichung mit den allgemeinen Koeffizienten angegeben sind. Man beachte, dass die Differenzengleichung für jede Eingangsfolge

$(x_k)_{k=0}^\infty$ eine eindeutige Lösung hat, die sukzessive beginnend mit y_0 berechnet werden kann ($n \geq 2$):

$$\begin{aligned} y_0 &= b_n x_0, \\ y_1 &= -a_{n-1} y_0 + b_n x_1 + b_{n-1} x_0, \\ y_2 &= -a_{n-1} y_1 - a_{n-2} y_0 + b_n x_2 + b_{n-1} x_1 + b_{n-2} x_0, \dots \end{aligned}$$

Über die eindeutige Lösung der Differenzengleichung wird eine lineare Abbildung L definiert, die einer Eingangsfolge die Ausgangsfolge zuordnet: $L((x_k)_{k=0}^\infty) = (y_k)_{k=0}^\infty$. Die Abbildung ist auf dem Vektorraum der (reellen) Folgen definiert, und sie bildet in den gleichen Vektorraum ab. Tatsächlich ist L eine **lineare** Abbildung, da jede Linearkombination von Folgen auf die Linearkombination ihrer Bilder abgebildet wird. Das passt zur Bezeichnung **lineare** zeitinvariante Systeme. Das Attribut **zeitinvariant** rührt daher, dass eine (nach rechts) verschobene Eingangsfolge zu einer entsprechend verschobenen Ausgangsfolge führt.

In der Regelungstechnik werden die Ausgangsfolgen häufig mit der Z-Transformation berechnet, siehe Beispiel 2.29 auf Seite 249. Arbeitet man nicht mit den Differenzen diskreter Abtastwerte, sondern mit kontinuierlichen Signalen, so spricht man von linearen zeitinvarianten Übertragungssystemen, die häufig mittels Laplace-Transformation berechnet werden, siehe Kapitel 6.3.3.

Ein lineares Abtastsystem heißt **BIBO-stabil** („bounded input – bounded output“) genau dann, wenn zu jeder beschränkten Eingangsfolge $(x_k)_{k=0}^\infty$ auch die zugehörige Ausgangsfolge $(y_k)_{k=0}^\infty := L((x_k)_{k=0}^\infty)$ beschränkt ist.

Wir zeigen mit dem Prinzip gleichgradiger Beschränktheit, dass die BIBO-Stabilität gleichbedeutend ist mit der Stetigkeit bzw. Beschränktheit der linearen Abbildung $L : l^\infty \rightarrow l^\infty$:

Satz 3.44 (BIBO-Stabilität und Stetigkeit bei Abtastsystemen)

Ein lineares zeitinvariantes Abtastsystem ist genau dann BIBO-stabil, wenn es eine von den Eingangsfolgen unabhängige Konstante C gibt, so dass für jedes Paar aus einer beschränkten Eingangsfolge $(x_k)_{k=0}^\infty$ und der zugehörigen Ausgangsfolge $(y_k)_{k=0}^\infty = L((x_k)_{k=0}^\infty)$ gilt:

$$\|(y_k)_{k=0}^\infty\|_\infty := \sup_{k \in \mathbb{N}_0} |y_k| \leq C \|(x_k)_{k=0}^\infty\|_\infty := C \sup_{k \in \mathbb{N}_0} |x_k|, \quad (3.45)$$

d. h., L ist beschränkt mit $\|L\|_{[l^\infty, l^\infty]} \leq C$.

Beweis: Falls eine Konstante C wie im Satz angegeben existiert, dann folgt sofort die Definition der BIBO-Stabilität. Die umgekehrte Richtung ist wesentlich schwieriger. Ihr Beweis ist eine Konsequenz aus dem Prinzip gleichgradiger Beschränktheit. Wir betrachten dazu eine Folge linearer Funktionale $L_m : l^\infty \rightarrow \mathbb{R}$ auf dem Banach-Raum

der beschränkten Folgen l^∞ mit Werten im Banach-Raum der reellen Zahlen, die über $L_m((x_k)_{k=0}^\infty) := L((x_k)_{k=0}^\infty)_m = y_m$ definiert ist. L_m liefert also das m -te Folgenglied der Ausgangsfolge.

Wir zeigen zunächst, dass jedes dieser Funktionale L_m beschränkt (und damit stetig) ist. Zu zeigen ist also, dass es eine Konstante C_m gibt, so dass

$$|L_m((x_k)_{k=0}^\infty)| \leq C_m \|(x_k)_{k=0}^\infty\|_\infty \text{ für alle Folgen } (x_k)_{k=0}^\infty \in l^\infty.$$

Aufgrund der Differenzengleichung gilt:

$$\begin{aligned} |L_m((x_k)_{k=0}^\infty)| &= |y_m| = |-a_{n-1}y_{m-1} - \cdots - a_0y_{m-n} + b_nx_m + b_{n-1}x_{m-1} + \cdots + b_0x_{m-n}| \\ &\leq \max\{|y_0|, \dots, |y_{m-1}|\} \sum_{k=0}^{n-1} |a_k| + \|(x_k)_{k=0}^\infty\|_\infty \sum_{k=0}^n |b_k|. \end{aligned}$$

Wir schätzen nun den größten Wert von $|y_0|, \dots, |y_{m-1}|$ entsprechend ab und fahren iterativ fort, bis der größte Wert $|y_0| = |b_n||x_0| \leq |b_n| \|(x_k)_{k=0}^\infty\|_\infty$ ist. Grob abgeschätzt ergibt sich

$$\begin{aligned} |L_m((x_k)_{k=0}^\infty)| &\leq \left[\sum_{i=0}^m \left(\sum_{k=0}^{n-1} |a_k| \right)^i \right] \left[\sum_{k=0}^n |b_k| \right] \|(x_k)_{k=0}^\infty\|_\infty \\ &=: C_m \|(x_k)_{k=0}^\infty\|_\infty. \end{aligned}$$

Die linearen Funktionale L_m sind also einzeln beschränkt und damit Elemente des Dualraums von l^∞ . Wir wissen noch nichts über das Verhalten der Folge der Konstanten C_m . Um hier weiterzukommen, benötigen wir die Voraussetzung der Stabilität.

Stabilität bedeutet, dass für jede Eingangsfolge $(x_k)_{k=0}^\infty \in l^\infty$ die zugehörige Betragsfolge der Ausgangsfolge $(|L_n((x_k)_{k=0}^\infty)|)_{n=0}^\infty$ beschränkt ist. Damit besagt das Prinzip der gleichgradigen Beschränktheit, dass die Menge $\{\|L_n\|_{[l^\infty, \mathbb{R}]}\}$ der Operatornormen beschränkt ist. Es gibt also eine von n und der Eingangsfolge unabhängige Schranke $C \in \mathbb{R}$, so dass

$$|y_n| = |L_n((x_k)_{k=0}^\infty)| \leq C \|(x_k)_{k=0}^\infty\|_\infty$$

ist. Damit haben wir $\|(y_n)_{n=0}^\infty\|_\infty \leq C \|(x_k)_{k=0}^\infty\|_\infty$ bewiesen. $\square$

Weitere positive Anwendungen des Prinzips gleichgradiger Beschränktheit sind Konvergenzuntersuchungen. Sei dazu $(L_n)_{n=1}^\infty$ eine Folge von linearen beschränkten Abbildungen von einem Banach-Raum X in sich selbst, $L_n : X \rightarrow X$. Falls für jedes $\vec{x} \in X$ gilt:

$$\lim_{n \rightarrow \infty} \|L_n(\vec{x}) - \vec{x}\|_X = 0, \quad (3.46)$$

so nennen wir die Folge einen **Approximationsprozess** auf X . Hier haben wir die Vorstellung, dass z. B. $L_n(\vec{x})$ ein einfacheres Objekt als $\vec{x}$ ist, z. B. ein Polynom vom Grad n , wobei $\vec{x}$ eine stetige Funktion sein kann. Hat man so einen Approximationsprozess, dann sind die Operatornormen nach dem Prinzip der gleichgradigen Beschränktheit gemeinsam beschränkt. Denn für jedes einzelne $\vec{x} \in X$ konvergiert die Folge $\|L_n(\vec{x})\|_X$ gegen $\|\vec{x}\|_X$ und ist damit beschränkt. Hat man umgekehrt die Beschränktheit der Operatornormen, so kann man daraus unter Zusatzbedingungen auf (3.46) schließen. Ohne Zusatzbedingungen kann es nicht funktionieren, da z. B. $L_n(\vec{x}) := \vec{0}$ keinen Approximationsprozess bilden. Als Zusatzeigenschaft genügt es aber, wenn (3.46) nicht für alle $\vec{x}$, sondern nur für alle $\vec{x} \in U$ für einen Untervektorraum $U \subset X$ gilt, der **dicht** in X ist. U ist dicht in X genau dann, wenn es zu jedem $\vec{x} \in X$ und jedem $\varepsilon > 0$ ein $\vec{u} \in U$ gibt, so dass $\|\vec{x} - \vec{u}\|_X < \varepsilon$. Man kommt also mit Vektoren aus U beliebig nahe an jedes Element von X heran:

Folgerung 3.4 (Satz von Banach-Steinhaus)

Sei $X \neq \{\vec{0}\}$ ein Banach-Raum. Eine Folge $(L_n)_{n=1}^\infty$ von linearen beschränkten Abbildungen $L_n : X \rightarrow X$ ist genau dann ein Approximationsprozess, wenn

- die Folge der Operatornormen beschränkt ist mit $\|L_n\|_{[X,X]} \leq M$ für alle $n \in \mathbb{N}$
- und es einen in X dichten Untervektorraum U gibt mit

$$\lim_{n \rightarrow \infty} \|L_n(\vec{u}) - \vec{u}\|_X = 0 \text{ für alle } \vec{u} \in U.$$

Beweis: Die schwierige Richtung, dass ein Approximationsprozess die beiden Bedingungen der Folgerung erfüllt, haben wir bereits auf das Prinzip der gleichgradigen Beschränktheit zurückgeführt (Konvergenz für alle $\vec{u} \in U$ ist trivial, wenn sie für alle $\vec{x} \in X$ vorliegt). Jetzt müssen wir umgekehrt zeigen, dass die beiden Bedingungen zu einem Approximationsprozess führen. Sei dazu $\vec{x} \in X$ und $\varepsilon > 0$. Dann gibt es ein $\vec{u} \in U$ und ein $n_0 \in \mathbb{N}$, so dass

$$\begin{aligned} \|\vec{x} - \vec{u}\|_X &< \frac{\varepsilon}{2[M+1]}, \\ \|L_n(\vec{u}) - \vec{u}\|_X &< \frac{\varepsilon}{2} \text{ für alle } n > n_0. \end{aligned}$$

Zusammen erhalten wir für $n > n_0$:

$$\begin{aligned} \|L_n(\vec{x}) - \vec{x}\|_X &= \|L_n(\vec{x} - \vec{u}) - (\vec{x} - \vec{u}) + L_n(\vec{u}) - \vec{u}\|_X \\ &\leq [\|L_n\|_{[X,X]} + 1]\|\vec{x} - \vec{u}\|_X + \|L_n(\vec{u}) - \vec{u}\|_X \leq [M+1]\frac{\varepsilon}{2[M+1]} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Da $\varepsilon > 0$ beliebig gewählt werden kann, ist die Konvergenz für alle $\vec{x} \in X$ bewiesen. $\square$

Mit dem Satz von Banach-Steinhaus kann man z. B. zeigen, dass Fourier-Reihen in den L^p -Räumen für $1 < p < \infty$ konvergieren, vgl. Kapitel 6.1.5.

Stetige Abbildungen haben genau die Eigenschaft, dass die Urbilder offener Mengen ebenfalls offen sind (siehe Definition 2.19). Der folgende Satz kümmert sich um die umgekehrte Richtung (Heuser, 2006, S. 241 ff.):

Satz 3.45 (Satz von Banach-Schauder, Satz über die offene Abbildung)

Seien $X \neq \{\vec{0}\}$ und Y Banach-Räume und L eine beschränkte (d. h. stetige) lineare Abbildung von X in Y . Falls L surjektiv ist (d. h., jedes Element von Y kommt als Bild vor), dann bildet L offene Mengen auf offene Mengen (hinsichtlich den von den Normen induzierten Topologien) ab.

Eine beschränkte bijektive Abbildung L von einem Banach-Raum X auf einen Banach-Raum Y besitzt eine lineare Umkehrabbildung L^{-1} . Als direkte Konsequenz des Satzes von Banach-Schauder ist L^{-1} stetig (d. h. beschränkt), da hinsichtlich L^{-1} die Urbilder offener Mengen wieder offen sind.

Der nächste Hauptsatz widmet sich der Fortsetzbarkeit beschränkter Funktionale:

Satz 3.46 (Satz von Hahn-Banach)

Sei X ein normierter Vektorraum und U ein Unterraum von X . Ist $F : U \subset X \rightarrow \mathbb{R}$ ein lineares Funktional, so dass eine Konstante $C \in [0, \infty[$ existiert mit

$$|F(\vec{x})| \leq C \|\vec{x}\|_X \text{ für alle } \vec{x} \in U,$$

dann kann F von U auf ganz X unter Beibehaltung der Schranke fortgesetzt werden: Es gibt ein lineares Funktional $\tilde{F} : X \rightarrow \mathbb{R}$ mit

$$\begin{aligned} F(\vec{x}) &= \tilde{F}(\vec{x}) \text{ für alle } \vec{x} \in U, \\ |\tilde{F}(\vec{x})| &\leq C \|\vec{x}\|_X \text{ für alle } \vec{x} \in X. \end{aligned}$$

Im Beweis des Satzes werden aus der Definition der Norm nur die absolute Homogenität und die Dreiecksungleichung verwendet. Daher findet man ihn auch in allgemeinerer Form, bei der die Normen durch schwächere Halbnormen oder sublineare Funktionale ersetzt sind. Ein Beweis steht z. B. in (Heuser, 2006, S. 228 f.).

In Hilbert-Räumen haben beschränkte lineare Funktionale eine besonders einfache Darstellung, siehe z. B. (Heuser, 2006, S. 319 f.):

Satz 3.47 (Riesz'scher Darstellungssatz)

Sei $H \neq \{\vec{0}\}$ ein Hilbert-Raum mit dem Skalarprodukt $(\cdot, \cdot)_H$ und F ein beschränktes lineares Funktional auf H . Dann gibt es genau ein Element $\vec{f} \in H$, so dass

$$F(\vec{x}) = (\vec{f}, \vec{x})_H \text{ für alle } \vec{x} \in H.$$

Damit gilt außerdem $\|F\|_{H^*} = \|\vec{f}\|_H$.

Da umgekehrt durch einen Vektor $\vec{f} \in H$ ein beschränktes lineares Funktional $F(\vec{x}) := (\vec{f}, \vec{x})_H$ definiert wird, gibt es eine bijektive, beschränkte lineare Abbildung zwischen H und dem Dualraum H^* . In diesem Sinne entspricht der Hilbert-Raum seinem Dualraum.

Der Darstellungssatz hat praktische Anwendungen, z. B. beim Lösen von Differenzialgleichungen mit der Finite-Elemente-Methode (siehe Kapitel 5.5.3). Dazu benötigt man aber Ableitungen in L^p -Räumen.

3.7.6 Sobolev-Räume

Um sinnvoll in den L^p -Räumen einen Ableitungsbegriff einzuführen, verallgemeinern wir die klassische Definition der Ableitung mittels partieller Integration. Sei φ eine auf $[a, b]$ beliebig oft differenzierbare Funktion (Testfunktion) mit $\varphi^{(k)}(a) = \varphi^{(k)}(b) = 0$ für alle $k \in \mathbb{N}_0$, dann gilt für eine n -mal stetig differenzierbare Funktion f :

$$\begin{aligned} & \int_a^b f^{(n)}(x) \varphi(x) dx \\ &= [f^{(n-1)}(x) \varphi(x)]_a^b - \int_a^b f^{(n-1)}(x) \varphi'(x) dx = (-1)^1 \int_a^b f^{(n-1)}(x) \varphi'(x) dx \\ &= -[f^{(n-2)}(x) \varphi'(x)]_a^b + \int_a^b f^{(n-2)}(x) \varphi^{(2)}(x) dx = (-1)^2 \int_a^b f^{(n-2)}(x) \varphi^{(2)}(x) dx \\ &= \dots = (-1)^n \int_a^b f(x) \varphi^{(n)}(x) dx. \end{aligned}$$

Die rechte Seite macht auch dann Sinn, wenn f gar nicht differenzierbar ist. Diese wird jetzt genutzt, um eine schwache Ableitung zu definieren. Später verwenden wir im Kasten auf Seite 898 den gleichen Trick, um Ableitungen für verallgemeinerte Funktionen (Distributionen) einzuführen.

Definition 3.36 (Schwache Ableitung)

Eine Lebesgue-integrierbare Funktion f heißt auf einem Intervall $[a, b]$ **schwach differenzierbar** mit **schwacher Ableitung** g genau dann, wenn g Lebesgue-

integrierbar ist und für alle Testfunktionen φ , d. h. für alle auf $[a, b]$ beliebig oft differenzierbaren Funktionen φ mit $\varphi^{(k)}(a) = \varphi^{(k)}(b) = 0$ für alle $k \in \mathbb{N}_0$, gilt:

$$\int_a^b g(x) \varphi(x) dx = - \int_a^b f(x) \varphi'(x) dx.$$

Entsprechend heißt g eine schwache n -te Ableitung von f auf $[a, b]$ genau dann, wenn für alle Testfunktionen φ gilt:

$$\int_a^b g(x) \varphi(x) dx = (-1)^n \int_a^b f(x) \varphi^{(n)}(x) dx.$$

Hat man zwei schwache n -te Ableitungen g und h einer Funktion f , dann gilt laut Definition $\int_a^b [g(x) - h(x)] \varphi(x) dx = 0$. Man kann sich überlegen, dass diese Gleichung nur dann für alle Testfunktionen φ erfüllt sein kann, falls $g = h$ f. ü. Damit sind schwache Ableitungen in diesem Sinne eindeutig. Hinsichtlich der Notation unterscheidet man nicht zwischen den „normalen“ und den schwachen Ableitungen. Existiert eine klassische Ableitung, dann stimmt diese mit der entsprechenden schwachen überein. Das haben wir eingangs mit der partiellen Integration nachgerechnet.

Definition 3.37 (Sobolev-Räume)

Sei $1 \leq p < \infty$. Der Vektorraum der Funktionen $f \in L^p[a, b]$, die k -te schwache Ableitungen $f^{(k)}$ für $1 \leq k \leq n$ besitzen, die ebenfalls in $L^p[a, b]$ enthalten sind, heißt **Sobolev-Raum** $W^{n,p}[a, b]$. Er besitzt die Norm

$$\|f\|_{n,p} := \left[\sum_{k=0}^n \|f^{(k)}\|_p^p \right]^{\frac{1}{p}}.$$

Tatsächlich sind die Sobolev-Räume vollständige normierte Vektorräume, also Banach-Räume. Nachdem wir intensiver auf Funktionen mit mehreren Variablen eingegangen sind, geben wir in Kapitel 4.4.5.2 eine umfassendere Definition der L^p - und Sobolev-Räume und fassen weitere Eigenschaften zusammen.

3.8 Aufgaben

Aufgabe 3.1

Im Rechteck mit den Eckpunkten $O = (0, 0)$, $A = (2, 0)$, $B = (2, 8)$ und $C = (0, 8)$ sei P der Mittelpunkt der Strecke $\overline{BC}$ und Q der Mittelpunkt der Strecke $\overline{AB}$. Wie lautet die Komponentendarstellung der Vektoren $\vec{OP}$, $\vec{OQ}$, $\vec{AC}$, $\vec{QP} \in \mathbb{R}^2$?

Aufgabe 3.2

Welche Werte kann a_1 annehmen, wenn $\vec{a} = (a_1, 3, -2)^\top$ den Betrag $|\vec{a}| = 4$ besitzt?

Aufgabe 3.3

Gegeben sind $\vec{a} = (1, -1, 1)^\top$, $\vec{b} = (2, 0, 1)^\top$ und $\vec{c} = (1, 0, -1)^\top$. Berechnen Sie:

- a) die Skalarprodukte $\vec{a} \cdot \vec{b}$, $\vec{b} \cdot \vec{c}$ und $\vec{a} \cdot \vec{c}$,
 b) die Vektorprodukte $\vec{a} \times \vec{c}$, $\vec{b} \times \vec{c}$ und $(\vec{a} + \vec{b}) \times \vec{c}$, c) das Spatprodukt $[\vec{a} \times \vec{c}] \cdot \vec{b}$.

Aufgabe 3.4

Die Summe von $\vec{a}$ mit einem Vielfachen des Vektors $\vec{b}$ soll senkrecht zum Vektor $\vec{c}$ stehen. Wie muss das Vielfache gewählt werden? Berechnen Sie eine allgemeine Lösung, und wenden Sie das Ergebnis auf die konkreten Vektoren $\vec{a} = (1, 1, 1)^\top$, $\vec{b} = (0, 1, -1)^\top$ und $\vec{c} = (-1, 3, 5)^\top$ an.

Aufgabe 3.5

Berechnen Sie alle Vektoren, die senkrecht zum Vektor $\vec{a} = (0, 1, 2)^\top$ stehen.

Aufgabe 3.6

Ermitteln Sie die Projektion $\vec{u}$ und das Lot $\vec{v}$ von $\vec{a} = (1, 2, 3)^\top$ auf $\vec{b} = (3, 2, 1)^\top$.

Aufgabe 3.7

Bestimmen Sie die Hesse'sche Normalform der Geraden $g: 8x + 15y + 34 = 0$ in $\mathbb{R}^2$.

Aufgabe 3.8

Gegeben sei die Gerade $g: 3x + 4y + 15 = 0$ in $\mathbb{R}^2$. Welche der Punkte $P_1 = (-5, 0)$, $P_2 = (-9, 3)$ und $P_3 = (1, 0)$ liegen auf g ? Berechnen Sie den (kürzesten) Abstand der übrigen Punkte zur Geraden.

Aufgabe 3.9

Durch $P_1 = (3, 2, 2)$ und $P_2 = (6, 3, 4)$ verläuft eine Gerade g . Bestimmen Sie:

- a) eine Parameterdarstellung von g ,
 b) die Koordinaten des Schnittpunkts von g mit der y - z -Ebene,
 c) die Koordinaten des Fußpunkts des Lots von $P_0 = (1, 1, 1)$ auf g ,
 d) den Abstand d des Punkts P_0 von g .

Aufgabe 3.10

Beweisen Sie die folgenden Regeln für das Vektorprodukt:

- a) $\vec{a} + \vec{b} + \vec{c} = \vec{0} \implies \vec{a} \times \vec{b} = \vec{b} \times \vec{c} = \vec{c} \times \vec{a}$,
 b) $\vec{a} \perp \vec{b} \implies \vec{a} \times (\vec{a} \times (\vec{a} \times (\vec{a} \times \vec{b}))) = |\vec{a}|^4 \vec{b}$.

Aufgabe 3.11

Berechnen Sie den Schnittpunkt der drei Ebenen $x + y - z = 1$, $x - 4y + z = -2$, $x + y + z = 3$. Eine weitere Ebene soll durch diesen Schnittpunkt und parallel zur Ebene $x + 2y + 3z = 0$ verlaufen. Bestimmen Sie die Punkt-Richtungsform von E .

Aufgabe 3.12

Durch $P_1 = (0, 0, 1)$, $P_2 = (1, -1, 0)$ und $P_3 = (1, 1, 1)$ verläuft im $\mathbb{R}^3$ die Ebene E .

- Bestimmen Sie die Punkt-Richtungsform von E .
- Berechnen Sie eine (Koordinaten-)Gleichung von E .
- Ermitteln Sie die Hesse'sche Normalform von E .

Aufgabe 3.13

Bestimmen Sie zu den Ebenen $E_1 : 4x + 3y - 9z = 6$ und $E_2 : x + 12y - 6z = 9$ die Punkt-Richtungsform der Schnittgeraden g . Berechnen Sie außerdem den Winkel zwischen 0° und 90° zwischen den Normalenvektoren der beiden Ebenen, den wir als Winkel zwischen den beiden Ebenen auffassen.

Aufgabe 3.14

Zeigen Sie, dass die folgenden Vektoren in $(\mathbb{C}^3, +; \mathbb{C}, \cdot)$ linear abhängig sind:

$$\text{a) } \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{b) } \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad \text{c) } \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Aufgabe 3.15

Wir untersuchen die Menge L aller auf $]0, \infty[$ definierten Logarithmenfunktionen einschließlich der Nullfunktion, d. h.

$$f \in L \iff f(x) = 0 \quad \text{oder} \quad f(x) = \log_a(x), \quad x \in]0, \infty[, \quad \text{mit} \quad a > 0, \quad a \neq 1.$$

Zeigen Sie, dass L einen Unterraum des Vektorraums der auf $]0, \infty[$ definierten reellen Funktionen bildet.

Aufgabe 3.16

Untersuchen Sie die folgenden Vektoren auf lineare Unabhängigkeit im Vektorraum $(\mathbb{R}^3, +; \mathbb{R}, \cdot)$ bzw. im Vektorraum der reellen Funktionen auf $\mathbb{R}$, und berechnen Sie die lineare Hülle, d. h. den von den Vektoren erzeugten Unterraum:

- $\vec{a} = (-1, 1, 1)^\top$, $\vec{b} = (1, -1, -1)^\top$.
- $\vec{a} = (1, 1, -1)^\top$, $\vec{b} = (-1, 0, 1)^\top$, $\vec{c} = (2, 1, -3)^\top$.
- $f(x) = e^x$, $g(x) = e^{-x}$. d) $f(x) = e^x$, $g(x) = e^{-x}$, $h(x) = \cosh(x)$.

Aufgabe 3.17

In dieser Aufgabe beschäftigen wir uns mit Koordinatentransformationen. Dazu untersuchen wir, wie sich die Darstellung eines Vektors bezüglich einer Basis beim Übergang zu einer anderen Basis ändert. Wir beschränken uns hier auf zwei Basen des $\mathbb{R}^3$, die wir $A = \{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ und $B = \{\vec{b}_1, \vec{b}_2, \vec{b}_3\}$ nennen. Ein beliebiger Vektor $\vec{c} \in \mathbb{R}^3$ besitzt jeweils eine eindeutige Linearkombination bzgl. A und B :

$$\vec{c} = x_1 \vec{a}_1 + x_2 \vec{a}_2 + x_3 \vec{a}_3 \quad \text{bzw.} \quad \vec{c} = y_1 \vec{b}_1 + y_2 \vec{b}_2 + y_3 \vec{b}_3,$$

wobei die Faktoren x_k bzw. y_k die Koordinaten bezüglich der jeweiligen Basis sind.

- a) Bestimmen Sie die Matrix $\mathbf{M}_{AB}$, die die Koordinaten $(x_1, x_2, x_3)^\top$ von $\vec{c}$ bezüglich der Basis A in die Koordinaten $(y_1, y_2, y_3)^\top$ von $\vec{c}$ bezüglich B transformiert.
- b) Bestimmen Sie die Matrix $\mathbf{M}_{BA}$, die die Koordinaten von $\vec{c}$ bezüglich der Basis B in die Koordinaten von $\vec{c}$ bezüglich A umrechnet.
- c) Wie lauten die Matrizen $\mathbf{M}_{AB}$ und $\mathbf{M}_{BA}$ für die konkreten Basisvektoren $\vec{a}_1 = (1, 0, 0)^\top$, $\vec{a}_2 = (0, 1, 0)^\top$, $\vec{a}_3 = (0, 0, 1)^\top$ und $\vec{b}_1 = (1, 0, 0)^\top$, $\vec{b}_2 = (1, 0, 1)^\top$, $\vec{b}_3 = (0, 1, 1)^\top$? Berechnen Sie die Darstellung der Vektoren $\vec{c}_1 = (1, 2, 3)_A^\top$ und $\vec{c}_2 = (0, 1, 2)_A^\top$ als Linearkombination der Basis B .

Aufgabe 3.18

Zeigen Sie, dass $\vec{a} = (3, 4, 0)^\top$, $\vec{b} = (0, 0, 3)^\top$ und $\vec{c} = (0, 1, 0)^\top$ im $\mathbb{R}^3$ linear unabhängig sind. Berechnen Sie eine Orthonormalbasis (Gram-Schmidt'sches Verfahren).

Aufgabe 3.19

Bestimmen Sie den Rang der Matrizen und die Dimension des Nullraums:

$$\mathbf{A} = \begin{bmatrix} 4 & 9 & 5 \\ 1 & 5 & 3 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & 1 & 0 & 4 \\ 3 & 1 & 0 & 2 \\ 3 & 0 & 6 & 5 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

Aufgabe 3.20

Es sei $L: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ die lineare Abbildung mit

$$L(\vec{e}_1) := -2\vec{e}_2, \quad L(\vec{e}_2) := 3\vec{e}_1 + \vec{e}_3, \quad L(\vec{e}_3) := \vec{0}.$$

- a) Geben Sie die Abbildungsmatrix $\mathbf{A}$ von L an, und berechnen Sie $L((1, 2, 3)^\top)$.
- b) Bestimmen Sie $\text{Kern}(L)$ und $\text{Bild}(L)$, und verifizieren Sie den Dimensionssatz.

Aufgabe 3.21

Bestimmen Sie alle Eigenwerte und Eigenvektoren der Matrizen

$$\text{a) } \mathbf{A} = \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix}, \quad \text{b) } \mathbf{A} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix}.$$

Aufgabe 3.22

- a) Bestimmen Sie die Eigenwerte und jeweils einen zugehörigen Eigenvektor für die Matrix $\mathbf{A}$. Bilden die normierten Eigenvektoren ein Orthonormalsystem?

$$\mathbf{A} := \begin{bmatrix} -1 & -2 & 0 \\ -2 & 1 & -2 \\ 0 & -2 & -1 \end{bmatrix}.$$

- b) Zeigen Sie, dass 1 Eigenwert von $\mathbf{A} := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ mit $a + c = 1$ und $b + d = 1$ ist.



4 Funktionen mit mehreren Variablen

Übersicht

4.1	Grenzwerte und Stetigkeit	592
4.2	Ableitungen von reellwertigen Funktionen mit mehreren Variablen	597
4.3	Extremwertrechnung	617
4.4	Integralrechnung mit mehreren Variablen	649
4.5	Vektoranalysis	672
4.6	Aufgaben	701

Bislang haben wir Funktionen $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ betrachtet. Ein reeller Definitionsbereich und Wertebereich ist aber bei Vorgängen in der Wirklichkeit eher selten. Die Regel ist, dass Abhängigkeiten von vielen Parametern bestehen und auch viele Größen beeinflusst werden.

Bei einem Weg-Zeit-Diagramm erhält man die zurückgelegte Wegstrecke (Distanz) $s(t)$ als Funktion der Zeit. Kennt man den Startpunkt, so weiß man aber nicht, welche Richtungen gewählt wurden. Möchte man den genauen Aufenthaltsort nach t Sekunden kennen, so benötigt man eine Funktion, die einem Zeitpunkt t sowohl eine x - als auch eine y -Koordinate und vielleicht sogar eine z -Koordinate zuordnet, d. h., $\vec{s}(t)$ ist nun für jeden Zeitpunkt ein Vektor, $\vec{s}$ ist eine vektorwertige Funktion. Die Vektoren hängen hier aber weiterhin nur von der einen reellen Variable t ab. Möchte man aber beispielsweise ein magnetisches Feld darstellen, so ordnet man jedem Raumpunkt einen Vektor zu, der an dieser Stelle das Feld beschreibt. Zusätzlich hängt es von der Zeit ab, so dass eine Funktion von den vier Variablen x, y, z und t entsteht. Auch bei solchen Funktionen interessiert man sich für momentane Änderungen, so dass wir in diesem Kapitel die Differenzialrechnung entsprechend erweitern.

Während wir bislang bei der Integralrechnung nur die Größe von Flächen unter Funktionsgraphen bestimmt haben, entstehen bei Funktionen von zwei Variablen zwischen der x - y -Ebene und dem Funktionsgraphen Körper im dreidimensionalen Raum.

Deren Volumen kann man ebenfalls über Integrale ausrechnen. Auch das sehen wir uns in diesem Kapitel an.

Vielleicht wirkt die Verwendung vieler Variablen auf Sie zunächst etwas abschreckend. Die gute Nachricht ist aber, dass wir prinzipiell keine neuen Differenziations- und Integrationstechniken außer den bereits bekannten für Funktionen mit einer Variable benötigen. Betrachtet man alle Variablen bis auf eine als Konstanten, so können wir bereits die resultierende Funktion mit einer Variable ableiten und integrieren. Und mit dieser Reduktion auf den bekannten Fall rechnet man mit Funktionen mehrerer Variablen.

Beispiel 4.1

Wir beschreiben die Oberfläche einer Kugel mit Radius 1 über Funktionswerte. Die Oberfläche ist die Punktmenge

$$\{(x, y, z) : x^2 + y^2 + z^2 = 1\} = \{(x, y, z) : x^2 + y^2 \leq 1, z = \pm\sqrt{1 - x^2 - y^2}\}.$$

Zu jedem Punkt (x, y) in der Einheitskreisscheibe $D := \{(x, y) : x^2 + y^2 \leq 1\}$ erhalten wir so einen Punkt auf der oberen und einen auf der unteren Halbkugeloberfläche. Diese Zuordnung können wir über die beiden Funktionen $f_1 : D \rightarrow \mathbb{R}$, $f_1(x, y) := \sqrt{1 - x^2 - y^2}$ (siehe Abbildung 4.1) und $f_2 : D \rightarrow \mathbb{R}$, $f_2 := -f_1$, ausdrücken. ■

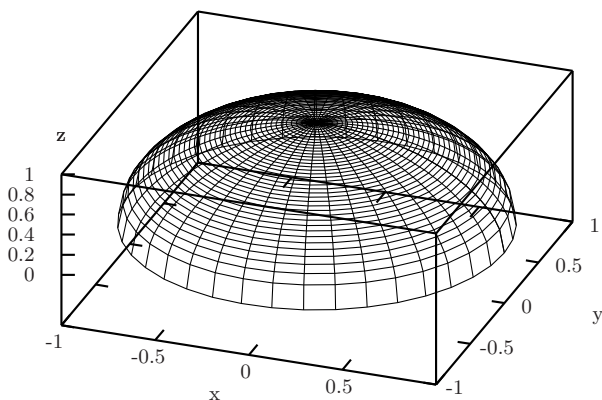


Abb. 4.1: $f_1(x, y) := \sqrt{1 - x^2 - y^2}$

Eigentlich müssten wir für das Argument $(x, y) \in D$ die Funktionswerte im Beispiel mit $f_1((x, y))$ bzw. $f_2((x, y))$ bezeichnen. Die zusätzlichen Vektorklammern lässt man aber üblicherweise weg.

Beispiel 4.2

Sei $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$. Unter einer **Höhenlinie** zur Höhe $c \in \mathbb{R}$ verstehen wir die Punktmenge $\{(x, y) \in D : f(x, y) = c\}$. Wir skizzieren die Höhenlinie der Funktion $f(x, y) := -x^2 + y^2 - 2y$. Dazu lösen wir die Gleichung $f(x, y) = c$ nach y auf:

$$c = f(x, y) = -x^2 + y^2 - 2y \iff y^2 - 2y - x^2 - c = 0, \quad y = 1 \pm \sqrt{1 + x^2 + c}.$$

Die Höhenlinie für c ist darstellbar über die Vereinigung der Funktionsgraphen zu $g_1(x) = 1 + \sqrt{1 + x^2 + c}$ und $g_2(x) = 1 - \sqrt{1 + x^2 + c}$ (siehe Abbildung 4.2). Falls $c < -1$ ist, sind diese definiert für $x^2 \geq -c - 1$, also für $|x| \geq \sqrt{-c - 1}$. Falls $c \geq -1$ ist, sind sie auf $\mathbb{R}$ definiert. ■

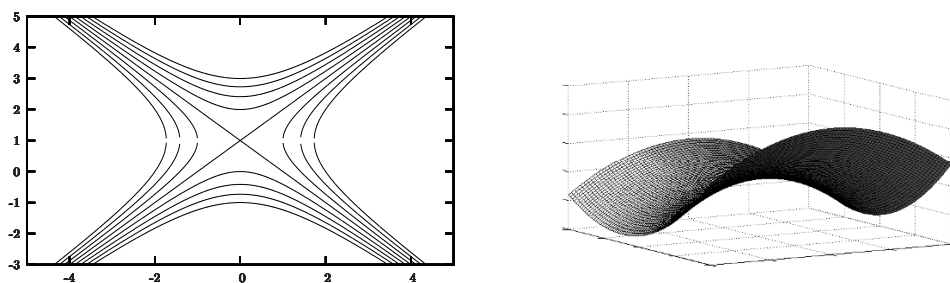


Abb. 4.2: Links: Höhenlinien zu $-x^2 + y^2 - 2y = c$ für $c = -4, -3, \dots, 4$; rechts: zugehöriger Funktionsgraph

Hintergrund: Curryng

Wir werden mit einer Funktion $f(x, y, z) : \mathbb{R}^3 \rightarrow \mathbb{R}$ 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 $f_3 := [[f_1(x)](y)] : \mathbb{R} \rightarrow \mathbb{R}$ 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 $f_2 := [f_1(x)] : y \rightarrow f_3$. Dabei ist $f_3 : \mathbb{R} \rightarrow \mathbb{R}$ 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 .

4.1 Grenzwerte und Stetigkeit

Wir übertragen in diesem Kapitel die Differenzialrechnung von Funktionen mit einer reellen Variable aus Kapitel 2 auf Funktionen $\vec{f} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\vec{f}(\vec{x}) = (f_1(\vec{x}), \dots, f_m(\vec{x}))$. Dabei schreiben wir statt f das Vektorsymbol $\vec{f}$, wenn $m > 1$ zugelassen ist, also die Funktion vektorwertig wird. $\vec{f}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \dots, f_m(\vec{x}))$ hat die reellwertigen Komponentenfunktionen $f_1, \dots, f_m$.

- Statt des Betrages einer reellen Zahl verwenden wir den Betrag (die Euklid'sche Norm) eines Vektors $\vec{x} \in \mathbb{R}^n$, den wir in bereits auf Seite 486 über das Standardskalarprodukt in $\mathbb{R}^n$ definiert haben:

$$|\vec{x}| := \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \sqrt{\sum_{k=1}^n x_k^2},$$

also z. B. $|(1, 2, 3)| = \sqrt{14}$. Damit lassen sich alle Aussagen übertragen, bei denen Grenzwerte verwendet werden. Der Betrag entspricht auch genau der l^2 -Norm aus Kapitel 3.7.1 (siehe insbesondere Beispiel 3.73 auf Seite 560), für die wir bereits (mit den folgenden Begriffen verträglich) Folgenkonvergenz und Stetigkeit definiert haben. Den Abschnitt über Funktionalanalysis setzen wir hier aber nicht voraus, und wir halten die Darstellung elementar.

- In der Regel genügt es, $m = 1$ (also reellwertige Funktionen) zu betrachten. Ist $m > 1$, so kann man die Komponentenfunktionen $f_1, \dots, f_m$ einzeln untersuchen.

Einen Grenzwertbegriff für Folgen von Vektoren erhalten wir über die Definition 2.5 des Folgen Grenzwerts für Folgen reeller (oder komplexer) Zahlen (siehe Seite 217):

Definition 4.1 (Konvergente Folge)

Eine Folge $(\vec{a}_k)_{k=1}^\infty$ mit Gliedern $\vec{a}_k = (a_{k,1}, \dots, a_{k,n}) \in \mathbb{R}^n$ heißt **konvergent** gegen einen Grenzvektor $\vec{a} \in \mathbb{R}^n$ genau dann, wenn die Folge $(|\vec{a}_k - \vec{a}|)_{k=1}^\infty$ der Abstände zu $\vec{a}$ eine reelle Nullfolge ist, d. h.

$$0 = \lim_{k \rightarrow \infty} |\vec{a}_k - \vec{a}| = \lim_{k \rightarrow \infty} \sqrt{\sum_{i=1}^n (a_{k,i} - a_i)^2}.$$

Schreibweise: $\lim_{k \rightarrow \infty} \vec{a}_k = \vec{a}$.

Man erkennt, dass die Konvergenz der Folge $(\vec{a}_k)_{k=1}^\infty$ gleichbedeutend ist mit der Konvergenz aller Komponentenfolgen $(a_{k,i})_{k=1}^\infty$:

$$\begin{aligned} 0 = \lim_{k \rightarrow \infty} \sqrt{\sum_{i=1}^n (a_{k,i} - a_i)^2} &\iff 0 = \lim_{k \rightarrow \infty} \sum_{i=1}^n (a_{k,i} - a_i)^2 \\ \iff 0 = \lim_{k \rightarrow \infty} (a_{k,i} - a_i)^2, 1 \leq i \leq n &\iff \lim_{k \rightarrow \infty} a_{k,i} = a_i, 1 \leq i \leq n. \end{aligned} \quad (4.1)$$

Das erste Äquivalenzzeichen der zweiten Zeile erklärt sich so: Von rechts nach links wird eine endliche Summe konvergenter Folgen gebildet, die gegen die Summe der Einzelgrenzwerte (hier 0) konvergiert. Von links nach rechts wird benutzt, dass $(a_{k,i} - a_i)^2 \geq 0$ ist. Konvergiert die Summe gegen null, so geht das daher nur, wenn die einzelnen, nicht-negativen Summanden gegen null streben.

Beispiel 4.3

Die Folge $\left((2 + \frac{1}{k}, 2 - \frac{2}{k})\right)_{k=1}^\infty$ konvergiert gegen $(2, 2)$. ■

Bei Grenzwerten war es wichtig, dass die Funktionen auch in der Umgebung der betrachteten Stelle erklärt sind. Daher haben wir in Definition 2.11 auf Seite 256 den Begriff eines inneren Punktes einer Menge $E \subset \mathbb{R}$ geprägt. Einem inneren Punkt kann man sich in der Menge sowohl von rechts als auch von links nähern. Das verallgemeinern wir auf beliebige Richtungen:

Definition 4.2 (Offene Menge, innerer Punkt, Randpunkt)

- Sei $E \subset \mathbb{R}^n$. Ein Punkt $\vec{x} \in E$ heißt **innerer Punkt** von E genau dann, falls ein $\delta = \delta(\vec{x}) > 0$ existiert, so dass die δ -Umgebung $\{\vec{y} \in \mathbb{R}^n : |\vec{x} - \vec{y}| < \delta\}$ vollständig in E enthalten ist (siehe Abbildung 4.3).
- Eine Menge $E \subset \mathbb{R}^n$ heißt **offen** genau dann, falls jeder Punkt von E ein innerer Punkt ist.
- Ein Punkt $\vec{x} \in \mathbb{R}^n$ heißt **Randpunkt** einer Menge $E \subset \mathbb{R}^n$ genau dann, wenn in jeder δ -Umgebung von $\vec{x}$ sowohl ein Punkt $\vec{y} \in E$ als auch ein Punkt $\vec{y} \notin E$ liegt (siehe Abbildung 4.3). Der **Rand** von E ist die Menge der Randpunkte von E .

Man beachte, dass dies eine Erweiterung der Definition 2.11 ist: Für $n = 1$ ist z. B. $]a, b[\subset \mathbb{R}$ eine offene Menge.

Die mathematische Definition des Randes einer Menge entspricht genau der Anschauung.

Definition 4.3 (Grenzwert einer Funktion)

Sei $f : D \rightarrow \mathbb{R}$ mit $D \subset \mathbb{R}^n$ und $\vec{x}_0 \in \mathbb{R}^n$, so dass $\vec{x}_0$ innerer Punkt der Menge $D \cup \{\vec{x}_0\}$ ist (d. h., $\vec{x}_0$ muss nicht in D sein, aber f muss in einer vollständigen Umgebung von $\vec{x}_0$ mit Ausnahme der Stelle $\vec{x}_0$ erklärt sein).

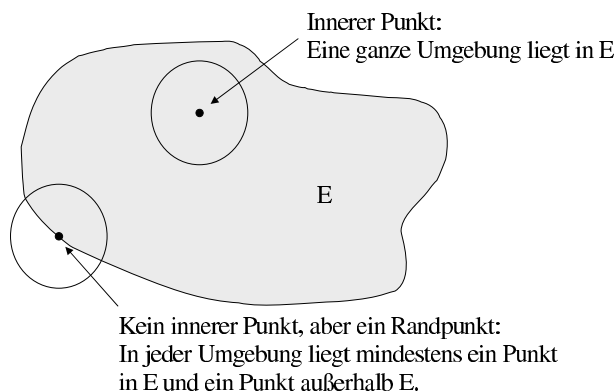


Abb. 4.3: Innerer Punkt und Randpunkt

Die Funktion f heißt **konvergent** gegen $L \in \mathbb{R}$ für $\vec{x} \rightarrow \vec{x}_0$ genau dann, wenn zu jedem noch so kleinen $\varepsilon > 0$ ein $\delta = \delta(\varepsilon, \vec{x}_0)$ existiert, so dass für alle von $\vec{x}_0$ verschiedenen Punkte $\vec{x} \in D$, die nicht weiter als δ von $\vec{x}_0$ entfernt sind, gilt: Die Funktionswerte sind nicht weiter als ε von L entfernt. Also:

$$0 < |\vec{x} - \vec{x}_0| < \delta \implies |f(\vec{x}) - L| < \varepsilon.$$

Schreibweise: $\lim_{\vec{x} \rightarrow \vec{x}_0} f(\vec{x}) = L$.

Für $n = 1$ ist dies genau die Grenzwertdefinition 2.14 für $x \rightarrow x_0$. Für $n > 1$ unterscheidet sich die Definition von einem iterierten Grenzwert, also z. B. für $n = 2$ von $\lim_{x \rightarrow x_0} [\lim_{y \rightarrow y_0} f(x, y)] = L$. Aus einem solchen hintereinander gesetzten Grenzwert folgt nicht immer $\lim_{\vec{x} \rightarrow \vec{x}_0} f(\vec{x}) = L$. Das sehen wir am nächsten Beispiel (vgl. (4.2)).

Wie im Eindimensionalen gilt das Übertragungsprinzip, mit dem man Grenzwerte von Funktionen zurückführen kann auf Grenzwerte von Folgen (vgl. Satz 2.22 auf Seite 266):

Satz 4.1 (Übertragungsprinzip)

Sei $f : D \rightarrow \mathbb{R}$ und $\vec{x}_0 \in D \subset \mathbb{R}^n$ ein innerer Punkt. Äquivalent sind:

- a) $\lim_{\vec{x} \rightarrow \vec{x}_0} f(\vec{x}) = L$.
- b) Für alle Folgen $(\vec{x}_k)_{k=1}^{\infty}$, die gegen $\vec{x}_0$ konvergieren (genauer: $\vec{x}_k \in D \setminus \{\vec{x}_0\}$ und $\lim_{k \rightarrow \infty} \vec{x}_k = \vec{x}_0$), gilt: $\lim_{k \rightarrow \infty} f(\vec{x}_k) = L$.

Man hat also genau dann an der Stelle $\vec{x}_0$ Konvergenz gegen L , wenn man für jede Folge, die gegen $\vec{x}_0$ strebt, Konvergenz der zugehörigen Funktionswerte gegen L hat. Egal, wie man sich der Stelle $\vec{x}_0$ nähert, die zugehörigen Funktionswerte müssen gegen L streben. Es ist schwierig, mit dem Übertragungsprinzip Konvergenz zu zeigen.

Dagegen ist es einfach, die Konvergenz zu widerlegen. Man braucht nur zwei gegen $\vec{x}_0$ konvergente Folgen, für die die zugehörigen Funktionswerte nicht gegen die gleiche Zahl streben.

Beispiel 4.4

Die Funktion

$$f(x, y) := \begin{cases} \frac{xy}{x^2+y^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0) \end{cases}$$

besitzt in $(0, 0)$ keinen Grenzwert: Wir betrachten die Nullfolge $(\vec{x}_k)_{k=1}^\infty = (\frac{1}{k}(1, 1))_{k=1}^\infty$. Damit ist $f(\vec{x}_k) = \frac{1/k^2}{2/k^2} = \frac{1}{2}$ und $\lim_{k \rightarrow \infty} f(\vec{x}_k) = \frac{1}{2}$. Wählen wir aber als Nullfolge $(\vec{y}_k)_{k=1}^\infty = (\frac{1}{k}(0, 1))_{k=1}^\infty$, so ist $f(\vec{y}_k) = 0$ und $\lim_{k \rightarrow \infty} f(\vec{y}_k) = 0 = f(0, 0) \neq \frac{1}{2}$. f besitzt keinen Grenzwert im Punkt $(0, 0)$.

Insbesondere sehen wir auch, dass iterierte Grenzwerte etwas anderes als ein Grenzwert in $\mathbb{R}^2$ sein können: Für festes x ist

$$\lim_{y \rightarrow 0} f(x, y) = \begin{cases} \lim_{y \rightarrow 0} \frac{xy}{x^2+y^2}, & x \neq 0 \\ 0, & x = 0 \end{cases} = 0. \quad (4.2)$$

Damit ist $\lim_{x \rightarrow 0} \left[\lim_{y \rightarrow 0} \frac{xy}{x^2+y^2} \right] = 0$, obwohl der Grenzwert in $(0, 0)$ nicht existiert. Würde aber der Grenzwert existieren, dann würden die iterierten Grenzwerte wegen des Übertragungsprinzips ebenfalls zu diesem Wert führen. Dann darf also mit einer einzelnen Variable gerechnet werden, indem die andere als konstant angesehen wird. Das werden wir beim Ableiten und bei der Integration ausnutzen. ■

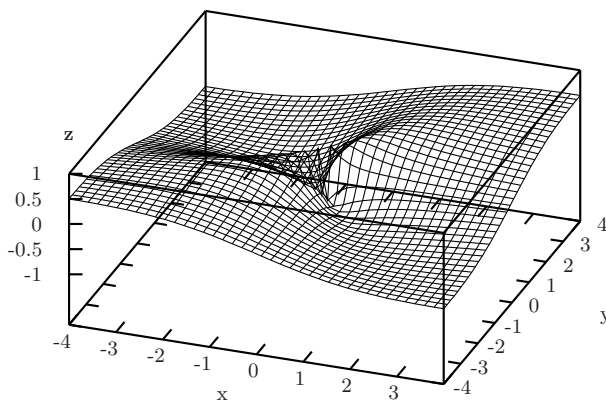


Abb. 4.4: $\frac{xy}{x^2+y^2}$

Wir definieren die Stetigkeit jetzt wie in der Charakterisierung auf Seite 272:

Definition 4.4 (Stetigkeit)

Sei $f : D \rightarrow \mathbb{R}$ mit $D \subset \mathbb{R}^n$ und $\vec{x}_0 \in D$.

- Die Funktion f heißt **stetig** in $\vec{x}_0$ genau dann, wenn zu jedem noch so kleinen $\varepsilon > 0$ ein $\delta = \delta(\vec{x}_0, \varepsilon) > 0$ existiert, so dass für alle $\vec{x} \in D$, die nicht weiter als δ von $\vec{x}$ entfernt sind (d. h. $|\vec{x} - \vec{x}_0| < \delta$), gilt: $|f(\vec{x}) - f(\vec{x}_0)| < \varepsilon$.
- f heißt genau dann stetig auf D , wenn f in jedem Punkt $\vec{x} \in D$ stetig ist.
- f heißt stetig ergänzbar in einem Punkt $\vec{x}_0 \in D$ mit Wert L genau dann, wenn die Funktion g stetig in $\vec{x}_0$ ist, wobei

$$g(\vec{x}) := \begin{cases} f(\vec{x}), & \vec{x} \neq \vec{x}_0, \\ L, & \vec{x} = \vec{x}_0. \end{cases}$$

In der Definition haben wir nicht gefordert, dass $\vec{x}_0$ ein innerer Punkt von D ist. Damit haben wir direkt auch die Stetigkeit an den Rändern eines Definitionsbereichs erklärt. In $\mathbb{R}^n$ (für $n > 1$) reicht es im Gegensatz zu $\mathbb{R}$ nicht, von rechts- und linksseitigen Grenzwerten an den Rändern eines Definitionsbereichs zu sprechen, da man hier auch aus beliebigen anderen Richtungen kommen kann.

Ist aber zusätzlich $\vec{x}_0$ ein innerer Punkt, so können wir die Definition der Stetigkeit mit dem Grenzwertbegriff umschreiben (siehe Definition 2.17 auf Seite 270):

Folgerung 4.1 (Stetigkeit)

Ist $\vec{x}_0 \in D \subset \mathbb{R}^n$ ein innerer Punkt, so ist $f : D \rightarrow \mathbb{R}$ stetig in $\vec{x}_0$ genau dann, wenn

$$\lim_{\vec{x} \rightarrow \vec{x}_0} f(\vec{x}) = f(\vec{x}_0).$$

Damit ist die Funktion f aus dem vorangehenden Beispiel 4.4 nicht stetig in $(0, 0)$.

Wie bei Funktionen einer reellen Variable sind die Summe, das Produkt und die Verkettung stetiger Funktionen stetig.

Die vorangehend definierten Begriffe übertragen sich auf vektorwertige Funktionen $\vec{f} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$, indem man sie für jede Komponentenfunktion $f_k : D \rightarrow \mathbb{R}$ verlangt.

4.2 Ableitungen von reellwertigen Funktionen mit mehreren Variablen

4.2.1 Ableitungsbegriffe

Zunächst halten wir alle Variablen bis auf eine fest und rechnen mit der verbliebenen so, als wäre es die Einzige. Zum Beispiel können wir $f(x, y, z) := x^3 + y^2 z$ bei festem y und z (Konstanten) nach x differenzieren und erhalten $3x^2$. Dies ist eine partielle Ableitung nach x .

Definition 4.5 (Partielle Ableitung, Gradient)

Sei $f = f(x_1, \dots, x_n) : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ und $\vec{x}_0 = (x_{0,1}, x_{0,2}, \dots, x_{0,n})$ ein innerer Punkt von D .

- Die **partielle Ableitung** erster Ordnung von f bezüglich der Variable x_k (für ein $k \in \{1, 2, \dots, n\}$) im Punkt $\vec{x}_0$ ist definiert als Grenzwert

$$\frac{\partial f}{\partial x_k}(\vec{x}_0) = f_{x_k}(\vec{x}_0) := \lim_{h \rightarrow 0} \frac{f(x_{0,1}, \dots, x_{0,k-1}, x_{0,k} + h, x_{0,k+1}, \dots, x_{0,n}) - f(x_{0,1}, \dots, x_{0,k}, \dots, x_{0,n})}{h},$$

sofern dieser existiert.

- Existieren in $\vec{x}_0$ alle partiellen Ableitungen, so kann man diese in einem Vektor, dem **Gradienten**, zusammenfassen:

$$\text{grad } f(\vec{x}_0) := \nabla f(\vec{x}_0) := \left(\frac{\partial f}{\partial x_1}(\vec{x}_0), \dots, \frac{\partial f}{\partial x_n}(\vec{x}_0) \right).$$

Man nennt ∇ den **Nabla-Operator**.

Der Nabla-Operator ist nur eine andere, kürzere Schreibweise für grad . Man beachte, dass dies bei festen Werten für die von der partiellen Ableitung nicht betroffenen Variablen x_l , $l \neq k$, der bereits bekannte Ableitungsbegriff ist: $\frac{\partial f}{\partial x_k}(\vec{x}_0) = \frac{df}{dx_k}(\vec{x}_0)$. Man betrachtet also alle Variablen, nach denen man nicht differenziert, als Konstanten. Dadurch erhält man eine Funktion mit einer Variable, für die man die Steigung berechnen kann. Bei Funktionen mit mehreren Variablen schreibt man als Konvention die Ableitungen mit dem Symbol ∂ , bei einer Variable mit d . Diese Unterscheidung vereinfacht das Lesen, wäre aber mathematisch gar nicht nötig.

In Abbildung 4.5 ist eine Funktion $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dargestellt, indem die Funktionswerte $z = f(x, y)$ als z -Koordinate aufgetragen sind. Hält man $y = y_0$ als Konstante fest, so erhält man den von links oben nach rechts unten eingezeichneten Funktionsgraphen. An

der Stelle x_0 hat dieser die Steigung der entsprechend eingezeichneten Gerade, die der partiellen Ableitung $\frac{\partial f}{\partial x}(x_0, y_0)$ entspricht. Entsprechend ist $\frac{\partial f}{\partial y}(x_0, y_0)$ die Steigung der zweiten Gerade, die die Tangente an den von links unten nach rechts oben verlaufenden Funktionsgraphen zu $f(x_0, y)$ als Funktion von y ist. Partielle Ableitungen geben also die Steigung in Richtung der Koordinatenachsen an.

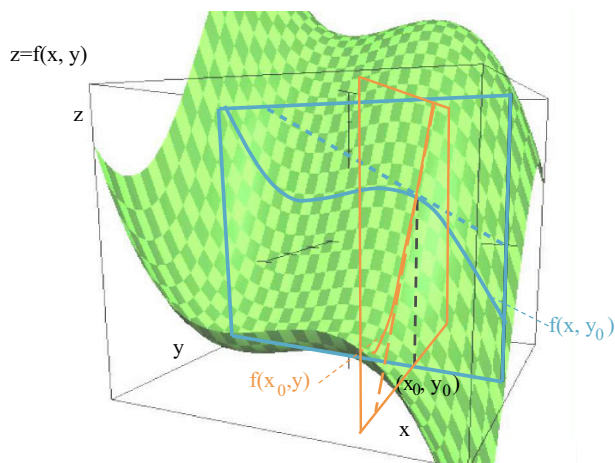


Abb. 4.5: Die partielle Ableitung nach x in (x_0, y_0) ist die Steigung der von links oben nach rechts unten verlaufenden blau gestrichelten, die partielle Ableitung nach y die Steigung der von links unten nach rechts oben verlaufenden orange gestrichelten Gerade. Das Totale Differenzial beschreibt die Ebene, die durch die beiden Geraden aufgespannt wird.

Beispiel 4.5

a) $\text{grad}(x^2 + x \sin y) = \nabla(x^2 + x \sin y) = \left(\frac{\partial}{\partial x}(x^2 + x \sin y), \frac{\partial}{\partial y}(x^2 + x \sin y) \right)$
 $= (2x + \sin y, x \cos y).$

b) Sei $f(x, y, z) := \frac{x-y^2}{yz+1}$ für $yz \neq -1$:

$$\begin{aligned} \frac{\partial f}{\partial x}(x, y, z) &= \frac{1}{yz+1}, & \frac{\partial f}{\partial y}(x, y, z) &= \frac{-2y(yz+1) - (x-y^2)z}{(yz+1)^2}, \\ \frac{\partial f}{\partial z}(x, y, z) &= \frac{-(x-y^2)y}{(yz+1)^2}, & (\text{grad } f)(0, 0, 0) &= (1, 0, 0). \end{aligned}$$

■

Da die partiellen Ableitungen als Ableitungen einer Funktion mit einer Variable aufgefasst werden können, erhält man sofort Ableitungsregeln:

Lemma 4.1 (Ableitungsregeln für den Gradienten)

Seien $D \subset \mathbb{R}^n$ offen und $f, g : D \rightarrow \mathbb{R}$ partiell differenzierbar nach allen Variablen in jedem $\vec{x} \in D$. Dann gelten die folgenden Rechenregeln auf D :

a) $\text{grad}(f + g) = \left(\frac{\partial(f+g)}{\partial x_1}, \dots, \frac{\partial(f+g)}{\partial x_n} \right) = \text{grad } f + \text{grad } g,$

b) $\text{grad}(f \cdot g) = \left(\frac{\partial(f \cdot g)}{\partial x_1}, \dots, \frac{\partial(f \cdot g)}{\partial x_n} \right) = \left(f \cdot \frac{\partial g}{\partial x_1} + g \cdot \frac{\partial f}{\partial x_1}, \dots, f \cdot \frac{\partial g}{\partial x_n} + g \cdot \frac{\partial f}{\partial x_n} \right)$
 $= f \cdot \text{grad } g + g \cdot \text{grad } f.$

Dabei ist $f \cdot \text{grad } g$ die Multiplikation der reellen Zahl $f(\vec{x})$ (als Skalar) mit dem Vektor $(\text{grad } g)(\vec{x})$.

Beispiel 4.6

Wir betrachten wieder die Funktion

$$f(x, y) := \begin{cases} \frac{xy}{x^2+y^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0), \end{cases}$$

von der wir bereits wissen, dass sie in $(0, 0)$ nicht stetig ist. Für $y = 0$ und $x \in \mathbb{R}$ ist $f(x, 0) = 0$, $\frac{\partial f}{\partial x}(x, 0) = 0$, entsprechend ist $\frac{\partial f}{\partial y}(0, y) = 0$. Beide partiellen Ableitungen existieren also in $(0, 0)$, obwohl f hier nicht stetig ist! ■

Leider gibt es also Funktionen $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$, bei denen alle partiellen Ableitungen an einer Stelle existieren, die aber dort nicht stetig sind. Dies liegt daran, dass wir hier nur entlang der Koordinatenachsen differenzieren, aber nicht sehen, was in anderen Richtungen geschieht. Wir brauchen einen stärkeren Ableitungsbegriff, aus dem wie in einer Dimension die Stetigkeit folgt.

Eine Funktion f mit nur einer Variable x ist differenzierbar an einer Stelle x_0 mit Ableitung $f'(x_0)$ genau dann, wenn

$$\begin{aligned} 0 &= \lim_{h \rightarrow 0} \left| \frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0) \right| = \lim_{h \rightarrow 0} \underbrace{\left| \frac{h}{|h|} \right|}_{=1} \left| \frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0) \right| \\ &= \lim_{h \rightarrow 0} \left| \frac{f(x_0 + h) - f(x_0)}{|h|} - f'(x_0) \frac{h}{|h|} \right|. \end{aligned}$$

Die letzte Formulierung des Grenzwerts lässt sich nun auch für Vektoren $\vec{h}$ hinschreiben, da nur durch die Länge der Vektoren und nicht durch die Vektoren selbst dividiert wird. Das wäre nicht definiert.

Definition 4.6 (Totales Differenzial)

Sei $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ und $\vec{x}_0$ ein innerer Punkt von D .

- Die Funktion f heißt genau dann **(total) differenzierbar** in $\vec{x}_0$, wenn dort die partiellen Ableitungen existieren und zusätzlich gilt:

$$\lim_{\vec{h} \rightarrow \vec{0}} \left[\frac{f(\vec{x}_0 + \vec{h}) - f(\vec{x}_0)}{|\vec{h}|} - \sum_{k=1}^n \frac{\partial f}{\partial x_k}(\vec{x}_0) \frac{h_k}{|\vec{h}|} \right] = 0. \quad (4.3)$$

■ Für festes $\vec{x}_0$ heißt dann die Funktion $Df : \mathbb{R}^n \rightarrow \mathbb{R}$, definiert als

$$Df(h_1, \dots, h_n) := \sum_{k=1}^n \frac{\partial f}{\partial x_k}(\vec{x}_0) h_k = (\text{grad } f(\vec{x}_0)) \cdot \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix},$$

das **Totale Differenzial** von f in $\vec{x}_0$.

Mit den Begriffen Totales Differenzial und Gradient können wir (4.3) umschreiben:

$$\begin{aligned} \lim_{\vec{h} \rightarrow \vec{0}} \left[\frac{f(\vec{x}_0 + \vec{h}) - f(\vec{x}_0)}{|\vec{h}|} - (\text{grad } f(\vec{x}_0)) \cdot \frac{1}{|\vec{h}|} \vec{h} \right] \\ = \lim_{\vec{h} \rightarrow \vec{0}} \left[\frac{f(\vec{x}_0 + \vec{h}) - f(\vec{x}_0)}{|\vec{h}|} - Df \left(\frac{1}{|\vec{h}|} \vec{h} \right) \right] = 0. \end{aligned}$$

Mit dieser Darstellung kann man die Stetigkeit von f in $\vec{x}_0$ zeigen, indem man die eckige Klammer mit $|\vec{h}|$ multipliziert (vgl. Satz 2.31 auf Seite 294).

In Definition 2.25 auf Seite 308 haben wir das Differenzial einer Funktion mit einer Variable definiert. Dabei handelt es sich um das Totale Differenzial im Spezialfall $n = 1$, wobei df eine lineare Funktion $df : \mathbb{R} \rightarrow \mathbb{R}$ mit $df(\Delta x) = f'(x_0)\Delta x$ ist, die eine Gerade mit Steigung $f'(x_0)$ durch den Koordinatenursprung beschreibt. In der Nähe von x_0 ist

$$f(x_0 + \Delta x) \approx f(x_0) + f'(x_0)\Delta x = f(x_0) + df(\Delta x).$$

Nun ist $Df : \mathbb{R}^n \rightarrow \mathbb{R}$ und beschreibt eine Hyperebene des $\mathbb{R}^{n+1}$ durch den Koordinatenursprung. Bei einer Funktion $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ist diese Hyperebene tatsächlich eine Ebene, und der Begriff stimmt mit der Umgangssprache überein. Bei einer Funktion $f : \mathbb{R}^n \rightarrow \mathbb{R}$ handelt es sich bei $\{(h_1, \dots, h_n, y) : y = Df(h_1, \dots, h_n)\}$ allgemeiner um einen Unterraum des $\mathbb{R}^{n+1}$ der Dimension n . In Richtung der Koordinatenachsen hat die Ebene die Steigungen

$$\frac{\partial}{\partial h_k} Df(h_1, \dots, h_n) = \frac{\partial f}{\partial x_k}(\vec{x}_0),$$

die die partiellen Ableitungen von f in $\vec{x}_0$ sind. In Richtung der Koordinatenachsen stimmen damit die Steigungen der Hyperebene mit denen von f in $\vec{x}_0$ überein, die Hyperebene liegt tangential am Funktionsgraphen von f an der Stelle $\vec{x}_0$ an, wenn man sie so verschiebt, dass sie durch den Punkt $(\vec{x}_0, f(\vec{x}_0))$ verläuft. Es handelt sich dann um eine Tangentialebene:

Definition 4.7 (Tangentialebene)

Existieren in $\vec{x}_0 = (x_{0,1}, x_{0,2}, \dots, x_{0,n})$ partielle Ableitungen einer Funktion $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ nach allen Variablen x_i , $i = 1, 2, \dots, n$, so wird die Ebene $E :=$

$$\left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ y \end{pmatrix} : y = f(\vec{x}_0) + \frac{\partial f}{\partial x_1}(\vec{x}_0)(x_1 - x_{0,1}) + \dots + \frac{\partial f}{\partial x_n}(\vec{x}_0)(x_n - x_{0,n}) \right\}$$

als **Tangentialebene** an die Funktion f in $\vec{x}_0$ bezeichnet.

Mit der Tangentialebene verbinden wir die Vorstellung, dass wir f in der Nähe von $\vec{x}_0$ durch diese annähern können:

$$\begin{aligned} f(\vec{x}_0 + (\Delta x_1, \Delta x_2, \dots, \Delta x_n)) &\approx f(\vec{x}_0) + \text{grad } f(\vec{x}_0) \cdot \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \vdots \\ \Delta x_n \end{pmatrix} \\ &= f(\vec{x}_0) + Df(\Delta x_1, \Delta x_2, \dots, \Delta x_n). \end{aligned}$$

Allerdings gelingt dies nur, falls f in $\vec{x}_0$ total differenzierbar ist und man mit der Tangentialebene das Verhalten von f nicht nur in Richtung der Koordinatenachsen wiedergibt: Wählt man für $\vec{h}$ in (4.3) speziell $h\vec{e}$, wobei $\vec{e} \in \mathbb{R}^n$ ein Richtungsvektor mit normierter Länge $|\vec{e}| = 1$ (Einheitsvektor) sei, so existiert die **Richtungsableitung erster Ordnung** im Punkt $\vec{x}_0$ in Richtung $\vec{e}$, definiert als

$$\frac{\partial}{\partial \vec{e}} f(\vec{x}_0) := \lim_{h \rightarrow 0} \frac{f(\vec{x}_0 + h\vec{e}) - f(\vec{x}_0)}{h}.$$

Denn aus (4.3) erhalten wir:

$$\begin{aligned} 0 &= \lim_{h \rightarrow 0} \frac{f(\vec{x}_0 + h\vec{e}) - f(\vec{x}_0) - h \text{grad } f(\vec{x}_0) \cdot \vec{e}}{|h\vec{e}|} \\ &= \lim_{h \rightarrow 0} \left| \frac{f(\vec{x}_0 + h\vec{e}) - f(\vec{x}_0)}{h} - \text{grad } f(\vec{x}_0) \cdot \vec{e} \right|, \end{aligned}$$

$$\frac{\partial}{\partial \vec{e}} f(\vec{x}_0) = \sum_{k=1}^n \frac{\partial f}{\partial x_k}(\vec{x}_0) e_k = \text{grad } f(\vec{x}_0) \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = Df(e_1, \dots, e_n).$$

Das Totale Differenzial beschreibt damit eine Ebene, die zum Funktionsgraphen tangential bezüglich jeder vorgegebenen Richtung $\vec{e}$ ist. Alle Geraden durch $\vec{x}_0$, die in der Tangentialebene liegen, sind Tangenten an den Funktionsgraphen an dieser Stelle. Ihre Steigung entspricht genau der zugehörigen Richtungsableitung.

Beispiel 4.7

Die Tangentialebene an die Funktion

$$z = f(x, y) = x^2 + 2xy + 4$$

im Punkt $\vec{x}_0 = (1, -2) \in D$ ist durch die Gleichung

$$z = f(1, -2) + \left[\frac{\partial f}{\partial x}(1, -2) \right] (x - 1) + \left[\frac{\partial f}{\partial y}(1, -2) \right] (y + 2)$$

beschrieben, also mit $\frac{\partial f}{\partial x}(x, y) = 2x + 2y$ und $\frac{\partial f}{\partial y}(x, y) = 2x$ ist

$$z = 1 - 2(x - 1) + 2(y + 2). \quad \blacksquare$$

Beispiel 4.8 (Charakteristiken einer partiellen Differenzialgleichung)

Häufig sucht man Funktionen, über deren Ableitungsverhalten man aufgrund von Naturgesetzen etwas weiß. Wie sehen beispielsweise die (total differenzierbaren) Funktionen $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 (siehe Kapitel 3.2.1.1). 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 $g : \mathbb{R} \rightarrow \mathbb{R}$, 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. ■

In welcher Richtung $\vec{e}$ hat eine Funktion $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ an einer Stelle $\vec{x}_0$ die größte Steigung $\frac{\partial}{\partial \vec{e}} f(\vec{x}_0)$? Ist α der Winkel zwischen den Vektoren $\text{grad } f(\vec{x}_0)$ und $\vec{e}$, so gilt unter Verwendung des Skalarprodukts (siehe Kapitel 3.1.2):

$$\frac{\partial}{\partial \vec{e}} f(\vec{x}_0) = \text{grad } f(\vec{x}_0) \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = |\text{grad } f(\vec{x}_0)| \cdot |\vec{e}| \cdot \cos \alpha = |\text{grad } f(\vec{x}_0)| \cdot \cos \alpha.$$

Die größte Richtungsableitung ergibt sich für $\alpha = 0$ bzw. $\cos \alpha = 1$. Die zugehörige Richtung ist $\vec{e} = |\text{grad } f(\vec{x}_0)|^{-1} \text{grad } f(\vec{x}_0)$. Das ist für viele praktische Anwendungen eine sehr wichtige Beobachtung, die wir durch einen Satz würdigen:

Satz 4.2 (Gradient als Richtung des größten Anstiegs)

Sei f total differenzierbar in $\vec{x}_0$. Der Gradient $\text{grad } f(\vec{x}_0) \neq \vec{0}$ zeigt in die Richtung des größten (steilsten) Anstiegs von f . In entgegengesetzter Richtung $-\text{grad } f(\vec{x}_0)$ hat f den steilsten Abstieg.

Beispiel 4.9 (Gradientenverfahren)

Ein Algorithmus, mit dem man ein (lokales) Minimum einer Funktion $y = f(x)$ iterativ bestimmen kann, ist das **Verfahren des steilsten Abstiegs** (Gradientenverfahren, steepest descent), vgl. (Dobner und Engelmann, 2002, Band 2, S. 102–103). Dabei sucht man ausgehend von einem Startpunkt $\vec{x}_0$ einen nächsten Punkt $\vec{x}_1$ in Richtung des Vektors des steilsten Abstiegs $\vec{v} = -\text{grad } f(\vec{x}_0) (\neq \vec{0})$. Dazu minimiert man die Funktion

$$g(t) := f(\vec{x}_0 + t \cdot \vec{v}) \quad \text{mit} \quad \vec{v} = -\text{grad } f(\vec{x}_0),$$

indem man einen Wert $t_0 > 0$ über die notwendige Bedingung $g'(t_0) = 0$ bestimmt. Nun kann man den Algorithmus mit dem neuen Startpunkt $\vec{x}_1 = \vec{x}_0 + t_0 \cdot \vec{v}$ wiederholen. Nach einiger Zeit wird $\text{grad } f(\vec{x}_k) \approx \vec{0}$ (vgl. Satz 4.7 auf Seite 618) oder man erzielt keine weiteren signifikanten Verbesserungen. Damit hat man annähernd ein lokales Minimum gefunden. Entsprechend kann ein (lokales) Maximum gesucht werden, indem man Schritte in Richtung des jeweils steilsten Anstiegs verwendet. ■

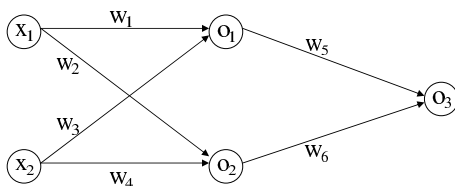


Abb. 4.6: Neuronales Netz

Beispiel 4.10 (Mit dem Gradientenverfahren lernendes Neuronales Netz)

Neuronale Netze bilden das Verhalten von vernetzten Nervenzellen nach. Sie berechnen dazu aus Eingabewerten einen oder mehrere Ausgabewerte. In Abbildung 4.6 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 $\lim_{x \rightarrow -\infty} g(x) = 0$ und $\lim_{x \rightarrow \infty} g(x) = 1$ 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 $o_3 = o_3(w_1, \dots, w_6)$ 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

$$\begin{aligned} o_3(w_1, \dots, w_6) &= g(w_5 o_1(w_1, w_3) + w_6 o_2(w_2, w_4)) \\ &= g(w_5 g(w_1 x_1 + w_3 x_2) + w_6 g(w_2 x_1 + w_4 x_2)). \end{aligned}$$

Wir berechnen die partiellen Ableitungen

$$\frac{\partial o_3}{\partial w_5}(w_1, \dots, w_6) = g'(w_5 o_1 + w_6 o_2) o_1 = g(w_5 o_1 + w_6 o_2) [1 - g(w_5 o_1 + w_6 o_2)] o_1$$

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

und kennen damit $\text{grad } o_3(w_1, \dots, w_6)$. 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 $f(w_1, \dots, w_6) := (o - o_3(w_1, \dots, w_6))^2$ minimieren. Dazu kann man von einer Startbelegung der Gewichte ausgehend sukzessive Schritte in Richtung des negativen Gradienten $-\text{grad } f(w_1, \dots, w_6) = -2(o - o_3(w_1, \dots, w_6))[-\text{grad } o_3(w_1, \dots, w_6)]$ machen. Bei jedem Schritt addiert man zum aktuellen Vektor der Gewichte den Wert $\lambda 2(o - o_3) \text{grad } o_3$, wobei λ die Schrittweite bezeichnet, die auch Lernrate genannt wird. Wir minimieren den Fehler also mit einem Gradientenverfahren, wobei im Gegensatz zu Beispiel 4.9 mit einer Schrittweite gearbeitet wird. 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 4.11 (Eine andere Sicht auf das Newton-Verfahren)

Wir können das eindimensionale Newton-Verfahren von Seite 306, mit dem wir näherungsweise Nullstellen einer differenzierbaren Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ bestimmt haben, als modifiziertes Gradientenverfahren auffassen. Sofern Nullstellen vorliegen, ist die Nullstellenbestimmung 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 zu gehen ist, wird allerdings einfacher als in Beispiel 4.9 berechnet, indem die Funktion durch die Tangente angenähert wird. ■

Existiert das Totale Differenzial, so hat man über den Gradienten sofort alle Richtungsableitungen. Der über einen Grenzwert definierte Begriff des Totalen Differenzials

ist allerdings für die Anwendung sperrig. Glücklicherweise gibt es aber den folgenden Satz, den wir ohne Beweis zitieren:

Satz 4.3 (Hinreichende Bedingung: Existenz des Totalen Differenzials)

Sei $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$, D offen, so dass alle partiellen Ableitungen erster Ordnung stetig sind auf D , dann existiert das Totale Differenzial Df für jeden Punkt aus D .

In der Praxis kommt man daher ganz gut mit dem Begriff des Gradienten aus, ohne den Begriff des Totalen Differenzials bemühen zu müssen. Man muss lediglich sicherstellen, dass die partiellen Ableitungen stetig sind.

Wir haben bislang das Totale Differenzial nur für reellwertige Funktionen betrachtet.

Definition 4.8 (Jacobi-Matrix)

*Vektorwertige Funktionen $\vec{f} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ heißen genau dann **total differenzierbar**, wenn alle Komponentenfunktionen $f_1, f_2, \dots, f_m$ total differenzierbar sind. Die Matrix*

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{x}) & \frac{\partial f_1}{\partial x_2}(\vec{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\vec{x}) \\ \frac{\partial f_2}{\partial x_1}(\vec{x}) & \frac{\partial f_2}{\partial x_2}(\vec{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\vec{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\vec{x}) & \frac{\partial f_m}{\partial x_2}(\vec{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\vec{x}) \end{bmatrix},$$

*deren Zeilen die Gradienten $\text{grad } f_1(\vec{x}), \dots, \text{grad } f_m(\vec{x})$ der Komponentenfunktionen sind, heißt **Jacobi-Matrix** von $\vec{f}$ in $\vec{x}$ und wird mit $\vec{f}'(\vec{x})$ bezeichnet (Jacobi: 1804–1851). Die Determinante dieser Matrix heißt **Funktionaldeterminante**.*

Die Funktionaldeterminante werden wir später bei der Integration über Teilmengen des $\mathbb{R}^n$ benötigen.

Bei vektorwertigen Funktionen kann man einzeln mit den Komponentenfunktionen rechnen. Man kann partielle Ableitungen aber auch unter Ausnutzung der Matrizen-Schreibweise berechnen. Dabei übertragen sich die Ableitungsregeln sinngemäß:

Beispiel 4.12

Wir betrachten eine Situation, die wir später bei Differenzialgleichungssystemen noch benötigen werden.

$$\vec{f}(x) := \mathbf{W}(x)\vec{y}(x), \quad \mathbf{W}(x) := \begin{bmatrix} w_{1,1}(x) & w_{1,2}(x) \\ w_{2,1}(x) & w_{2,2}(x) \end{bmatrix}, \quad \vec{y}(x) := \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix}.$$

$\vec{f} : \mathbb{R} \rightarrow \mathbb{R}^2$, wobei die Funktionen $w_{k,i}$, y_1 und y_2 reelle Funktionen einer reellen Variable seien. Für die Jacobi-Matrix $\vec{f}'(x) = \begin{bmatrix} f'_1(x) \\ f'_2(x) \end{bmatrix} =: \frac{d}{dx} \vec{f}(x)$ gilt:

$$\begin{aligned} \vec{f}'(x) &= \frac{d}{dx} [\mathbf{W}(x) \vec{y}(x)] = \frac{d}{dx} \begin{pmatrix} w_{1,1}(x)y_1(x) + w_{1,2}(x)y_2(x) \\ w_{2,1}(x)y_1(x) + w_{2,2}(x)y_2(x) \end{pmatrix} \\ &= \begin{pmatrix} w'_{1,1}(x)y_1(x) + w'_{1,2}(x)y_2(x) + w_{1,1}(x)y'_1(x) + w_{1,2}(x)y'_2(x) \\ w'_{2,1}(x)y_1(x) + w'_{2,2}(x)y_2(x) + w_{2,1}(x)y'_1(x) + w_{2,2}(x)y'_2(x) \end{pmatrix} \\ &= \left[\frac{d}{dx} \mathbf{W}(x) \right] \vec{y}(x) + \mathbf{W}(x) \frac{d}{dx} \vec{y}(x). \end{aligned} \quad (4.4)$$

Hier verwenden wir $\frac{d}{dx}$ auch für Vektoren und Matrizen und verstehen darunter die komponentenweise Ableitung nach der einzigen Variable x . Bei partiellen Ableitungen würden wir analog $\frac{\partial}{\partial x}$ schreiben. ■

Die Kettenregel $(f \circ g)'(x) = \frac{d}{dx}(f(g(x))) = f'(g(x))g'(x)$ für reelle Funktionen mit einer reellen Variable (siehe Satz 2.33) geht nun über in diese allgemeinere Gestalt:

Satz 4.4 (Kettenregel)

Ist $\vec{g} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$, so dass im inneren Punkt $\vec{x}_0 \in D$ alle partiellen Ableitungen von $\vec{g}$ existieren (wir können also die Jacobi-Matrix aufschreiben). Weiter sei $f : E \subset \mathbb{R}^m \rightarrow \mathbb{R}$, so dass $\vec{g}(D) \subset E$, $\vec{g}(\vec{x}_0)$ ein innerer Punkt von E ist und alle partiellen Ableitungen von f in $\vec{g}(\vec{x}_0)$ existieren. Dann gilt für $1 \leq k \leq n$:

$$\left[\frac{\partial}{\partial x_k} (f \circ \vec{g}) \right] (\vec{x}_0) = (\text{grad } f)(\vec{g}(\vec{x}_0)) \cdot \begin{pmatrix} \frac{\partial g_1}{\partial x_k}(\vec{x}_0) \\ \frac{\partial g_2}{\partial x_k}(\vec{x}_0) \\ \vdots \\ \frac{\partial g_m}{\partial x_k}(\vec{x}_0) \end{pmatrix},$$

d. h., wir erhalten die partielle Ableitung über ein Skalarprodukt. Unter Verwendung der Jacobi-Matrix sieht das so aus:

$$[\text{grad}(f \circ \vec{g})](\vec{x}_0) = (\text{grad } f)(\vec{g}(\vec{x}_0)) \cdot \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\vec{x}_0) & \cdots & \frac{\partial g_1}{\partial x_n}(\vec{x}_0) \\ \vdots & & \vdots \\ \frac{\partial g_m}{\partial x_1}(\vec{x}_0) & \cdots & \frac{\partial g_m}{\partial x_n}(\vec{x}_0) \end{bmatrix}.$$

Hier wird der Zeilenvektor der partiellen Ableitungen $\text{grad } f$ an der Stelle $\vec{g}(\vec{x}_0)$ berechnet und mit der Jacobi-Matrix der ersten partiellen Ableitungen von $\vec{g}$ an der Stelle $\vec{x}_0$ multipliziert. Diese Kettenregel sieht genauso aus wie die altbekannte von Seite

297. Man muss nur die Ableitungen durch die entsprechenden Begriffe für Funktionen mit mehreren Variablen austauschen, so dass die Regel mit Vektoren und Matrizen geschrieben wird.

Beispiel 4.13

Wir verketten eine Funktion $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ mit $g:]0, \infty[\subset \mathbb{R} \rightarrow \mathbb{R}^3$:

$$f(x_1, x_2, x_3) := \exp(x_1 + 3x_2 - x_3), \quad \vec{g}(t) := (t, t^2, \sqrt{t}) = (g_1(t), g_2(t), g_3(t)).$$

Damit ist

$$\begin{aligned} [f \circ \vec{g}]'(t) &= (\text{grad } f)(\vec{g}(t)) \cdot \left(\frac{\partial g_1}{\partial t}(t), \frac{\partial g_2}{\partial t}(t), \frac{\partial g_3}{\partial t}(t) \right)^\top \\ &= \exp(t + 3t^2 - \sqrt{t}) (1, 3, -1) \begin{pmatrix} 1 \\ 2t \\ \frac{1}{2\sqrt{t}} \end{pmatrix} = \exp(t + 3t^2 - \sqrt{t}) \left[1 + 6t - \frac{1}{2\sqrt{t}} \right]. \end{aligned}$$

Wegen $[f \circ \vec{g}](t) = \exp(t + 3t^2 - \sqrt{t})$ folgt mit der „eindimensionalen“ Kettenregel ebenfalls $[f \circ \vec{g}]'(t) = \exp(t + 3t^2 - \sqrt{t}) \left(1 + 6t - \frac{1}{2\sqrt{t}} \right)$. ■

Beispiel 4.14

Sei $f(x_1, x_2) := x_1^{x_2}$ und $\vec{g}(t) := (t, t)$. Dann ist wegen $\frac{\partial}{\partial x_1} x_1^{x_2} = x_1^{x_2-1}$ und $\frac{\partial}{\partial x_2} x_1^{x_2} = \frac{\partial}{\partial x_2} \exp(x_2 \ln x_1) = x_1^{x_2} \ln x_1$:

$$\frac{d}{dt} t^t = [f \circ \vec{g}]'(t) = (\text{grad } f)(\vec{g}(t)) \cdot \begin{pmatrix} \frac{\partial g_1}{\partial t}(t) \\ \frac{\partial g_2}{\partial t}(t) \end{pmatrix} = (t \cdot t^{t-1}, t^t \ln(t)) \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = t^t(1 + \ln t).$$

Dieses Ergebnis ergibt sich auch bei „eindimensionaler“ Rechnung, siehe (2.22) auf Seite 304. ■

Die Kettenregel ist auch für äußere Funktionen $\vec{f}: E \subset \mathbb{R}^m \rightarrow \mathbb{R}^p$ anwendbar, indem man sie für die Komponentenfunktionen $f_1, \dots, f_p$ benutzt. Es entsteht die Formel

$$\begin{aligned} & \begin{bmatrix} \frac{\partial [f_1 \circ \vec{g}]}{\partial x_1}(\vec{x}_0) & \dots & \frac{\partial [f_1 \circ \vec{g}]}{\partial x_n}(\vec{x}_0) \\ \vdots & & \vdots \\ \frac{\partial [f_p \circ \vec{g}]}{\partial x_1}(\vec{x}_0) & \dots & \frac{\partial [f_p \circ \vec{g}]}{\partial x_n}(\vec{x}_0) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{g}(\vec{x}_0)) & \dots & \frac{\partial f_1}{\partial x_n}(\vec{g}(\vec{x}_0)) \\ \vdots & & \vdots \\ \frac{\partial f_p}{\partial x_1}(\vec{g}(\vec{x}_0)) & \dots & \frac{\partial f_p}{\partial x_n}(\vec{g}(\vec{x}_0)) \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\vec{x}_0) & \dots & \frac{\partial g_1}{\partial x_n}(\vec{x}_0) \\ \vdots & & \vdots \\ \frac{\partial g_m}{\partial x_1}(\vec{x}_0) & \dots & \frac{\partial g_m}{\partial x_n}(\vec{x}_0) \end{bmatrix}. \end{aligned}$$

Zum Abschluss noch eine Bemerkung zu komplexwertigen Funktionen einer komplexen Variable. Eine Funktion $f : \mathbb{C} \rightarrow \mathbb{C}$ kann durch Auftrennen der Variable z und des Funktionswerts $f(z)$ in Real- und Imaginärteil auch als Funktion $\vec{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ betrachtet werden über $f(z) = f(x + jy) = f_1(x, y) + jf_2(x, y)$. Auf Seite 419 haben wir ganz kurz die komplexe Ableitung angesprochen, die über einen komplexen Differenzenquotienten erklärt ist. Man kann nachrechnen, dass im Falle komplexer Differenzierbarkeit an einer Stelle $z_0 = x_0 + jy_0$ die partiellen Ableitungen von $\vec{f}$ existieren und sie zusätzlich die **Cauchy-Riemann-Differenzialgleichungen** erfüllen:

$$\frac{\partial}{\partial x} f_1(x_0, y_0) = \frac{\partial}{\partial y} f_2(x_0, y_0), \quad \frac{\partial}{\partial y} f_1(x_0, y_0) = -\frac{\partial}{\partial x} f_2(x_0, y_0). \quad (4.5)$$

4.2.2 Implizite Differenziation und implizite Funktion

Wir wollen in diesem Abschnitt eine differenzierbare Funktion $f :]a, b[\rightarrow \mathbb{R}$ ableiten, die nicht explizit gegeben ist, sondern implizit über eine Gleichung

$$F(x, f(x)) = c$$

definiert ist. Dabei ist c eine Konstante und $F : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ eine Funktion, deren partielle Ableitungen auf einer offenen Menge D existieren. Damit wir die Paare $(x, f(x))$ in $F(x, y)$ einsetzen können, muss zudem die Menge $]a, b[\times \{f(x) : x \in]a, b[\}$ in D enthalten sein.

Die Voraussetzungen sind jetzt so gewählt, dass wir die Kettenregel (Satz 4.4) anwenden können:

$$0 = \frac{d}{dx} F(x, f(x)) = (\text{grad } F)(x, f(x)) \cdot \begin{pmatrix} 1 \\ f'(x) \end{pmatrix} = \frac{\partial F}{\partial x}(x, f(x)) + f'(x) \cdot \frac{\partial F}{\partial y}(x, f(x)).$$

Wir möchten diese Gleichung nach $f'(x)$ auflösen. Dazu benötigen wir als weitere Voraussetzung, dass $\frac{\partial F}{\partial y}(x, f(x)) \neq 0$ ist. Sei also $x_0 \in]a, b[$ eine Stelle mit $\frac{\partial F}{\partial y}(x_0, f(x_0)) \neq 0$, dann gilt:

$$f'(x_0) = -\frac{\frac{\partial F}{\partial x}(x_0, f(x_0))}{\frac{\partial F}{\partial y}(x_0, f(x_0))}.$$

Ist $\frac{\partial F}{\partial y}$ stetig, so ist der Nenner nicht nur für x_0 , sondern sogar in einer Umgebung $]x_0 - \delta, x_0 + \delta[$ ungleich null, so dass hier die Ableitung ausgerechnet werden kann.

Beispiel 4.15

Wir können den oberen Rand des Einheitskreises über die Funktion $f(x) = \sqrt{1 - x^2}$, $x \in [-1, 1]$ beschreiben. Diese erfüllt die Gleichung $x^2 + f(x)^2 = 1$. Mit $F(x, y) := x^2 + y^2$ wenden wir die Regel zur impliziten Differenziation an. Zunächst ist $\text{grad } f(x, y) = (2x, 2y)$. Damit erhalten wir für $x \in]-1, 1[$:

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

■

Die Ableitung im vorangehenden Beispiel hätten wir auch direkt mit der klassischen Kettenregel berechnen können. Die implizite Differenziation wird dann wichtig, wenn man die Gleichung nicht unmittelbar nach f auflösen kann. Dabei stellt sich die Frage, ob und wo es für eine Gleichung $F(x, y) = c$ eine Funktion f gibt, so dass für y die Funktionswerte $y = f(x)$ eingesetzt werden können, so dass die Gleichung erfüllt ist. Ein passendes Kriterium ist der **Satz über die implizite Funktion**: Sei F stetig differenzierbar auf D und die Gleichung für das Zahlenpaar (x_0, y_0) erfüllt (d. h. $F(x_0, y_0) = c$). Dann genügt die oben verwendete Bedingung $\frac{\partial F}{\partial y}(x_0, y_0) \neq 0$, damit in einer (eventuell kleinen) Umgebung $]x_0 - \delta, x_0 + \delta[$ von x_0 eine (eindeutige) stetig differenzierbare Funktion f mit $f(x_0) = y_0$ existiert, so dass $F(x, f(x)) = c$ für alle $x \in]x_0 - \delta, x_0 + \delta[$. Die Ableitung dieser Funktion f haben wir oben berechnet.

Der Satz über die implizite Funktion gilt auch dann, wenn statt der reellen Variablen x und y Vektoren $\vec{x} \in \mathbb{R}^m$ und $\vec{y} \in \mathbb{R}^n$ verwendet werden, $F = \vec{F}$ in den $\mathbb{R}^n$ abbildet und eine Funktion $\vec{f}$ mit m Variablen und Werten im $\mathbb{R}^n$ gesucht ist. Die partiellen Ableitungen nach x und y werden dann durch entsprechende Jakobi-Matrizen ersetzt.

4.2.3 Höhere Ableitungen

Partielle Ableitungen lassen sich als Funktionen auffassen, die dann selbst wieder partiell abgeleitet werden können. So entstehen partielle Ableitungen höherer Ordnung, wobei die Ableitungsordnung die Anzahl aller vorzunehmenden einzelnen partiellen Ableitungen ist. Dabei schreiben wir (am Beispiel einer Funktion $f(x, y)$ und Ableitungen zweiter und dritter Ordnung)

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x} \frac{\partial f}{\partial x}, \quad \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y} \frac{\partial f}{\partial x}, \quad \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x} \frac{\partial f}{\partial y}, \quad \frac{\partial^3 f}{\partial y \partial x^2} = \frac{\partial}{\partial y} \frac{\partial^2 f}{\partial x^2}.$$

Die Ableitungsoperatoren $\frac{\partial}{\partial x}$ und $\frac{\partial}{\partial y}$ sind bei dieser Schreibweise also nacheinander von rechts nach links abzarbeiten. Dagegen wertet man die folgende Notation von links nach rechts (und damit ebenfalls vom Funktionssymbol f nach außen) aus:

$$f_{xy} := (f_x)_y := \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial f_x}{\partial y}.$$

Hier stellt sich die Frage, ob die Reihenfolge der Durchführung der partiellen Ableitungen eine Rolle spielt.

Beispiel 4.16

$$\frac{\partial^2}{\partial x \partial y} \sin(yx^2) = \frac{\partial}{\partial x} [x^2 \cos(yx^2)] = 2x \cos(yx^2) - 2x^3 y \sin(yx^2).$$

Andererseits ist

$$\frac{\partial^2}{\partial y \partial x} \sin(yx^2) = \frac{\partial}{\partial y} [2yx \cos(yx^2)] = 2x \cos(yx^2) - 2x^3 y \sin(yx^2). \quad \blacksquare$$

Die Reihenfolge der Ableitungen spielt zumindest in diesem Beispiel keine Rolle. Allgemeiner gibt der folgende Satz eine hinreichende Bedingung für die Vertauschbarkeit der partiellen Ableitungen:

Satz 4.5 (Hermann Amandus Schwarz (1843–1921))

Sei $D \subset \mathbb{R}^n$ und $f : D \rightarrow \mathbb{R}$, so dass alle partiellen Ableitungen von f bis zur Ordnung 2 von f in einer Umgebung des inneren Punkts $\vec{x}_0 \in D$ existieren. Weiterhin mögen die zweiten partiellen Ableitung in $\vec{x}_0$ stetig sein. Dann gilt für $l, k \in \{1, \dots, n\}$:

$$\frac{\partial}{\partial x_l} \frac{\partial f}{\partial x_k}(\vec{x}_0) = \frac{\partial}{\partial x_k} \frac{\partial f}{\partial x_l}(\vec{x}_0).$$

Die Reihenfolge der partiellen Ableitungen ist also vertauschbar (analog für vektorwertige Funktionen).

Man muss sich schon etwas Mühe geben, um ein Beispiel zu konstruieren, bei dem die Voraussetzungen des Satzes nicht erfüllt und die partiellen Ableitungen nicht vertauschbar sind. Gängig ist das folgende:

Beispiel 4.17

$$f(x, y) := \begin{cases} \frac{xy(x^2 - y^2)}{x^2 + y^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0). \end{cases}$$

Wir rechnen nach, dass diese Funktion stetig auf $\mathbb{R}^2$ ist. Dabei ist nur die Stetigkeit an der Stelle $(0, 0)$ fraglich: Zu $\varepsilon > 0$ wählen wir $\delta := \sqrt{\frac{2}{3}}\varepsilon$, und sei $(x, y) \in \mathbb{R}^2$ mit $|(x, y) - (0, 0)| = \sqrt{x^2 + y^2} < \delta$. Ist $y = 0$, so ist $|f(x, y) - 0| = 0 < \varepsilon$. Ist $y \neq 0$:

$$\left| \frac{xy(x^2 - y^2)}{x^2 + y^2} - 0 \right| = |xy| \left| 1 - \frac{2y^2}{x^2 + y^2} \right| \leq |xy| \left[1 + \frac{2y^2}{y^2} \right] = 3|xy|.$$

Nun benutzen wir $0 \leq (|x| - |y|)^2 = x^2 - 2|xy| + y^2 \implies |xy| \leq \frac{1}{2}[x^2 + y^2]$, um weiter abzuschätzen:

$$\left| \frac{xy(x^2 - y^2)}{x^2 + y^2} - 0 \right| \leq 3|xy| \leq \frac{3}{2}[\sqrt{x^2 + y^2}]^2 = \frac{3}{2}|(x, y)|^2 < \frac{3}{2}\delta^2 = \varepsilon.$$

Nachdem wir uns von der Stetigkeit überzeugt haben, berechnen wir nun die partiellen Ableitungen erster Ordnung mit der Quotientenregel: Für $(x, y) \neq (0, 0)$ ist

$$\begin{aligned} \frac{\partial f}{\partial x}(x, y) &= \frac{(3x^2y - y^3)(x^2 + y^2) - (x^3y - xy^3)2x}{(x^2 + y^2)^2} \\ &= \frac{3x^4y - x^2y^3 + 3x^2y^3 - y^5 - 2x^4y + 2x^2y^3}{(x^2 + y^2)^2} = \frac{x^4y + 4x^2y^3 - y^5}{(x^2 + y^2)^2}, \end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial y}(x, y) &= \frac{(x^3 - 3xy^2)(x^2 + y^2) - (x^3y - xy^3)2y}{(x^2 + y^2)^2} \\ &= \frac{x^5 + x^3y^2 - 3x^3y^2 - 3xy^4 - 2x^3y^2 + 2xy^4}{(x^2 + y^2)^2} = \frac{x^5 - 4x^3y^2 - xy^4}{(x^2 + y^2)^2}.\end{aligned}$$

Im Punkt $(0, 0)$ ist

$$\frac{\partial f}{\partial x}(0, 0) = \lim_{h \rightarrow 0} \frac{1}{h} [f(h, 0) - f(0, 0)] = \lim_{h \rightarrow 0} \frac{1}{h} [0 - 0] = 0, \quad \frac{\partial f}{\partial y}(0, 0) = \lim_{h \rightarrow 0} \frac{1}{h} [0 - 0] = 0.$$

Jetzt sehen wir uns die gemischten zweiten partiellen Ableitungen im Nullpunkt an:

$$\begin{aligned}\frac{\partial^2 f}{\partial x \partial y}(0, 0) &= \lim_{h \rightarrow 0} \frac{1}{h} \left[\frac{\partial f}{\partial y}(h, 0) - \frac{\partial f}{\partial y}(0, 0) \right] = \lim_{h \rightarrow 0} \frac{1}{h} \left[\frac{h^5}{h^4} \right] = 1, \\ \frac{\partial^2 f}{\partial y \partial x}(0, 0) &= \lim_{h \rightarrow 0} \frac{1}{h} \left[\frac{\partial f}{\partial x}(0, h) - \frac{\partial f}{\partial x}(0, 0) \right] = \lim_{h \rightarrow 0} \frac{1}{h} \left[-\frac{h^5}{h^4} \right] = -1.\end{aligned}$$

Die gemischten zweiten partiellen Ableitungen im Nullpunkt sind also unterschiedlich, die Voraussetzungen des Satzes 4.5 können nicht erfüllt sein. Das Beispiel ist tatsächlich so konstruiert, dass zweite Ableitungen im Nullpunkt unstetig sind. Der Regelfall in Anwendungen ist dagegen aber, dass die Ableitungen vertauschbar sind. ■

Vertauschungsprobleme gibt es nach Satz 4.5 nicht, wenn die partiellen Ableitungen stetig sind. Das würdigen wir mit einem weiteren Begriff:

Definition 4.9 (k -fache stetige Differenzierbarkeit)

Sei $D \subset \mathbb{R}^n$ offen und $f : D \rightarrow \mathbb{R}$. f heißt genau dann **k -mal stetig differenzierbar** auf D , wenn alle partiellen Ableitungen bis zur Ordnung k existieren und selbst als Funktionen stetig auf D sind. Eine vektorwertige Funktion $\vec{f} : D \rightarrow \mathbb{R}^m$ heißt (k -mal) stetig differenzierbar auf D genau dann, wenn jede Komponentenfunktion (k -mal) stetig differenzierbar auf D ist.

Man beachte: Ist f stetig differenzierbar (also 1-mal stetig differenzierbar), so existiert nach Satz 4.3 das Totale Differenzial in jedem Punkt aus D .

Definition 4.10 (Hesse-Matrix)

Sei $D \subset \mathbb{R}^n$ und $\vec{x}_0$ ein innerer Punkt von D . Ist die Funktion $f : D \rightarrow \mathbb{R}$ in $\vec{x}_0$ zweimal partiell nach allen Variablen differenzierbar, so nennt man

$$\mathbf{H}(\vec{x}_0) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\vec{x}_0) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\vec{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\vec{x}_0) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\vec{x}_0) & \frac{\partial^2 f}{\partial x_2^2}(\vec{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\vec{x}_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\vec{x}_0) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\vec{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\vec{x}_0) \end{bmatrix},$$

die **Hesse-Matrix** der Funktion f an der Stelle $\vec{x}_0$.

Bemerkungen:

- Die ersten partiellen Ableitungen haben wir als Vektor (Gradient) geschrieben, die zweiten bilden eine Matrix.
- Die Hesse-Matrix entspricht der zweiten Ableitung einer Funktion mit einer Variable und wird im Rahmen einer hinreichenden Bedingung für das Identifizieren von Extremstellen benutzt (siehe Kapitel 4.3).
- Ist die Funktion f zweimal stetig differenzierbar, so ist die Reihenfolge der Ableitungen vertauschbar, und die Hesse-Matrix wird symmetrisch.

Viele Aussagen der Analysis über Funktionen mit einer Variable lassen sich auf Funktionen mit mehreren Variablen übertragen. Ein Beispiel ist der Satz von Taylor (vgl. Satz 2.54 auf Seite 384), bei dem wir auch hier eine Funktion f über ein Polynom annähern, das an einer Stelle $\vec{x}$ gleiche partielle Ableitungen wie f besitzt:

Satz 4.6 (Satz von Taylor)

Sei $D \subset \mathbb{R}^n$ offen und $f : D \rightarrow \mathbb{R}$ eine Funktion, die $(k+1)$ -mal stetig differenzierbar ist. Zu zwei Punkten $\vec{x}$ und $\vec{y} \in D$, für die auch die komplette Verbindungsstrecke in D liegt, existiert eine Zwischenstelle $\vec{\xi}$ auf dieser Verbindungsstrecke, so dass

$$\begin{aligned}
 f(\vec{y}) = f(\vec{x}) &+ \frac{1}{1!} \sum_{l_1=1}^n \frac{\partial f(\vec{x})}{\partial x_{l_1}} (y_{l_1} - x_{l_1}) + \frac{1}{2!} \sum_{l_1, l_2=1}^n \frac{\partial^2 f(\vec{x})}{\partial x_{l_1} \partial x_{l_2}} (y_{l_1} - x_{l_1})(y_{l_2} - x_{l_2}) \\
 &+ \cdots + \frac{1}{k!} \sum_{l_1, l_2, \dots, l_k=1}^n \frac{\partial^k f(\vec{x})}{\partial x_{l_1} \dots \partial x_{l_k}} (y_{l_1} - x_{l_1})(y_{l_2} - x_{l_2}) \dots (y_{l_k} - x_{l_k}) \\
 &+ \frac{1}{(k+1)!} \sum_{l_1, l_2, \dots, l_{k+1}=1}^n \frac{\partial^{k+1} f(\vec{\xi})}{\partial x_{l_1} \dots \partial x_{l_{k+1}}} (y_{l_1} - x_{l_1})(y_{l_2} - x_{l_2}) \dots (y_{l_{k+1}} - x_{l_{k+1}}).
 \end{aligned}$$

Der Beweis ist eine Reduktion auf die bekannte eindimensionale Taylor-Formel auf Seite 384. Diese wendet man unter Zuhilfenahme der mehrdimensionalen Kettenregel auf die Funktion $g(t) := f((1-t)\vec{x} + t\vec{y})$ an, die auf einem offenen Intervall, in dem $[0, 1]$ liegt, $(k+1)$ -mal stetig differenzierbar ist.

Bemerkung: Beim Satz von Taylor muss die Verbindungsstrecke zum Entwicklungspunkt im Definitionsbereich liegen. Man nennt eine Menge $D \subset \mathbb{R}^n$ **konvex** genau dann, wenn sie zu je zwei beliebigen Elementen auch ihre Verbindungsstrecke vollständig enthält. In Definition 2.38 auf Seite 388 haben wir auch linksgekrümmte Funktionen $f :]a, b[\rightarrow \mathbb{R}$ konvex genannt. Tatsächlich ist die Fläche $\{(x, y) \in \mathbb{R}^2 : x \in]a, b[, y \geq f(x)\}$ oberhalb des Funktionsgraphen von f eine konvexe Menge.

Hintergrund: Krümmungsverhalten

Die Hesse-Matrix $\mathbf{H}(\vec{x})$ übernimmt im mehrdimensionalen Satz von Taylor mit dem Term $\frac{1}{2!}[\vec{y} - \vec{x}]^\top \mathbf{H}(\vec{x})[\vec{y} - \vec{x}]$ die Rolle der zweiten Ableitung im Term $\frac{f''(x)}{2!}(y - x)^2$ des entsprechenden eindimensionalen Satzes. Damit beschreibt $\mathbf{H}(\vec{x})$ wie die zweite Ableitung in einer Dimension das Krümmungsverhalten an der Stelle $\vec{x}$. Im Gegensatz zum eindimensionalen Fall können wir aber die Krümmung in unterschiedlichen Richtungen betrachten. Bei beispielsweise zwei Variablen ist die (vorzeichenbehaftete) Größe der Krümmung in Richtung $(\cos(\alpha), \sin(\alpha))$ die Zahl

$$(\cos(\alpha), \sin(\alpha)) \cdot \mathbf{H}(\vec{x}) \cdot \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix}.$$

Später zeigen wir mit Lemma 7.2 für symmetrische Matrizen auf Seite 988, dass die größte Krümmung in Richtung eines Eigenvektors zum größten Eigenwert von $\mathbf{H}(\vec{x})$ vorliegt. Die kleinste Krümmung erhalten wir in Richtung eines Eigenvektors zum kleinsten Eigenwert.

4.2.4 Fehlerrechnung*

Die **Fehlerrechnung** ist eine weitere wichtige Anwendung der mehrdimensionalen Differenzialrechnung, insbesondere des Differenzials (vgl. Seite 309). Wir orientieren uns an der Darstellung von (Dobner und Engelmann, 2002, Band 2, S. 115–119).

Wir möchten einen Funktionswert

$$y = f(x_1, \dots, x_n)$$

berechnen, kennen aber statt der exakten Eingangsdaten $\vec{x} = (x_1, \dots, x_n)$ z. B. durch Messung nur fehlerbehaftete Eingangsdaten $\vec{z} = (z_1, \dots, z_n)$. Auch bei exakten Eingangsdaten entstehen durch Rundungsfehler fehlerbehaftete Zwischenergebnisse. Da im Computer nur Dualzahlen mit endlich vielen Stellen gespeichert werden, lassen sich Rundungsfehler in realen Anwendungen nicht vermeiden. Wir berechnen also $\tilde{y} = f(z_1, \dots, z_n)$ anstatt y . Weiter sei ein absolutes Fehlerniveau als obere Schranke

$$|x_i - z_i| \leq \Delta x_i, \quad i = 1, \dots, n$$

bekannt. Gesucht ist eine Schranke Δy mit $|\tilde{y} - y| \leq \Delta y$, d. h. ein absolutes Fehlerniveau in y .

Ist f stetig differenzierbar, so kann das Problem mit dem Satz von Taylor 4.6 für $k = 0$ (dafür wird der Satz von Taylor zu einem mehrdimensionalen Mittelwertsatz)

gelöst werden. Es gilt $\vec{z} = \vec{x} + (\vec{z} - \vec{x}) = \vec{x} + \vec{h}$ mit dem Fehlervektor $\vec{h} = (h_1, \dots, h_n)$ wobei $h_i = z_i - x_i$, $i = 1, \dots, n$. Der Satz von Taylor liefert dann

$$\tilde{y} = f(\vec{z}) = f(\vec{x} + \vec{h}) = f(\vec{x}) + \sum_{i=1}^n \frac{\partial f(\vec{\xi})}{\partial x_i} h_i = y + \sum_{i=1}^n \frac{\partial f(\vec{\xi})}{\partial x_i} (z_i - x_i),$$

wobei $\vec{\xi} = \vec{x} + t(\vec{z} - \vec{x})$, $0 < t < 1$, eine Zwischenstelle zwischen $\vec{x}$ und $\vec{z}$ ist. Da $y = f(\vec{x})$ die exakte Ausgangsgröße ist, folgt mit der Dreiecksungleichung die Abschätzung

$$|\tilde{y} - y| \leq \sum_{i=1}^n \left| \frac{\partial f(\vec{\xi})}{\partial x_i} \right| |z_i - x_i| \leq \sum_{i=1}^n \left| \frac{\partial f(\vec{\xi})}{\partial x_i} \right| \Delta x_i.$$

Die Punkte $\vec{\xi}$ und $\vec{x}$ liegen in dem n -dimensionalen **Unsicherheitsintervall**

$$I = [z_1 - \Delta x_1, z_1 + \Delta x_1] \times [z_2 - \Delta x_1, z_2 + \Delta x_2] \times \dots \times [z_n - \Delta x_n, z_n + \Delta x_n].$$

Mit mehrdimensionalen Intervallen werden wir später ab Seite 649 intensiver arbeiten.

Verwendet man in der Abschätzung die Maxima der Ableitungen über dem Unsicherheitsintervall I der Daten, so erhält man die **Maximalfehlerabschätzung**

$$|\tilde{y} - y| \leq (\Delta y)_{\max} := \sum_{i=1}^n \left(\max_{\vec{x} \in I} \left| \frac{\partial f(\vec{x})}{\partial x_i} \right| \right) \cdot \Delta x_i. \quad (4.6)$$

Die Abschätzung (4.6) ist schwierig anzuwenden, da man Maxima für die Beträge der Ableitungen kennen muss. Daher rechnet man näherungsweise mit den Werten der Ableitungen in der Mitte $\vec{z}$ des Unsicherheitsintervalls. Dabei nimmt man an, dass die Niveaus der Eingangsfehler Δx_i , $i = 1, \dots, n$, klein sind und die Ableitungen $\frac{\partial f(\vec{x})}{\partial x_i}$ auf dem kleinen Unsicherheitsintervall I nahezu konstant sind. Dann werden sich die Maxima der Ableitungen nur wenig von den Werten der Ableitungen in $\vec{z}$ unterscheiden. Diese Vereinfachungen führen zur **linearisierten Fehlerschätzung**

$$(\Delta y)_{\max} \approx (\Delta y)_{\text{lin}} := \sum_{i=1}^n \left| \frac{\partial f(\vec{z})}{\partial x_i} \right| \cdot \Delta x_i. \quad (4.7)$$

Der tatsächliche Fehler kann zwar im Gegensatz zur Abschätzung (4.6) größer als der durch (4.7) geschätzte Wert sein, im Allgemeinen ist diese Fehlerschranke aber realistisch.

Für den **relativen Fehler** muss man $\frac{|y - \tilde{y}|}{|y|}$ abschätzen. Im Sinne der linearisierten Fehlerabschätzung wird der Zähler durch (4.7) ersetzt. Da man den exakten Wert von y nicht kennt, setzt man im Nenner $\tilde{y}$ ein. Damit erhält man den **relativen linearisierten Fehler**

$$\frac{|y - \tilde{y}|}{|y|} \approx \frac{|(\Delta y)_{\text{lin}}|}{|\tilde{y}|} := \sum_{i=1}^n \left| \frac{\partial f(\vec{z})}{\partial x_i} \right| \cdot \frac{\Delta x_i}{|\tilde{y}|} = \sum_{i=1}^n \left[\left| \frac{\partial f(\vec{z})}{\partial x_i} \right| \cdot \frac{|z_i|}{|\tilde{y}|} \right] \cdot \frac{\Delta x_i}{|z_i|},$$

der sich auf die relativen Größen bezieht. Wir betrachten die Anwendung der Fehlerrechnung an zwei einfachen Beispielen.

Beispiel 4.18

- a) Die Fläche einer rechteckigen Tischlerplatte mit Länge a und Breite b berechnet sich zu

$$F(a, b) = a \cdot b.$$

Die Längenmessungen sind mit einem Fehler von 3 % behaftet, d. h. es gilt

$$a - \tilde{a} = \pm 0,03 \tilde{a}, \quad b - \tilde{b} = \pm 0,03 \tilde{b},$$

wobei a, b die exakten und $\tilde{a}, \tilde{b}$ die gemessenen Längen bedeuten. Wir fragen, wie sich diese Fehler auf das Ergebnis der Flächenberechnung auswirken. Die „wahren“ Werte a und b liegen im Unsicherheitsintervall

$$I = [\tilde{a} - 0,03 \tilde{a}, \tilde{a} + 0,03 \tilde{a}] \times [\tilde{b} - 0,03 \tilde{b}, \tilde{b} + 0,03 \tilde{b}].$$

Mit dem absoluten Fehlerniveau $\Delta a = 0,03 \tilde{a}$, $\Delta b = 0,03 \tilde{b}$ ergibt die linearisierte Fehlerschätzung

$$\begin{aligned} (\Delta F)_{\text{lin}} &= \left| \frac{\partial F}{\partial a}(\tilde{a}, \tilde{b}) \right| \cdot \Delta a + \left| \frac{\partial F}{\partial b}(\tilde{a}, \tilde{b}) \right| \cdot \Delta b \\ &= \tilde{b} \cdot 0,03 \tilde{a} + \tilde{a} \cdot 0,03 \tilde{b} = 0,06 \cdot \tilde{a} \tilde{b}. \end{aligned}$$

Der mit $a \cdot b$ berechnete Flächeninhalt ist mit einem Fehler von 6 % behaftet.

- b) Eine von Messdaten x_1, x_2, x_3 abhängige Größe y werde mit Hilfe der Formel

$$y = f(x_1, x_2, x_3) = \frac{x_1^2 \sqrt{x_2}}{x_3}$$

berechnet, wobei für die x_i unsichere Daten vorliegen:

$$x_1 = 3,0 \pm 0,1, \quad x_2 = 4,0 \pm 0,2, \quad x_3 = 1,0 \pm 0,1.$$

Wir bestimmen den Näherungswert $\tilde{y}$ für y und seine Genauigkeit mit Hilfe der Maximalfehlerabschätzung sowie der linearisierten Fehlerschätzung. Der Unsicherheitsbereich der Daten ist durch das dreidimensionale Intervall

$$I = \{(x_1, x_2, x_3) : 2,9 \leq x_1 \leq 3,1, \quad 3,8 \leq x_2 \leq 4,2, \quad 0,9 \leq x_3 \leq 1,1\}$$

gegeben. Der Vektor der Messwerte lautet $\vec{z} = (3,0; 4,0; 1,0)$ mit dem Vektor der Fehlerniveaus $\vec{\Delta x} = (\Delta x_1, \Delta x_2, \Delta x_3) = (0,1; 0,2; 0,1)$. Den Näherungswert $\tilde{y}$ erhält man durch Einsetzen von $\vec{z}$, also $\tilde{y} = f(\vec{z}) = f(3, 4, 1) = 18$.

Zur Fehlerabschätzung bestimmen wir die partiellen Ableitungen sowie die Werte der Ableitungen, die in (4.6) bzw. (4.7) eingehen. Diese sind in Tabelle 4.1 zusammengefasst. Für die Maximalfehlerabschätzung (4.6) sind die Maxima der Ableitungen über I zu verwenden, die hier leicht zu berechnen sind. Man erhält

$$(\Delta y)_{\text{max}} = \max_{\vec{x} \in I} \left| \frac{\partial f}{\partial x_1}(\vec{x}) \right| \cdot \Delta x_1 + \max_{\vec{x} \in I} \left| \frac{\partial f}{\partial x_2}(\vec{x}) \right| \cdot \Delta x_2 + \max_{\vec{x} \in I} \left| \frac{\partial f}{\partial x_3}(\vec{x}) \right| \cdot \Delta x_3$$

Tab. 4.1: Werte der partiellen Ableitungen zu Beispiel 4.18 b)

Partielle Ableitung	Maximaler Betrag der Ableitung über I	Betrag der Ab- leitung in $\vec{z}$
$\frac{\partial f}{\partial x_1}(x_1, x_2, x_3) = \frac{2x_1\sqrt{x_2}}{x_3^2}$	$\frac{2 \cdot 3,1 \cdot \sqrt{4,2}}{0,9} = 14,118$	$\frac{2 \cdot 3 \cdot \sqrt{4}}{1} = 12$
$\frac{\partial f}{\partial x_2}(x_1, x_2, x_3) = \frac{x_1}{2x_3\sqrt{x_2}}$	$\frac{3,1^2}{2 \cdot 0,9 \cdot \sqrt{3,8}} = 2,739$	$\frac{3^2}{2 \cdot 1 \cdot \sqrt{4}} = 2,25$
$\frac{\partial f}{\partial x_3}(x_1, x_2, x_3) = -\frac{x_1^2\sqrt{x_2}}{x_3^3}$	$\frac{3,1 \cdot \sqrt{4,2}}{0,9^2} = 24,314$	$\frac{3^2 \cdot \sqrt{4}}{1} = 18$

$$\approx 14,118 \cdot 0,1 + 2,739 \cdot 0,2 + 24,314 \cdot 0,1 = 4,391.$$

Der Wert 4,3914 ist eine obere Schranke für den maximalen Fehler, die vom tatsächlichen Fehler nicht überschritten werden kann. Verwendet man die linearisierte Fehlerschätzung, so gilt

$$\begin{aligned} (\Delta y)_{\text{lin}} &= \left| \frac{\partial f}{\partial x_1}(\vec{z}) \right| \cdot \Delta x_1 + \left| \frac{\partial f}{\partial x_2}(\vec{z}) \right| \cdot \Delta x_2 + \left| \frac{\partial f}{\partial x_3}(\vec{z}) \right| \cdot \Delta x_3 \\ &= 12 \cdot 0,1 + 2,25 \cdot 0,2 + 18 \cdot 0,1 = 3,45. \end{aligned}$$

Der Wert 3,45 stellt eine realistische Schätzung des möglichen Fehlers dar, wie man durch Vergleich mit der Maximalfehlerabschätzung sieht. Der wahre Fehler kann allerdings größere Werte annehmen. Im Fall der einfachen Formel für y kann das Minimum $y_{\min}$ und Maximum $y_{\max}$ über I direkt bestimmt werden:

$$18 - 3,096 \approx \frac{2,9^2 \cdot \sqrt{3,8}}{1,1} = y_{\min} \leq y \leq y_{\max} = \frac{3,1^2 \cdot \sqrt{4,2}}{0,9} \approx 18 + 3,883.$$

Diese Ungleichung zeigt, dass der tatsächliche Fehler den Wert 3,883 nicht überschreiten kann. Der durch die Maximalfehlerabschätzung berechnete Wert 4,391 erweist sich als eine etwas zu pessimistische obere Schranke. Der mittels linearisierter Fehlerschätzung bestimmte Wert 3,45 liefert die Größenordnung des Fehlers, kann aber sehr wohl von diesem übertroffen werden. ■

4.3 Extremwertrechnung

Hier beschäftigen wir uns mit Extremwerten einer reellwertigen Funktion, die mehrere reelle Variablen hat. Bei vektorwertigen Funktionen $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ mit $m > 1$ kann man die Funktionswerte nicht der Größe nach vergleichen, so dass man auch nicht nach Extremwerten suchen kann.

4.3.1 Lokale und globale Extrema

Zur Vollständigkeit erweitern wir die Definition 2.20 von Seite 280:

Definition 4.11 (Extremstellen)

Sei $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$.

- f hat genau dann in $\vec{x}_0 \in D$ ein **globales Maximum**, wenn $f(\vec{x}) \leq f(\vec{x}_0)$ für alle $\vec{x} \in D$. Gilt stattdessen sogar $f(\vec{x}) < f(\vec{x}_0)$ für $\vec{x} \neq \vec{x}_0$, so spricht man von einem **strikten globalen Maximum**.
- f hat genau dann in $\vec{x}_0 \in D$ ein **globales Minimum**, wenn $f(\vec{x}) \geq f(\vec{x}_0)$ für alle $\vec{x} \in D$. Gilt stattdessen sogar $f(\vec{x}) > f(\vec{x}_0)$ für $\vec{x} \neq \vec{x}_0$, so spricht man von einem **strikten globalen Minimum**.
- f hat genau dann in $\vec{x}_0 \in D$ ein **lokales Maximum (relatives Maximum)**, wenn für ein $\delta > 0$ gilt: $f(\vec{x}) \leq f(\vec{x}_0)$ für alle $\vec{x} \in D$ mit $|\vec{x} - \vec{x}_0| < \delta$. Gilt hier für $\vec{x} \neq \vec{x}_0$ sogar $f(\vec{x}) < f(\vec{x}_0)$, spricht man von einem **strikten lokalen Maximum**.
- f hat genau dann in $\vec{x}_0 \in D$ ein **lokales Minimum (relatives Minimum)**, wenn für ein $\delta > 0$ gilt: $f(\vec{x}) \geq f(\vec{x}_0)$ für alle $\vec{x} \in D$ mit $|\vec{x} - \vec{x}_0| < \delta$. Gilt hier für $\vec{x} \neq \vec{x}_0$ sogar $f(\vec{x}) > f(\vec{x}_0)$, spricht man von einem **strikten lokalen Minimum**.

Wie zuvor verwenden wir die Bezeichnung Extremum sowohl für ein Minimum als auch für ein Maximum. Wir können nun direkt den Satz von Fermat (Seite 314) auf Funktionen mit Definitionsbereich aus $\mathbb{R}^n$ verallgemeinern:

Satz 4.7 (Notwendige Bedingung für ein lokales Extremum)

Hat $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ im inneren Punkt $\vec{x}_0 \in D$ ein lokales Extremum und ist f dort partiell differenzierbar nach allen Variablen, dann gilt: $(\text{grad } f)(\vec{x}_0) = \vec{0}$.

Beweis: Da f in $\vec{x}_0 = (x_{0,1}, x_{0,2}, \dots, x_{0,n}) \in D$ ein lokales Extremum besitzt, haben insbesondere die Funktionen $g_k(x) := f(x_{0,1}, \dots, x_{0,k-1}, x, x_{0,k+1}, \dots, x_{0,n})$ in $x_{0,k}$ ein lokales Extremum. Damit folgt mit dem Satz von Fermat für Funktionen mit einer Variable, dass $g'_k(x_{0,k}) = 0$, oder anders ausgedrückt: $(\text{grad } f)(\vec{x}_0) = \vec{0}$. $\square$

Beispiel 4.19

Wir suchen lokale Extrema der Funktion $f_1 : \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\} \rightarrow \mathbb{R}$, $f_1(x, y) := \sqrt{1 - x^2 - y^2}$, aus Abbildung 4.1. Aus der notwendigen Bedingung

$$(0, 0) = \text{grad } f(x, y) = \left(\frac{-2x}{2\sqrt{1 - x^2 - y^2}}, \frac{-2y}{2\sqrt{1 - x^2 - y^2}} \right)$$

folgt $x = y = 0$. Ein lokales Extremum kann also nur im Punkt $(0, 0)$ vorliegen. Dies ist hier (sogar) das globale Maximum. ■

Beispiel 4.20 (Spektralnorm)

Sei $\mathbf{A}$ eine reelle $(m \times n)$ -Matrix. Mit $\vec{x} \in \mathbb{R}^n$ bezeichnen wir hier Spaltenvektoren. Wir berechnen die Spektralnorm (siehe Seite 571)

$$\|\mathbf{A}\|_2 := \sup_{0 \neq \vec{x} \in \mathbb{R}^n} \frac{|\mathbf{A}\vec{x}|}{|\vec{x}|} = \sup_{0 \neq \vec{x} \in \mathbb{R}^n} \frac{\|\mathbf{A}\vec{x}\|_2}{\|\vec{x}\|_2},$$

indem wir ein Maximum der Funktion

$$f(\vec{x}) := \left[\frac{|\mathbf{A}\vec{x}|}{|\vec{x}|} \right]^2 = \frac{\vec{x}^\top \mathbf{A}^\top \mathbf{A} \vec{x}}{\vec{x}^\top \vec{x}}$$

suchen. Man kann nachrechnen, dass die Menge der Funktionswerte 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 dieser Funktion angenommen wird. Darauf gehen wir hier nicht weiter ein. Stattdessen benutzen wir die notwendige Bedingung für ein lokales Extremum. Dazu berechnen wir die partiellen Ableitungen. Mittels Kettenregel ist

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

Damit und mit der Quotientenregel erhalten wir:

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

Aus der notwendigen Bedingung $\text{grad } f(\vec{x}) = \vec{0}$ 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 s , so ist $s \geq 0$ und insbesondere reell, denn

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

also $s = \frac{|\mathbf{A}\vec{y}|^2}{|\vec{y}|^2} \geq 0$. Außerdem ist $\|\mathbf{A}\|_2 \geq \frac{|\mathbf{A}\vec{y}|}{|\vec{y}|} = \sqrt{s}$. 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 $\|\mathbf{A}\|_2^2 = \left(\frac{|\mathbf{A}\vec{x}_0|}{|\vec{x}_0|}\right)^2$ insbesondere der größte Eigenwert von $\mathbf{A}^\top \mathbf{A}$.

Bemerkung: Ist $m = n$ und $\mathbf{A} \in \mathbb{R}^{n \times n}$ z. B. eine symmetrische Matrix, dann ist $\mathbf{A}$ nach Satz 3.41 auf Seite 552 diagonalisierbar zu $\mathbf{A} = \mathbf{X}\mathbf{D}\mathbf{X}^\top$, wobei $\mathbf{D}$ die Diagonalmatrix der Eigenwerte von $\mathbf{A}$ ist, und $\mathbf{X}^{-1} = \mathbf{X}^\top$. Damit gilt

$$\mathbf{A}^\top \mathbf{A} = \mathbf{A}^2 = \mathbf{X}\mathbf{D}\mathbf{X}^\top \mathbf{X}\mathbf{D}\mathbf{X}^\top = \mathbf{X}\mathbf{D}(\mathbf{X}^{-1}\mathbf{X})\mathbf{D}\mathbf{X}^\top = \mathbf{X}\mathbf{D}^2\mathbf{X}^\top,$$

so dass die quadrierten Eigenwerte von $\mathbf{A}$ die Eigenwerte von $\mathbf{A}^2$ sind, die auf der Diagonale von $\mathbf{D}^2$ stehen. Umgekehrt ist zu jedem Eigenwert s von $\mathbf{A}^2$ (als Diagonalelement von $\mathbf{D}^2$) auch $\sqrt{s}$ oder $-\sqrt{s}$ ein Eigenwert von $\mathbf{A}$ (als Diagonalelement von $\mathbf{D}$). Damit ist dann $\|\mathbf{A}\|_2$ gleich dem betragsmäßig größten Eigenwert (dem Spektralradius) von $\mathbf{A}$. ■

Beispiel 4.21

Wir werden in Kapitel 6.1 Signale in ihre Frequenzbestandteile zerlegen, d. h. als Überlagerung vieler Sinus- und Kosinusfunktionen mit unterschiedlichen Kreisfrequenzen schreiben. So lässt sich die komplexe Wechselstromrechnung (siehe Kapitel 1.6.5) auf beliebige periodische Spannungen und Ströme übertragen. Die Amplituden der einzelnen Sinus- und Kosinusfunktionen, die wir Fourier-Koeffizienten nennen werden, erhält man über eine Optimierungsaufgabe unter Verwendung des Gradienten.

Hat man eine 2π -periodische, auf $[-\pi, \pi]$ integrierbare Funktion f (z. B. eine periodische Spannung), so kann man diese als Überlagerung von Funktionen $a_k \cos(kt)$ und $b_k \sin(kt)$, $k \in \mathbb{Z}$ darstellen. Die Amplituden a_k und b_k heißen Fourier-Koeffizienten und werden in Kapitel 6.1 so eingeführt, dass der quadratische Fehler

$$\begin{aligned} g(a_0, \dots, a_n, b_1, \dots, b_n) &:= \int_{-\pi}^{\pi} \left[f(x) - \left(a_0 + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)) \right) \right]^2 dx \\ &= \int_{-\pi}^{\pi} \left[a_0 + \left(-f(x) + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)) \right) \right]^2 dx \end{aligned}$$

(die Abweichung zwischen der Funktion f und einer endlichen Summe dieser Sinus- und Kosinus-Funktionen, siehe (6.1)) minimal wird, die Funktion f also bestmöglich durch die Summe angenähert wird. $g : \mathbb{R}^{2n+1} \rightarrow \mathbb{R}$ ist eine Funktion von $2n+1$ Variablen, die hier nicht $x_1, x_2, \dots, x_{2n+1}$, sondern $a_0, a_1, \dots, a_n, b_1, \dots, b_n$ heißen.

Wir suchen also insbesondere ein lokales Minimum dieser Funktion. Aus der notwendigen Bedingung $\text{grad } g(a_0, \dots, a_n, b_1, \dots, b_n) = \vec{0}$ ergibt sich für a_0 :

$$\frac{\partial g}{\partial a_0}(a_0, \dots, a_n, b_1, \dots, b_n)$$

$$\begin{aligned}
&= \frac{\partial}{\partial a_0} \int_{-\pi}^{\pi} a_0^2 + 2a_0 \left[-f(x) + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)) \right] \\
&\quad + \left[-f(x) + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)) \right]^2 dx \\
&= \frac{\partial}{\partial a_0} a_0^2 \int_{-\pi}^{\pi} 1 dx + \frac{\partial}{\partial a_0} 2a_0 \int_{-\pi}^{\pi} \left[-f(x) + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)) \right] dx \\
&\quad + \frac{\partial}{\partial a_0} \int_{-\pi}^{\pi} \left[-f(x) + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)) \right]^2 dx \\
&= 2a_0 2\pi + 2 \int_{-\pi}^{\pi} \left[-f(x) + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)) \right] dx \\
&= 2a_0 2\pi - 2 \int_{-\pi}^{\pi} f(x) dx,
\end{aligned}$$

da $\int_{-\pi}^{\pi} \cos(kx) dx = \int_{-\pi}^{\pi} \sin(kx) dx = 0$. Aus $0 = \frac{\partial g}{\partial a_0}(a_0, \dots, a_n, b_1, \dots, b_n)$ folgt damit

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx. \quad (4.8)$$

Genauso erhält man (mit etwas mehr Aufwand bei der Integration) die Darstellung der anderen Fourier-Koeffizienten. Wir rechnen diese auf Seite 827 aus. ■

Um eine hinreichende Bedingung für lokale Extremstellen zu erhalten, müssen wir analog zur Kurvendiskussion beim eindimensionalen Definitionsbereich (siehe Folgerung 2.7 auf Seite 388) zweite Ableitungen betrachten, nämlich jetzt die Hesse-Matrix. Zur Vorbereitung sehen wir uns eine zweimal stetig differenzierbare Funktion $f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$ an einer lokalen Extremstelle (x_0, y_0) an. Taylor-Entwicklung (siehe Satz 4.6) liefert:

$$\begin{aligned}
f(x_0 + h, y_0 + \delta) &= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0) \cdot (x_0 + h - x_0) + \frac{\partial f}{\partial y}(x_0, y_0) \cdot (y_0 + \delta - y_0) \\
&\quad + \frac{1}{2!} \left[\frac{\partial^2 f}{\partial x^2}(\vec{\xi}) \cdot h^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(\vec{\xi}) \cdot h\delta + \frac{\partial^2 f}{\partial y^2}(\vec{\xi}) \cdot \delta^2 \right],
\end{aligned}$$

wobei die Stelle $\vec{\xi}$ auf der Verbindungsstrecke zwischen (x_0, y_0) und $(x_0 + h, y_0 + \delta)$ liegt. Nach Satz 4.7 sind die partiellen Ableitungen an der Stelle (x_0, y_0) null, und wir erhalten

$$\begin{aligned}
f(x_0 + h, y_0 + \delta) - f(x_0, y_0) &= \frac{1}{2} \left[\frac{\partial^2 f}{\partial x^2}(\vec{\xi}) h^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(\vec{\xi}) h\delta + \frac{\partial^2 f}{\partial y^2}(\vec{\xi}) \delta^2 \right] \\
&= \frac{1}{2} (h, \delta) \cdot \begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(\vec{\xi}) & \frac{\partial^2 f}{\partial x \partial y}(\vec{\xi}) \\ \frac{\partial^2 f}{\partial x \partial y}(\vec{\xi}) & \frac{\partial^2 f}{\partial y^2}(\vec{\xi}) \end{bmatrix} \cdot \begin{pmatrix} h \\ \delta \end{pmatrix}. \quad (4.9)
\end{aligned}$$

Wir haben an der Stelle (x_0, y_0) ein (striktes) lokales Minimum, wenn für kleine Werte von $|h|$ und $|\delta|$ mit $(h, \delta) \neq (0, 0)$ das Produkt (4.9) stets größer als null ist. Entsprechend liegt ein (striktes) lokales Maximum vor, wenn (4.9) kleiner als null ist. Unschön ist, dass die Matrix in (4.9) von der nicht explizit bekannten Stelle $\vec{\xi}$ und damit aber auch von h und δ abhängt. Wegen der zweimaligen stetigen Differenzierbarkeit von f sind die Einträge aber stetige Funktionen des Parameters $\vec{\xi}$. Wählt man h und δ nahe genug bei null und damit $\vec{\xi}$ nah bei (x_0, y_0) , kann man die Matrix durch die Hesse-Matrix an der Stelle (x_0, y_0) ersetzen, um eine hinreichende Bedingung zu erhalten. Wir stellen diese mit dem folgenden Begriff etwas allgemeiner dar:

Definition 4.12 (Positiv definite Matrix)

Eine symmetrische reelle $(n \times n)$ -Matrix $\mathbf{A}$ heißt genau dann **positiv definit**, wenn

$$(h_1, h_2, \dots, h_n) \mathbf{A} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix} > 0$$

für alle Vektoren $\vec{0} \neq \vec{h} \in \mathbb{R}^n$; sie heißt genau dann **negativ definit**, wenn für alle Vektoren $\vec{0} \neq \vec{h} \in \mathbb{R}^n$ gilt:

$$(h_1, h_2, \dots, h_n) \mathbf{A} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix} < 0.$$

Beispiel 4.22

Die symmetrische Matrix $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ ist positiv definit, da für alle Vektoren $\vec{h} \neq \vec{0}$ gilt:

$$\begin{aligned} (h_1, h_2) \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} &= h_1(h_1 + h_2) + h_2(h_1 + 2h_2) \\ &= h_1^2 + 2h_1h_2 + 2h_2^2 = (h_1 + h_2)^2 + h_2^2 > 0. \end{aligned}$$

Die symmetrische Matrix $\begin{bmatrix} -1 & 1 \\ 1 & -2 \end{bmatrix}$ ist negativ definit, da für alle $\vec{h} \neq \vec{0}$ gilt:

$$(h_1, h_2) \begin{bmatrix} -1 & 1 \\ 1 & -2 \end{bmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = h_1(-h_1 + h_2) + h_2(h_1 - 2h_2)$$

$$= -h_1^2 + 2h_1h_2 - 2h_2^2 = -(h_1 - h_2)^2 - h_2^2 < 0. \blacksquare$$

Eine symmetrische Matrix $\mathbf{A}$ hat nach Satz 3.36 auf Seite 545 ausschließlich reelle Eigenwerte. Ist $\mathbf{A}$ positiv definit, so ist zwangsläufig jeder dieser reellen Eigenwerte positiv. Denn ist s ein (reeller) Eigenwert zu einem Eigenvektor $\vec{d}$, so gilt: $\vec{d}^\top \mathbf{A} \vec{d} = s \vec{d}^\top \vec{d} = s |\vec{d}|^2 > 0$. Damit muss $s > 0$ sein. Eine notwendige Bedingung für die positive Definitheit ist damit, dass alle Eigenwerte positiv sind. Diese Bedingung ist auch hinreichend. Sind alle Eigenwerte einer symmetrischen, reellen Matrix $\mathbf{A}$ positiv, so lässt sich $\mathbf{A}$ nach Satz 3.41 auf Seite 552 mit einer orthogonalen Transformationsmatrix $\mathbf{X}$ diagonalisieren: $\mathbf{D} = \mathbf{X}^\top \mathbf{A} \mathbf{X}$. Dabei ist $\mathbf{D}$ eine Diagonalmatrix mit den positiven Eigenwerten $d_1, d_2, \dots, d_n$ als Diagonalelemente. Damit ergibt sich für jeden Vektor $\vec{h} \neq \vec{0}$:

$$\vec{h}^\top \mathbf{A} \vec{h} = \vec{h}^\top \mathbf{X} \mathbf{D} \mathbf{X}^\top \vec{h} = (\mathbf{X}^\top \vec{h})^\top \underbrace{\mathbf{D} (\mathbf{X}^\top \vec{h})}_{=: \vec{a}} = \vec{a}^\top \mathbf{D} \vec{a} = \sum_{k=1}^n d_k a_k^2 > 0,$$

da $\mathbf{X}^\top$ invertierbar ist und somit wegen $\vec{h} \neq \vec{0}$ auch $\vec{a} \neq \vec{0}$ ist.

Völlig analog ist die negative Definitheit damit äquivalent, dass alle Eigenwerte negativ sind. Damit haben wir bewiesen:

Satz 4.8 (Charakterisierung positiv und negativ definiter Matrizen)

Eine symmetrische, reelle Matrix ist genau dann positiv (negativ) definit, wenn alle ihre Eigenwerte positiv (negativ) sind.

Beispiel 4.23 (Iteratives Lösen eines linearen Gleichungssystems)

Um eine Lösung $\vec{x} \in \mathbb{R}^n$ eines linearen Gleichungssystems $\mathbf{A} \vec{x} = \vec{b}$ mit $\mathbf{A} \in \mathbb{R}^{n \times n}$ und $\vec{b} \in \mathbb{R}^n$ zu bestimmen, betrachten wir die Funktion

$$f(\vec{x}) := \frac{1}{2} \vec{x}^\top \cdot \mathbf{A} \cdot \vec{x} - \vec{b}^\top \cdot \vec{x}.$$

Dabei fassen wir $\vec{x}$ und $\vec{b}$ als Spaltenvektoren auf, so dass die Produkte Matrixmultiplikationen (bzw. Standardskalarprodukte zwischen Vektoren) sind. Dann lässt sich der Gradient elementar ausrechnen zu

$$\text{grad } f(\vec{x}) = \mathbf{A} \cdot \vec{x} - \vec{b}.$$

Wenn wir die Stelle $\vec{x}$ eines lokalen Extremums von f finden, dann ist

$$\vec{0} = \text{grad } f(\vec{x}) = \mathbf{A} \cdot \vec{x} - \vec{b} \iff \mathbf{A} \cdot \vec{x} = \vec{b},$$

und wir haben eine Lösung des Gleichungssystems gefunden. Damit eignet sich das Gradientenabstiegsverfahren aus Beispiel 4.9 (Seite 603) auch zum (numerischen) Lösen von Gleichungssystemen. In Richtung des (bei einem Minimum negativen) Gradienten gehen wir dabei ausgehend von einer Stelle $\vec{x}^{(k)}$ nicht ein beliebiges Stück zu einer Stelle $\vec{x}^{(k+1)}$, sondern wir suchen ein lokales Extremum von f eingeschränkt auf die durch den Gradienten

$$\vec{r}^{(k)} := \text{grad } f(\vec{x}^{(k)}) = \mathbf{A} \cdot \vec{x}^{(k)} - \vec{b}$$

vorgegebenen Gerade $\vec{x}^{(k)} + r \vec{r}^{(k)}$. Den verwendeten Richtungsvektor $\vec{r}^{(k)}$ nennt man **Residuum**, er beschreibt den verbleibenden Rest zwischen $\mathbf{A} \cdot \vec{x}^{(k)}$ und $\vec{b}$.

Aufgrund der notwendigen Bedingung für ein lokales Extremum hinsichtlich des Parameters r ermitteln wir eine Nullstelle der Ableitung nach r :

$$0 = \frac{d}{dr} f(\vec{x}^{(k)} + r \vec{r}^{(k)}) = \left[\mathbf{A}(\vec{x}^{(k)} + r \vec{r}^{(k)}) - \vec{b} \right]^\top \cdot \vec{r}^{(k)} = \left[\vec{r}^{(k)} + r \mathbf{A} \vec{r}^{(k)} \right]^\top \cdot \vec{r}^{(k)},$$

so dass

$$r_k := r = - \frac{(\vec{r}^{(k)})^\top \cdot \vec{r}^{(k)}}{(\mathbf{A} \vec{r}^{(k)})^\top \cdot \vec{r}^{(k)}}. \quad (4.10)$$

Damit ist $\vec{x}^{(k+1)} := \vec{x}^{(k)} + r_k \vec{r}^{(k)}$ berechnet.

Für die speziellen, aber in der Praxis häufig auftretenden symmetrischen, positiv-definiten Matrizen $\mathbf{A}$ aus Definition 4.12 kann man sogar mit dem **Verfahren der konjugierten Gradienten (CG-Verfahren)** sicherstellen, dass eine Lösung von $\mathbf{A} \vec{x} = \vec{b}$ nach n Abstiegsschritten gefunden wird. Über die Matrix $\mathbf{A}$ lässt sich ein Skalarprodukt $\bullet$ über $\vec{a} \bullet \vec{b} := \vec{a}^\top \cdot (\mathbf{A} \vec{b})$ definieren. Dieses haben wir bereits im Nenner von (4.10) verwendet. Die Anforderungen an die Matrix sind erforderlich, damit tatsächlich auf diese Weise ein Skalarprodukt definiert ist.

Das CG-Verfahren startet wie das Gradientenverfahren mit der Suchrichtung $\vec{r}^{(1)} = \text{grad } f(\vec{x}^{(1)}) = \mathbf{A} \cdot \vec{x}^{(1)} - \vec{b}$. Die Iterationsvorschrift lautet dann

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} + s_k \vec{s}^{(k)}$$

mit $\vec{s}^{(1)} := \vec{r}^{(1)}$ und für $k > 1$

$$\vec{s}^{(k)} := \text{grad } f(\vec{x}^{(k)}) - \frac{\text{grad } f(\vec{x}^{(k)}) \mathbf{A} \vec{s}^{(k-1)}}{(\vec{s}^{(k-1)})^\top \mathbf{A} \vec{s}^{(k-1)}} \vec{s}^{(k-1)} = \vec{r}^{(k)} - \frac{(\vec{r}^{(k)})^\top \mathbf{A} \vec{s}^{(k-1)}}{(\vec{s}^{(k-1)})^\top \mathbf{A} \vec{s}^{(k-1)}} \vec{s}^{(k-1)}$$

und (vgl (4.10))

$$s_k = - \frac{(\vec{r}^{(k)})^\top \cdot \vec{r}^{(k)}}{(\mathbf{A} \vec{s}^{(k)})^\top \cdot \vec{s}^{(k)}}.$$

Dabei ist $\vec{r}^{(k)} := \text{grad } f(\vec{x}^{(k)}) = \mathbf{A} \cdot \vec{x}^{(k)} - \vec{b}$ wie zuvor das Residuum für $\vec{x}^{(k)}$. Allerdings ergibt sich nun eine andere Folge von Stellen $\vec{x}^{(k)}$, so dass die Werte von $\vec{r}^{(k)}$ andere als beim reinen Gradientenverfahren sein können. Wir erhalten $\vec{r}^{(k)}$ über

$$\vec{r}^{(k)} = \mathbf{A} \cdot \vec{x}^{(k)} - \vec{b} = \mathbf{A} \cdot (\vec{x}^{(k-1)} + s_{k-1} \vec{s}^{(k-1)}) - \vec{b} = \vec{r}^{(k-1)} + s_{k-1} \mathbf{A} \vec{s}^{(k-1)}.$$

Es lässt sich mit Vollständiger Induktion zeigen, dass die Suchrichtungen $\vec{s}^{(k)}$, $1 \leq k \leq n$, paarweise orthogonal bezüglich des Skalarprodukts $\bullet$ sind. Nach n Schritten gibt es keine weiteren orthogonalen Richtungen, und es lässt sich zeigen, dass spätestens dann eine Lösung des Gleichungssystems gefunden ist. Damit das Verfahren auch bei ungenauer Arithmetik anwendbar wird, wird es üblicher Weise in Verbindung mit Vorkonditionierung (siehe Seite 573) eingesetzt. ■

Wir haben zuvor für den Fall $n = 2$ gezeigt, dass es für ein lokales Minimum (lokales Maximum) hinreichend ist, wenn die Hesse-Matrix an der betrachteten Stelle positiv (negativ) definit ist. Das lässt sich verallgemeinern:

Satz 4.9 (Hinreichende Bedingung für ein lokales Extremum)

Sei $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ im inneren Punkt $\vec{x}_0 \in D$ zweimal stetig differenzierbar mit $(\text{grad } f)(\vec{x}_0) = \vec{0}$. Dann gilt:

- a) Ist die Hesse-Matrix $\mathbf{H}(\vec{x}_0)$ negativ definit, so hat f in $\vec{x}_0$ ein lokales Maximum.
- b) Ist $\mathbf{H}(\vec{x}_0)$ positiv definit, so hat f in $\vec{x}_0$ ein lokales Minimum.

Diesen Satz kann man über die mehrdimensionale Taylor-Entwicklung wie zuvor im Fall $n = 2$ beweisen. Man beachte, dass aufgrund der vorausgesetzten zweimaligen stetigen Differenzierbarkeit die Hesse-Matrix nach Satz 4.5 symmetrisch ist, so dass Definition 4.12 anwendbar ist.

Um tatsächlich auszurechnen, ob eine Matrix positiv oder negativ definit ist, benötigt man ein einfaches Kriterium. Schließlich möchte man nicht mit allen denkbaren Vektoren multiplizieren und auch nicht die Eigenwerte bestimmen, da dazu Nullstellen von Polynomen berechnet werden müssten.

Lemma 4.2 (Sylvester-Kriterium für Definitheit)

Sei $\mathbf{A}$ eine symmetrische $(n \times n)$ -Matrix mit Elementen $a_{k,l}$, $1 \leq k \leq n$, $1 \leq l \leq n$. $\mathbf{A}$ ist genau dann positiv definit, falls alle Hauptabschnittsdeterminanten

$$H_1 := \det[a_{1,1}], H_2 := \det \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}, H_3 := \det \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}, \dots,$$

$H_n := \det \mathbf{A}$ positiv sind. $\mathbf{A}$ ist genau dann negativ definit, falls die Hauptabschnittsdeterminanten ein alternierendes Vorzeichen besitzen, wobei $H_1 = \det[a_{1,1}] < 0$ ist, also $H_1 < 0$, $H_2 > 0$, $H_3 < 0$, $H_4 > 0, \dots$

Achtung: Bei einer negativ definiten Matrix alterniert das Vorzeichen beginnend mit $H_1 < 0$ (**nicht** mit $H_1 > 0$). Wir haben **keine** Analogie zur positiven Definitheit im Sinne, dass alle Determinanten kleiner null sind.

Beweis: (Skizze) Wir beweisen die benötigte Richtung des Sylvester-Kriteriums, dass aus dem Verhalten der Determinanten die positive oder negative Definitheit folgt, exemplarisch für $n = 2$:

$$\begin{aligned} (h_1, h_2) \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{1,2} & a_{2,2} \end{bmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} &= (a_{1,1}h_1 + a_{1,2}h_2, a_{1,2}h_1 + a_{2,2}h_2) \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \\ &= a_{1,1}h_1^2 + a_{1,2}h_1h_2 + a_{1,2}h_1h_2 + a_{2,2}h_2^2 = a_{1,1}h_1^2 + 2a_{1,2}h_1h_2 + a_{2,2}h_2^2. \end{aligned}$$

Den letzten Ausdruck kann man im Fall $a_{1,1} \neq 0$ geschickt umschreiben zu

$$a_{1,1} \left(h_1 + \frac{h_2 a_{1,2}}{a_{1,1}} \right)^2 + \frac{a_{1,1} a_{2,2} - a_{1,2}^2}{a_{1,1}} h_2^2.$$

Hier kann man die Bedingungen ablesen: Ist $H_1 = a_{1,1} > 0$ und $a_{1,1}a_{2,2} - a_{1,2}^2 = \det \mathbf{A} = H_2 > 0$, so ist der Ausdruck für $\vec{h} \neq \vec{0}$ größer null und damit $\mathbf{A}$ positiv definit. Ist $H_1 = a_{1,1} < 0$ und $a_{1,1}a_{2,2} - a_{1,2}^2 = \det \mathbf{A} = H_2 > 0$, so ist der Ausdruck kleiner null und damit $\mathbf{A}$ negativ definit.

Ist umgekehrt eine symmetrische, reelle Matrix positiv (oder negativ) definit, dann gilt das auch für alle Matrizen $\mathbf{A}_k$, $1 \leq k \leq n$, die bei der Berechnung der Hauptabschnittsdeterminante verwendet werden. Jede dieser Matrizen kann wieder wegen Satz 3.41 zu einer Diagonalmatrix $\mathbf{D}_k = \mathbf{X}_k^\top \mathbf{A}_k \mathbf{X}_k = \mathbf{X}_k^{-1} \mathbf{A}_k \mathbf{X}_k$ mit den Eigenwerten von $\mathbf{A}_k$ als Diagonalelemente transformiert werden. Sind $\mathbf{A}$ und damit $\mathbf{A}_k$ positiv definit, so sind alle Eigenwerte positiv und $\det(\mathbf{D}_k) > 0$ als Produkt der Eigenwerte. Daher ist auch

$$\det(\mathbf{A}_k) = \det(\mathbf{X}_k \mathbf{D}_k \mathbf{X}_k^{-1}) = \frac{\det(\mathbf{X}_k)}{\det(\mathbf{X}_k)} \det(\mathbf{D}_k) > 0.$$

Sind $\mathbf{A}$ und damit $\mathbf{A}_k$ negativ definit, so sind alle Eigenwerte negativ und $(-1)^k \det(\mathbf{D}_k) > 0$, so dass auch $(-1)^k \det(\mathbf{A}_k) > 0$ ist. $\square$

Man kann sich das Kriterium in Anlehnung an den letzten Teil des Beweises so merken: Bei einer positiv definiten Diagonalmatrix sind die Hauptabschnittsdeterminanten das Produkt der positiven Diagonalelemente (Eigenwerte) und damit alle größer null. Bei einer negativ definiten Diagonalmatrix sind alle Diagonalelemente (Eigenwerte) kleiner null. Die Hauptabschnittsdeterminanten sind dann abwechselnd das Produkt von ungeradzahlig und gradzahlig vielen negativen Zahlen. So entsteht das alternierende Vorzeichen.

Hat man eine zweimal differenzierbare Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ mit $f'(x_0) = 0$, so besteht die (1×1) -Hesse-Matrix $\mathbf{H}(x_0) = [f''(x_0)]$ nur aus der zweiten Ableitung. Die für ein Maximum hinreichende Bedingung $f''(x_0) < 0$ aus Folgerung 2.7 ist identisch mit $H_1 < 0$ (also negativ definite Hesse-Matrix), und die für ein Minimum hinreichende Bedingung $f''(x_0) > 0$ ist identisch mit $H_1 > 0$ (positiv definite Hesse-Matrix).

Beispiel 4.24

Für $f(x, y) := x^3 + y^3 - 3x - 27y + 24$ ergeben sich aus der notwendigen Bedingung $\vec{0} = \text{grad } f(\vec{x}) = (3x^2 - 3, 3y^2 - 27)$ die vier Kandidaten $(\pm 1, \pm 3)$ für lokale Extrema.

Die Hesse-Matrix ist $\mathbf{H}(\vec{x}) = \begin{bmatrix} 6x & 0 \\ 0 & 6y \end{bmatrix}$, so dass $H_1 = 6x$, $H_2 = 36xy$. Damit haben

wir im Punkt $(1, 3)$ ein lokales Minimum, in $(-1, 3)$ und $(1, -3)$ keine Aussage (denn wir wissen nicht, ob $\mathbf{H}$ hier positiv oder negativ definit ist) und in $(-1, -3)$ ein lokales Maximum. ■

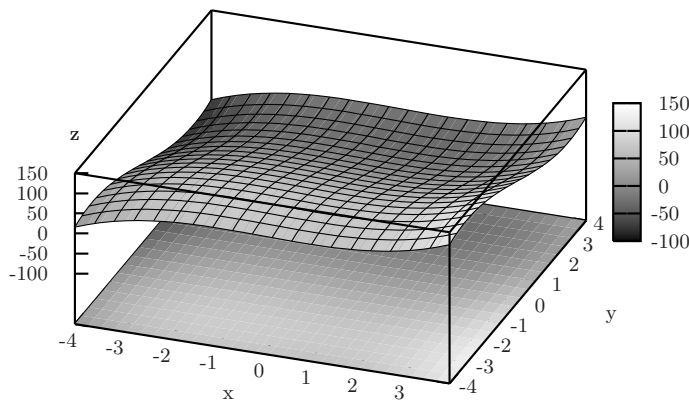


Abb. 4.7: $f(x, y) := x^3 + y^3 - 3x - 27y + 24$

Hintergrund: Cholesky-Zerlegung

Für symmetrische, positiv definite Matrizen kann man vergleichsweise effizient eine Darstellung als Produkt einer unteren mit einer oberen Dreiecksmatrix ausrechnen. Solche LR-Zerlegungen haben wir in Kapitel 1.7.6 auf Seite 176 besprochen. Dort haben wir sie mittels Gauß-Umformungen berechnet. Das geht für symmetrische, positiv definite Matrizen einfacher und effizienter. Hat man eine LR-Zerlegung, so kann man leicht durch fortgesetztes Einsetzen zugehörige Gleichungssysteme lösen.

Wir wissen bereits mit Satz 3.41 auf Seite 552, dass jede symmetrische reelle Matrix $\mathbf{A}$ eine Darstellung $\mathbf{A} = \mathbf{X}\mathbf{D}\mathbf{X}^\top$ mit einer Diagonalmatrix $\mathbf{D}$ besitzt, wobei die Diagonalelemente in $\mathbf{D}$ die Eigenwerte von $\mathbf{A}$ sind. Hier ist $\mathbf{X}$ eine orthogonale Matrix, die aus Eigenvektoren besteht. Da nun $\mathbf{A}$ zusätzlich positiv definit ist, sind alle Eigenwerte und damit Diagonalelemente größer als null. Für eine LR-Zerlegung benötigen wir aber keine orthogonale Matrix $\mathbf{X}$, sondern wir suchen eine untere Dreiecksmatrix $\mathbf{L}$:

Satz 4.10 (Cholesky-Zerlegung)

Jede symmetrische und positiv definite Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ besitzt die (eindeutige) Cholesky-Zerlegung

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{L}^\top,$$

wobei $\mathbf{L}$ eine untere Dreiecksmatrix und $\mathbf{D}$ eine Diagonalmatrix wie folgt ist:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ l_{2,1} & 1 & & 0 \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \dots & l_{n,n-1} & 1 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & d_n \end{bmatrix}$$

mit $d_k > 0$, $k = 1, \dots, n$.

Umgekehrt ist jede Matrix der Form $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{L}^\top$ (mit $\mathbf{L}$ und $\mathbf{D}$ wie zuvor) symmetrisch und positiv definit.

Beweis: Die Symmetrie und positive Definitheit von $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{L}^\top$ folgt sofort aus den entsprechenden Definitionen.

Wir beweisen die eigentliche Aussage, indem wir den Algorithmus angeben, mit dem die Cholesky-Zerlegung berechnet wird. Um sicherzustellen, dass wir dabei nicht durch null teilen, benötigen wir die Voraussetzungen in Verbindung mit dem Sylvester-Kriterium Lemma 4.2, nach dem alle Hauptabschnittsdeterminanten größer null sind.

Es soll nun

$$(\mathbf{L} \cdot \mathbf{D}) \cdot \mathbf{L}^\top = \mathbf{A}$$

gelten, d. h. mit ausmultiplizierter Matrix $\mathbf{L} \cdot \mathbf{D}$ und für $n = 4$:

$$\begin{bmatrix} d_1 & 0 & 0 & 0 \\ l_{2,1}d_1 & d_2 & 0 & 0 \\ l_{3,1}d_1 & l_{3,2}d_2 & d_3 & 0 \\ l_{4,1}d_1 & l_{4,2}d_2 & l_{4,3}d_3 & d_4 \end{bmatrix} \cdot \begin{bmatrix} 1 & l_{2,1} & l_{3,1} & l_{4,1} \\ 0 & 1 & l_{3,2} & l_{4,2} \\ 0 & 0 & 1 & l_{4,3} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{bmatrix}.$$

An dieser Darstellung sehen wir, dass wir mit der Cholesky-Zerlegung auch eine LR-Zerlegung finden.

Nun lesen wir sukzessive die gesuchten Einträge von $\mathbf{L}$ und $\mathbf{D}$ ab:

- Wir multiplizieren die erste Zeile von $\mathbf{L} \cdot \mathbf{D}$ mit der ersten Spalte von $\mathbf{L}^\top$ und erhalten

$$d_1 = a_{1,1}.$$

Entsprechend multiplizieren wir die Zeilen 2 bis n mit der ersten Spalte von $\mathbf{L}^\top$:

$$l_{k,1} = \frac{a_{k,1}}{d_1} \text{ für } 2 \leq k \leq n.$$

Wir dürfen hier durch $d_1 = a_{1,1}$ teilen, da die erste Hauptabschnittsdeterminante von $\mathbf{A}$ gleich $a_{1,1} > 0$ ist. Man beachte, dass wegen der Symmetrie von $\mathbf{A}$ jetzt auch die Produkte der ersten Zeile von $\mathbf{L} \cdot \mathbf{D}$ mit den Spalten 2 bis n von $\mathbf{L}^\top$ die richtigen Ergebnisse liefern, d. h., die berechneten Werte d_1 und $l_{k,1}$ führen zu keinen Widersprüchen. Das gilt entsprechend auch in den folgenden Schritten.

- Wir multiplizieren die zweite Zeile von $\mathbf{L} \cdot \mathbf{D}$ mit der zweiten Spalte von $\mathbf{L}^\top$ und erhalten

$$d_2 = a_{2,2} - l_{2,1}^2 d_1,$$

wobei die Werte der rechten Seite bereits bekannt sind. Wir überlegen uns, dass $d_2 \neq 0$ ist. Das Produkt der bereits berechneten 2×2 -Abschnittsmatrizen ergibt den entsprechenden Ausschnitt von $\mathbf{A}$:

$$\begin{bmatrix} d_1 & 0 \\ l_{2,1}d_1 & d_2 \end{bmatrix} \cdot \begin{bmatrix} 1 & l_{2,1} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}.$$

Falls $d_2 = 0$ wäre, hätte die Produktmatrix die Determinante 0. Das kann aber nicht sein, da die Hauptabschnittsdeterminanten größer null sind. Wir dürfen also im Folgenden durch d_2 dividieren.

Multiplikation der Zeilen 3 bis n mit der zweiten Spalte von $\mathbf{L}^\top$ führt zu

$$l_{k,2} = \frac{1}{d_2} (a_{k,2} - l_{k,1}d_1l_{2,1}) \text{ für } 3 \leq k \leq n.$$

- So verfährt man weiter und erhält im m -ten Schritt durch Multiplikation der m -ten Zeile von $\mathbf{L} \cdot \mathbf{D}$ mit der m -ten Spalte von $\mathbf{L}^\top$:

$$d_m = a_{m,m} - \sum_{k=1}^{m-1} l_{m,k}^2 d_k.$$

Mit dem gleichen Argument wie zuvor ist $d_m \neq 0$, da auch die m -te Hauptabschnittsdeterminante von $\mathbf{A}$ größer als null ist.

Für die Zeilen $m+1$ bis n ergibt sich wie zuvor

$$l_{k,m} = \frac{1}{d_m} \left(a_{k,m} - \sum_{r=1}^{m-1} l_{k,r} d_r l_{m,r} \right) \text{ für } m+1 \leq k \leq n.$$

Wir haben jetzt die gewünschte Darstellung, wobei $d_k \neq 0$, $1 \leq k \leq n$. Die Hauptabschnittsdeterminanten von $\mathbf{L} \cdot \mathbf{D} \cdot \mathbf{L}^\top$ sind positiv. Bei blockweiser Matrixmultiplikation ergibt sich die m -te Hauptabschnittsdeterminante als Produkt der m -ten Hauptab-

schnittsdeterminanten von $\mathbf{L}$, $\mathbf{D}$ und $\mathbf{L}^\top$, welche für $\mathbf{L}$ und $\mathbf{L}^\top$ gleich sind. Ihr Produkt ist also positiv. Damit muss auch die m -te Hauptabschnittsdeterminante von $\mathbf{D}$ positiv sein. Das gelingt für $1 \leq m \leq n$ nur, wenn alle Diagonalelemente positiv sind. $\square$

Der im Beweis verwendete Algorithmus lautet:
Berechne für $1 \leq m \leq n$ sukzessive:

$$\begin{aligned} d_m &= a_{m,m} - \sum_{k=1}^{m-1} l_{m,k}^2 d_k, \\ l_{k,m} &= \frac{1}{d_m} \left(a_{k,m} - \sum_{r=1}^{m-1} l_{k,r} d_r l_{m,r} \right) \text{ für } m+1 \leq k \leq n. \end{aligned}$$

Findet man bei der Durchführung des Algorithmus für eine symmetrische Matrix ein Diagonalelement $d_m \leq 0$, dann sind die Voraussetzungen der Cholesky-Zerlegung nicht erfüllt, d. h., die Matrix ist nicht positiv definit.

Beispiel 4.25

Wir berechnen eine Cholesky-Zerlegung für

$$\mathbf{A} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 8 & 0 \\ 1 & 0 & 16 \end{bmatrix}.$$

- Für $m = 1$ erhalten wir: $d_1 = 4$, $l_{2,1} = \frac{1}{4} \cdot 2 = \frac{1}{2}$, $l_{3,1} = \frac{1}{4}$.
- Für $m = 2$ ergibt sich: $d_2 = 8 - \left(\frac{1}{2}\right)^2 4 = 7$, $l_{3,2} = \frac{1}{7} \left(0 - \frac{1}{4} \cdot 4 \cdot \frac{1}{2}\right) = -\frac{1}{14}$.
- Schließlich erhalten wir für $m = 3$: $d_3 = 16 - \left(\frac{4}{4^2} + \frac{7}{14^2}\right) = 16 - \frac{4}{14} = \frac{110}{7}$.

Damit haben wir die Zerlegung berechnet:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{4} & -\frac{1}{14} & 1 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & \frac{110}{7} \end{bmatrix}, \quad \mathbf{L}^\top = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ 0 & 1 & -\frac{1}{14} \\ 0 & 0 & 1 \end{bmatrix}.$$

■

Folgerung 4.2 (Alternative Formulierung der Cholesky-Zerlegung)

Jede positiv definite, symmetrische Matrix $\mathbf{A}$ besitzt eine eindeutige Darstellung $\mathbf{A} = \mathbf{Q} \cdot \mathbf{Q}^\top$, wobei $\mathbf{Q}$ eine untere Dreiecksmatrix mit positiven Diagonalelementen ist.

Umgekehrt ist natürlich jede Matrix $\mathbf{A} = \mathbf{Q} \cdot \mathbf{Q}^\top$ symmetrisch und positiv definit.

Beweis: Wir wissen, dass $\mathbf{A}$ die eindeutige Darstellung $\mathbf{L}\mathbf{D}\mathbf{L}^\top$ aus Satz 4.10 besitzt. Sei $\sqrt{\mathbf{D}}$ die Diagonalmatrix mit den Diagonalelementen $\sqrt{d_1}, \dots, \sqrt{d_n}$, d. h. $\mathbf{D} = \sqrt{\mathbf{D}} \cdot \sqrt{\mathbf{D}}$.

Mit $\mathbf{Q} := \mathbf{L} \cdot \sqrt{\mathbf{D}}$ ergibt sich sofort die Darstellung $\mathbf{A} = \mathbf{Q} \cdot \mathbf{Q}^\top$. Diese ist auch eindeutig. Denn hat $\mathbf{A}$ eine solche Darstellung, dann kann man die m -te Spalte von $\mathbf{Q}$ durch das Diagonalelement dieser Spalte dividieren. Das lässt sich durch Multiplikation mit der Matrix der positiven Diagonalelemente rückgängig machen:

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \frac{q_{2,1}}{q_{1,1}} & 1 & & 0 \\ \vdots & \ddots & \ddots & 0 \\ \frac{q_{n,1}}{q_{1,1}} & \dots & \frac{q_{n,n-1}}{q_{n-1,n-1}} & 1 \end{bmatrix} \cdot \begin{bmatrix} q_{1,1} & & & \\ & q_{2,2} & & \\ & & \ddots & \\ & & & q_{n,n} \end{bmatrix} = \mathbf{L} \cdot \sqrt{\mathbf{D}}.$$

Da die Darstellung $\mathbf{A} = \mathbf{LDL}^\top$ eindeutig ist, ist somit auch $\mathbf{Q}$ eindeutig festgelegt. $\square$

Beispiel 4.26

Im vorangehenden Beispiel haben wir die Zerlegung

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{Q}^\top = (\mathbf{L}\sqrt{\mathbf{D}}) \cdot (\mathbf{L}\sqrt{\mathbf{D}})^\top = \begin{bmatrix} 2 & 0 & 0 \\ 1 & \sqrt{7} & 0 \\ \frac{1}{2} & -\frac{1}{2\sqrt{7}} & \sqrt{\frac{110}{7}} \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 & \frac{1}{2} \\ 0 & \sqrt{7} & -\frac{1}{2\sqrt{7}} \\ 0 & 0 & \sqrt{\frac{110}{7}} \end{bmatrix}$$

berechnet. $\blacksquare$

Zusammenfassend sind folgende Aussagen für eine reelle Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ äquivalent:

- $\mathbf{A}$ ist symmetrisch und positiv definit.
- Die reellen Eigenwerte der symmetrischen Matrix $\mathbf{A}$ sind alle positiv.
- Alle Hauptabschnittsdeterminanten der symmetrischen Matrix $\mathbf{A}$ sind positiv.
- $\mathbf{A} = \mathbf{Q} \cdot \mathbf{Q}^\top$, wobei $\mathbf{Q}$ eine reelle untere Dreiecksmatrix mit positiven Diagonalelementen ist.
- $\mathbf{A} = \mathbf{L} \cdot \mathbf{D} \cdot \mathbf{L}^\top$, wobei $\mathbf{L}$ eine reelle untere Dreiecksmatrix mit Einsen als Diagonalelementen und $\mathbf{D}$ eine Diagonalmatrix mit positiven Diagonalelementen ist.

4.3.2 Extrema unter Nebenbedingungen*

Man hat es häufig mit Minima und Maxima unter zusätzlichen Nebenbedingungen zu tun. Das gilt insbesondere für die klassische Mechanik. Dort werden die Nebenbedingungen **Zwangsbedingungen** genannt, die die Bewegungsfreiheit eines Körpers einschränken. Ein Beispiel ist der Einfluss der Seillänge bei einem Pendel.

Wir sehen uns die Problemstellung an einem noch einfacheren Beispiel an. Entlang eines Wegs durchwandern wir ein Gebirge. Uns interessieren jetzt nicht allgemein die

Täler und Gipfel des Gebirges, sondern die Stellen, an denen wir einen Tief- oder Hochpunkt auf unserem Weg erreicht haben.

Das Gebirge sei über eine Funktion $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ beschrieben. Der Weg ist in der x - y -Ebene durch die Punkte (x, y) gegeben, für die $g(x, y) = 0$ gilt. Dabei ist $g : D \rightarrow \mathbb{R}$ eine geeignete Funktion. Durch g werden also die x - und y -Koordinaten unserer Wanderung bestimmt, unsere Höhe ist dann $z = f(x, y)$. Uns interessieren jetzt die lokalen Maxima und Minima von f unter dieser Nebenbedingung $g(x, y) = 0$, da dies Gipfel und Täler auf unserem Weg sind. Zunächst betrachten wir den Spezialfall, dass $g(x, y) = 0$ sich nach y auflösen lässt zu $y = h(x)$, also lautet die Nebenbedingung $g(x, y) := y - h(x) = 0$, und der Wanderweg ist der Funktionsgraph von h . Dies ist tatsächlich nur ein Spezialfall, da z. B. der Einheitskreis, der über die Nebenbedingung $g(x, y) := x^2 + y^2 - 1 = 0$ beschrieben wird, so nicht auflösbar ist (vgl. Abschnitt 4.2.2). Bei auflösbarem g ist eine notwendige Bedingung für eine Extremstelle (x, y) laut Satz 4.7:

$$\frac{d}{dx}f(x, h(x)) = \text{grad } f(x, h(x)) \cdot \begin{pmatrix} 1 \\ h'(x) \end{pmatrix} = 0. \quad (4.11)$$

Beispiel 4.27

Wir betrachten $f(x, y) := x^3 + y^3 - 3x - 27y + 24$ und $g(x, y) := y - x$. Die Nebenbedingung $g(x, y) = 0 \iff y = x$ ist nach y auflösbar (mit $y = h(x) := x$) und beschreibt einen „Weg“, der auf die x - y -Ebene projiziert die Hauptdiagonale im kartesischen Koordinatensystem ist.

$$\frac{d}{dx}f(x, h(x)) = \frac{d}{dx}f(x, x) = \frac{d}{dx}[2x^3 - 30x + 24] = 6x^2 - 30.$$

Damit muss $x = \pm\sqrt{5}$ sein, und wegen der Nebenbedingung $y = x$ sind nur $(\sqrt{5}, \sqrt{5})$ und $(-\sqrt{5}, -\sqrt{5})$ Kandidaten für Extremstellen. ■

Wir betrachten nun einen anderen Ansatz, der auch dann noch funktioniert, wenn die Nebenbedingung später nicht auflösbar ist. Wir starten aber weiterhin mit der einfachen Nebenbedingung $g(x, y) = y - h(x)$. Der Gradient $\text{grad } f$ zeigt als Vektor in Richtung des steilsten Anstiegs von f (siehe Satz 4.2). Erreichen wir auf dem Wanderweg $g(x, y) = 0$ einen Gipfel oder ein Tal, so gibt es keinen Anstieg oder Abstieg in Wanderrichtung, d. h., $\text{grad } f$ muss senkrecht zur Wanderrichtung stehen. (Anders ausgedrückt: Die Richtungsableitung von f in Wanderrichtung ist $\vec{0}$.)

Am Punkt $(x, h(x))$ des Wanderwegs gehen wir in Richtung $(1, h'(x))$, denn $h'(x)$ ist die Steigung der Tangenten an den Funktionsgraphen von h an der Stelle x . Diese Richtung steht senkrecht zum Gradienten genau dann, wenn das Skalarprodukt beider Vektoren 0 ergibt (siehe Kapitel 3.1.2), also genau dann, wenn die Bedingung (4.11) erfüllt ist. Senkrecht zu $(1, h'(x))$ steht z. B. der Vektor $(-h'(x), 1) = \left(\frac{\partial g}{\partial x}(x, y), \frac{\partial g}{\partial y}(x, y) \right) =$

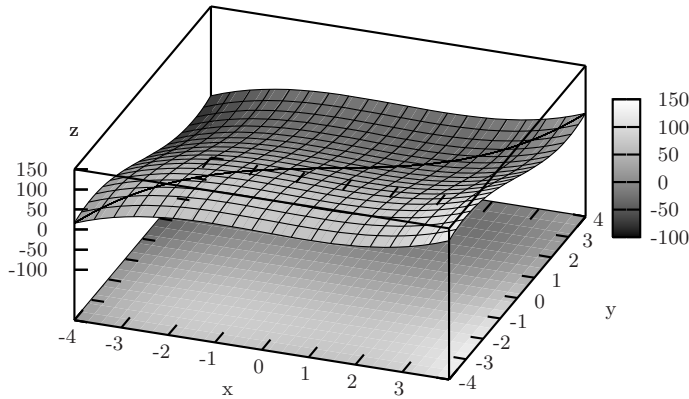


Abb. 4.8: $f(x, y) := x^3 + y^3 - 3x - 27y + 24$ mit „Wanderweg“ $g(x, y) = y - x = 0$

$\text{grad } g(x, y)$ (siehe Abbildung 4.9). $\text{grad } f$ muss sich an einer Extremstelle als Vielfaches dieses Vektors schreiben lassen. Damit erhalten wir als notwendige Bedingung: Es existiert ein $\lambda \in \mathbb{R}$, so dass

$$[\text{grad } f](x, y) + \lambda \text{grad } g(x, y) = (0, 0) \text{ bzw. } \text{grad}(f + \lambda g)(x, y) = \vec{0},$$

also

$$\begin{aligned} \frac{\partial f}{\partial x}(x, y) + \lambda \frac{\partial g}{\partial x}(x, y) &= 0, \\ \frac{\partial f}{\partial y}(x, y) + \lambda \frac{\partial g}{\partial y}(x, y) &= 0. \end{aligned}$$

λ nennt man in diesem Zusammenhang **Lagrange-Multiplikator**. Damit findet man als Lösungen dieses (eventuell nicht-linearen) Gleichungssystems eine Menge von Kandidaten für Extremstellen, die noch vom Parameter λ abhängen. Diese kann mittels der Nebenbedingung $g(x, y) = 0$ weiter eingeschränkt werden, indem über sie λ bestimmt wird.

Man beachte, dass zum Hinschreiben dieser notwendigen Bedingung nun kein Auflösen von $g(x, y) = 0$ nach y erforderlich ist.

Beispiel 4.28

Wir betrachten wie oben $f(x, y) := x^3 + y^3 - 3x - 27y + 24$ und $g(x, y) = y - x$ und erhalten das Gleichungssystem $\text{grad } f(x, y) + \lambda \text{grad } g(x, y) = (0, 0)$, das explizit lautet:

$$\begin{aligned} 3x^2 - 3 - \lambda &= 0 \\ 3y^2 - 27 + \lambda &= 0. \end{aligned}$$

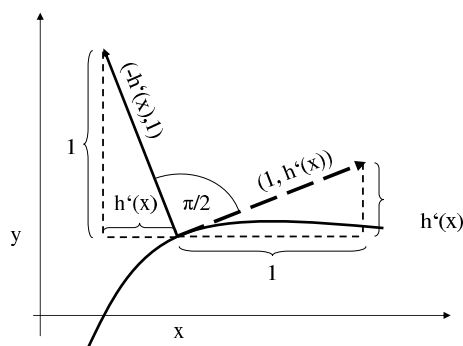


Abb. 4.9: Konstruktion eines Vektors senkrecht zur Tangenten an $h(x)$

Damit:

$$x = \pm \sqrt{1 + \frac{\lambda}{3}}, y = \pm \sqrt{9 - \frac{\lambda}{3}}.$$

Aus der Nebenbedingung $y = x$ folgt, dass das Vorzeichen von x und y gleich ist und

$$1 + \frac{\lambda}{3} = 9 - \frac{\lambda}{3} \implies \frac{2}{3}\lambda = 8 \implies \lambda = 12.$$

Kandidaten für Extrema sind auch bei diesem Ansatz $(-\sqrt{5}, -\sqrt{5})$ und $(\sqrt{5}, \sqrt{5})$. Durch Einsetzen sieht man, dass bei $(-\sqrt{5}, -\sqrt{5})$ ein lokales Maximum und bei $(\sqrt{5}, \sqrt{5})$ ein lokales Minimum liegt. ■

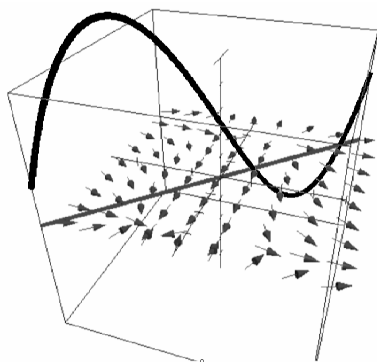


Abb. 4.10: Dort, wo $\text{grad } f(x, y)$ (Pfeile) senkrecht zum Weg $y = x$ ist, können Extrema vorliegen.

Tatsächlich funktioniert dieser Ansatz unter Verwendung impliziter Differenziation (vgl. Abschnitt 4.2.2) auch für Nebenbedingungen $g(x, y) = 0$, die nicht auflösbar sind, und man kann mehrere Nebenbedingungen gleichzeitig haben, die alle erfüllt sein sollen. Man hat dann m Funktionen $g_1, \dots, g_m : D \rightarrow \mathbb{R}$. Die Nebenbedingung lautet $g_1(\vec{x}) = 0$ und $g_2(\vec{x}) = 0 \dots$ und $g_m(\vec{x}) = 0$, also $\vec{g}(\vec{x}) = \vec{0}$ für $\vec{g} : D \rightarrow \mathbb{R}^m$. Daraus gewinnt man die notwendige Bedingung

$$\text{grad}[f + \lambda_1 g_1 + \lambda_2 g_2 + \dots + \lambda_m g_m](\vec{x}) = \vec{0}.$$

Genauer gilt der allgemeine Satz:

Satz 4.11 (Lagrange-Multiplikatoren)

Seien $m < n$, $D \subset \mathbb{R}^n$ eine offene Menge, $f : D \rightarrow \mathbb{R}$, $\vec{g} : D \rightarrow \mathbb{R}^m$. Weiter seien f und alle Komponentenfunktionen von $\vec{g}$ stetig differenzierbar auf D . Die Funktion $\vec{g}$ möge die Nebenbedingung ausdrücken, unter der wir ein lokales Extremum von f suchen, genauer suchen wir eine Extremstelle von f über der Menge $N := \{\vec{x} \in D : \vec{g}(\vec{x}) = \vec{0}\}$, d. h., wir suchen Stellen $\vec{x}_0 \in N$, so dass es ein $\delta > 0$ gibt mit $f(\vec{x}) \geq (\leq) f(\vec{x}_0)$ für alle $\vec{x} \in N$ mit $|\vec{x} - \vec{x}_0| < \delta$.

Sei $\vec{x}_0 \in N$, so dass der Rang der $(m \times n)$ -Matrix $\left[\frac{\partial g_l}{\partial x_k}(\vec{x}_0) \right]_{l=1, \dots, m; k=1, \dots, n}$ den Wert m hat (d. h., die m -Zeilen sind linear unabhängig).

Dann folgt aus a) als notwendige Bedingung die Aussage b):

- a) f hat ein lokales Extremum in $\vec{x}_0$ unter der Nebenbedingung N .
 b) Es existieren Zahlen (**Lagrange-Multiplikatoren**) $\lambda_1, \dots, \lambda_m \in \mathbb{R}$, so dass

$$\text{grad}[f + \lambda_1 g_1 + \lambda_2 g_2 + \dots + \lambda_m g_m](\vec{x}_0) = \vec{0}, \quad (4.12)$$

d. h., für alle $1 \leq k \leq n$ gilt:

$$\frac{\partial f}{\partial x_k}(\vec{x}_0) + \lambda_1 \frac{\partial g_1}{\partial x_k}(\vec{x}_0) + \dots + \lambda_m \frac{\partial g_m}{\partial x_k}(\vec{x}_0) = 0. \quad (4.13)$$

Bemerkungen:

- Sucht man lokale Extrema ohne Nebenbedingung, so kann man $g(\vec{x}) := 0$ setzen. Alle $\vec{x} \in D$ erfüllen damit die triviale Nebenbedingung $g(\vec{x}) = 0$. Die notwendige Bedingung (4.12) geht dann über in die bekannte Bedingung $0 = \text{grad}[f + \lambda g](\vec{x}_0) = \text{grad } f(\vec{x}_0)$.
- Hat man nur eine Nebenbedingung ($m = 1$), also eine Funktion $g : D \rightarrow \mathbb{R}$, so ist die Matrix $\left[\frac{\partial g}{\partial x_k}(\vec{x}_0) \right]_{k=1, \dots, n}$ der Gradient von g . Die zunächst kompliziert aussehende Rang-Bedingung lautet dann einfach $\text{grad } g(\vec{x}_0) \neq \vec{0}$.

Beispiel 4.29

Als Verpackung soll ein Quader mit Kantenlängen x, y und z gefunden werden, dessen Volumen 1 m^3 ist und dessen Oberfläche möglichst klein ist. Wir suchen das Minimum der Funktion

$$f(x, y, z) := 2xy + 2yz + 2xz,$$

die die Flächeninhalte aller sechs Seiten summiert, unter der Nebenbedingung

$$g(x, y, z) := xyz - 1 = 0,$$

die das Volumen festlegt. Damit erhalten wir das nicht-lineare Gleichungssystem

$$\begin{aligned} 2y + 2z + \lambda yz &= 0 \\ \wedge \quad 2x + 2z + \lambda xz &= 0 \\ \wedge \quad 2y + 2x + \lambda xy &= 0 \end{aligned}$$

zusammen mit der Nebenbedingung $xyz - 1 = 0$, wegen der $x \neq 0$, $y \neq 0$ und $z \neq 0$ ist. Für alle Punkte, die die Nebenbedingung erfüllen, ist damit $\text{grad } g(x, y, z) = (yz, xz, xy) \neq \vec{0}$. Die Rangbedingung aus Satz 4.11 ist also erfüllt. Multiplizieren wir die erste Gleichung mit x , die zweite mit y und die dritte mit z und nutzen wir die Nebenbedingung aus, so erhalten wir das Gleichungssystem

$$\begin{aligned} 2xy + 2xz &= -\lambda \\ \wedge \quad 2xy + 2yz &= -\lambda \\ \wedge \quad 2xz + 2yz &= -\lambda. \end{aligned}$$

Um die Werte für xy , xz und yz zu erhalten, können wir es als lineares Gleichungssystem lösen:

$$\begin{aligned} 2xy + 2xz &= -\lambda & 2xy + 2yz &= -\lambda \\ \wedge \quad -2xz + 2yz &= 0 & \implies \wedge \quad -2xz + 2yz &= 0 \\ \wedge \quad 2xz + 2yz &= -\lambda & \wedge \quad +4yz &= -\lambda \\ \implies \wedge \quad 2xy &= -\frac{\lambda}{2} & \wedge \quad -2xz &= \frac{\lambda}{2} \\ \wedge \quad +4yz &= -\lambda. \end{aligned}$$

Also $xy = xz = yz = -\frac{\lambda}{4}$. Mit der Nebenbedingung erhalten wir daraus: $-\frac{\lambda}{4}z = -\frac{\lambda}{4}y = -\frac{\lambda}{4}x = xyz = 1$, also $x = y = z = -\frac{4}{\lambda}$. Aus der Nebenbedingung erhalten wir zudem den Wert für λ : $-\frac{4^3}{\lambda^3} = 1 \implies \lambda = -4$. Nur für $x = y = z = 1$ kann ein Extremum vorliegen – und die Oberfläche ist tatsächlich minimal für den Würfel mit Kantenlänge 1 m. ■

Beispiel 4.30

Wir bestimmen Kandidaten für alle Extremalstellen der Funktion

$$f(x, y, z) = x^2 + y^2 + z$$

unter den Nebenbedingungen

$$g_1(x, y, z) = x + y - z - 1 = 0 \text{ und } g_2(x, y, z) = x^2 + y^2 - \frac{1}{8} = 0.$$

Mit den Lagrange-Multiplikatoren λ_1, λ_2 erhält man das Gleichungssystem

$$\begin{aligned} 2x + \lambda_1 + 2\lambda_2 x &= 0, \\ 2y + \lambda_1 + 2\lambda_2 y &= 0, \\ 1 - \lambda_1 &= 0. \end{aligned}$$

Hinzu kommen die beiden Nebenbedingungen. Aus der dritten Gleichung liest man $\lambda_1 = 1$ ab. Mit $\lambda_1 = 1$ erhält man aus den ersten beiden Gleichungen

$$2x(1 + \lambda_2) + 1 = 0 \quad \text{und} \quad 2y(1 + \lambda_2) + 1 = 0.$$

Damit ist $\lambda_2 = -1$ ausgeschlossen, da die Gleichungen für diesen Wert zu $1 = 0$ werden. Für $\lambda \neq -1$ folgt sofort $x = y = -\frac{1}{2(1+\lambda_2)}$. Insgesamt haben wir also:

$$\lambda_1 = 1, \quad \lambda_2 \neq -1 \quad \text{und} \quad x = y.$$

Setzen wir $x = y$ in die Nebenbedingungen g_1, g_2 ein, so erhalten wir $2x - z - 1 = 0$ und $2x^2 - \frac{1}{8} = 0$ bzw.

$$x_{1,2} = \pm \frac{1}{4} \quad \text{und das zugehörige} \quad z_{1,2} = 2x_{1,2} - 1.$$

Damit können Extrema unter der Nebenbedingung nur an den beiden Stellen $\vec{x}_1 = (\frac{1}{4}, \frac{1}{4}, -\frac{1}{2})$ und $\vec{x}_2 = (-\frac{1}{4}, -\frac{1}{4}, -\frac{3}{2})$ liegen. In der Tat handelt es sich bei $\vec{x}_1$ mit $f(\vec{x}_1) = -\frac{3}{8}$ um ein lokales Maximum und bei $\vec{x}_2$ mit $f(\vec{x}_2) = -\frac{11}{8}$ um ein lokales Minimum unter Nebenbedingung. ■

Beispiel 4.31 (Mehrdimensionales Newton-Verfahren)

Bei der Nullstellensuche mit dem eindimensionalen Newton-Verfahren (Seite 306, vgl. Beispiel 4.11 auf Seite 605) 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 . Die Nullstelle x_{k+1} ist ein Minimum von $[f(x_k) + f'(x_k)(x - x_k)]^2$. Entsprechend kann man mit einer mehrdimensionalen Taylor-Entwicklung für $f : \mathbb{R}^n \rightarrow \mathbb{R}$ die Nullstellensuche von f bzw. allgemeiner die Suche nach einem lokalen Minimum der Funktion f^2 gestalten: Statt f^2 minimiere $[f(\vec{x}_k) + \text{grad } f(\vec{x}_k) \cdot (\vec{x} - \vec{x}_k)]^2$. Diese quadrierte Taylor-Entwicklung von f 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. Im Inneren der Umgebung ist die notwendige Bedingung $\text{grad}([f(\vec{x}_k) + \text{grad } f(\vec{x}_k)(\vec{x} - \vec{x}_k)]^2) = \vec{0}$. Man beachte, dass es im Inneren der Umgebung aber kein lokales Minimum geben muss. Daher wird der Rand in die Suche einbezogen. Auf dem Rand lässt sich eine notwendige Bedingung für ein lokales Minimum mit einem Lagrange-Multiplikator λ über die Nebenbedingung

$0 = g(\vec{x}) = |\vec{x} - \vec{x}_k|^2 - [\delta(\vec{x}_k)]^2$ formulieren. Ein lokales Minimum auf dem Rand im Punkt $\vec{x}$ kann nur vorliegen, falls ein $\lambda \in \mathbb{R}$ existiert, so dass

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

gilt. Für $\lambda = 0$ ist der Fall eines Minimums im Inneren der Umgebung eingeschlossen. Der resultierende iterative Algorithmus, der ein auf einer Umgebung gefundenes Minimum (der Annäherung an f^2) als Ausgangspunkt für die nächste lokale Minimumsuche nimmt, heißt **Levenberg-Marquardt-Verfahren**. Es sucht ein lokales Minimum von f^2 und damit gegebenenfalls eine Nullstelle von f . ■

4.3.3 Lineare Optimierung*

Im vorangehenden Unterkapitel haben wir den Satz über die Lagrange-Multiplikatoren benutzt, um Kandidaten für Extremstellen unter Nebenbedingungen zu erhalten. In diesem Abschnitt betrachten wir eine ähnliche Aufgabenstellung, bei der sowohl die Funktion, für die wir ein Maximum suchen, als auch die Funktionen, über die die Nebenbedingungen formuliert sind, linear sind. Für diesen wichtigen Spezialfall **lineare Optimierung** (auch als **lineare Programmierung** bezeichnet) gibt es leistungsfähige Lösungsalgorithmen, vgl. z. B. (Hochstättler, 2010, Kapitel 8), mit denen nicht nur viele betriebswirtschaftliche Aufgaben gelöst werden.

Ausgehend von Leonid Witaljewitsch Kantorowitsch (1912–1986, lineare Optimierung 1939, Wirtschaftsnobelpreis 1975) wurde die lineare Optimierung von Georg Dantzig weiterentwickelt. Er hat 1947 die Grundzüge des Simplex-Verfahrens vorgestellt, das wir weiter unten erklären. Es ist bis heute das meistgenutzte Verfahren zur linearen Optimierung.

Gegeben seien $\vec{p} \in \mathbb{R}^n$, $\vec{s} \in \mathbb{R}^m$ mit nicht-negativen Komponenten $s_k \geq 0$ und $\mathbf{A} \in \mathbb{R}^{m \times n}$. Gesucht sind Zahlen $x_1, \dots, x_n \in [0, \infty[$, für die die lineare **Zielfunktion** („objective function“)

$$p_1 x_1 + p_2 x_2 + \dots + p_n x_n = \vec{p} \cdot \vec{x}$$

maximal wird, wobei gleichzeitig die Nebenbedingungen

$$(\mathbf{A}\vec{x})_k \leq s_k, \quad 1 \leq k \leq m,$$

für die m Komponenten des Vektors $\mathbf{A}\vec{x}$ erfüllt sein sollen. Es ist üblich, die Nebenbedingungen mit der Abkürzung s. t. (für „subject to“) anzukündigen.

Als Beispiel betrachten wir die (lineare) Funktion $2x_1 + 3x_2$, die für $x_1, x_2 \geq 0$ unter den (linearen) Nebenbedingungen (s. t.)

$$\begin{aligned} x_1 + 2x_2 &\leq 4 \text{ und} \\ 3x_1 + x_2 &\leq 6 \end{aligned}$$

zu maximieren sei. Hier ist

$$\vec{p} = (2, 3), \quad \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix} \quad \text{und} \quad \vec{s} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}.$$

Anschaulich kann man das Problem auch so formulieren: Es werden zwei Produkte P_1 und P_2 verkauft. Der Gewinn des ersten Produkts ist zwei Geldeinheiten pro Mengeneinheit, der des zweiten drei Geldeinheiten. Bezeichnen wir mit x_1 die verkaufte Menge von P_1 und mit x_2 die verkaufte Menge von P_2 , dann ist der zu maximierende Gewinn $2x_1 + 3x_2$. Die Produkte können aber nicht in beliebigen Stückzahlen hergestellt werden. P_2 möge den doppelten Lagerplatz wie P_1 benötigen. Die Lagerkapazität wird mit $x_1 + 2x_2 \leq 4$ angegeben. Ähnlich kann es eine Beschränkung $3x_1 + x_2 \leq 6$ hinsichtlich der zur Verfügung stehenden Rohstoffe geben.

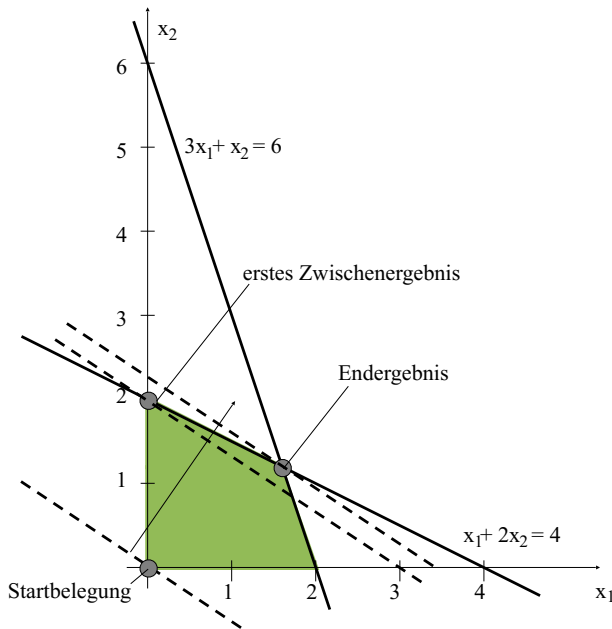


Abb. 4.11: Die gestrichelten Geraden haben die Gleichung $2x_1 + 3x_2 = z$, wobei der größte Wert z gesucht ist, für den es eine solche Gerade mit mindestens einem Punkt im grünen, durch die Nebenbedingungen berandeten Gebiet gibt. Der Wert z wird umso größer, je weiter man eine gestrichelte Gerade nach rechts oben schiebt. Der größte Wert unter den Nebenbedingungen wird in dem Eckpunkt angenommen, der als Endergebnis markiert ist. Außerdem sind der Startwert und ein Zwischenergebnis des Simplex-Algorithmus eingezeichnet.

Die Zielfunktion und die Funktionen auf den linken Seiten der Nebenbedingungen sind linear. Das ist prinzipiell eine viel einfachere Situation als die, die wir mit Lagrange-Multiplikatoren lösen. Tatsächlich besteht in Anwendungen die Schwierigkeit aber in einer sehr großen Anzahl n von Variablen und m von Nebenbedingungen.

Wir haben das Optimierungsproblem in der **Standardform** eingeführt. Viele Optimierungsaufgaben mit linearen Zielfunktionen und Nebenbedingungen lassen sich in diese Form überführen:

- Aus einem Minimierungsproblem wird durch Multiplikation der Zielfunktion mit -1 ein Maximierungsproblem (und umgekehrt).

- Wegen $x_1, \dots, x_n \in [0, \infty[$ sind alle Variablen nicht-negativ. Möchte man z. B. für x_1 auch negative Zahlen zulassen, so kann man eine weitere Variable $x_{n+1} \geq 0$ einführen und ersetzt x_1 überall durch $x_1 - x_{n+1}$.
- Der Vektor $\vec{s}$ hat in der Standardform nur nicht-negative Komponenten. Das ist für den Simplex-Algorithmus wichtig. Ist eine Komponente im gegebenen Optimierungsproblem negativ, so erfüllt der Nullvektor die Nebenbedingungen nicht. Diesen benötigen wir aber für den Simplex-Algorithmus als Startpunkt. Daher wird in einer Vorverarbeitung mit ähnlichen Gauß-Umformungen wie beim eigentlichen Simplex-Verfahren (s. u.) ein anderer Startpunkt gesucht und gleichzeitig $\vec{s}$ in einen Vektor mit nicht-negativen Komponenten überführt (falls dies möglich ist). Dieses Verfahren, auf das wir hier nicht weiter eingehen, heißt **dualer Simplex-Algorithmus** und ist dem Simplex-Algorithmus sehr ähnlich. Lässt sich das Problem negativer Komponenten in $\vec{s}$ damit umgehen, dann lassen sich auch mit „ $\geq$ “ formulierte Nebenbedingungen durch Multiplikation mit -1 in Bedingungen mit „ $\leq$ “ überführen. So sind auch Nebenbedingungen mit unteren und oberen Schranken („ $\geq$ “ und „ $\leq$ “) sowie Gleichungen möglich und in die Standardform überführbar.

Die Nebenbedingungen beschreiben jeweils alle Punkte im $\mathbb{R}^n$, die auf einer Seite einer Hyperebene liegen. Im Beispiel sind das die Punkte unterhalb der beiden Geraden $x_1 + 2x_2 = 4$ und $3x_1 + x_2 = 6$. Da wir uns auf den nicht-negativen Bereich beschränken, kommen weitere Hyperebenen hinzu, deren Punkte in jeweils einer gemeinsamen Komponente null sind. Diese Hyperebenen beschreiben zusammen den Rand des Bereichs der Punkte mit nicht-negativen Komponenten. Im Beispiel sind das die x_1 - und die x_2 -Achse. In Abbildung 4.11 ist die so entstehende Menge zulässiger Punkte grau unterlegt.

Die Nebenbedingungen mögen so gewählt sein, dass ein Maximum auf einer beschränkten konvexen Menge gesucht wird. Dadurch wird sichergestellt, dass es tatsächlich ein Maximum gibt. Wenn beispielsweise nicht jede Variable in den Nebenbedingungen vorkommt, so können die nicht vorkommenden Variablen beliebig große Werte annehmen und damit (bei positiven Koeffizienten $p_1, \dots, p_n$) die lineare Zielfunktion, die ohne Nebenbedingungen keine Extrema besitzt, ebenfalls beliebig groß werden lassen. Das müssen wir ausschließen. Damit gibt es dann insbesondere keine Nullspalten in **A**.

Ist die Menge zulässiger Punkte beschränkt, dann liegt ein Extremum (mindestens) in einem Eckpunkt, in dem sich n dieser Hyperebenen schneiden. Das liegt daran, dass auch $\vec{p} \cdot \vec{x} = c$ eine Hyperebene beschreibt, wobei durch Parallelverschiebung c größer oder kleiner wird (entsprechend dem Abstand zum Nullpunkt bei der Hesse'schen Normalform einer Ebene, siehe Seite 463). Ein Extremum für c wird also angenommen, wenn eine weitere Verschiebung die Nebenbedingungen verletzen würde – und das passiert nur am Rand und insbesondere in Eckpunkten. Unendlich viele Lösungen kann

es geben, wenn die Zielfunktion Hyperebenen beschreibt, die parallel zur Hyperebene einer Nebenbedingung liegen.

Mit Ungleichungen lässt sich nicht gut rechnen. Auch beim Satz über die Lagrange-Multiplikatoren sind die Nebenbedingungen Gleichungen. Daher werden die Ungleichungen zu Gleichungen umgeformt. Dies geschieht über zusätzliche Variablen. Mit diesen sogenannten **Schlupfvariablen** $y_1 \geq 0, \dots, y_m \geq 0$, die den Abstand zu den Werten s_k angeben, lässt sich das Optimierungsproblem so umformulieren:

Gesucht sind Zahlen $x_1, \dots, x_n, y_1, \dots, y_m \in [0, \infty[$, für die $p_1 x_1 + p_2 x_2 + \dots + p_n x_n = \vec{p} \cdot \vec{x}$ maximal wird, wobei gleichzeitig die folgenden m Nebenbedingungen erfüllt sind:

$$(\mathbf{A}\vec{x})_k + y_k = s_k, \quad 1 \leq k \leq m.$$

Mit $\vec{v} = (x_1, \dots, x_n, y_1, \dots, y_m)^\top$ und $\mathbf{B} \in \mathbb{R}^{m \times (n+m)}$, wobei die ersten n Spalten von $\mathbf{B}$ genau die Matrix $\mathbf{A}$ bilden und die restlichen m Spalten eine $m \times m$ -Einheitsmatrix $\mathbf{E}_m$ darstellen (der Rang der Matrix $\mathbf{B}$ ist m), lassen sich die Nebenbedingungen umschreiben zu

$$\mathbf{B}\vec{v} = \vec{s}.$$

Im Beispiel ist also

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}.$$

Zur Abgrenzung gegen die Schlupfvariablen nennt man die ursprünglichen Variablen $x_1, \dots, x_n$ **Strukturvariablen**.

Ein Maximum tritt in einer Ecke auf, also in einem Punkt, in dem sich n Hyperebenen des $\mathbb{R}^n$ (im Beispiel sind das $n = 2$ Geraden im $\mathbb{R}^2$) schneiden, wobei die Hyperebenen durch (a) die m Nebenbedingungen sowie durch (b) die n Bedingungen $x_1 \geq 0, \dots, x_n \geq 0$ gegeben sind. Ein Punkt liegt genau dann auf einer dieser Hyperebenen, wenn im Fall (a) die entsprechende Schlupfvariable null ist und im Fall (b) die entsprechende Strukturvariable null ist. Liegt also ein Punkt $(x_1, \dots, x_n)$ gleichzeitig auf n Hyperebenen, dann gibt es eine Lösung $\vec{v} = (x_1, \dots, x_n, y_1, \dots, y_m)^\top$ von $\mathbf{B}\vec{v} = \vec{s}$, bei der mindestens n Komponenten null sind. Wir müssen uns also nur Lösungen anschauen, bei denen n Variablen null sind. Diese Lösungen heißen **Basislösungen** und die von null verschiedenen Variablen **Basisvariablen**. Die Begriffe werden üblicherweise ohne die Einschränkung verwendet, dass die Variablenwerte nicht-negativ sein sollen. Man spricht dagegen von einer **zulässigen Basislösung**, wenn zusätzlich alle Variablenwerte nicht-negativ sind. Generell heißt eine nicht-negative Variablenbelegung eine **zulässige Lösung** („feasible solution“), falls sie alle Nebenbedingungen erfüllt.

Jetzt wäre es aber bei großem m sehr zeitaufwändig, die maximal $\binom{n+m}{m}$ verschiedenen Gleichungssysteme alle zu lösen (Kombinationen von m Spalten aus einer Menge von $n + m$ Spalten). Der **Simplex-Algorithmus** geht hier geschickter vor und ist häufig viel schneller, hat aber im schlechtesten Fall dennoch eine exponentielle und damit ebenfalls schlechte Laufzeit bezogen auf die Anzahl der Variablen. Er beginnt mit der Basislösung $x_1 = \dots = x_n = 0$, $y_1 = s_1, \dots, y_m = s_m$, d. h., die letzten m Spalten von $\mathbf{B}$ gelten als ausgewählt, die zugehörigen Schlupfvariablen sind zunächst die Basisvariablen. Es handelt sich hier sogar um eine zulässige Basislösung, da wir die bei den Nebenbedingungen verwendeten Obergrenzen, also die Komponenten von $\vec{s}$, als nicht-negativ vorausgesetzt haben. Ausgehend von dieser initialen zulässigen Basislösung werden die ausgewählten Spalten (also die Basisvariablen) durch paarweisen Tausch so lange zu anderen zulässigen Basislösungen variiert, bis ein Maximum gefunden ist. In jedem Schritt wird dabei eine der m ausgewählten Spalten (Basisvariablen) durch eine nicht ausgewählte so ersetzt, dass dadurch der Wert der Zielfunktion um einen möglichst großen, aber in jedem Fall nicht-negativen Wert wächst. Der Austausch geschieht mittels Gauß-Umformungen unter Verwendung eines Pivot-Elements. Darauf gehen wir weiter unten genauer ein. Man hat also stets eine „aktive Menge“ von m ausgewählten Variablen (die restlichen sind null), deren Zusammensetzung schrittweise verändert wird. Falls jede nicht berücksichtigte Variable bei Aufnahme in die „aktive Menge“ eine echte Verkleinerung des zu maximierenden Werts bewirken würde, ist das Verfahren beendet. Der Algorithmus ist so aufgebaut, dass in jedem Schritt direkt abgelesen werden kann, ob ein Maximum erreicht ist oder welche Variablen auszutauschen sind. Außerdem wird automatisch der Wert des Maximums berechnet.

Wir sehen uns den Simplex-Algorithmus in unserem Beispiel an. Dabei erklären wir, warum die einzelnen Verarbeitungsschritte tatsächlich zu einer Lösung führen. Für den Algorithmus verwenden wir ein Schema, in dessen Zeilen wir die mit den Schlupfvariablen formulierten Nebenbedingungen eintragen. Zudem bezeichnen wir mit $z := 2x_1 + 3x_2$ den Wert der zu maximierenden Zielfunktion, so dass wir durch Umstellen als weitere Gleichung $1z - 2x_1 - 3x_2 = 0$ erhalten. Die Koeffizienten 1, -2 und -3 schreiben wir in die letzte Zeile des Schemas, ebenso die Null dort auf die rechte Seite. Häufig findet man auch eine Variante, bei der dies zur ersten Zeile wird. Da in dieser Gleichung keine Schlupfvariablen vorkommen, sind dort die entsprechenden Koeffizienten für y_1 und y_2 null. In der letzten Zeile treten somit die mit -1 multiplizierten Koeffizienten der zu maximierenden Zielfunktion auf.

z	x_1	x_2	y_1	y_2	
0	1	2	1	0	4
0	3	1	0	1	6
1	-2	-3	0	0	0

Dieses Schema entspricht genau dem linearen Gleichungssystem

$$\left[\begin{array}{ccccc|c} 0 & 1 & 2 & 1 & 0 & 4 \\ 0 & 3 & 1 & 0 & 1 & 6 \\ 1 & -2 & -3 & 0 & 0 & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{ccccc|c} 0 & 1 & 2 & 1 & 0 & \\ 0 & 3 & 1 & 0 & 1 & \\ 1 & -2 & -3 & 0 & 0 & \end{array} \right] \begin{pmatrix} z \\ x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \\ 0 \end{pmatrix}.$$

Auf dieses Gleichungssystem wenden wir die üblichen Gauß'schen Zeilenumformungen an, die die Lösungsmenge nicht verändern. Allerdings suchen wir nicht alle Lösungen, sondern nur eine optimale zulässige, bei der n Variablen null sind. Das geschieht dadurch, dass wir sukzessive Basisvariablen auswählen, indem wir in den zugehörigen Spalten des Schemas verschiedene Standard-Einheitsvektoren erzeugen. Das ist beim Start bereits gegeben, da initial die Schlupfvariablen ausgewählt sind, also y_1 und y_2 . Zu allen Werten der nicht ausgewählten Variablen gibt es damit eine Lösung des linearen Gleichungssystems. Das gilt also insbesondere, wenn wir alle nicht ausgewählten Variablen mit null belegen, wie wir es für die Basislösungen benötigen. Allerdings können die Variablenwerte der Lösung durchaus negativ und damit nicht zulässig sein. Deshalb ist eine besondere Strategie für den Abauf der Gauß-Schritte erforderlich, bei der man von einer zulässigen Lösung zur nächsten gelangt. Das entspricht nicht dem üblichen Vorgehen beim Lösen von Gleichungssystemen, bei dem sukzessive eine Dreiecks- oder eine Diagonalmatrix erzeugt wird und mit dem man hier gegebenenfalls unendlich viele Lösungen finden würde.

Im Beispiel sind zunächst $x_1 = x_2 = 0$ gewählt. Das ist der Startpunkt in Abbildung 4.11 und eine Lösung des Problems, die durch den Algorithmus iterativ verbessert wird. Aus den ersten beiden Gleichungen des Schemas ergeben sich jetzt direkt $y_1 = 4$ und $y_2 = 6$, die Startlösung ist also auch zulässig (da die Komponenten von $\vec{s}$ nicht-negativ sind).

Der zur aktuellen Auswahl der Basisvariablen gehörende Wert z der Zielfunktion ist stets unten rechts im Schema ablesbar. Denn die Spalten der Basisvariablen besitzen in der letzten Zeile eine Null. Bei Belegung der nicht ausgewählten Variablen mit null wird aus der letzten Zeile die Gleichung „ z gleich Wert unten rechts“. Das ist ein wesentlicher Trick des Algorithmus. Die Eigenschaft ist im Startzustand sofort verifiziert: Sind alle Strukturvariablen null, so bleibt nur $z = 0$.

Außerdem stehen in der letzten Spalte die Lösungswerte der ausgewählten Basisvariablen, die zum Funktionswert unten rechts führen.

Jetzt suchen wir den betragsmäßig größten negativen Wert (als den kleinsten Wert) der letzten Zeile in den Spalten der nicht ausgewählten Variablen. Dieser ist hier -3 . Der aktuelle Funktionswert $z = 0$ lässt sich zu $z = 3x_2$ vergrößern, indem wir die Variable $x_2 \geq 0$, die zu -3 gehört, auswählen. Allerdings müssen wir noch herausfinden,

ob wir y_1 oder y_2 gegen x_2 austauschen. Dazu berechnen wir die Quotienten der Werte der rechten Spalte durch die Werte der Spalte zu x_2 , falls letztere größer null sind.

Man beachte, dass es Einträge größer null geben muss. Denn anderenfalls wären alle Werte der Spalte zu x_2 kleiner oder gleich null, und x_2 kann beliebig groß gewählt werden – im Widerspruch zur Annahme, dass wir ein Maximum auf einer beschränkten Menge suchen.

Wir wählen die Zeile zum kleinsten zuvor berechneten Quotienten aus. Wir erzeugen nun mittels Zeilenumformungen in der Spalte zu x_2 einen Einheitsvektor mit einer 1 in der gewählten Zeile. Damit tauschen wir gegen die ausgewählte Variable, zu der zuvor dieser Einheitsvektor gehörte.

Im Beispiel sind die Quotienten $4/2 = 2$ und $6/1 = 6$. Damit wird die erste Zeile ausgewählt. Das Element dieser Zeile in der Spalte x_2 heißt **Pivot**-Element und ist hier 2. Gauß'sche Zeilenumformungen liefern:

z	x_1	x_2	y_1	y_2	
0	$\frac{1}{2}$	1	$\frac{1}{2}$	0	2
0	$\frac{5}{2}$	0	$-\frac{1}{2}$	1	4
1	$-\frac{1}{2}$	0	$\frac{3}{2}$	0	6

Jetzt sind die Variablen x_2 und y_2 ausgewählt. In ihren Spalten stehen Standard-Einheitsvektoren. Für die jetzigen Werte $x_1 = y_1 = 0$ (da diese Variablen nicht ausgewählt sind), $x_2 = 2$ und $y_2 = 4$ nimmt die Zielfunktion den Wert 6 an (siehe erstes Zwischenergebnis in Abbildung 4.11). Das kann man (wie oben beschrieben) direkt an der letzten Zeile rechts außen ablesen:

$$1 \cdot z - \frac{1}{2} \cdot 0 + 0 \cdot x_2 + \frac{3}{2} \cdot 0 + 0 \cdot y_2 = 6.$$

Die Zeilenauswahl haben wir so über den kleinsten Quotienten vorgenommen, dass bei den anschließenden Zeilenumformungen in der rechten Spalte keine negativen Werte entstehen. Die Bedingung an $\vec{s}$, dass alle Komponenten nicht-negativ sind, sorgt initial für eine nicht-negative rechte Spalte. Wir erzeugen dann in der Spalte zu x_2 Nullen, indem wir zunächst die ausgewählte Zeile so normieren, dass in der betreffenden Spalte eine Eins steht. In der rechten Spalte steht dann besagter kleinster Quotient, den wir jetzt α nennen. Nun können wir Vielfache der Zeile von den anderen subtrahieren:

- Hat eine Zeile in der Spalte zu x_2 einen Nulleintrag, dann muss sie nicht mit Gauß-Umformungen geändert werden.
- Hat eine Zeile in der Spalte zu x_2 einen negativen Eintrag, so können wir ein positives Vielfaches der ausgewählten Zeile addieren, um eine Null zu erzeugen. Insbesondere addieren wir zum Wert der rechten Spalte etwas Positives.

- Hat eine Zeile in der Spalte zu x_2 einen positiven Eintrag β , so subtrahieren wir die bereits zuvor normierte ausgewählte Zeile (mit Eintrag eins in Spalte x_2) mal β . Vom Eintrag γ in der rechten Spalte subtrahieren wir $\beta \cdot \alpha$ und erhalten $\gamma - \beta \cdot \alpha$. Nach Auswahl der Zeile ist $\gamma/\beta \geq \alpha$, also $\gamma/\beta - \alpha \geq 0$. Multiplizieren wir jetzt mit $\beta > 0$, so sehen wir, dass in der rechten Spalte tatsächlich ein nicht-negativer Wert steht.

Damit werden dann auch die Variablenwerte nicht-negativ, schließlich entsprechen die Werte der ausgewählten Variablen wegen der Einheitsvektoren in ihren zugehörigen Spalten direkt den Einträgen der rechten Spalte. So bleibt diese Randbedingung des Optimierungsproblems eingehalten, und wir hangeln uns von einer zulässigen Basislösung zur nächsten.

Die Variablenwahl über den betragsmäßig größten negativen Eintrag δ der letzten Zeile (in den Spalten der nicht ausgewählten Variablen) führt dazu, dass durch die anschließende Gauß-Umformung unten rechts ein (hoffentlich großer) nicht-negativer Wert $|\delta|\alpha \geq 0$ addiert wird, wobei allerdings $\alpha \geq 0$ als kleinster Quotient bestimmt wurde. Die Wahl von δ ist die Strategie (Heuristik), möglichst schnell in die Nähe des Maximums zu kommen. Bei jedem Schritt des Verfahrens werden insbesondere immer nur bessere oder gleich gute Lösungen gefunden.

Jetzt wiederholt sich das Vorgehen, bis alle Einträge zu nicht ausgewählten Variablen in der letzten Zeile nicht-negativ sind. Ist dieses Abbruchkriterium erreicht, dann ist zumindest ein lokales Maximum gefunden. Denn jeder weitere Variablentausch würde dazu führen, dass vom Maximumkandidaten ein nicht-negativer Wert $|\delta|\alpha$ abgezogen wird. Tatsächlich verfängt sich der Algorithmus aber nicht in einem Suboptimum. Da nur Gauß-Operationen durchgeführt werden, gilt auch beim Erreichen der Abbruchbedingung – unabhängig von der Belegung nicht ausgewählter Variablen mit Nullen – die letzte Gleichung. Für jede zulässige Belegung der Struktur- und Schlupfvariablen liefert sie den Wert z der Zielfunktion. Dabei ist z gleich dem Wert unten rechts minus der nicht-negativen Summe aus nicht-negativen Faktoren (Abbruchbedingung) mal nicht-negativen Variablenwerten (zulässige Lösung). Damit kann z nicht mehr größer werden, ein globales Maximum ist gefunden. Damit sieht der Algorithmus so aus:

procedure SIMPLEX(Tableau)

while Es existieren negative Einträge in der letzten Zeile des Tableaus **do**

 Sei k die Spaltennummer des kleinsten Eintrags der letzten Zeile.

 Berechne Quotienten der Einträge der rechten Spalte durch
 positive Einträge der Spalte s .

 Sei l die Zeilennummer zum kleinsten Quotienten.

 Erzeuge mittels Gauß-Umformungen einen Einheitsvektor
 in Spalte k mit einer Eins in Zeile l .

return Maximalwert = Wert unten rechts.

Eine Schwierigkeit bei der iterativen Durchführung des Verfahrens ist noch, dass prinzipiell Endlosschleifen entstehen können, bei denen sich eine Variablenauswahl immer wiederholt. Das kann geschehen, wenn der rechts unten stehende aktuelle Wert von z z. B. bei $\alpha = 0$ gleich bleibt und sich nicht erhöht. In diesem Fall liegen sogenannte **degenerierte Basislösungen** vor. Daher muss man bei der Auswahl solche Variablen ausschließen, die zu einer bereits zuvor betrachteten Variablenauswahl führen würden.

Im nächsten Schritt des Beispiels wird x_1 ausgewählt, da $-1/2$ der betragsmäßig größte negative Wert der letzten Zeile zu den nicht ausgewählten Variablen ist und wir die Basisvariablen-Auswahl x_1, x_2 noch nicht betrachtet haben. Die Quotienten der Zeilen sind damit 4 und $8/5$, so dass die zweite Zeile passend zum kleinsten Quotienten gewählt wird. Das Pivot-Element ist jetzt $5/2$:

z	x_1	x_2	y_1	y_2	
0	0	1	$\frac{3}{5}$	$-\frac{1}{5}$	$\frac{6}{5}$
0	1	0	$-\frac{1}{5}$	$\frac{2}{5}$	$\frac{8}{5}$
1	0	0	$\frac{7}{5}$	$\frac{1}{5}$	$\frac{34}{5}$

Das Abbruchkriterium ist erreicht. Das Maximum $\frac{34}{5}$ wird angenommen für $x_1 = \frac{8}{5}$, $x_2 = \frac{6}{5}$ mit den Werten $y_1 = y_2 = 0$ für die Schlupfvariablen (siehe Endergebnis in Abbildung 4.11).

Da sich die erste Spalte (Spalte für z) bei der Durchführung des Algorithmus nicht verändert, kann man diese auch weglassen.

Ein im schlechtesten Fall besseres Laufzeitverhalten als der Simplex-Algorithmus hat das neuere **Innere-Punkte-Verfahren**, das unter Verwendung des Newton-Verfahrens gegen eine Lösung (in polynomialer Laufzeit) konvergiert.

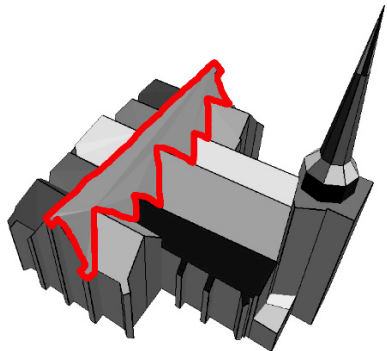


Abb. 4.12: Die markierte Dachfläche ist nicht eben. Die für die Darstellung verwendeten Dreiecke sind unterschiedlich geneigt und daher aufgrund der Beleuchtung deutlich sichtbar.

Beispiel 4.32 (Planarisierung mittels linearer Optimierung)

Bei unserer Arbeit mit 3-D-Stadtmodellen haben wir das Problem, dass eine durch ein Polygon (Streckenzug) berandete Dachfläche oft nicht in einer Ebene liegt, siehe Abbildung 4.12. Um dies zu korrigieren, modifizieren wir die z -Koordinaten (Höhenwerte)

der Polygonknoten (Ecken der Streckenzüge) möglichst minimal mittels eines linearen Programms. Sei $\{(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)\}$ die Menge aller Knoten der Dachpolygone eines Gebäudes, wobei benachbarte Polygone gemeinsame Knoten besitzen. Wir führen die Variablen $h_1^+ \geq 0, h_1^- \geq 0, \dots, h_n^- \geq 0$ mit dem Ziel ein, dass alle Dachpolygone mit der neuen Knotenmenge $\{(x_1, y_1, z_1 + h_1^+ - h_1^-), \dots, (x_n, y_n, z_n + h_n^+ - h_n^-)\}$ möglichst eben werden. Um die Abweichungen klein zu halten, wollen wir die Zielfunktion $\sum_{k=1}^n (h_k^+ + h_k^-)$ minimieren bzw. $\sum_{k=1}^n (-h_k^+ - h_k^-)$ maximieren. Anstelle eines nicht-linearen Betrags verwenden wir also die Differenz zweier nicht-negativer Variablen. Als Nebenbedingungen sollen die Höhenabweichungen jeder Dachfläche zu einer passenden Referenzebene kleiner als ein Schwellwert $\varepsilon > 0$ sein. Die Referenzebene wird über drei Polygonpunkte der Dachfläche definiert. Wir formulieren exemplarisch die Nebenbedingung für eine Dachfläche. Um die Notation einfach zu halten und keine weiteren Indizes einzuführen, möge diese Beispielfläche ebenfalls exemplarisch durch ein Polygon mit den Knoten $(x_1, y_1, z_1 + h_1^+ - h_1^-), \dots, (x_m, y_m, z_m + h_m^+ - h_m^-)$ berandet sein. Die ersten drei Knoten mögen zudem in der x - y -Ebene nicht auf einer Geraden liegen, wir definieren darüber die Referenzebene. Dann sind die x - y -Koordinaten jedes weiteren Knotens zu $4 \leq k \leq m$ darstellbar als

$$(x_k, y_k) = (x_2, y_2) + r_k(x_1 - x_2, y_1 - y_2) + s_k(x_3 - x_2, y_3 - y_2)$$

für Zahlen r_k und s_k . Das Polygon ist genau dann eben, wenn für $4 \leq k \leq m$ gilt:

$$\begin{aligned} z_k + h_k^+ - h_k^- &= z_2 + h_2^+ - h_2^- + r_k(z_1 + h_1^+ - h_1^- - z_2 - h_2^+ + h_2^-) \\ &\quad + s_k(z_3 + h_3^+ - h_3^- - z_2 - h_2^+ + h_2^-). \end{aligned}$$

Statt der Gleichheit wollen wir eine kleine Abweichung erlauben. Sei $d_k^+ - d_k^-$ die Differenz der beiden Seiten:

$$\begin{aligned} d_k^+ - d_k^- &= -z_k - h_k^+ + h_k^- + z_2 + h_2^+ - h_2^- + r_k(z_1 + h_1^+ - h_1^- - z_2 - h_2^+ + h_2^-) \\ &\quad + s_k(z_3 + h_3^+ - h_3^- - z_2 - h_2^+ + h_2^-). \end{aligned}$$

Diese lineare Gleichung für die Variablen $d_k^+ \geq 0$ und $d_k^- \geq 0$ wird als Nebenbedingung aufgenommen (die in die Standardform mit Ungleichungen überführt werden kann, s. o.). Jetzt fordern wir als weitere Nebenbedingungen $d_k^+ \leq \varepsilon$ und $d_k^- \leq \varepsilon$. Entsprechende Bedingungen sind für jedes Dachpolygon aufzustellen. Praktische Tests haben zu verfeinerten und weiteren Bedingungen geführt, siehe Goebels, Pohle-Fröhlich und Rethmann (2016). ■

Kommerzielle Softwarepakete wie CPLEX oder Gurobi und nicht-kommerzielle Pakete wie SCIP/SoPlex berechnen nicht nur lineare Programme. Sie lösen auch ganzzahlige oder gemischt-ganzzahlige lineare Optimierungsprobleme, bei denen als zusätzliche Nebenbedingung noch hinzukommt, dass einige oder alle Variablenwerte ganzzahlig

sein müssen. Der Simplex-Algorithmus liefert allerdings in der Regel keine ganzzahligen Lösungen, so dass die Programme aus Simplex-Ergebnissen (mit viel Aufwand und vielen Tricks) eine ganzzahlige Lösung gewinnen müssen. Eine gängige Technik ist beispielsweise **Branch-and-Bound**. Dabei wird das gegebene Maximierungsproblem sukzessive (z. B. durch Belegung oder Einschränkung von Variablen) in Teilprobleme aufgeteilt („branch“). Vergleichsweise schnell kann mit dem Simplex-Algorithmus eine eventuell nicht-ganzzahlige Lösung eines Teilproblems berechnet werden. Der zugehörige Wert der Zielfunktion ist offensichtlich eine obere Schranke M für das zu berechnende Maximum des ganzzahligen Teilproblems. Wurde zuvor bereits eine zulässige Lösung mit einem Maximum größergleich M gefunden, so kann das aktuelle Teilproblem ignoriert werden („bound“), es führt zu keiner besseren Lösung. Eine weitere Technik besteht im sukzessiven Hinzufügen weiterer Nebenbedingungen (zusätzlicher Ungleichungen, die „**Cuts**“ genannt werden). Durch die neuen Nebenbedingungen dürfen keine zulässigen Lösungen verloren gehen. Sie sollen aber die Menge der nicht-ganzzahligen Lösungen weiter einschränken. Die sogenannten **Gomory-Cuts** können automatisch aus den Zeilen des Simplex-Tableaus abgelesen werden. Dazu muss man sich nur überlegen, was die entsprechenden Gleichungen für ganzzahlige Lösungen bedeuten.

Mit solchen Techniken können kombinatorische Probleme geschickter als durch Ausprobieren gelöst werden (vgl. Kapitel 7.2.3).

Beispiel 4.33 (Sudoku)

Mit einem ganzzahligen linearen Programm, bei dem die Variablen $x_{i,k,l} \in \{0,1\}$, $i, k, l \in \{1, \dots, 9\}$, nur die binären Werte 0 und 1 annehmen dürfen, lässt sich ein 9×9 -Sudoku-Rätsel lösen. Hier ist eine dünn mit einzelnen Ziffern besetzte 9×9 -Matrix vorgegeben. Die unbesetzten Plätze sind so mit Ziffern zu belegen, dass jede Ziffer nur einmal pro Zeile, einmal pro Spalte und einmal in jedem der neun 3×3 -Kästchen verwendet wird, die nebeneinander die Matrix überdecken. Der Index i bestimmt die Zeile, j die Spalte in der Matrix, l bezeichnet eine Ziffer für diese Position. Steht an der Stelle $(2, 3)$ die Ziffer 7, so ist $x_{2,3,7} = 1$, $x_{2,3,l} = 0$ für alle $l \neq 7$. Bei einem Sudoku sind einige Variablen vorgelegt, die Werte der anderen müssen gefunden werden. Dies geschieht durch Lösen der Aufgabe:

$$\begin{aligned}
 &\text{Maximiere} && \sum_{i=1}^9 \sum_{k=1}^9 \sum_{l=1}^9 x_{i,k,l} \quad (\text{falls es eine Lösung gibt, ist das Maximum 81}) \\
 &\text{s. t.} && \sum_{l=1}^9 x_{i,k,l} \leq 1 \text{ für alle } i, k \in \{1, \dots, 9\} \quad (\text{maximal eine Ziffer pro Feld}), \\
 &&& \sum_{i=1}^9 x_{i,k,l} \leq 1 \text{ für alle } k, l \in \{1, \dots, 9\} \quad (\text{keine Ziffer doppelt pro Spalte}),
 \end{aligned}$$

$$\sum_{k=1}^9 x_{i,k,l} \leq 1 \text{ für alle } i, l \in \{1, \dots, 9\} \text{ (keine Ziffer doppelt pro Zeile),}$$

$$\sum_{r \in \{2,5,8\}} \sum_{s \in \{2,5,8\}} \sum_{i=r-1}^{r+1} \sum_{k=s-1}^{s+1} x_{i,k,l} \leq 1 \text{ für alle } l \in \{1, \dots, 9\}$$

(keine Ziffer doppelt pro Kästchen). ■

4.4 Integralrechnung mit mehreren Variablen

Während wir bislang mit Integralen den Inhalt von Flächen unter Funktionsgraphen bestimmt haben, berechnen wir nun das Volumen von Körpern in höheren Raumdimensionen. Beispielsweise ist das Integral einer stetigen Funktion $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ über eine geeignete Menge D das Volumen des Körpers zwischen der Fläche D in der x - y -Ebene und dem Funktionsgraphen, der als Fläche in $\mathbb{R}^3$ aufgefasst werden kann. Auch wenn es unsere Vorstellung sprengt, so erlauben wir statt $\mathbb{R}^2$ direkt allgemeinere Definitionsbereiche, die in $\mathbb{R}^n$ liegen. Sehr angenehm ist, dass das bislang diskutierte Integral nicht nur der Spezialfall für $n = 1$ ist, sondern sich die Berechnung von Integralen für Funktionen mit n Variablen oft auch darauf zurückführen lässt.

4.4.1 Integration über mehrdimensionale Intervalle

Wir haben bisher Funktionen über Intervalle $[a, b] \subset \mathbb{R}$ integriert. Wir verallgemeinern nun den Integrationsbegriff auf Integrationsbereiche $E \subset \mathbb{R}^n$, über die wir reellwertige Funktionen $f : E \rightarrow \mathbb{R}$ integrieren werden. Das Vorgehen entspricht dabei genau dem in Kapitel 2.6.

Die hier vorgenommene Erweiterung des Integrationsbegriffs hat vielfältige Anwendungen, z. B. bei der Beschreibung elektrischer Felder (vgl. Satz 4.20) oder bei der Berechnung von Trägheitsmomenten von Körpern mit endlicher Ausdehnung. Auch kann man damit die Größe von Oberflächen bestimmen.

Definition 4.13 (Intervalle in $\mathbb{R}^n$)

Seien $\vec{a}, \vec{b} \in \mathbb{R}^n$ mit $a_k < b_k$ für $1 \leq k \leq n$.

■ Die Menge

$$\begin{aligned} [\vec{a}, \vec{b}] &:= [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \\ &:= \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n : a_k \leq x_k \leq b_k, 1 \leq k \leq n\} \end{aligned}$$

heißt **n -dimensionales (abgeschlossenes) Intervall**. Es ist das kartesische Produkt von n eindimensionalen Intervallen und wird in der Schreibweise $[\vec{a}, \vec{b}]$ über zwei gegenüberliegende Eckpunkte $\vec{a}$ und $\vec{b}$ beschrieben.

■ **Das Produkt**

$$v([\vec{a}, \vec{b}]) := (b_1 - a_1) \cdot (b_2 - a_2) \cdots (b_n - a_n)$$

heißt der **elementare Inhalt** des Intervalls $[\vec{a}, \vec{b}]$.

■ **Eine Zerlegung Z des Intervalls $[\vec{a}, \vec{b}]$ ist ein Tupel $(Z_1, Z_2, \dots, Z_n)$ von Zerlegungen Z_k der Intervalle $[a_k, b_k]$ (vgl. Definition 2.29 auf Seite 329).**

Für $n = 1$ ist der elementare Inhalt die Länge des Intervalls $[a_1, b_1]$, für $n = 2$ entspricht der elementare Inhalt von $[a_1, b_1] \times [a_2, b_2]$ der Fläche des Rechtecks mit Seitenlängen $b_1 - a_1$ und $b_2 - a_2$, und für $n = 3$ erhalten wir das Volumen des Quaders usw.

Beispiel 4.34

$$[(1, 0, 2), (3, 2, 5)] = [1, 3] \times [0, 2] \times [2, 5] \text{ und } v([(1, 0, 2), (3, 2, 5)]) = (3-1) \cdot (2-0) \cdot (5-2).$$

■

Ist $Z_k = \{x_{k,0}, x_{k,1}, \dots, x_{k,N_k}\}$, so zerlegt Z das Intervall $[\vec{a}, \vec{b}]$ in $N_1 \cdot N_2 \cdots N_n$ Teilintervalle. Über diesen kann jeweils das Supremum und Infimum von f ermittelt werden (bei einer stetigen Funktion f ist das ein größter und ein kleinster Funktionswert), so dass wir analog zu Funktionen mit einer Variable Ober- und Untersummen bilden können:

Definition 4.14 (Unter- und Obersummen)

Seien $[\vec{a}, \vec{b}] \subset \mathbb{R}^n$ und Z eine Zerlegung von $[\vec{a}, \vec{b}]$ wie in Definition 4.13. Weiter sei $f : [\vec{a}, \vec{b}] \rightarrow \mathbb{R}$ eine **beschränkte Funktion**, d. h., der reelle Wertebereich ist beschränkt.

$$s_Z := \sum_{l_1=1}^{N_1} \sum_{l_2=1}^{N_2} \cdots \sum_{l_n=1}^{N_n} \left[\inf \{f(\vec{x}) : \vec{x} \in [(x_{1,l_1-1}, \dots, x_{n,l_n-1}), (x_{1,l_1}, \dots, x_{n,l_n})]\} \right] \cdot v([(x_{1,l_1-1}, \dots, x_{n,l_n-1}), (x_{1,l_1}, \dots, x_{n,l_n})])$$

heißt eine **Untersumme** von f bezüglich der Zerlegung Z und

$$S_Z := \sum_{l_1=1}^{N_1} \sum_{l_2=1}^{N_2} \cdots \sum_{l_n=1}^{N_n} \left[\sup \{f(\vec{x}) : \vec{x} \in [(x_{1,l_1-1}, \dots, x_{n,l_n-1}), (x_{1,l_1}, \dots, x_{n,l_n})]\} \right] \cdot v([(x_{1,l_1-1}, \dots, x_{n,l_n-1}), (x_{1,l_1}, \dots, x_{n,l_n})])$$

heißt eine **Obersumme** von f bezüglich der Zerlegung Z .

In der Untersumme wird auf jedem der zuvor erwähnten Teilintervalle eine größte untere Schranke der Funktionswerte mit dem Volumen des jeweiligen Teilintervalls multipliziert, bei der Obersumme wird dieses Volumen mit einer kleinsten oberen Schranke der Funktionswerte multipliziert. Das geht, da f nach Voraussetzung mit einem Wert M beschränkt ist ($|f(\vec{x})| \leq M$). Offensichtlich gilt damit

$$-Mv([\vec{a}, \vec{b}]) \leq s_Z \leq S_Z \leq Mv([\vec{a}, \vec{b}]),$$

und beide Zahlen sind eine Approximation an den gesuchten Rauminhalt, der zwischen s_Z und S_Z liegt.

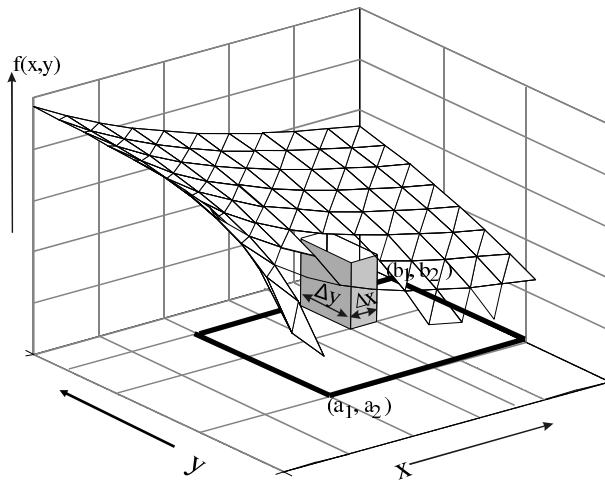


Abb. 4.13: Zur Konstruktion einer Obersumme für $\int_{[\vec{a}, \vec{b}]} f(x, y) d(x, y)$

Definition 4.15 (Riemann-Integral)

Seien $f : [\vec{a}, \vec{b}] \subset \mathbb{R}^n \rightarrow \mathbb{R}$ eine beschränkte Funktion und $\mathcal{Z}$ die Menge aller Zerlegungen des Intervalls $[\vec{a}, \vec{b}]$. Die kleinste obere Schranke der Untersummen

$$\underline{I} := \sup\{s_Z : Z \in \mathcal{Z}\}$$

heißt **Riemann-Unterintegral** von f auf $[\vec{a}, \vec{b}]$. Die größte untere Schranke der Obersummen

$$\bar{I} := \inf\{S_Z : Z \in \mathcal{Z}\}$$

heißt **Riemann-Oberintegral** von f auf $[\vec{a}, \vec{b}]$. Falls $\underline{I} = \bar{I}$, heißt f auf $[\vec{a}, \vec{b}]$ **Riemann-integrierbar** (kurz: **integrierbar**), und die Zahl $\underline{I} = \bar{I}$ heißt das **Riemann-Integral** (oder kurz **Integral**) von f auf $[\vec{a}, \vec{b}]$. Bezeichnung:

$$\int_{[\vec{a}, \vec{b}]} f(\vec{x}) d\vec{x} = \int \dots \int_{[\vec{a}, \vec{b}]} f(\vec{x}) d\vec{x},$$

wobei häufig n Integralzeichen bei einem Integrationsintervall in $\mathbb{R}^n$ geschrieben werden.

Das mehrfache Schreiben des Integralsymbols gibt bereits einen Hinweis darauf, dass wir das Integral später durch mehrere Integrale zu einer Variable lösen werden.

Beispiel 4.35

$$\iint_{[0,1] \times [2,5]} 42 \, d(x, y) = 42(1-0)(5-2) = 126. \quad \blacksquare$$

Wie in Kapitel 2.6 kann man statt der Ober- und Untersummen auch Zwischensummen betrachten. Dabei summiert man über alle Teilintervalle der Zerlegung und multipliziert den elementaren Inhalt des jeweiligen Teilintervalls mit einem Funktionswert zu einer Stelle (an einem Zwischenpunkt) aus diesem Intervall. Das Riemann'sche Integrabilitätskriterium Satz 2.42 von Seite 333 gilt hier völlig analog. Eine Funktion ist genau dann Riemann-integrierbar, wenn für immer feiner werdende Zerlegungen und für jede zugehörige Zwischenpunktwahl die Zwischensummen gegen eine Zahl streben, die dann gleich dem Integral ist.

Auch die Eigenschaften des Integrals von Funktionen mit einer Variable übertragen sich. Insbesondere ist das Integral linear, d. h., für $f : [\vec{a}, \vec{b}] \subset \mathbb{R}^n \rightarrow \mathbb{R}$, $g : [\vec{a}, \vec{b}] \subset \mathbb{R}^n \rightarrow \mathbb{R}$, $c, d \in \mathbb{R}$ gilt, sofern die Integrale existieren:

$$\int_{[\vec{a}, \vec{b}]} c \cdot f(\vec{x}) + d \cdot g(\vec{x}) \, d\vec{x} = c \int_{[\vec{a}, \vec{b}]} f(\vec{x}) \, d\vec{x} + d \int_{[\vec{a}, \vec{b}]} g(\vec{x}) \, d\vec{x}.$$

Ebenso sind auf $[\vec{a}, \vec{b}]$ stetige Funktionen integrierbar (vgl. Satz 2.43 auf Seite 335).

Um nun Integrale über Teilmengen des $\mathbb{R}^n$ auszurechnen, ist der Satz von Fubini entscheidend. Mit ihm lassen sich die Integrale auf mehrere Integrale von Funktionen mit einer Variable zurückführen, die mit den Mitteln aus Kapitel 2.6 ausgerechnet werden können. Um zu sehen, wie dies funktioniert, ersetzen wir ein Integral über $[a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$ näherungsweise durch eine Zwischensumme zur Zerlegung $Z = (Z_1, Z_2)$ mit $Z_1 = \{x_{1,0}, x_{1,1}, \dots, x_{1,N_1}\}$ und $Z_2 = \{x_{2,0}, x_{2,1}, \dots, x_{2,N_2}\}$. Als Zwischenstellen verwenden wir $(x_{1,k}, x_{2,l}) \in [x_{1,k-1}, x_{1,k}] \times [x_{2,l-1}, x_{2,l}]$, $k = 1, \dots, N_1$, $l = 1, \dots, N_2$. Dann sehen wir Zwischensummen zu Integralen über Funktionen einer Variable:

$$\begin{aligned} \iint_{[a_1, b_1] \times [a_2, b_2]} f(x_1, x_2) \, d(x_1, x_2) &\approx \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} (x_{1,k} - x_{1,k-1})(x_{2,l} - x_{2,l-1}) f(x_{1,k}, x_{2,l}) \\ &\approx \sum_{k=1}^{N_1} (x_{1,k} - x_{1,k-1}) \int_{a_2}^{b_2} f(x_{1,k}, x_2) \, dx_2 \approx \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f(x_1, x_2) \, dx_2 \right] dx_1. \end{aligned}$$

Vertauschen wir die beiden Summen, erhalten wir stattdessen

$$\iint_{[a_1, b_1] \times [a_2, b_2]} f(x_1, x_2) \, d(x_1, x_2) \approx \int_{a_2}^{b_2} \left[\int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \right] dx_2.$$

Ist z.B. f stetig auf $[\vec{a}, \vec{b}]$, so kann man zeigen, dass alle durch Integration entstehenden Funktionen ebenfalls stetig und damit integrierbar sind. Zusammen mit dem Riemann'schen-Integrabilitätskriterium kann man dann das vorangehende Argument zu einem Beweis des folgenden Satzes ausbauen:

Satz 4.12 (Fubini)

Sei $f : [\vec{a}, \vec{b}] \subset \mathbb{R}^n \rightarrow \mathbb{R}$ stetig. Dann gilt:

$$\int_{[\vec{a}, \vec{b}]} f(\vec{x}) d\vec{x} = \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} \left[\dots \left[\int_{a_n}^{b_n} f(x_1, x_2, \dots, x_n) dx_n \right] \dots \right] dx_2 \right] dx_1.$$

Die Integrationsreihenfolge kann dabei beliebig getauscht werden.

Beim Ausrechnen des innersten Integrals müssen die Variablen $x_1, x_2, \dots, x_{n-1}$ als konstante Werte betrachtet werden. Das ist analog zum Ausrechnen einer partiellen Ableitung. Nach dem Ausrechnen hat man eine Funktion der Variablen $x_1, x_2, \dots, x_{n-1}$. Das ist insofern anders als beim Berechnen einer partiellen Ableitung, als nun eine Variable nicht mehr vorkommt. Für das nächste Integral sind die Variablen $x_1, x_2, \dots, x_{n-2}$ als Konstanten zu interpretieren. Hat man schließlich alle n Integrale berechnet, so sind alle Variablen verschwunden, und das Ergebnis ist eine reelle Zahl.

Achtung: Wichtig beim Aufbrechen des Integrals ist, dass die unteren Integrationsgrenzen der Einzelintegrale alle kleiner oder gleich der jeweiligen oberen Grenze sind. Anderenfalls handelt man sich ein falsches Vorzeichen ein.

Beispiel 4.36

- a) Das Intervall $[0, 1] \times [0, 2]$ kann auch durch die Eckpunkte als $[(0, 0), (1, 2)]$ ausgedrückt werden. Wir integrieren darüber die Funktion $x^2 - 2xy + 3$:

$$\begin{aligned} \iint_{[0,1] \times [0,2]} x^2 - 2xy + 3 d(x, y) &= \int_0^1 \left[\int_0^2 x^2 - 2xy + 3 dy \right] dx \\ &= \int_0^1 [x^2 y - xy^2 + 3y]_{y=0}^{y=2} dx = \int_0^1 2x^2 - 4x + 6 dx \\ &= \left[\frac{2}{3}x^3 - \frac{4}{2}x^2 + 6x \right]_0^1 = \frac{2}{3} - 2 + 6 = \frac{14}{3}. \end{aligned}$$

- b) Auf $[1, 2] \times [0, 1] = [(1, 0), (2, 1)]$ wird $x^3 \cdot \exp(x^2 y)$ integriert:

$$\begin{aligned} \iint_{[1,2] \times [0,1]} x^3 \cdot \exp(x^2 y) d(x, y) &= \int_1^2 \int_0^1 x^3 \cdot \exp(x^2 y) dy dx \\ &= \int_1^2 [x \exp(x^2 y)]_{y=0}^{y=1} dx = \int_1^2 x \exp(x^2) - x dx = \left[\frac{1}{2} \exp(x^2) - \frac{1}{2} x^2 \right]_1^2 \\ &= \frac{e^4}{2} - 2 - \frac{e}{2} + \frac{1}{2} = \frac{e^4 - e - 3}{2}. \quad \blacksquare \end{aligned}$$

Achtung: Integralzeichen $\int_{a_1}^{b_1}$ und die Angabe dx_1 einer Integrationsvariable bilden ein Klammerpaar, dessen Angaben zueinander passen müssen. Wählt man eine andere Integrationsreihenfolge, so ist diese Klammerung konsistent zu ändern:

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_2 dx_1 = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) dx_1 dx_2.$$

Im Allgemeinen ist $\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) dx_2 dx_1$ eine andere Zahl oder gar nicht definiert.

Nach dem Satz von Schwarz (Satz 4.5) kann die Reihenfolge partieller Ableitungen beliebig gewählt werden, wenn die partiellen Ableitungen stetig sind. Er ist ein Satz zur Vertauschung von Grenzwerten, nämlich der partiellen Ableitungen. Der Satz von Fubini erlaubt in ähnlicher Weise das Vertauschen der Reihenfolge von iterierten Integralen, die ebenfalls über einen Grenzprozess definiert sind. Geht es um das Vertauschen von Integralen, so ist die folgende Version des Satzes praktikabler, die unter geeigneten Bedingungen sogar für unbeschränkte Intervalle gilt und die Stetigkeit des Integranden nicht benötigt. Wir werden sie später noch in Beweisen (z. B. beim Faltungssatz, siehe Seite 837) verwenden, für die Ingenieur-Praxis ist nur wichtig, dass man in vielen Situationen die Integrationsreihenfolge ändern darf.

Satz 4.13 (Satz von Fubini-Tonelli zum Vertauschen von Integralen *)

Sei $f : [a, b] \times [c, d] \rightarrow \mathbb{R}$, so dass $f(x, y)$ (und damit auch $|f(x, y)|$) bei jedem festen $x \in [a, b]$ als Funktion von y auf $[c, d]$ integrierbar ist und $\int_c^d f(x, y) dy$, $\int_c^d |f(x, y)| dy$ als Funktionen von x auf $[a, b]$ integrierbar sind. Dann darf die Integrationsreihenfolge vertauscht werden, es gilt:

$$\int_a^b \left[\int_c^d f(x, y) dy \right] dx = \int_c^d \left[\int_a^b f(x, y) dx \right] dy = \iint_{[a, b] \times [c, d]} f(x, y) d(x, y).$$

Der Satz gilt auch, wenn alternativ $f(x, y)$ bei jedem festen $y \in [c, d]$ als Funktion von x auf $[a, b]$ integrierbar ist und $\int_a^b f(x, y) dx$, $\int_a^b |f(x, y)| dx$ als Funktionen von y auf $[c, d]$ integrierbar sind.

Die Voraussetzung, dass $\int_a^b |f(x, y)| dx$ als Funktion von y oder $\int_c^d |f(x, y)| dy$ als Funktion von x integrierbar sein müssen, kann nicht weggelassen werden:

Beispiel 4.37

Wir betrachten (siehe Abbildung 4.14)

$$f(x, y) := \begin{cases} \frac{1}{y^2} & \text{für } 0 < x < y < 1, \\ -\frac{1}{x^2} & \text{für } 0 < y < x < 1, \\ 0 & \text{sonst.} \end{cases} \quad (4.14)$$

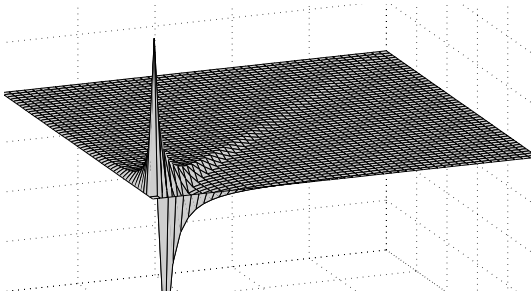


Abb. 4.14: In einer Umgebung des Nullpunkts unbeschränkte Funktion (4.14) als Beispiel zu den Voraussetzungen des Satzes von Fubini

- Für jedes $0 < y < 1$ ist

$$\int_0^1 f(x, y) dx = \int_0^y \frac{1}{y^2} dx - \int_y^1 \frac{1}{x^2} dx = \frac{1}{y} - \left[-\frac{1}{x} \right]_y^1 = \frac{1}{y} + 1 - \frac{1}{y} = 1.$$

Damit ist $\int_0^1 \int_0^1 f(x, y) dx dy = \int_0^1 dy = 1$. Man beachte aber, dass

$$\int_0^1 |f(x, y)| dx = \int_0^y \frac{1}{y^2} dx + \int_y^1 \frac{1}{x^2} dx = \frac{2}{y} - 1$$

nicht integrierbar bezüglich y auf $[0, 1]$ ist.

- Für jedes $0 < x < 1$ ist

$$\int_0^1 f(x, y) dy = \int_0^x -\frac{1}{x^2} dy + \int_x^1 \frac{1}{y^2} dy = -\frac{1}{x} - 1 + \frac{1}{x} = -1.$$

Damit ist

$$\int_0^1 \int_0^1 f(x, y) dy dx = \int_0^1 -1 dy = -1 \neq 1 = \int_0^1 \int_0^1 f(x, y) dx dy.$$

Glücklicherweise sind die Voraussetzungen des Satzes von Fubini ja auch nicht erfüllt. Es ist auch

$$\int_0^1 |f(x, y)| dy = \int_0^x \frac{1}{x^2} dy + \int_x^1 \frac{1}{y^2} dy = \frac{2}{x} - 1.$$

nicht integrierbar bezüglich x auf $[0, 1]$. ■

Zerfällt der Integrand in einzelne Faktoren, die jeweils nur von einer Variable abhängen, so kann man ein Mehrfachintegral als Produkt von Integralen zu einer Variable ausrechnen:

$$\int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f(x)g(y) dy \right] dx = \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} g(y) dy \right] f(x) dx = \left[\int_{a_2}^{b_2} g(y) dy \right] \left[\int_{a_1}^{b_1} f(x) dx \right],$$

also z. B.

$$\int_0^\pi \int_0^1 \sin(x) \exp(y) dy dx = \int_0^\pi \sin(x) dx \cdot \int_0^1 \exp(y) dy = 2(e - 1).$$

Achtung: Das ist ein Spezialfall. So darf nur vorgegangen werden, wenn f nicht auch noch von y und g nicht auch noch von x abhängt. Dass man sonst einen Fehler macht, sieht man z. B. daran, dass das Ergebnis der Integration noch von Variablen abhängen würde. Es muss aber eine Zahl und keine Funktion herauskommen.

Häufig hat man es mit Parameterintegralen zu tun. Dabei hängt der Integrand von einer weiteren Variable ab, nach der nicht integriert wird. Das ist beispielsweise bei der Faltung und den Integraltransformationen (Fourier- und Laplace-Transformation) der Fall, die wir in Kapitel 6 behandeln.

Satz 4.14 (Ableitung von Parameterintegralen)

Sei f stetig auf dem Intervall $[a, b] \times [c, d] \subset \mathbb{R}^2$ und hier ebenfalls stetig partiell nach x differenzierbar. Dann gilt die **Leibniz'sche Regel für Parameterintegrale**:

$$\frac{d}{dx} \int_c^d f(x, y) dy = \int_c^d \frac{\partial}{\partial x} f(x, y) dy,$$

d. h., Differenziation und Integration dürfen vertauscht werden.

Beweis: Zum Beweis benutzen wir den hier anwendbaren Satz von Fubini. Dazu seien $x_0, x_0 + h \in [a, b]$ für $h \neq 0$. Wir betrachten den Differenzenquotienten der zu berechnenden Ableitung und schreiben ihn mit dem Fundamentalsatz der Differenzial- und Integralrechnung (Seite 340) um:

$$\begin{aligned} \frac{1}{h} \left[\int_c^d f(x_0 + h, y) dy - \int_c^d f(x_0, y) dy \right] &= \frac{1}{h} \int_c^d f(x_0 + h, y) - f(x_0, y) dy \\ &= \frac{1}{h} \int_c^d \int_{x_0}^{x_0+h} \frac{\partial}{\partial x} f(x, y) dx dy = \frac{1}{h} \int_{x_0}^{x_0+h} \int_c^d \frac{\partial}{\partial x} f(x, y) dy dx, \end{aligned}$$

wobei wir die Integrationsgrenzen mit dem Satz von Fubini vertauscht haben. Da $\frac{\partial}{\partial x} f(x, y)$ auf dem Intervall $[a, b] \times [c, d]$ stetig (und sogar gleichmäßig stetig) ist, kann man sich mit der Definition der Stetigkeit überlegen, dass die Funktion $\int_c^d \frac{\partial}{\partial x} f(x, y) dy$ (mit der Variable x) auf $[a, b]$ stetig ist. Zusammen mit dem Mittelwertsatz der Integralrechnung (siehe Seite 337) folgt daher

$$\lim_{h \rightarrow 0} \frac{1}{h} \int_{x_0}^{x_0+h} \int_c^d \frac{\partial}{\partial x} f(x, y) dy dx = \int_c^d \frac{\partial}{\partial x} f(x_0, y) dy$$

und damit die Aussage des Satzes. □

Bisweilen hilft dieser Satz beim Ausrechnen von gewöhnlichen Integralen:

Beispiel 4.38

Die Funktion $\frac{x^2-1}{\ln(x)}$ ist stetig auf $[0, 1]$, wenn man sie für $x = 0$ mit dem Wert 0 und für $x = 1$ mit dem Wert 2 fortsetzt, denn insbesondere gilt mit dem Satz von L'Hospital für den Typ 0/0:

$$\lim_{x \rightarrow 1-} \frac{x^2 - 1}{\ln(x)} = \lim_{x \rightarrow 1-} \frac{2x}{\frac{1}{x}} = 2.$$

Wir wollen $\int_0^1 \frac{x^2-1}{\ln(x)} dx$ berechnen. Dazu betrachten wir für $a \geq 0$ aber die Funktion $\frac{x^a-1}{\ln(x)}$, die ebenso stetig fortgesetzt werden kann.

$$\begin{aligned} I(a) &:= \int_0^1 \frac{x^a - 1}{\ln(x)} dx, \\ I'(a) &= \int_0^1 \frac{\partial}{\partial a} \frac{x^a - 1}{\ln(x)} dx = \int_0^1 x^a dx = \left[\frac{1}{a+1} x^{a+1} \right]_0^1 = \frac{1}{a+1}. \end{aligned}$$

Damit wissen wir durch Bildung der Stammfunktion, dass $I(a) = \ln(a+1) + c$ ist. Offensichtlich ist $I(0) = 0$, da für $a = 0$ der Integrand null wird. Damit ist $c = 0$ bestimmt. Gesucht war ursprünglich $I(2) = \ln(3)$. Der Trick besteht hier darin, dass $\frac{\partial}{\partial a} \frac{x^a-1}{\ln(x)} = x^a$ ist und damit der Nenner wegfällt. Durch die Ableitung nach dem Parameter wird die zu integrierende Funktion also viel einfacher. ■

Im Beispiel haben wir einen Parameter a eingebaut und dann $I(a)$ berechnet, wobei wir ursprünglich lediglich an $I(2)$ interessiert waren. Das ist ein allgemeiner Rechentrick. Der Parameter a ermöglicht ein Ableiten, wobei $I'(a)$ mit der Regel für Parameterintegrale bestimmt werden kann. Die Hoffnung dabei ist, dass der abgeleitete Integrand einfach zu integrieren ist. Anschließende Bildung der Stammfunktion führt bis auf eine Konstante zu $I(a)$. Die Konstante bestimmt man durch Einsetzen eines Werts für a , für den $I(a)$ leicht berechenbar ist. Damit ist $I(a)$ allgemein bestimmt, und wir verfügen auch über den speziell gesuchten Wert – im Beispiel $I(2)$.

4.4.2 Integration über Normalbereiche

Die Integration von Funktionen über allgemeine Mengen wird auf die Integration über Intervalle zurückgeführt.

Definition 4.16 (Allgemeiner Integrationsbereich)

Sei $E \subset \mathbb{R}^n$ eine **beschränkte Menge**, d. h., die Menge der reellen Beträge der Elemente von E ist beschränkt. Soll eine Funktion f über E integriert werden, dann wählt man ein E umschließendes Intervall $[\vec{a}, \vec{b}]$ (also $E \subset [\vec{a}, \vec{b}]$) und setzt

$$g(\vec{x}) := \begin{cases} f(\vec{x}) & : \vec{x} \in E, \\ 0 & : \vec{x} \in [\vec{a}, \vec{b}] \setminus E. \end{cases}$$

Damit ist (unabhängig von der Wahl des Intervalls) das Integral von f über E definiert als

$$\int_E f(\vec{x}) d\vec{x} := \int_{[\vec{a}, \vec{b}]} g(\vec{x}) d\vec{x}, \quad (4.15)$$

sofern die rechte Seite (im Sinne von Definition 4.15) existiert.

Beispiel 4.39

Es sei B ein fester Körper in $\mathbb{R}^3$ mit einer Dichte $\rho(x, y, z)$ (Masse pro Volumen), die vom jeweiligen Punkt (x, y, z) des Körpers abhängt. Dann ist $V = \iiint_B 1 d(x, y, z)$ das Volumen von B , die Masse von B ist $M = \iiint_B \rho(x, y, z) d(x, y, z)$, und die Koordinaten des Schwerpunkts (x_s, y_s, z_s) von B lauten:

$$x_s = \frac{1}{M} \iiint_B x \cdot \rho(x, y, z) d(x, y, z), \quad y_s = \frac{1}{M} \iiint_B y \cdot \rho(x, y, z) d(x, y, z)$$

und $z_s = \frac{1}{M} \iiint_B z \cdot \rho(x, y, z) d(x, y, z)$. ■

Im Beispiel haben wir durch Integration der konstanten Funktion $f(x, y, z) = 1$ das Volumen eines Objektes angegeben. Um das zu verstehen, betrachten wir den einfacheren Fall einer Fläche in $\mathbb{R}^2$, über die wir die Funktion $f(x, y) = 1$ integrieren. Wir erhalten das Volumen eines Objektes mit der gegebenen Fläche als Grundfläche und der Höhe 1. Die Maßzahl des Volumens ist damit gleich der Maßzahl des Flächeninhalts. Alternativ können wir diese auch über ein Integral einer Variable ausrechnen, wenn es sich um eine Fläche unter einem Funktionsgraphen handelt.

Man kann nicht erwarten, dass Integrale über beliebige beschränkte Mengen E wie beispielsweise die Mandelbrotmenge (siehe Seite 214) existieren. Die Mengen bzw. ihr Rand müssen glatt genug sein, damit die Funktion g aus Definition 4.16 integrierbar wird. Ein praktisch wichtiger Spezialfall sind Mengen E , bei denen man den Rand als Graphen integrierbarer Funktionen ausdrücken kann:

Definition 4.17 (Normalbereiche in $\mathbb{R}^2$)

- Die Funktionen $f_1(x), f_2(x) : [a_1, b_1] \rightarrow \mathbb{R}$ seien integrierbar, und es gelte $f_1(x) \leq f_2(x)$ für alle $x \in [a_1, b_1]$. Die Menge

$$B_x := \{(x, y) : a_1 \leq x \leq b_1, f_1(x) \leq y \leq f_2(x)\}$$

heißt **x-Normalbereich** (vgl. Abbildung 4.15).

- Die Funktionen $g_1(y), g_2(y) : [a_2, b_2] \rightarrow \mathbb{R}$ seien integrierbar, und es gelte $g_1(y) \leq g_2(y)$ für alle $y \in [a_2, b_2]$. Die Menge

$$B_y := \{(x, y) : a_2 \leq y \leq b_2, g_1(y) \leq x \leq g_2(y)\}$$

heißt **y-Normalbereich** (vgl. Abbildung 4.16).

Normalbereiche in höheren Dimensionen kann man entsprechend definieren, z. B. in $\mathbb{R}^3$, indem man zwei Funktionen auf einem Normalbereich $B_x \subset \mathbb{R}^2$ definiert, die die Werte in z -Richtung begrenzen. Mit dem Satz von Fubini folgt nun aus (4.15):

Satz 4.15 (Integration über Normalbereiche)

a) Seien $B_x := \{(x, y) : a_1 \leq x \leq b_1, f_1(x) \leq y \leq f_2(x)\}$ ein x -Normalbereich und $f(x, y) : B_x \rightarrow \mathbb{R}$ (die Funktion f hat nichts mit den Funktionen f_1 und f_2 des Normalbereichs zu tun) integrierbar, so dass zudem alle auftretenden Integrale im Sinne der Voraussetzungen von Satz 4.13 existieren. Dann gilt:

$$\iint_{B_x} f(x, y) d(x, y) = \int_{a_1}^{b_1} \left[\int_{f_1(x)}^{f_2(x)} f(x, y) dy \right] dx.$$

b) Seien $B_y := \{(x, y) : a_2 \leq y \leq b_2, g_1(y) \leq x \leq g_2(y)\}$ ein y -Normalbereich und $f(x, y) : B_y \rightarrow \mathbb{R}$ integrierbar, so dass wie zuvor die Voraussetzungen von Satz 4.13 erfüllt sind. Dann gilt:

$$\iint_{B_y} f(x, y) d(x, y) = \int_{a_2}^{b_2} \left[\int_{g_1(y)}^{g_2(y)} f(x, y) dx \right] dy.$$

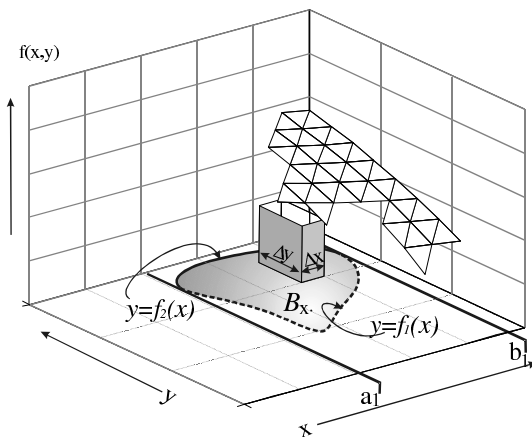


Abb. 4.15: Integration über einen x -Normalbereich B_x

Beweis: Wir zeigen exemplarisch den Teil a). Dazu sei $[\vec{a}, \vec{b}]$ ein Intervall mit $B_x \subset [\vec{a}, \vec{b}]$. Weiter sei wie zuvor $g : [\vec{a}, \vec{b}] \rightarrow \mathbb{R}$ mit $g(x, y) = f(x, y)$ für alle $(x, y) \in B_x$ und $g(x, y) = 0$ für $(x, y) \notin B_x$. Mit dem Satz von Fubini erhalten wir

$$\iint_{B_x} f(x, y) d(x, y) = \iint_{[\vec{a}, \vec{b}]} g(x, y) d(x, y) = \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} g(x, y) dy \right] dx$$

$$= \int_{a_1}^{b_1} \left[\int_{f_1(x)}^{f_2(x)} g(x, y) dy \right] dx = \int_{a_1}^{b_1} \left[\int_{f_1(x)}^{f_2(x)} f(x, y) dy \right] dx,$$

da die Funktion g außerhalb des inneren Integrationsbereichs $[f_1(x), f_2(x)]$ gleich null ist, innerhalb dieses Bereichs aber mit f übereinstimmt. $\square$

Bei drei Variablen (mehr Variablen analog) diskutiert man entsprechend Integrale vom Typ

$$\int_{a_1}^{b_1} \left[\int_{f_1(x)}^{f_2(x)} \left[\int_{g_1(x,y)}^{g_2(x,y)} f(x, y, z) dz \right] dy \right] dx.$$

Auch hier muss man darauf achten, dass die jeweiligen unteren Integrationsgrenzen nicht größer als die entsprechenden oberen Grenzen sind. Außerdem dürfen Integrationsgrenzen nur von den Variablen abhängen, nach denen in den weiter außen liegenden Integralen integriert wird.

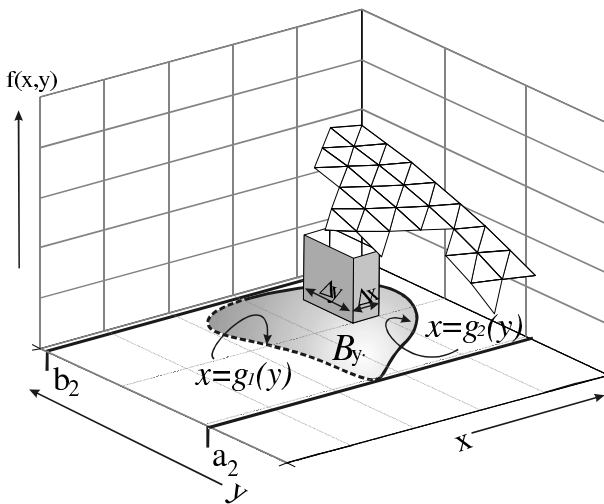


Abb. 4.16: Integration über einen y -Normalbereich B_y

Beispiel 4.40

Die Funktion

$$f(x, y) := 3x^2 - 6xy + 9$$

ist zu integrieren über der Menge $E \subset \mathbb{R}^2$, die durch die Graphen zu $f_1(x) := x^2 - 3$ und $f_2(x) := -x^2 + 2$ berandet ist. Zunächst bestimmen wir E genauer, indem wir die Schnittpunkte von f_1 und f_2 berechnen:

$$-x^2 + 2 = x^2 - 3 \implies 2x^2 = 5 \implies x = \pm \sqrt{\frac{5}{2}}.$$

Damit ist $E = \left\{ (x, y) : -\sqrt{\frac{5}{2}} \leq x \leq \sqrt{\frac{5}{2}}, x^2 - 3 \leq y \leq -x^2 + 2 \right\}$. (Man beachte, dass $f_1(x) \leq f_2(x)$ ist. Das sehen wir, indem wir z. B. $x = 0 \in \left[-\sqrt{\frac{5}{2}}, \sqrt{\frac{5}{2}}\right]$ einsetzen.)

$$\begin{aligned}
 \iint_E f(x, y) d(x, y) &= \int_{-\sqrt{\frac{5}{2}}}^{\sqrt{\frac{5}{2}}} \left[\int_{x^2-3}^{-x^2+2} 3x^2 - 6xy + 9 dy \right] dx \\
 &= 3 \int_{-\sqrt{\frac{5}{2}}}^{\sqrt{\frac{5}{2}}} [x^2 y - xy^2 + 3y]_{y=x^2-3}^{y=-x^2+2} dx = 3 \int_{-\sqrt{\frac{5}{2}}}^{\sqrt{\frac{5}{2}}} [-x^4 + 2x^2 - x(x^4 - 4x^2 + 4) - 3x^2 + 6] \\
 &\quad - [x^4 - 3x^2 - x(x^4 - 6x^2 + 9) + 3x^2 - 9] dx \\
 &= \int_{-\sqrt{\frac{5}{2}}}^{\sqrt{\frac{5}{2}}} -6x^4 - 6x^3 - 3x^2 + 15x + 45 dx = \left[-\frac{6}{5}x^5 - \frac{3}{2}x^4 - x^3 + \frac{15}{2}x^2 + 45x \right]_{-\sqrt{\frac{5}{2}}}^{\sqrt{\frac{5}{2}}} \\
 &= -\frac{12}{5}\sqrt{\frac{5}{2}}^5 - 2\sqrt{\frac{5}{2}}^3 + 90\sqrt{\frac{5}{2}} = \sqrt{\frac{5}{2}} [-15 - 5 + 90] = 70\sqrt{\frac{5}{2}}.
 \end{aligned}$$

■

4.4.3 Substitutionsregel

Bei kugel- oder zylinderförmigen Integrationsbereichen wird die Berechnung von Integralen einfacher, wenn man eine Koordinatentransformation durchführt, so dass man anschließend über ein Intervall integrieren kann. Dies geschieht mit einer Substitutionsregel, die eine Verallgemeinerung der bereits bekannten Substitutionsregel in der Form (2.54) auf Seite 347 ist. Danach gilt für eine streng monotone und damit invertierbare Funktion g mit $x = g(t)$ und $dx = g'(t) dt$:

$$\int_a^b f(x) dx = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(t)) g'(t) dt.$$

Jetzt übertragen wir diese Situation in höhere Dimensionen. Dabei ist $\vec{g} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$. Die Rolle von g' nimmt dann die Funktionaldeterminante (Determinante der Jacobi-Matrix, siehe Definition 4.8) ein.

Satz 4.16 (Substitutionsregel)

Sei $E \subset \mathbb{R}^n$ beschränkt und offen. Weiter sei $\overline{E}$ die Menge, die durch Vereinigung von E mit dem Rand von E entsteht (siehe Seite 593). Die Funktion $\vec{g} : \overline{E} \rightarrow \mathbb{R}^n$ sei stetig auf $\overline{E}$, stetig differenzierbar auf E und invertierbar (injektiv) auf E . Weiter sei die Funktionaldeterminante

$$\det[\vec{g}'(\vec{x})] = \det \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\vec{x}) & \cdots & \frac{\partial g_1}{\partial x_n}(\vec{x}) \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1}(\vec{x}) & \cdots & \frac{\partial g_n}{\partial x_n}(\vec{x}) \end{bmatrix} \neq 0$$

für alle $\vec{x} \in E$. Außerdem sei $\det[\vec{g}'(\vec{x})]$ als reellwertige Funktion von $\vec{x}$ stetig auf E und stetig fortsetzbar auf $\overline{E}$. Weiter sei $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ stetig und beschränkt, außerdem sei $\vec{g}(\overline{E}) \subset D$. Für jeden Normalbereich (vgl. Definition 4.17) $A \subset \vec{g}(E)$ gilt:

$$\int_A f(\vec{x}) d\vec{x} = \int_{\vec{g}^{-1}(A) \subset E} f(\vec{g}(\vec{t})) \cdot |\det \vec{g}'(\vec{t})| d\vec{t}. \quad (4.16)$$

Insbesondere existieren beide Integrale.

Bemerkungen:

- Diese Regel wird in der Praxis überwiegend für wenige Funktionen $\vec{g}$ eingesetzt, die Koordinatentransformationen beschreiben. Damit muss die Regel kaum in dieser komplizierten Form angewendet werden. Die gängigen Koordinatentransformationen sehen wir uns im nächsten Abschnitt an.
- Es wird der Betrag der Determinante benutzt. Das ist nötig, da man bei einem Integrationsbereich aus $\mathbb{R}^n$ für $n > 1$ nicht wie bei $\mathbb{R}^1$ einen Vorzeichenwechsel über das Vertauschen von Integrationsgrenzen definieren kann (siehe Satz 4.15). Im Falle von Funktionen mit einer Variable ergibt sich kein Unterschied zur alten Substitutionsregel Satz 2.50 auf Seite 347:

- Ist $g(t)$ streng monoton steigend, so ist $g'(t) \geq 0$ und

$$\begin{aligned} \int_a^b f(x) dx &= \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(t))g'(t) dt = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(t))|g'(t)| dt \\ &= \int_{g^{-1}([a,b])} f(g(t))|g'(t)| dt, \end{aligned}$$

wobei $g^{-1}(a) \leq g^{-1}(b)$ und $g^{-1}([a,b]) = [g^{-1}(a), g^{-1}(b)]$.

- Ist $g(t)$ streng monoton fallend, so ist $g'(t) \leq 0$, und wir erhalten

$$\begin{aligned} \int_a^b f(x) dx &= \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(t))g'(t) dt = - \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(t))|g'(t)| dt \\ &= \int_{g^{-1}(b)}^{g^{-1}(a)} f(g(t))|g'(t)| dt = \int_{g^{-1}([a,b])} f(g(t))|g'(t)| dt, \end{aligned}$$

wobei $g^{-1}(b) \leq g^{-1}(a)$ und $g^{-1}([a,b]) = [g^{-1}(b), g^{-1}(a)]$.

- Wählt man $f(\vec{x}) = 1$, so gilt für das Volumen des Integrationsbereichs A :

$$\int_A 1 d\vec{x} = \int_{\vec{g}^{-1}(A)} |\det \vec{g}'(\vec{t})| d\vec{t}.$$

Damit beschreibt $|\det \vec{g}'(\vec{t})|$ lokal die Verzerrung von A unter $\vec{g}^{-1}$ (vgl. dazu die Interpretation der Determinante als Fläche eines Parallelogramms in Abbildung 1.60 auf Seite 183 oder eines Parallelepipeds über das Spatprodukt, siehe Seite 448). Am Beispiel einer konkreten Substitution mit Polarkoordinaten wird klarer, wie sich die Verzerrung auswirkt, siehe dazu Beispiel 4.41 auf Seite 664.

4.4.4 Polar-, Zylinder- und Kugelkoordinaten

Die häufigste Anwendung der Substitutionsregel sind Koordinatentransformationen. Dabei hat man eine Funktion $\vec{g} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ oder $\vec{g} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, die verschiedene Koordinatendarstellungen ineinander umrechnet. Sie überführt beispielsweise einen kreisförmigen oder einen kugelförmigen Integrationsbereich in ein Intervall.

4.4.4.1 Polarkoordinaten

Polarkoordinaten haben wir bereits zur Darstellung der komplexen Zahlen und bei der Dreh-Matrix kennengelernt (siehe Kapitel 1.6 und (1.48) auf Seite 160). In $\mathbb{R}^2$ können wir einen Punkt (x, y) über den Betrag $r \geq 0$ und Winkel φ zur x-Achse schreiben: $(x, y) = (r \cos \varphi, r \sin \varphi) = (\operatorname{Re}(re^{j\varphi}), \operatorname{Im}(re^{j\varphi}))$. Ist $\vec{t} = (t_1, t_2) = (r, \varphi)$ die Darstellung eines Punktes in Polarkoordinaten, so bekommen wir mittels

$$\vec{g}(\vec{t}) = \vec{g}(t_1, t_2) = \vec{g}(r, \varphi) = (r \cos \varphi, r \sin \varphi)$$

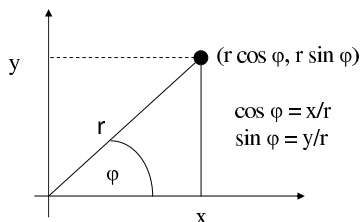
die Euklid'sche Darstellung des Punktes. Es ist

$$|\det[\vec{g}'(\vec{t})]| = \left| \det \begin{bmatrix} \cos \varphi & -r \sin \varphi \\ \sin \varphi & r \cos \varphi \end{bmatrix} \right| = |r \cos^2(\varphi) + r \sin^2(\varphi)| = r.$$

Wir erhalten also den Radius der Polarkoordinatendarstellung. Für $r > 0$ ist die Funktionaldeterminante damit von null verschieden (siehe Voraussetzung in Satz 4.16).

Wollen wir die Substitutionsregel mit $\vec{g}$ anwenden, so müssen wir eigentlich $r = 0$ ausschließen. Außerdem ist die Polarkoordinatendarstellung nicht eindeutig für $r = 0$ oder $\varphi = 2\pi$, denn $(0, 0) = (r \cos \varphi, r \sin \varphi)$ gilt für $r = 0$ und jeden Winkel φ . Für $\varphi = 2\pi$ erhält man die gleichen Werte wie für $\varphi = 0$. Da $\vec{g}$ invertierbar sein muss, wäre dies ein weiterer Grund, $r = 0$ und zusätzlich $\varphi = 2\pi$ auszuschließen. Aber Teilmengen des $\mathbb{R}^n$, die in mindestens einer Richtung keine echt positive Ausdehnung haben, liefern keinen Beitrag zum Integral und können vernachlässigt werden. Wir erlauben also auch die Werte $r = 0$ und $\varphi = 2\pi$ und erhalten als Anwendung der Substitutionsregel:

Polarkoordinaten:



Geometrische Kugelkoordinaten: Astronomische Kugelkoordinaten:

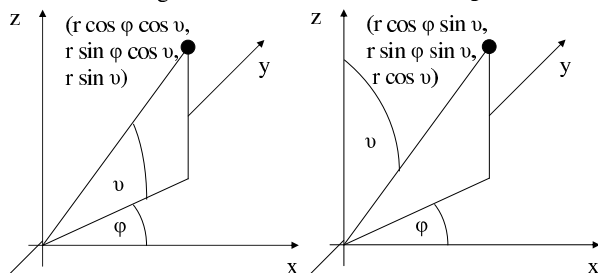


Abb. 4.17: Koordinatentransformationen

Ist $A \subset \mathbb{R}^2$ ein Normalbereich und sei $B := \{(r, \varphi) : r \geq 0, 0 \leq \varphi \leq 2\pi, \text{ und es gibt ein } (x, y) \in A : (x, y) = \vec{g}(r, \varphi)\}$ die entsprechende Menge in Polarkoordinatendarstellung, dann ist

$$\iint_A f(x, y) d(x, y) = \iint_B f(r \cos \varphi, r \sin \varphi) r d(r, \varphi).$$

Die Substitution lässt sich also beschreiben mit

$$x = r \cos \varphi, y = r \sin \varphi \text{ und } d(x, y) = r d(r, \varphi),$$

wobei man zusätzlich den Integrationsbereich überführen muss. Ist beispielsweise $A = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq R^2\}$, so ist $B = \{(r, \varphi) \in \mathbb{R}^2 : 0 \leq r \leq R, 0 \leq \varphi \leq 2\pi\}$.

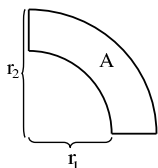


Abb. 4.18: A eignet sich zur Integration mittels Polarkoordinaten.

Beispiel 4.41

Wir integrieren $f(x, y) = \arctan^2\left(\frac{y}{x}\right)$ über den Viertelkreisring (siehe Abbildung 4.18)

$$A := \left\{ (x, y) = r(\cos \varphi, \sin \varphi) : r_1 \leq r \leq r_2, 0 \leq \varphi \leq \frac{\pi}{2} \right\}.$$

Dem Viertelkreisring in der x - y -Ebene entspricht das Rechteck $[r_1, r_2] \times [0, \frac{\pi}{2}]$ in der r - φ -Ebene. Mit der Funktionaldeterminante r ergibt sich

$$\begin{aligned} \iint_A \arctan^2\left(\frac{y}{x}\right) d(x, y) &= \int_0^{\frac{\pi}{2}} \left(\int_{r_1}^{r_2} \left[\arctan\left(\frac{r \cdot \sin \varphi}{r \cdot \cos \varphi}\right) \right]^2 \cdot r dr \right) d\varphi \\ &= \int_0^{\frac{\pi}{2}} \left(\int_{r_1}^{r_2} \varphi^2 \cdot r dr \right) d\varphi = \int_0^{\frac{\pi}{2}} \left[\frac{1}{2} r^2 \cdot \varphi^2 \right]_{r=r_1}^{r=r_2} d\varphi = \frac{\pi^3}{48} (r_2^2 - r_1^2). \end{aligned}$$

Wir sehen uns die Rolle der Funktionaldeterminante an: Das linke Randstück $\{(x, y) = r_1(\cos(\varphi), \sin(\varphi)) : 0 \leq \varphi \leq \pi/2\}$ hat eine Länge $r_1\pi/2$, ein Streifen der (kleinen) Breite ε , der an diesem Rand anliegt, ungefähr den Flächeninhalt $\varepsilon r_1\pi/2$. Durch die Transformation wird das Randstück in die Kante $\{(r_1, \varphi) : 0 \leq \varphi \leq \pi/2\}$ eines Rechtecks überführt, die die Länge $\pi/2$ hat. Ein zugehöriger kleiner Streifen hat den Flächeninhalt $\varepsilon\pi/2$. Um die gleiche Länge bzw. den gleichen Flächeninhalt zu erhalten, muss also mit der Funktionaldeterminante $r = r_1$ multipliziert werden. ■

4.4.4.2 Zylinderkoordinaten

Betrachtet man in $\mathbb{R}^3$ als Integrationsbereich einen Zylinder, so kann man für zwei Variablen x und y die Polarkoordinaten verwenden und eine Variable z untransformiert belassen. Man nennt diese Darstellung **Zylinderkoordinaten**. Ist $\vec{t} = (t_1, t_2, t_3) = (r, \varphi, z)$ ein Punkt in Zylinderkoordinatenschreibweise, so bekommen wir mittels

$$\vec{g}(\vec{t}) = \vec{g}(t_1, t_2, t_3) = \vec{g}(r, \varphi, z) = (r \cos \varphi, r \sin \varphi, z)$$

die Euklid'sche Darstellung des Punktes (x, y, z) zurück.

$$|\det[\vec{g}'(\vec{t})]| = \left| \det \begin{bmatrix} \cos \varphi & -r \sin \varphi & 0 \\ \sin \varphi & r \cos \varphi & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = r.$$

Als Anwendung der Substitutionsregel erhalten wir nun: Ist $A \subset \mathbb{R}^3$ ein Normalbereich und sei $B := \{(r, \varphi, z) : r \geq 0, 0 \leq \varphi \leq 2\pi, \text{ und es gibt ein } (x, y, z) \in A : (x, y, z) = \vec{g}(r, \varphi, z)\}$ die entsprechende Menge in Zylinderkoordinatendarstellung, dann ist

$$\iiint_A f(x, y, z) d(x, y, z) = \iiint_B f(r \cos \varphi, r \sin \varphi, z) r d(r, \varphi, z).$$

Die Substitution lautet also

$$x = r \cos \varphi, y = r \sin \varphi, z = z \text{ und } d(x, y, z) = r d(r, \varphi, z).$$

Beispiel 4.42

Wir berechnen das Volumen des Zylinders $A := \{(x, y, z) : x^2 + y^2 \leq R^2, 0 \leq z \leq H\}$:

$$\int_D 1 d(x, y, z) = \int_0^R \int_0^{2\pi} \int_0^H 1 \cdot r dz d\varphi dr = H 2\pi \frac{1}{2} R^2 = H\pi R^2. \quad \blacksquare$$

Beispiel 4.43

Wie schon zuvor (siehe (2.63) auf Seite 372) drehen wir den Funktionsgraphen zu $y = x^2 + 3$ um die y -Achse und berechnen das Volumen des Rotationskörpers zwischen $y = 3$ und $y = 4$. Dazu integrieren wir über $A := \{(x, y, z) : 3 \leq y \leq 4, \sqrt{x^2 + z^2} \leq \sqrt{y-3}\}$. Dies ist ein „Zylinder“ mit variablem Radius $R(y) = \sqrt{y-3}$:

$$\begin{aligned} \int_D 1 d(x, y, z) &= \int_3^4 \int_0^{2\pi} \int_0^{\sqrt{y-3}} r dr d\varphi dy = 2\pi \int_3^4 \left[\frac{1}{2} r^2 \right]_{r=0}^{r=\sqrt{y-3}} dy \\ &= \pi \int_3^4 y - 3 dy = \pi \left[\frac{y^2}{2} - 3y \right]_3^4 = \frac{\pi}{2}. \quad \blacksquare \end{aligned}$$

4.4.4.3 Kugelkoordinaten

Von den Längen- und Breitengraden kennen wir die Kugelkoordinaten. Jeder Punkt der Erdoberfläche ist eindeutig charakterisiert über zwei Winkel. Der nullte Breitengrad ist der Äquator. Betrachtet man einen Strahl, der im Erdmittelpunkt beginnt und der einen Winkel $0 \leq \vartheta \leq 90^\circ$ zur Äquatorebene nach oben hat, so schneidet er die Erdoberfläche auf dem nördlichen ϑ -ten Breitengrad (siehe Abbildung 4.19). Entsprechend erhält man die südlichen Breitengrade über negative Winkel ϑ . Der nullte Längengrad verläuft durch Greenwich bei London. Von dort ausgehend ist die östliche und westliche Erdhälfte in Längengrade zu Winkeln $0 \leq \varphi \leq 180^\circ$ unterteilt. Man kann ohne die Ost-West-Unterteilung auch Winkel $0 \leq \varphi < 360^\circ$ verwenden. Die Winkel φ und ϑ bilden dann die Kugelkoordinaten für die Erdoberfläche. Lässt man statt des Erdradius beliebige Radien r zu, so kann man jeden Punkt (x, y, z) des $\mathbb{R}^3$ beschreiben mittels der geometrischen Kugelkoordinaten

$$(r \cos(\varphi) \cos(\vartheta), r \sin(\varphi) \cos(\vartheta), r \sin(\vartheta))$$

mit $r \geq 0$, $0 \leq \varphi \leq 2\pi$, $-\frac{\pi}{2} \leq \vartheta \leq \frac{\pi}{2}$. Diese Darstellung ist eindeutig für $r > 0$, $0 \leq \varphi < 2\pi$ und $-\frac{\pi}{2} < \vartheta < \frac{\pi}{2}$. Wie zuvor erlauben wir aber auch $r = 0$, $\varphi = 2\pi$ und jetzt $\vartheta = \pm \frac{\pi}{2}$. Ist $\vec{t} = (t_1, t_2, t_3) = (r, \varphi, \vartheta)$ die Darstellung eines Punktes in Kugelkoordinaten, so erhalten wir die Euklid'sche Darstellung des Punktes mittels

$$\vec{g}(\vec{t}) = \vec{g}(r, \varphi, \vartheta) = (r \cos(\varphi) \cos(\vartheta), r \sin(\varphi) \cos(\vartheta), r \sin(\vartheta)).$$

Als Entwicklung der Determinante nach der letzten Zeile unter Ausnutzung der trigonometrischen Form des Satzes von Pythagoras (1.20) auf Seite 118 erhalten wir

$$\begin{aligned} |\det[\vec{g}'(\vec{t})]| &= \left| \det \begin{bmatrix} \cos(\varphi) \cos(\vartheta) & -r \sin(\varphi) \cos(\vartheta) & -r \cos(\varphi) \sin(\vartheta) \\ \sin(\varphi) \cos(\vartheta) & r \cos(\varphi) \cos(\vartheta) & -r \sin(\varphi) \sin(\vartheta) \\ \sin(\vartheta) & 0 & r \cos(\vartheta) \end{bmatrix} \right| \\ &= |\sin(\vartheta)[r^2 \sin^2(\varphi) \cos(\vartheta) \sin(\vartheta) + r^2 \cos^2(\varphi) \cos(\vartheta) \sin(\vartheta)] \\ &\quad + r \cos(\vartheta)[r \cos^2(\varphi) \cos^2(\vartheta) + r \sin^2(\varphi) \cos^2(\vartheta)]| \\ &= |r^2 \sin^2(\vartheta) \cos(\vartheta) + r^2 \cos^3(\vartheta)| = r^2 \cos(\vartheta). \end{aligned}$$

Für $r > 0$ und $-\frac{\pi}{2} < \vartheta < \frac{\pi}{2}$ ist auch diese Funktionaldeterminante ungleich null (wie in Satz 4.16 gefordert), und wir erhalten mit der Substitutionsregel:

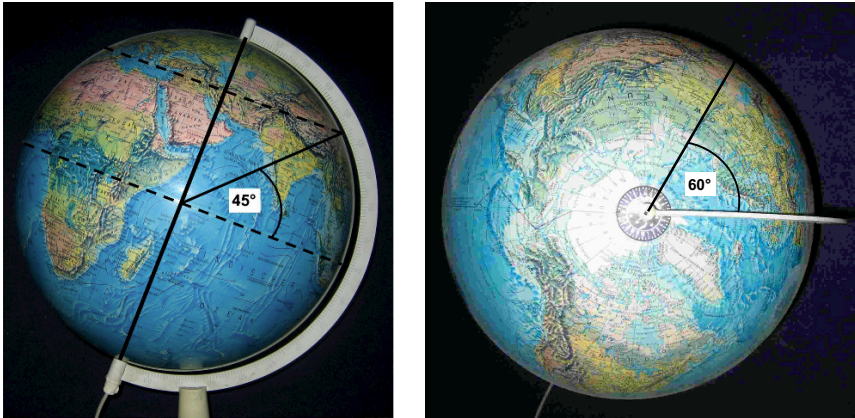


Abb. 4.19: Links: Breitengrad zum Winkel $\vartheta = 45^\circ$, rechts: Längengrad zum Winkel $\varphi = 60^\circ$

Ist $A \subset \mathbb{R}^3$ ein Normalbereich und sei $B := \{(r, \varphi, \vartheta) : r \geq 0, 0 \leq \varphi \leq 2\pi, -\frac{\pi}{2} \leq \vartheta \leq \frac{\pi}{2}, \text{ und es gibt ein } (x, y, z) \in A : (x, y, z) = \vec{g}(r, \varphi, \vartheta)\}$ die entsprechende Menge in Kugelkoordinatendarstellung, dann ist

$$\iiint_A f(x, y, z) d(x, y, z) = \iiint_B f(r \cos(\varphi) \cos(\vartheta), r \sin(\varphi) \cos(\vartheta), r \sin(\vartheta)) r^2 \cos(\vartheta) d(r, \varphi, \vartheta).$$

Die Substitution lautet also

$$x = r \cos(\varphi) \cos(\vartheta), y = r \sin(\varphi) \cos(\vartheta), z = r \sin(\vartheta) \text{ und } d(x, y, z) = r^2 \cos \vartheta d(r, \varphi, \vartheta).$$

Wieder müssen die Integrationsbereiche passend sein. Ist beispielsweise $A = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq R^2, z \geq 0\}$, so ist $B = \{(r, \varphi, \vartheta) \in \mathbb{R}^3 : 0 \leq r \leq R, 0 \leq \varphi \leq 2\pi, 0 \leq \vartheta \leq \frac{\pi}{2}\}$.

Achtung: Bei geometrischen Kugelkoordinaten ist ϑ der Winkel zwischen $(x, y, 0)$ und (x, y, z) . Dagegen verwendet man bei astronomischen Kugelkoordinaten den Winkel zwischen (x, y, z) und $(0, 0, z)$, siehe Abbildung 4.17. Dann betrachtet man

$$(r \cos(\varphi) \sin(\vartheta), r \sin(\varphi) \sin(\vartheta), r \cos(\vartheta)), r \geq 0, 0 \leq \varphi \leq 2\pi, 0 \leq \vartheta \leq \pi,$$

und erhält als Funktionaldeterminante $r^2 \sin \vartheta$.

Beispiel 4.44

Durch den Schnitt der Kugel mit Radius $R > 0$ um den Nullpunkt mit dem ersten Oktanten $\{(x, y, z) : x \geq 0, y \geq 0, z \geq 0\}$ erhält man die Achtelkugel A . Wir bestimmen den Schwerpunkt von A , wobei wir konstante Dichte $\rho = 1$ annehmen. Zunächst berechnen wir mit Hilfe von Kugelkoordinaten (r, φ, ϑ) die Masse:

$$\begin{aligned} M &= \iiint_A 1 d(x, y, z) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \int_0^R r^2 \cos \vartheta dr d\varphi d\vartheta \\ &= \int_0^{\frac{\pi}{2}} \cos \vartheta d\vartheta \cdot \int_0^{\frac{\pi}{2}} 1 d\varphi \cdot \int_0^R r^2 dr = [\sin \vartheta]_0^{\frac{\pi}{2}} \cdot \frac{\pi}{2} \cdot \frac{R^3}{3} = \frac{\pi}{6} R^3. \end{aligned}$$

Für die erste Koordinate des Schwerpunkts $\vec{x}_s = (x_s, y_s, z_s)^\top$ erhalten wir:

$$\begin{aligned} x_s &= \frac{1}{M} \iiint_A x d(x, y, z) = \frac{1}{M} \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \int_0^R r \cdot \cos(\varphi) \cos(\vartheta) \cdot r^2 \cos(\vartheta) dr d\varphi d\vartheta \\ &= \frac{1}{M} \int_0^{\frac{\pi}{2}} \cos^2 \vartheta d\vartheta \int_0^{\frac{\pi}{2}} \cos \varphi d\varphi \int_0^R r^3 dr = \frac{1}{M} \int_0^{\frac{\pi}{2}} \cos^2 \vartheta d\vartheta \cdot 1 \cdot \frac{R^4}{4} \\ &\stackrel{(2.52)}{=} \frac{1}{M} \cdot \frac{1}{4} R^4 \left[\frac{1}{2} \vartheta + \frac{1}{2} \cos(\vartheta) \sin(\vartheta) \right]_{\vartheta=0}^{\frac{\pi}{2}} = \frac{1}{M} \cdot \frac{1}{4} R^4 \cdot \frac{\pi}{4} = \frac{3}{8} R. \end{aligned}$$

Entsprechend erhalten wir für die weiteren Koordinaten $y_s = z_s = \frac{3}{8} R$. ■

4.4.5 Lebesgue-Integral, L^p - und Sobolev-Räume*

4.4.5.1 Lebesgue-Integral

Völlig analog zum Aufbau des Lebesgue-Integrals für Funktionen mit einer Variable in Kapitel 2.6.8 wird auch das Lebesgue-Integral für Funktionen mit mehreren Variablen definiert. Statt der eindimensionalen Intervalle $]a, b[$ werden dabei mehrdimensionale

Intervalle $]a_1, b_1[\times]a_2, b_2[\times \cdots \times]a_n, b_n[$ mit dem elementaren Inhalt $(b_1 - a_1)(b_2 - a_2) \cdots (b_n - a_n)$ verwendet. Damit erhält man die Lebesgue- σ -Algebra der Lebesgue-messbaren Teilmengen des $\mathbb{R}^n$. Wie zuvor heißt eine Funktion $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ **Lebesgue-messbar** genau dann, wenn D in dieser σ -Algebra liegt und das ebenso für alle Mengen $\{\vec{x} \in \mathbb{R}^n : f(\vec{x}) \geq a\}$ gilt. Das Lebesgue-Integral wird damit exakt so über einfache Funktionen aufgebaut wie in Kapitel 2.6.8.3 beschrieben. Auch die Eigenschaften des Integrals gelten unverändert.

Wie beim Riemann-Integral lassen sich Integrale über Teilmengen des $\mathbb{R}^n$ in einzelne, nacheinander zu berechnende Integrale zerlegen:

Satz 4.17 (Satz von Fubini für Lebesgue-Integrale)

Seien $A \subset \mathbb{R}^n$ und $B \subset \mathbb{R}^m$ Lebesgue-messbare Mengen. Dann ist auch $A \times B \in \mathbb{R}^{n+m}$ Lebesgue-messbar. Weiter sei $f : A \times B \rightarrow \mathbb{R} \cup \{\pm\infty\}$ eine messbare Funktion, die auch die „Funktionswerte“ $\pm\infty$ annehmen darf.

Ist f Lebesgue-integrierbar auf $A \times B$, dann ist $f(\vec{x}, \vec{y})$ für fast alle $\vec{x} \in A$ eine Lebesgue-integrierbare Funktion auf B in der Variablen $\vec{y}$. Außerdem ist dann $\int_B f(\vec{x}, \vec{y}) d\vec{y}$ Lebesgue-integrierbar bezüglich der Variablen $\vec{x}$ auf A , und es gilt:

$$L\text{-} \int_{A \times B} f(\vec{x}, \vec{y}) d(\vec{x}, \vec{y}) = L\text{-} \int_A \left[L\text{-} \int_B f(\vec{x}, \vec{y}) d\vec{y} \right] d\vec{x}.$$

Auch darf unter den Voraussetzungen des Satzes (analog zu den Riemann-Integralen) die Reihenfolge der Integrale beliebig gewählt werden.

4.4.5.2 L^p - und Sobolev-Räume

Ebenfalls analog zu den Lebesgue- und Sobolev-Räumen für Funktionen auf reellen Intervallen (siehe Abschnitte 3.7.3 und 3.7.6) können wir diese Räume auch für Funktionen auf Teilmengen des $\mathbb{R}^n$ definieren. Für eine messbare Menge $\Omega \subset \mathbb{R}^n$ ist für $1 \leq p \leq \infty$ der L^p -Raum definiert über

$$L^p(\Omega) := \{f : \Omega \rightarrow \mathbb{R} \cup \{-\infty, \infty\} : f \text{ ist Lebesgue-messbar und } \|f\|_{L^p(\Omega)} < \infty\}$$

mit der Norm

$$\|f\|_{L^p(\Omega)} := \left[\int_{\Omega} |f(\vec{x})|^p d\vec{x} \right]^{1/p}$$

für $1 \leq p < \infty$. Ist $p = \infty$, so wird mit $\|\cdot\|_{L^\infty(\Omega)}$ das wesentliche Supremum auf Ω bezeichnet. Die Definition ist analog zum eindimensionalen Fall auf Seite 563.

$L^2(\Omega)$ ist ein Hilbert-Raum mit Skalarprodukt $(f, g)_{L^2(\Omega)} := \int_{\Omega} f(\vec{x})g(\vec{x}) d\vec{x}$. Im Folgenden sei $\Omega \subset \mathbb{R}^n$ offen, so dass wir keine Probleme mit der Definition der parti-

ellen Ableitung am Rand von Ω bekommen. Außerdem benötigen wir den Begriff der kompakten Menge $K \subset \mathbb{R}^n$.

Definition 4.18 (Kompakte Menge)

Eine Menge $K \subset \mathbb{R}^n$ heißt **abgeschlossen** genau dann, wenn ihr Komplement offen ist, also nur aus inneren Punkten besteht. Sie heißt **kompakt** genau dann, wenn sie abgeschlossen und beschränkt ist.

Zum Beispiel sind Intervalle $[\vec{a}, \vec{b}] = [a_1, b_1] \times \cdots \times [a_n, b_n]$ kompakt. Man kann zeigen, dass der Grenzwert jeder konvergenten Folge, deren Glieder in einer abgeschlossenen Menge liegen, ebenfalls in dieser Menge liegt. Hat man gar eine Folge mit Gliedern aus einer kompakten Menge, dann hat sie wegen der Beschränktheit eine konvergente Teilfolge – und ihr Grenzwert liegt in der kompakten Menge.

Die Sobolev-Räume basieren wie in einer Dimension auf dem Begriff der schwachen Ableitung. Dazu sei $L_{1,\text{loc}}(\Omega)$ die Menge der auf einer messbaren Menge $\Omega \subset \mathbb{R}^n$ definierten messbaren Funktionen $f : \Omega \rightarrow \mathbb{R}$, die lokal integrierbar sind, d. h.: $f \in L_1(K)$ für jede kompakte Menge $K \subset \Omega$.

Definition 4.19 (schwache Ableitung)

Seien $\Omega \subset \mathbb{R}^n$ messbar, $f \in L_{1,\text{loc}}(\Omega)$, $\alpha \in \mathbb{N}_0^n$ und $a := \sum_{k=1}^n \alpha_k$. Falls es ein $g \in L_{1,\text{loc}}(\Omega)$ gibt mit

$$\int_{\Omega} g(\vec{x}) \varphi(\vec{x}) d\vec{x} = (-1)^a \int_{\Omega} f(\vec{x}) \frac{\partial^a}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} \varphi(\vec{x}) d\vec{x}$$

für alle unendlich oft partiell nach allen Variablen differenzierbaren Testfunktionen φ , die außerhalb einer kompakten Menge $K \subset \Omega$ gleich der Nullfunktion sind, so heißt g die **schwache Ableitung** der Ordnung α von f . Wir führen keine neue Bezeichnung ein, sondern schreiben auch hier: $g = \frac{\partial^a}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f$.

Da die Testfunktionen außerhalb einer kompakten Menge gleich null sind, muss man dort auch nichts über f wissen. Daher wird nicht die Lebesgue-Integrierbarkeit von f auf Ω gefordert, sondern nur lokale Integrierbarkeit.

Die Notation ist sinnvoll: Ist f entsprechend oft stetig differenzierbar, so stimmen die klassischen partiellen Ableitungen mit den schwachen überein.

Definition 4.20 (Sobolev-Räume)

Seien $\Omega \subset \mathbb{R}^n$ messbar, $s \in \mathbb{N}_0$ und $1 \leq p \leq \infty$. Die Räume

$$W^{s,p}(\Omega) := \left\{ f \in L^p(\Omega) : \frac{\partial^a}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f \in L^p(\Omega) \text{ existiert für jedes } \alpha \in \mathbb{N}_0^n \right\}$$

$$\text{mit } a := \sum_{k=1}^n \alpha_k \leq s \text{ im Sinne von Definition 4.19}\},$$

versehen mit der über die **Halbnormen** ($0 \leq i \leq s$)

$$|f|_{i,p,\Omega} := \begin{cases} \left(\sum_{\alpha \in \mathbb{N}_0^n, \sum_{k=1}^n \alpha_k = i} \left\| \frac{\partial^i}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f \right\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}} & \text{für } 1 \leq p < \infty \\ \sum_{\alpha \in \mathbb{N}_0^n, \sum_{k=1}^n \alpha_k = i} \left\| \frac{\partial^i}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f \right\|_{L^\infty(\Omega)} & \text{für } p = \infty \end{cases}$$

definierten Norm (die entsprechenden Eigenschaften lassen sich beweisen)

$$\|f\|_{s,p,\Omega} := \begin{cases} \left(\sum_{0 \leq i \leq s} |f|_{i,p,\Omega}^p \right)^{\frac{1}{p}} & \text{für } 1 \leq p < \infty \\ \sum_{0 \leq i \leq s} |f|_{i,\infty,\Omega} & \text{für } p = \infty, \end{cases} \quad (4.17)$$

heißen **Sobolev-Räume**. Weiter benutzen wir die Schreibweise ($\alpha \in \mathbb{N}_0^n$)

$$|f|_{\alpha,p,\Omega} := \left\| \frac{\partial^{\sum_{k=1}^n \alpha_k}}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f \right\|_{L^p(\Omega)}.$$

Es sind durchaus auch geringfügig andere Definitionen der Sobolev-(Halb-)Normen üblich, die aber mit der hier benutzten äquivalent sind. So existiert z. B. ein $c > 0$, so dass für $1 \leq p < \infty$ gilt:

$$c \sum_{\alpha \in \mathbb{N}_0^n, \sum_{k=1}^n \alpha_k = i} \left\| \frac{\partial^i}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f \right\|_{L^p(\Omega)} \leq |f|_{i,p,\Omega} \leq \sum_{\alpha \in \mathbb{N}_0^n, \sum_{k=1}^n \alpha_k = i} \left\| \frac{\partial^i}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f \right\|_{L^p(\Omega)}.$$

Sobolev-Räume werden häufig benutzt, um partielle Differenzialgleichungen mit Randbedingungen in einer abgeschwächten und damit besser lösbaren Form zu beschreiben. Bei partiellen Differenzialgleichungen sind Funktionen gesucht, deren partielle Ableitungen vorgegebene Gleichungen auf einem Gebiet des $\mathbb{R}^n$ erfüllen müssen, vgl. Kapitel 5. Dazu kommen meist noch Bedingungen auf dem Rand des Gebiets, die ebenfalls erfüllt sein müssen. Zur Darstellung einfacher Randbedingungen benötigt man zu einem Sobolev-Raum einen Untervektorraum, in dem die Funktionen auf dem Rand gleich null sind. Da aber in den L^p -Räumen die Funktionen nur bis auf eine Menge vom Lebesgue-Maß null eindeutig sind, ist es zwecklos zu sagen, dass $f(x) = 0$ für alle x aus dem Rand $\partial\Omega$ von Ω sein soll, da der Rand das Lebesgue-Maß null hat. Vielmehr definiert man als $W_0^{s,p}(\Omega)$ den Untervektorraum von $W^{s,p}(\Omega)$, der entsteht, wenn man alle unendlich oft differenzierbaren Testfunktionen φ aufnimmt, die außerhalb einer kompakten Menge $K \subset \Omega$ gleich null sind, und dann diese Menge vervollständigt. Das heißt, man wählt den „kleinsten“ Untervektorraum, in dem diese

Testfunktionen und auch alle bezüglich $W^{s,p}$ gebildeten Grenzwerte aller konvergenten Folgen aus Funktionen des Untervektorraums liegen.

Wenn Sobolev-Räume über nicht-offene Mengen Ω gebildet werden, dann verstehen wir darunter den entsprechenden Sobolev-Raum über das Innere der Menge.

Aus den Definitionen folgt, dass zu $f \in W^{s,p}(\Omega)$ und $\alpha \in \mathbb{N}_0^n$ mit $\sum_{k=1}^n \alpha_k \leq s$ die Funktion $\frac{\partial^{\sum_{k=1}^n \alpha_k}}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f$ zu $W^{s-\sum_{k=1}^n \alpha_k, p}(\Omega)$ gehört. Das folgende Lemma findet man beispielsweise in (Adams, 1975, S. 45, 47, 54, 52, 159):

Lemma 4.3 (Eigenschaften von Sobolev-Räumen)

Seien $\Omega \subset \mathbb{R}^n$ messbar, $s \in \mathbb{N}_0$ und $1 \leq p \leq \infty$.

- Die Räume $W^{s,p}(\Omega)$ sind Banach-Räume. Insbesondere gilt dies auch für $W_0^{s,p}(\Omega)$ und für die Räume $W_0^{s,p}(\Omega) \cap W^{k,p}(\Omega)$, $k > s$.
- Die Räume $W^{s,2}(\Omega)$ und $W_0^{s,2}(\Omega)$ sind Hilbert-Räume mit dem Skalarprodukt

$$(f, g)_{s,2,\Omega} := \sum_{\alpha \in \mathbb{N}_0^n, \sum_{k=1}^n \alpha_k \leq s} \left(\frac{\partial^{\sum_{k=1}^n \alpha_k}}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f, \frac{\partial^{\sum_{k=1}^n \alpha_k}}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} g \right)_{L^2(\Omega)}.$$

- **Satz von Meyers und Serrin:** Die s -mal stetig partiell auf Ω differenzierbaren Funktionen liegen dicht in $W^{s,p}(\Omega)$, d. h., man findet zu jeder Funktion f aus $W^{s,p}(\Omega)$ und jedem $\varepsilon > 0$ eine s -mal stetig partiell differenzierbare Funktion φ , so dass $\|f - \varphi\|_{s,p,\Omega} < \varepsilon$.
- **Poincaré-Ungleichung:** Sei Ω beschränkt und $p < \infty$. Dann existiert eine Konstante C , so dass

$$|f|_{s,p,\Omega} \leq \|f\|_{s,p,\Omega} \leq C|f|_{s,p,\Omega} \text{ für alle } f \in W_0^{s,p}(\Omega).$$

Die Halbnorm ist äquivalent zur Norm. Das gilt aber nicht allgemein für Funktionen aus $W^{s,p}(\Omega)$, sondern nur wenn die Funktionen am Rand zu null werden im Sinne der Zugehörigkeit zu $W_0^{s,p}(\Omega)$.

4.5 Vektoranalysis

In der Physik hat man es häufig mit Feldern zu tun. Ein Feld ordnet einem Punkt des Raums einen Wert zu, ist also eine Funktion. Der zugeordnete Wert kann eine reelle Zahl sein (z. B. beim elektrischen Potenzial). Man spricht dann von einem **Skalarfeld**. Der Wert kann aber auch ein Vektor sein, der z. B. eine Kraft beschreibt, die auf eine elektrische Ladung oder einen Körper an der entsprechenden Stelle wirkt. In diesem Fall spricht man von einem **Vektorfeld**. In einem solchen Vektorfeld kann es Quellen,

Senken und Wirbel geben. Diese anschaulichen Phänomene lassen sich mit partiellen Ableitungen beschreiben.

4.5.1 Vektorfelder

Definition 4.21 (Vektorfeld)

Sei $\vec{V} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ eine vektorwertige Funktion. $\vec{V}$ heißt (auch) **Vektorfeld**. Man nennt zur Abgrenzung Funktionen $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ (auch) **skalare Funktionen**.

Anschaulich wird der Begriff des Vektorfelds bei strömenden Flüssigkeiten. Hier kann man jedem Punkt eine Geschwindigkeit zuordnen. Das ist ein Vektorpfeil, der in Richtung der Strömung zeigt und dessen Länge (Betrag) die Geschwindigkeit angibt.

Beispiel 4.45 (Feldstärke)

Elektrische Ladungen üben aufeinander eine Kraft aus, sie erzeugen ein elektrisches Feld. Die elektrische Feldstärke $\vec{E}$ ist die Kraft $\vec{F}$ pro Probeladung q , die in einem elektrischen Feld auf eine sehr kleine positive Probeladung q ausgeübt wird. Die Probeladung muss sehr klein sein, damit sie das Feld nicht beeinflusst. Man definiert daher an einer Stelle (x, y, z) mit der Kraft $\vec{F}(q)$, die an dieser Stelle auf q wirkt:

$$\vec{E}(x, y, z) = \lim_{q \rightarrow 0+} \frac{1}{q} \vec{F}(q).$$

$\vec{E}$ ist ein Vektorfeld. Eine positive Punktladung Q erzeugt ein elektrisches Feld, dessen Funktionswerte als Vektoren radial vom Punkt der Ladung wegzeigen. In einem Punkt P , der den Abstand r von der Punktladung hat, ist $|\vec{E}| = \frac{Q}{4\pi\epsilon_0 r}$, wobei ϵ_0 eine elektrische Feldkonstante ist. ■

Beispiel 4.46 (Optischer Fluss)

Der optische Fluss ist (für einen festen Zeitpunkt t) ein Vektorfeld von Geschwindigkeitsvektoren, das in der Analyse einer Videosequenz verwendet wird, um die Bewegung von Pixeln von einem Bild zum nächsten zu beschreiben, siehe Abbildung 4.20. Die Funktion $G(x, y, t) : [a, b] \times [c, d] \times [0, \infty[\rightarrow [0, 1]$ möge den Grauwert des Bildes zum Zeitpunkt t an der Bildposition (x, y) beschreiben. Unter der Annahme, dass der Gradient von G existiert, gilt zum festen Zeitpunkt t für den optischen Fluss $V(x, y) = (V_1(x, y), V_2(x, y))$ die **Horn-Schunck-Gleichung**

$$\frac{\partial G}{\partial x}(x, y, t) \cdot V_1(x, y) + \frac{\partial G}{\partial y}(x, y, t) \cdot V_2(x, y) + \frac{\partial G}{\partial t}(x, y, t) = 0.$$

Durch diese lineare Gleichung mit den zwei Unbekannten $V_1(x, y)$ und $V_2(x, y)$ ist das Vektorfeld leider noch nicht eindeutig beschrieben, und man benötigt zusätzliche Bedingungen wie z. B., dass der optische Fluss lokal nahezu in die gleiche Richtung zeigt. ■

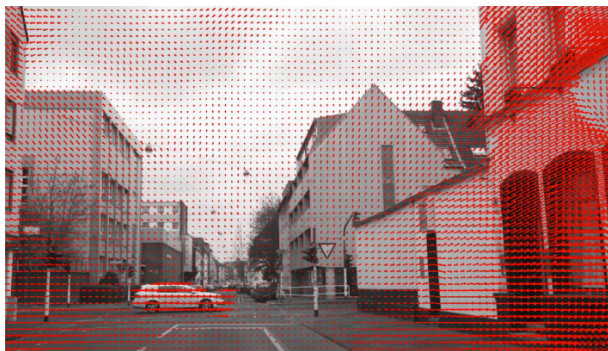


Abb. 4.20: Die roten Pfeile beschreiben den optischen Fluss von diesem Bild eines Videos zum nächsten. Sie definieren damit ein Vektorfeld.

4.5.2 Kurven

Felder werden häufig über Feldlinien beschrieben. Dazu benötigen wir Kurven.

Definition 4.22 (Parametrisierte Kurve)

Sei $[\alpha, \beta] \subset \mathbb{R}$ und $\vec{x}(t) = (x_1(t), x_2(t), \dots, x_n(t)) : [\alpha, \beta] \rightarrow \mathbb{R}^n$ eine stetige, vektorwertige Funktion mit einer reellen Variable. Eine **Kurve** $K \subset \mathbb{R}^n$ ist eine (orientierte) Punktmenge $K = \{\vec{x}(t) : t \in [\alpha, \beta]\}$. Die Orientierung (Durchlaufrichtung) entsteht, indem t von α nach β läuft.

- $(\vec{x}(t), [\alpha, \beta])$ heißt **Parameterdarstellung** der Kurve, $\vec{x}(\alpha)$ heißt **Anfangspunkt**, und $\vec{x}(\beta)$ heißt **Endpunkt** der Kurve.
- Stimmen Anfangs- und Endpunkt überein, heißt die Kurve **geschlossen**.
- K heißt genau dann **einfach**, wenn mit $\vec{x}(t)$ für $t \in [\alpha, \beta[$ kein Punkt mehrfach durchlaufen wird, wenn also $\vec{x}(t)$ auf $[\alpha, \beta[$ injektiv ist.
- Eine einfache, geschlossene Kurve heißt **Jordan-Kurve**.
- Eine Kurve heißt **(stückweise) stetig differenzierbar** genau dann, wenn die Komponenten $x_1(t), \dots, x_n(t)$ (stückweise) stetig differenzierbar sind.

Beispiel 4.47

- a) $(\vec{x}(t) := (\cos t, \sin t), [0, 2\pi])$ beschreibt den Einheitskreis um den Nullpunkt und ist eine Jordan-Kurve.
- b) Sei $f : [\alpha, \beta] \rightarrow \mathbb{R}$ stetig. Dann beschreibt die Kurve $(\vec{x}(t) := (t, f(t)), [\alpha, \beta])$ den Funktionsgraphen von f vom Anfangspunkt $(\alpha, f(\alpha))$ zum Endpunkt $(\beta, f(\beta))$ und ist eine einfache Kurve.

■

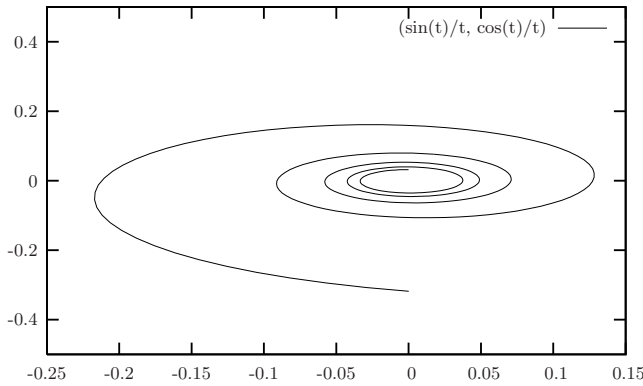


Abb. 4.21: Eine Kurve $\left(\left(\frac{\sin t}{t}, \frac{\cos t}{t}\right), [\pi, 10\pi]\right)$

Eine Kurve mit $\vec{x}(t) = (x_1(t), x_2(t)) : [\alpha, \beta] \rightarrow \mathbb{R}^2$ kann auch als komplexwertige Funktion $f(t) := x_1(t) + j \cdot x_2(t)$ aufgefasst werden und ist damit z. B. der Fourier-Analyse zugänglich, die wir in Kapitel 6 behandeln. Aus der Kurve mit Parameterdarstellung $(\vec{x}(t) := (\cos t, \sin t), [0, 2\pi])$ wird so $f : [0, 2\pi] \rightarrow \mathbb{C}$ mit $f(\varphi) := e^{j\varphi}$.

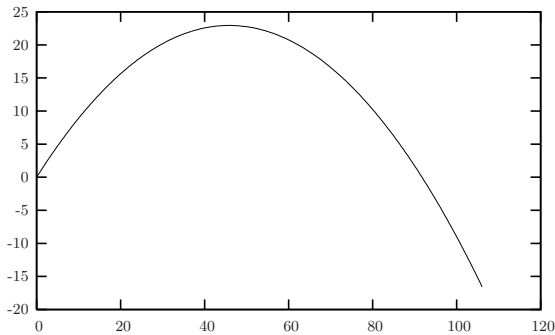


Abb. 4.22: Eine Wurfparabel $\left((v_0 \cos \alpha)t, (v_0 \sin \alpha)t - \frac{a}{2}t^2\right), [0, 5]$ mit $v_0 = 30$ und $\alpha = \pi/4$

Beispiel 4.48

Wir beschreiben eine Wurfparabel als Kurve (siehe Abbildung 4.22). An der Stelle $(0, 0)$ wird im Winkel α ein Ball mit der Anfangsgeschwindigkeit v_0 geworfen. In x -Richtung bewegt er sich (unter Vernachlässigung des Luftwiderstands) mit konstanter Geschwindigkeit $v_0 \cos \alpha$. Nach t Sekunden befindet er sich bei $x(t) = (v_0 \cos \alpha)t$.

In y -Richtung ist die Anfangsgeschwindigkeit $v_0 \sin \alpha$. Allerdings wird diese Bewegung durch den freien Fall des Balls überlagert. Die Beschleunigung ist $-a \approx -9,81 \frac{\text{m}}{\text{s}^2}$, so dass nach t Sekunden die Geschwindigkeit $-at$ addiert werden muss: $v_0 \sin \alpha - at$. Damit ist $y(t) = \int_0^t v_0 \sin \alpha - au \, du = (v_0 \sin \alpha)t - \frac{a}{2}t^2$. Damit erhalten wir für den Wurf bis zum Zeitpunkt T die Kurve $((x(t), y(t)), [0, T])$. ■

Definition 4.23 (Tangente an eine Kurve)

Sei K eine stetig differenzierbare Kurve mit der Parameterdarstellung $(\vec{x}(t), [\alpha, \beta])$ und $\vec{x}'(t) \neq \vec{0}$ für alle $t \in [\alpha, \beta]$. Die Funktion $T : [\alpha, \beta] \rightarrow \mathbb{R}^n$ mit

$$T(t) := \frac{(x'_1(t), x'_2(t), \dots, x'_n(t))}{|(x'_1(t), x'_2(t), \dots, x'_n(t))|} = \frac{(x'_1(t), x'_2(t), \dots, x'_n(t))}{\sqrt{(x'_1(t))^2 + (x'_2(t))^2 + \dots + (x'_n(t))^2}}$$

heißt **Tangenteneinheitsvektor**. Die Gerade

$$\{\vec{y} \in \mathbb{R}^n : \vec{y} = \vec{x}(t) + rT(t), r \in \mathbb{R}\}$$

heißt die **Tangente** an K im Punkt $\vec{x}(t)$.

$T(t)$ ist also ein Vektor, der parallel zur Tangente an die Kurve im Punkt $\vec{x}(t)$ verläuft, so dass man mit ihm die Tangente beschreiben kann.

Beispiel 4.49

- a) Wir betrachten wieder den Funktionsgraphen einer differenzierbaren Funktion $f : [\alpha, \beta] \rightarrow \mathbb{R}$ als Kurve mit Parametrisierung $((t, f(t)), [\alpha, \beta])$. Dann wird über $T(t) = \frac{(1, f'(t))}{\sqrt{1+(f'(t))^2}}$ mittels $(t, f(t)) + rT(t)$, $r \in \mathbb{R}$, die Tangente an den Funktionsgraphen an der Stelle t berechnet.
- b) Der Tangenteneinheitsvektor des Kreises $(\vec{x}(t) := (\cos t, \sin t), [0, 2\pi])$ ist

$$T(t) = \frac{(-\sin t, \cos t)}{\sqrt{(-\sin t)^2 + \cos^2 t}} = (-\sin t, \cos t). \quad \blacksquare$$

Definition 4.24 (Kurvenlänge)

Sei K eine stetig differenzierbare Kurve mit $(\vec{x}(t), [\alpha, \beta])$ als Parameterdarstellung. Der Zahlenwert

$$\int_{\alpha}^{\beta} \sqrt{\sum_{k=1}^n [x'_k(t)]^2} dt$$

heißt die **(Bogen-)Länge** der Kurve K .

Diese Länge stimmt mit dem intuitiven Verständnis der Kurvenlänge überein. Das sieht man, indem man die Kurve in kleine Streckenzüge zerlegt und deren Längen aufsummiert. Sei $Z = \{\alpha = t_0, t_1, t_2, \dots, t_m = \beta\}$ eine Zerlegung des Intervalls $[\alpha, \beta]$ (siehe Kapitel 2.6). Die Punkte $\vec{x}(t_k)$ werden durch Strecken verbunden. Der Abstand a_i (Streckenlänge) der Punkte t_i und t_{i+1} ist

$$a_i = \sqrt{[x_1(t_{i+1}) - x_1(t_i)]^2 + \dots + [x_n(t_{i+1}) - x_n(t_i)]^2}.$$

Nach Mittelwertsatz (Seite 315) gibt es aber Stellen $\xi_{i,1}, \dots, \xi_{i,n} \in]t_i, t_{i+1}[$ mit

$$a_i = \sqrt{[x'_1(\xi_{i,1})(t_{i+1} - t_i)]^2 + \dots + [x'_n(\xi_{i,n})(t_{i+1} - t_i)]^2}$$

$$= (t_{i+1} - t_i) \sqrt{[x'_1(\xi_{i,1})]^2 + \cdots + [x'_n(\xi_{i,n})]^2}.$$

Als Gesamtlänge des Streckenzugs erhalten wir

$$\sum_{i=0}^{m-1} (t_{i+1} - t_i) \sqrt{[x'_1(\xi_{i,1})]^2 + \cdots + [x'_n(\xi_{i,n})]^2} = \sum_{i=0}^{m-1} (t_{i+1} - t_i) \sqrt{\sum_{k=1}^n [x'_k(\xi_{i,k})]^2}.$$

Aufgrund der Stetigkeit der Ableitungen ist dies ungefähr gleich

$$\sum_{i=0}^{m-1} (t_{i+1} - t_i) \sqrt{\sum_{k=1}^n [x'_k(t_i)]^2}$$

und damit gleich einer Riemann-Zwischensumme des Integrals aus der Definition 4.24, die für $m \rightarrow \infty$ gegen das Integral strebt.

Beispiel 4.50

Wir berechnen die Bogenlänge des Einheitskreises $(\vec{x}(t) := (\cos t, \sin t), [0, 2\pi])$, also den Umfang des Einheitskreises:

$$\int_0^{2\pi} \sqrt{(-\sin t)^2 + \cos^2 t} dt = \int_0^{2\pi} 1 dt = 2\pi. \quad \blacksquare$$

Wir haben die Länge nur für stetig differenzierbare Kurven im $\mathbb{R}^n$ mit Parameterdarstellung $(\vec{x}(t), [\alpha, \beta])$ eingeführt. Auf die Differenzierbarkeit kann man verzichten, allerdings reicht Stetigkeit alleine nicht aus.

Definition 4.25 (Rektifizierbare Kurve)

Sei $[\alpha, \beta] \subset \mathbb{R}$ und $\vec{x}(t) = (x_1(t), x_2(t), \dots, x_n(t)) : [\alpha, \beta] \rightarrow \mathbb{R}^n$ eine stetige, vektorwertige Funktion, über die eine Kurve $K = \{\vec{x}(t) : t \in [\alpha, \beta]\}$ parametrisiert ist. Die Kurve K heißt **rektifizierbar** und die Parameterfunktion $\vec{x}(t)$ von **beschränkter Variation** genau dann, wenn

$$L := \sup \left\{ \sum_{k=1}^m |\vec{x}(t_k) - \vec{x}(t_{k-1})| : (t_0, t_1, \dots, t_m) \in \mathcal{Z} \right\}$$

existiert. Dabei ist $\mathcal{Z}$ die Menge der Zerlegungen des Intervalls $[\alpha, \beta]$. Ist die Kurve rektifizierbar, so heißt die Zahl L die **Länge der Kurve**.

Hier wird die kleinste obere Schranke über Summen zu allen Zerlegungen $\alpha = t_0 < t_1 < \cdots < t_m = \beta$ des Intervalls $[\alpha, \beta]$ mit beliebiger endlicher Anzahl m von Zerlegungsintervallen $[t_{k-1}, t_k]$ betrachtet (vgl. Definition des Integrals). Für jede einzelne Zerlegung entspricht die Summe der Länge eines Kantenzugs, der die zu den Zerlegungspunkten gehörenden Punkte der Kurve verbindet. Die Länge des Kantenzugs ist offensichtlich kleiner oder gleich der Länge der Kurve.

Tatsächlich haben wir genau so die Länge

$$L = \int_{\alpha}^{\beta} \sqrt{\sum_{i=1}^m [\vec{x}'_i(t)]^2} dt$$

der vorangehenden Definition hergeleitet.

Faraday (1791–1867) führte zur Beschreibung elektrischer Felder den Begriff der Feldlinie ein:

Definition 4.26 (Feldlinie)

Eine Feldlinie eines Vektorfelds $\vec{V}$ ist eine Kurve, deren Tangenteneinheitsvektor in jedem Punkt der Kurve in Richtung des Funktionswerts von $\vec{V}$ an dieser Stelle zeigt.

Eine Feldlinie folgt sozusagen den Vektoren (Funktionswerten) des Vektorfelds. Die Feldlinien eines Magnetfelds werden sichtbar, wenn man Eisenspäne verstreut. Diese richten sich entlang der Feldlinien aus. Das elektrische Feld einer positiv geladenen Kugel hat Feldlinien, die radial von der Kugel in den Raum weggehen.

4.5.3 Quellen, Senken und Wirbel in Vektorfeldern

Mittels partieller Ableitungen kann man Eigenschaften eines Vektorfelds charakterisieren.

Definition 4.27 (Divergenz)

Ist ein Vektorfeld $\vec{V}$ (auf geeignetem Definitionsbereich $D \subset \mathbb{R}^n$) differenzierbar, so heißt

$$\operatorname{div} \vec{V}(\vec{x}) := (\nabla \cdot \vec{V})(\vec{x}) := \sum_{k=1}^n \frac{\partial V_k}{\partial x_k}(\vec{x})$$

Divergenz des Vektorfelds $\vec{V}$.

Bemerkungen:

- Den Begriff Divergenz haben wir schon bei Grenzwerten verwendet. Eine Folge (oder Funktion) heißt divergent, wenn der Grenzwert nicht existiert. Die hier definierte Divergenz ist ein völlig anderer Begriff. Aus dem Zusammenhang ist stets klar, was gemeint ist.
- Für eine **skalare** Funktion f ist $\operatorname{grad} f = \nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)$. Hier betrachten wir aber eine vektorwertige Funktion $\vec{V}$. Formal kann man die Divergenz verstehen als Skalarprodukt von $\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \dots, \frac{\partial}{\partial x_n} \right)$ mit $\vec{V} = (V_1, V_2, \dots, V_n)$.

Die Divergenz gibt an, ob ein Vektorfeld Quellen oder Senken hat. In den Quellen beginnen Feldlinien, die in den Senken enden (siehe Abbildung 4.23). Quellen oder Senken

liegen in einem Raumabschnitt vor, wenn dort die Divergenz nicht konstant null ist. Dies verstehen wir später mit dem Satz von Gauß. In Kapitel 4.5.7 sehen wir, dass bei positiver Divergenz „mehr aus einem Volumenelement heraus- als hereinströmt“, hier also eine Quelle liegen muss. Umgekehrt findet man eine negative Divergenz bei Senken. Ist die Divergenz überall gleich 0, so sprechen wir von einem **quellenfreien Feld**. Eine Herleitung dieses Zusammenhangs, die ohne explizite Verwendung des Satzes von Gauß auskommt (ihn aber nachempfunden), findet man z. B. in (Papula, 2008, Band 3, S. 71–74).

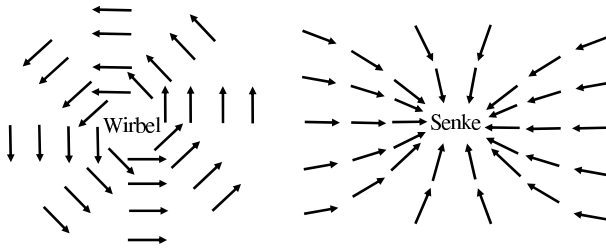


Abb. 4.23: Darstellung zweier Vektorfelder ($n = 2$): Die Funktionswerte zu einigen Punkten der Ebene sind als Vektorpfeile dargestellt.

Lemma 4.4 (Regeln für die Divergenz)

Seien $D \subset \mathbb{R}^n$ offen, $\vec{V}, \vec{W} : D \rightarrow \mathbb{R}^n$, $f : D \rightarrow \mathbb{R}$ differenzierbar. Dann gelten die folgenden Rechenregeln:

- a) $\operatorname{div}(\vec{V} + \vec{W}) = \operatorname{div} \vec{V} + \operatorname{div} \vec{W}$,
- b) $\operatorname{div}(f \cdot \vec{V}) = \operatorname{grad}(f) \cdot \vec{V} + f \operatorname{div} \vec{V}$ (beachte: Skalarprodukt + Skalar mal Skalar).

Beweis: Alle Aussagen ergeben sich direkt aus den Ableitungsregeln für Funktionen mit einer Variable. □

Definition 4.28 (Rotation)

Sei $\vec{V}(\vec{x}) : D \subset \mathbb{R}^3 \rightarrow \mathbb{R}^3$ ein differenzierbares Vektorfeld (auf geeignetem Definitionsbereich D). Die Rotation von $\vec{V}$ ist definiert als das Vektorfeld

$$\operatorname{rot} \vec{V}(\vec{x}) := \begin{pmatrix} \frac{\partial V_3}{\partial x_2}(\vec{x}) - \frac{\partial V_2}{\partial x_3}(\vec{x}) \\ \frac{\partial V_1}{\partial x_3}(\vec{x}) - \frac{\partial V_3}{\partial x_1}(\vec{x}) \\ \frac{\partial V_2}{\partial x_1}(\vec{x}) - \frac{\partial V_1}{\partial x_2}(\vec{x}) \end{pmatrix}.$$

Die Rotation ergibt sich über das formale Vektorprodukt (siehe Seite 444)

$$\nabla \times \vec{V} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} \end{pmatrix} \times \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} = \begin{pmatrix} \det \begin{bmatrix} \frac{\partial}{\partial x_2} & V_2 \\ \frac{\partial}{\partial x_3} & V_3 \end{bmatrix} \\ \det \begin{bmatrix} \frac{\partial}{\partial x_3} & V_3 \\ \frac{\partial}{\partial x_1} & V_1 \end{bmatrix} \\ \det \begin{bmatrix} \frac{\partial}{\partial x_1} & V_1 \\ \frac{\partial}{\partial x_2} & V_2 \end{bmatrix} \end{pmatrix}.$$

Die Rotation ist ein Maß für Wirbel im Vektorfeld. Ist die Rotation ungleich dem Nullvektor, so hat das Vektorfeld Wirbel, d. h. geschlossene Feldlinien. Darauf gehen wir in Kapitel 4.5.7 mit dem Satz von Stokes ein. Ist die Rotation überall gleich $\vec{0}$, so sprechen wir von einem **wirbelfreien Feld**. Der Name „Rotation“ ist vom Phänomen abgeleitet, dass bei Flüssen in Ufernähe durch die Reibung mit dem Ufer kleine Wirbel entstehen, die im Fluss schwimmende Teilchen in eine Drehung (Rotation) versetzen.

Lemma 4.5 (Regeln für die Rotation)

Seien $f : D \subset \mathbb{R}^3 \rightarrow \mathbb{R}$ und $\vec{V}, \vec{W} : D \rightarrow \mathbb{R}^3$ differenzierbar, wobei D ein geeigneter Definitionsbereich sei. Dann gelten die folgenden Rechenregeln:

- a) $\text{rot}(\vec{V} + \vec{W}) = \text{rot } \vec{V} + \text{rot } \vec{W}$,
- b) $\text{rot}(f\vec{V}) = f \text{rot } \vec{V} + \vec{V} \times \text{grad } f$,
- c) $\text{div}(\vec{V} \times \vec{W}) = \vec{W} \cdot \text{rot } \vec{V} - \vec{V} \cdot \text{rot } \vec{W}$.
- d) Sind sogar f und $\vec{V}$ zweimal stetig differenzierbar, so gilt:
 - i) $\text{rot grad } f(\vec{x}) = \vec{0}$, insbesondere ist $\text{grad } f$ ein wirbelfreies Vektorfeld,
 - ii) $\text{div rot } \vec{V}(\vec{x}) = 0$, insbesondere ist $\text{rot } \vec{V}$ ein quellenfreies Vektorfeld,
 - iii) $\text{rot rot } \vec{V} = \text{grad}(\text{div } \vec{V}) - \nabla^2(\vec{V})$, wobei $\nabla^2 := \Delta := \sum_{k=1}^n \frac{\partial^2}{\partial x_k^2}$ der **Laplace-Operator** bzw. Δ -Operator ist, der auf jede Komponente von $\vec{V}$ angewendet wird.

Die Regeln lassen sich unmittelbar mit den Ableitungsregeln nachrechnen. Je nachdem, ob die Funktionen reell- oder vektorwertig sind, muss im Lemma das Produkt mit einem Skalar oder das Skalarprodukt „ $\cdot$ “ verwendet werden. Daneben wird auch das Vektorprodukt „ $\times$ “ eingesetzt.

4.5.4 Kurvenintegrale

Bei einem Kurvenintegral berechnet man die Wirkung eines Vektorfelds entlang der Kurve. So kann man die Arbeit ermitteln, die man aufbringen muss, um einen Körper entlang einer Kurve durch ein Kraftfeld zu bewegen.

Nähert man eine Kurve in einem Kraftfeld durch viele kleine gerade Streckenstücke an, wie wir es bereits bei der Berechnung der Kurvenlänge getan haben (siehe Seite 676), so ergibt sich die Arbeit, indem man die Arbeit für jedes Streckenstück berechnet und dann aufsummiert. Für ein Streckenstück ist die Arbeit das Skalarprodukt des Streckenstücks (als Vektor) mit dem Vektor der Kraft, sofern diese auf dem kleinen Streckenstück konstant ist. Das Skalarprodukt berechnet das Produkt der Streckenlänge mit der Länge der Komponente der Kraft in Streckenrichtung. Eine Komponente der Kraft senkrecht zum Streckenstück hat keinen Einfluss. Lässt man nun die Längen der kleinen Streckenstücke gegen null gehen, um die Kurve exakt zu erhalten, so wird die Summation zur Integration, und man kommt von den Streckenstücken zur Ableitung der Kurve (ebenfalls wie bei der Berechnung der Kurvenlänge). Damit erhalten wir die folgende Definition:

Definition 4.29 (Kurvenintegral)

Sei K eine (stückweise) stetig differenzierbare Kurve mit der Parameterdarstellung $(\vec{x}(t), [\alpha, \beta])$. Weiter sei $\vec{V} : K \rightarrow \mathbb{R}^n$ ein stetiges Vektorfeld auf K mit Komponentenfunktionen $V_1, \dots, V_n$. Das **Kurvenintegral** von $\vec{V}$ längs K ist definiert als

$$\begin{aligned} \int_K \vec{V} d\vec{x} &:= \int_K V_1(\vec{x}) dx_1 + V_2(\vec{x}) dx_2 + \dots + V_n(\vec{x}) dx_n := \int_{\alpha}^{\beta} \vec{V}(\vec{x}(t)) \cdot \frac{\partial}{\partial t} \vec{x}(t) dt \\ &= \int_{\alpha}^{\beta} V_1(\vec{x}(t))x'_1(t) + V_2(\vec{x}(t))x'_2(t) + \dots + V_n(\vec{x}(t))x'_n(t) dt. \end{aligned}$$

Ist die Kurve K geschlossen, so schreibt man auch $\oint_K \vec{V} d\vec{x} := \int_K \vec{V} d\vec{x}$.

Man beachte, dass dies trotz der beiden neuen Schreibweisen ein normales (eindimensionales) Riemann-Integral ist! Es wird über das Skalarprodukt des Vektorfelds mit dem nicht-normierten Tangentenvektor der Kurve integriert.

Ist $\vec{V}$ beispielsweise unabhängig von $\vec{x}$ ein konstanter Vektor, so erhalten wir mit partieller Integration

$$\begin{aligned} &\int_{\alpha}^{\beta} V_1 x'_1(t) + V_2 x'_2(t) + \dots + V_n x'_n(t) dt \\ &= [V_1 x_1(t) + V_2 x_2(t) + \dots + V_n x_n(t)]_{\alpha}^{\beta} - \int_{\alpha}^{\beta} \vec{0} \cdot \vec{x}(t) dt = \vec{V} \cdot [\vec{x}(\beta) - \vec{x}(\alpha)], \end{aligned}$$

und dies ist das Produkt der Länge des Vektors vom End- zum Anfangspunkt der Kurve mit der in dieser Richtung wirkenden Komponente von $\vec{V}$.

Die Spannung zwischen zwei Punkten in einem elektrischen Feld erhält man über ein Kurvenintegral von einem zum anderen Punkt entlang einer Kurve K : $\int_K \vec{E} d\vec{x}$.

Beispiel 4.51

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 $V_1(x_1, x_2) := -x_2$, $V_2(x_1, x_2) := x_1$:

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

■

Durchläuft man eine Kurve in umgekehrter Richtung, also z. B. unter Verwendung der Parameterdarstellung $(\vec{x}(\alpha + \beta - t), [\alpha, \beta])$, so ändert sich das Vorzeichen des Integrals (Substitution $u = \alpha + \beta - t$, $du = -dt$):

$$\begin{aligned} &\int_\alpha^\beta \vec{V}(\vec{x}(\alpha + \beta - t)) \cdot \frac{\partial}{\partial t} \vec{x}(\alpha + \beta - t) dt \\ &= \int_\alpha^\beta V_1(\vec{x}(\alpha + \beta - t))(-x'_1(\alpha + \beta - t)) + V_2(\vec{x}(\alpha + \beta - t))(-x'_2(\alpha + \beta - t)) \\ &\quad + \cdots + V_n(\vec{x}(\alpha + \beta - t))(-x'_n(\alpha + \beta - t)) dt \\ &= \int_\beta^\alpha V_1(\vec{x}(u))x'_1(u) + V_2(\vec{x}(u))x'_2(u) + \cdots + V_n(\vec{x}(u))x'_n(u) du = - \int_K \vec{V} d\vec{x}. \end{aligned}$$

Wählt man statt $(\vec{x}(t), [\alpha, \beta])$ eine stückweise stetige Parametrisierung $(\vec{y}(t), [\gamma, \delta])$, mit der die gleiche Punktmenge in der gleichen Richtung durchlaufen wird, so erhält man den gleichen Wert des Kurvenintegrals. Der Wert eines Kurvenintegrals hängt nicht von der konkreten Parametrisierung der Kurve ab. Statt dies nachzurechnen, untersuchen wir, wann selbst unterschiedliche Kurven zum gleichen Wert führen.

Definition 4.30 (Gebiet)

Eine Menge $G \subset \mathbb{R}^n$ heißt ein **Gebiet** genau dann, wenn die folgenden Bedingungen erfüllt sind:

- G ist offen (siehe Definition 4.2 auf Seite 593).
- G ist **zusammenhängend**, d. h., je zwei Punkte aus G lassen sich durch einen Polygonzug (aneinandergesetzte Strecken) verbinden, der vollständig in G liegt (siehe Abbildung 4.24).

Definition 4.31 (Wegunabhängigkeit)

Sei $G \subset \mathbb{R}^n$ ein Gebiet und $\vec{V} : G \rightarrow \mathbb{R}^n$ ein stetiges Vektorfeld. Man spricht genau dann von **Wegunabhängigkeit** in Verbindung mit G und $\vec{V}$, wenn alle Kurveninte-

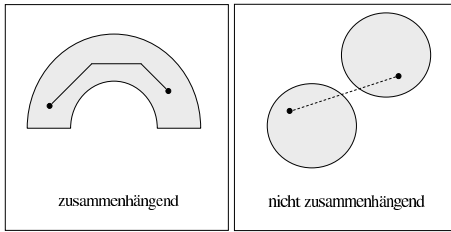


Abb. 4.24: Zusammenhang

grale zu stückweise stetig differenzierbaren Kurven, die in G einen gleichen Anfangs- mit einem gleichen Endpunkt verbinden, den gleichen Wert haben.

Wegunabhängigkeit liegt genau dann vor, wenn das Integral über jede geschlossene, stückweise stetig differenzierbare Kurve in G null ergibt. Hat man nämlich zwei Kurven, die einen Punkt $\vec{a}$ mit $\vec{b}$ verbinden, so erhält man daraus eine geschlossene Kurve, indem man die zweite in umgekehrter Richtung durchläuft und an die erste anfügt. Wechselt man die Durchlaufrichtung, ändert sich ja das Vorzeichen des Integrals.

Satz 4.18 (Potenzial)

Sei $G \subset \mathbb{R}^n$ ein Gebiet und $\vec{V} : G \rightarrow \mathbb{R}^n$ ein stetiges Vektorfeld. Wegunabhängigkeit liegt vor genau dann, falls es eine stetig differenzierbare Funktion $\varphi : G \rightarrow \mathbb{R}$ gibt mit $\vec{V}(\vec{x}) = \text{grad } \varphi(\vec{x})$. φ heißt ein **Potenzial** von $\vec{V}$ auf G .

Beweis: (Skizze) Der Beweis ist nicht schwer, aber länglich. Um einen Eindruck zu vermitteln, zeigen wir nur die Richtung „ $\Leftarrow$ “. Seien also φ ein Potenzial und $\vec{a}, \vec{b}$ zwei beliebige Punkte aus G . Für jede stückweise stetig differenzierbare Kurve K , die $\vec{a}$ mit $\vec{b}$ verbindet, gilt unter Verwendung von Kettenregel Satz 4.4 (Seite 607, vgl. Beispiel 4.13) und Hauptsatz Satz 2.48 (Seite 340):

$$\begin{aligned} \int_K \vec{V} d\vec{x} &= \int_{\alpha}^{\beta} \sum_{k=1}^n V_k(\vec{x}(t)) x'_k(t) dt \stackrel{\varphi \text{ ist Potenzial}}{=} \int_{\alpha}^{\beta} \sum_{k=1}^n \frac{\partial \varphi}{\partial x_k}(\vec{x}(t)) x'_k(t) dt \\ &\stackrel{\text{Satz 4.4}}{=} \int_{\alpha}^{\beta} \frac{d}{dt} [\varphi(\vec{x}(t))] dt \stackrel{\text{Satz 2.48}}{=} \varphi(\vec{x}(\beta)) - \varphi(\vec{x}(\alpha)). \end{aligned} \quad (4.18)$$

Unabhängig von K ergibt sich also stets der gleiche Wert.

Bei der umgekehrten Richtung wird das Potenzial über Kurvenintegrale konstruiert. Dies geschieht ähnlich zum Beweis von Satz 5.6 auf Seite 744, bei dem später Potenziale zur Lösung von Differenzialgleichungen genutzt werden. $\square$

Wenn wir Wegunabhängigkeit haben, ergibt sich wegen (4.18) der Wert eines Kurvenintegrals als Differenz der Werte des Potenzials am End- und Ausgangspunkt der Kurve.

Das elektrische Potenzial hat nicht nur zufällig den gleichen Namen wie der mathematische Begriff. Es handelt sich dabei (bis auf das Vorzeichen) um ein Potenzial im Sinne der Mathematik.

Jetzt sehen wir uns den Zusammenhang zwischen Wegunabhängigkeit und Wirbelfreiheit an. Anschaulich ist klar, dass man bei einem Integral entlang einer geschlossenen Feldlinie eines Wirbels nicht den Wert null erhält und damit keine Wegunabhängigkeit vorliegt. Zur exakten Formulierung benötigen wir einen weiteren Begriff:

Definition 4.32 (Einfach zusammenhängendes Gebiet)

Ein Gebiet G heißt einfach zusammenhängend genau dann, wenn man jede geschlossene Kurve in G zu einem Punkt zusammenziehen kann, ohne G zu verlassen.

Ist eine einfach zusammenhängende Menge G Teilmenge des $\mathbb{R}^2$, so gibt es keine Löcher in G , ist $G \subset \mathbb{R}^3$, so gibt es z. B. keine ausgesparten Kanäle von Seite zu Seite des Gebiets oder geschlossene Kanäle im Inneren.

Auf einem einfach zusammenhängenden Gebiet der $\mathbb{R}^3$ beweist man für ein stetig differenzierbares Vektorfeld $\vec{V}$, dass aus $\text{rot } \vec{V} = \vec{0}$ die Existenz eines Potenzials φ mit $\vec{V} = \text{grad } \varphi$ folgt (vgl. Seite 744). Aus der Wirbelfreiheit folgt so die Existenz eines Potenzials und damit die Wegunabhängigkeit.

Liegt umgekehrt Wegunabhängigkeit vor, so besitzt $\vec{V}$ nach Satz 4.18 ein Potenzial φ . Nach Lemma 4.5 ist $\text{rot } \vec{V}(\vec{x}) = \text{rot grad } \varphi(\vec{x}) = \vec{0}$, so dass $\vec{V}$ wirbelfrei ist. Fassen wir beide Argumente zusammen, so erhalten wir:

Lemma 4.6 (Wegunabhängigkeit und Wirbelfreiheit)

Für ein stetig differenzierbares Vektorfeld $\vec{V}$ auf einem einfach zusammenhängenden Gebiet $G \subset \mathbb{R}^3$ ist Wegunabhängigkeit äquivalent zur Wirbelfreiheit von $\vec{V}$.

Beim Fehlen sich ändernder Magnetfelder ist die elektrische Feldstärke $\vec{E}$ ein wirbelfreies Vektorfeld (siehe Kasten auf Seite 699 und insbesondere (4.27)) mit $\vec{E} = -\text{grad } \varphi$, wobei φ das elektrische Potenzial ist. Die Spannung zwischen zwei Punkten P_1 und P_2 erhält man über ein (beliebiges) Kurvenintegral über eine Kurve K von P_1 nach P_2 und damit als Potentialdifferenz:

$$\int_K \vec{E} d\vec{x} = -[\varphi(P_2) - \varphi(P_1)] = \varphi(P_1) - \varphi(P_2).$$

Unter den Voraussetzungen von Lemma 4.6 sind Wegunabhängigkeit und Wirbelfreiheit gleichbedeutend. Insbesondere hat ein wirbelfreies Feld $\vec{V}$ dann ein Potenzial φ , also $\vec{V} = \text{grad } \varphi$. Ist das Feld zusätzlich quellenfrei, so ist $0 = \text{div } \vec{V} = \text{div grad } \varphi = \Delta \varphi$

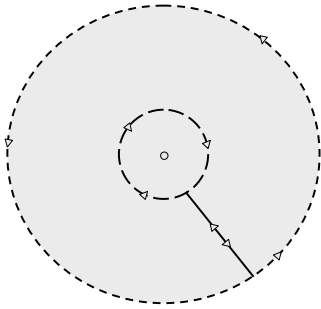


Abb. 4.25: Zum Residuensatz: Gegeben sei eine auf einem Gebiet, das die ausgefüllte Menge umfasst, komplex differenzierbare Funktion. Dabei können Punkte in der Bildmitte ausgespart sein. Das Integral entlang des kompletten gekennzeichneten Wegs ist null. Die beiden durchgezeichneten Wegstücke führen zu Integralen mit entgegengesetzten Vorzeichen, die sich wegheben. Damit hat das Kurvenintegral über den inneren Kreis den negativen Wert des Integrals über den äußeren Kreis. Durchläuft man beide Kreise positiv orientiert, haben die Integrale den gleichen Wert.

mit dem Laplace-Operator $\Delta := \sum_{k=1}^3 \frac{\partial^2}{\partial x_k^2}$ (vgl. Seite 680). Ein wirbel- und quellenfreies Feld hat damit ein Potenzial φ , das die **Laplace-Differenzialgleichung**

$$\Delta\varphi(x_1, x_2, x_3) = \frac{\partial^2 \varphi}{\partial x_1^2}(x_1, x_2, x_3) + \frac{\partial^2 \varphi}{\partial x_2^2}(x_1, x_2, x_3) + \frac{\partial^2 \varphi}{\partial x_3^2}(x_1, x_2, x_3) = 0 \quad (4.19)$$

erfüllt. Das Potenzial φ ist also eine Lösung dieser Gleichung, in der eine Bedingung an zweite partielle Ableitungen von φ gestellt wird. Daher handelt es sich hier wieder um eine **partielle Differenzialgleichung** (siehe auch Beispiel 4.8 auf Seite 602). 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 4.52

Jetzt betrachten wir die Jordan-Kurve K mit der Parametrisierung $(\vec{x}(t) := (\cos t, \sin t, 0), [0, 2\pi])$ und das Vektorfeld $\vec{V}(x_1, x_2, x_3) = (-x_2, x_1, 0)$ (vgl. Beispiel 4.51):

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

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

Hintergrund: Integrale von Funktionen mit einer komplexen Variable

Wir haben bislang (und werden im Folgenden) nur Funktionen mit einer reellen Variable integrieren. Integriert man von einer Stelle a zu einer Stelle $b \in \mathbb{R}$, so zerlegt man die Strecke von a nach b in (unendlich viele) kleine Teilintervalle, multipliziert jeweils einen Funktionswert mit einer Teilintervalllänge und summiert die reellen Produkte. Möchte man eine komplexwertige Funktion f mit der komplexen Variable $z = x + jy$ von einem Punkt $z_1 \in \mathbb{C}$ zu einem Punkt $z_2 \in \mathbb{C}$ integrieren, so kann man analog vorgehen und die Strecke zwischen z_1 und z_2 zerlegen. Die Differenz der Endpunkte eines Teilintervalls ist eine komplexe Zahl, die mit einem komplexen Funktionswert multipliziert wird. Auf-

summiert ergibt sich dann eine komplexe Zahl als Integral. Neben der direkten Strecke kann man aber in der komplexen Ebene auch entlang anderer Kurven K von z_1 zu z_2 gelangen. Daher verwendet man Kurvenintegrale für das Vektorfeld $\vec{f}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ mit $f_1(x, y) = \operatorname{Re}(f(x + jy))$ und $f_2(x, y) = \operatorname{Im}(f(x + jy))$. Damit ist das komplexe Integral erklärt über

$$\int_K f(z) dz := \int_K \begin{pmatrix} f_1(x, y) \\ -f_2(x, y) \end{pmatrix} d(x, y) + j \int_K \begin{pmatrix} f_2(x, y) \\ f_1(x, y) \end{pmatrix} d(x, y).$$

Die Definition wird verständlich, wenn wir die Punkte einer Kurve $K = ((x(t), y(t)), [\alpha, \beta])$ als komplexe Zahlen $x(t) + jy(t)$ auffassen und damit das Kurvenintegral überführen in ein Integral über das Produkt $f(x(t) + jy(t))[x'(t) + jy'(t)]$:

$$\begin{aligned} \int_K f(z) dz &= \int_{\alpha}^{\beta} f_1(x(t), y(t))x'(t) - f_2(x(t), y(t))y'(t) dt \\ &\quad + j \int_{\alpha}^{\beta} f_2(x(t), y(t))x'(t) + f_1(x(t), y(t))y'(t) dt \\ &= \int_{\alpha}^{\beta} [f_1(x(t), y(t)) + jf_2(x(t), y(t))][x'(t) + jy'(t)] dt \\ &= \int_{\alpha}^{\beta} f(x(t) + jy(t))[x'(t) + jy'(t)] dt. \end{aligned}$$

Hier haben wir statt separater reeller Integrale für Real- und Imaginärteil ein Integral geschrieben. Es ist aber genau so zu verstehen, dass man in Real- und Imaginärteil aufspaltet und separat integriert.

Beispiel 4.53

Entlang der positiv-orientierten Kurve K in der komplexen Ebene, die durch das Rechteck mit den Eckpunkten $1 + j$, $-1 + j$, $-1 - j$ und $1 - j$ gegeben ist, berechnen wir $\oint_K \frac{1}{z} dz$.

$$\begin{aligned} \oint_K \frac{1}{z} dz &= \int_1^{-1} \frac{1}{t+j} \cdot 1 dt + \int_1^{-1} \frac{1}{-1+jt} \cdot j dt + \int_{-1}^1 \frac{1}{t-j} \cdot 1 dt + \int_{-1}^1 \frac{1}{1+jt} \cdot j dt \\ &= \int_{-1}^1 -\frac{1}{t+j} - \frac{j}{-1+jt} + \frac{1}{t-j} + \frac{j}{1+jt} dt \\ &= \int_{-1}^1 -2\frac{t-j}{t^2+1} + 2\frac{t+j}{t^2+1} dt = 4j \int_{-1}^1 \frac{1}{t^2+1} dt \\ &= 4j[\arctan(1) - \arctan(-1)] = 4j \left[\frac{\pi}{4} - \frac{-\pi}{4} \right] = 2\pi j. \end{aligned}$$

Man beachte, dass der Integrand auf der Kurve definiert ist. Die Polstelle 0 liegt jedoch im Inneren der Kurve. ■

Die Funktionentheorie ist die mathematische Disziplin, die sich mit Funktionen einer komplexen Variable beschäftigt. Wir gehen in Kürze auf einige ihrer zentralen Sätze ein: Wenn auf dem betrachteten Definitionsbereich die Funktion f komplex differenzier-

bar ist (siehe Seite 419), dann führen die Cauchy-Riemann-Differenzialgleichungen (4.5) (siehe Seite 609) in Verbindung mit Satz 4.18 dazu, dass Wegunabhängigkeit vorliegt, und wir können von K unabhängig $\int_{z_1}^{z_2} f(z) dz$ schreiben. Insbesondere ist damit das Integral über eine geschlossene Kurve, die in einem einfach zusammenhängenden Gebiet liegt, auf dem die Funktion f komplex differenzierbar ist, gleich null. Das ist der **Cauchy-Integralsatz**. Ist die Kurve zusätzlich eine Jordan-Kurve (Definition 4.22), die im mathematisch positiven Sinn durchlaufen wird, so gilt für jeden Punkt z_0 im Inneren der Kurve:

$$f^{(n)}(z_0) = \frac{n!}{2\pi j} \oint_K \frac{f(z)}{(z - z_0)^{n+1}} dz.$$

Der neue Integrand $\frac{f(z)}{(z - z_0)^{n+1}}$ ist in z_0 nicht definiert, daher kann das Integral einen Wert ungleich null annehmen. Man erhält den Wert der n -ten komplexen Ableitung an der Stelle z_0 . Insbesondere ist damit eine auf dem einfach zusammenhängenden Gebiet einmal komplex differenzierbare Funktion automatisch beliebig oft differenzierbar. Die Darstellung für $f^{(n)}(z_0)$ heißt die **Cauchy-Integralformel**. Mit ihr erhalten wir für $n = 0$ und $f(z) = 1$ den Wert im obigen Beispiel schneller:

$$\oint_K \frac{1}{z} dz = \frac{2\pi j}{0!} f(0) = 2\pi j.$$

Ist eine Funktion auf einem einfach zusammenhängenden Gebiet differenzierbar bis auf einige Stellen $z_1, z_2, \dots, z_n$ im Inneren einer geschlossenen Kurve K in diesem Gebiet, so kann man das Integral über K zurückführen auf eine Summe von Integralen über geschlossene Kurven um die einzelnen Stellen z_k . Denn mit dem Cauchy-Integralsatz kann aus den einzelnen Integralen das Ausgangsintegral erhalten werden, indem man Integrale über geschlossene Kurven ergänzt, die allesamt den Wert null ergeben. Das ist in Abbildung 4.25 für eine Stelle z_1 (als Punkt in der Mitte) dargestellt.

Diese Reduktion auf Integrale um einzelne Stellen ist Gegenstand des **Residuensatzes**. Die einzelnen Summanden können gegebenenfalls mit der Cauchy-Integralformel unter Berücksichtigung der Anzahl der Kurvendurchläufe berechnet werden.

Beispiel 4.54

Entlang des positiv orientierten Kreises $|z| = 3$ integrieren wir $f(z) = \frac{z^3+2}{z(z-2)}$. Innerhalb des Kreises liegen die beiden Pole $z = 0$ und $z = 2$. Integrieren wir über positiv orientierte Kurven K_0 und K_2 , innerhalb derer jeweils nur entweder 0 oder 2 liegt, so ist

$$\oint_K \frac{z^3+2}{z(z-2)} dz = \oint_{K_0} \frac{\frac{z^3+2}{z-2}}{z} dz + \oint_{K_1} \frac{\frac{z^3+2}{z}}{z-2} dz = \frac{2\pi j}{0!} \frac{0^3+2}{0-2} + \frac{2\pi j}{0!} \frac{2^3+2}{2} = 8\pi j.$$

■

Eine Integrationstechnik für Funktionen mit einer reellen Variable besteht darin, dass man den Integranden zu einer Funktion mit einer komplexen Variable fortsetzt und dann entlang einer geschlossenen Kurve integriert, in der ein Intervall der reellen Achse enthalten ist. Der Integralwert für das reelle Teilstück entspricht dem gesuchten reellen Integral.

4.5.5 Satz von Green*

Wir nähern uns nun einigen zentralen Sätzen der Vektoranalysis und beginnen in diesem Abschnitt mit dem Satz von Green. Der **Satz von Green** stellt für ebene Flächen $E \subset \mathbb{R}^2$ einen Zusammenhang zwischen einem Integral über die Fläche und einem Kurvenintegral über die Randkurve ∂E der Fläche dar. Die Informationen des Randes reichen aus, um das Integral über die komplette Fläche anzugeben. Damit dieser Zusammenhang gilt, muss die Fläche einfach strukturiert sein und die Funktion $\vec{V}(x, y) = (V_1(x, y), V_2(x, y))$ stetig differenzierbar sein:

$$\iint_E \left[\frac{\partial V_2}{\partial x}(x, y) - \frac{\partial V_1}{\partial y}(x, y) \right] d(x, y) = \oint_{\partial E} \vec{V} d\vec{x}.$$

Mit dem Satz von Stokes werden wir später eine allgemeinere Aussage für Flächen in $\mathbb{R}^3$ erhalten.

Der Sachverhalt dieser Aussage ist gar nicht so überraschend, denn auch mit dem Hauptsatz der Differenzial- und Integralrechnung reduziert man ein Integral auf die Berechnung von Werten des Randes (siehe Seite 342): $\int_a^b f'(x) dx = f(b) - f(a)$. Der Rand des Intervalls $[a, b]$ wird durch die beiden Stellen a und b gebildet.

Mit vollständigen Voraussetzungen lautet der Satz so:

Satz 4.19 (Satz von Green, auch Satz von Gauß in der Ebene)

Sei $E \subset \mathbb{R}^2$ ein Gebiet, so dass E zusammen mit dem Rand von E ein zweidimensionaler Normalbereich (siehe Definition 4.17) ist, d. h., jede Gerade parallel zur x -Achse betritt das Gebiet höchstens einmal und verlässt es höchstens einmal. Entsprechendes gilt für jede Gerade parallel zur y -Achse. Interpretiert man jeweils (separat) die Eintritts- und Austrittspunkte als Funktionsgraphen, so müssen die zugehörigen Funktionen (dies sind f_1 , f_2 und g_1 , g_2 in Definition 4.17) hinreichend glatt sein (z. B. stetig differenzierbar).

Weiter sei $D \subset \mathbb{R}^2$, so dass E und die Randkurve von E in D liegen und zudem D offen ist (siehe Definition 4.2), d. h., zu jedem Punkt aus D liegt auch eine komplette Umgebung in D . Weiter sei $\vec{V} : D \rightarrow \mathbb{R}^2$ ein stetig differenzierbares Vektorfeld. Dann gilt:

$$\iint_E \left[\frac{\partial V_2}{\partial x}(x, y) - \frac{\partial V_1}{\partial y}(x, y) \right] d(x, y) = \oint_{\partial E} \vec{V} d\vec{x}. \quad (4.20)$$

Dabei bezeichnet ∂E eine Kurve, die dem Rand von E entspricht und die im Gegenuhreizersinn, d. h. im mathematisch positiven Sinn durchlaufen wird.

Bemerkungen:

- Die linke Seite von (4.20) ist ein „normales Integral“ einer Funktion von E nach $\mathbb{R}$.

- Die rechte Seite von (4.20) ist ein Kurvenintegral, wobei man den Rand von E stückweise als Kurve im Gegenuhrzeigersinn parametrisieren muss.
- Wird dagegen die Kurve im Uhrzeigersinn durchlaufen, unterscheiden sich die beiden Integrale genau im Vorzeichen.
- Es wird gefordert, dass die Randfunktionen des Normalbereichs stetig differenzierbar sein sollen. Diese Bedingung ist z. B. für den Einheitskreis verletzt. Dennoch ist dieses Gebiet glatt genug, damit der Satz von Green gilt.

Beweis: (Skizze) Der Satz 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 (Seite 340) 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 vier Kurven, die über $\vec{x}_1(t) := (t, a_2)$, $t \in [a_1, b_1]$, $\vec{x}_2(t) := (b_1, t)$, $t \in [a_2, b_2]$, $\vec{x}_3(t) := (a_1 + b_1 - t, b_2)$, $t \in [a_1, b_1]$ und $\vec{x}_4(t) := (a_1, a_2 + b_2 - t)$, $t \in [a_2, b_2]$ parametrisiert sind. Für ein stetig differenzierbares Vektorfeld $\vec{V} = (V_1, V_2)$ gilt mit dem Satz von Fubini (Seite 653):

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

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

Beispiel 4.55

Sei $E :=]0, 1[\times]0, 1[$ und $\vec{V} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ mit $\vec{V}(x, y) = (x^4 - y^3, x^3 - y^3)$. Dann ist wegen des Satzes von Green

$$\begin{aligned}
 \oint_{\partial E} (x^4 - y^3) dx + (x^3 - y^3) dy &= \iint_E \left[\frac{\partial V_2}{\partial x}(x, y) - \frac{\partial V_1}{\partial y}(x, y) \right] d(x, y) \\
 &= \iint_E 3x^2 + 3y^2 d(x, y) = 3 \int_0^1 \int_0^1 x^2 + y^2 dy dx = 3 \int_0^1 x^2 + \frac{1}{3} dx = 2.
 \end{aligned}$$

Wir kontrollieren das Ergebnis, indem wir das Kurvenintegral über eine Parametrisierung der vier Kanten ausrechnen:

$$\begin{aligned}
 \oint_{\partial E} (x^4 - y^3) dx + (x^3 - y^3) dy &= \int_0^1 (t^4 - 0^3, t^3 - 0^3) \begin{pmatrix} 1 \\ 0 \end{pmatrix} dt \\
 &+ \int_0^1 (1^4 - t^3, 1^3 - t^3) \begin{pmatrix} 0 \\ 1 \end{pmatrix} dt + \int_0^1 ((1-t)^4 - 1^3, (1-t)^3 - 1^3) \begin{pmatrix} -1 \\ 0 \end{pmatrix} dt \\
 &+ \int_0^1 (0^4 - (1-t)^3, 0^3 - (1-t)^3) \begin{pmatrix} 0 \\ -1 \end{pmatrix} dt \\
 &= \int_0^1 t^4 + 1 - t^3 + 1 - (1-t)^4 + (1-t)^3 dt = \frac{1}{5} + 1 - \frac{1}{4} + 1 - \frac{1}{5} + \frac{1}{4} = 2. \quad \blacksquare
 \end{aligned}$$

Beispiel 4.56

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

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

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

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

4.5.6 Flächenintegrale *

Hat man ein Vektorfeld, das eine Strömung beschreibt, so interessiert man sich häufig für die Flussmenge durch eine bestimmte Fläche. Um diese zu berechnen, verwendet man ein Flächenintegral. Hier werden die Strömungsanteile „aufintegriert“, die senkrecht zur Fläche stehen.

Definition 4.33 (Parameterdarstellung einer Fläche)

Seien $S \subset \mathbb{R}^2$ ein Gebiet und $\vec{F}(u, v) := (x(u, v), y(u, v), z(u, v)) : S \rightarrow \mathbb{R}^3$ stetig differenzierbar, so dass für jeden Punkt $(u, v) \in S$ die Jacobi-Matrix (siehe Seite 606) von $\vec{F}$ den Rang 2 besitzt:

$$\text{Rang} \begin{bmatrix} \frac{\partial x}{\partial u}(u, v) & \frac{\partial x}{\partial v}(u, v) \\ \frac{\partial y}{\partial u}(u, v) & \frac{\partial y}{\partial v}(u, v) \\ \frac{\partial z}{\partial u}(u, v) & \frac{\partial z}{\partial v}(u, v) \end{bmatrix} = 2.$$

Dann heißt die Menge

$$F := \{(x, y, z) \in \mathbb{R}^3 : x = x(u, v), y = y(u, v), z = z(u, v), (u, v) \in S\}$$

eine **Fläche** im $\mathbb{R}^3$, und $(\vec{F}, S)$ heißt eine **Parameterdarstellung** dieser Fläche.

Bei einer Kurve hat man einen Parameter, eine Fläche ist über zwei Variablen parametrisiert.

Definition 4.34 (Oberflächenintegral)

Gegeben sei eine Fläche F in $\mathbb{R}^3$ mit Parameterdarstellung $(\vec{F}, S)$. Außerdem sei $g : F \rightarrow \mathbb{R}$ stetig. Das Integral

$$\int_F g \, d\sigma := \iint_S g(\vec{F}(u, v)) \left| \frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v) \right| d(u, v)$$

heißt **Oberflächenintegral** von g über F (bezüglich der Parameterdarstellung $(\vec{F}, S)$), falls es existiert.

In Definition 3.5 auf Seite 444 wurde das Vektorprodukt für Vektoren $\vec{x}, \vec{y} \in \mathbb{R}^3$ eingeführt über

$$\vec{x} \times \vec{y} = (x_2 y_3 - y_2 x_3, x_3 y_1 - y_3 x_1, x_1 y_2 - y_1 x_2).$$

Sind $\vec{x}$ und $\vec{y}$ linear unabhängig, so steht der dabei resultierende Vektor senkrecht auf der von $\vec{x}$ und $\vec{y}$ aufgespannten Fläche. Aufgrund der Rang-Bedingung gilt das insbesondere für die linear unabhängigen (Tangenten-)Vektoren $\frac{\partial \vec{F}}{\partial u}(u, v)$ und $\frac{\partial \vec{F}}{\partial v}(u, v)$. Außerdem beschreibt $\left| \frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v) \right|$ die Fläche des durch die Tangentenvektoren im Punkt $\vec{F}(u, v)$ aufgespannten Parallelogramms (siehe Seite 444) und drückt das Verzerrungsverhältnis durch die Parametrisierung aus. Damit wird der berechnete Wert insbesondere unabhängig von der Parametrisierung.

Um die Definition 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

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

$1 \leq i \leq n$, $1 \leq k \leq m$, 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

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

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.

Betrachtet man speziell die Funktion $g(\vec{x}) = 1$, so liefert das Oberflächenintegral $\int_F d\sigma$ den Flächeninhalt der parametrisierten Fläche.

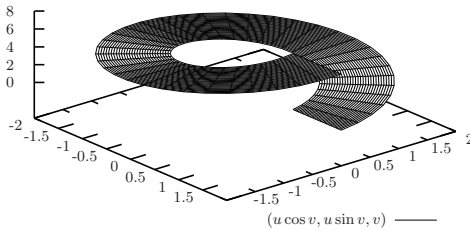


Abb. 4.26: Wendelfläche $\vec{F}(u, v) = (u \cos v, u \sin v, v)$, $S =]1, 2[\times]0, 2\pi[$

Beispiel 4.57

Wir berechnen den Flächeninhalt der Wendelfläche $(\vec{F}, S)$ mit $\vec{F}(u, v) = \begin{pmatrix} u \cos v \\ u \sin v \\ v \end{pmatrix}$

und $S =]1, 2[\times]0, 2\pi[$ (siehe Abbildung 4.26).

Zunächst ist mit (1.20)

$$\begin{aligned} \left| \frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v) \right| &= \left| \begin{pmatrix} \cos v \\ \sin v \\ 0 \end{pmatrix} \times \begin{pmatrix} -u \sin v \\ u \cos v \\ 1 \end{pmatrix} \right| = \left| \begin{pmatrix} \sin v \\ -\cos v \\ u \end{pmatrix} \right| \\ &= \sqrt{1 + u^2}. \end{aligned}$$

Damit erhalten wir für den Flächeninhalt:

$$\int_F d\sigma = \iint_S \sqrt{1 + u^2} d(u, v) = \int_1^2 \int_0^{2\pi} \sqrt{1 + u^2} dv du = 2\pi \int_1^2 \sqrt{1 + u^2} du$$

$$= \pi \left[u\sqrt{1+u^2} + \ln \left(u + \sqrt{1+u^2} \right) \right]_1^2 = \pi \left[2\sqrt{5} - \sqrt{2} + \ln \left(\frac{2+\sqrt{5}}{1+\sqrt{2}} \right) \right].$$

Dass es sich beim letzten Schritt um eine Stammfunktion handelt, lässt sich durch Ableiten verifizieren:

$$\begin{aligned} & \frac{d}{du} \frac{1}{2} \left[u\sqrt{1+u^2} + \ln \left(u + \sqrt{1+u^2} \right) \right] \\ &= \frac{1}{2} \left[\sqrt{1+u^2} + u \frac{2u}{2\sqrt{1+u^2}} + \frac{1}{u + \sqrt{1+u^2}} \left(1 + \frac{2u}{2\sqrt{1+u^2}} \right) \right] \\ &= \frac{1}{2} \left[\sqrt{1+u^2} + \frac{(u^2+1)(u+\sqrt{1+u^2})}{\sqrt{1+u^2}(u+\sqrt{1+u^2})} \right] = \sqrt{1+u^2}. \quad \blacksquare \end{aligned}$$

Um die **Flussmenge** eines Vektorfelds $\vec{V}$ durch eine Fläche zu bestimmen, ist nur der Anteil der Strömung relevant, der senkrecht zur Fläche steht. Ein Anteil parallel zur Fläche spielt keine Rolle. Man erhält diesen Anteil an einer Stelle (x, y, z) der Fläche, indem man das Skalarprodukt mit einem Vektor $\vec{N} \in \mathbb{R}^3$ bildet, der senkrecht zur Fläche steht und den Betrag 1 hat. Nach Kapitel 3.1.2 ist

$$\vec{V}(x, y, z) \cdot \vec{N}(x, y, z) = |\vec{V}(x, y, z)| \cos \varphi,$$

wobei φ der Winkel zwischen $\vec{N}$ und $\vec{V}$ an der Stelle (x, y, z) ist. Interessiert man sich für die Flussmenge durch die Fläche F , so erhält man diese über das Oberflächenintegral zu $g(x, y, z) := \vec{V}(x, y, z) \cdot \vec{N}(x, y, z)$:

$$\int_F \vec{V}(x, y, z) \cdot \vec{N}(x, y, z) d\sigma := \iint_S \vec{V}(\vec{F}(u, v)) \cdot \vec{N}(\vec{F}(u, v)) \left| \frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v) \right| d(u, v).$$

Wenn wir nun noch den Normalenvektor $\vec{N}$ über das Vektorprodukt

$$\vec{N}(\vec{F}(u, v)) = \frac{\frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v)}{\left| \frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v) \right|}$$

(siehe Kapitel 3.1.3) berechnen, erhalten wir das **Flussintegral**

$$\int_F \vec{V}(x, y, z) \cdot \vec{N}(x, y, z) d\sigma = \iint_S \vec{V}(\vec{F}(u, v)) \cdot \left(\frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v) \right) d(u, v). \quad (4.21)$$

Für den Wert des Integrals spielt die Orientierung des Normalenvektors eine Rolle. Hier ist gegebenenfalls das Vorzeichen passend zur Aufgabenstellung anzupassen.

4.5.7 Die Sätze von Gauß und Stokes*

Diese Sätze sind insbesondere für die Elektrotechnik interessant. Mit den Maxwell'schen Gleichungen (siehe Kasten auf Seite 699) lassen sich dort viele Zusammenhänge darstellen. Diese Gleichungen existieren in einer Integralform und einer Differenzialform. Beide lassen sich mit den Sätzen von Gauß und Stokes (die hier im Folgenden ohne Beweis angegeben werden) ineinander überführen.

Beim Satz von Green kann man die Integration über eine ebene Fläche auf ein Integral über den Rand der Fläche zurückführen (und umgekehrt). Die Sätze von Gauß und Stokes sind sehr ähnliche Aussagen. Beim Satz von Gauß überführt man Volumenintegrale in Oberflächenintegrale (und umgekehrt). Der Satz von Stokes vergleicht Flächen- mit Randintegralen. Allerdings müssen hier die Flächen nicht eben sein wie beim Satz von Green.

4.5.7.1 Satz von Gauß

Mit dem Satz von Gauß kann man Quellen und Senken eines Vektorfelds über die Divergenz charakterisieren. In Kurzfassung besagt der Satz von Gauß, dass für ein (geeignetes) Gebiet $E \subset \mathbb{R}^3$ und eine stetig differenzierbare Funktion $\vec{V} : D \rightarrow \mathbb{R}^3$, die auf einer genügend großen Menge D definiert ist, ein Integral der Divergenz von $\vec{V}$ über E durch ein Oberflächenintegral über die Randfläche ∂E von E ausgerechnet werden kann (und umgekehrt), wobei das Oberflächenintegral den Fluss des Vektorfelds durch die Oberfläche angibt:

$$\iiint_E \operatorname{div} \vec{V}(x, y, z) \, d(x, y, z) = \int_{\partial E} \vec{V} \cdot \vec{N} \, d\sigma.$$

Mit ausführlichen Voraussetzungen sieht der Satz dann so aus:

Satz 4.20 (Satz von Gauß, der Divergenz-Satz)

Sei $E \subset \mathbb{R}^3$ ein Gebiet, so dass E zusammen mit dem Rand von E ein dreidimensionaler Normalbereich (siehe Bemerkung zu Definition 4.17) ist, so dass die Randfunktionen des Normalbereichs stetig differenzierbar sind. Weiter sei $D \subset \mathbb{R}^3$, so dass E und die Randfläche von E in D liegen und zudem D offen ist (siehe Definition 4.2). Weiter sei $\vec{V} : D \rightarrow \mathbb{R}^3$ ein stetig differenzierbares Vektorfeld. Dann gilt:

$$\iiint_E \operatorname{div} \vec{V}(x, y, z) \, d(x, y, z) = \int_{\partial E} \vec{V} \cdot \vec{N} \, d\sigma. \quad (4.22)$$

Dabei bezeichnet ∂E die Oberfläche von E , und $\int_{\partial E} d\sigma$ ist das zugehörige Oberflächenintegral. An jeder Stelle der Oberfläche ist $\vec{N}$ die **äußere Normale**. Dies ist ein Vektor der Länge 1, der senkrecht zur Oberfläche steht und aus E hinauszeigt.

Bemerkungen:

- Die linke Seite von (4.22) ist ein „normales Integral“ einer reellwertigen Funktion $\operatorname{div}(x, y, z)$ von drei Variablen x, y und z , das wir z. B. mit dem Satz von Fubini berechnen können.
- Die rechte Seite von (4.22) ist ein Oberflächenintegral (siehe Definition 4.34) und genauer ein Flussintegral (4.21). Dabei wird der Rand ∂E stückweise als Fläche parametrisiert (was geht, da E ein Normalbereich ist). Für jede Parametrisierung $(\vec{F}, S)$ eines Teils der Oberfläche ∂E ist das Flussintegral

$$\int_{\partial E} \vec{V} \cdot \vec{N} \, d\sigma = \iint_S s(F) \cdot \vec{V}(\vec{F}(u, v)) \cdot \left(\frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v) \right) d(u, v).$$

Dabei ist $s(F)$ entweder als 1 oder -1 zu wählen, so dass der Vektor $\frac{\partial \vec{F}}{\partial u}(u, v) \times \frac{\partial \vec{F}}{\partial v}(u, v)$ aus E heraus nach außen zeigt. Man integriert jetzt über ein Skalarprodukt, dessen Vorzeichen durch $s(F)$ verändert wird. Damit ist der Integrand eine reellwertige Funktion. Schließlich werden die Integrale der Parametrisierungen der Teiloberflächen addiert.

- Auch dieser Satz wird mit dem Hauptsatz (Seite 340) und dem Satz von Fubini (Seite 653) bewiesen. Falls das Gebiet E ein Quader ist, wird der Satz genau wie in der Beweisskizze zum Satz von Green auf Seite 689 gezeigt. Tatsächlich unterscheiden sich die Sätze auch nur in der Raumdimension, wenn man die unterschiedliche Definition von Kurven- und Oberflächenintegralen berücksichtigt. Der Integrand auf der linken Seite des Satzes von Green wird zu einer „zweidimensionalen Divergenz“, wenn man das Produkt mit einem Normalenvektor senkrecht zur Randkurve herausrechnet, siehe z. B. (Göllmann, 2017, Band 2, S. 181).

Beispiel 4.58

Mit dem Satz von Gauß berechnen wir das Oberflächenintegral $\int_{\partial E} \vec{V} \cdot \vec{N} \, d\sigma$, wobei $\vec{V}(x, y, z) := (x^2, y^2, z^2)$ und $E =]0, 1[\times]0, 1[\times]0, 1[$ ist:

$$\int_{\partial E} \vec{V} \cdot \vec{N} \, d\sigma = \iiint_E \operatorname{div} \vec{V}(x, y, z) \, d(x, y, z) = \int_0^1 \int_0^1 \int_0^1 2x + 2y + 2z \, dz \, dy \, dx = 3.$$

■

Mit dem Satz von Gauß können wir nun die Definition der Divergenz besser verstehen. Wir betrachten eine Kugel K_r mit Oberfläche F_r und Radius r (d. h. Volumen $\frac{4}{3}\pi r^3$) um eine Stelle (x, y, z) . Mit $\vec{N}$ sei der Normalenvektor bezeichnet, der an einer Stelle der Kugeloberfläche senkrecht nach außen zeigt. Für eine stetige Funktion $\operatorname{div} \vec{V}$ kann der Funktionswert an der Stelle (x, y, z) auch als Grenzwert eines Integrals geschrieben werden (vgl. (2.49) auf Seite 338), den wir mit dem Satz von Gauß anders schreiben können:

$$\operatorname{div} \vec{V}(x, y, z) = \lim_{r \rightarrow 0+} \frac{1}{\frac{4}{3}\pi r^3} \iiint_{K_r} \operatorname{div} \vec{V}(u, v, w) \, d(u, v, w) = \lim_{r \rightarrow 0+} \frac{1}{\frac{4}{3}\pi r^3} \int_{F_r} \vec{V} \cdot \vec{N} \, d\sigma.$$

Hier betrachtet man die Flussmenge durch eine geschlossene Kugeloberfläche (pro Volumen der Kugel), die sich im Grenzwert auf einen Punkt zusammenzieht (vgl. (4.21)). Hat das Vektorfeld $\vec{V}$ in (x, y, z) eine „Quelle“ (z. B. eine Punktladung), so ist die Divergenz positiv. Der Fluss aus der Kugel nach außen ist größer als der Fluss durch die Kugeloberfläche nach innen. Hat das Vektorfeld in (x, y, z) eine „Senke“, so ist die Divergenz negativ. Der Fluss durch die Kugeloberfläche in die Kugel hinein ist größer als der durch die Oberfläche nach außen. Hat das Vektorfeld weder Quelle noch Senke, so ist die Divergenz null. Dann ist der Fluss durch die Oberfläche ausgeglichen, innerhalb der Kugel kommt nichts hinzu und nichts verschwindet (vgl. Abbildung 4.23, Seite 679).

Hintergrund: Der Satz von Gauß und die Laplace-Differenzialgleichung

Mit dem Satz von Gauß gewinnen wir auch eine neue Sicht auf die Laplace- oder Poisson-Differenzialgleichung von Seite 685, 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 (siehe (4.19)), oder sonst als Lösung einer Poisson-Gleichung. Wählen wir in (4.22) die Funktion $\vec{V}(x, y, z) = \text{grad } f(x, y, z) = \nabla f(x, y, z)$ für eine geeignete reellwertige Funktion f (z. B. wäre mit dem elektrischen Potenzial φ die Wahl $f := -\varphi$ möglich, und die elektrische Feldstärke wäre $\vec{E} = \vec{V} = -\text{grad } \varphi$), so ergibt sich die Green'sche Formel

$$\begin{aligned} \iiint_E \Delta f(x, y, z) d(x, y, z) &\stackrel{\text{vgl. Seite 680}}{=} \iiint_E \text{div grad } f(x, y, z) d(x, y, z) \\ &= \iiint_E \text{div } \vec{V}(x, y, z) d(x, y, z) \stackrel{(4.22)}{=} \int_{\partial E} \vec{V} \cdot \vec{N} d\sigma \\ &= \int_{\partial E} \text{grad } f \cdot \vec{N} d\sigma \stackrel{\text{vgl. Seite 601}}{=} \int_{\partial E} \frac{\partial f}{\partial \vec{N}} d\sigma. \end{aligned}$$

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 Felds eine harmonische Funktion. Hier muss notwendigerweise 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. Üblicherweise 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 $\vec{x}_0 \neq \vec{0}$ (Strauss, 1995, S. 198):

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

Hier hat man allerdings die Schwierigkeit, dass die Funktion in jeder Umgebung von $\vec{x}_0$ unbeschränkt ist, so dass $\iiint_E g(\vec{x}) G(\vec{x}, \vec{x}_0) d\vec{x}$ 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.

4.5.7.2 Satz von Stokes

So wie der Satz von Gauß den Zusammenhang zwischen Divergenz und Quellen und Senken eines Felds beschreibt, drückt der Satz von Stokes den Zusammenhang zwischen Rotation und Wirbeln aus. In Kurzfassung besagt der Satz von Stokes, dass der Fluss der Rotation eines Vektorfelds durch eine Oberfläche F (vgl. (4.21)) durch ein Kurvenintegral bezüglich des Randes von F ausgerechnet werden kann (und umgekehrt):

$$\int_F (\operatorname{rot} \vec{V}) \cdot \vec{N} \, d\sigma = \oint_{\partial F} \vec{V} \, d\vec{x}.$$

Die vollständige Information über die Rotation eines Vektorfelds $\vec{V}$ senkrecht zur Fläche F steckt bereits in den Funktionswerten von $\vec{V}$ auf dem Rand ∂F der Fläche. Mit ausführlichen Voraussetzungen sieht der Satz so aus:

Satz 4.21 (Satz von Stokes)

Sei $(\vec{F}, S)$ die Parameterdarstellung einer Fläche $F \subset \mathbb{R}^3$, so dass die Parameterfunktion $\vec{F}$ zweimal stetig differenzierbar auf einer offenen Menge ist, die S und den Rand von S umfasst. Außerdem sei S ein zweidimensionaler Normalbereich mit stetig differenzierbaren Randfunktionen. Der Rand ∂F der Fläche F sei stückweise als Kurve parametrisierbar. Sei $D \subset \mathbb{R}^3$ offen und Definitionsbereich des stetig differenzierbaren Vektorfelds $\vec{V} : D \rightarrow \mathbb{R}^3$. Die Fläche F möge vollständig im Definitionsbereich von $\vec{V}$ liegen: $F \subset D$. Dann gilt:

$$\int_F (\operatorname{rot} \vec{V}) \cdot \vec{N} \, d\sigma = \oint_{\partial F} \vec{V} \, d\vec{x}. \quad (4.23)$$

Dabei ist das Vorzeichen des Normalenvektors $\vec{N}$ so zu wählen, dass die Vorzeichen beider Seiten übereinstimmen. Durchläuft man die Randkurve im Gegenuhrzeigersinn, so zeigt der Normalenvektor relativ dazu nach oben.

Bemerkungen:

- Die linke Seite von (4.23) ist ein Oberflächenintegral. Dabei wird das Skalarprodukt der Rotation von $\vec{F}$ mit dem Normalenvektor der Fläche integriert. Man erhält so den Fluss von $\operatorname{rot} \vec{V}$ durch die Fläche (vgl. (4.21)), den man den **Wirbelfluss** nennt.
- Die rechte Seite ist ein Kurvenintegral entlang der Randkurve der Fläche F .
- Der Satz von Green folgt aus dem Satz von Stokes, wenn man eine Fläche betrachtet, deren Punkte alle die z-Koordinate 0 haben. Der Satz von Green ist also ein Spezialfall des Satzes von Stokes. So kann man analog zum Beispiel 4.56 mit dem Satz von Stokes den Flächeninhalt einer Fläche $F \subset \mathbb{R}^3$ über ein Randintegral berechnen.

Beispiel 4.59

Mit dem Satz von Stokes berechnen wir $\oint_{\partial F} \vec{V} \, d\vec{x}$, wobei $\partial F = \{(x, y, z) : x = \cos t, y = \sin t, z = 1, t \in [0, 2\pi]\}$ und $\vec{V}(x, y, z) = (2y, -2x, z^2x)$. Die berandete Fläche ist eine

ebene Kreisscheibe mit Normalenvektor $\vec{N} = (0, 0, 1)$, Radius 1 und Flächeninhalt π .
 $\text{rot } \vec{V}(x, y, z) = (0, -z^2, -4)$. Damit:

$$\int_F (\text{rot } \vec{V}) \cdot \vec{N} d\sigma = \int_F (0, -z^2, -4) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} d\sigma = \int_F -4 d\sigma = -4\pi.$$

Wir erhalten das gleiche Ergebnis durch direktes Ausrechnen des Kurvenintegrals:

$$\begin{aligned} \oint_{\partial F} \vec{V} d\vec{x} &= \int_0^{2\pi} \vec{V}(\cos t, \sin t, 1) \cdot \begin{pmatrix} -\sin t \\ \cos t \\ 0 \end{pmatrix} dt \\ &= \int_0^{2\pi} (2 \sin t, -2 \cos t, \cos t) \cdot \begin{pmatrix} -\sin t \\ \cos t \\ 0 \end{pmatrix} dt = \int_0^{2\pi} -2 dt = -4\pi. \quad \blacksquare \end{aligned}$$

Der Satz von Stokes liefert eine Interpretation der Rotation (vgl. Abbildung 4.23 auf Seite 679). Wir betrachten eine Kreisscheibe K_r mit Rand ∂K_r und Radius r um den Punkt (x, y, z) . $\vec{N}$ sei der Normalenvektor zu dieser Kreisscheibe. Für eine stetige Funktion $\text{rot } \vec{V}$ kann der Funktionswert an der Stelle (x, y, z) auch als Grenzwert eines Integrals geschrieben werden, den wir mit dem Satz von Stokes anders schreiben können:

$$\text{rot}(\vec{V}(x, y, z)) \cdot \vec{N} = \lim_{r \rightarrow 0+} \frac{1}{\pi r^2} \int_{K_r} (\text{rot } \vec{V}) \cdot \vec{N} d\sigma = \lim_{r \rightarrow 0+} \frac{1}{\pi r^2} \oint_{\partial K_r} \vec{V} d\vec{x}.$$

Ist der Anteil der Rotation in Richtung des Normalenvektors ungleich null, so sind auch die Kurvenintegrale für kleine Kreise K_r der rechten Seite ungleich null. Die Vektoranteile des Felds heben sich entlang der Kurve nicht auf. Das ist ein Indiz für geschlossene Feldlinien, also für Wirbel. Diese Beobachtung passt zu der bereits zuvor gemachten Feststellung, dass Wegunabhängigkeit des Kurvenintegrals genau dann vorliegt, wenn die Rotation null ist.

Hintergrund: Die Maxwell'schen Gleichungen

An dieser Stelle geschieht keine Herleitung der nach dem englischen Physiker James Clark Maxwell (1831–1879) benannten Gleichungen. Wir wollen lediglich auf den Zusammenhang mit den Sätzen von Gauß und Stokes aufmerksam machen. Die **Maxwell'schen Gleichungen** beschreiben Eigenschaften der folgenden Vektorfelder, die Abbildungen aus $\mathbb{R}^3$ in $\mathbb{R}^3$ sind:

- $\vec{E}$: elektrische Feldstärke, siehe Kapitel 4.5.1.

- $\vec{D}$: elektrische Erregung bzw. Verschiebungsdichte. Ein elektrisches Feld trennt positive und negative Ladungen in einer Metallplatte, die sich im Feld befindet. Die elektrische Erregung zeigt in Richtung der elektrischen Feldstärke, und ihr Betrag ist gleich der Flächenladungsdichte (Ladung pro Fläche), die entsteht, wenn man zwei aufeinanderliegende Metallplatten, die sich senkrecht zum elektrischen Feld befinden, in Feldrichtung trennt. In einem isotropen, homogenen Medium ist die elektrische Erregung proportional zur Feldstärke.
- $\vec{S}$: elektrische Stromdichte. $\vec{S}$ zeigt in Richtung der Bewegung positiver Ladungsträger. $|\vec{S}|$ ist ein Grenzwert von Stromstärke pro senkrecht zum Strom stehender Fläche, wobei man die Größe der Fläche gegen null gehen lässt.
- $\vec{B}$: magnetische Flussdichte (Induktion). $\vec{B}$ zeigt in Richtung des magnetischen Felds. Der Betrag der Größe ist über die Kraft $\vec{F}$ definiert, die auf einen stromdurchflossenen, geraden Leiter der Länge l vom Magnetfeld ausgeübt wird, wenn der Leiter senkrecht zum Magnetfeld steht. Fließt der Strom i durch den Leiter, so ergibt sich der Betrag von $\vec{B}$ aus dem Quotienten $\frac{\vec{F}}{l \cdot |i|}$, indem man die Seitenlänge l der Leiterschleife und die Stromstärke $|i|$ gegen null gehen lässt.
- $\vec{H}$: magnetische Erregung bzw. Feldstärke. In einem homogenen, isotropen Material (mit Ausnahme von ferro- und ferrimagnetischen Materialien) ist die magnetische Erregung proportional zur magnetischen Flussdichte: $\vec{B} = \mu \vec{H}$.

Zudem benötigt man die Raumladungsdichte ρ als reellwertige (skalare) Funktion, die die Verteilung von elektrischer Ladung im Raum beschreibt. Ist K_r eine Kugel mit Radius r (und Volumen $\frac{4}{3}\pi r^3$) um den Punkt (x, y, z) , so ist

$$\rho(x, y, z) := \lim_{r \rightarrow 0+} \frac{\text{Ladung der Kugel } K_r}{\frac{4}{3}\pi r^3}$$

und gibt damit die Ladungsdichte im Punkt (x, y, z) an.

In Differenzialform lauten die Maxwell'schen Gleichungen (ohne Bedingungen an das Verhalten an Grenzflächen):

$$\operatorname{div} \vec{B} = 0 \quad (4.24) \qquad \operatorname{rot} \vec{H} = \vec{S} + \frac{\partial}{\partial t} \vec{D} \quad (4.26)$$

$$\operatorname{div} \vec{D} = \rho \quad (4.25) \qquad \operatorname{rot} \vec{E} = -\frac{\partial}{\partial t} \vec{B}. \quad (4.27)$$

Die erste Gleichung besagt, dass die magnetische Flussdichte keine Quellen oder Senken hat. Nach (4.25) ist dies für die elektrische Erregung (und damit auch für die elektrische Feldstärke) anders. Hier sind Ladungen (beschrieben durch die Raumladungsdichte ρ) die Quellen und Senken. Die Gleichungen (4.26) und (4.27) beschreiben Wirbel der magnetischen bzw. elektrischen Feldstärke. Hier gehen zeitliche Änderungen der Vektorfelder $\vec{B}$ und $\vec{D}$ ein. Zeitliche Änderungen der magnetischen Flussdichte $\vec{B}$ führen gemäß (4.27) zu Wirbeln im elektrischen Feld. Zeitliche Änderungen im elektrischen Feld führen nach (4.26) zu Wirbeln der magnetischen Flussdichte, sofern $\vec{H}$ und $\vec{B}$ sowie $\vec{E}$ und $\vec{D}$ proportional sind. Mathematisch müssen wir eine Variable für die Zeit hinzufügen, z. B. $\vec{B}(x, y, z, t) : D \subset \mathbb{R}^4 \rightarrow \mathbb{R}^3$. Für feste Zeitpunkte t hat man dann weiterhin ein Vektorfeld.

Mit dem Satz von Gauß können (unter geeigneten Voraussetzungen) die Gleichungen (4.24) und (4.25) in die folgende Integralform gebracht werden:

$$0 = \iiint_A \operatorname{div} \vec{B}(x, y, z) d(x, y, z) = \int_{\partial A} \vec{B} \cdot \vec{N} d\sigma \quad (4.28)$$

$$\iiint_A \rho d(x, y, z) = \iiint_A \operatorname{div} \vec{D}(x, y, z) d(x, y, z) = \int_{\partial A} \vec{D} \cdot \vec{N} d\sigma. \quad (4.29)$$

Die magnetische Flussdichte hat keine Quellen und Senken. Alle Feldlinien, die A betreten, verlassen A wieder. Das Integral über die Oberfläche von A in (4.28) ist daher null. In der Gleichung (4.29) steht auf der linken Seite mit $\iiint_A \rho d(x, y, z)$ die Gesamtladung im Volumen A . Diese bestimmt die elektrische Erregung durch die Hüllenfläche von A .

Die Gleichungen (4.26) und (4.27) werden mit dem Satz von Stokes in die Integralform für eine geeignete Fläche A überführt, sofern Integration und Ableitung vertauscht werden dürfen (was man z. B. mit dem Satz von Fubini bewerkstelligen kann, vgl. Beweis zu Lemma 6.2 b) auf Seite 835):

$$\begin{aligned} \oint_{\partial A} \vec{H} d\vec{x} &= \int_A (\operatorname{rot} \vec{H}) \cdot \vec{N} d\sigma = \int_A \vec{S} \cdot \vec{N} + \left(\frac{\partial}{\partial t} \vec{D} \right) \cdot \vec{N} d\sigma \\ &= \int_A \vec{S} \cdot \vec{N} d\sigma + \frac{\partial}{\partial t} \int_A \vec{D} \cdot \vec{N} d\sigma \\ \oint_{\partial A} \vec{E} d\vec{x} &= \int_A (\operatorname{rot} \vec{E}) \cdot \vec{N} d\sigma = \int_A \left(-\frac{\partial}{\partial t} \vec{B} \right) \cdot \vec{N} d\sigma = -\frac{\partial}{\partial t} \int_A \vec{B} \cdot \vec{N} d\sigma. \end{aligned}$$

4.6 Aufgaben

Aufgabe 4.1

Berechnen Sie die folgenden Grenzwerte in $\mathbb{R}^n$:

$$\text{a) } \lim_{k \rightarrow \infty} \left(\frac{\sin k}{k}, \left(1 + \frac{1}{k} \right)^k \right), \quad \text{b) } \lim_{k \rightarrow \infty} \left(k \sin \frac{1}{k}, 1, \frac{k^2 + 2k + 1}{2k^2} \right).$$

Aufgabe 4.2

Man berechne die gemischten zweiten partiellen Ableitungen $\frac{\partial^2 f}{\partial x \partial y}$ und $\frac{\partial^2 f}{\partial y \partial x}$ der Funktion $f(x, y) = (x^2 + y^2) \sin \frac{1}{x^2 + y^2} + \sin(y) + 2x$ für $(x, y) \neq (0, 0)$ und zeige, dass diese nicht stetig ergänzbar in $(0, 0)$ sind.

Hinweis: Man betrachte einen Grenzwert entlang der Gerade $y = x$.

Aufgabe 4.3

Man berechne alle partiellen Ableitungen erster Ordnung der Funktionen

- a) $f(x, y) = x^2y + \sin(y) + \pi$, b) $f(x, y, z) = \sqrt{x^2 + y^2 + z^2}$,
 c) $f(x, y, z) = (xy)^z$ für $x, y > 0$.

Aufgabe 4.4

Man berechne die folgenden partiellen Ableitungen:

- a) $\frac{\partial^2}{\partial y \partial x} [x + x^2y + 3y^2]$, b) $\frac{\partial^2}{\partial x^2} 3e^{2x+y^2}$, c) $\frac{\partial^2}{\partial y \partial x} [3y \cos(x^2y)]$,
 d) $\frac{\partial^2}{\partial x^2} [y \sin(x^2 + y^2) + 3x^2 + 1]$.

Aufgabe 4.5

Bestimmen Sie den Gradienten der Funktion $f(x, y) := \frac{x}{\sqrt{x^2 + y^2}} - 3$ an der Stelle $(2, 3)$, also den Wert $[\text{grad } f](2, 3)$. Berechnen Sie damit die Richtungsableitung in Richtung des Vektors $(3, -2)$ an der Stelle $(2, 3)$.

Aufgabe 4.6

Berechnen Sie den Gradienten von f unter Verwendung der Kettenregel:

$$f(x, y, z) = g(\vec{h}(x, y, z)), \quad g(u, v, w) := 3u^2 + 4v + 16w, \quad \vec{h}(x, y, z) := (x^2 + 2z, y^3, x^2 - z^2).$$

Aufgabe 4.7

Man bestimme alle Stellen, an denen die Funktion $f(x, y) = (x + 1)^2 + y^2 + 5$ lokale Extrema besitzt.

Aufgabe 4.8

Man bestimme alle (aber maximal zwei) Stellen, an denen die Funktion $f(x, y) = y^2x - 4x$ lokale Extremstellen besitzen kann.

Aufgabe 4.9

Man bestimme das absolute Minimum der Funktion $f(x, y, z) = 3x^2 + 3y^2 + z^2 + 1$ unter der Nebenbedingung $x + y + z = 1$.

Aufgabe 4.10

a) Wir suchen ein Minimum der Funktion

$$f(x, y, z) := \sin(x^2 + y^2 + z^2 - 1) \exp(x)$$

unter der Nebenbedingung $g(x, y, z) := (x - 5)^2 + (y - 3)^2 + (z - 1)^2 = 0$. Stellen Sie dazu nach dem Satz über die Langrange-Multiplikatoren vier verschiedene Gleichungen auf, die notwendigerweise erfüllt sein müssen, wenn in (x, y, z) ein Minimum unter der Nebenbedingung liegt. Das Lösen des Gleichungssystems ist hier **nicht** verlangt.

b) Geben Sie analog zu a) das Gleichungssystem für die Funktion

$$f(x, y, z) := \exp(x^2 + y^2 + z^2 - 1) \sin(x)$$

und die Nebenbedingung $g(x, y, z) := (x - 3)^2 + (y - 2)^2 + (z - 1)^2 = 0$ an.

c) Wo hat die Funktion $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ mit $f(x, y, z) := x - y + 2z$ lokale Extremstellen unter der Nebenbedingung $x^2 + y^2 + 2z^2 = 2$, d. h. $g(x, y, z) = 0$ mit

$$g(x, y, z) := x^2 + y^2 + 2z^2 - 2 : \mathbb{R}^3 \rightarrow \mathbb{R}?$$

Die Nebenbedingung stellt einen Ellipsoid dar.

Aufgabe 4.11

Man berechne folgende Integrale:

- a) $\iint_{[0,2] \times [1,2]} 1 + x + y \, d(x, y),$ b) $\iint_{[0,2] \times [1,2]} y \sin(\pi xy) \, d(x, y),$
 c) $\iiint_{[0,1] \times [0,1] \times [0,1]} xyz \, d(x, y, z),$ d) $\iiint_{[0,1] \times [0,1] \times [0,1]} x^3 + y^2 + z \, d(x, y, z),$
 e) $\iint_{[0, \frac{\pi}{2}] \times [0, \frac{\pi}{2}]} \cos(y) \sin(x) \, d(x, y),$ f) $\iint_{[1,2] \times [0,1]} x^3 \cdot \exp(x^2 y) \, d(x, y).$

Aufgabe 4.12

a) B sei das durch die Geraden $x = 0$, $y = 0$ und $y = -\frac{1}{2}x + 1$ begrenzte Dreieck. Man berechne $\iint_B 2x^2 y \, d(x, y)$.

b) Es sei B der von den Kurven $y = -x$, $y = x^2$, $x = 1$ im ersten und vierten Quadranten berandete Bereich. Man berechne $\iint_B x^2 + 2xy + 1 \, d(x, y)$.

Aufgabe 4.13

Man berechne mittels Substitution $x = ar \cos(\varphi) \cos(\vartheta)$, $y = br \sin(\varphi) \cos(\vartheta)$, $z = cr \sin(\vartheta)$ das Volumen eines Ellipsoids mit Halbachsen $a, b, c > 0$:

$$\iiint_{\{(x,y,z): x^2/a^2 + y^2/b^2 + z^2/c^2 \leq 1\}} 1 \, d(x, y, z).$$

Aufgabe 4.14

Man gebe eine Formel für die kürzeste Entfernung zwischen zwei Punkten $(R, \varphi_1, \vartheta_1)$ und $(R, \varphi_2, \vartheta_2)$ auf der Oberfläche einer Kugel mit Radius R an. Die beiden Punkte in sind Kugelkoordinatendarstellung (Radius, Längen- und Breitengrad) gegeben. **Hinweis:** Verwenden Sie das Skalarprodukt zur Berechnung des Winkels zwischen den Strecken vom Ursprung zu den beiden Punkten in Euklid'scher Darstellung.



5 Gewöhnliche Differenzialgleichungen

Übersicht

5.1	Einführung	705
5.2	Lösungsmethoden für Differenzialgleichungen erster Ordnung	721
5.3	Lineare Differenzialgleichungssysteme	748
5.4	Lineare Differenzialgleichungen höherer Ordnung	770
5.5	Ausblick: Partielle Differenzialgleichungen und Finite-Elemente-Methode *	788
5.6	Aufgaben	809

Differenzialgleichungen sind von fundamentaler Bedeutung für die Ingenieurmathematik, da viele physikalische Gesetze durch Differenzialgleichungen formuliert sind. Denn oft verhalten sich Größen proportional zu Änderungsraten, also zu Ableitungen. Durch Lösen von Differenzialgleichungen werden wir u. a. die folgenden Fragen beantworten:

- Wie biegt sich das Seil einer Hängebrücke (Kettenlinie) durch?
- Wie alt ist ein fossiler Knochen?
- Wie können wir Kaffee durch Zugabe von Milch schnell abkühlen?
- Wie entwickeln sich die Ströme beim Einschalten einer Spannung in einem elektrischen Netzwerk?

Mit der Herleitung von Differenzialgleichungen aus physikalischen oder technischen Aufgabenstellungen beschäftigt sich die **Mathematische Modellierung**.

5.1 Einführung

In diesem Kapitel erarbeiten wir Lösungsmethoden für gegebene Differenzialgleichungen aus der Praxis. Wir betrachten Methoden, die auf geschlossene Lösungen führen.

Das sind Lösungen, die man mit Formeln exakt angeben kann. Die meisten technisch interessanten Differenzialgleichungen sind jedoch nicht geschlossen lösbar. Hier kommen Näherungsverfahren zur Anwendung, die wir auch kurz ansprechen.

5.1.1 Beispiele für Differenzialgleichungen aus Physik und Technik

Differenzialgleichungen sind Gleichungen, in denen gesuchte Funktionen (die häufig mit y bezeichnet werden) und Ableitungen dieser Funktionen auftreten. Wir beginnen mit einigen Beispielen.

Beispiel 5.1

Beim freien Fall eines Steins ohne Berücksichtigung des Luftwiderstands sei $y(t)$ die zurückgelegte Strecke. Die Fallbeschleunigung $y''(t)$ ist gleich der Gravitationskonstante $g = 9,81 \text{ m/s}^2$, d. h., es gilt die Differenzialgleichung

$$y''(t) = g.$$

Durch Integration erhalten wir die Fallgeschwindigkeit

$$y'(t) = gt + v_0$$

mit v_0 als Anfangsgeschwindigkeit des Steins zur Zeit $t = 0$. Nochmalige Integration ergibt das Weg-Zeit-Gesetz

$$y(t) = \frac{1}{2}gt^2 + v_0t + s_0$$

mit s_0 als Anfangshöhe zur Zeit $t = 0$. Bei diesem einfachen Beispiel können wir die Differenzialgleichung durch elementare Integration lösen. Kompliziertere Beispiele erfordern speziellere Methoden, die wir uns in diesem Kapitel ansehen. ■

Beispiel 5.2

In einem Stromkreis (R/L-Kreis), in dem eine Spannungsquelle mit Spannung $u(t)$, ein Widerstand R und eine Induktivität hintereinandergeschaltet sind, ergibt sich für die Stromstärke aus der Kirchhoff'schen Maschengleichung und den Bauteilgesetzen von Ohm und Faraday die Differenzialgleichung

$$Li'(t) + Ri(t) = u(t). \quad (5.1)$$

Für sinus- oder kosinusförmige Spannungen $u(t) = \hat{u} \cos(\omega t + \varphi_u)$ kann man die Stromstärke über die komplexe Wechselstromrechnung ermitteln (siehe Kapitel 1.6.5). Für andere Spannungsfunktionen (z. B. beim Einschalten) muss man aber die Differenzialgleichung lösen. ■

Eine sehr wichtige Rolle beim Lösen von Differenzialgleichungen spielt die Exponentialfunktion $\exp(x) = e^x$, die gleich ihrer Ableitung ist. Sie erfüllt also die Differenzialgleichung

$$y'(x) - y(x) = 0.$$

Betrachten wir als Erweiterung die Funktion $y(x) = ce^{ax}$ für Konstanten $a, c \in \mathbb{R}$, so ist $y'(x) = \frac{d}{dx}(ce^{ax}) = ace^{ax} = ay(x)$, und $y(x)$ erfüllt damit die Differenzialgleichung

$$y'(x) - ay(x) = 0. \quad (5.2)$$

Für $u(t) = 0$ erhalten wir somit eine Lösung $i(t) = ce^{-\frac{R}{L}t}$ der Gleichung (5.1). Gibt man die Stromstärke $i(0) = i_0$ als Anfangsbedingung zum Zeitpunkt $t = 0$ vor, so erhalten wir dazu den Stromverlauf $i(t) = i_0 e^{-\frac{R}{L}t}$.

Beispiel 5.3

Die Zerfallsrate einer radioaktiven Substanz ist proportional zur vorhandenen Menge der Substanz. Wir bezeichnen mit $y(t)$ die Menge der Substanz zum Zeitpunkt t . Die Ableitung $\frac{d}{dt}y(t)$ ist die Änderungsrate der Menge und entspricht der Zerfallsrate. Damit lautet das Zerfallsgesetz

$$y'(t) \sim y(t),$$

d. h., es gibt eine Konstante $k \in \mathbb{R}$ mit

$$y'(t) = k \cdot y(t). \quad (5.3)$$

Dies ist die Differenzialgleichung für den radioaktiven Zerfall. Die materialabhängige Konstante k ist negativ, da ja die Menge mit der Zeit abnimmt. Beispielsweise gilt für das Radium-Isotop ^{226}Ra die Zerfallskonstante $k \approx -1,4 \cdot 10^{-11} \text{ s}^{-1}$.

Liegt zur Zeit $t = 0$ die Menge $y_0 = y(0)$ der Substanz vor, so erhält man $y(t)$ als Lösung eines sogenannten **Anfangswertproblems**

$$y'(t) = k \cdot y(t), \quad y(0) = y_0.$$

Wie im vorherigen Beispiel erfüllt die Funktion ce^{kt} die Differenzialgleichung. Wählt man speziell $c = y_0$, so ist auch die Anfangsbedingung $y(0) = y_0$ erfüllt. ■

Als nächstes Beispiel betrachten wir ein einfaches Feder-Masse-Dämpfer-System.

Beispiel 5.4

Die Beschleunigung $x''(t)$ der Masse m in Abbildung 5.1 ist bestimmt durch die durch die Feder und den Dämpfer ausgeübten Kräfte $F_c(t)$ und $F_b(t)$:

$$m x''(t) = F_c(t) + F_b(t). \quad (5.4)$$

Die beiden Kräfte sind durch die Feder mit Federkonstante c und Nullposition u sowie durch den Dämpfer mit Reibungskoeffizient b festgelegt:

$$F_c(t) = c(u - x(t)) \text{ (lineares Federgesetz) und } F_b(t) = -b x'(t) \text{ (lineare Dämpfung).}$$

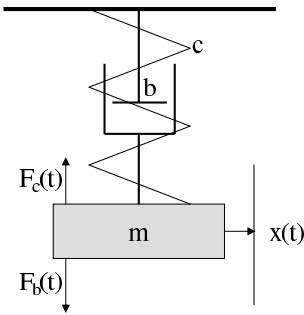


Abb. 5.1: Feder-Masse-Dämpfer-System

Einsetzen in die Gleichung (5.4) ergibt die Differenzialgleichung des Feder-Masse-Dämpfer-Systems für die zeitabhängige Auslenkung $x(t)$ der Masse

$$m x''(t) + b x'(t) + c x(t) = c u.$$

Hierbei handelt es sich um eine lineare Differenzialgleichung zweiter Ordnung (wegen der zweiten Ableitung x''), die wir auf Seite 774 für die Nullposition $u = 0$ lösen. ■

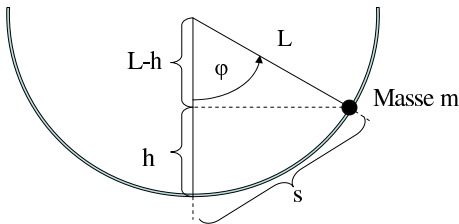


Abb. 5.2: Mathematisches Pendel

Beispiel 5.5

An einem Pendel der Länge L schwingt eine Masse m . Die Auslenkung lässt sich sowohl durch den zeitabhängigen Winkel $\varphi(t)$ als auch durch die Länge des Bogens $s = s(t)$ (siehe Abbildung 5.2) messen. Um die Auslenkung in Abhängigkeit der Zeit zu beschreiben, leiten wir eine Differenzialgleichung her und starten mit dem Gesetz der Energieerhaltung, d. h.

$$E_{\text{kin}} + E_{\text{pot}} = \text{const.},$$

d. h., die Summe aus kinetischer und potentieller Energie der bewegten Masse ist konstant. Dabei gilt

$$E_{\text{kin}} = \frac{1}{2} m \cdot v^2(t) \quad \text{und} \quad E_{\text{pot}} = m \cdot g \cdot h,$$

wobei $g = 9,81 \text{ m/s}^2$ die Gravitationskonstante, h die jeweilige Höhe der Masse und v ihre Geschwindigkeit ist. Aus Abbildung 5.2 erhalten wir:

$$s(t) = L \cdot \varphi(t), \quad h(t) = L - L \cos(\varphi(t)) = L(1 - \cos(\varphi(t))).$$

Wir setzen die Gleichungen in den Energieerhaltungssatz ein:

$$\begin{aligned} \frac{1}{2} m \cdot s'(t)^2 + m \cdot g \cdot h(t) &= \text{const.} \\ \iff \frac{1}{2} m \cdot L^2 \varphi'(t)^2 + m \cdot g \cdot L - m \cdot g \cdot L \cos(\varphi(t)) &= \text{const.} \end{aligned}$$

Ableiten dieser Beziehung ergibt $m \cdot L^2 \varphi'(t) \varphi''(t) + m \cdot g \cdot L \sin(\varphi(t)) \cdot \varphi'(t) = 0$.
Division durch $m L^2 \varphi'(t)$ liefert

$$\varphi''(t) + \frac{g}{L} \sin(\varphi(t)) = 0.$$

Es handelt sich hier um eine nicht-lineare Differenzialgleichung zweiter Ordnung (nicht-linear, da die gesuchte Funktion im Argument des Sinus vorkommt, zweiter Ordnung, da zweite Ableitungen vorkommen), die Differenzialgleichung des **mathematischen Pendels**. Nicht-lineare Differenzialgleichungen höherer Ordnung werden wir hier nicht betrachten. Ist die Auslenkung aber klein, so ist $\sin(\varphi(t)) \approx \varphi(t)$, so dass wir zu einer linearen Differenzialgleichung zweiter Ordnung

$$\varphi''(t) + \frac{g}{L} \varphi(t) = 0$$

gelangen. Dies ist eine homogene Schwingungsgleichung, siehe Kapitel 5.4.3. ■

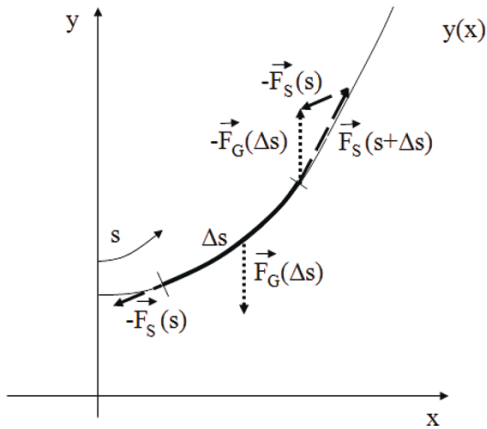


Abb. 5.3: Seilkurve, Kettenlinie

Beispiel 5.6

In diesem Beispiel beschreiben wir, welche Linie ein an zwei Stellen befestigtes Seil beschreibt. Das Seil sei vollkommen biegsam und undehnbar. Außerdem sei sein Gewicht pro Längeneinheit konstant gleich m .

Wir betrachten den Verlauf des Seils als Funktionsgraph zu $y(x)$ von der tiefsten Stelle bei $x = 0$ bis zur Aufhängung. Auf ein beliebiges Seilstück der Länge Δs wirkt die Gravitationskraft $\vec{F}_G(\Delta s) = (0, -gm\Delta s)$ nach unten, wobei g die Erdbeschleunigung ist. Die Gravitationskraft wird kompensiert durch eine Kraft tangential zum Seil in Richtung der Aufhängung. Die Größe und Richtung dieser Kraft ist abhängig von der Position. Wir messen die Position nicht mittels der x -Koordinate, sondern über die Entfernung $s = s(x)$ vom Punkt der Kurve an der Stelle x entlang der Seilkurve bis zum tiefsten Punkt an der Stelle 0. Damit nennen wir die von s abhängende Kraft $\vec{F}_S(s) = (f_x(s), f_y(s))$. Auf ein Seilstück der Länge Δs , das an der Position s beginnt, wirkt so die Kraft $\vec{F}_S(s + \Delta s) - \vec{F}_S(s)$. Damit ist $\vec{F}_G(\Delta s) + \vec{F}_S(s + \Delta s) - \vec{F}_S(s) = \vec{0}$. Division durch Δs und Grenzübergang $\Delta s \rightarrow 0+$ ergibt:

$$\vec{0} = \lim_{\Delta s \rightarrow 0+} \frac{\vec{F}_G(\Delta s)}{\Delta s} + \frac{\vec{F}_S(s + \Delta s) - \vec{F}_S(s)}{\Delta s} = (0, -gm) + \frac{d}{ds} \vec{F}_S(s).$$

Damit ist $f'_x(s) = 0$ und $f'_y(s) = gm$. Integrieren wir beide Gleichungen, so erhalten wir $f_x(s) = c$ und $f_y(s) = gms + d$. Da an der Stelle $s = 0$ die y -Komponente der Seil-Kraft verschwindet, ist $d = 0$. Die Kraft $\vec{F}_S$ wirkt tangential zur Seilkurve. Für die Steigung der Tangente gilt mit den Komponenten der Kraft

$$\frac{d}{dx} y(x) = \frac{f_y(s(x))}{f_x(s(x))} = \frac{gms(x)}{c}.$$

Differenziation der Gleichung nach x liefert

$$\frac{d^2}{dx^2} y(x) = \frac{gm}{c} \frac{d}{dx} s(x) = \frac{gm}{c} \frac{d}{dx} \int_0^x \sqrt{1 + \left(\frac{d}{du} y(u) \right)^2} du = \frac{gm}{c} \sqrt{1 + (y'(x))^2}.$$

Dabei haben wir die Formel zur Längenberechnung einer Kurve von Seite 676 für die Kurve $((u, y(u)), [0, x])$ benutzt. Die Differenzialgleichung der Seilkurve oder der Kettenlinie lautet für eine Konstante $k \in \mathbb{R}$:

$$y''(x) = k \sqrt{1 + (y'(x))^2}.$$

Diese Gleichung lösen wir später auf Seite 738. ■

5.1.2 Grundbegriffe

Um den Begriff Differenzialgleichung exakt zu definieren, benutzen wir Funktionen F und f mit mehreren Variablen, mit denen wir uns nach Kapitel 4 gut auskennen. Dabei entspricht die erste Variable von F bzw. f der Variable der gesuchten Lösungsfunktion, in die weiteren Variablen werden die gesuchte Funktion und ihre Ableitungen eingesetzt. So lassen sich Differenzialgleichungen ganz allgemein formulieren und Eigenschaften der Gleichungen (wie z. B. eindeutige Lösbarkeit) mit Eigenschaften der Funktionen F und f in Verbindung bringen.

Definition 5.1 (Gewöhnliche Differenzialgleichung)

Seien $n \in \mathbb{N}$ (die maximale Ordnung der auftretenden Ableitungen), $D \subset \mathbb{R}^{n+2}$, $G \subset \mathbb{R}^{n+1}$, $F : D \rightarrow \mathbb{R}$ und $f : G \rightarrow \mathbb{R}$.

- Eine Bestimmungsgleichung für Funktionen $y = y(x)$ der Form

$$F(x, y(x), y'(x), \dots, y^{(n)}(x)) = 0 \quad (5.5)$$

heißt **implizite gewöhnliche Differenzialgleichung n -ter Ordnung**.

- Lässt sich eine implizite Differenzialgleichung überführen in die Form

$$y^{(n)}(x) = f(x, y(x), y'(x), \dots, y^{(n-1)}(x))$$

für eine geeignete Funktion f , so spricht man von einer **expliziten gewöhnlichen Differenzialgleichung n -ter Ordnung**.

- Eine auf einem Intervall $]a, b[$ n -mal differenzierbare Funktion $y :]a, b[\rightarrow \mathbb{R}$ heißt genau dann **Lösung** der impliziten Differenzialgleichung auf $]a, b[$, wenn die folgenden Bedingungen für alle $x \in]a, b[$ erfüllt sind:

i) $(x, y(x), y'(x), \dots, y^{(n)}(x)) \in D$.

ii) $F(x, y(x), y'(x), \dots, y^{(n)}(x)) = 0$.

y ist genau dann Lösung der expliziten Differenzialgleichung auf $]a, b[$, wenn für alle $x \in]a, b[$ gilt:

i) $(x, y(x), y'(x), \dots, y^{(n-1)}(x)) \in G$,

ii) $y^{(n)}(x) = f(x, y(x), y'(x), \dots, y^{(n-1)}(x))$.

Die Bezeichnung „gewöhnlich“ bedeutet, dass Funktionen einer (reellen) Variable gesucht sind. Im Gegensatz dazu werden bei **partiellen Differenzialgleichungen** Funktionen mit mehreren Variablen gesucht, z. B. eine Funktion $u = u(x, t)$, die die Gleichung

$$\frac{\partial^2 u}{\partial t^2}(x, t) = c \frac{\partial^2 u}{\partial x^2}(x, t)$$

erfüllt (siehe Seite 788). Dies ist eine Differenzialgleichung zweiter Ordnung, da zweite partielle Ableitungen vorkommen. Der Name „partielle Differenzialgleichung“ leitet sich von der Verwendung partieller Ableitungen ab. Abgesehen von einem kurzen Ausblick in Kapitel 5.5 werden wir nicht auf das umfangreiche Spezialgebiet der partiellen Differenzialgleichungen eingehen.

Beispiel 5.7 (Anwendung der Definition)

Wir betrachten die Eingangsbeispiele vor dem Hintergrund der Definition:

- a) Die Differenzialgleichung des freien Falls $y''(t) = g$ kann geschrieben werden als $F(t, y(t), y'(t), y''(t)) = 0$ mit $F(t, u, v, w) := w(t) - g$. In expliziter Form ist $y''(t) = f(t, y(t), y'(t))$ mit $f(t, u, v) := g$.
- b) Im Beispiel (5.1) eines R/L-Kreises ist $F(t, v, w) := Lw + Rv - u(t)$. In expliziter Form lautet die Gleichung $\frac{d}{dt}i(t) = -\frac{R}{L}i(t) + \frac{u(t)}{L}$, also $\frac{d}{dt}i(t) = f(t, i(t))$ mit $f(t, v) := -\frac{R}{L}v + \frac{u(t)}{L}$.
- c) Die Gleichung $y'(t) - ky(t) = 0$ des radioaktiven Zerfalls lässt sich $F(t, y(t), y'(t)) = 0$ mit $F(t, v, w) := w - kv$ oder explizit $y'(t) = f(t, y(t))$ mit $f(t, v) := kv$.
- d) Beim Feder-Masse-Dämpfer-System können wir die Gleichung $mx''(t) + bx'(t) + cx(t) = cu(t)$ schreiben als $F(t, x(t), x'(t), x''(t)) = 0$ mit $F(t, v, w, z) := mz + bw + cv - cu(t)$ oder als $x''(t) = f(t, x(t), x'(t))$ mit $f(t, v, w) := -\frac{b}{m}w - \frac{c}{m}v + \frac{c}{m}u(t)$.
- e) Die Gleichung $\varphi''(t) + \frac{g}{L}\sin\varphi(t) = 0$ des mathematischen Pendels ist äquivalent zu $F(t, \varphi(t), \varphi'(t), \varphi''(t)) = 0$ mit $F(t, v, w, z) := z + \frac{g}{L}\sin v$ bzw. $\varphi''(t) = f(t, \varphi(t), \varphi'(t))$ mit $f(t, v, w) := -\frac{g}{L}\sin v$.
- f) Die über $y''(x) = k\sqrt{1 + (y'(x))^2}$ beschriebene Kettenlinie kann auch so dargestellt werden: $F(x, y(x), y'(x), y''(x)) = 0$ mit $F(x, v, w, z) := z - k\sqrt{1 + w^2}$ oder $y''(x) = f(x, y(x), y'(x))$ mit $f(x, v, w) := k\sqrt{1 + w^2}$. ■

Beispiel 5.8

Gesucht sind alle Lösungen von $y'(x) = f(x, y(x))$ mit $f(x, y) = 1$, also von $y'(x) = 1$. Offensichtlich erfüllt jede Funktion $y(x) = x + c$ mit $c \in \mathbb{R}$ die Differenzialgleichung. Haben wir damit bereits alle Lösungen gefunden? Sind y_1, y_2 zwei beliebige Lösungen der Differenzialgleichung, so gilt: $\frac{d}{dx}[y_1(x) - y_2(x)] = y_1'(x) - y_2'(x) = 1 - 1 = 0$. Aus dem Mittelwertsatz der Differenzialrechnung folgt (siehe Folgerung 2.5 auf Seite 387), dass $y_1 - y_2$ konstant ist. Lösungen unterscheiden sich nur durch eine Konstante, wir haben alle Lösungen gefunden. ■

Beispiel 5.9

Gesucht sind alle Lösungen von $y'(x) = f(x, y(x))$ mit $f(x, y) = 2x$, d. h. $y'(x) = 2x$ (siehe Abbildung 5.4). Gesucht sind also Stammfunktionen von $2x$. Alle Stammfunktionen und damit alle Lösungen der Gleichung sind: $y(x) = \int 2x dx = x^2 + c$. ■

Häufig kennt man zusätzlich zur Differenzialgleichung noch Nebenbedingungen, z. B. eine Ausgangsstromstärke $i(0)$ oder eine Ausgangsposition s_0 zusammen mit einer Ausgangsgeschwindigkeit v_0 . Man sucht dann nach Lösungen, die diese Nebenbedingung erfüllen:

Definition 5.2 (Anfangswertproblem)

Gegeben sei eine explizite Differenzialgleichung

$$y^{(n)}(x) = f(x, y(x), y'(x), \dots, y^{(n-1)}(x))$$

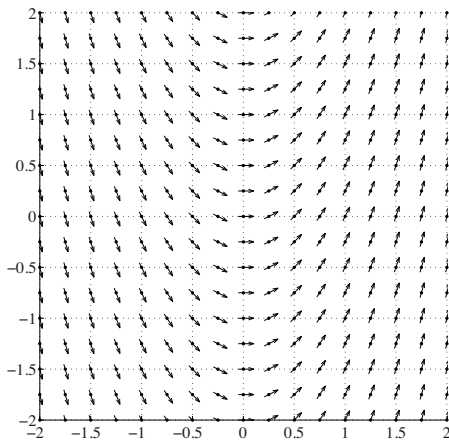


Abb. 5.4: $y'(x) = f(x, y(x))$ mit $f(x, y) = 2x$, die Pfeile haben genau die von der rechten Seite der Differentialgleichung geforderte Steigung an der Stelle (x, y) .

mit $n \in \mathbb{N}$, $G \subset \mathbb{R}^{n+1}$ und $f : G \rightarrow \mathbb{R}$. Weiter seien zu einem $x_0 \in \mathbb{R}$ Werte für $y(x_0), y'(x_0), \dots, y^{(n-1)}(x_0)$ vorgegeben, so dass $(x_0, y(x_0), y'(x_0), \dots, y^{(n-1)}(x_0)) \in G$. Dies ist eine **Anfangsbedingung**.

Gesucht ist eine Lösung $y :]a, b[\rightarrow \mathbb{R}$ der Differentialgleichung, die an der Stelle $x_0 \in]a, b[$ die Anfangsbedingung erfüllt. Diese Aufgabe heißt **Anfangswertproblem**.

Die Bezeichnung „Anfangsbedingung“ ist etwas willkürlich, da es sich bei x_0 um irgendeine Stelle handeln kann. Häufig ist aber $x_0 = t_0 = 0$ der Zeitpunkt, für den man initiale Bedingungen wie z. B. eine Ausgangsspannung kennt. Man ist dann an einer Lösung für $x > x_0$ interessiert. Gibt man dagegen Funktionswerte an verschiedenen Stellen vor, so spricht man von einer **Randbedingung** (vgl. auch Kasten auf Seite 696).

Bei der Anfangsbedingung sind alle Werte vorgegeben, die man benötigt, um die rechte Seite der Differentialgleichung für $x = x_0$ auszurechnen, d. h., man hat alle Daten, um hier den Wert von f zu bestimmen.

Beispiel 5.10 (Anfangsbedingungen legen Freiheitsgrade fest)

Die folgenden Differentialgleichungen werden durch Integration gelöst. Bei jeder Integration kommt eine Konstante als freier Parameter in die Lösung. Die Ordnung der Differentialgleichung bestimmt, wie oft wir integrieren müssen und legt damit die Anzahl der Parameter fest. Es gibt also unendlich viele Lösungen, die sich durch die Wahl der Parameter (Freiheitsgrade) unterscheiden. Jede Vorgabe durch eine Anfangs- oder Randbedingung eliminiert einen freien Parameter:

- a) $\frac{d}{dt}x(t) = 10$ ist eine lineare Differentialgleichung erster Ordnung ohne Bedingung: Wir erwarten eine freie Konstante. $\frac{dx(t)}{dt} = 10 \implies x(t) = 10t + C$.

- b) Bei $\frac{d^2}{dt^2}x(t) = 2t$, $x(0) = 2$, handelt es sich um eine lineare Differenzialgleichung zweiter Ordnung mit einer Anfangsbedingung: Wir erwarten eine freie Konstante.

$$\frac{d^2x(t)}{dt^2} = 2t \implies \frac{dx(t)}{dt} = t^2 + c \implies x(t) = \frac{1}{3}t^3 + c \cdot t + d.$$

Die Anfangsbedingung liefert $x(0) = d = 2$. Damit lautet die gesuchte Lösung $x(t) = \frac{1}{3}t^3 + c \cdot t + 2$.

- c) $\frac{d^2}{dt^2}x(t) = \sin(3t)$, $x\left(\frac{\pi}{6}\right) = 0$, $x\left(\frac{\pi}{2}\right) = \frac{2}{9}$, ist eine lineare Differenzialgleichung zweiter Ordnung mit zwei Randwerten: Wir erwarten keine freie Konstante.

$$\frac{d^2x(t)}{dt^2} = \sin(3t) \implies \frac{d}{dt}x(t) = -\frac{1}{3}\cos(3t) + c \implies x(t) = -\frac{1}{9}\sin(3t) + c \cdot t + d.$$

Die Randbedingung liefert $x\left(\frac{\pi}{6}\right) = 0$ und $x\left(\frac{\pi}{2}\right) = \frac{2}{9}$, d. h.

$$\begin{aligned} 0 &= -\frac{1}{9}\sin\left(\frac{\pi}{2}\right) + c\frac{\pi}{6} + d \implies \frac{\pi}{6}c + d = \frac{1}{9}, \\ \frac{2}{9} &= -\frac{1}{9}\sin\left(\frac{3\pi}{2}\right) + c\frac{\pi}{2} + d \implies \frac{\pi}{2}c + d = \frac{1}{9}. \end{aligned}$$

Die Lösung des Gleichungssystems lautet $c = 0$, $d = \frac{1}{9}$, und somit ist $x(t) = -\frac{1}{9}\sin(3t) + \frac{1}{9}$.

- d) $\frac{d}{dt}x(t) + 2t = 0$, $x(1) = 7$: Bei dieser linearen Differenzialgleichung erster Ordnung mit einer Anfangsbedingung erwarten wir keine freie Konstante.

$$\frac{d}{dt}x(t) = -2t \implies x(t) = -t^2 + c.$$

Wegen $x(1) = -1 + c = 7$ ist $c = 8$. Damit: $x(t) = -t^2 + 8$. ■

Bei der Modellierung einer technischen Aufgabenstellung durch ein Anfangswertproblem spricht man von einem **korrekt gestellten Problem**, wenn es eine eindeutige Lösung gibt (damit beschäftigen wir uns im nächsten Abschnitt) und wenn sich bei einer kleinen Änderung der Daten (Anfangswerte, Inhomogenität) auch die eindeutige Lösung nur geringfügig ändert. Die letzte Eigenschaft heißt **Stabilität**. Man benötigt sie, da die Daten häufig gemessen werden und damit nur näherungsweise zu bestimmen sind. Außerdem macht man bei numerischen Lösungen mittels Computer Diskretisierungs- und Rundungsfehler, so dass man auf die Gutmütigkeit der Lösung angewiesen ist. Ein bekanntes nicht-stabiles Problem ist die Berechnung des Wetters. Kleine Ursachen („Flügel Schlag eines Schmetterlings“) können mittel- bis langfristig zu einem völlig anderen Wetter führen. Gutmütigkeit von Funktionen bezüglich kleiner Änderungen ihrer Argumentwerte haben wir schon früher diskutiert. Der Begriff war nicht Stabilität, sondern Stetigkeit (siehe Seite 274).

5.1.3 Konstruktion einer Lösung, Existenz und Eindeutigkeit

Nur in einfachen Fällen lassen sich Differenzialgleichungen mittels Integration lösen. Bevor wir uns praktische Lösungsverfahren ansehen, untersuchen wir zunächst generell die Lösbarkeit.

Eine Näherungslösung für ein Anfangswertproblem erster Ordnung erhält man so: Ist $y(x)$ eine Lösung der Differenzialgleichung, die durch den Punkt (x_0, y_0) geht, für die also $y(x_0) = y_0$ gilt, so hat $y(x)$ hier die Steigung $y'(x_0) = f(x_0, y_0)$. Die Lösung $y(x)$ hat also eine Tangente mit Steigung $f(x_0, y_0)$ an der Stelle x_0 .

Zeichnet man zu jedem Punkt (x, y) einen Pfeil (als Richtungsvektor, man spricht auch von einem **Linienelement**, siehe Abbildung 5.4) in Richtung der Tangente, so kann man Lösungen durch Verfolgen dieser Vektoren ablesen. Die Pfeile bilden ein Vektorfeld (siehe Kapitel 4.5.1), das sogenannte **Richtungsfeld**. Lösungen erhält man als Feldlinien.

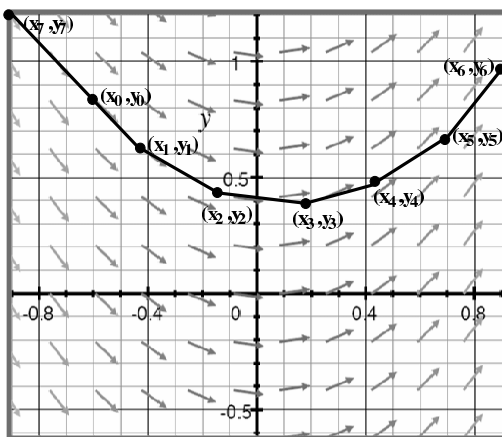


Abb. 5.5: Euler-Polygonzugmethode

Die **Euler-Cauchy-Polygonzugmethode** macht sich dies zunutze. Am Anfangswert (x_0, y_0) beginnend zeichnet man eine kurze Strecke mit Steigung $f(x_0, y_0)$ nach rechts. Am so gewonnenen neuen Punkt (anderer Endpunkt der Strecke) (x_1, y_1) schließt man eine weitere Strecke mit Steigung $f(x_1, y_1)$ an, die zum Punkt (x_2, y_2) führt usw. Entsprechend konstruiert man den Graphen auch nach links und erhält so die Näherung für den Funktionsgraphen einer Lösung des Anfangswertproblems (siehe Abbildung 5.5).

Dieses einfache Verfahren kann auch zur numerischen Berechnung von Lösungen mit dem Computer genutzt werden, falls z. B. exakte Ansätze nicht funktionieren. Man berechnet beginnend beim Punkt x_0 approximative Funktionswerte der Lösung über die Endpunkte der jeweiligen Strecken:

$$y_{k+1} = y_k + (x_{k+1} - x_k)f(x_k, y_k).$$

In der Praxis werden bessere Verfahren eingesetzt, z. B., das Runge-Kutta-Verfahren, das eine Weiterentwicklung der Polygonzugmethode ist, siehe Kapitel 5.1.5.

Peano (1858–1932) hat die Polygonzugmethode aufgegriffen, um die Existenz von Lösungen für Anfangswertprobleme zu beweisen. Verlangt man zusätzliche Eigenschaften von f und G , so erhält man darüber hinaus die Eindeutigkeit der Lösung. Diese ist nicht selbstverständlich:

Beispiel 5.11

Das Anfangswertproblem $y'(x) = f(x, y(x))$, $y(0) = 0$ mit (vgl. Definition 1.29 auf Seite 89)

$$f(x, y) := 2x \operatorname{sign}(y), \quad \operatorname{sign}(y) := \begin{cases} 1, & y > 0 \\ 0, & y = 0 \\ -1, & y < 0 \end{cases},$$

hat neun Lösungen $y(x)$, die für $x < 0$ und für $x \geq 0$ jeweils mit einer der Funktionen $y_1(x) = x^2$, $y_2(x) = -x^2$, $y_3(x) = 0$ übereinstimmen (siehe Abbildung 5.6). ■

Die eindeutige Lösbarkeit eines Anfangswertproblems hängt im Wesentlichen von der Funktion $f(x, y)$ ab. Dies besagt der Satz von Picard (1856–1941) und Lindelöf (1870–1946):

Satz 5.1 (Existenz- und Eindeutigkeitssatz, Variante 1)

Gegeben sei das Anfangswertproblem

$$y'(x) = f(x, y), \quad y(x_0) = y_0.$$

Ist $f(x, y)$ in einer Umgebung des Punktes (x_0, y_0) stetig (siehe Definition 4.4 auf Seite 596) und existiert dort die partielle Ableitung $\frac{\partial f}{\partial y}(x, y)$ und ist beschränkt, so existiert auf einer hinreichend kleinen Umgebung von x_0 genau eine Lösung $y(x)$.

Bemerkungen:

- Die Differenzialgleichung $y'(x) = f(x, y(x)) := \sqrt{y(x)}$ ist nur für $y \geq 0$ definiert. In jeder Umgebung eines Punktes $(x_0, 0)$ liegen Punkte mit negativer y -Koordinate, so dass in $(0, 0)$ die Voraussetzung des Satzes nicht erfüllt ist und man keine Lösung erwarten kann, die die Anfangsbedingung $y(x_0) = 0$ erfüllt.
- Wichtig ist auch, dass die Differenzialgleichung explizit vorliegt: Die implizite Differenzialgleichung

$$F(x, y(x), y'(x)) = x^2 + y^2(x) + (y'(x))^2 + 1 = 0$$

hat keine Lösung, obwohl $F(x, v, w) = x^2 + v^2 + w^2 + 1$ beliebig oft stetig differenzierbar ist. Es gilt nämlich $F(x, v, w) \geq 1$ für alle x, v, w , so dass der Wert null nicht angenommen werden kann.

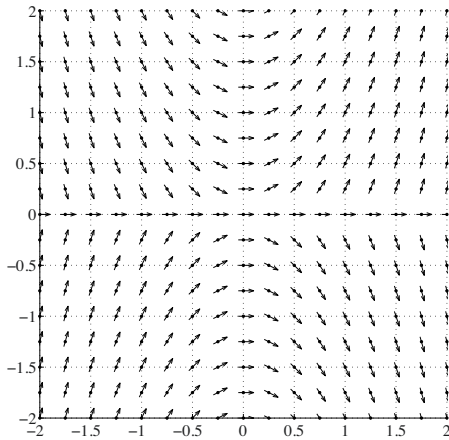


Abb. 5.6: Fehlende Eindeutigkeit: Im Punkt $(0, 0)$ können die Lösungen verzweigen.

- Die partielle Ableitung $\frac{\partial}{\partial y} f(x, y)$ der Funktion $f(x, y) = 2x \operatorname{sign}(y)$ existiert an den Stellen $(x, 0)$, $x \neq 0$, nicht und ist damit auch nicht in einer Umgebung des Punktes $(0, 0)$ erklärt. Die fehlende Eindeutigkeit ist daher kein Widerspruch zum Satz.
- Diese Fassung des Satzes sagt nichts über den maximalen Definitionsbereich der Lösungen aus. Außerdem gibt es auch dann noch eindeutige Lösungen, wenn f nicht partiell differenzierbar ist. Wir zitieren daher eine zweite Fassung des Satzes, die allerdings zusätzliche Begriffe benötigt.

Definition 5.3 (Lipschitz-Bedingung*)

Sei $\emptyset \neq G \subset \mathbb{R}^2$ und $f : G \rightarrow \mathbb{R}$. f erfüllt genau dann auf G eine **Lipschitz-Bedingung**, wenn es eine Konstante $L > 0$ gibt, so dass für alle Punkte $(x, y_1), (x, y_2) \in G$ gilt:

$$|f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2|.$$

Für ein festes x folgt aus der Lipschitz-Bedingung insbesondere die (gleichmäßige) Stetigkeit der Funktion $g(y) := f(x, y)$.

Satz 5.2 (Existenz- und Eindeutigkeitsatz, Variante 2*)

Die Funktion f genüge auf dem Gebiet G (siehe Definition 4.30 auf Seite 682) einer Lipschitz-Bedingung und sei stetig. Dann hat das Anfangswertproblem

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0,$$

für jeden Anfangswert $(x_0, y_0) \in G$ eine eindeutige Lösung $y = y(x)$, die sich beidseitig bis zum Rand von G erstreckt.

Eine Beweisskizze kann z. B. in (Meyberg und Vachenauer, 1997, Band 2, S. 54) nachgelesen werden.

Bemerkungen:

- Satz 5.1 folgt aus dieser Formulierung, indem man G als kleine Umgebung von (x_0, y_0) wählt. Die Lipschitz-Bedingung folgt dann aus dem Mittelwertsatz und der Beschränktheit der partiellen Ableitung nach y mit einer Schranke M :

$$|f(x, y_1) - f(x, y_2)| = \left| \frac{\partial f}{\partial y}(x, \xi) \right| |y_1 - y_2| \leq M |y_1 - y_2|.$$

- Im Beispiel $y'(x) = f(x, y) := 2x \operatorname{sign}(y)$ gab es mehrere Lösungen des Anfangswertproblems zu $y(0) = 0$. Damit muss eine Voraussetzung von Satz 5.2 verletzt sein. $G = \mathbb{R}^2$ ist ein Gebiet, aber die Lipschitz-Bedingung ist verletzt: Wählen wir $x = 1$ und $y_1 > 0$, $y_2 < 0$, dann ist

$$|f(1, y_1) - f(1, y_2)| = |2 + 2| = 4.$$

Zu einem beliebigen $L \in \mathbb{R}$ setzen wir konkret $y_1 = \frac{1}{L}$, $y_2 = -\frac{1}{L}$ und haben

$$L|y_1 - y_2| = 2.$$

Wegen $4 \not\leq 2$ findet sich damit kein L , für das die Lipschitz-Bedingung für $x = 1$ erfüllt ist. Entsprechend gilt die Lipschitz-Bedingung für kein $x \neq 0$.

- Die hier verwendete Lipschitz-Bedingung erstreckt sich auf ganz G und ist damit keine lokale Bedingung wie die Existenz und Beschränktheit der partiellen Ableitung in Satz 5.1. Allerdings kann sie abgeschwächt werden zu einer lokalen Lipschitz-Bedingung. Dabei muss es zu jedem Punkt $(x, y) \in D$ eine Umgebung geben, in der f einer Lipschitz-Bedingung genügt. Für jeden Punkt (x, y) darf dabei eine andere Konstante $L = L(x, y)$ verwendet werden.

5.1.4 Iterationsverfahren von Picard und Lindelöf

Es ist durchaus möglich, dass eine eindeutige Lösung $y(x)$ für ein Anfangswertproblem existiert, diese aber nicht in geschlossener Form dargestellt werden kann. In diesem Fall kann man die eindeutige Lösung nahe bei x_0 über das folgende Iterationsverfahren (Fixpunktverfahren) von Picard und Lindelöf (z. B. mit dem Computer) approximieren. Durch Integration erhält man aus dem Anfangswertproblem

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0,$$

die **Integralgleichung**

$$y(x) = y_0 + \int_{x_0}^x y'(t) dt = y_0 + \int_{x_0}^x f(t, y(t)) dt,$$

die als Iterationsvorschrift verwendet wird:

$$\begin{aligned} y_0(x) &:= y_0, \\ y_n(x) &:= y_0 + \int_{x_0}^x f(t, y_{n-1}(t)) dt, \quad n = 1, 2, 3, \dots \end{aligned}$$

Man kann nun zeigen (vgl. Satz 2.38 auf Seite 318), dass die so konstruierte Folge $y_n(x)$ für $n \rightarrow \infty$ konvergiert. Die Grenzfunktion ist ein Fixpunkt (den Begriff haben wir zuvor im Kontext von Zahlen statt von Funktionen in Definition 2.28 formuliert): Setzt man sie in die rechte Seite ein, erhält man sie als Ergebnis zurück. Sie löst damit die Integral- und somit auch die Differenzialgleichung.

Beispiel 5.12

Für das bereits bekannte Anfangswertproblem $y'(x) = y(x)$ mit $y(0) = 1$ erhält man mit der Startnäherung $y_0(x) = 1$ die sukzessiven Approximationen

$$\begin{aligned} y_1(x) &:= y_0 + \int_{x_0}^x f(t, y_0(t)) dt = 1 + \int_0^x 1 dt = 1 + x \\ y_2(x) &:= y_0 + \int_{x_0}^x f(t, y_1(t)) dt = 1 + \int_0^x (1 + t) dt = 1 + \left[t + \frac{t^2}{2} \right]_0^x = 1 + x + \frac{x^2}{2} \\ y_3(x) &:= y_0 + \int_{x_0}^x f(t, y_2(t)) dt = 1 + \int_0^x \left(1 + t + \frac{t^2}{2} \right) dt = 1 + x + \frac{x^2}{2} + \frac{x^3}{3 \cdot 2} \\ &\dots \\ y_n(x) &:= y_0 + \int_{x_0}^x f(t, y_{n-1}(t)) dt = \dots = 1 + x + \frac{x^2}{2} + \dots + \frac{x^n}{n!}, \end{aligned}$$

und für $n \rightarrow \infty$ folgt $y(x) = \lim_{n \rightarrow \infty} \left(1 + x + \frac{x^2}{2} + \dots + \frac{x^n}{n!} \right) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$. ■

5.1.5 Runge-Kutta-Verfahren

Wir haben die Existenz und Eindeutigkeit der Lösung einer gewöhnlichen Differenzialgleichung $y'(x) = f(x, y(x))$ mit Anfangsbedingung $y(x_0) = y_0$ mit Hilfe der Euler-Cauchy-Polygonzugmethode betrachtet. Die Idee dabei ist, mit der Anfangsbedingung zu starten und dann mit dem durch die Gleichung gegebenen Wissen über die Ableitung die Lösung $y(x)$ durch einen Polygonzug anzunähern:

$$y_{k+1} := y_k + (x_{k+1} - x_k) f(x_k, y_k) = y_k + h f(x_k, y_k),$$

wobei y_0 der Anfangswert und y_k der näherungsweise Funktionswert von y an einer Stelle x_k sei. Der Abstand der Stellen x_k sei außerdem konstant gleich $h > 0$. Wir können $h f(x_k, y_k) = (x_{k+1} - x_k) f(x_k, y_k)$ als Näherungswert des Integrals

$$\int_{x_k}^{x_{k+1}} f(x, y(x)) dx = \int_{x_k}^{x_{k+1}} y'(x) dx = y(x_{k+1}) - y(x_k)$$

auffassen. Wir erhielten statt der Näherungswerte y_k die exakten Funktionswerte $y(x_k)$, wenn wir dieses Integral nicht näherungsweise, sondern exakt berechnen würden:

$$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} f(x, y(x)) dx.$$

Eine analoge Darstellung $y_n(x) = y_0 + \int_{x_0}^x f(t, y_{n-1}(t)) dt$ wird beim Iterationsverfahren von Picard und Lindelöf eingesetzt, um ausgehend von einer gröberen Annäherung $y_{n-1}(x)$ an die Lösungsfunktion $y(x)$ durch Einsetzen in die rechte Seite und Integration eine bessere Näherung $y_n(x)$ zu erhalten. Diese wird dann wieder rechts eingesetzt usw. Durch Grenzwertbildung erhält man schließlich die exakte Lösung als Fixpunkt. Dabei wird in jedem Schritt exakt integriert, man erhält also jeweils eine komplette Näherungsfunktion und nicht nur einen einzelnen Funktionswert der Näherungslösung.

Bei der Euler-Cauchy-Polygonzugmethode wird dagegen das Integral auf sehr einfache Weise approximiert, und man erhält in jedem Schritt einen weiteren näherungsweisen Funktionswert y_{k+1} für $y(x_{k+1})$. Man diskutiert also keine Funktionenfolge von Näherungslösungen, sondern approximiert Funktionswerte der exakten Lösung. In etwa gilt: Halbiert man die Schrittweiten $h = x_{k+1} - x_k$, dann halbiert sich auch der Fehler $|y(x_k) - y_k|$.

Bei der Euler-Cauchy-Polygonzugmethode löst man also ein Anfangswertproblem mittels numerischer Integration. Wenn man sich hier mehr Mühe gibt, dann erhält man numerische Verfahren mit einem signifikant kleineren Fehler. Die Schwierigkeit dabei ist, dass wir $y(x)$ nur näherungsweise als y_k an der Stelle x_k kennen. Die anderen Funktionswerte auf dem Intervall $[x_k, x_{k+1}]$ sind nicht einmal näherungsweise bekannt.

Beim **Verfahren von Heun** wird das Integral dennoch näherungsweise mit der Trapezregel (siehe Kapitel 2.6.5) berechnet, es wird also auf dem Intervall $[x_k, x_{k+1}]$ die Sekante integriert:

$$\int_{x_k}^{x_{k+1}} f(x, y(x)) dx \approx \frac{h}{2} [f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1}))],$$

so dass wir Näherungswerte berechnen können über

$$y_{k+1} := y_k + \frac{h}{2} [f(x_k, y_k) + f(x_{k+1}, y_{k+1})].$$

Jetzt steht aber leider y_{k+1} sowohl auf der linken als auch auf der rechten Seite. Daher berechnet man zunächst in einer ersten Stufe eine vorläufige Näherungslösung (Prädiktor) $y_{k+1}^{(P)}$ über einen Schritt der Euler-Cauchy-Polygonzugmethode:

$$y_{k+1}^{(P)} := y_k + hf(x_k, y_k).$$

Der Prädiktor wird dann in einer zweiten Stufe zu y_{k+1} „korrigiert“:

$$y_{k+1} := y_k + \frac{h}{2} [f(x_k, y_k) + f(x_{k+1}, y_{k+1}^{(P)})].$$

Wegen der beiden Stufen des Verfahrens von Heun spricht man von einem Prädiktor-Korrektor-Verfahren. Bei Halbierung der Schrittweiten h reduziert sich der Fehler $|y(x_k) - y_k|$ ungefähr um den Faktor $\frac{1}{4}$.

Der Fehler lässt sich sogar um den Faktor $\frac{1}{16}$ reduzieren, indem man die Trapezregel durch die Simpson-Regel (siehe ebenfalls Abschnitt 2.6.5) ersetzt. Das entsprechende Verfahren ist das **Runge-Kutta-Verfahren**. Zunächst erhalten wir mit der Simpson-Regel für die Schrittweite $\frac{h}{2}$ den Ansatz

$$y_{k+1} := y_k + \frac{h}{6} \left[f(x_k, y_k) + 4f\left(x_k + \frac{h}{2}, y_{k+1/2}\right) + f(x_{k+1}, y_{k+1}) \right].$$

Jetzt ist aber auf der rechten Seite nicht nur y_{k+1} , sondern auch $y_{k+1/2}$ nicht bekannt. Daher werden auch hier Näherungen berechnet:

$$\begin{aligned} F_1 &:= f(x_k, y_k), \\ F_2 &:= f\left(x_k + \frac{h}{2}, y_k + \frac{h}{2}F_1\right), \\ F_3 &:= f\left(x_k + \frac{h}{2}, y_k + \frac{h}{2}F_2\right), \\ F_4 &:= f(x_{k+1}, y_k + hF_3). \end{aligned}$$

Hier sind F_2 und F_3 Näherungswerte für $f(x_k + \frac{h}{2}, y_{k+1/2})$, F_4 ist ein Näherungswert für $f(x_{k+1}, y_{k+1})$. Damit lautet die Rechenvorschrift des Runge-Kutta-Verfahrens:

$$y_{k+1} := y_k + \frac{h}{6} [F_1 + 2F_2 + 2F_3 + F_4].$$

In Kapitel 5.5.3 sehen wir uns weitere numerische Verfahren an.

5.2 Lösungsmethoden für Differenzialgleichungen erster Ordnung

Nachdem wir uns im letzten Abschnitt Verfahren zur Berechnung von Näherungslösungen angesehen haben, wollen wir jetzt wieder exakte Lösungen herleiten. Das gelingt z. B. mit Potenzreihen, da wir sie gliedweise ableiten dürfen (siehe Satz 2.58 auf Seite 409). Dies ist allerdings sehr aufwändig:

Beispiel 5.13

Wir betrachten erneut die Gleichung $y' = ay$ (vgl. (5.2)) mit einer Konstante $a \in \mathbb{R}$ und mit Anfangsbedingung $y(0) = 1$. Wir kennen bereits die Lösung $y(x) = e^{ax}$. Die Lösung konnten wir zu Beginn des Kapitels raten, da wir die Eigenschaften der Exponentialfunktion in der Differenzialgleichung erkannt haben. Im letzten Abschnitt haben wir mit dem Iterationsverfahren von Picard und Lindelöf im Fall $a = 1$ die Lösung berechnet.

Nun leiten wir die Lösung mit einem Potenzreihenansatz her. Dieser funktioniert wie das Iterationsverfahren auch für Gleichungen, bei denen man die Lösung nicht sofort sieht. Falls eine Lösung existiert, die als Potenzreihe um 0 (dort ist der Anfangswert gegeben) entwickelbar ist, gilt mit gliedweiser Ableitung:

$$y(x) = \sum_{k=0}^{\infty} a_k x^k, \quad y'(x) = \sum_{k=1}^{\infty} k a_k x^{k-1} = \sum_{k=0}^{\infty} (k+1) a_{k+1} x^k.$$

Wegen $y(0) = 1$ ist $a_0 = 1$. Einsetzen der Reihen in die Differenzialgleichung, d. h.

$$\sum_{k=0}^{\infty} (k+1) a_{k+1} x^k = a \sum_{k=0}^{\infty} a_k x^k,$$

und Koeffizientenvergleich liefern für $k \geq 0$ die Iterationsvorschrift $(k+1)a_{k+1} = aa_k$, also

$$a_{k+1} = \frac{a}{k+1} a_k = \frac{a}{k+1} \frac{a}{k} a_{k-1} = \cdots = \frac{a^{k+1}}{(k+1)!} a_0 = \frac{a^{k+1}}{(k+1)!}.$$

Damit ist $y(x) = \sum_{k=0}^{\infty} \frac{(ax)^k}{k!} = e^{ax}$. ■

In der Regel erweist sich dieser Potenzreihenansatz, aber auch das Iterationsverfahren von Picard und Lindelöf als sehr schwierig zum Berechnen exakter Lösungen. Für viele Varianten der Differenzialgleichung

$$y'(x) = f(x, y(x))$$

gibt es wesentlich einfachere Lösungsverfahren. Einige sehen wir uns nun an.

5.2.1 Lineare Differenzialgleichungen erster Ordnung

Hier betrachten wir Differenzialgleichungen der Form

$$y'(x) = g(x)y(x) + h(x). \quad (5.6)$$

Wir identifizieren die einzelnen Bestandteile:

$$\underbrace{y'(x)}_{\text{Ableitung der gesuchten Funktion}} = \underbrace{g(x)y(x)}_{\substack{g(x) \text{ mal} \\ \text{gesuchte Funktion}}} + \underbrace{h(x)}_{\substack{\text{Inhomogenität, hier steht} \\ \text{keine gesuchte Funktion } y}}$$

Ist $h(x) = 0$ für alle x , so spricht man von einer **homogenen linearen Differenzialgleichung erster Ordnung**, ansonsten von einer **inhomogenen**. Bereits in der Einleitung dieses Kapitels (Seite 707) haben wir die homogenen linearen Differenzialgleichungen (5.2) und (5.3) gelöst, bei denen $g(x)$ eine Konstante war.

Mit $f(x, y) := g(x)y + h(x)$ lautet die Gleichung $y'(x) = f(x, y(x))$. Sind $g(x)$ und $h(x)$ stetig auf einem Intervall $]a, b[$, so ist f auf $G :=]a, b[\times \mathbb{R}$ stetig. Ist zusätzlich g beschränkt mit $|g(x)| \leq L$ für alle $x \in]a, b[$, so erfüllt f eine Lipschitz-Bedingung:

$$|f(x, y_1) - f(x, y_2)| = |g(x)||y_1 - y_2| \leq L|y_1 - y_2|$$

für alle $(x, y_1), (x, y_2) \in G$. Damit hat jedes Anfangswertproblem zu dieser Differenzialgleichung nach dem Existenz- und Eindeutigkeitssatz (Satz 5.2) eine eindeutige Lösung. Um diese zu berechnen, betrachtet man zunächst das zu (5.6) gehörende homogene Problem, indem man $h(x) = 0$ setzt. Man erhält **homogene Lösungen**, mit denen man anschließend eine Lösung des inhomogenen Problems berechnen kann.

5.2.1.1 Homogene lineare Differenzialgleichungen erster Ordnung

In diesem Abschnitt untersuchen wir die homogene Differenzialgleichung

$$y'(x) = g(x)y(x).$$

Um zu einer allgemeinen Lösung zu gelangen, rechnen wir formal mit einer Lösung y_h , deren Eigenschaften wir aber noch gar nicht kennen. So erhalten wir aber einen Kandidaten, den wir anschließend noch prüfen müssen. Sofern die folgenden Rechenschritte erlaubt sind (insbesondere $y_h(x) > 0$ für alle betrachteten x oder $y_h(x) < 0$ für alle diese x), ergibt sich mit der Substitution $u = y_h(x)$, $du = y'_h(x) dx$:

$$\begin{aligned} y'_h(x) = g(x)y_h(x) &\implies \frac{y'_h(x)}{y_h(x)} = g(x) \implies \int \frac{y'_h(x)}{y_h(x)} dx = \int g(x) dx + c \\ &\implies \int \frac{1}{u} du = \int g(x) dx + c \implies \ln |u| = \int g(x) dx + c \\ &\implies \ln |y_h(x)| = \int g(x) dx + c \implies y_h(x) = \pm e^c \exp \left(\int g(x) dx \right). \end{aligned}$$

Damit haben wir eine mögliche Lösung gefunden. Der hier verwendete Ansatz heißt **Trennung der Variablen** und wird in Kapitel 5.2.2.1 ausführlicher beschrieben.

Satz 5.3 (Formel für homogene Lösungen)

Sei $g :]a, b[\rightarrow \mathbb{R}$ stetig. Dann ist

$$y_h(x) = k \exp \left(\int g(x) dx \right) \quad (5.7)$$

eine Lösung der Differenzialgleichung $y'(x) - g(x)y(x) = 0$ auf $]a, b[$, wobei $k \in \mathbb{R}$ beliebig gewählt werden kann.

Achtung: Es werden hier häufig Vorzeichenfehler gemacht. Die Formel gilt, wenn auf der rechten Seite der Differenzialgleichung $g(x)y(x)$ oder auf der linken $-g(x)y(x)$ steht. Lautet dagegen die Differenzialgleichung $y'(x) + g(x)y(x) = 0$, so ist die homogene Lösung $y_h(x) = k \exp\left(-\int g(x) dx\right)$.

Beweis: Nach Kettenregel (siehe Seite 297) und Hauptsatz (Seite 340) ist

$$y'_h(x) = kg(x) \exp\left(\int g(x) dx\right) = g(x)y_h(x).$$

Damit erfüllt (5.7) die homogene Gleichung. □

Die Stammfunktion $\int g(x) dx$ ist nur bis auf eine additive Konstante c eindeutig bestimmt. Da diese aber in (5.7) im Exponenten der Exponentialfunktion steht, führt sie zu einem Faktor e^c , der mit der frei wählbaren Konstante k verrechnet werden kann. Bei der Berechnung einer homogenen Lösung kann man also irgendeine Stammfunktion verwenden und muss nicht zusätzlich eine allgemeine Konstante addieren.

Satz 5.4 (Vektorraum der Lösungen)

Die homogenen Lösungen der Gleichung (5.6) bilden einen eindimensionalen Vektorraum von Funktionen.

Beweis: Wir zeigen zunächst, dass die Lösungen einen Vektorraum bilden. Sie liegen in einem Vektorraum von Funktionen (siehe Seite 469), und wir müssen nur zeigen, dass sie einen Unterraum formen. Dazu muss die Menge abgeschlossen gegenüber Addition und Multiplikation mit einem Skalar sein: Sind $y_{h,1}$ und $y_{h,2}$ homogene Lösungen auf $]a, b[$ und sind $c_1, c_2 \in \mathbb{R}$, dann ist auch

$$c_1 y_{h,1} + c_2 y_{h,2}$$

eine homogene Lösung. Dies sieht man sofort durch Einsetzen in die Differenzialgleichung. Damit liegt ein Vektorraum vor. Dieser ist aber nur eindimensional:

- Jede Lösung hat die Gestalt (5.7): Sei $x_0 \in]a, b[$, und sei $z_h(x)$ eine Lösung. Wir wissen, dass nach (5.7) auch $y_h(x) = z_h(x_0) \exp\left(\int_{x_0}^x g(t) dt\right)$ eine Lösung ist, die die gleiche Anfangsbedingung in x_0 erfüllt. Da das Anfangswertproblem bei einer stetigen Koeffizientenfunktion g eine eindeutige Lösung hat, müssen beide Lösungen gleich sein: $y_h = z_h$.
- Zwei Lösungen unterscheiden sich nur durch einen Faktor. Denn beide Lösungen haben die Gestalt (5.7).
- Damit ist $\{\exp\left(\int g(x) dx\right)\}$ eine einelementige Basis des Lösungsraums. □

Beispiel 5.14

Bei der einfachsten homogenen linearen Differenzialgleichung $y'(x) = ay(x)$ ist $g(x) = a$ konstant. Bereits zuvor haben wir die Lösung über die Eigenschaften der Exponentialfunktion und über einen Potenzreihenansatz gefunden: $y_h = c \exp(ax)$. Diese allgemeine Lösung erhält man auch über die Formel (5.7):

$$y_h(x) = k_0 \exp \left(\int a \, dx \right) = k_0 \exp(ax + c) = k_1 \exp(ax). \quad \blacksquare$$

Beispiel 5.15

Das Gesetz $y'(t) - k \cdot y(t) = 0$ für den radioaktiven Zerfall (siehe (5.3)) ist eine homogene lineare Differenzialgleichung mit Lösung $y(t) = c \exp(kt)$. Liegt zum Zeitpunkt $t = 0$ die Anfangsmenge $y(0) = y_0$ vor, so lautet die Lösung des Anfangswertproblems

$$y(t) = y_0 e^{kt}.$$

Bei der Altersbestimmung nach der **Radiocarbon-** oder ^{14}C -**Methode** nutzt man aus, dass in der Natur das radioaktive Isotop ^{14}C und das nicht zerfallende ^{12}C in etwa dem Verhältnis $\frac{^{14}\text{C}}{^{12}\text{C}} \approx 10^{-12}$ vorkommen. Das gilt auch für lebende Organismen. Stirbt ein Organismus, so wird ab diesem Zeitpunkt kein ^{14}C mehr aufgenommen, da dies z. B. durch Atmung oder Nahrungsaufnahme geschieht. Das radioaktive ^{14}C zerfällt, während die Menge des nicht-radioaktiven ^{12}C unverändert bleibt. Aus dem Verhältnis der Mengen der beiden Isotope ^{14}C und ^{12}C wird auf das Alter eines Fossils geschlossen.

Wir bestimmen das Alter eines fossilen Knochens, der noch 20 % der normalen Menge des Isotops ^{14}C enthält. Die Halbwertszeit von ^{14}C beträgt 5 730 Jahre. Daraus berechnen wir die Zerfallskonstante k : Sei $y_0 > 0$ die am Anfang vorhandene Menge, so ist

$$\frac{y_0}{2} = y(5\,730) \iff \frac{y_0}{2} = y_0 \cdot e^{5\,730 \cdot k} \iff \ln \frac{1}{2} = 5\,730 \cdot k \iff k = \frac{-\ln 2}{5\,730} \approx -0,000121.$$

Das Zerfallsgesetz lautet

$$y(t) = y_0 e^{-0,000121 \cdot t}.$$

In unserem Beispiel ist die noch vorhandene Menge ^{14}C gleich einem Fünftel der Anfangsmenge:

$$\frac{1}{5} y_0 = y(T) = y_0 \cdot e^{-0,000121 \cdot T} \iff -\ln 5 = -0,000121 \cdot T \iff T = \frac{\ln 5}{0,000121} \approx 13\,301.$$

Das Alter des Knochens wird zwischen 13 000 und 14 000 Jahre geschätzt. ■

5.2.1.2 Inhomogene lineare Differenzialgleichungen erster Ordnung

Mit den homogenen Lösungen lassen sich **partikuläre Lösungen** (spezielle Lösungen) der inhomogenen Ausgangsgleichung (5.6)

$$y'(x) = g(x)y(x) + h(x)$$

ermitteln.

Als Ansatz verändern wir dazu eine nicht-triviale homogene Lösung y_h so, dass auf der rechten Seite nicht 0, sondern die gewünschte Inhomogenität $h(x)$ herauskommt. Jede homogene Lösung können wir schreiben als $ky_h(x)$ mit einer Konstante $k \in \mathbb{R}$. Nun probieren wir aus, ob eine partikuläre Lösung der Differenzialgleichung $y'(x) = g(x)y(x) + h(x)$ die Gestalt

$$y_p(x) = K(x) \cdot y_h(x)$$

mit $K(x)$ als differenzierbare Funktion von x hat. Wir variieren also die Konstante k der homogenen Lösung in Abhängigkeit von x . Man nennt diesen Ansatz daher **Variation der Konstanten**. Zu diesem Zeitpunkt kann man nicht wissen, ob eine partikuläre Lösung tatsächlich diesen Aufbau hat. Das ist Ausprobieren. Durch Einsetzen in die Differenzialgleichung lässt sich $K(x)$ berechnen: Es gilt mit der Produktregel der Differenziation

$$y_p'(x) = K'(x)y_h(x) + K(x)y_h'(x) = K'(x)y_h(x) + K(x)g(x)y_h(x),$$

so dass wir beim Einsetzen in die Differenzialgleichung erhalten:

$$K'(x)y_h(x) + \underbrace{g(x)K(x)y_h(x) - g(x)K(x)y_h(x)}_{=0} = h(x) \iff K'(x) = \frac{h(x)}{y_h(x)}.$$

Mit $K(x) = \int \frac{h(x)}{y_h(x)} dx$ erhalten wir also eine partikuläre Lösung der inhomogenen Gleichung zu

$$y_p(x) = y_h(x) \int \frac{h(x)}{y_h(x)} dx.$$

Wir fassen das in einem Satz zusammen:

Satz 5.5 (Formel für eine partikuläre Lösung)

Seien $g, h :]a, b[\rightarrow \mathbb{R}$ stetig und $y_h \neq 0$ eine Lösung der homogenen Differenzialgleichung $y'(x) - g(x)y(x) = 0$ auf $]a, b[$. Dann ist

$$y_p(x) = y_h(x) \int \frac{h(x)}{y_h(x)} dx \quad (5.8)$$

eine (partikuläre) Lösung der inhomogenen Differenzialgleichung $y'(x) - g(x)y(x) = h(x)$ auf $]a, b[$.

Die Stammfunktion $\int \frac{h(x)}{y_h(x)} dx$ ist nur bis auf eine Konstante $c \in \mathbb{R}$ eindeutig. Die Konstante liefert einen Summanden $y_h(x) \cdot c$, der Lösung der homogenen Gleichung ist. Eine partikuläre Lösung erhält man insbesondere auch für $c = 0$.

Beweis: Zunächst beachte man, dass die homogenen Lösungen einen eindimensionalen Vektorraum bilden, dessen Elemente nach (5.7) die Gestalt $y_h(t) = k \exp(\int g(x) dx) \neq 0$ für $k \neq 0$ haben. Damit ist entweder y_h die Nullfunktion, oder y_h ist an keiner Stelle gleich null, so dass wir durch y_h dividieren können.

Um zu sehen, dass y_p tatsächlich eine Lösung ist, setzen wir y_p in die Differenzialgleichung (5.6) ein:

$$\begin{aligned} y_p'(x) - g(x)y_p(x) &= y_h'(x) \int \frac{h(x)}{y_h(x)} dx + y_h(x) \frac{h(x)}{y_h(x)} - g(x)y_h(x) \int \frac{h(x)}{y_h(x)} dx \\ &= (y_h'(x) - g(x)y_h(x)) \int \frac{h(x)}{y_h(x)} dx + h(x) = h(x), \end{aligned}$$

da y_h als homogene Lösung für $y_h'(x) - g(x)y_h(x) = 0$ sorgt. $\square$

Hat man eine partikuläre Lösung y_p der Gleichung (5.6) gefunden, erhält man weitere Lösungen über den Ansatz $y_p + y_h$, wobei y_h eine beliebige homogene Lösung ist. Das sieht man direkt durch Einsetzen in die Differenzialgleichung.

Lemma 5.1 (Allgemeine Lösung)

Hat man eine partikuläre Lösung y_p der Gleichung (5.6) gefunden, erhält man alle Lösungen über den Ansatz $y_p + y_h$, wobei y_h eine beliebige homogene Lösung ist.

Man benötigt also nur eine einzige partikuläre Lösung, um auf diese Art alle Lösungen anzugeben. Verwendet man bei der Berechnung einer Stammfunktion in der Formel (5.8) zur partikulären Lösung eine Konstante $c \neq 0$, so erhält man damit als Summanden eine homogene Lösung, die mit der beliebigen homogenen Lösung aus dem Lemma zusammengefasst werden kann.

Beweis: Wir müssen nur noch zeigen, dass jede Lösung von (5.6) die Gestalt $y_p + y_h$ für eine homogene Lösung y_h hat. Dazu sei y eine beliebige Lösung. Die Differenz $y - y_p$ ist dann aber eine homogene Lösung $y_h := y - y_p$, und $y = y_p + y_h$. $\square$

Da der Lösungsraum der homogenen Differenzialgleichung eindimensional ist, kann man mit einer homogenen Lösung $y_h \neq 0$ und einer partikulären Lösung y_p die Lösungsmenge der inhomogenen Differenzialgleichung vollständig beschreiben:

$$\{y : y(x) = y_p(x) + cy_h(x) \text{ für ein } c \in \mathbb{R}\}.$$

Die Konstante c entspricht einem Freiheitsgrad. Gibt man über eine Anfangsbedingung einen Funktionswert der Lösung vor, wird c dadurch eindeutig bestimmt. Insbesondere ist der Freiheitsgrad nötig, damit man zu jeder Anfangsbedingung eine Lösung findet.

Haben wir ein Anfangswertproblem mit der Anfangsbedingung $y(x_0) = y_0$, so erhalten wir mit der partikulären Lösung $y_p(x) = y_h(x) \int_{x_0}^x \frac{h(t)}{y_h(t)} dt$ die nach dem Existenz- und Eindeutigkeitsatz eindeutige Lösung

$$y(x) = y_p(x) + \frac{y_0}{y_h(x_0)} y_h(x) = y_h(x) \left[\int_{x_0}^x \frac{h(t)}{y_h(t)} dt + \frac{1}{y_h(x_0)} y_0 \right]. \quad (5.9)$$

dieses Anfangswertproblems. Allerdings verwendet man im Allgemeinen nicht diese Formel, sondern bestimmt die Konstante der allgemeinen Lösung durch Einsetzen von x_0 so, dass der Anfangswert angenommen wird. So gehen wir auch in einigen der folgenden Beispielen vor.

Beispiel 5.16

Wir betrachten $y'(x) + ay(x) = b$, $a, b \in \mathbb{R}$, und erhalten

$$\begin{aligned} y_h(x) &= k \exp(-ax), \\ y_p(x) &= k \exp(-ax) \int \frac{b}{k \exp(-ax)} dx = b \exp(-ax) \int \exp(ax) dx \\ &= \frac{b}{a} \exp(-ax) [\exp(ax) + c_0] = \frac{b}{a} + c_1 \exp(-ax). \end{aligned} \quad (5.10)$$

Da wir nicht eine spezielle Stammfunktion verwendet, sondern die allgemeine Konstante c_0 angegeben haben, ist die partikuläre Lösung direkt auch die allgemeine Lösung, bei der beliebige homogene Lösungen $c_1 \exp(-ax)$, $c_1 \in \mathbb{R}$, addiert werden. ■

Beispiel 5.17

Jetzt können wir insbesondere auch das Eingangsbeispiel des R/L-Kreises (5.1)

$$L \frac{d}{dt} i(t) + Ri(t) = u(t).$$

lösen. Es ist $g(t) = -\frac{R}{L}$ und $h(t) = \frac{u(t)}{L}$ und $i_h(t) = k \exp\left(-\frac{R}{L}t\right)$,

$$i(t) = i_p(t) + c_1 i_h(t) = \exp\left(-\frac{R}{L}t\right) \int \frac{\frac{u(t)}{L}}{\exp\left(-\frac{R}{L}t\right)} dt + c_1 \exp\left(-\frac{R}{L}t\right).$$

Legen wir eine konstante Spannung $u(t) = u_0$ an, so bekommen wir die Lösung

$$i(t) = \exp\left(-\frac{R}{L}t\right) \frac{u_0}{L} \int \exp\left(\frac{R}{L}t\right) dt + c_1 \exp\left(-\frac{R}{L}t\right) = \frac{u_0}{R} + c_2 \exp\left(-\frac{R}{L}t\right).$$

Bei einer Anfangsbedingung $i(0) = 0$ ist $i(t) = \frac{u_0}{R} [1 - \exp\left(-\frac{R}{L}t\right)]$. ■

Beispiel 5.18

Das Anfangswertproblem $y'(x) = x^2 y(x) + x^2$, $y(0) = 2$, besitzt die homogene Lösung:

$$y_h(x) = k_0 \exp\left(\int x^2 dx\right) = k_1 \exp\left(\frac{x^3}{3}\right).$$

Damit gewinnen wir die partikuläre Lösung

$$y_p(x) = y_h(x) \int \frac{x^2}{y_h(x)} dx = k_1 \exp\left(\frac{x^3}{3}\right) \int^x \frac{t^2}{k_1 \exp\left(\frac{t^3}{3}\right)} dt.$$

Mit der Substitution $u = t^3/3$, $du = t^2 dt$, erhalten wir daraus

$$\begin{aligned} y_p(x) &= \exp\left(\frac{x^3}{3}\right) \int^{x^3/3} \frac{1}{\exp(u)} du = -\exp\left(\frac{x^3}{3}\right) \exp(-u) \Big|^{x^3/3} \\ &= -\exp\left(\frac{x^3}{3}\right) \exp\left(-\frac{x^3}{3}\right) + c \exp\left(\frac{x^3}{3}\right) = -1 + c \exp\left(\frac{x^3}{3}\right). \end{aligned}$$

Da wir die Konstante c bei der Stammfunktion angegeben haben, ist dies zugleich die allgemeine Lösung für $c \in \mathbb{R}$. Wegen der Anfangsbedingung ist die gesuchte Lösung $y(x) = -1 + 3 \exp(x^3/3)$ und erfüllt die Differenzialgleichung auf $\mathbb{R}$. Eine Probe bestätigt das: $y'(x) = 3x^2 \exp(x^3/3) = x^2 y(x) + x^2$. ■

Beispiel 5.19 (Newton'sches Abkühlungsgesetz)

In diesem Beispiel geht es um praktische Lebenshilfe mittels linearer Differenzialgleichungen. Um einen Kaffee abzukühlen, geben wir ein Teil Milch auf zwei Teile Kaffee hinzu und warten eine kurze Zeit, z. B. zwei Minuten. Wird der Kaffee kälter, wenn man erst die Milch hinzufügt und dann wartet, oder wird er kälter, wenn man erst wartet und dann die Milch hinzugibt?

Wir berechnen den Temperaturunterschied der beiden Ansätze. Dazu benötigen wir das Newton'sche Abkühlungsgesetz:

$$\frac{d}{dt}T(t) = k \cdot (T(t) - T_u), \quad T(0) = T_0.$$

Hier ist $T(t)$ die Temperatur des Kaffees in Abhängigkeit von der Zeit $t > 0$. T_0 ist die Ausgangstemperatur des Kaffees, T_u ist die Umgebungstemperatur. k ist eine negative Proportionalitätskonstante. Je größer die Differenz zur Umgebungstemperatur ist, desto schneller kühlt der Kaffee ab. Die allgemeine Lösung dieser inhomogenen linearen Differenzialgleichung $\frac{d}{dt}T(t) - kT(t) = -kT_u$ mit konstantem Koeffizienten ist (siehe (5.10))

$$T(t) = \frac{-kT_u}{-k} + c \exp(kt) = T_u + c \exp(kt).$$

Aus der Anfangsbedingung $T_0 = T(0) = T_u + c$ erhalten wir die Konstante $c = T_0 - T_u$ und damit

$$T(t) = T_u + (T_0 - T_u) \exp(kt).$$

Entspricht also insbesondere die Ausgangstemperatur T_0 des Kaffees genau der Umgebungstemperatur T_u , so bleibt die Temperatur des Kaffees konstant die Umgebungstemperatur. Sonst drückt die Exponentialfunktion die Temperaturänderung mit der Zeit aus.

Geben wir vor der Wartezeit t Milch mit der Temperatur T_M hinzu, so erhalten wir als neue Ausgangstemperatur des Kaffees $\frac{2T_0+T_M}{3}$ und damit (unter der Annahme, dass sich die Proportionalitätskonstante k nicht ändert) insgesamt die Temperatur

$$T_V(t) = T_u + \left(\frac{2T_0 + T_M}{3} - T_u \right) \exp(kt).$$

Geben wir erst nach der Wartezeit t die Milch hinzu, so benutzen wir das Abkühlungsgesetz für die gegebene Ausgangstemperatur T_0 und mischen dann das Ergebnis mit der Temperatur der Milch:

$$T_N(t) = \frac{1}{3} [T_M + 2(T_u + (T_0 - T_u) \exp(kt))].$$

Nun ist

$$\begin{aligned} T_N(t) &< T_V(t) \\ \iff \frac{1 - \exp(kt)}{3} T_M &< \frac{1}{3} T_u - \frac{2}{3} (T_0 - T_u) \exp(kt) + \left(\frac{2}{3} T_0 - T_u \right) \exp(kt) \\ \iff T_M &< \frac{3}{1 - \exp(kt)} \left[\frac{1}{3} T_u - \frac{1}{3} T_u \exp(kt) \right] \iff T_M < T_u. \end{aligned}$$

Man beachte, dass wegen $k < 0$ der Faktor $\frac{1 - \exp(kt)}{3}$ positiv ist, so dass bei Multiplikation mit dem Kehrwert das „<“-Zeichen erhalten bleibt.

Ist die Milch kälter als die Umgebungstemperatur, so kühlt der Kaffee schneller ab, wenn man erst wartet. Ist die Milch wärmer, so kühlt das Getränk schneller, wenn erst die Milch hinzugegeben wird. ■

5.2.1.3 Inhomogene lineare Differenzialgleichung mit konstantem Koeffizienten

Hier sehen wir uns Gleichungen vom Typ

$$y'(x) - ay(x) = h(x)$$

an, die wir bereits allgemein lösen können. Allerdings müssen wir dazu das Integral in (5.8) ausrechnen. Für gängige Inhomogenitäten (rechte Seiten) $h(x)$ kann man stattdessen Kandidaten $y_p(x)$ für partikuläre Lösungen einer Tabelle (siehe Tabelle 5.1) entnehmen, die eine ähnliche Struktur wie die Inhomogenität haben. Man nennt dieses Vorgehen **Ansatz vom Typ der rechten Seite**, da man je nach Inhomogenität $h(x)$ der Tabelle einen Kandidaten y_p für die Lösung entnimmt, bei dem noch Konstanten

zu bestimmen sind. Diese erhält man durch Einsetzen von y_p in die Differenzialgleichung. Der Ansatz funktioniert, da die rechten Seiten und mit ihnen die Kandidaten y_p eine Gestalt haben, die durch Einsetzen in die linke Seite der Differenzialgleichung erhalten bleibt. Damit stehen auf der linken und auf der rechten Seite Terme der gleichen Gestalt. Die auf der linken Seite vorkommenden Konstanten lassen sich dadurch mittels Koeffizientenvergleich über ein eindeutig lösbares lineares Gleichungssystem bestimmen. Der Fall $h(x) = A e^{\mu x}$ in Tabelle 5.1 fällt etwas aus der Rolle, da man

Tab. 5.1: Ansatzfunktionen für die partikuläre Lösung von $y'(x) = ay(x) + h(x)$

Inhomogenität $h(x)$	Ansatz $y_p(x)$
Polynom vom Grad n	$y_p(x) = C_0 + C_1 x + \dots + C_n x^n$, unabhängig von der Inhomogenität verwende man alle Konstanten
$h(x) = A \sin(\omega x) + B \cos(\omega x)$	$y_p(x) = C_1 \sin(\omega x) + C_2 \cos(\omega x)$, auch bei $A = 0$ oder $B = 0$ sind beide Konstanten nötig
$h(x) = A e^{\mu x}$	$y_p(x) = \begin{cases} C \cdot e^{\mu x}, & \mu \neq a \\ C \cdot x \cdot e^{\mu x}, & \mu = a \end{cases}$ (Resonanzfall)

im Fall $\mu = a$ einen Kandidaten $C \cdot x \cdot e^{\mu x}$ verwendet, der nicht direkt der Inhomogenität h entspricht. Das liegt daran, dass für $\mu = a$ die Funktion $A e^{\mu x}$ eine homogene Lösung der Differenzialgleichung ist. Daher muss man die Funktion etwas modifizieren, damit die rechte Seite und nicht null beim Einsetzen in die Differenzialgleichung herauskommt.

Ist die Inhomogenität $h(x)$ eine Summe mehrerer Funktionen, so wählt man den Lösungsansatz für y_p als Summe der Lösungsansätze für die einzelnen Funktionen. Das klappt, da das Ableiten linear ist. Entsprechend geht man bei Produkten $h(x) = h_1(x) \cdot h_2(x)$ vor. Kann man über die Tabelle 5.1 zu h_1 und h_2 Lösungsansätze y_{p_1} und y_{p_2} finden, so rechnet man mit dem Kandidaten $y_p(x) = y_{p_1}(x) \cdot y_{p_2}(x)$. Dass das funktioniert, ist nicht ganz so offensichtlich. Nach der Produktregel ist $y'_p(x) = y'_{p_1}(x)y_{p_2}(x) + y_{p_1}(x)y'_{p_2}(x)$, aber auch $y'_{p_1}(x)$ ist vom gleichen Typ wie y_{p_1} und $y'_{p_2}(x)$ ist vom gleichen Typ wie $y_{p_2}(x)$, so dass y'_p ebenfalls vom Typ $y_{p_1} \cdot y_{p_2}$ bzw. $h_1 \cdot h_2$ ist.

Beispiel 5.20

Wir bestimmen die allgemeine Lösung der Differenzialgleichung

$$y' + 2y = x^2 + 2.$$

Die homogene Gleichung besitzt die allgemeine Lösung

$$y_h(x) = C \cdot e^{-2x}, \quad C \in \mathbb{R}.$$

Der Ansatz für eine partikuläre Lösung lautet gemäß Tabelle 5.1

$$y_p(x) = C_2 x^2 + C_1 x + C_0 \quad \text{mit} \quad C_2, C_1, C_0 \in \mathbb{R}.$$

Setzen wir y_p mit der Ableitung $y_p'(x) = 2C_2 x + C_1$ in die Differenzialgleichung ein, erhalten wir:

$$\begin{aligned} 2C_2 x + C_1 + 2C_2 x^2 + 2C_1 x + 2C_0 &= x^2 + 2 \\ \iff x^2 (2C_2) + x (2C_1 + 2C_2) + C_1 + 2C_0 &= x^2 + 2, \end{aligned}$$

und der Koeffizientenvergleich ergibt das lineare Gleichungssystem für C_0, C_1, C_2

$$\begin{aligned} 2C_2 &= 1 \implies C_2 = \frac{1}{2} \\ 2C_1 + 2C_2 &= 0 \implies C_1 = -\frac{1}{2} \\ C_1 + 2C_0 &= 2 \implies C_0 = \frac{5}{4} \end{aligned}$$

und somit $y_p(x) = \frac{1}{2}x^2 - \frac{1}{2}x + \frac{5}{4}$. Die allgemeine Lösung lautet dann

$$y(x) = y_h(x) + y_p(x) = C \cdot e^{-2x} + \frac{1}{2}x^2 - \frac{1}{2}x + \frac{5}{4}, \quad C \in \mathbb{R}. \quad \blacksquare$$

Beispiel 5.21

Für die Lösung der Differenzialgleichung

$$y' + 9y = -2 \sin x$$

berechnen wir zunächst die Lösung der homogenen Gleichung $y' + 9y = 0$ zu

$$y_h(x) = C \cdot e^{-9x}, \quad C \in \mathbb{R}.$$

Die Lösung der inhomogenen Gleichung bestimmen wir mit dem Ansatz vom Typ der rechten Seite

$$y_p(x) = A \cdot \sin(x) + B \cdot \cos(x).$$

Einsetzen in die Differenzialgleichung liefert

$$\begin{aligned} A \cos(x) - B \sin(x) + 9A \sin(x) + 9B \cos(x) &= -2 \sin(x) \\ \iff (9A - B) \sin(x) + (A + 9B) \cos(x) &= -2 \sin(x). \end{aligned}$$

Auch hier ist ein Koeffizientenvergleich möglich. Man kann nämlich zeigen, dass die Funktionen 1 , $\sin x$ und $\cos x$ linear unabhängig sind und damit die Nullfunktion nur auf triviale Weise linear kombiniert werden kann. Der Koeffizientenvergleich ergibt $9A - B = -2$ und $A + 9B = 0$, also $A = -9B$, $B = \frac{1}{41}$ und $A = -\frac{9}{41}$, so dass $y_p(x) = -\frac{9}{41} \sin(x) + \frac{1}{41} \cos(x)$. Die allgemeine Lösung lautet dann

$$y(x) = y_h(x) + y_p(x) = C \cdot e^{-9x} - \frac{9}{41} \sin(x) + \frac{1}{41} \cos(x), \quad C \in \mathbb{R}. \quad \blacksquare$$

Beispiel 5.22

Wir kommen noch einmal auf den R/L-Kreis mit der Differenzialgleichung (5.1)

$$\frac{d}{dt}i(t) + \frac{R}{L}i(t) = \frac{u(t)}{L}$$

mit Anfangsbedingung $i(0) = 0$ zurück und legen nun eine Spannung $u(t) = u_0 \sin(\omega t)$ mit $\omega > 0$ an. Die allgemeine Lösung $i_h = c \exp\left(-\frac{R}{L}t\right)$ der homogenen Differenzialgleichung haben wir bereits zuvor berechnet.

Für eine partikuläre Lösung der inhomogenen Gleichung verwendet man den Ansatz vom Typ der rechten Seite

$$i_p(t) = A \cos(\omega t) + B \sin(\omega t) \text{ mit Ableitung } i_p'(t) = -A\omega \sin(\omega t) + B\omega \cos(\omega t).$$

Wir setzen $i_p(t)$ in die inhomogene Differenzialgleichung ein und erhalten

$$\begin{aligned} -A\omega \sin(\omega t) + B\omega \cos(\omega t) + \frac{R}{L} \cdot (A \cos(\omega t) + B \sin(\omega t)) &= \frac{u_0}{L} \cdot \sin(\omega t) \\ \iff \left(\frac{R}{L}B - A\omega\right) \cdot \sin(\omega t) + \left(\frac{R}{L}A + B\omega\right) \cdot \cos(\omega t) &= \frac{u_0}{L} \cdot \sin(\omega t). \end{aligned}$$

Da $\sin(\omega t)$ und $\cos(\omega t)$ linear unabhängig sind, ergibt ein Koeffizientenvergleich:

$$\begin{aligned} \frac{R}{L}A + B\omega &= 0 \implies B = -\frac{R}{L\omega}A \\ \frac{R}{L}B - A\omega &= \frac{u_0}{L} \implies \left(-\frac{R^2}{L^2\omega} - \omega\right) \cdot A = \frac{u_0}{L} \implies -\frac{R^2 + L^2\omega^2}{L^2\omega} \cdot A = \frac{u_0}{L} \end{aligned}$$

und weiter $A = \frac{-\omega L u_0}{R^2 + L^2 \omega^2}$ und $B = \frac{R u_0}{R^2 + L^2 \omega^2}$. Damit haben wir eine partikuläre Lösung

$$i_p(t) = \frac{u_0}{R^2 + L^2 \omega^2} (-\omega L \cos(\omega t) + R \sin(\omega t)),$$

und die allgemeine Lösung lautet

$$i(t) = C e^{-\frac{R}{L}t} + \frac{u_0}{R^2 + L^2 \omega^2} (-\omega L \cos(\omega t) + R \sin(\omega t)).$$

Aus der Anfangsbedingung $i(0) = 0$ erhalten wir die Konstante C :

$$C \cdot 1 + \frac{u_0}{R^2 + L^2 \omega^2} \cdot (-\omega L) = 0 \implies C = \frac{\omega L u_0}{R^2 + L^2 \omega^2},$$

und die Lösung des Anfangswertproblems lautet

$$i(t) = \frac{u_0}{R^2 + L^2 \omega^2} \left[\underbrace{\omega L \cdot e^{-\frac{R}{L}t}}_{\text{abklingender Anteil}} \underbrace{-\omega L \cos(\omega t) + R \sin(\omega t)}_{\text{stationäre Schwingung}} \right].$$

Nach einer Einschwingphase dominiert die stationäre Schwingung, und der abklingende Anteil kann vernachlässigt werden, d. h., es gilt $i(t) \approx i_p(t)$ für große t .

Diese Situation nach der Einschwingphase erhält man auch über die komplexe Wechselstromrechnung, siehe Kapitel 1.6.5. Mit der komplex erweiterten Spannung

$$\underline{u}(t) = u_0 \underbrace{\left[\cos\left(\omega t - \frac{\pi}{2}\right) + j \sin\left(\omega t - \frac{\pi}{2}\right) \right]}_{\sin(\omega t)}$$

gilt $\underline{u}(t) = (R + j\omega L)\underline{i}(t)$, also

$$\begin{aligned} i(t) &= \operatorname{Re}(\underline{i}(t)) = \operatorname{Re}\left(\frac{1}{R + j\omega L}\underline{u}(t)\right) = \operatorname{Re}\left(\frac{R - j\omega L}{R^2 + \omega^2 L^2}\underline{u}(t)\right) \\ &= \frac{u_0}{R^2 + \omega^2 L^2} \left[R \cos\left(\omega t - \frac{\pi}{2}\right) + \omega L \sin\left(\omega t - \frac{\pi}{2}\right) \right] \\ &= \frac{u_0}{R^2 + \omega^2 L^2} [R \sin(\omega t) - \omega L \cos(\omega t)] = i_p(t). \end{aligned}$$

Die komplexe Wechselstromrechnung berechnet eine spezielle Lösung der Differenzialgleichung für $C = 0$ in der Formel der allgemeinen Lösung. Die Konstante C wird durch einen Anfangswert $i(0)$ bestimmt. Damit funktioniert die komplexe Wechselstromrechnung in der Einschwingphase nicht für jeden Anfangswert.

Um den stationären Anteil der Schwingung etwas kompakter in der Form $A \sin(\omega t - \varphi)$ wie beim Rechnen mit Zeigerdiagrammen darzustellen, formen wir um:

$$i_p(t) = \frac{u_0}{\sqrt{R^2 + L^2 \omega^2}} \left(\frac{R}{\sqrt{R^2 + L^2 \omega^2}} \sin(\omega t) - \frac{\omega L}{\sqrt{R^2 + L^2 \omega^2}} \cos(\omega t) \right).$$

In einem rechtwinkligen Dreieck mit Gegenkathete $L\omega$ und Ankathete R zum Winkel φ ist

$$\cos \varphi = \frac{R}{\sqrt{R^2 + L^2 \omega^2}}, \quad \sin \varphi = \frac{\omega L}{\sqrt{R^2 + L^2 \omega^2}}, \quad \tan \varphi = \frac{\omega L}{R}.$$

Mit diesem Winkel $\varphi \in]0, \frac{\pi}{2}]$ erhalten wir

$$i_p(t) = \frac{u_0}{\sqrt{R^2 + L^2 \omega^2}} [\cos(\varphi) \sin(\omega t) - \sin(\varphi) \cos(\omega t)] \stackrel{(1.28)}{=} \frac{u_0}{\sqrt{R^2 + L^2 \omega^2}} \sin(\omega t - \varphi).$$

Legt man eine sinusförmige Spannung an, so reagiert das System mit einer Schwingung der Kreisfrequenz ω der anliegenden Spannung, frequenzabhängiger Amplitude $A(\omega)$ und Phasenverzögerung $\varphi(\omega)$ mit

$$A(\omega) = \frac{u_0}{\sqrt{R^2 + L^2 \omega^2}}, \quad \varphi(\omega) = \arctan\left(\frac{\omega L}{R}\right).$$

Je größer ω ist, desto kleiner ist die Amplitude, d. h., das System ist ein **Tiefpass** für den Strom. Für kleine $\omega \approx 0$ ist die Amplitude des Stroms ungefähr $\frac{u_0}{R}$. Das entspricht dem Ohm'schen Gesetz. Wegen $\varphi \approx 0$ schwingen Spannung und Strom etwa in Phase. Die Induktivität der Spule hat bei kleiner Frequenz geringen Einfluss. ■

5.2.2 Nicht-lineare Differenzialgleichungen erster Ordnung

Bei den zuvor betrachteten linearen Differenzialgleichungen kann man sofort wie angegeben die Lösung berechnen. Bei anderen Typen von Differenzialgleichungen geht das leider nicht. Hier gibt es einen Zoo unterschiedlicher Verfahren und keine einheitliche Theorie. Je nach Aussehen der Differenzialgleichung wählt man dabei ein Verfahren aus, das speziell nur für diesen Typ von Gleichung funktioniert. Man muss zunächst den Typ der Gleichung bestimmen, bevor man mit der Lösung beginnen kann.

Bei allen Verfahren geht man generell davon aus, dass es eine Lösung gibt und rechnet mit ihr formal, bis man eine Darstellung der Lösung gefunden hat. Anschließend muss man das Ergebnis (und damit die Annahme, dass es eine Lösung gibt) durch Einsetzen in die Differenzialgleichung verifizieren und das Lösungsintervall bestimmen.

5.2.2.1 Trennung der Variablen (Separation)

In diesem Abschnitt betrachten wir Differenzialgleichungen, die die Gestalt

$$y'(x) = f(x, y(x)) = f_1(x) \cdot f_2(y(x)).$$

für geeignete Funktionen f_1 und f_2 haben. Den folgenden Ansatz haben wir bereits bei der Herleitung der Formel (5.7) für homogene Lösungen einer linearen Differenzialgleichung erster Ordnung verwendet.

Angenommen, es gibt eine Lösung $y(x)$, für diese gilt dann (falls $f_2(y(t)) \neq 0$):

$$\begin{aligned} \frac{y'(t)}{f_2(y(t))} = f_1(t) &\quad \implies \quad \int_{x_0}^x \frac{y'(t)}{f_2(y(t))} dt = \int_{x_0}^x f_1(t) dt \\ v=y(t), \frac{dv}{dt}=y'(t) &\quad \implies \quad \int_{y(x_0)}^{y(x)} \frac{1}{f_2(v)} dv = \int_{x_0}^x f_1(t) dt. \end{aligned}$$

Die Integrale müssen nun ausgerechnet und das Ergebnis nach $y(x)$ aufgelöst werden. Der Ansatz wird mit **Trennung der Variablen** bezeichnet, da nach der Umformung die gesuchte Funktion y bzw. $y(x)$ nur auf der linken Seite auftritt, während die rechte Seite nur von der Variable x , aber nicht von y abhängt. In diesem Sinne haben wir x und y getrennt.

Beispiel 5.23

Wir lösen die Differenzialgleichung

$$x(y(x))^2 - 1 + (x^2 - 1)y(x)y'(x) = 0.$$

Ist $x \neq \pm 1$ und $y(x) \neq 0$, können wir sie umformen zu

$$y'(x) = \frac{-x(y(x))^2 - 1}{(x^2 - 1)y(x)} = -\frac{x}{x^2 - 1} \cdot \frac{y(x)^2 - 1}{y(x)} =: f_1(x) \cdot f_2(y(x)).$$

Damit erhalten wir einen Lösungskandidaten über

$$\int_{y(x_0)}^{y(x)} \frac{v}{v^2 - 1} dv = \int_{x_0}^x -\frac{t}{t^2 - 1} dt. \quad (5.11)$$

Man beachte, dass wir hier den Fall $f_2(y(t)) = 0$, also $y(t) = \pm 1$, ausgeschlossen haben. Damit verlieren wir zwei Lösungen der Ausgangsgleichung.

Mittels Substitution $u = t^2 - 1$, $du = 2t dt$ ist:

$$\int^x \frac{t}{t^2 - 1} dt = \frac{1}{2} \int^{x^2 - 1} \frac{1}{u} du = \frac{1}{2} \ln |x^2 - 1| + c,$$

d. h., $\frac{1}{2} \ln |x^2 - 1|$ ist eine Stammfunktion zu $\frac{x}{x^2 - 1}$. Benutzen wir das für beide Seiten von (5.11), so erhalten wir

$$\ln |y(x)^2 - 1| + \ln |x^2 - 1| = c.$$

Jetzt müssen wir die Gleichung nach y auflösen:

$$\begin{aligned} \exp(\ln |y(x)^2 - 1| + \ln |x^2 - 1|) &= \exp(c) \iff |y(x)^2 - 1| = \frac{c'}{|x^2 - 1|} \\ \iff y(x)^2 &= 1 \pm \frac{c'}{|x^2 - 1|} \iff y(x) = \pm \sqrt{1 \pm \frac{c'}{|x^2 - 1|}}. \end{aligned}$$

Das Ergebnis ist unter Berücksichtigung des Definitionsbereichs zu verifizieren. ■

Rein formal kann man die Trennung der Variablen mit anschließender Integration und Substitution auch so schreiben:

$$\frac{dy}{dx} = f_1(x) \cdot f_2(y) \implies \frac{dy}{f_2(y)} = f_1(x) dx, \text{ d. h. } \int \frac{1}{f_2(y)} dy = \int f_1(x) dx + c.$$

Die Multiplikation mit dx ist lediglich eine formale Merkregel, die nach der Definition der Symbole nicht erklärt ist. Zudem wird nach der Variable y integriert. $y = y(x)$ ist aber eigentlich eine Funktion. Die Schreibweise drückt bereits die vorgenommene Substitution aus, nur dass keine neue Variable eingeführt wird. Nach der Berechnung der Stammfunktion wird y wieder als Funktion $y(x)$ interpretiert. Dies entspricht der Rücksubstitution bzw. dem Einsetzen der oberen Grenze. Im Folgenden verzichten wir auf diese Kurzschreibweise.

Beispiel 5.24

Zu lösen ist die Differenzialgleichung

$$y'(x) = \sqrt{1 - y^2(x)}.$$

Wir setzen $f_1(x) = 1$ und $f_2(y) = \sqrt{1 - y^2}$. Für eine Lösung y mit $y(x) \neq \pm 1$ (beachte: Aufgrund der Gleichung ist nur $-1 \leq y(x) \leq 1$ verlangt) erhalten wir die Gleichung

$$\int^x f_1(t) dt = \int^{y(x)} \frac{1}{f_2(v)} dv + c_1 \implies \int^x 1 dt = \int^{y(x)} \frac{1}{\sqrt{1 - v^2}} dv + c_1$$

$$\implies x + c_2 = \arcsin(y(x)) + c_3 \implies y(x) = \sin(x + c_4).$$

Man beachte, dass $\frac{d}{dx} \arcsin x = 1/\sqrt{1-x^2}$ nur für $-1 < x < 1$ gilt. Entsprechend kann der berechnete Kandidat $\sin(x + c_4)$ nicht auf $\mathbb{R}$ Lösung sein. Einsetzen des Kandidaten in die Differenzialgleichung liefert wegen $1 - \sin^2 x = \cos^2 x$

$$\cos(x + c_4) = \sqrt{1 - \sin^2(x + c_4)} = \sqrt{\cos^2(x + c_4)} = |\cos(x + c_4)|.$$

Das kann aber nur gelten, wenn Kosinus nicht-negative Werte annimmt, also z. B. auf $[-\frac{\pi}{2}, \frac{\pi}{2}]$, d. h., für ein Intervall $I \subset [-\frac{\pi}{2} - c_4, \frac{\pi}{2} - c_4]$ haben wir eine Lösung gefunden.

Achtung: Auch $y(x) = 1$ ist eine Lösung, diese hatten wir durch unseren Ansatz zuvor aber ausgeschlossen. ■

Beispiel 5.25

Nun betrachten wir

$$y'(x) = e^{2y(x)} \cos x.$$

Mit $f_1(x) = \cos x$ und $f_2(y) = e^{2y}$ erhalten wir für eine Lösung y die Gleichung (beachte: Jetzt ist $f_2(y) \neq 0$, so dass wir hier keine Lösungen verlieren)

$$\begin{aligned} \int^x f_1(t) dt &= \int^{y(x)} \frac{1}{f_2(v)} dv + c_0 \implies \int^x \cos t dt = \int^{y(x)} e^{-2v} dv + c_0 \\ \implies \sin(x) + c_1 &= -\frac{1}{2} e^{-2v} \Big|^{y(x)} \implies \sin(x) + c_1 = -\frac{1}{2} e^{-2y(x)} + c_2 \\ \implies y(x) &= -\frac{1}{2} \ln(-2 \sin(x) + c_3). \end{aligned}$$

Damit haben wir für jedes $c > -2$ eine Lösung $y(x) = -\frac{1}{2} \ln(c - 2 \sin x)$ gefunden. Auf dem Lösungsintervall muss $2 \sin x < c$ gelten. Für $c > 2$ ist die Lösung auf ganz $\mathbb{R}$ definiert. Probe:

$$y'(x) = -\frac{1}{2} \frac{1}{c - 2 \sin x} (-2 \cos x), \text{ andererseits: } e^{2y(x)} \cos x = \frac{1}{c - 2 \sin x} \cos x. \quad \blacksquare$$



Abb. 5.7: Hochspannungsleitungen, gezeichnet mit dem Kosinushyperbolikus

Beispiel 5.26 (Kettenlinie)

Die Differenzialgleichung $y''(x) = k\sqrt{1 + (y'(x))^2}$, die wir zu Beginn des Kapitels auf Seite 710 für die Kettenlinie hergeleitet haben, lässt sich mittels Trennung der Variablen lösen. Dazu berechnen wir zunächst $z(x) := y'(x)$:

$$z'(x) = k\sqrt{1 + (z(x))^2}, \text{ so dass } \int^{z(x)} \frac{1}{\sqrt{1 + v^2}} dv = kx + c,$$

also $\operatorname{arsinh}(z(x)) = kx + c$ und damit $z(x) = \sinh(kx + c)$. Damit erhalten wir

$$y(x) = \int^x z(v) dv = \int^x \sinh(kv + c) dv = \frac{1}{k} \cosh(kv + c) + C.$$

Mit dem Kosinushyperbolicus kann man also die Form eines an den Enden befestigten Seils beschreiben, siehe Abbildung 5.7. ■

Beispiel 5.27 (Freier Fall mit Luftwiderstand)

Wir berechnen die Fallgeschwindigkeit $v(t)$ eines Körpers der Masse m unter Berücksichtigung des Luftwiderstands, vgl. (Westermann, 2008, S. 495). Die auf den Körper wirkenden Kräfte sind:

- die Schwerkraft $m \cdot g$, mit $g = 9,81 \text{ m/s}^2$,
- der Luftwiderstand $k \cdot v^2(t)$ mit Reibungskoeffizient $k > 0$.

Bei einer Startgeschwindigkeit $v(0) = 0$ erhält man aus dem Newton'schen Gesetz „Kraft = Masse mal Beschleunigung“ das Anfangswertproblem

$$m \cdot v'(t) = m \cdot g - k \cdot v^2(t), \quad v(0) = 0.$$

Wir lösen diese Differenzialgleichung mittels Trennung der Variablen:

$$\int_{v(0)}^{v(t)} \frac{m}{m \cdot g - k \cdot u^2} du = \int_0^t 1 du = t. \quad (5.12)$$

Für das linke Integral benutzen wir die Substitution $w = \sqrt{\frac{k}{mg}} \cdot u$:

$$\begin{aligned} & \int_{v(0)}^{v(t)} \frac{1}{g - \frac{k}{m} \cdot u^2} du \stackrel{v(0)=0}{=} \int_0^{v(t)} \frac{1}{g - \frac{k}{m} \cdot u^2} du \\ &= \frac{1}{g} \int_0^{v(t)} \frac{1}{1 - \left(\sqrt{\frac{k}{mg}} \cdot u\right)^2} du = \frac{1}{g\sqrt{\frac{k}{mg}}} \int_0^{\sqrt{\frac{k}{mg}} \cdot v(t)} \frac{1}{1 - w^2} dw \\ &= \sqrt{\frac{m}{kg}} \left[\operatorname{artanh} \left(\sqrt{\frac{k}{mg}} \cdot v(t) \right) - \operatorname{artanh}(0) \right] = \sqrt{\frac{m}{kg}} \operatorname{artanh} \left(\sqrt{\frac{k}{mg}} \cdot v(t) \right). \end{aligned}$$

Eingesetzt in (5.12) erhalten wir $\operatorname{artanh}\left(\sqrt{\frac{k}{mg}} \cdot v(t)\right) = \sqrt{\frac{kg}{m}} \cdot t$.

Um die Gleichung nach $v(t)$ aufzulösen, verwenden wir den Tangenshyperbolikus:

$$\sqrt{\frac{k}{mg}} \cdot v(t) = \tanh\left(\sqrt{\frac{kg}{m}} \cdot t\right),$$

und wir erhalten die zeitabhängige Geschwindigkeit $v(t) = \sqrt{\frac{mg}{k}} \cdot \tanh\left(\sqrt{\frac{kg}{m}} \cdot t\right)$.

Wegen $\lim_{x \rightarrow \infty} \tanh(x) = \lim_{x \rightarrow \infty} \frac{e^x - e^{-x}}{e^x + e^{-x}} = \lim_{x \rightarrow \infty} \frac{1 - e^{-2x}}{1 + e^{-2x}} = \frac{1 - 0}{1 + 0} = 1$ finden wir

$$v_{\max} := \lim_{t \rightarrow \infty} v(t) = \sqrt{\frac{mg}{k}} \cdot \lim_{t \rightarrow \infty} \tanh\left(\sqrt{\frac{kg}{m}} \cdot t\right) = \sqrt{\frac{mg}{k}}$$

als Grenzggeschwindigkeit, d. h., für große Zeitpunkte t wird der Körper praktisch nicht weiter beschleunigt. Reibungskraft und Schwerkraft kompensieren sich gegenseitig. Der Körper fällt nicht schneller als $v_{\max}$. ■

5.2.2.2 Substitution

Beim Lösungsverfahren „Substitution“ wird eine Differenzialgleichung durch Einführen einer gesuchten Funktion z , die von y abhängt, in eine neue Differenzialgleichung überführt, in der dann nur z , aber nicht mehr y vorkommt. Aus deren Lösung für z kann dann y ermittelt werden. Das haben wir bei der Differenzialgleichung zur Kettenlinie auf Seite 738 mit $z(x) := y'(x)$ getan, um eine Differenzialgleichung erster Ordnung zu erhalten. Substitution wird aber vor allem bei Differenzialgleichungen verwendet, deren rechte Seite die Gestalt

$$f(x, y) = g(ax + by + c)$$

mit $a, b, c \in \mathbb{R}$ hat. Wir nehmen an, dass $y(x)$ eine Lösung von $y'(x) = f(x, y(x)) = g(ax + by(x) + c)$ ist. Dann setzen wir

$$z(x) := ax + by(x) + c$$

und erhalten die neue Gleichung, bei der die Funktion z gesucht ist:

$$z'(x) = a + by'(x) = a + bg(z(x)) = 1 \cdot [a + bg(z(x))].$$

Die Gleichung können wir nun mittels Trennung der Variablen lösen:

$$\int_{x_0}^x 1 \, dt = x - x_0 = \int_{z(x_0)}^{z(x)} \frac{1}{a + bg(v)} \, dv.$$

Diese Gleichung muss jetzt für eine konkrete Funktion g nach $z(x)$ aufgelöst werden. Aus $z(x)$ erhalten wir danach $y(x) = \frac{1}{b}[z(x) - ax - c]$.

Beispiel 5.28

Wir lösen die Differenzialgleichung

$$y'(x) = (x + y(x))^2 - 1.$$

Wir setzen $z(x) := x + y(x)$ und erhalten $z'(x) = 1 + y'(x) = 1 + (x + y(x))^2 - 1 = z^2(x)$, also für $z(x) \neq 0$: $1 = z'(x)/z^2(x)$,

$$\int_{x_0}^x 1 \, dt = x - x_0 = \int_{z(x_0)}^{z(x)} \frac{1}{t^2} \, dt = -t^{-1} \Big|_{z(x_0)}^{z(x)} = -\frac{1}{z(x)} + \frac{1}{z(x_0)}.$$

Damit ist also $z(x) = -\frac{1}{x+c}$, und als Kandidaten erhalten wir $y(x) = -x - \frac{1}{x+c}$.

Achtung: Wir haben zuvor $z(x) = 0$, also $y(x) = -x$, ausgeschlossen. Dabei handelt es sich aber auch um eine Lösung. ■

Beispiel 5.29

Bei der Differenzialgleichung (vgl. Beispiel 5.28)

$$y'(x) = (x + y(x))^2$$

substituieren wir $z(x) := x + y(x)$ und erhalten damit folgende Differenzialgleichung in z :

$$z'(x) = 1 + y'(x) = 1 + (x + y(x))^2 = 1 + z^2.$$

Diese wird wieder mittels Trennung der Variablen gelöst. Dabei ist $f_1(x) = 1$ und $f_2(z) = 1 + z^2$. Wir erhalten für eine Lösung z die Gleichung (beachte: $f_2(z) > 0$, wir verlieren keine Lösungen)

$$\begin{aligned} \int^x f_1(t) \, dt &= \int^{z(x)} \frac{1}{f_2(v)} \, dv + c_0 \implies \int^x 1 \, dt = \int^{z(x)} \frac{1}{1+v^2} \, dv + c_0 \\ \implies x + c_1 &= \arctan(z(x)) + c_2 \implies z(x) = \tan(x + c). \end{aligned}$$

Damit ist $y(x) = z(x) - x = \tan(x + c) - x$ auf einem geeigneten Intervall, auf dem $\tan(x + c)$ definiert ist, z. B. $]-\frac{\pi}{2} - c, \frac{\pi}{2} - c[$. Probe:

$$\begin{aligned} y'(x) &= \frac{d}{dx} [\tan(x + c) - x] = \frac{\cos^2(x + c) + \sin^2(x + c)}{\cos^2(x + c)} - 1 \\ &= 1 + \frac{\sin^2(x + c)}{\cos^2(x + c)} - 1 = \tan^2(x + c). \end{aligned}$$

Andererseits: $(x + y(x))^2 = z(x)^2 = \tan^2(x + c)$. ■

5.2.2.3 Homogene Differenzialgleichungen *

Eine **homogene Differenzialgleichung** besitzt eine Darstellung

$$y'(x) = f(x, y(x)) = g\left(\frac{y(x)}{x}\right).$$

Dieser Typ einer Differenzialgleichung muss keine homogene lineare Differenzialgleichung sein und umgekehrt. Der Name „homogen“ wird leider für zwei unterschiedliche Dinge benutzt und erschließt sich aus dem Zusammenhang.

Die Gleichung lässt sich wie im vorangehenden Abschnitt beschrieben mittels Substitution lösen, so dass man danach den Lösungsansatz zur Trennung der Variablen benutzen kann: Wir nehmen wieder an, dass $y(x)$ eine Lösung ist und substituieren $z(x) := \frac{y(x)}{x}$. Dann ist

$$g(z(x)) = y'(x) = z(x) + xz'(x), \quad z'(x) = \frac{g(z(x)) - z(x)}{x},$$

so dass die Variablen getrennt werden können:

$$\ln|x| - \ln|x_0| = \int_{x_0}^x \frac{1}{t} dt = \int_{z(x_0)}^{z(x)} \frac{1}{g(t) - t} dt.$$

Beispiel 5.30

Wir betrachten die Differenzialgleichung

$$xy(x)y'(x) = x^2 + y^2(x), \text{ d. h. } y'(x) = \frac{x}{y(x)} + \frac{y(x)}{x}.$$

Mit $z(x) := \frac{y(x)}{x}$ ist $g(z(x)) = \frac{1}{z(x)} + z(x)$, also

$$\ln|x| + c_1 = \int_{z(x_0)}^{z(x)} \frac{1}{\frac{1}{t} + t - t} dt = \frac{1}{2} t^2 \Big|_{z(x_0)}^{z(x)} = \frac{1}{2} z^2(x) + c_2.$$

Damit ist

$$y^2(x) = x^2 z^2(x) = x^2 [2 \ln|x| + c] = x^2 [\ln(x^2) + c], \quad y(x) = \pm \sqrt{x^2 [\ln(x^2) + c]}. \quad \blacksquare$$

5.2.2.4 Bernoulli-Differenzialgleichungen *

Eine **Bernoulli-Differenzialgleichung** hat die Form

$$y'(x) + g(x)y(x) = h(x)[y(x)]^\alpha, \quad 1 \neq \alpha \in \mathbb{R}.$$

Diesen Typ kann man mittels Division durch $[y(x)]^\alpha$ und Substitution $z(x) = [y(x)]^{1-\alpha}$ auf eine lineare Differenzialgleichung zurückführen (für $\alpha = 1$ liegt bereits eine homogene lineare Differenzialgleichung vor). Sei also wieder $y(x)$ eine Lösung der Gleichung

$$\frac{y'(x)}{[y(x)]^\alpha} + g(x)[y(x)]^{1-\alpha} = h(x).$$

Wir setzen $z(x) := [y(x)]^{1-\alpha}$, dann ist nach Kettenregel $z'(x) = (1-\alpha) \frac{y'(x)}{[y(x)]^\alpha}$. Damit erhalten wir die lineare Differenzialgleichung

$$\frac{1}{1-\alpha} z'(x) + g(x)z(x) = h(x).$$

Mit ihrer Lösung findet man $y(x) = z(x)^{\frac{1}{1-\alpha}}$.

Beispiel 5.31

Die Differenzialgleichung

$$y'(x) - 3xy(x) = xy^2(x)$$

ist vom Bernoulli-Typ mit $g(x) = -3x$, $h(x) = x$, $\alpha = 2$. Die zugehörige lineare Differenzialgleichung lautet

$$\frac{1}{1-2} z'(x) - 3xz(x) = x \implies z'(x) + 3xz(x) = -x.$$

Eine zugehörige homogene Lösung ist (siehe Kapitel 5.2.1)

$$z_h(x) = k_0 \exp\left(-\int 3x dx\right) = k_1 \exp\left(-\frac{3}{2}x^2\right).$$

Eine partikuläre Lösung ist (mit $v = 3t^2/2$, $dv = 3t dt$)

$$\begin{aligned} z_p(x) &= z_h(x) \int_{x_0}^x \frac{-t}{z_h(t)} dt = -z_h(x) \int_{x_0}^x \frac{1}{k_1} t \exp\left(\frac{3}{2}t^2\right) dt \\ &= -\frac{1}{k_1} z_h(x) \int_{\frac{3}{2}x_0^2}^{\frac{3}{2}x^2} \frac{1}{3} \exp(v) dv = -\frac{1}{k_1} z_h(x) \left(\frac{1}{3} \exp\left(\frac{3}{2}x^2\right) + c\right) \\ &= -\frac{1}{3} - c \exp\left(-\frac{3}{2}x^2\right). \end{aligned}$$

Für $c \in \mathbb{R}$ haben wir damit alle Lösungen, und wir müssen wieder zu y zurückkehren: $z(x) = [y(x)]^{1-2} = \frac{1}{y(x)}$, d. h.

$$y(x) = \frac{1}{z(x)} = \left(-\frac{1}{3} - c \exp\left(-\frac{3}{2}x^2\right)\right)^{-1}. \quad \blacksquare$$

5.2.2.5 Ricatti-Differenzialgleichungen *

Hier betrachtet man den Fall der Bernoulli-Differenzialgleichung für $\alpha = 2$ mit einer zusätzlichen Inhomogenität. Es entsteht eine **Ricatti-Differenzialgleichung**

$$y'(x) = P(x)[y(x)]^2 + Q(x)y(x) + R(x).$$

Man kommt bei der Lösung weiter, wenn mindestens eine Lösung bekannt ist. Diese kann man über den Ansatz $y_1(x) = kx^\beta$, $\beta \in \mathbb{R}$, raten. Dann setzt man $y(x) = y_1(x) + [z(x)]^{-1}$ (das ist die Substitution $z(x) := \frac{1}{y(x) - y_1(x)}$) und sucht eine Lösung für z . Einsetzen in die Differenzialgleichung ergibt:

$$y_1'(x) - \frac{z'(x)}{z^2(x)} = P(x) \left(y_1^2(x) + 2 \frac{y_1(x)}{z(x)} + \frac{1}{[z(x)]^2} \right) + Q(x) \left(y_1(x) + \frac{1}{z(x)} \right) + R(x).$$

Da y_1 eine Lösung der Differenzialgleichung ist, können wir $R(x)$ entfernen:

$$\begin{aligned} -\frac{z'(x)}{z^2(x)} &= P(x) \left(2 \frac{y_1(x)}{z(x)} + \frac{1}{z^2(x)} \right) + \frac{Q(x)}{z(x)} \\ \implies z'(x) &= -P(x)(2z(x)y_1(x) + 1) - Q(x)z(x) \\ \implies z'(x) &= -[2P(x)y_1(x) + Q(x)]z(x) - P(x). \end{aligned}$$

Damit sind wir zu einer einfach lösbaren linearen Differenzialgleichung gelangt, aus ihrer Lösung gewinnt man durch Rücksubstitution die gesuchte Funktion y .

Beispiel 5.32

Für $x > 0$ betrachten wir die Gleichung

$$y'(x) = [y(x)]^2 - 3 \frac{y(x)}{x} + \frac{1}{x^2}$$

mit $P(x) = 1$, $Q(x) = -\frac{3}{x}$ und $R(x) = \frac{1}{x^2}$. Durch Einsetzen sieht man, dass $y_1(x) = \frac{1}{x}$ eine Lösung ist. Damit ist die lineare Differenzialgleichung

$$z'(x) = - \left[2y_1(x) - \frac{3}{x} \right] z(x) - 1 = \frac{1}{x} z(x) - 1$$

zu lösen. Ihre allgemeine Lösung ist $z(x) = cx - x \ln x$. Rücksubstitution ergibt damit

$$y(x) = \frac{1}{x} + \frac{1}{cx - x \ln x} = \frac{c - \ln(x) + 1}{x(c - \ln x)}.$$

■

5.2.2.6 Exakte Differenzialgleichungen*

In diesem Abschnitt sehen wir uns eine Klasse von Differenzialgleichungen an, die mit dem Potenzialbegriff der Vektoranalysis gelöst werden können. Die damit verbundene Theorie ist etwas anspruchsvoller, dafür lassen sich diese Gleichungen dann aber recht leicht lösen.

Definition 5.4 (Exakte Differenzialgleichung)

Eine Differenzialgleichung vom Typ

$$P(x, y(x)) + Q(x, y(x))y'(x) = 0 \quad (5.13)$$

heißt **exakt** genau dann, wenn es eine Funktion $F = F(u, v)$ als Abbildung einer Teilmenge von $\mathbb{R}^2$ nach $\mathbb{R}$ gibt, so dass für alle (u, v) aus dem Definitionsbereich gilt:

$$\frac{\partial}{\partial u} F(u, v) = P(u, v), \quad (5.14)$$

$$\frac{\partial}{\partial v} F(u, v) = Q(u, v). \quad (5.15)$$

Die Funktion F heißt ein **Potenzial**.

Potenziale kennen wir bereits aus dem Kapitel über Vektoranalysis, siehe Seite 683.

Satz 5.6 (Charakterisierung exakter Differenzialgleichungen)

Sind P und Q stetig differenzierbar (der Einfachheit wegen auf $\mathbb{R}^2$), so gilt: Die Differenzialgleichung ist exakt genau dann, wenn

$$\frac{\partial}{\partial v} P(u, v) = \frac{\partial}{\partial u} Q(u, v). \quad (5.16)$$

Beweis: Ist die Differenzialgleichung exakt, so ist nach Satz von H. A. Schwarz (siehe Seite 611)

$$\frac{\partial}{\partial v} P(u, v) = \frac{\partial}{\partial v} \frac{\partial}{\partial u} F(u, v) = \frac{\partial}{\partial u} \frac{\partial}{\partial v} F(u, v) = \frac{\partial}{\partial u} Q(u, v).$$

Für die Umkehrung kann man die Funktion F konstruieren (kein Tippfehler: Der erste Integrand hängt von v ab, der zweite jedoch nicht von u , sondern von u_0):

$$F(u, v) = \int_{u_0}^u P(t, v) dt + \int_{v_0}^v Q(u_0, t) dt + c. \quad (5.17)$$

Diese Konstruktion fällt nicht vom Himmel. Man nutzt die Verbindung zwischen der Existenz eines Potenzials und der Wegunabhängigkeit eines Kurvenintegrals aus (Seite 683). Hinter der Formel für F verbirgt sich ein Kurvenintegral entlang eines L-förmigen Wegs (siehe Seite 681). Das müssen wir aber gar nicht wissen, um nachzurechnen, dass F die Gleichungen (5.14) und (5.15) erfüllt. Zunächst gilt mit dem Hauptsatz (Seite 340): $\frac{\partial}{\partial u} F(u, v) = P(u, v) + 0$. Die Eigenschaft $\frac{\partial}{\partial v} F(u, v) = Q(u, v)$ ist etwas schwieriger zu zeigen, da man Grenzwerte vertauschen muss:

$$\begin{aligned} \frac{\partial}{\partial v} F(u, v) &= \frac{\partial}{\partial v} \int_{u_0}^u P(t, v) dt + Q(u_0, v) \\ &= \lim_{h \rightarrow 0} \int_{u_0}^u \frac{P(t, v+h) - P(t, v)}{h} dt + Q(u_0, v) \end{aligned}$$

$$= \lim_{h \rightarrow 0} \frac{1}{h} \int_{u_0}^u \int_0^h \frac{\partial}{\partial v} P(t, v+w) dw dt + Q(u_0, v).$$

Nun verwenden wir den Satz von Fubini (Satz 4.12, Seite 653) (beachte: $\frac{\partial}{\partial v} P(t, v+w)$ ist stetig) und erhalten mit (5.16) und dem Hauptsatz (Satz 2.48):

$$\begin{aligned} \frac{\partial}{\partial v} F(u, v) &\stackrel{\text{Satz 4.12}}{=} \lim_{h \rightarrow 0} \frac{1}{h} \int_0^h \int_{u_0}^u \frac{\partial}{\partial v} P(t, v+w) dt dw + Q(u_0, v) \\ &\stackrel{(5.16)}{=} \lim_{h \rightarrow 0} \frac{1}{h} \int_0^h \int_{u_0}^u \frac{\partial}{\partial t} Q(t, v+w) dt dw + Q(u_0, v) \\ &\stackrel{\text{Satz 2.48}}{=} \lim_{h \rightarrow 0} \frac{1}{h} \int_0^h [Q(u, v+w) - Q(u_0, v+w)] dw + Q(u_0, v) \\ &\stackrel{(2.49)}{=} Q(u, v) - Q(u_0, v) + Q(u_0, v) = Q(u, v), \end{aligned}$$

wobei wir mit (2.49) von Seite 338 auch die Stetigkeit von Q verwendet haben. □

Satz 5.7 (Lösung einer exakten Differenzialgleichung)

Jede Lösung der Gleichung

$$F(x, y(x)) = c$$

ist auch Lösung der Differenzialgleichung (5.13) mit Potenzial $F(u, v)$.

Beweis: Den Satz kann man leicht mit der Kettenregel für Funktionen mehrerer Veränderlicher (Seite 607) nachrechnen: Wir leiten beide Seiten nach x ab und erhalten

$$\begin{aligned} 0 &= \frac{\partial}{\partial x} F(x, y(x)) = \left(\frac{\partial F}{\partial u}(x, y(x)), \frac{\partial F}{\partial v}(x, y(x)) \right) \cdot \begin{pmatrix} 1 \\ y'(x) \end{pmatrix} \\ &= (P(x, y(x)), Q(x, y(x))) \cdot \begin{pmatrix} 1 \\ y'(x) \end{pmatrix} = P(x, y(x)) + Q(x, y(x))y'(x). \end{aligned}$$

Damit erfüllt y die Gleichung (5.13). □

Mit der Bedingung (5.16) kann man schnell prüfen, ob eine Differenzialgleichung exakt ist. In diesem Fall lässt sich ein Potenzial mit (5.17) berechnen und die Differenzialgleichung dann mit Satz 5.7 lösen.

Beispiel 5.33

Für die Differenzialgleichung

$$x + 3y^2(x)y'(x) = 0$$

ist $P(u, v) = u$ und $Q(u, v) = 3v^2$. Die Gleichung ist exakt, da $\frac{\partial P(u, v)}{\partial v} = 0 = \frac{\partial Q(u, v)}{\partial u}$ ist. Mit (5.17) erhalten wir ein zugehöriges Potenzial

$$F(u, v) = \int_{u_0}^u t \, dt + \int_{v_0}^v 3t^2 \, dt + c_0 = \frac{1}{2}u^2 + v^3 + c_1.$$

Damit ist nach Satz 5.7 eine Lösung y von $\frac{x^2}{2} + y^3(x) = c$ gesucht, also $y(x) = \left(c - \frac{x^2}{2}\right)^{1/3}$.

Man kommt auch etwas „pragmatischer“ ans Ziel: Wegen $u = P(u, v) = \frac{d}{du}F(u, v)$ folgt $F(u, v) = \frac{1}{2}u^2 + f(v)$ mit einer Integrationskonstante bezüglich der Variable u , die als Funktion $f(v)$ nur von v abhängt. Entsprechend folgt aus $3v^2 = Q(u, v) = \frac{d}{dv}F(u, v)$, dass $F(u, v) = v^3 + g(u)$. Zusammengesetzt ist damit $F(u, v) = \frac{1}{2}u^2 + v^3$ ein Potenzial, so dass wir wie oben aus $F(x, y(x)) = c$ die Lösung erhalten. Die Funktion F kann man damit häufig auch durch gedankliches Hochintegrieren raten. ■

5.2.2.7 Integrierender Faktor *

Trifft man auf eine Differenzialgleichung, die nicht exakt ist, kann man sie eventuell mit einem Faktor exakt machen und dann als exakte Differenzialgleichung lösen. Hier betrachten wir wieder Differenzialgleichungen vom Typ

$$P(x, y(x)) + Q(x, y(x))y'(x) = 0,$$

wobei es eine Funktion $\mu = \mu(u, v)$ geben möge, so dass die damit erweiterte Gleichung

$$\mu(x, y(x))P(x, y(x)) + \mu(x, y(x))Q(x, y(x))y'(x) = 0$$

exakt ist. μ heißt **integrierender Faktor** oder **Euler'scher Multiplikator**. Wird die entstehende exakte Differenzialgleichung gelöst, hat man auch die Lösung der Ausgangsgleichung, sofern der integrierende Faktor überall von null verschieden ist.

Häufig hängt ein integrierender Faktor nur von der Variable u oder nur von v ab. Daher kann man z. B. den Ansatz $\mu(u, v) = \mu(u)$ versuchen.

Beispiel 5.34

Die Differenzialgleichung

$$x^2 + y - xy'(x) = 0 \text{ mit } P(u, v) = u^2 + v \text{ und } Q(u, v) = -u$$

ist nicht exakt, da

$$\frac{\partial}{\partial v}P(u, v) = 1 \neq \frac{\partial}{\partial u}Q(u, v) = -1.$$

Hier klappt der Ansatz über den integrierenden Faktor $\mu(u)$: Wegen

$$\frac{\partial}{\partial u}F(u, v) = \mu(u)(u^2 + v) \text{ und } \frac{\partial}{\partial v}F(u, v) = -\mu(u)u$$

liefert der Satz von Schwarz (siehe Seite 611) die Bedingung

$$\begin{aligned}\frac{\partial^2}{\partial v \partial u} F(u, v) &= \frac{\partial^2}{\partial u \partial v} F(u, v) \implies \mu(u) = -\mu'(u)u - \mu(u) \\ &\implies \mu'(u) = -\frac{2\mu(u)}{u} \implies \mu(u) = cu^{-2}.\end{aligned}$$

Multiplikation der Gleichung mit dem Faktor x^{-2} führt auf die exakte Differenzialgleichung

$$0 = x^{-2}(x^2 + y) - x^{-2}xy'(x) = 1 + x^{-2}y - x^{-1}y'(x),$$

die wir durch Bestimmung des Potenzials lösen: Aus $\frac{\partial}{\partial x} F(x, y) = 1 + x^{-2}y$ folgt $F(x, y) = x - \frac{y}{x} + f(y)$, und aus $\frac{\partial}{\partial y} F(x, y) = -\frac{1}{x}$ folgt $F(x, y) = -\frac{y}{x} + g(x)$, so dass wir aus $F(x, y(x)) = x - \frac{y(x)}{x} = c$ die Lösung $y(x) = x^2 - cx$ erhalten. ■

Etwas formaler können wir zu integrierenden Faktoren in Spezialfällen auch so gelangen:

- Falls $Q(u, v) \neq 0$ (für alle relevanten u, v) ist und die Funktion

$$h(u, v) := \frac{\frac{\partial}{\partial v} P(u, v) - \frac{\partial}{\partial u} Q(u, v)}{Q(u, v)} = h(u)$$

nur von u und nicht von v abhängig ist, erhält man den integrierenden Faktor

$$\mu(u, v) = \mu(u) = \exp\left(\int h(u) du\right).$$

- Falls $P(u, v) \neq 0$ und die Funktion

$$h(u, v) := \frac{\frac{\partial}{\partial v} P(u, v) - \frac{\partial}{\partial u} Q(u, v)}{P(u, v)} = h(v)$$

nur von v und nicht von u abhängig ist, erhält man den integrierenden Faktor

$$\mu(u, v) = \mu(v) = \exp\left(-\int h(v) dv\right).$$

Wir rechnen den ersten Fall nach. Zu zeigen ist nach Satz 5.6

$$\frac{\partial}{\partial v} [\mu(u)P(u, v)] = \frac{\partial}{\partial u} [\mu(u)Q(u, v)].$$

Da $\mu(u)$ nicht von v abhängt, ist die linke Seite gleich $\mu(u)\frac{\partial}{\partial v} P(u, v)$. Für die rechte Seite gilt:

$$\begin{aligned}\frac{\partial}{\partial u} [\mu(u)Q(u, v)] &= \mu'(u)Q(u, v) + \mu(u)\frac{\partial}{\partial u} Q(u, v) \\ &= h(u)\mu(u)Q(u, v) + \mu(u)\frac{\partial}{\partial u} Q(u, v) \\ &= \mu(u) \left[\frac{\frac{\partial}{\partial v} P(u, v) - \frac{\partial}{\partial u} Q(u, v)}{Q(u, v)} Q(u, v) + \frac{\partial}{\partial u} Q(u, v) \right] = \mu(u)\frac{\partial}{\partial v} P(u, v).\end{aligned}$$

Damit stimmen beide Seiten überein.

Beispiel 5.35

Den integrierenden Faktor aus Beispiel 5.34 erhalten wir auch so:

$$\frac{\frac{\partial}{\partial v}P(u, v) - \frac{\partial}{\partial u}Q(u, v)}{Q(u, v)} = -\frac{2}{u} =: h(u),$$

$$\mu(u) = \exp\left(\int h(u) du\right) = c \exp(-2 \ln |u|) = c|u|^{-2} = \frac{c}{u^2}.$$

Multiplikation der Gleichung mit dem Faktor x^{-2} führt wie zuvor auf eine exakte Differenzialgleichung. ■

5.3 Lineare Differenzialgleichungssysteme

In diesem Kapitel wird ein Lösungsverfahren für lineare Differenzialgleichungssysteme erster Ordnung mit konstanten Koeffizienten ausgehend von einer einfachen Aufgabenstellung aus der Elektrotechnik betrachtet. Die Idee dabei ist, Eigenvektoren und Eigenwerte mit den Eigenschaften der Exponentialfunktion zu verbinden.

5.3.1 Motivation: Eine Schaltung mit Induktivitäten

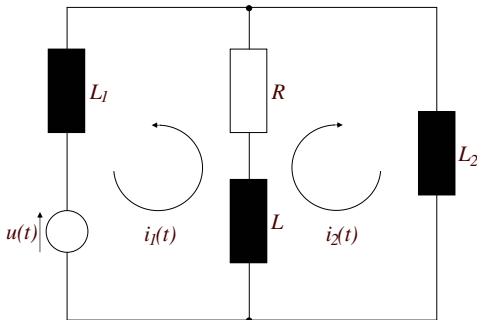


Abb. 5.8: Über eine Induktivität und einen Ohm'schen Widerstand gekoppelte Leiterschleifen

Beispiel 5.36

In Abhängigkeit von der zeitveränderlichen Spannung $u(t)$ interessieren uns die Ströme $i_1(t)$ und $i_2(t)$ in Abbildung 5.8. Zu deren Berechnung wenden wir die Kirchhoff'schen Regeln an. Der Strom durch den Widerstand R beträgt nach der Knotenregel $i_1(t) + i_2(t)$. Nach der Maschenregel addieren sich alle Spannungen einer Masche zu null. Demnach erhalten wir

$$L_1 \frac{d}{dt} i_1(t) + L \frac{d}{dt} (i_1(t) + i_2(t)) + R(i_1(t) + i_2(t)) = u(t)$$

$$\wedge \quad L_2 \frac{d}{dt} i_2(t) + L \frac{d}{dt} (i_1(t) + i_2(t)) + R(i_1(t) + i_2(t)) = 0,$$

wobei für die Rechnung die Richtungen der Spannungen an den Spulen und am Ohm'schen Widerstand mit den eingezeichneten Stromrichtungen übereinstimmen.

Wir suchen nach solchen Lösungen, die zum Zeitpunkt $t = 0$ die Bedingung $i_1(0) = i_2(0) = 0$ erfüllen.

Ist $u(t)$ eine Überlagerung von sinus- oder kosinusförmigen Spannungen $u(t) = \hat{u} \cos(\omega t + \varphi_u)$, so kann man (abgesehen von einer Einschwingphase) die Stromstärken über die komplexe Wechselstromrechnung ermitteln (siehe Seite 144). Hier betrachten wir aber einen beliebigen Spannungsverlauf $u(t)$.

Um die Darstellung im Folgenden übersichtlicher zu gestalten, wählen wir als Werte für die Induktivitäten $L_1 := L_2 := L := 1 \text{ H}$ und $R := 1 \Omega$. Außerdem benennen wir die Variable t in x und die Funktionen i_1 und i_2 in y_1 und y_2 um, so dass wir die in der Theorie der Differenzialgleichungen übliche Notation erhalten. Damit ist das folgende Problem zu lösen: $y_1(0) = y_2(0) = 0$ und

$$\begin{aligned} 2y_1'(x) + y_2'(x) + y_1(x) + y_2(x) &= u(x) \\ \wedge \quad 2y_2'(x) + y_1'(x) + y_1(x) + y_2(x) &= 0. \end{aligned}$$

■

5.3.2 Grundbegriffe

Definition 5.5 (Differenzialgleichungssystem)

Ein System von n Gleichungen, das die unbekannten Funktionen $y_1(x), y_2(x), \dots, y_n(x)$ sowie deren Ableitungen $y_1'(x), y_1''(x), \dots, y_1^{(m_1)}(x), \dots, y_n'(x), y_n''(x), \dots, y_n^{(m_n)}(x)$ enthält, heißt **Differenzialgleichungssystem**. Die höchste auftretende Ableitung wird als **Ordnung** des Differenzialgleichungssystems bezeichnet.

Zuvor haben wir jeweils nur eine einzige Gleichung, also den Fall $n = 1$, diskutiert. Viele Methoden lassen sich aber direkt von einzelnen Differenzialgleichungen auf Systeme mit mehreren Gleichungen übertragen, vgl. Kasten.

Hintergrund: Runge Kutta-Verfahren für ein Systeme erster Ordnung

Das Runge-Kutta-Verfahren aus Kapitel 5.1.5 eignet sich auch zur numerischen Lösung von Differenzialgleichungssystemen. Betrachten wir statt der Gleichung $y'(x) = f(x, y(x))$ das System

$$\begin{aligned} y_1'(x) &= f_1(x, y_1(x), y_2(x)) \\ y_2'(x) &= f_2(x, y_1(x), y_2(x)), \end{aligned}$$

so erhält man ausgehend von einer vorgegebenen Anfangsbedingung $y_1(x_0) = y_{1,0}$, $y_2(x_0) = y_{2,0}$ eine Näherungslösung über $(i = 1, 2)$

$$\begin{aligned} F_{i,1} &:= f_i(x_k, y_{1,k}, y_{2,k}), \\ F_{i,2} &:= f_i\left(x_k + \frac{h}{2}, y_{1,k} + \frac{h}{2}F_{1,1}, y_{2,k} + \frac{h}{2}F_{2,1}\right), \\ F_{i,3} &:= f_i\left(x_k + \frac{h}{2}, y_{1,k} + \frac{h}{2}F_{1,2}, y_{2,k} + \frac{h}{2}F_{2,2}\right), \\ F_{i,4} &:= f_i(x_{k+1}, y_{1,k} + hF_{1,3}, y_{2,k} + hF_{2,3}). \end{aligned}$$

mit der Iterationsvorschrift

$$y_{i,k+1} := y_{i,k} + \frac{h}{6} [F_{i,1} + 2F_{i,2} + 2F_{i,3} + F_{i,4}].$$

Entsprechend kann das Verfahren von 2 auf mehr Gleichungen erweitert werden.

Wir betrachten im Folgenden nur Systeme erster Ordnung – und hier auch nur den Spezialfall linearer Systeme mit konstanten Koeffizienten, zu dem es ein elegantes Lösungsverfahren gibt.

Definition 5.6 (Lineares Differenzialgleichungssystem erster Ordnung)

Gegeben seien eine reelle $(n \times n)$ -Matrix $\mathbf{A}$ und ein Vektor $\vec{b}$ von Funktionen:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}, \quad \vec{b}(x) = \begin{pmatrix} b_1(x) \\ b_2(x) \\ \vdots \\ b_n(x) \end{pmatrix}.$$

Dabei seien die Funktionen $b_1, \dots, b_n : I \rightarrow \mathbb{R}$ auf einem (ggf. unbeschränkten) Intervall $I \subset \mathbb{R}$ definiert.

Gesucht sind auf I differenzierbare Funktionen $y_1(x), y_2(x), \dots, y_n(x)$ mit

$$\begin{pmatrix} y_1'(x) \\ y_2'(x) \\ \vdots \\ y_n'(x) \end{pmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix} \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{pmatrix} + \begin{pmatrix} b_1(x) \\ b_2(x) \\ \vdots \\ b_n(x) \end{pmatrix},$$

also $\frac{d}{dx}\vec{y}(x) = \mathbf{A}\vec{y}(x) + \vec{b}(x)$. Mit anderen Worten:

$$\begin{aligned} y_1'(x) &= a_{1,1}y_1(x) + a_{1,2}y_2(x) + \cdots + a_{1,n}y_n(x) + b_1(x) \\ \wedge \quad y_2'(x) &= a_{2,1}y_1(x) + a_{2,2}y_2(x) + \cdots + a_{2,n}y_n(x) + b_2(x) \end{aligned}$$

$$\wedge \quad y'_n(x) = a_{n,1}y_1(x) + a_{n,2}y_2(x) + \cdots + a_{n,n}y_n(x) + b_n(x).$$

Das Problem heißt ein (gewöhnliches) **lineares Differenzialgleichungssystem erster Ordnung mit konstanten Koeffizienten**.

- Für $\vec{b}(x) = \vec{0}$ heißt das Differenzialgleichungssystem **homogen**. Anderenfalls heißt es **inhomogen**.
- Eine (spezielle) Lösung eines inhomogenen Differenzialgleichungssystems heißt **partikuläre Lösung**. Eine Lösung eines homogenen Differenzialgleichungssystems wird auch als **homogene Lösung** bezeichnet.
- Stellt man zusätzlich die **Anfangsbedingung**

$$\vec{y}(x_0) = \vec{y}_0$$

für ein $x_0 \in I$ und $\vec{y}_0 \in \mathbb{R}^n$, so spricht man von einem **Anfangswertproblem**.

Das Differenzialgleichungssystem aus der Definition heißt außerdem **explizit**, da auf der rechten Seite im Vektor $\vec{y}$ keine Ableitungen auftreten. Vielmehr steht in jeder Gleichung nur die Ableitung einer Komponentenfunktion auf der linken Seite. Lineare Differenzialgleichungssysteme, bei denen die Ableitungen $y'_i(x)$ von gesuchten Lösungen dagegen (linear) von anderen Ableitungen $y'_i(x)$ abhängen, können in die von der Definition geforderte explizite Gestalt gebracht werden. Dazu berechnet man mit dem Gauß'schen-Eliminationsverfahren oder durch Einsetzen eine Lösung für die Variablen $y'_1, \dots, y'_n$. Für unser Beispiel sieht das z. B. so aus:

Beispiel 5.37

Wir bringen das aus der Schaltung abgeleitete System

$$2y'_1(x) + y'_2(x) + y_1(x) + y_2(x) = u(x) \quad \wedge \quad 2y'_2(x) + y'_1(x) + y_1(x) + y_2(x) = 0$$

in die in Definition 5.6 geforderte Struktur: Subtrahiert man die zweite Gleichung von der ersten, so erhält man die äquivalente Aufgabenstellung

$$y'_1(x) - y'_2(x) = u(x) \quad \wedge \quad 2y'_2(x) + y'_1(x) + y_1(x) + y_2(x) = 0.$$

Setzt man die erste Gleichung für $y'_1(x)$ in die zweite ein, wird dort die Abhängigkeit von $y'_1(x)$ eliminiert:

$$y'_1(x) = y'_2(x) + u(x) \quad \wedge \quad 3y'_2(x) = -y_1(x) - y_2(x) - u(x).$$

Schließlich:

$$y'_1(x) = -\frac{1}{3}y_1(x) - \frac{1}{3}y_2(x) + \frac{2}{3}u(x)$$

$$\wedge \quad y_2'(x) = -\frac{1}{3}y_1(x) - \frac{1}{3}y_2(x) - \frac{1}{3}u(x).$$

In der Matrixschreibweise lautet das Differenzialgleichungssystem

$$\frac{d}{dx}\vec{y}(x) = \mathbf{A}\vec{y} + \begin{pmatrix} \frac{2}{3}u(x) \\ -\frac{1}{3}u(x) \end{pmatrix} \quad \text{mit } \mathbf{A} := \begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} \end{bmatrix}. \quad (5.18) \quad \blacksquare$$

Satz 5.8 (Eindeutige Lösung des Anfangswertproblems)

Gegeben sei ein lineares Differenzialgleichungssystem erster Ordnung mit konstanten Koeffizienten wie in Definition 5.6 und mit auf I stetigen Funktionen $b_1(x), \dots, b_n(x)$ sowie $x_0 \in I$ und $\vec{y}_0 \in \mathbb{R}^n$. Dann existiert auf I eine eindeutige Lösung des Anfangswertproblems.

Dieser Satz folgt aus der allgemeineren Theorie für nicht notwendigerweise lineare Systeme, die analog zum bereits diskutierten Fall des Anfangswertproblems einer gewöhnlichen Differenzialgleichung aufgebaut ist (vgl. insbesondere Satz 5.2 auf Seite 717).

Wir wissen jetzt, dass es Lösungen gibt, aber noch nicht, wie man sie berechnet. Bei linearen Systemen besteht ein naheliegender Lösungsansatz darin, durch Gauß-Operationen, Einsetzen und zusätzlich durch Differenzieren die Abhängigkeiten einer gesuchten Funktion $y_k(x)$ von anderen Funktionen $y_l(x)$ zu eliminieren. Dabei entstehen durch das Ableiten aber Differenzialgleichungen höherer Ordnung, die wir zum jetzigen Zeitpunkt noch nicht lösen können.

Beispiel 5.38

Gegeben sei das lineare System

$$\begin{aligned} y_1'(x) &= y_1(x) + y_2(x) + x \\ y_2'(x) &= y_1(x) - y_2(x), \end{aligned} \quad \text{d. h.} \quad \frac{d}{dx}\vec{y}(x) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \vec{y}(x) + \begin{pmatrix} x \\ 0 \end{pmatrix}.$$

Umformung der ersten Gleichung und Differenziation nach x ergibt

$$y_2(x) = y_1'(x) - y_1(x) - x \quad \text{und} \quad y_2'(x) = y_1''(x) - y_1'(x) - 1.$$

Einsetzen in die zweite Gleichung liefert die folgende Differenzialgleichung für $y_1(x)$:

$$y_1''(x) - y_1'(x) - 1 = y_1(x) - (y_1'(x) - y_1(x) - x) \iff y_1''(x) - 2y_1'(x) = 1 + x.$$

Damit ist die Abhängigkeit von y_2 beseitigt, und es ist nun eine Gleichung höherer Ordnung für y_1 zu lösen (siehe Kapitel 5.4). Dann kann man y_2 durch Einsetzen von y_1 ermitteln. ■

Die Eliminationsmethode ist zwar einfach, wenn man Differenzialgleichungen höherer Ordnung lösen kann, in der Praxis aber nur für Systeme mit zwei oder drei Gleichungen sinnvoll anwendbar. Wir werden eine viel elegantere Lösungstechnik kennenlernen. Bei einem inhomogenen Differenzialgleichungssystem wird dabei wieder zunächst das entsprechende homogene System betrachtet und die Menge der zugehörigen homogenen Lösungen berechnet. Mit deren Hilfe kann dann in einem zweiten Schritt eine partikuläre Lösung gewonnen werden, die zugleich eine Anfangsbedingung erfüllt.

5.3.3 Homogene Lösungen

Eine Summe von Vielfachen homogener Lösungen ist selbst wieder eine homogene Lösung. Alle homogenen Lösungen kann man so aus n einzelnen (unabhängigen) Lösungen zusammenbauen. Genauer gilt:

Folgerung 5.1 (Vektorraum der homogenen Lösungen)

Gegeben sei ein homogenes lineares Differenzialgleichungssystem erster Ordnung mit konstanten Koeffizienten wie in Definition 5.6 (ohne Anfangsbedingung). Die Menge $\mathcal{M}$ der Lösungen $\vec{y}$ dieser Aufgabe bildet (bezüglich der üblichen Verknüpfungen „+“ und „ $\cdot$ “) einen reellen Vektorraum $(\mathcal{M}, +; \mathbb{R}, \cdot)$ der Dimension n von vektorwertigen Funktionen auf $\mathbb{R}$.

Hier wird $I = \mathbb{R}$ verwendet, da die Inhomogenität $\vec{b}(x) = \vec{0}$ offensichtlich auf ganz $\mathbb{R}$ definiert und stetig ist, so dass die homogenen Lösungen auf ganz $\mathbb{R}$ erklärt sind.

Die Eigenschaften eines Vektorraums rechnet man wie bei einzelnen linearen Differenzialgleichungen leicht nach (vgl. Seite 724). Bezüglich der Bestimmung der Dimension kann man ausnutzen:

Lemma 5.2 (Lineare Unabhängigkeit homogener Lösungen)

Gegeben sei ein homogenes lineares Differenzialgleichungssystem erster Ordnung mit konstanten Koeffizienten wie in Definition 5.6. Für (homogene) Lösungen $\vec{y}_k$ dieses Systems sind folgende Aussagen äquivalent:

- a) Die Funktionen $\vec{y}_k : \mathbb{R} \rightarrow \mathbb{R}^n$, $1 \leq k \leq m$, sind linear unabhängig.
- b) Für ein $x_0 \in \mathbb{R}$ sind die Vektoren $\vec{y}_k(x_0) \in \mathbb{R}^n$ linear unabhängig.
- c) Für jedes $x_0 \in \mathbb{R}$ sind die Vektoren $\vec{y}_k(x_0) \in \mathbb{R}^n$ linear unabhängig.

Das Lemma führt die lineare Unabhängigkeit von vektorwertigen Funktionen auf die lineare Unabhängigkeit von Vektoren des $\mathbb{R}^n$ zurück, also auf einen Begriff, der viel einfacher zu handhaben ist. Das klappt nicht für beliebige Funktionen, sondern nur speziell für homogene Lösungen.

Beweis: ■ Die Aussage a) folgt aus b) mit der Definition der linearen Unabhängigkeit: Die Funktionen $\vec{y}_k$ heißen linear unabhängig, wenn aus

$$\sum_{i=1}^m r_i \vec{y}_i(x) = \vec{0}$$

für alle $x \in \mathbb{R}$ folgt, dass $r_1 = \dots = r_m = 0$ (siehe Seite 477). Dies ergibt sich wegen b) aber bereits für ein spezielles $x = x_0$.

- Die Implikation c) $\implies$ b) ist trivial, denn da die Unabhängigkeit der Vektoren für alle x_0 gilt, gilt sie insbesondere für ein spezielles.
- Es bleibt zu zeigen, dass a) die Aussage c) impliziert (und damit als Ringschluss automatisch auch b)). Hier verwenden wir die Eigenschaft, dass die $\vec{y}_k$ homogene Lösungen sind: Gilt c) für irgendein $x_0 \in \mathbb{R}$ nicht, so gibt es zu diesem x_0 Skalare $r_1, \dots, r_m$, die nicht alle gleich null sind, mit $\sum_{k=1}^m r_k \vec{y}_k(x_0) = \vec{0}$. Die Funktion $\sum_{k=1}^m r_k \vec{y}_k(x)$ ist aber eine homogene Lösung des Anfangswertproblems zu $\vec{y}(x_0) = \vec{0}$, das gemäß Satz 5.8 eindeutig lösbar ist. Die Nullfunktion ist auch eine Lösung, so dass $\sum_{k=1}^m r_k \vec{y}_k(x) = \vec{0}$ für alle $x \in \mathbb{R}$ ist – im Widerspruch zu a), also zur linearen Unabhängigkeit der Funktionen $\vec{y}_k$. Also gilt c) für alle $x_0 \in \mathbb{R}$. $\square$

Unter Verwendung von Lemma 5.2 kann man nun die Dimension des Lösungsraums ablesen: Nach Satz 5.8 gibt es zu jeder Anfangsbedingung eine Lösung. Wählt man als Anfangswerte die Elemente einer Basis des $(\mathbb{R}^n, +; \mathbb{R}, \cdot)$, hat man nach Lemma 5.2 n linear unabhängige Lösungen gefunden. Eine größere Anzahl von linear unabhängigen Lösungen kann es wegen Lemma 5.2 b) auch nicht geben. Die Dimension ist also n .

Definition 5.7 (Fundamentalsystem)

Gegeben sei ein homogenes lineares Differenzialgleichungssystem erster Ordnung mit konstanten Koeffizienten wie in Definition 5.6. Eine Menge von n linear unabhängigen Lösungen heißt ein **Fundamentalsystem** des Differenzialgleichungssystems. Die Matrix $\mathbf{W}$ ($= \mathbf{W}(x)$), deren Spalten gerade die n Lösungen sind, heißt eine **Wronski-Matrix** des Systems.

Achtung: In der Definition wird nicht von **der** Wronski-Matrix gesprochen. Je nach Wahl der n linear unabhängigen Lösungen entsteht eine andere Matrix!

Jede homogene Lösung ist eine Linearkombination von Elementen des Fundamentalsystems bzw. von Spalten der Wronski-Matrix, da diese ja eine Basis des Vektorraums der homogenen Lösungen bilden. Jede homogene Lösung lässt sich also schreiben als Multiplikation einer Wronski-Matrix mit einem Vektor aus Konstanten.

Jetzt wird es Zeit, dass wir tatsächlich homogene Lösungen und damit Wronski-Matrizen berechnen. Das ist überraschend einfach, wenn man genügend Eigenwerte s und zugehörige Eigenvektoren $\vec{d}$ der Matrix $\mathbf{A}$ des Differenzialgleichungssystems hat. Ein Eigenvektor $\vec{d} \in \mathbb{C}^n$ (also $\vec{d} \neq \vec{0}$) zum Eigenwert $s \in \mathbb{C}$ erfüllt nach Definition 3.28

die Gleichung $\mathbf{A}\vec{d} = s\vec{d}$. Das folgende Lemma beschreibt den entscheidenden Trick, mit dem man lineare Differenzialgleichungssysteme löst:

Lemma 5.3 (Exponentialansatz)

Gegeben sei ein homogenes lineares Differenzialgleichungssystem erster Ordnung wie in Definition 5.6. Sei s ein reeller Eigenwert zu $\mathbf{A}$ und $\vec{d}$ ein reeller Eigenvektor zu s . Dann ist eine homogene Lösung gegeben über

$$\vec{y}(x) := \vec{d}e^{sx}.$$

Dies entspricht genau der bei einer linearen Differenzialgleichung $y'(x) = a_{1,1}y(x)$ berechneten homogenen Lösung $d_1e^{a_{1,1}x}$ (siehe Kapitel 5.2.1). Den folgenden kleinen Beweis sollte man zum Verständnis nachvollziehen:

Beweis: Einerseits nutzen wir aus, dass $\frac{d}{dx}e^{sx} = se^{sx}$:

$$\frac{d}{dx}\vec{y}(x) = \begin{pmatrix} \frac{d}{dx}d_1e^{sx} \\ \vdots \\ \frac{d}{dx}d_ne^{sx} \end{pmatrix} = s\vec{y}(x),$$

andererseits ist $\vec{d}$ ein Eigenvektor zum Eigenwert s :

$$\mathbf{A}\vec{y}(x) = (\mathbf{A}\vec{d})e^{sx} = s\vec{d}e^{sx} = s\vec{y}(x),$$

so dass wir $\frac{d}{dx}\vec{y}(x) = \mathbf{A}\vec{y}(x)$ gezeigt haben. $\square$

Wir betrachten bei Differenzialgleichungssystemen ausschließlich reelle Matrizen $\mathbf{A}$, damit ist es keine Einschränkung, dass im Lemma zu einem reellen Eigenwert auch ein reeller Eigenvektor verlangt wird:

Lemma 5.4 (Existenz reeller Eigenvektoren)

Zu jedem reellen Eigenwert einer reellen Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ gibt es reelle Eigenvektoren.

Beweis: Ist s ein Eigenwert, so erhalten wir Eigenvektoren als nicht-triviale Lösungen des unterbestimmten reellen Gleichungssystems $(\mathbf{A} - s\mathbf{E})\vec{d} = \vec{0}$. Es ist $\det(\mathbf{A} - s\mathbf{E}) = 0$ und damit $\text{Rang}(\mathbf{A} - s\mathbf{E}) < n$. Die Lösungen dieses homogenen Gleichungssystems bilden nach Satz 3.33 auf Seite 527 einen reellen Vektorraum mit Dimension $n - \text{Rang}(\mathbf{A} - s\mathbf{E}) \geq 1$. $\square$

In Lemma 3.22 auf Seite 542 haben wir gezeigt, dass Eigenvektoren $\vec{d}$ zu verschiedenen Eigenwerten s linear unabhängig sind. Damit sind für verschiedene Eigenvektoren $\vec{d}$ auch die Funktionen $\vec{d}e^{sx}$ linear unabhängig wegen Lemma 5.2 für $x_0 = 0$.

Folgerung 5.2 (Existenz eines Fundamentalsystems)

Gibt es zu einem linearen Differenzialgleichungssystem erster Ordnung wie in Definition 5.6 genau n verschiedene reelle Eigenwerte, lässt sich daraus ein Fundamentalsystem berechnen.

Diese Situation haben wir im Beispiel 5.37:

Beispiel 5.39

Zunächst bestimmen wir die Eigenwerte von $\mathbf{A}$ aus (5.18):

$$\det \begin{bmatrix} -\frac{1}{3} - s & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} - s \end{bmatrix} = \left(-\frac{1}{3} - s\right)^2 - \frac{1}{9} = \frac{2}{3}s + s^2 = s \left(\frac{2}{3} + s\right).$$

Die Nullstellen dieses charakteristischen Polynoms und damit die Eigenwerte sind 0 und $-\frac{2}{3}$. Man beachte, dass der Eigenwert 0 zulässig ist (im Gegensatz dazu ist $\vec{0}$ als Eigenvektor nicht definiert).

Bestimmung von Eigenvektoren $\vec{d}$ zum Eigenwert $-\frac{2}{3}$:

$$\begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = -\frac{2}{3} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \text{ d. h. } \quad \wedge \quad \begin{aligned} \frac{1}{3}d_1 - \frac{1}{3}d_2 &= 0 \\ -\frac{1}{3}d_1 + \frac{1}{3}d_2 &= 0, \end{aligned}$$

also $d_1 = d_2$, so dass wir einen Eigenvektor $(1, 1)^\top$ erhalten ($r(1, 1)^\top$ beschreibt für $r \in \mathbb{R} \setminus \{0\}$ alle zugehörigen Eigenvektoren). Analog findet man zum Eigenwert 0 wegen

$$-\frac{1}{3}d_1 - \frac{1}{3}d_2 = 0 \quad \wedge \quad -\frac{1}{3}d_1 - \frac{1}{3}d_2 = 0$$

einen Eigenvektor $(1, -1)^\top$ (genauer ist nun jeder Vektor $r(1, -1)^\top$ für $r \in \mathbb{R} \setminus \{0\}$ ein zugehöriger Eigenvektor).

Für die beiden ausgewählten Eigenvektoren erhalten wir eine Wronski-Matrix

$$\mathbf{W}(x) = \begin{bmatrix} 1 \cdot e^{0x} & 1 \cdot e^{-\frac{2}{3}x} \\ -1 \cdot e^{0x} & 1 \cdot e^{-\frac{2}{3}x} \end{bmatrix} = \begin{bmatrix} 1 & e^{-\frac{2}{3}x} \\ -1 & e^{-\frac{2}{3}x} \end{bmatrix}. \quad \blacksquare$$

Beispiel 5.40

Wir berechnen die allgemeine Lösung des homogenen Systems

$$\begin{aligned} y_1'(x) &= 3y_1(x) + 3y_2(x) \\ y_2'(x) &= 3y_1(x) - 5y_2(x) \end{aligned} \quad \text{bzw.} \quad \vec{y}'(t) = \mathbf{A} \cdot \vec{y}(t), \quad \mathbf{A} := \begin{bmatrix} 3 & 3 \\ 3 & -5 \end{bmatrix}.$$

Das charakteristische Polynom $p(s)$ der Matrix $\mathbf{A}$ lautet

$$p(s) = \det(\mathbf{A} - s\mathbf{E}) = \begin{vmatrix} 3-s & 3 \\ 3 & -5-s \end{vmatrix} = (3-s)(-5-s) - 3 \cdot 3 = s^2 + 2s - 24$$

mit Nullstellen $s = -1 \pm \sqrt{\frac{4}{4} + 24}$, d. h. $s = 4$ oder $s = -6$. Wir bestimmen einen Eigenvektor zu $s = 4$:

$$\left[\begin{array}{cc|c} 3-4 & 3 & 0 \\ & 3 & -5-4 \\ & & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} -1 & 3 & 0 \\ & 3 & -9 \\ & & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} -1 & 3 & 0 \\ & 0 & 0 \\ & & 0 \end{array} \right],$$

d. h. $x_1 = 3x_2$. $\vec{d}_1 = (3, 1)^\top$ ist ein Eigenvektor zu $s = 4$. Dieser ist natürlich nicht eindeutig bestimmt, wir hätten auch jedes skalare Vielfache $r\vec{d}_1$, $r \neq 0$, auswählen können. Wir bestimmen einen Eigenvektor zu $s = -6$:

$$\left[\begin{array}{cc|c} 3+6 & 3 & 0 \\ & 3 & -5+6 \\ & & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} 9 & 3 & 0 \\ & 3 & 1 \\ & & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} 3 & 1 & 0 \\ & 0 & 0 \\ & & 0 \end{array} \right],$$

d. h. $3x_1 = -x_2$. Ein Eigenvektor ist z. B. $\vec{d}_2 = (1, -3)^\top$. Die allgemeine Lösung lautet

$$\vec{y}(x) = C_1 \begin{pmatrix} 3 \\ 1 \end{pmatrix} e^{4x} + C_2 \begin{pmatrix} 1 \\ -3 \end{pmatrix} e^{-6x}, \quad C_1, C_2 \in \mathbb{R}; \quad \begin{bmatrix} 3 \cdot e^{4x} & 1 \cdot e^{-6x} \\ 1 \cdot e^{4x} & -3 \cdot e^{-6x} \end{bmatrix}$$

ist eine Wronski-Matrix. ■

Hintergrund: Wronski-Matrix mittels Matrixdiagonalisierung

Wir haben in diesem Abschnitt homogene Lösungen in der Situation berechnet, dass die Matrix $\mathbf{A}$ zu einer reellen Diagonalmatrix $\mathbf{D}$ mit Diagonalelementen $s_1, s_2, \dots, s_n$ ähnlich ist (siehe Kapitel 3.6.2). Es gibt also eine Transformationsmatrix $\mathbf{X}$ (deren Spalten Eigenvektoren sind) mit $\mathbf{A} = \mathbf{X}\mathbf{D}\mathbf{X}^{-1}$:

$$\begin{aligned} \vec{y}'(x) = \mathbf{A}\vec{y}(x) &\Longleftrightarrow \vec{y}'(x) = \mathbf{X}\mathbf{D}\mathbf{X}^{-1}\vec{y}(x) \Longleftrightarrow \mathbf{X}^{-1}\vec{y}'(x) = \mathbf{D}\mathbf{X}^{-1}\vec{y}(x) \\ &\Longleftrightarrow [\mathbf{X}^{-1}\vec{y}]'(x) = \mathbf{D}[\mathbf{X}^{-1}\vec{y}(x)] \end{aligned} \quad (5.19)$$

mit $\vec{u}(x) := \mathbf{X}^{-1}\vec{y}(x)$ entsteht die Gleichung $\vec{u}'(x) = \mathbf{D}\vec{u}(x)$. Diese hat offensichtlich eine Wronski-Matrix $\mathbf{W}$, bei der nur die Hauptdiagonale besetzt ist – und zwar mit den Werten $e^{s_1 x}, e^{s_2 x}, \dots, e^{s_n x}$. Damit erhalten wir aufgrund der Definition von $\vec{u}$ eine Wronski-Matrix des Ausgangsproblems über $\mathbf{X} \cdot \mathbf{W}$, deren Spalten Eigenvektoren sind, die mit den Faktoren $e^{s_k x}$ multipliziert werden.

5.3.4 Partikuläre Lösungen

Basierend auf den homogenen Lösungen des Differenzialgleichungssystems konstruieren wir nun eine partikuläre Lösung, also eine Lösung des inhomogenen Systems, die zusätzlich eine geforderte Anfangsbedingung erfüllt.

Lemma 5.5 (Berechnung einer partikulären Lösung)

Sei $\mathbf{W}(x)$ eine Wronski-Matrix des Differenzialgleichungssystems erster Ordnung aus Definition 5.6 und seien $b_1(x), b_2(x), \dots, b_n(x)$ stetig auf I sowie $x_0 \in I$. Dann ist

$$\vec{y}_p(x) = \mathbf{W}(x) \left(\int_{x_0}^x \mathbf{W}^{-1}(t) \vec{b}(t) dt + \mathbf{W}^{-1}(x_0) \vec{y}_0 \right)$$

partikuläre Lösung des (inhomogenen) Systems mit Anfangsbedingung $\vec{y}(x_0) = \vec{y}_0$.

Diese Aussage entspricht genau der Berechnung der Lösung des Anfangswertproblems für $n = 1$ (siehe (5.9) auf Seite 728)

$$y(x) = y_h(x) \left[\int_{x_0}^x \frac{h(t)}{y_h(t)} dt + \frac{1}{y_h(x_0)} y_0 \right],$$

wobei statt der einzelnen homogenen Lösung y_h nun eine Wronski-Matrix mit n linear unabhängigen homogenen Lösungen verwendet wird und die Inhomogenität statt $h(x)$ nun $\vec{b}(x)$ heißt. Bevor wir die Formel beweisen, müssen wir noch die Notation erläutern. Hier wird eine vektorwertige Funktion integriert. Darunter versteht man den Vektor der entsteht, wenn man die einzelnen reellwertigen Komponentenfunktionen integriert, also

$$\int_a^b \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} dx := \begin{pmatrix} \int_a^b f_1(x) dx \\ \int_a^b f_2(x) dx \end{pmatrix}.$$

Die Ableitung von Funktionen $\vec{f}: \mathbb{R} \rightarrow \mathbb{R}^n$ geschieht ebenfalls wieder separat für jede Komponentenfunktion, vgl. (4.4) auf Seite 607.

Beweis: Zunächst beachte man, dass die Spalten von $\mathbf{W}(t)$ für jedes $t \in I$ wegen Lemma 5.2 linear unabhängig sind, so dass $\mathbf{W}(t)$ für jedes $t \in I$ invertierbar ist. Damit können wir also über $\mathbf{W}^{-1}(t)$ verfügen. Außerdem sind die Spalten der Matrix $\mathbf{W}(x)$ homogene Lösungen $\vec{y}_1(x), \dots, \vec{y}_n(x)$. Mit $\mathbf{W}(x) = [\vec{y}_1(x) \dots \vec{y}_n(x)]$ erhält man somit

$$\begin{aligned} \frac{d}{dx} \mathbf{W}(x) &= \left[\frac{d}{dx} \vec{y}_1(x) \dots \frac{d}{dx} \vec{y}_n(x) \right] = [\mathbf{A} \vec{y}_1(x) \dots \mathbf{A} \vec{y}_n(x)] \\ &= \mathbf{A} [\vec{y}_1(x) \dots \vec{y}_n(x)] = \mathbf{A} \mathbf{W}(x). \end{aligned} \quad (5.20)$$

$\vec{y}_p$ erfüllt offensichtlich die Anfangsbedingung, da das Integral für $x = x_0$ den Wert $\vec{0}$ annimmt und sich $\mathbf{W}$ und $\mathbf{W}^{-1}$ aufheben. Außerdem gilt nach Hauptsatz der Differenzial- und Integralrechnung (Satz 2.48 auf Seite 340, beachte, dass alle Komponenten von $\mathbf{W}$ und damit auch von $\mathbf{W}^{-1}$ stetig sind, vgl. Beispiel 5.41):

$$\begin{aligned} \frac{d}{dx} \vec{y}_p(x) &\stackrel{(4.4)}{=} \left(\frac{d}{dx} \mathbf{W}(x) \right) \left(\int_{x_0}^x \mathbf{W}^{-1}(t) \vec{b}(t) dt + \mathbf{W}^{-1}(x_0) \vec{y}_0 \right) \\ &\quad + \mathbf{W}(x) \frac{d}{dx} \left(\int_{x_0}^x \mathbf{W}^{-1}(t) \vec{b}(t) dt + \mathbf{W}^{-1}(x_0) \vec{y}_0 \right) \end{aligned}$$

$$\begin{aligned}
& \stackrel{\text{Satz 2.48}}{=} \left(\frac{d}{dx} \mathbf{W}(x) \right) \left(\int_{x_0}^x \mathbf{W}^{-1}(t) \vec{b}(t) dt + \mathbf{W}^{-1}(x_0) \vec{y}_0 \right) + \mathbf{W}(x) \mathbf{W}^{-1}(x) \vec{b}(x) \\
& \stackrel{(5.20)}{=} \mathbf{A} \mathbf{W}(x) \left(\int_{x_0}^x \mathbf{W}^{-1}(t) \vec{b}(t) dt + \mathbf{W}^{-1}(x_0) \vec{y}_0 \right) + \vec{b}(x), \\
& = \mathbf{A} \vec{y}_p(x) + \vec{b}(x).
\end{aligned}$$

Damit erfüllt die angegebene Funktion $\vec{y}_p$ die Differenzialgleichung. $\square$

Beispiel 5.41

Wir berechnen partikuläre Lösungen für das System (5.18). Dazu bestimmen wir $\mathbf{W}^{-1}(x)$ mittels des Gauß'schen Eliminationsverfahrens aus $\mathbf{W}(x)$. Addiert man die erste Zeile zur zweiten und multipliziert man diese dann mit $\frac{1}{2}$, ergibt sich:

$$\left[\begin{array}{cc|cc} 1 & e^{-\frac{2}{3}x} & 1 & 0 \\ -1 & e^{-\frac{2}{3}x} & 0 & 1 \end{array} \right] \Rightarrow \left[\begin{array}{cc|cc} 1 & e^{-\frac{2}{3}x} & 1 & 0 \\ 0 & e^{-\frac{2}{3}x} & \frac{1}{2} & \frac{1}{2} \end{array} \right].$$

Zieht man von der ersten Zeile die zweite ab und multipliziert man die zweite Zeile mit $e^{\frac{2}{3}x}$, so erhält man schließlich

$$\left[\begin{array}{cc|cc} 1 & 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & \frac{1}{2}e^{\frac{2}{3}x} & \frac{1}{2}e^{\frac{2}{3}x} \end{array} \right], \quad W^{-1}(x) = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2}e^{\frac{2}{3}x} & \frac{1}{2}e^{\frac{2}{3}x} \end{bmatrix}.$$

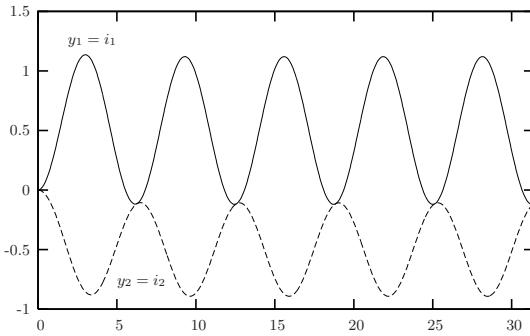
Insbesondere sehen wir, dass durch die Gauß-Umformungen stetige Komponentenfunktionen von $\mathbf{W}^{-1}$ entstehen.

Die partikuläre Lösung zur Anfangsbedingung $\vec{y}(0) = \vec{0}$ ($x_0 = 0$) erhält man für stetiges u damit zu

$$\begin{aligned}
\vec{y}_p(x) &= \begin{bmatrix} 1 & e^{-\frac{2}{3}x} \\ -1 & e^{-\frac{2}{3}x} \end{bmatrix} \left(\int_0^x \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2}e^{\frac{2}{3}t} & \frac{1}{2}e^{\frac{2}{3}t} \end{bmatrix} \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix} u(t) dt + \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \vec{0} \right) \\
&= \begin{bmatrix} 1 & e^{-\frac{2}{3}x} \\ -1 & e^{-\frac{2}{3}x} \end{bmatrix} \begin{pmatrix} \int_0^x \left(\frac{1}{3} + \frac{1}{6} \right) u(t) dt \\ \int_0^x \left(\frac{1}{3}e^{\frac{2}{3}t} - \frac{1}{6}e^{\frac{2}{3}t} \right) u(t) dt \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2} \int_0^x u(t) dt + \frac{1}{6} e^{-\frac{2}{3}x} \int_0^x e^{\frac{2}{3}t} u(t) dt \\ -\frac{1}{2} \int_0^x u(t) dt + \frac{1}{6} e^{-\frac{2}{3}x} \int_0^x e^{\frac{2}{3}t} u(t) dt \end{pmatrix}.
\end{aligned}$$

Wir berechnen die Lösung für die konkrete Spannung $u(t) := \sin t$ (für die man abgesehen von der Anfangsbedingung auch eine komplexe Wechselstromrechnung hätte durchführen können, siehe Kapitel 1.6.5). Zunächst ist

$$\begin{aligned}
\int_0^x e^{\frac{2}{3}t} \sin t dt &= \left[\frac{3}{2} e^{\frac{2}{3}t} \sin t \right]_0^x - \int_0^x \frac{3}{2} e^{\frac{2}{3}t} \cos t dt \\
&= \frac{3}{2} e^{\frac{2}{3}x} \sin(x) - \frac{9}{4} \left[e^{\frac{2}{3}t} \cos t \right]_0^x - \frac{9}{4} \int_0^x e^{\frac{2}{3}t} \sin t dt,
\end{aligned}$$

Abb. 5.9: Lösung zu $u(t) = \sin t$

so dass $\int_0^x e^{\frac{2}{3}t} \sin t \, dt = \frac{6}{13} e^{\frac{2}{3}x} \sin(x) - \frac{9}{13} e^{\frac{2}{3}x} \cos(x) + \frac{9}{13}$.

Damit können wir die Lösung berechnen:

$$\begin{aligned} y_1(x) &= \frac{1}{2} \int_0^x \sin t \, dt + \frac{1}{6} e^{-\frac{2}{3}x} \int_0^x e^{\frac{2}{3}t} \sin t \, dt \\ &= \frac{1}{2} [-\cos t]_0^x + \frac{1}{13} \sin x - \frac{3}{26} \cos x + \frac{3}{26} e^{-\frac{2}{3}x} = \frac{1}{13} \sin x - \frac{8}{13} \cos x + \frac{3}{26} e^{-\frac{2}{3}x} + \frac{1}{2}. \end{aligned}$$

Analog erhält man $y_2(x) = \frac{1}{13} \sin(x) + \frac{5}{13} \cos(x) + \frac{3}{26} e^{-\frac{2}{3}x} - \frac{1}{2}$.

In der ursprünglichen Notation der Physik:

$$\begin{aligned} i_1(t) &= \frac{1}{13} \sin(t) - \frac{8}{13} \cos(t) + \frac{3}{26} e^{-\frac{2}{3}t} + \frac{1}{2} \\ i_2(t) &= \frac{1}{13} \sin(t) + \frac{5}{13} \cos(t) + \frac{3}{26} e^{-\frac{2}{3}t} - \frac{1}{2}. \end{aligned}$$

Diese Lösungen sind auf $[0, \infty[$ beschränkt. Die Exponentialfunktion beschreibt das Einschwingen. Diesen schnell abfallenden Anteil der Lösung, der der Anfangsbedingung geschuldet ist, erhält man mit der komplexen Wechselstromrechnung nicht.

Legt man zum Zeitpunkt $t = 0$ eine konstante Spannung $u(t) = 1$ V an, erhält man

$$i_1(t) = \frac{1}{2}t + \frac{1}{4} - \frac{1}{4}e^{-\frac{2}{3}t}, \quad i_2(t) = -\frac{1}{2}t + \frac{1}{4} - \frac{1}{4}e^{-\frac{2}{3}t}.$$

Hier sieht man, dass mit der Zeit t die Ströme nur (nahezu) linear anwachsen, obwohl die idealen Induktivitäten die Spannungsquelle bei Anlegen einer Gleichspannung vermeintlich kurzschließen. Aufgrund der Eigeninduktivität der Spulen wird der Strom aber auf einen (nahezu) linearen Anstieg begrenzt. ■

Bei der Ermittlung einer partikulären Lösung $\vec{y}_p(x)$ des inhomogenen Systems

$$\vec{y}'(x) = \mathbf{A} \vec{y}(x) + \vec{b}(x)$$

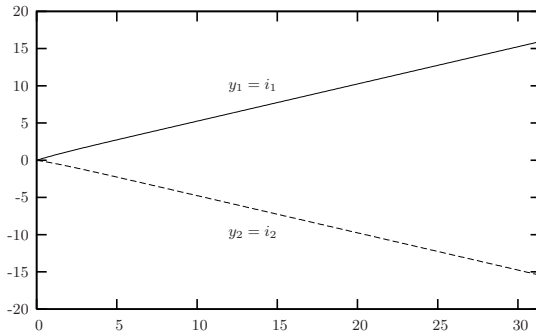


Abb. 5.10: Lösung zu $u(t) = 1$ V für $t \geq 0$

führt häufig ein „**Ansatz vom Typ der rechten Seite $\vec{b}(x)$** “ zum Ziel, ähnlich wie bei den linearen Differenzialgleichungen erster Ordnung (siehe Kapitel 5.2.1.3). Man erhält für Inhomogenitäten, deren Typ beim Einsetzen in das Differenzialgleichungssystem erhalten bleibt, einen allgemeinen Kandidaten für eine partikuläre Lösung und muss nur noch Konstanten bestimmen. Dadurch erspart man sich nicht nur das Ausrechnen des Integrals in Lemma 5.5, sondern auch das Invertieren der Wronski-Matrix.

Beispiel 5.42

Wir berechnen die allgemeine Lösung des inhomogenen Systems

$$\vec{y}'(x) = \mathbf{A} \cdot \vec{y}(x) + \vec{b}(x) \quad \text{mit} \quad \mathbf{A} := \begin{bmatrix} -2 & 3 \\ 2 & -1 \end{bmatrix}, \quad \vec{b}(x) = \begin{pmatrix} x \\ e^{-x} \end{pmatrix}.$$

Das charakteristische Polynom

$$p(s) = \det(\mathbf{A} - s\mathbf{E}) = \begin{vmatrix} -2-s & 3 \\ 2 & -1-s \end{vmatrix} = (1+s)(2+s) - 6 = s^2 + 3s - 4$$

hat die Nullstellen $-\frac{3}{2} \pm \sqrt{\frac{9}{4} + 4}$, d. h. $s = -4$ und $s = 1$. Wegen

$$\left[\begin{array}{cc|c} -2+4 & 3 & 0 \\ 2 & -1+4 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} 2 & 3 & 0 \\ 2 & 3 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} 2 & 3 & 0 \\ 0 & 0 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 \end{array} \right]$$

ist z. B. $\vec{d}_1 = (1, -\frac{2}{3})^\top$ ein Eigenvektor zum Eigenwert -4 . Für den Eigenwert $s = 1$ erhalten wir wegen

$$\left[\begin{array}{cc|c} -2-1 & 3 & 0 \\ 2 & -1-1 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} -3 & 3 & 0 \\ 2 & -2 & 0 \end{array} \right] \iff \left[\begin{array}{cc|c} -1 & 1 & 0 \\ 0 & 0 & 0 \end{array} \right]$$

einen Eigenvektor $\vec{d}_2 = (1, 1)^\top$. Die allgemeine Lösung des homogenen Systems lautet damit

$$\vec{y}_h(x) = C_1 \cdot \begin{pmatrix} 1 \\ -\frac{2}{3} \end{pmatrix} e^{-4x} + C_2 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^x, \quad C_1, C_2 \in \mathbb{R}.$$

Für eine partikuläre Lösung $\vec{y}_p(x)$ verwenden wir den Ansatz vom Typ der rechten Seite, bei dem wir einen Kandidaten $\vec{y}_p(x)$ mit den Funktionen von $\vec{b}(x)$ zusammensetzen (Polynom von Grad 1 und e^{-x}):

$$\vec{y}_p(x) = \begin{pmatrix} a_0 + a_1x + a_2e^{-x} \\ b_0 + b_1x + b_2e^{-x} \end{pmatrix}.$$

Differenzieren und Einsetzen in das inhomogene System liefert

$$\begin{aligned} \begin{pmatrix} a_1 - a_2 e^{-x} \\ b_1 - b_2 e^{-x} \end{pmatrix} &= \begin{bmatrix} -2 & 3 \\ 2 & -1 \end{bmatrix} \begin{pmatrix} a_0 + a_1x + a_2e^{-x} \\ b_0 + b_1x + b_2e^{-x} \end{pmatrix} + \begin{pmatrix} x \\ e^{-x} \end{pmatrix} \\ &= \begin{pmatrix} -2a_0 - 2a_1x - 2a_2e^{-x} + 3b_0 + 3b_1x + 3b_2e^{-x} + x \\ 2a_0 + 2a_1x + 2a_2e^{-x} - b_0 - b_1x - b_2e^{-x} + e^{-x} \end{pmatrix} \end{aligned}$$

bzw.

$$\begin{aligned} (a_1 + 2a_0 - 3b_0) + (2a_1 - 3b_1)x + (a_2 - 3b_2)e^{-x} &= x, \\ (b_1 - 2a_0 + b_0) + (b_1 - 2a_1)x - 2a_2e^{-x} &= e^{-x}. \end{aligned}$$

Nun ist ein Koeffizientenvergleich möglich, da die Funktionen 1, x und e^{-x} linear unabhängig sind. Er liefert das lineare Gleichungssystem

$$\begin{aligned} 2a_0 + a_1 - 3b_0 &= 0 & \wedge & & b_1 - 2a_0 + b_0 &= 0 \\ \wedge & & 2a_1 - 3b_1 &= 1 & \wedge & b_1 - 2a_1 &= 0 \\ \wedge & & a_2 - 3b_2 &= 0 & \wedge & -2a_2 &= 1 \end{aligned}$$

mit Lösungen $a_2 = -\frac{1}{2}$, $b_2 = -\frac{1}{6}$, $b_1 = -\frac{1}{2}$, $a_1 = -\frac{1}{4}$, $b_0 = -\frac{3}{8}$ und $a_0 = -\frac{7}{16}$. Wir erhalten die partikuläre Lösung

$$\vec{y}_p(x) = \begin{pmatrix} -\frac{7}{16} - \frac{1}{4}x - \frac{1}{2}e^{-x} \\ -\frac{3}{8} - \frac{1}{2}x - \frac{1}{6}e^{-x} \end{pmatrix}.$$

Die allgemeine Lösung lautet

$$\vec{y}(x) = C_1 \cdot \begin{pmatrix} 1 \\ -\frac{2}{3} \end{pmatrix} e^{-4x} + C_2 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^x + \begin{pmatrix} -\frac{7}{16} - \frac{1}{4}x - \frac{1}{2}e^{-x} \\ -\frac{3}{8} - \frac{1}{2}x - \frac{1}{6}e^{-x} \end{pmatrix}, \quad C_1, C_2 \in \mathbb{R}. \quad \blacksquare$$

5.3.5 Komplexe und mehrfache Eigenwerte*

Bei dem zuvor verwendeten Lösungsweg können zwei Schwierigkeiten einzeln oder gemeinsam auftreten, so dass die Wronski-Matrix anders bestimmt werden muss.

- Das charakteristische Polynom $\det(\mathbf{A} - s\mathbf{E})$ hat keine n reellen Nullstellen. Hier muss komplex gerechnet werden.
- Es gibt im Fall mehrfacher Nullstellen keine n paarweise verschiedenen Eigenwerte und damit eventuell nicht genug linear unabhängige Eigenvektoren. Hier hilft der Übergang von Eigenvektoren zu Hauptvektoren.

5.3.5.1 Komplexe Eigenwerte

Nach dem Fundamentalsatz der Algebra (siehe Seite 147) hat ein Polynom (mit reellen oder komplexen Koeffizienten) genau n komplexe Nullstellen unter Berücksichtigung ihrer Vielfachheit. Das haben wir bereits bei der Partialbruchzerlegung in Kapitel 2.6.4.3 ausgenutzt. Auch wenn komplexe Nullstellen auftreten, lassen sich reelle Lösungen des homogenen Differenzialgleichungssystems gewinnen.

Im Beispiel (5.18) ist $\mathbf{A}$ eine reelle, symmetrische Matrix. Solche Matrizen haben nach Satz 3.36 ausschließlich reelle Eigenwerte, so dass wir keine komplexwertigen Lösungen betrachten mussten.

Wir benutzen jetzt komplexwertige Funktionen einer reellen Variable. Alle gängigen Operationen auf diesen Funktionen sind komponentenweise separat für den Real- und für den Imaginärteil (sowie separat für alle Komponenten der Vektoren) erklärt. Wie bereits in Kapitel 1.6 eingeführt und auf Seite 419 begründet sei ($z = u + jv$)

$$e^z = e^{u+jv} := e^u e^{jv} := e^u (\cos v + j \sin v).$$

Insbesondere gilt für $s \in \mathbb{C}$ und $x \in \mathbb{R}$ wegen $sx = x\operatorname{Re}(s) + jx\operatorname{Im}(s)$:

$$\begin{aligned} \frac{d}{dx} e^{sx} &= \frac{d}{dx} \left[e^{x\operatorname{Re}(s)} (\cos(x\operatorname{Im}(s)) + j \sin(x\operatorname{Im}(s))) \right] \\ &= \operatorname{Re}(s) e^{x\operatorname{Re}(s)} [\cos(x\operatorname{Im}(s)) + j \sin(x\operatorname{Im}(s))] \\ &\quad + e^{x\operatorname{Re}(s)} [(-\operatorname{Im}(s) \sin(x\operatorname{Im}(s)) + j \operatorname{Im}(s) \cos(x\operatorname{Im}(s)))] \\ &= \operatorname{Re}(s) e^{x\operatorname{Re}(s)} [\cos(x\operatorname{Im}(s)) + j \sin(x\operatorname{Im}(s))] \\ &\quad + j \operatorname{Im}(s) e^{x\operatorname{Re}(s)} [(j \sin(x\operatorname{Im}(s)) + \cos(x\operatorname{Im}(s)))] \\ &= s e^{x\operatorname{Re}(s)} [\cos(x\operatorname{Im}(s)) + j \sin(x\operatorname{Im}(s))] = s e^{sx}. \end{aligned} \quad (5.21)$$

Lemma 5.6 (Reelle Lösungen bei komplexen Eigenwerten)

Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ die reelle Matrix des Differenzialgleichungssystems erster Ordnung aus Definition 5.6. Ist $s \in \mathbb{C}$ ein echt komplexer Eigenwert und $\vec{d} \in \mathbb{C}^n$, $\vec{d} = \vec{d}_1 + j\vec{d}_2$, zugehöriger Eigenvektor von $\mathbf{A}$, dann sind $\operatorname{Re}(\vec{d}e^{sx})$ und $\operatorname{Im}(\vec{d}e^{sx})$ homogene Lösungen.

Mit s ist auch $\bar{s}$ Eigenwert mit Eigenvektor $\bar{\vec{d}} = \vec{d}_1 - j\vec{d}_2$. Die zugehörigen homogenen Lösungen $\operatorname{Re}(\vec{d} e^{sx})$ und $\operatorname{Im}(\vec{d} e^{sx})$ stimmen (bis auf Vorzeichen) mit denen zu s überein und liefern damit keinen weiteren Beitrag zu einem reellen Fundamentalsystem.

Hat die Matrix $\mathbf{A}$ genau n verschiedene komplexe Eigenwerte, so erhält man ein reelles Fundamentalsystem, indem man neben den Lösungen zu reellen Eigenwerten zu jedem Paar $s, \bar{s}$ echt komplexer Eigenwerte Real- und Imaginärteil der Lösung zum Eigenwert s aufnimmt.

Beweis: Wir zeigen, dass $\operatorname{Re}(\vec{d} e^{sx})$ eine homogene Lösung ist. Da Real- und Imaginärteil separat differenziert werden, ist

$$\frac{d}{dx} \operatorname{Re}(\vec{d} e^{sx}) = \operatorname{Re}\left(\frac{d}{dx} \vec{d} e^{sx}\right) \stackrel{(5.21)}{=} \operatorname{Re}(s \vec{d} e^{sx}) = \operatorname{Re}(\mathbf{A} \vec{d} e^{sx}) = \mathbf{A} \operatorname{Re}(\vec{d} e^{sx}),$$

da $\vec{d}$ Eigenvektor zum Eigenwert s und $\mathbf{A}$ reell ist. Entsprechend sieht man, dass auch $\operatorname{Im}(\vec{d} e^{sx})$ das homogene System löst.

Nach Lemma 3.23 (Seite 544) ist mit jedem Eigenwert s auch $\bar{s}$ Eigenwert mit Eigenvektor $\bar{\vec{d}}$. Wegen

$$\begin{aligned} \vec{d} e^{sx} &= (\vec{d}_1 + j\vec{d}_2) e^{(s_1 + js_2)x} = (\vec{d}_1 + j\vec{d}_2) e^{s_1 x} [\cos(s_2 x) + j \sin(s_2 x)] \\ &= \vec{d}_1 e^{s_1 x} \cos(s_2 x) - \vec{d}_2 e^{s_1 x} \sin(s_2 x) + j \left[\vec{d}_1 e^{s_1 x} \sin(s_2 x) + \vec{d}_2 e^{s_1 x} \cos(s_2 x) \right] \\ \bar{\vec{d}} e^{\bar{s}x} &= (\vec{d}_1 - j\vec{d}_2) e^{(s_1 - js_2)x} = (\vec{d}_1 - j\vec{d}_2) e^{s_1 x} [\cos(s_2 x) - j \sin(s_2 x)] \\ &= \vec{d}_1 e^{s_1 x} \cos(s_2 x) - \vec{d}_2 e^{s_1 x} \sin(s_2 x) - j \left[\vec{d}_1 e^{s_1 x} \sin(s_2 x) + \vec{d}_2 e^{s_1 x} \cos(s_2 x) \right] \end{aligned}$$

stimmen die Realteile überein, und die Imaginärteile unterscheiden sich lediglich durch ein Vorzeichen. Wir müssen noch zeigen, dass man durch Auswahl der Real- und Imaginärteile ein reelles Fundamentalsystem erhält. Da alle so entstehenden n Funktionen Lösungen sind, müssen wir nur ihre lineare Unabhängigkeit nachweisen. Nach Lemma 5.2 genügt es, die lineare Unabhängigkeit ihrer Werte an der Stelle $x_0 = 0$ zu untersuchen. Dazu betrachten wir eine reelle Linearkombination, die null ergibt. In der Linearkombination seien die Funktionswerte der Lösungen zu den echt komplexen Eigenwerten s und $\bar{s}$ an der Stelle $x_0 = 0$ vertreten durch $(a, b \in \mathbb{R})$

$$\begin{aligned} a \cdot \operatorname{Re}(\vec{d} e^{s_0}) + b \cdot \operatorname{Im}(\vec{d} e^{s_0}) &= a \cdot \operatorname{Re}(\vec{d}) + b \cdot \operatorname{Im}(\vec{d}) = a\vec{d}_1 + b\vec{d}_2 \\ &= \left(\frac{a}{2} - j\frac{b}{2}\right) (\vec{d}_1 + j\vec{d}_2) + \left(\frac{a}{2} + j\frac{b}{2}\right) (\vec{d}_1 - j\vec{d}_2). \end{aligned}$$

Nun sind $\vec{d} = \vec{d}_1 + j\vec{d}_2$ und $\bar{\vec{d}} = \vec{d}_1 - j\vec{d}_2$ Eigenvektoren zu unterschiedlichen Eigenwerten s und $\bar{s}$. Schreibt man so die reelle Linearkombination für alle Paare $s, \bar{s}$ um, erhält man eine komplexe Linearkombination von komplexen Eigenvektoren, die nach

Lemma 3.22 komplex linear unabhängig sind. Damit müssen alle komplexen Faktoren null sein. Aus $\frac{a}{2} - j\frac{b}{2} = 0$ und $\frac{a}{2} + j\frac{b}{2} = 0$ folgt $a = b = 0$. In der reellen Linearkombination sind damit auch alle Faktoren null, die Lösungen sind reell linear unabhängig. $\square$

Beispiel 5.43

Wir berechnen die allgemeine Lösung des homogenen Systems

$$\vec{y}'(x) = \mathbf{A} \cdot \vec{y}(x) \quad \text{mit} \quad \mathbf{A} := \begin{bmatrix} 1 & -1 \\ 4 & 1 \end{bmatrix}.$$

Das charakteristische Polynom

$$p(s) = \det(\mathbf{A} - s\mathbf{E}) = \begin{vmatrix} 1-s & -1 \\ 4 & 1-s \end{vmatrix} = (1-s)^2 + 4 = s^2 - 2s + 5$$

hat die Nullstellen $s = 1 \pm \sqrt{\frac{4}{4} - 5} = 1 \pm 2j$. Wir berechnen einen Eigenvektor zu $s = 1 + 2j$:

$$\left[\begin{array}{cc|c} 1 - (1 + 2j) & -1 & 0 \\ 4 & 1 - (1 + 2j) & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} -2j & -1 & 0 \\ 4 & -2j & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{cc|c} 2j & 1 & 0 \\ 0 & 0 & 0 \end{array} \right]$$

d. h. $x_2 = -2j x_1$. Damit ist beispielsweise $\vec{d}_1 = (1, -2j)^\top$ ein Eigenvektor. Dieser reicht bereits, um ein reelles Fundamentalsystem zu berechnen. Dieses ergibt sich über den Real- und Imaginärteil der komplexen Lösung: Mit $\vec{y}_1(x) = (1, -2j)^\top e^x \cdot e^{2jx}$ sind $\vec{y}_r(x) := \operatorname{Re}(\vec{y}_1(x))$ und $\vec{y}_i(x) := \operatorname{Im}(\vec{y}_1(x))$ Lösungen des homogenen Systems:

$$\begin{aligned} \vec{y}_r(x) &= \operatorname{Re} \left(\begin{pmatrix} 1 \\ -2j \end{pmatrix} e^x \cdot e^{2jx} \right) = e^x \cdot \operatorname{Re} \left(\begin{pmatrix} \cos(2x) + j \sin(2x) \\ 2 \sin(2x) - 2j \cos(2x) \end{pmatrix} \right) \\ &= e^x \cdot \begin{pmatrix} \cos(2x) \\ 2 \sin(2x) \end{pmatrix}, \\ \vec{y}_i(x) &= \operatorname{Im} \left(\begin{pmatrix} 1 \\ -2j \end{pmatrix} e^x \cdot e^{2jx} \right) = e^x \cdot \operatorname{Im} \left(\begin{pmatrix} \cos(2x) + j \sin(2x) \\ 2 \sin(2x) - 2j \cos(2x) \end{pmatrix} \right) \\ &= e^x \cdot \begin{pmatrix} \sin(2x) \\ -2 \cos(2x) \end{pmatrix}. \end{aligned}$$

Somit erhalten wir die allgemeine (reelle) Lösung

$$\vec{y}(x) = C_1 \cdot \begin{pmatrix} \cos(2x) \\ 2 \sin(2x) \end{pmatrix} e^x + C_2 \cdot \begin{pmatrix} \sin(2x) \\ -2 \cos(2x) \end{pmatrix} e^x, \quad C_1, C_2 \in \mathbb{R}. \quad \blacksquare$$

Beispiel 5.44

Wir berechnen die allgemeine Lösung des homogenen (3×3) -Systems

$$\vec{y}'(x) = \mathbf{A} \cdot \vec{y}(x) \quad \text{mit} \quad \mathbf{A} := \begin{bmatrix} 2 & 0 & 2 \\ 0 & -2 & -2 \\ -2 & 2 & 0 \end{bmatrix}.$$

Das charakteristische Polynom

$$\begin{aligned} p(s) = \det(\mathbf{A} - s\mathbf{E}) &= \begin{vmatrix} 2-s & 0 & 2 \\ 0 & -2-s & -2 \\ -2 & 2 & -s \end{vmatrix} \\ &= (2-s) \cdot \begin{vmatrix} -2-s & -2 \\ 2 & -s \end{vmatrix} + 2 \begin{vmatrix} 0 & -2-s \\ -2 & 2 \end{vmatrix} = -s(s^2 + 4) \end{aligned}$$

hat die Nullstellen 0 , $2j$ und $-2j$. Einen Eigenvektor zu $s = 0$ erhalten wir über

$$\left[\begin{array}{ccc|c} 2 & 0 & 2 & 0 \\ 0 & -2 & -2 & 0 \\ -2 & 2 & 0 & 0 \end{array} \right] \iff \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \iff \begin{array}{l} x_1 = -x_3, \\ x_2 = -x_3, \end{array}$$

z. B. als $\vec{d}_1 = (1, 1, -1)^\top$. Zum Eigenvektor zu $s = 2j$:

$$\begin{aligned} \left[\begin{array}{ccc|c} 2-2j & 0 & 2 & 0 \\ 0 & -2-2j & -2 & 0 \\ -2 & 2 & -2j & 0 \end{array} \right] &\iff \left[\begin{array}{ccc|c} 1 & 0 & \frac{1}{1-j} & 0 \\ 0 & 1 & \frac{1}{1+j} & 0 \\ -1 & 1 & -j & 0 \end{array} \right] \iff \left[\begin{array}{ccc|c} 1 & 0 & \frac{1+j}{2} & 0 \\ 0 & 1 & \frac{1-j}{2} & 0 \\ -1 & 1 & -j & 0 \end{array} \right] \\ &\iff \left[\begin{array}{ccc|c} 1 & 0 & \frac{1+j}{2} & 0 \\ 0 & 1 & \frac{1-j}{2} & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \iff \begin{array}{l} x_1 = -\frac{1+j}{2} x_3, \\ x_2 = -\frac{1-j}{2} x_3, \end{array} \end{aligned}$$

also ist z. B. $\vec{d}_2 = (j, 1, -1-j)^\top$ ein zugehöriger Eigenvektor. Für $s = -2j$ ist ein Eigenvektor $\vec{d}_3 = \vec{d}_2 = (-j, 1, -1+j)^\top$. Wir erhalten die allgemeine (komplexe) Lösung

$$\vec{y}(x) = C_1 \cdot \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} + C_2 \cdot \begin{pmatrix} j \\ 1 \\ -1-j \end{pmatrix} e^{2jx} + C_3 \cdot \begin{pmatrix} -j \\ 1 \\ -1+j \end{pmatrix} e^{-2jx}.$$

Mit

$$\vec{y}_2(x) = \begin{pmatrix} j \\ 1 \\ -1-j \end{pmatrix} (\cos(2x) + j \sin(2x)) = \begin{pmatrix} -\sin(2x) + j \cos(2x) \\ \cos(2x) + j \sin(2x) \\ \sin(2x) - \cos(2x) - j(\sin(2x) + \cos(2x)) \end{pmatrix}$$

erhalten wir die reellen homogenen Lösungen

$$\operatorname{Re}(\vec{y}_2(x)) = \begin{pmatrix} -\sin(2x) \\ \cos(2x) \\ \sin(2x) - \cos(2x) \end{pmatrix}, \quad \operatorname{Im}(\vec{y}_2(x)) = \begin{pmatrix} \cos(2x) \\ \sin(2x) \\ -\sin(2x) - \cos(2x) \end{pmatrix}$$

und bekommen damit die allgemeine reelle Lösung ($C_1, C_2, C_3 \in \mathbb{R}$)

$$\vec{y}(x) = C_1 \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} + C_2 \begin{pmatrix} -\sin(2x) \\ \cos(2x) \\ \sin(2x) - \cos(2x) \end{pmatrix} + C_3 \begin{pmatrix} \cos(2x) \\ \sin(2x) \\ -\sin(2x) - \cos(2x) \end{pmatrix}. \quad \blacksquare$$

5.3.5.2 Mehrfache Eigenwerte: Hauptvektoren

Ist ein Eigenwert mehrfache Nullstelle des charakteristischen Polynoms, lassen sich ggf. keine n linear unabhängigen Eigenvektoren finden. Dann kann eine Wronski-Matrix nicht wie zuvor beschrieben gebaut werden. Man erhält nicht alle homogenen Lösungen. Hier hilft die Verwendung von Hauptvektoren bzw. die Jordan-Normalform der Matrix des Differenzialgleichungssystems (siehe Kapitel 3.6.3). Hat man mit ihrer Hilfe eine Wronski-Matrix bestimmt, ergibt sich daraus eine partikuläre Lösung wie in Abschnitt 5.3.4.

Bei einer reellen, symmetrischen Matrix $\mathbf{A}$ wie im Beispiel (5.18) gibt es diese Probleme nicht, da die Eigenwerte nicht nur alle reell sind, sondern es auch stets n linear unabhängige Eigenvektoren gibt (die sogar eine Orthogonalbasis bilden, siehe Satz 3.41 auf Seite 552).

Hat man einen Hauptvektor $\vec{d}_r$ der Stufe r (d.h. $(\mathbf{A} - s\mathbf{E})^r \vec{d}_r = \vec{0}$ und $(\mathbf{A} - s\mathbf{E})^{r-1} \vec{d}_r \neq \vec{0}$) und die sich daraus ergebenden Hauptvektoren $\vec{d}_i$ der Stufen $1 \leq i < r$ mit $\vec{d}_i = (\mathbf{A} - s\mathbf{E})^{r-i} \vec{d}_r$ gefunden, so hat man gleichzeitig die folgenden linear unabhängigen komplexwertigen homogenen Lösungen für das Differenzialgleichungssystem $\vec{y}'(t) = \mathbf{A}\vec{y}(t)$:

$$\begin{aligned} \vec{y}_1(x) &:= e^{sx} \vec{d}_1, \quad \vec{y}_2(x) := e^{sx} [x \vec{d}_1 + \vec{d}_2], \quad \vec{y}_3(x) := e^{sx} \left[\frac{x^2}{2} \vec{d}_1 + x \vec{d}_2 + \vec{d}_3 \right], \\ \dots, \quad \vec{y}_r(x) &:= e^{sx} \left[\frac{x^{r-1}}{(r-1)!} \vec{d}_1 + \frac{x^{r-2}}{(r-2)!} \vec{d}_2 + \dots + \vec{d}_r \right]. \end{aligned} \quad (5.22)$$

Wir zeigen exemplarisch, dass $\vec{y}_2(x)$ tatsächlich eine homogene Lösung ist. Dazu nutzen wir aus, dass

$$\vec{d}_1 = (\mathbf{A} - s\mathbf{E})\vec{d}_2 \iff s\vec{d}_2 + \vec{d}_1 = \mathbf{A}\vec{d}_2. \quad (5.23)$$

Damit erhalten wir einerseits

$$\begin{aligned} \vec{y}_2'(x) &= se^{sx}(x\vec{d}_1 + \vec{d}_2) + e^{sx}\vec{d}_1 \\ &= sxe^{sx}\vec{d}_1 + e^{sx}(s\vec{d}_2 + \vec{d}_1) \stackrel{(5.23)}{=} sxe^{sx}\vec{d}_1 + e^{sx}\mathbf{A}\vec{d}_2. \end{aligned}$$

Andererseits gilt, da $\vec{d}_1$ ein Eigenvektor ist:

$$\mathbf{A}\vec{y}_2(x) = e^{sx}x\mathbf{A}\vec{d}_1 + e^{sx}\mathbf{A}\vec{d}_2 = sxe^{sx}\vec{d}_1 + e^{sx}\mathbf{A}\vec{d}_2.$$

Damit ist die Gleichung $\vec{y}_2'(x) = \mathbf{A}\vec{y}_2(x)$ erfüllt.

Über das Schema (5.22) lassen sich n linear unabhängige homogene (komplexwertige) Lösungen gewinnen, indem man alle Eigenwerte und zu jedem Eigenwert linear unabhängige Ketten von Hauptvektoren berücksichtigt. Ein Fundamentalsystem kann aus den Real- und Imaginärteilen aller so bestimmten Lösungen ausgewählt werden.

Dieser Lösungsansatz „fällt nicht vom Himmel“, sondern resultiert analog zu (5.19) aus der Überführung der Matrix $\mathbf{A}$ in eine möglichst einfache ähnliche Matrix $\mathbf{T} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$, die hier eine Jordan-Normalform ist (siehe Kapitel 3.6.3).

Beispiel 5.45

a) Wir berechnen die allgemeine Lösung des homogenen (3×3) -Systems

$$\vec{y}'(x) = \mathbf{A} \cdot \vec{y}(x) \quad \text{mit} \quad \mathbf{A} := \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Das (nach der ersten Zeile entwickelte) charakteristische Polynom

$$\begin{aligned} \det(\mathbf{A} - s\mathbf{E}) &= \begin{vmatrix} 1-s & 1 & 1 \\ 1 & 1-s & 1 \\ 1 & 1 & 1-s \end{vmatrix} \\ &= (1-s) \cdot \begin{vmatrix} 1-s & 1 \\ 1 & 1-s \end{vmatrix} - \begin{vmatrix} 1 & 1 \\ 1 & 1-s \end{vmatrix} + \begin{vmatrix} 1 & 1-s \\ 1 & 1 \end{vmatrix} \\ &= (1-s)((1-s)^2 - 1) - (1-s-1) + (1-(1-s)) = s^2(3-s) \end{aligned}$$

hat die Nullstellen $s = 0$ (doppelt) und $s = 3$. Wir bestimmen Eigenvektoren zu $s = 0$:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{array} \right] \iff \left[\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \iff x_1 + x_2 + x_3 = 0.$$

Aus dieser Gleichung können die zwei linear unabhängigen Eigenvektoren

$$\vec{d}_1 = (1, -1, 0)^\top \quad \text{und} \quad \vec{d}_2 = (1, 0, -1)^\top$$

abgelesen werden. Damit haben wir Glück und müssen keine Hauptvektoren der Stufe $k > 1$ zum Eigenwert $s = 0$ bestimmen. Einen Eigenvektor zu $s = 3$ erhalten wir vermöge

$$\begin{aligned} \left[\begin{array}{ccc|c} -2 & 1 & 1 & 0 \\ 1 & -2 & 1 & 0 \\ 1 & 1 & -2 & 0 \end{array} \right] &\Longleftrightarrow \left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1 & 1 & -2 & 0 \end{array} \right] \Longleftrightarrow \left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 3 & -3 & 0 \end{array} \right] \\ &\Longleftrightarrow \begin{aligned} x_1 - 2x_2 + x_3 &= 0, \\ 3x_2 - 3x_3 &= 0 \end{aligned} \end{aligned}$$

z. B. als $\vec{d}_3 = (1, 1, 1)^\top$. Wir erhalten die allgemeine Lösung

$$\vec{y}(x) = C_1 \cdot \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + C_2 \cdot \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + C_3 \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} e^{3x}, \quad C_1, C_2, C_3 \in \mathbb{R}.$$

b) Wir berechnen ein Fundamentalsystem für das lineare System

$$\vec{y}'(t) = \begin{bmatrix} r & 1 & 0 \\ 0 & r & 1 \\ 0 & 0 & r \end{bmatrix} \vec{y}(t).$$

Dabei ist $r \in \mathbb{R}$ ein fester Parameter. Da es sich um eine Dreiecksmatrix handelt, können wir das charakteristische Polynom

$$\det(\mathbf{A} - s\mathbf{E}) = \det \begin{bmatrix} r-s & 1 & 0 \\ 0 & r-s & 1 \\ 0 & 0 & r-s \end{bmatrix}$$

direkt ablesen: $p(s) = (r-s)^3$. Damit haben wir nur den Eigenwert r , der dreifache Nullstelle des charakteristischen Polynoms ist. Wir bestimmen homogene Lösungen über Hauptvektoren. Dazu:

$$(\mathbf{A} - r\mathbf{E})^2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$(\mathbf{A} - r\mathbf{E})^3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Jeder Vektor $\vec{d}_3 \in \mathbb{R}^3$ ist Lösung von $(\mathbf{A} - r\mathbf{E})^3 \vec{d}_3 = \vec{0}$. Aber z. B. $\vec{d}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} =: \vec{e}_3$

erfüllt $(\mathbf{A} - r\mathbf{E})^2 \vec{d}_3 \neq \vec{0}$ und ist Hauptvektor der Stufe 3. Damit erhalten wir Hauptvektoren der Stufen 2 und 1 über

$$\begin{aligned} \vec{d}_2 &:= (\mathbf{A} - r\mathbf{E}) \vec{d}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} =: \vec{e}_2, \\ \vec{d}_1 &:= (\mathbf{A} - r\mathbf{E}) \vec{d}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} =: \vec{e}_1. \end{aligned}$$

Die Matrix $\mathbf{A}$ ist bereits in Jordan-Normalform mit Einsen oberhalb der Hauptdiagonalen gegeben. Wir können sie mit der Transformationsmatrix $\mathbf{X}$, in der wir die Spalten $\vec{e}_3$, $\vec{e}_2$ und $\vec{e}_1$ eintragen, in eine Jordan-Normalform mit Einsen unterhalb der Hauptdiagonalen überführen (vgl. Seite 555).

Mit den Hauptvektoren bekommen wir ein Fundamentalsystem mit den linear unabhängigen homogenen Lösungen

$$\begin{aligned} \vec{y}_1(x) &:= e^{rx} \vec{d}_1 = e^{rx} \vec{e}_1, & \vec{y}_2(x) &:= e^{rx} [x \vec{d}_1 + \vec{d}_2] = e^{rx} [x \vec{e}_1 + \vec{e}_2], \\ \vec{y}_3(x) &:= e^{rx} \left[\frac{x^2}{2} \vec{d}_1 + x \vec{d}_2 + \vec{d}_3 \right] = e^{rx} \left[\frac{x^2}{2} \vec{e}_1 + x \vec{e}_2 + \vec{e}_3 \right]. \end{aligned}$$

Diese ergeben eine Wronski-Matrix $\mathbf{W}(x) := \begin{bmatrix} e^{rx} & xe^{rx} & \frac{x^2}{2}e^{rx} \\ 0 & e^{rx} & xe^{rx} \\ 0 & 0 & e^{rx} \end{bmatrix}.$ ■

5.4 Lineare Differenzialgleichungen höherer Ordnung mit konstanten Koeffizienten

5.4.1 Lösung über ein lineares Differenzialgleichungssystem

Wir kehren hier wieder zurück zu einer einzelnen Differenzialgleichung. Während wir bislang nur Ableitungen erster Ordnung betrachtet haben, sehen wir uns nun lineare

Differenzialgleichungen höherer Ordnung mit konstanten Koeffizienten an. Wir betrachten hier Gleichungen der Form

$$y^{(n)}(x) + a_1 y^{(n-1)}(x) + a_2 y^{(n-2)}(x) + \cdots + a_n y(x) = q(x), \text{ also} \\ y^{(n)}(x) = -a_n y(x) - a_{n-1} y^{(1)}(x) - \cdots - a_1 y^{(n-1)}(x) + q(x).$$

Für $n = 1$ haben wir den Fall einer linearen Differenzialgleichung erster Ordnung mit konstantem Koeffizienten, den wir zuvor bereits ausführlich diskutiert haben.

Mit dem folgenden Satz 5.9 können die zugehörigen homogenen Lösungen ($q(x) = 0$) sehr einfach berechnet werden. Dazu sind lediglich die Nullstellen eines Polynoms zu ermitteln. Diesen Satz wollen wir nun mit unseren Kenntnissen über Differenzialgleichungssysteme herleiten. Als Anwender können Sie unbeschadet auch direkt zum Satz auf Seite 773 springen.

Wir überführen Differenzialgleichungen höherer Ordnung in ein Differenzialgleichungssystem, indem wir Hilfsfunktionen einsetzen, über die die höheren Ableitungen eliminiert werden. Dazu definieren wir diese Hilfsfunktionen:

$$y_0(x) := y(x), \quad y_1(x) := y'(x) = y'_0(x), \quad y_2(x) := y^{(2)}(x) = y'_1(x), \dots, \\ y_{n-1}(x) := y^{(n-1)}(x) = y'_{n-2}(x)$$

also $y^{(k)}(x) = y_k(x)$, $0 \leq k \leq n-1$ und $y'_k(x) = y_{k+1}(x)$, $0 \leq k \leq n-2$. Zusätzlich ergibt sich aus der Differenzialgleichung auch eine Gleichung für $y'_{n-1}(x)$:

$$y'_{n-1}(x) = -a_n y_0(x) - a_{n-1} y_1(x) - \cdots - a_1 y_{n-1}(x) + q(x).$$

Insgesamt erhalten wir das folgende lineare Differenzialgleichungssystem:

$$\begin{pmatrix} y'_0(x) \\ y'_1(x) \\ y'_2(x) \\ \vdots \\ y'_{n-1}(x) \end{pmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \begin{pmatrix} y_0(x) \\ y_1(x) \\ \vdots \\ y_{n-1}(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ q(x) \end{pmatrix}.$$

Falls wir dieses System lösen können, haben wir mit y_0 auch eine Lösung der Ausgangsgleichung. Zur Lösung des homogenen Differenzialgleichungssystems benötigen wir das charakteristische Polynom. Sei $\mathbf{A}$ die obige Matrix. Es ist

$$\det(\mathbf{A} - s\mathbf{E}) = \det \begin{bmatrix} -s & 1 & 0 & \cdots & 0 & 0 \\ 0 & -s & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -s & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 - s \end{bmatrix}.$$

Entwicklung nach der jeweils letzten Spalte liefert

$$\det(\mathbf{A} - s\mathbf{E}) = (-a_1 - s) \cdot$$

$$\begin{aligned} & \cdot \det \begin{bmatrix} -s & 1 & 0 & \cdots & 0 \\ 0 & -s & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -s \end{bmatrix} - \det \begin{bmatrix} -s & 1 & 0 & \cdots & 0 & 0 \\ 0 & -s & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -s & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_3 & -a_2 \end{bmatrix} \\ &= (-1)^{n-1} s^{n-1} \\ &= (-1)^n s^n + (-1)^n a_1 s^{n-1} - (-s)^{n-2} (-a_2) \\ &+ \det \begin{bmatrix} -s & 1 & 0 & \cdots & 0 & 0 \\ 0 & -s & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -s & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_4 & -a_3 \end{bmatrix} \\ &= \cdots = (-1)^n [s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_{n-1} s + a_n] \\ &= (-1)^n \left[s^n + \sum_{k=1}^n a_k s^{n-k} \right] =: p_n(s). \end{aligned} \tag{5.24}$$

Bis auf das Vorzeichen (das für die Nullstellen keine Rolle spielt) ergibt sich das charakteristische Polynom genau aus der Differenzialgleichung, indem wir im Term $y^{(n)}(x) + a_1 y^{(n-1)}(x) + a_2 y^{(n-2)}(x) + \cdots + a_n y^{(0)}(x)$ die Ableitungen $y^{(k)}(x)$ ersetzen durch s^k .

Achtung: $y(x) = y^{(0)}(x)$ wird insbesondere durch $s^0 = 1$ ersetzt. Hier passiert schnell ein Flüchtigkeitsfehler, wenn man $y(x)$ durch $s = s^1$ ersetzt.

Hintergrund: Begleitmatrix

Unabhängig von Differenzialgleichungen sind die Eigenwerte der Matrix $\mathbf{A}$ und ebenso der Matrix $\mathbf{A}^\top$ die Nullstellen des Polynoms (5.24). Das kann man allgemein nutzen, um umgekehrt Nullstellen von Polynomen über Eigenwerte zu berechnen, so dass effiziente numerische Verfahren zur Eigenwertbestimmung verwendet werden können. In diesem Zusammenhang heißt $\mathbf{A}^\top$ die **Begleitmatrix** des Polynoms $s^n + \sum_{k=1}^n a_k s^{n-k}$.

Über die Eigenwerte, d. h. Nullstellen dieses Polynoms, kann eine Lösung des homogenen Systems über Eigen- und Hauptvektoren wie im vorangehenden Kapitel dar-

gestellt berechnet werden, wobei man bei komplexen Eigenwerten zu Real- und Imaginärteilen übergeht.

Man kann zeigen, dass man hier für einen k -fachen Eigenwert s einen Hauptvektor $\vec{d}_k$ der Stufe k findet, zu dem man mittels $\vec{d}_{k-1} = (\mathbf{A} - s\mathbf{E})\vec{d}_k, \dots, \vec{d}_1 = (\mathbf{A} - s\mathbf{E})\vec{d}_2$ Hauptvektoren kleinerer Stufe erhält. Mit diesen bekommt man die (ggf. komplexen) linear unabhängigen homogenen Lösungen

$$\begin{aligned}\vec{y}_1(x) &:= e^{sx} \vec{d}_1 \\ \vec{y}_2(x) &:= e^{sx} \left[x \vec{d}_1 + \vec{d}_2 \right] \\ \vec{y}_3(x) &:= e^{sx} \left[\frac{x^2}{2} \vec{d}_1 + x \vec{d}_2 + \vec{d}_3 \right] \\ &\dots \\ \vec{y}_k(x) &:= e^{sx} \left[\frac{x^{k-1}}{(k-1)!} \vec{d}_1 + \frac{x^{k-2}}{(k-2)!} \vec{d}_2 + \dots \right].\end{aligned}$$

Wir sind nur an $y = y_0$ interessiert und damit nur an der ersten Komponente der Lösungen.

Die erste Komponente des Eigenvektors $\vec{d}_1$ ist von null verschieden. Denn wäre sie gleich null, so kann man aus der erfüllten ersten Gleichung von $(\mathbf{A} - s\mathbf{E})\vec{d}_1 = \vec{0}$ ablesen, dass auch die zweite Komponente null ist. Die zweite Gleichung erzwingt dann, dass die dritte Komponente null ist usw. Es müsste also $\vec{d}_1 = \vec{0}$ sein – im Widerspruch dazu, dass $\vec{0}$ als Eigenvektor nicht zugelassen ist.

Aus den ersten Komponenten der Lösungen können wir damit zu s die linear unabhängigen homogenen Lösungen $e^{sx}, xe^{sx}, \dots, x^{k-1}e^{sx}$ ablesen. Wir müssen dazu die Vektoren $\vec{d}_1, \dots, \vec{d}_k$ gar nicht kennen.

Satz 5.9 (Fundamentalsystem)

Gegeben sei die lineare, homogene Differenzialgleichung n -ter Ordnung

$$y^{(n)}(x) + a_1 y^{(n-1)}(x) + a_2 y^{(n-2)}(x) + \dots + a_n y(x) = 0$$

mit konstanten Koeffizienten. Dazu gehört das charakteristische Polynom (siehe (5.24))

$$p_n(s) = (-1)^n (s^n + a_1 s^{n-1} + a_2 s^{n-2} + \dots + a_n).$$

Sind $s_l, l = 1, \dots, r \leq n$ die (ggf. komplexen) k_l -fachen Nullstellen dieses Polynoms (also $\sum_{l=1}^r k_l = n$), so sind die Funktionen

$$e^{s_1 x}, x e^{s_1 x}, \dots, x^{k_1-1} e^{s_1 x}, e^{s_2 x}, \dots, x^{k_2-1} e^{s_2 x}, \dots, e^{s_r x}, \dots, x^{k_r-1} e^{s_r x}$$

linear unabhängige Lösungen dieser homogenen Differenzialgleichung. Jede homogene Lösung ist eine Linearkombination dieser Funktionen.

Dass es sich bei den angegebenen Funktionen um Lösungen handelt, verifiziert man auch, ohne ein Differenzialgleichungssystem zu bemühen: Ist z. B. s_1 eine Nullstelle des charakteristischen Polynoms, so ist

$$\begin{aligned} \frac{d^n}{dx^n} e^{s_1 x} + a_1 \frac{d^{n-1}}{dx^{n-1}} e^{s_1 x} + a_2 \frac{d^{n-2}}{dx^{n-2}} e^{s_1 x} + \cdots + a_n e^{s_1 x} \\ = s_1^n e^{s_1 x} + a_1 s_1^{n-1} e^{s_1 x} + a_2 s_1^{n-2} e^{s_1 x} + \cdots + a_n e^{s_1 x} = (-1)^n p_n(s_1) e^{s_1 x} = 0. \end{aligned}$$

Beispiel 5.46

Die homogene Differenzialgleichung

$$m x''(t) + b x'(t) + c x = 0$$

eines Feder-Masse-Dämpfer-Systems mit Nullposition $u = 0$, Masse $m > 0$ und Federkonstante $c > 0$ kann nun gelöst werden. Das charakteristische Polynom $p(s) = s^2 + \frac{b}{m}s + \frac{c}{m}$ hat die Nullstellen

$$s = -\frac{b}{2m} \pm \sqrt{\frac{b^2}{4m^2} - \frac{c}{m}} = -\frac{b}{2m} \pm \frac{\sqrt{b^2 - 4mc}}{2m}.$$

Im Fall $b^2 > 4mc$ gibt es zwei (negative) Nullstellen, und die allgemeine Lösung lautet

$$x(t) = c_1 \exp\left(\frac{-b - \sqrt{b^2 - 4mc}}{2m} t\right) + c_2 \exp\left(\frac{-b + \sqrt{b^2 - 4mc}}{2m} t\right). \quad (5.25)$$

Für diese Lösung gilt wegen der negativen Exponenten $\lim_{t \rightarrow \infty} x(t) = 0$, das System kehrt in seine Ruhelage zurück. ■

Eine gesuchte inhomogene Lösung kann man nun ohne die Invertierung der Wronski-Matrix berechnen. Es gilt der folgende Satz:

Satz 5.10 (Partikuläre Lösung)

Gegeben sei die lineare, inhomogene Differenzialgleichung n -ter Ordnung

$$y^{(n)}(x) + a_1 y^{(n-1)}(x) + a_2 y^{(n-2)}(x) + \cdots + a_n y(x) = q(x)$$

mit konstanten Koeffizienten und einer auf einem Intervall I stetigen Funktion q . Weiter sei $x_0 \in I$. Sei $y_0(x)$ die eindeutige Lösung der homogenen Differenzialgleichung mit Anfangsbedingung $y_0(x_0) = y_0'(x_0) = \cdots = y_0^{(n-2)}(x_0) = 0$ und $y_0^{(n-1)}(x_0) = 1$. Dann ist

$$y_p(x) = \int_{x_0}^x y_0(x + x_0 - t) q(t) dt$$

eine partikuläre Lösung der inhomogenen Gleichung n -ter Ordnung mit $y_p(x_0) = 0$. Alle partikulären Lösungen y haben die Darstellung $y_p(x) + y_h(x)$, wobei y_h eine beliebige homogene Lösung ist.

Erfahrungsgemäß wirkt diese Formel zunächst abschreckend, da hier zwei Variablen auftreten und integriert wird. In der Tat kann das Berechnen solcher Integrale aufwändig sein. Im folgenden Abschnitt 5.4.2 betrachten wir daher eine einfachere zu benutzende Technik, mit der man ebenfalls eine partikuläre Lösung erhält. Diese funktioniert allerdings nur bei bestimmten Inhomogenitäten q . Daher lohnt sich doch ein Blick auf den Satz 5.10:

Bei der Berechnung von y_p ist ein Integral bezüglich der Variable t zu berechnen. Dabei ist x eine Konstante, von der das Ergebnis der Integration abhängt. Insgesamt ergibt sich so eine Funktion in der Variable x .

Auf Seite 887 wird erklärt, wie man zu dieser Formel kommt. Hinter ihr verbirgt sich eine Verknüpfung von Funktionen, die man Faltung nennt. Diese werden wir später bei der Fourier- und Laplace-Transformation kennenlernen.

Nach Satz 5.9 lässt sich jede homogene Lösung als Linearkombination von n linear unabhängigen homogenen Lösungen $y_{h,1}(x), \dots, y_{h,n}(x)$ schreiben. Damit hat die inhomogene Differenzialgleichung die allgemeine Lösung

$$y(x) = y_p(x) + c_1 y_{h,1}(x) + c_2 y_{h,2}(x) + \dots + c_n y_{h,n}(x)$$

mit Freiheitsgraden $c_1, \dots, c_n \in \mathbb{R}$. Über eine Anfangsbedingung, bei der an einer Stelle der Funktionswert und die Funktionswerte der Ableitungen bis zur Ordnung $n - 1$ der Lösung vorgegeben sind, sind die Freiheitsgrade eindeutig bestimmt.

Ist $n = 1$, so erhalten wir natürlich mit den Methoden für eine lineare Differenzialgleichung erster Ordnung die gleichen Lösungen: Die Gleichung

$$y'(x) = ay(x) + h(x) \iff y'(x) - ay(x) = h(x)$$

mit konstantem Koeffizienten a hat nach (5.7) auf Seite 723 die allgemeine homogene Lösung $y_h(x) = c \exp(\int a dx) = c \exp(ax)$. Diese erhalten wir auch durch Bestimmung der Nullstelle $s = a$ des charakteristischen Polynoms $p(s) = s - a$ mit dem Ansatz $\exp(ax)$ aus Satz 5.9. Insbesondere ist $y_0(x) = \exp(-ax_0) \exp(ax) = \exp(a(x - x_0))$ die homogene Lösung, die die Anfangsbedingung $y_0(x_0) = 1$ für eine Stelle $x_0 \in \mathbb{R}$ erfüllt. Eine partikuläre Lösung ist nach (5.8) auf Seite 726

$$y_p(x) = \exp(ax) \int_{x_0}^x \frac{h(t)}{\exp(at)} dt = \int_{x_0}^x \exp(a[x - t]) h(t) dt = \int_{x_0}^x y_0(x + x_0 - t) h(t) dt.$$

Damit haben wir genau die Darstellung aus Satz 5.10 erhalten.

Beispiel 5.47

Wir betrachten die Differenzialgleichung

$$y^{(3)}(x) + 2y^{(2)}(x) + y'(x) = 2.$$

Damit haben wir die Koeffizienten $a_1 = 2$, $a_2 = 1$, $a_3 = 0$ und die Inhomogenität $q(x) = 2$. Zur Berechnung der homogenen Lösungen bestimmen wir die Eigenwerte als Nullstellen des charakteristischen Polynoms

$$p_3(s) = (-1)^3(s^3 + 2s^2 + s) = -s(s^2 + 2s + 1) = -s(s+1)^2.$$

Damit ist 0 einfache und -1 doppelte Nullstelle. Linear unabhängige Lösungen der homogenen Differenzialgleichung sind $e^{0x} = 1$, e^{-x} und xe^{-x} . Um nun die inhomogene Lösung zu berechnen, wenden wir den Satz 5.10 für die Stelle $x_0 = 0$ an. Dazu benötigen wir eine spezielle homogene Lösung $y_h(x) = c_1 + c_2e^{-x} + c_3xe^{-x}$, die $y_h(0) = y'_h(0) = 0$ und $y_h^{(2)}(0) = 1$ erfüllt. Dazu:

$$\begin{aligned} y'_h(x) &= -c_2e^{-x} + c_3e^{-x} - c_3xe^{-x}, \\ y_h^{(2)}(x) &= c_2e^{-x} - c_3e^{-x} - c_3e^{-x} + c_3xe^{-x}. \end{aligned}$$

Für die Koeffizienten ergibt sich also durch Einsetzen von $x = 0$ das lineare Gleichungssystem

$$\begin{aligned} c_1 + c_2 &= 0 \\ \wedge \quad -c_2 + c_3 &= 0 \\ \wedge \quad c_2 - 2c_3 &= 1. \end{aligned}$$

Addieren wir die zweite und dritte Gleichung, erhalten wir $c_3 = -1$ und damit $c_2 = -1$. Eingesetzt in die erste Gleichung ist $c_1 = 1$. Die homogene Lösung

$$y_0(x) = 1 - e^{-x} - xe^{-x}$$

erfüllt demnach die Bedingungen aus Satz 5.10 für $x_0 = 0$, und wir erhalten mittels partieller Integration:

$$\begin{aligned} y_p(x) &= \int_0^x y_0(x-t)q(t) dt = \int_0^x (1 - e^{-(x-t)} - (x-t)e^{-(x-t)})2 dt \\ &= 2 \left([t - e^{t-x} - xe^{t-x}]_{t=0}^{t=x} + \int_0^x te^{t-x} dt \right) \\ &= 2 \left(x - 1 - x + e^{-x} + xe^{-x} + [te^{t-x}]_{t=0}^{t=x} - \int_0^x e^{t-x} dt \right) \\ &= 2(-1 + e^{-x} + xe^{-x} + x - 1 + e^{-x}) = 4e^{-x} + 2xe^{-x} + 2x - 4. \end{aligned}$$

Alle Lösungen haben damit die Gestalt ($c_i \in \mathbb{R}$)

$$y(x) = 4e^{-x} + 2xe^{-x} + 2x - 4 + c_1 + c_2e^{-x} + c_3xe^{-x} = 2x + c_4 + c_5e^{-x} + c_6xe^{-x}. \quad (5.26)$$

■

Die Lösung dieses Beispiels lässt sich wegen der einfachen Inhomogenität $q(x) = 2$ viel einfacher mit dem folgenden Ansatz vom Typ der rechten Seite bestimmen.

5.4.2 Lösung mit einem Ansatz vom Typ der rechten Seite

Für spezielle Inhomogenitäten $q(x)$ führt auch hier ein Ansatz für $y_p(x)$ „vom Typ der rechten Seite“ schneller und einfacher zu einer partikulären Lösung von

$$\underbrace{y^{(n)}(x) + a_1 y^{(n-1)}(x) + \cdots + a_{n-2} y''(x) + a_{n-1} y'(x) + a_n y(x)}_{=: L_n[y](x)} = q(x) \quad (5.27)$$

als die Berechnung eines Faltungsintegrals nach Satz 5.10. Als abkürzende Schreibweise haben wir $L_n[y](x)$ als die Funktion definiert, die entsteht, wenn man eine Funktion y in die linke Seite der Differenzialgleichung einsetzt.

Wir erweitern das Vorgehen aus Kapitel 5.2.1.3 auf lineare Gleichungen höherer Ordnung. Es funktioniert weiterhin nur für Inhomogenitäten, deren Typ beim Einsetzen in die Differenzialgleichung erhalten bleibt. Dazu gehören auch hier Polynome, die Exponentialfunktion und die trigonometrischen Funktionen:

$$L_n[\text{Polynom vom Grad } m] = \text{Polynom vom Grad } m, \quad L_n[a \cdot e^{\lambda x}] = A \cdot e^{\lambda x},$$

$$L_n[a \cdot \sin(\omega x) + b \cdot \cos(\omega x)] = A \cdot \sin(\omega x) + B \cdot \cos(\omega x).$$

Die Grundidee besteht nun wieder darin, bei einer Inhomogenität $q(x)$, die sich aus solchen Funktionen zusammensetzt, die partikuläre Lösung $y_p(x)$ als Funktion desselben Typs zu suchen.

5.4.2.1 Einfacher Ansatz vom Typ der rechten Seite

Wir betrachten nun Inhomogenitäten des Typs

$$q(x) = p_m(x) \cdot e^{\alpha x} \cdot \cos(\beta x) \quad \text{oder} \quad q(x) = p_m(x) \cdot e^{\alpha x} \cdot \sin(\beta x) \quad (5.28)$$

mit $\alpha, \beta \in \mathbb{R}$ und einem reellen Polynom $p_m(x)$ vom Grad m .

Der **einfache Ansatz** für eine partikuläre Lösung der inhomogenen Differenzialgleichung (5.27) lautet

$$y_p(x) = r_m(x) \cdot e^{\alpha x} \cdot \cos(\beta x) + s_m(x) \cdot e^{\alpha x} \cdot \sin(\beta x) \quad (5.29)$$

mit reellen Polynomen $r_m(x)$ und $s_m(x)$ vom Grad m . Die Koeffizienten von $r_m(x)$ und $s_m(x)$ sind durch Einsetzen in die Differenzialgleichung zu bestimmen. Ist eine Inhomogenität die Summe beider Typen, so entsteht die Lösung als Summe der beiden Einzellösungen.

Dieser Ansatz ist zielführend, falls $s = \alpha + j\beta$ nicht Nullstelle des charakteristischen Polynoms der Differenzialgleichung ist. Dann sind $\operatorname{Re}(e^{sx}) = e^{\alpha x} \cdot \cos(\beta x)$ und $\operatorname{Im}(e^{sx}) = e^{\alpha x} \cdot \sin(\beta x)$ nicht Lösungen der homogenen Differenzialgleichung. Dadurch erhält man beim Einsetzen der Ansatzfunktion (5.29) in die linke Seite L_n der Differenzialgleichung mit der Produktregel eine Funktion, die die Struktur der Inhomogenität $q(x)$ mit gleichem Grad m des Polynomanteils hat. Über einen Koeffizientenvergleich dieser Funktion mit der Inhomogenität erhält man ein lineares Gleichungssystem für die freien Parameter, das eindeutig lösbar ist.

Funktioniert der einfache Ansatz nicht, weil wir eine Lösung der homogenen Differenzialgleichung verwendet haben, so spricht man vom Resonanzfall. Dann benutzt man einen erweiterten Ansatz, den wir anschließend besprechen.

Beispiel 5.48

Für die Lösung der inhomogenen Differenzialgleichung

$$y''(x) + 3y'(x) + 2y(x) = 8x$$

betrachten wir zunächst die homogene Differenzialgleichung, zu der das charakteristische Polynom $p(s) = s^2 + 3s + 2$ mit Nullstellen $s_1 = -1$ und $s_2 = -2$ gehört. Damit lautet die allgemeine homogene Lösung $y_h(x) = C_1 e^{-x} + C_2 e^{-2x}$, $C_1, C_2 \in \mathbb{R}$. Die Inhomogenität $q(x) = 8x$ ist vom Typ (5.28) mit $\alpha = 0$, $\beta = 0$, und $p_m(x)$ ist ein Polynom vom Grad $m = 1$. Wegen $\alpha + j\beta = 0 + 0j = 0 \neq s_1$ und $0 \neq s_2$ führt der Ansatz vom Typ der rechten Seite

$$y_p(x) = A + Bx$$

zum Ziel. Setzt man $y_p'(x) = B$ und $y_p''(x) = 0$ in die Differenzialgleichung ein, so folgt

$$3B + 2A + 2Bx = 8x \quad \implies \quad B = 4, \quad A = -6,$$

und eine partikuläre Lösung lautet $y_p(x) = -6 + 4x$. Man erhält die allgemeine Lösung der inhomogenen Differenzialgleichung

$$y(x) = C_1 e^{-x} + C_2 e^{-2x} + 4x - 6, \quad C_1, C_2 \in \mathbb{R}. \quad \blacksquare$$

Beispiel 5.49

Für die allgemeine Lösung der inhomogenen Differenzialgleichung

$$y''(x) + y'(x) - 2y(x) = 6 \sin(2x)$$

betrachten wir ebenfalls zunächst die homogene Gleichung. Die Nullstellen des charakteristischen Polynoms $p(s) = s^2 + s - 2$ sind $s_1 = 1$ und $s_2 = -2$, und wir erhalten die allgemeine Lösung der homogenen Gleichung

$$y_h(x) = C_1 e^x + C_2 e^{-2x}, \quad C_1, C_2 \in \mathbb{R}.$$

Die Inhomogenität $q(x) = 6 \sin(2x)$ ist vom Typ (5.28) mit $\alpha = 0$ und $\beta = 2$. Wegen $s_1 \neq 2j \neq s_2$ führt hier der einfache Ansatz zum Ziel. Setzen wir

$$y_p(x) = A \sin(2x) + B \cos(2x)$$

mit $y_p'(x) = 2A \cos(2x) - 2B \sin(2x)$, $y_p''(x) = -4A \sin(2x) - 4B \cos(2x)$ in die inhomogene Differenzialgleichung ein, erhalten wir

$$\begin{aligned} & -4A \sin(2x) - 4B \cos(2x) + 2A \cos(2x) - 2B \sin(2x) \\ & \quad - 2A \sin(2x) - 2B \cos(2x) = 6 \sin(2x) \\ \implies & (-6A - 2B) \sin(2x) + (2A - 6B) \cos(2x) = 6 \sin(2x) \\ \implies & -6A - 2B = 6 \quad \wedge \quad 2A - 6B = 0 \implies A = -\frac{9}{10} \quad \wedge \quad B = -\frac{3}{10}, \end{aligned}$$

also $y_p(x) = -\frac{9}{10} \sin(2x) - \frac{3}{10} \cos(2x)$. Die allgemeine Lösung lautet

$$y(x) = C_1 e^x + C_2 e^{-2x} - \frac{9}{10} \sin(2x) - \frac{3}{10} \cos(2x), \quad C_1, C_2 \in \mathbb{R}. \quad \blacksquare$$

Wenn wir versuchen, die Lösung (5.26) der Differenzialgleichung

$$y^{(3)}(x) + 2y^{(2)}(x) + y'(x) = 2$$

auf Seite 776 mit dem einfachen Ansatz der rechten Seite zu berechnen, dann scheitern wir: Wegen $q(x) = 2$ wählen wir $y_p(x) = c$ und $\alpha = \beta = 0$. Eingesetzt in die inhomogene Differenzialgleichung erhalten wir $0 = 2$. Das liegt daran, dass unser Ansatz gerade eine Lösung der homogenen Gleichung ist. Für diesen Fall benötigt man den folgenden erweiterten Ansatz.

5.4.2.2 Erweiterter Ansatz vom Typ der rechten Seite, Resonanzfall

Ist $s = \alpha + j\beta$ eine k -fache Nullstelle ($k \geq 1$) des charakteristischen Polynoms, so führt für Inhomogenitäten vom Typ

$$q(x) = p_m(x) \cdot e^{\alpha x} \cdot \cos(\beta x) \quad \text{oder} \quad q(x) = p_m(x) \cdot e^{\alpha x} \cdot \sin(\beta x),$$

mit $\alpha, \beta \in \mathbb{R}$ und einem reellen Polynom $p_m(x)$ vom Grad m , der einfache Ansatz (5.29) nicht zu Ziel. Denn mit ihm würde man für $m < k$ eine Lösung der homogenen Gleichung erhalten (vgl. Satz 5.9), für $m \geq k$ erhält man beim Einsetzen in L_n einen Polynomanteil vom Grad kleiner m . Hier spricht man von **Resonanz**. Man verwendet den **erweiterten Ansatz vom Typ der rechten Seite**

$$y_p(x) = x^k \cdot [r_m(x) \cdot e^{\alpha x} \cdot \cos(\beta x) + s_m(x) \cdot e^{\alpha x} \cdot \sin(\beta x)] \quad (5.30)$$

mit reellen Polynomen $r_m(x), s_m(x)$ vom Grad m , d. h., man multipliziert den einfachen Ansatz (5.29) mit x^k und benutzt so Polynome vom Grad $m+k$. Dadurch bleibt beim Einsetzen in die Differenzialgleichung eine Funktion vom Typ der rechten Seite übrig, und man kann wieder die Lösung mittels Koeffizientenvergleich berechnen.

Den Resonanzfall haben wir auch bei linearen Differenzialgleichungen erster Ordnung gesondert betrachtet, indem wir in Tabelle 5.1 auf Seite 731 eine entsprechende Fallunterscheidung passend zur Nullstelle a des charakteristischen Polynoms vorgenommen haben.

Beispiel 5.50

a) Wir berechnen die partikuläre Lösung der Differenzialgleichung

$$y^{(3)}(x) + 2y^{(2)}(x) + y'(x) = 2$$

aus Beispiel 5.47 mit dem erweiterten Ansatz viel schneller als mit Satz 5.10: Da $0 = \alpha + j\beta$ einfache Nullstelle des charakteristischen Polynoms $p(s) = -(s^3 + 2s^2 + s) = -s(s+1)^2$ ist, wählen wir statt $y_p(x) = c$ den Ansatz $y_p(x) = x^1 \cdot c$. Eingesetzt in die Differenzialgleichung erhalten wir $0 + 0 + c = 2$, also ist $y_p(x) = 2x$ eine partikuläre Lösung, die allgemeine Lösung (5.26) entsteht durch Addition homogener Lösungen.

b) Bei der inhomogenen Differenzialgleichung

$$y''(x) + 3y'(x) + 2y(x) = 2e^{-x}$$

ist $s = -1$ einfache Nullstelle ($k = 1$) des charakteristischen Polynoms. Die rechte Seite $2e^{-1 \cdot x}$, d. h. $\alpha = -1, \beta = 0$, macht daher den erweiterten Ansatz nötig:

$$y_p(x) = C \cdot x \cdot e^{-x}.$$

Setzt man $y_p'(x) = C e^{-x} - x C e^{-x}$ und $y_p''(x) = -C e^{-x} - C e^{-x} + x C e^{-x}$ in die Differenzialgleichung ein, so folgt

$$-2C e^{-x} + x C e^{-x} + 3C e^{-x} - 3x C e^{-x} + 2x C e^{-x} = 2e^{-x} \implies C = 2,$$

und eine partikuläre Lösung lautet $y_p(x) = 2x e^{-x}$. Die allgemeine Lösung der inhomogenen Differenzialgleichung ist

$$y(x) = C_1 e^{-2x} + C_2 e^{-x} + 2x e^{-x}, \quad C_1, C_2 \in \mathbb{R}. \quad \blacksquare$$

Beispiel 5.51

Bei der inhomogenen Differenzialgleichung

$$y''(x) + 3y'(x) + 2y(x) = x e^{-x}$$

hat die Inhomogenität $q(x) = x e^{-x}$ die Gestalt „Polynom vom Grad $m = 1$ multipliziert mit $e^{\alpha x} \cos(0 \cdot x)$ “, wobei $\alpha = -1$ einfache ($k = 1$) Nullstelle des charakteristischen Polynoms $s^2 + 3s + 2$ ist. Deshalb wählen wir den erweiterten Ansatz

$$y_p(x) = \underbrace{(C_0 + C_1 x)}_{r_1(x)} \cdot \underbrace{x}_{=x^k} \cdot e^{-x}.$$

Es gilt

$$\begin{aligned} y_p(x) &= [C_1 x^2 + C_0 x] e^{-x} \\ y_p'(x) &= (2C_1 x + C_0) e^{-x} - (C_1 x^2 + C_0 x) e^{-x} = [-C_1 x^2 + (2C_1 - C_0)x + C_0] e^{-x} \\ y_p''(x) &= [-2C_1 x + 2C_1 - C_0] e^{-x} - [-C_1 x^2 + (2C_1 - C_0)x + C_0] e^{-x} \\ &= [C_1 x^2 + (C_0 - 4C_1)x + 2C_1 - 2C_0] e^{-x}, \end{aligned}$$

und Einsetzen in die Differenzialgleichung mit Koeffizientenvergleich liefert

$$[0 \cdot x^2 + 2C_1 x + 2C_1 + C_0] e^{-x} = x e^{-x} \implies C_1 = \frac{1}{2}, \quad C_0 = -1.$$

Wir erhalten eine partikuläre Lösung $y_p(x) = \left(\frac{1}{2}x^2 - x\right) e^{-x}$. ■

5.4.2.3 Komplexer Ansatz vom Typ der rechten Seite

Verwendet man statt trigonometrischer Funktionen die komplexe Exponentialfunktion, so werden die Additionstheoreme zu Regeln der Potenzrechnung, und das Rechnen vereinfacht sich mitunter wesentlich. Wie bei der komplexen Wechselstromrechnung, bei der wir in Kapitel 1.6.5 Ströme und Spannungen komplex erweitert haben, kann man zur Bestimmung einer partikulären Lösung der Differenzialgleichung

$$L_n[y](x) = q_1(x)$$

die Inhomogenität $q_1(x)$ mit einer geeigneten Funktion $q_2(x)$ komplex erweitern:

$$L_n[y](x) = q_1(x) + j \cdot q_2(x).$$

Eine günstige komplexe Erweiterung von

$$L_n[y](x) = \cos(\omega x)$$

ist gegeben durch

$$L_n[y](x) = \cos(\omega x) + j \cdot \sin(\omega x) = e^{j\omega x}.$$

Hier sind alle Daten der linken Seite reell. Daher erhält man durch Bildung des Realteils einer komplexen Lösung (wie bei Differenzialgleichungssystemen) die eigentlich gesuchte Lösung des reellen Problems.

Beispiel 5.52

Wir bestimmen die Lösung der Differenzialgleichung $y'(x) + ay(x) = 4 \cos(2x)$. Die allgemeine Lösung der homogenen Differenzialgleichung lautet

$$y_h(x) = C \cdot e^{-a \cdot x}, \quad C \in \mathbb{R}.$$

Jetzt erweitern wir die Differenzialgleichung ins Komplexe durch

$$y'(x) + ay(x) = 4(\cos(2x) + j \cdot \sin(2x)) = 4e^{j \cdot 2x}.$$

Der (komplexe) Ansatz vom Typ der rechten Seite lautet nun $y_p^c(x) := A \cdot e^{2jx}$, und für die Ableitung gilt $(y_p^c)'(x) = 2jA e^{2jx}$. Einsetzen in die Differenzialgleichung liefert:

$$(2j + a) \cdot A e^{2jx} = 4e^{2jx} \quad \Longleftrightarrow \quad A = \frac{4}{2j + a} = \frac{4a - 8j}{a^2 + 4}.$$

Eine partikuläre Lösung lautet somit $y_p^c(x) = \frac{4a-8j}{a^2+4} \cdot e^{2jx}$. Der Realteil der komplexen Lösung ist eine partikuläre Lösung der reellen Differenzialgleichung:

$$y_p(x) = \operatorname{Re}(y_p^c(x)) = \frac{4a}{a^2 + 4} \cos(2x) + \frac{8}{a^2 + 4} \sin(2x).$$

Insgesamt erhalten wir die allgemeine Lösung

$$y(x) = C \cdot e^{-ax} + \frac{4a}{a^2 + 4} \cos(2x) + \frac{8}{a^2 + 4} \sin(2x), \quad C \in \mathbb{R}. \quad \blacksquare$$

5.4.3 Schwingungsgleichung*

In diesem Kapitel diskutieren wir als Anwendung der Differenzialgleichungen die mathematische Beschreibung von Schwingungen, wie sie häufig in der Maschinendynamik und Elektrotechnik auftreten.

Als mechanisches Beispiel betrachten wir wieder ein Schwingssystem, bestehend aus einem Massenpunkt $m > 0$, einer Feder (Federkonstante $c > 0$) und einem Dämpfungsglied (Dämpfungskonstante $b \geq 0$) wie auf Seite 708. Neu kommt eine periodische Anregung $F(t) = K_0 \cos(\omega t)$ hinzu. Die Funktion $x(t)$ beschreibt wieder die Auslenkung des Massenpunkts relativ zur Ruhelage und genügt der linearen Differenzialgleichung zweiter Ordnung

$$m x''(t) + b x'(t) + c x(t) = K_0 \cos(\omega t), \quad t \in [0, \infty[. \quad (5.31)$$

Durch zwei Anfangsbedingungen $x(0) = x_0$ und $x'(0) = x'_0$ ist eine eindeutige Lösung festgelegt.

Als weiteres Beispiel betrachten wir einen elektrischen Schwingkreis, bestehend aus der Kapazität C , dem Ohm'schen Widerstand R , der Induktivität L sowie einer Spannungsquelle mit einer periodischen Spannung $u(t) = u_0 \cos(\omega t)$.

Für den Kondensator gilt $i(t) = Cu'_C(t)$, für den Ohm'schen Widerstand $u_R(t) = Ri(t) = RCu'_C(t)$ und für die Spule $u_L(t) = Li'(t) = LCu''_C(t)$. Addieren wir alle Spannungen, erhalten wir $u(t) = u_C(t) + u_R(t) + u_L(t) = u_C(t) + RCu'_C(t) + LCu''_C(t)$, also

$$Lu''_C(t) + Ru'_C(t) + \frac{1}{C}u_C(t) = \frac{u_0}{C}\cos(\omega t), \quad t \in [0, \infty[. \quad (5.32)$$

Da die Gleichungen (5.31) und (5.32) formal übereinstimmen, diskutieren wir im Folgenden den Prototyp der Schwingungsgleichung

$$x''(t) + 2D\omega_0 x'(t) + \omega_0^2 x(t) = K_0 \cos(\omega t), \quad t \in [0, \infty[. \quad (5.33)$$

Die Konstante $D \geq 0$ heißt das **Lehr'sche Dämpfungsmaß**, die Konstante $\omega_0 \geq 0$ wird als Eigenfrequenz der ungedämpften Schwingung bezeichnet. Die Bedeutung beider Größen wird weiter unten verständlich. Im mechanischen Beispiel ist $\omega_0 = \sqrt{\frac{c}{m}} \geq 0$ und $2D\omega_0 = \frac{b}{m}$, so dass $D = \frac{b}{2m\omega_0} = \frac{b}{2\sqrt{cm}} \geq 0$. Beim elektrischen Schwingkreis ist $\omega_0 = \sqrt{\frac{1}{LC}} \geq 0$ und $2D\omega_0 = \frac{R}{L}$, also $D = \frac{R}{2L\omega_0} = \frac{R\sqrt{LC}}{2L} = \frac{R}{2}\sqrt{\frac{C}{L}} \geq 0$.

5.4.3.1 Die homogene Schwingungsgleichung

Betrachten wir zunächst die homogene Gleichung

$$x''(t) + 2D\omega_0 x'(t) + \omega_0^2 x(t) = 0, \quad t \in [0, \infty[. \quad (5.34)$$

Die Lösung dieser Gleichung haben wir mit (5.25) bereits berechnet. Die Gleichung hat das charakteristische Polynom $p(s) = s^2 + 2D\omega_0 s + \omega_0^2$ mit den Nullstellen

$$s = -D\omega_0 \pm \omega_0 \sqrt{D^2 - 1}. \quad (5.35)$$

Wir unterscheiden drei Fälle:

a) Schwingfall: Dies ist der Fall, dass $0 \leq D < 1$ ist. Damit sind die Nullstellen der charakteristischen Gleichung $s = -D\omega_0 \pm j\omega_0 \sqrt{1 - D^2}$ zueinander konjugiert komplex. Mit der Bezeichnung

$$\omega_D := \omega_0 \sqrt{1 - D^2} \quad (5.36)$$

erhält man das (komplexe) Fundamentalsystem $x_{\pm}(t) = e^{-D\omega_0 t} \cdot e^{\pm j\omega_D t}$, und durch Übergang zu Real- bzw. Imaginärteil entsteht hieraus das reelle Fundamentalsystem

$$x_1(t) = e^{-D\omega_0 t} \cos(\omega_D t), \quad x_2(t) = e^{-D\omega_0 t} \sin(\omega_D t).$$

Diese Lösungen stellen für $D > 0$ gedämpfte harmonische Schwingungen dar, deren Amplitude $e^{-D\omega_0 t}$ für $t \rightarrow \infty$ gegen 0 strebt, siehe Abbildung 5.11. Jetzt wird verständlich, dass D als Dämpfungsmaß bezeichnet wird, da D das Abfallen der Amplitude bestimmt. Die Kreisfrequenz der Kosinus- und Sinus-Terme ist ω_D . Daher nennt man ω_D die **Eigenfrequenz** der Schwingung. Ist das Dämpfungsmaß $D = 0$, so ist $e^{-D\omega_0 t} = 1$, und wir erhalten ungedämpfte harmonische Schwingungen

$$x_1(t) = \cos(\omega_0 t), \quad x_2(t) = \sin(\omega_0 t).$$

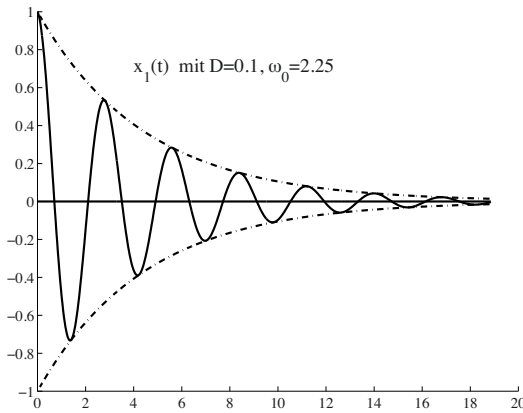


Abb. 5.11: Gedämpfte Schwingung
 $x_1(t) = e^{-D\omega_0 t} \cos(\omega_D t)$

Hier ist außerdem $\omega_D = \omega_0$, so dass die eingangs gemachte Bezeichnung von ω_0 als Eigenfrequenz der ungedämpften Schwingung verständlich wird.

b) Kriechfall: Im Fall $D > 1$ besitzt die charakteristische Gleichung zwei reelle Nullstellen

$$s = -D\omega_0 \pm \omega_0 \sqrt{D^2 - 1} < 0 \quad (\omega_0 \neq 0),$$

und man erhält das reelle Fundamentalsystem

$$x_{1,2}(t) = e^{(-D\omega_0 \pm \omega_0 \sqrt{D^2 - 1}) t}.$$

Wegen $s < 0$ klingen beide Lösungen $x(t)$ für große t exponentiell ab (siehe Abbildung 5.12). Dies ist der Fall, den wir beim Feder-Masse-Dämpfersystem auf Seite 774 berechnet haben.

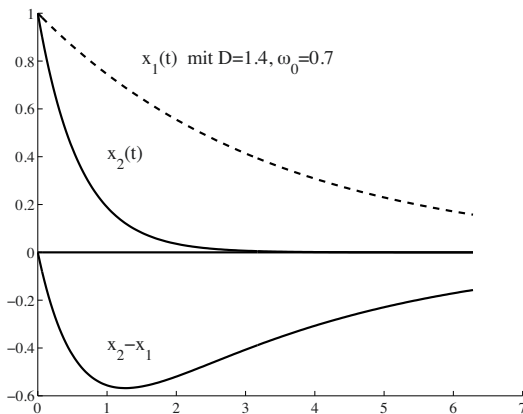


Abb. 5.12: Fundamentalsystem im Kriechfall

c) Aperiodischer Grenzfall: Für $D = 1$ erhält man aufgrund der doppelten Nullstelle $s = -D\omega_0 = -\omega_0$ das Fundamentalsystem

$$x_1(t) = e^{-\omega_0 t} \quad \text{und} \quad x_2(t) = te^{-\omega_0 t},$$

und somit qualitativ dasselbe Verhalten wie im Kriechfall, denn auch hier streben für $t \rightarrow \infty$ sowohl $x_1(t)$ als auch $x_2(t)$ gegen 0.

In allen drei Fällen ist die allgemeine Lösung gegeben durch

$$x(t) = c_1 x_1(t) + c_2 x_2(t) \quad \text{mit} \quad c_1, c_2 \in \mathbb{R}, \quad (5.37)$$

wobei $x_{1,2}(t)$ jeweils die unter a), b) und c) berechneten Funktionen sind.

5.4.3.2 Die inhomogene Schwingungsgleichung

Wir betrachten nun die inhomogene Gleichung mit periodischer Erregung

$$x''(t) + 2D\omega_0 x'(t) + \omega_0^2 x(t) = K_0 \cos(\omega t), \quad t \in [0, \infty[, \quad \omega > 0 \quad (5.38)$$

bzw. deren komplexe Form

$$x''(t) + 2D\omega_0 x'(t) + \omega_0^2 x(t) = K_0 e^{j\omega t}, \quad (5.39)$$

mit der leichter zu rechnen ist. Wie in Kapitel 5.4.2.3 erhalten wir aus der Lösung $x(t)$ von (5.39) durch Bildung des Realteils eine Lösung von (5.38). Wir unterscheiden nun zwei Fälle:

a) Normalfall: Als Normalfall bezeichnen wir die Situation, dass $e^{j\omega t}$ nicht Lösung der homogenen Gleichung ist.

- Dann können ω und ω_0 nicht gemeinsam gleich null sein. Denn für $\omega_0 = 0$ ist jede Konstante eine Lösung der homogenen Gleichung, und bei $\omega = 0$ ist auch die Inhomogenität $K_0 e^{j\omega t}$ eine Konstante.
- Zudem müssen im Fall $D = 0$ die Kreisfrequenzen ω und ω_0 verschieden sein (siehe homogene Lösung im Schwingfall).

Hier liefert der Ansatz vom Typ der rechten Seite $x(t) = C e^{j\omega t}$ nach Einsetzen in (5.39) und Kürzen durch $e^{j\omega t}$

$$(-\omega^2 + j2D\omega_0\omega + \omega_0^2)C = K_0,$$

d. h. mit konjugiert komplexer Erweiterung:

$$C = \frac{K_0}{(\omega_0^2 - \omega^2)^2 + 4D^2\omega_0^2\omega^2} [(\omega_0^2 - \omega^2) - j2D\omega_0\omega].$$

Der Nenner $N := (\omega_0^2 - \omega^2)^2 + 4D^2\omega_0^2\omega^2$ ist nach den Vorüberlegungen größer als null:

- Falls $D \neq 0$ ist, kann der Term $(\omega_0^2 - \omega^2)^2$ nur null sein, wenn $0 \neq \omega_0 = \omega$ gilt. Dann ist aber $4D^2\omega_0^2\omega^2 > 0$.
- Falls $D = 0$ ist, gilt für die verschiedenen, nicht-negativen Kreisfrequenzen ω und ω_0 auch $\omega_0^2 \neq \omega^2$, so dass $N = (\omega_0^2 - \omega^2)^2 > 0$ ist.

Damit erhalten wir über

$$x(t) = C(\cos(\omega t) + j \sin(\omega t)) = \frac{K_0[(\omega_0^2 - \omega^2) \cos(\omega t) + 2D\omega_0\omega \sin(\omega t)]}{N} + j \dots$$

eine (reelle) partikuläre Lösung von (5.38) mit zu

$$x_p(t) = \frac{K_0}{\sqrt{N}} \left[\frac{(\omega_0^2 - \omega^2)}{\sqrt{N}} \cos(\omega t) + \frac{2D\omega_0\omega}{\sqrt{N}} \sin(\omega t) \right].$$

Wir fassen nun die Terme in der Klammer zusammen, indem wir $\omega_0^2 - \omega^2$ als Ankathete und $2D\omega_0\omega$ als Gegenkathete in einem rechtwinkligen Dreieck mit Hypotenuse $\sqrt{N}$ auffassen und so die Faktoren $\frac{\omega_0^2 - \omega^2}{\sqrt{N}} = \cos(\varphi(\omega))$ und $\frac{2D\omega_0\omega}{\sqrt{N}} = \sin(\varphi(\omega))$ für einen Winkel $\varphi(\omega)$ schreiben. Mit dem Additionstheorem (1.27)

$$\cos(\omega t - \varphi(\omega)) = \cos(\varphi(\omega)) \cos(\omega t) + \sin(\varphi(\omega)) \sin(\omega t)$$

erhalten wir jetzt die Darstellung

$$x_p(t) = A(\omega) \cos(\omega t - \varphi(\omega))$$

mit der **Antwortamplitude**

$$A(\omega) := \frac{K_0}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4D^2\omega_0^2\omega^2}}. \quad (5.40)$$

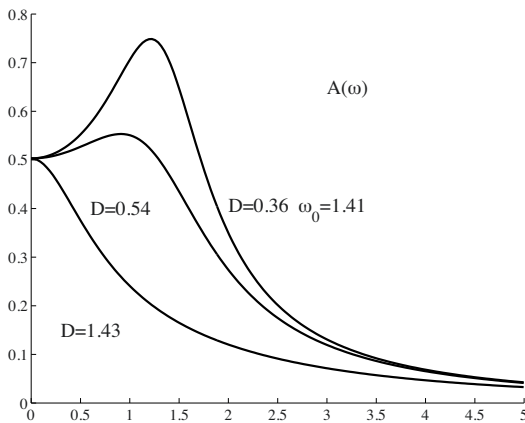
Das System schwingt mit der Kreisfrequenz ω des Erregers; es reagiert phasenverzögert um $\varphi(\omega)$ mit der Amplitude $A(\omega)$ (vgl. Abbildung 5.13). Insgesamt folgt für die allgemeine Lösung der inhomogenen Differenzialgleichung im Schwingfall $0 \leq D < 1$

$$\begin{aligned} x(t) &= x_h(t) + x_p(t) \\ &= \underbrace{e^{-D\omega_0 t} \cdot [C_1 \sin(\omega_D t) + C_2 \cos(\omega_D t)]}_{\text{zeitlich abklingende Schwingung}} + \underbrace{A(\omega) \cdot \sin(\omega t - \varphi(\omega))}_{\text{permanente Schwingung}}. \end{aligned}$$

Die Gesamtlösung setzt sich zusammen aus einer **transienten** (d. h. zeitlichen abklingenden) Schwingung der Kreisfrequenz ω_D , die von der homogenen Lösung x_h her stammt, und einer **permanenten** oder **stationären** Schwingung der Kreisfrequenz ω , der partikulären Lösung.

Wir berechnen nun die Kreisfrequenz ω_r , die zur maximalen Amplitude der permanenten Schwingung führt (vgl. Abbildung 5.13). Dazu benutzen wir die notwendige Bedingung, dass die Ableitung von $A(\omega)$ an der Stelle ω_r null sein muss. Es ist

$$0 = \frac{dA}{d\omega}(\omega_r) = -\frac{K_0}{2\sqrt{N}^3} (2(\omega_0^2 - \omega_r^2)(-2\omega_r) + 8D^2\omega_0^2\omega_r)$$

Abb. 5.13: Amplituden $A(\omega)$

genau dann, wenn $\omega_r = 0$ oder $\omega_r^2 = \omega_0^2(1 - 2D^2)$ ist.

Im Fall $D \geq \frac{1}{\sqrt{2}}$ (d. h. große Dämpfung) haben wir außer $\omega_r = 0$ keine weitere Extremalstelle. Für $0 < D < \frac{1}{\sqrt{2}}$ hingegen erhalten wir mit

$$\omega_r = \omega_0 \sqrt{1 - 2D^2}$$

die Anregungsfrequenz mit **maximaler Systemantwort**

$$A(\omega_r) = \frac{K_0}{\sqrt{(\omega_0^2 - \omega_0^2[1 - 2D^2])^2 + 4D^2\omega_0^2\omega_0^2[1 - 2D^2]}} = \frac{K_0}{2D\omega_0^2\sqrt{1 - D^2}}.$$

Bei Anregung eines gedämpften Systems mit ω_r können mitunter Amplituden erzeugt werden, die die Belastungsgrenze des Systems überschreiten, so dass Schäden auftreten. Dies bezeichnet man als (technische) **Resonanzkatastrophe**. Dieses Phänomen ist bei der Konstruktion von Bauwerken (wie z. B. Brücken und Hochhäuser) und Fahrzeugen (z. B. badisch-schwäbischer Kleinwagen) unbedingt zu berücksichtigen. Noch dramatischer macht sich dieser Effekt bei ungedämpften Schwingungen bemerkbar. Für $D = 0$ hätten wir $\omega_r = \omega_0$. Da aber $\omega = \omega_0$ im Normalfall ausgeschlossen ist, müssen wir dazu einen zweiten Fall betrachten.

b) Resonanzfall: Der Resonanzfall liegt vor, wenn $e^{j\omega t}$ Lösung der homogenen Gleichung ist. Aufgrund der berechneten homogenen Lösungen kann der Fall nur eintreten, wenn sowohl $D = 0$ (keine Dämpfung) als auch $\omega = \omega_0$ ist. Hier wird die ungedämpfte Schwingung mit ihrer Eigenfrequenz ω_0 angeregt. Für $\omega_0 \neq 0$ führt der modifizierte Ansatz vom Typ der rechten Seite

$$x(t) = Cte^{j\omega_0 t}$$

auf $x'(t) = (1 + j\omega_0 t)Ce^{j\omega_0 t}$, $x''(t) = (-\omega_0^2 t + 2j\omega_0)Ce^{j\omega_0 t}$, und Einsetzen in die Differenzialgleichung $x''(t) + \omega_0^2 x(t) = K_0 e^{j\omega_0 t}$ (beachte $D = 0$) liefert

$$(-\omega_0^2 t + 2j\omega_0 + \omega_0^2 t)C = K_0.$$

Damit erhalten wir $C = -\frac{K_0}{2\omega_0}j$ und nach Realteilbildung schließlich

$$x_p(t) = \frac{K_0}{2\omega_0}t \sin(\omega_0 t),$$

eine Schwingung, deren Amplitude $\frac{K_0}{2\omega_0}t$ für $t \rightarrow \infty$ über alle Grenzen wächst (siehe Abbildung 5.14). Insgesamt lautet die allgemeine Lösung mit dem Fundamentalsystem

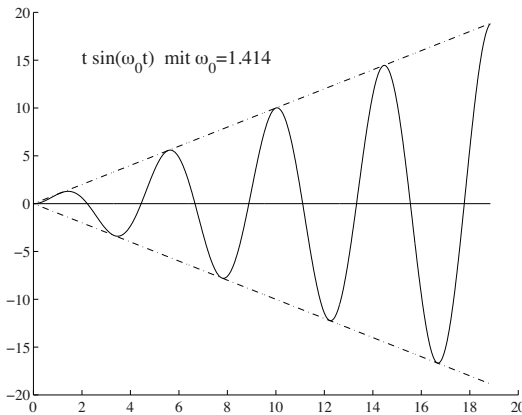


Abb. 5.14: Lösung im Resonanzfall

des homogenen Schwingfalls (da $\omega_D = \omega_0$ wegen $D = 0$)

$$x(t) = C_1 \cos(\omega_0 t) + C_2 \sin(\omega_0 t) + \frac{K_0}{2\omega_0}t \sin(\omega_0 t).$$

5.5 Ausblick: Partielle Differenzialgleichungen und Finite-Elemente-Methode*

5.5.1 Eine schwingende Saite: Wellengleichung

Wir betrachten eine Saite, die in einer Ebene schwingen soll. Zu einem Zeitpunkt t beschreibe $u(x, t)$ die Auslenkung der Saite an der Stelle x . Die Funktion u ist Lösung einer **partiellen Differenzialgleichung**. Im Gegensatz zu den zuvor diskutierten gewöhnlichen Differenzialgleichungen ist $u(x, t)$ eine Funktion von zwei Variablen, und es kommen partielle Ableitungen vor:

$$\frac{\partial^2 u}{\partial t^2}(x, t) = c \frac{\partial^2 u}{\partial x^2}(x, t).$$

Dabei ist $c > 0$ eine Konstante (Spannung dividiert durch Massendichte der Saite). Diese partielle Differenzialgleichung ist unter dem Namen **Wellengleichung** sehr bekannt.

An den Stellen $x = 0$ und $x = \pi$ sei die Saite eingespannt, d. h. $u(0, t) = u(\pi, t) = 0$ für alle Zeitpunkte t . Dies ist eine **Randbedingung**, da im Gegensatz zu den Anfangsbedingungen nun Funktionswerte an verschiedenen Stellen vorgegeben werden. Statt der Stelle π kann man auch einen anderen Punkt wählen, dann müssten wir aber eine zusätzliche Konstante berücksichtigen.

Um Lösungen zu finden, machen wir den Produktansatz

$$u(x, t) = v(x)w(t).$$

Jetzt suchen wir nur Lösungen, die diese Produktgestalt haben, andere Lösungen werden wir mit diesem Ansatz nicht finden. Die Auftrennung der Lösung nach Variablen ist ein etabliertes Lösungsverfahren für partielle Differenzialgleichungen.

Für jedes $k \in \mathbb{N}$ erfüllt $v(x) = \sin(kx)$ die Randbedingung. Denn da die Saite an den Enden befestigt ist, muss sie zu einem Zeitpunkt wie in Abbildung 5.15 aussehen. Mit einer solchen Wahl für v wird die Differenzialgleichung zu

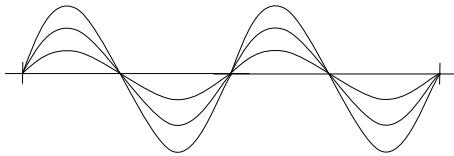


Abb. 5.15: Schwingende Saite $u(x, t)$ zu drei Zeitpunkten t

$$w''(t) \sin(kx) = cw(t)(-k^2 \sin(kx)),$$

wir suchen also eine Funktion w mit

$$w''(t) + ck^2 w(t) = 0.$$

Das ist aber die im vorangehenden Kapitel diskutierte homogene Schwingungsgleichung (5.34) mit $\omega_0 = \sqrt{ck}$ im dämpfungsfreien Fall ($D = 0$) mit Fundamentalsystem $\cos(\sqrt{ckt})$, $\sin(\sqrt{ckt})$. Damit haben wir für jedes $k \in \mathbb{N}$ die Lösungen

$$u_{k,1}(x, t) = \sin(kx) \cos(\sqrt{ckt}), \quad u_{k,2}(x, t) = \sin(kx) \sin(\sqrt{ckt})$$

unseres Randwertproblems gefunden. Auch Linearkombinationen der Lösungen sind wieder Lösungen, und selbst eine Reihe

$$\sum_{k=1}^{\infty} \sin(kx) (a_k \cos(\sqrt{ckt}) + b_k \sin(\sqrt{ckt}))$$

ist eine Lösung, sofern die Koeffizienten a_k und b_k geeignet gewählt sind, um die Konvergenz sicherzustellen. Welche dieser vielen Lösungen nun tatsächlich die Schwingung der Saite beschreibt, hängt von den vorgegebenen Auslenkungen zum Startzeitpunkt $t = 0$ ab. An einer festen Stelle $x = x_0$ beobachten wir dann zum Zeitpunkt t die Auslenkung $u(x_0, t) = f(\sqrt{ct})$ mit

$$f(t) = \sum_{k=1}^{\infty} (\tilde{a}_k \cos(kt) + \tilde{b}_k \sin(kt)) \quad (5.41)$$

und $\tilde{a}_k := \sin(kx_0)a_k$ und $\tilde{b}_k := \sin(kx_0)b_k$. Mit solchen Summen von Sinus- und Kosinusfunktionen beschäftigen wir uns in Kapitel 6 intensiv.

5.5.2 Partielle Differenzialgleichungen zweiter Ordnung

In Beispiel 4.8 auf Seite 602 haben wir schon die Lösungen einer partiellen Differenzialgleichung erster Ordnung diskutiert. Auch bei der Differenzialform der Maxwell'schen Gleichungen (siehe Kasten auf Seite 699) handelt es sich um partielle Differenzialgleichungen erster Ordnung. Dagegen sind die Laplace- und die Poisson-Gleichung (siehe Seite 685 und Kasten auf Seite 696) wie auch die im vorangehenden Kapitel diskutierte Wellengleichung Gleichungen zweiter Ordnung. Auf solche Gleichungen mit höchstens zweiten partiellen Ableitungen stößt man besonders häufig:

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

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

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 (siehe Seite 611), 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 tragen jetzt die Koeffizienten der zweiten Ableitungen in eine Matrix ein:

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

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

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

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

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.

- Falls $\det \mathbf{A} < 0$ ist ($ac < \frac{b^2}{4}$), so heißt die Gleichung **hyperbolisch**. Eine solche Gleichung kann man in die folgende Form ohne gemischte Ableitungen bringen:

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

Der Term mit den zweiten Ableitungen erinnert jetzt an die linke Seite der Gleichung $x^2 - y^2 = 1$, die eine Hyperbel als Lösungsmenge hat.

- Falls $\det \mathbf{A} = 0$ ist ($ac = \frac{b^2}{4}$), so heißt die Gleichung **parabolisch**, und man erhält eine Darstellung mit nur einer partiellen Ableitung zweiter Ordnung:

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

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

Bei der (zweidimensionalen) Poisson-Gleichung (siehe Seiten 685 und 696) ist $a = c = 1$ und $b = 0$. Damit ist $\det \mathbf{A} > 0$, es handelt sich um eine elliptische Gleichung. Elliptische Differenzialgleichungen werden in der Regel 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 Wellengleichung (siehe Kapitel 5.5.1) 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. Eine hyperbolische Differenzialgleichung beschreibt in der Regel ein zeitabhängiges Phänomen und besitzt üblicherweise 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.

In Spezialfällen lassen sich partielle Differenzialgleichungen exakt lösen. So verwenden wir beispielsweise für die Wellengleichung einen Separationsansatz, bei dem wir von einer zerfallenden Lösung ausgehen, die das Produkt von Funktionen nur einer Variablen ist. Eine andere Lösungstechnik ist die Methode der Charakteristiken, siehe Beispiel 4.8 auf Seite 602. Weitere Techniken, um partielle in gewöhnliche Differenzialgleichungen zu überführen, sind Substitution der Variablen und Integraltransforma-

tionen (siehe Kapitel 6 für gewöhnliche Differenzialgleichungen). Ein anderer Lösungsansatz besteht in der Überführung in Integralgleichungen. Umformulierungen zu Optimierungsproblemen (Variationsaufgaben) sind sowohl Ausgangspunkt für exakte als auch für numerische Lösungsansätze.

In der Regel werden partielle Differenzialgleichungen nicht exakt gelöst. Der einfachste Ansatz zur numerischen (näherungsweisen) Lösung besteht im Ersetzen der Ableitungen durch Differenzenquotienten. Zusammen mit Anfangs- oder Randbedingungen entsteht so ein Gleichungssystem für Funktionswerte der gesuchten Lösungsfunktion. Bei elliptischen Gleichungen hat sich die Finite-Elemente-Methode etabliert, mit der unterschiedliche Schrittweiten leicht zu realisieren sind und das sich damit auch an schwieriger berandete Definitionsgebiete der Gleichungen anpassen lässt. Näheres dazu finden Sie in den folgenden Kapiteln 5.5.3 und 5.5.4. Partielle Differenzialgleichungen, die auf physikalischen Erhaltungssätzen basieren, werden häufig mittels Finite-Volumen-Verfahren gelöst. Die Erhaltungssätze werden dann auf kleine Volumenelemente angewendet, wobei z. B. Zu- und Abfluss durch die Hülle des Elements betrachtet werden. Die Randelemente-Methode basiert auf einer Integralgleichungsformulierung und verwendet Ansatzfunktionen wie die Finite-Elemente-Methode.

5.5.3 Finite-Elemente-Methode

Die naheliegendsten Verfahren zum Lösen gewöhnlicher und partieller Differenzialgleichungen sind die Differenzenverfahren, bei denen Ableitungen durch Differenzenquotienten ersetzt werden. Bei komplizierten Randbedingungen ist aber das unmittelbare Aufstellen von Differenzenschemata schwierig. Hier hilft der Finite-Elemente-Ansatz. Wir beginnen den Abschnitt mit einer gewöhnlichen Randwertaufgabe, für die wir uns eine Differenzen- und Finite-Elemente-Diskretisierung ansehen. Dann gehen wir kurz auf funktionalanalytische Grundlagen für die Finite-Elemente-Methode ein. Schließlich betrachten wir im nächsten Abschnitt die Finite-Elemente-Lösung einer (immer noch einfachen, aber anwendungsnäheren) partiellen Differenzialgleichung mit zwei Variablen.

Gesucht ist eine Funktion $y : [0, 1] \rightarrow \mathbb{R}$ 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.

Zur näherungsweisen Bestimmung der Funktionswerte von y an endlich vielen Stellen des Intervalls ersetzen wir die Ableitungen durch Differenzenquotienten. Wir möchten

dabei y an den $n+1$ Stellen $t_k = \frac{k}{n}$, $k \in \{0, 1, \dots, n\}$ berechnen. Für $k \in \{1, \dots, n-1\}$ 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 lineares 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 durch Differenzenquotienten ersetzt, heißen **Differenzenverfahren**. Die Euler-Cauchy-Polygonzugmethode ist beispielsweise zugleich ein Differenzenverfahren:

$$\frac{y_{k+1} - y_k}{h} = f(x_k, y_k).$$

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 dann wie bei Differenzenverfahren nur die Lösung eines linearen Gleichungssystems gefunden werden. Das Gleichungssystem kann sogar als zu einem Differenzenverfahren gehörend verstanden werden. Allerdings ergibt es sich trotz eines eventuell komplizierten Rands automatisch.

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

$$\begin{aligned} y''(x) = f(x) &\implies y''(x)v(x) = f(x)v(x) \implies \int_0^1 y''(x)v(x) dx = \int_0^1 f(x)v(x) dx \\ &\implies [y'(x)v(x)]_0^1 - \int_0^1 y'(x)v'(x) dx = \int_0^1 f(x)v(x) dx \\ &\implies \int_0^1 y'(x)v'(x) dx = - \int_0^1 f(x)v(x) dx. \end{aligned} \tag{5.42}$$

Erfüllt y die Differenzialgleichung, so erfüllt y auch das „**schwache Problem**“ (5.42) für jede Testfunktion v . Wir zerlegen nun das Intervall $[0, 1]$ wieder in n Teilintervalle $[\frac{k-1}{n}, \frac{k}{n}]$, $k \in \{1, 2, \dots, n\}$ (auch unterschiedlich große Teilintervalle wären erlaubt) 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 $[\frac{k-1}{n}, \frac{k}{n}]$, so gilt:

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

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 Matrix des Gleichungssystems heißt **Steifigkeitsmatrix**. 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 Teilmengen führt zum Namen **Finite Elemente**, siehe Kapitel 5.5.4 und insbesondere Seite 808.

Ein großer Vorteil der schwachen Formulierung (5.42) ist, dass geringere Anforderungen an mögliche Lösungen gestellt werden als bei einer Differenzialgleichung (hier einmalige Differenzierbarkeit statt zweimaliger). Damit können schwache Probleme auch dann noch Lösungen besitzen, wenn die Differenzialgleichung streng genommen gar keine hat. Die Anforderungen lassen sich sogar noch weiter reduzieren, indem man schwache Ableitungen verwendet und Lösungen in Sobolev-Räumen (siehe Kapitel 4.4.5.2 für die jetzt verwendeten Notationen) sucht:

Im Beispiel kann zu einer Inhomogenität $f \in L^2(\Omega)$, $\Omega := (0, 1)$, eine Funktion $y \in W_0^{1,2}(\Omega)$ gesucht werden, so dass für alle Funktionen $v \in W_0^{1,2}(\Omega)$ gilt:

$$(y, v)_{1,2,\Omega} = - \int_0^1 f(x)v(x) dx.$$

Jetzt definiert die rechte Seite ein beschränktes lineares Funktional $f^*(v) := - \int_0^1 f(x)v(x) dx$ auf dem Hilbert-Raum $W_0^{1,2}(\Omega)$. Die Beschränktheit folgt mit der Hölder-Ungleichung (3.32 auf Seite 564) und der Poincaré-Ungleichung (Lemma 4.3 auf Seite 672):

$$\begin{aligned} \left| - \int_0^1 f(x)v(x) dx \right| &\leq \int_0^1 |f(x)v(x)| dx \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq \|f\|_{L^2(\Omega)} \|v\|_{1,2,\Omega} \leq \frac{\|f\|_{L^2(\Omega)}}{c} \sqrt{(v, v)_{1,2,\Omega}}. \end{aligned}$$

Hier kommt der Satz von Riesz ins Spiel (Satz 3.47, Seite 584). Er besagt, dass das Funktional f^* darstellbar ist als $(y, v)_{1,2,\Omega}$ für ein eindeutig bestimmtes Element $y \in W_0^{1,2}(\Omega)$. Das bedeutet, dass es eine eindeutige Lösung y des schwachen Problems gibt.

Die Idee der Finite-Elemente-Methode ist nun, das schwache Problem über einem endlich-dimensionalen Hilbert-Raum V_n , der ein Untervektorraum von $W_0^{1,2}(\Omega)$ ist, zu lösen. Ein entsprechender Algorithmus heißt **Ritz-Galerkin-Verfahren**. Dabei ist also eine Näherungslösung $y_n \in V_n$ des ursprünglichen schwachen Problems als exakte Lösung der neuen Aufgabe

$$(y_n, v_n)_{1,2,\Omega} = - \int_0^1 f(x)v_n(x) dx \text{ für alle } v_n \in V_n$$

zu berechnen. Auch hier gibt es nach dem Satz von Riesz eine eindeutige Lösung, die letztlich über ein Gleichungssystem gefunden werden kann. Wählt man V_n als Vektorraum der oben verwendeten Ansatzfunktionen, die stückweise Geraden entsprechen, dann entsteht auch das gleiche Gleichungssystem bzw. die Steifigkeitsmatrix wie oben.

Wir haben bei diesem Beispiel einen sehr einfachen Fall betrachtet. Im Allgemeinen führt die Differenzialgleichung nicht direkt auf das Skalarprodukt eines Sobolev-Raums. Aber auch dann erhält man oft Existenz und Eindeutigkeit von Lösungen mit einer erweiterten Fassung des Satzes von Riesz, dem Lemma von Lax und Milgram. Darüber hinaus kann man den Fehler zwischen Näherungslösung und exakter Lösung sehr gut abschätzen, das ist die Aussage des Lemmas von Céa, siehe Kasten.

Hintergrund: Funktionalanalytische Grundlagen der Finite-Elemente-Methode

Im Folgenden werden mit V und H reelle Hilbert-Räume bezeichnet. Die zugehörigen Skalarprodukte seien $(\cdot, \cdot)_V$ und $(\cdot, \cdot)_H$. Im zuvor betrachteten Beispiel ist $V = W_0^{1,2}(0, 1)$ und $H = L^2(0, 1)$.

Definition 5.8 (Bilinearform)

Sei $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ eine Bilinearform, d. h., für alle $r \in \mathbb{R}$ und $\vec{x}, \vec{y}, \vec{z} \in V$ gilt:

$$\begin{aligned} a(\vec{x} + \vec{y}, \vec{z}) &= a(\vec{x}, \vec{z}) + a(\vec{y}, \vec{z}), \\ a(\vec{x}, \vec{y} + \vec{z}) &= a(\vec{x}, \vec{y}) + a(\vec{x}, \vec{z}), \\ a(r\vec{x}, \vec{y}) &= ra(\vec{x}, \vec{y}) = a(\vec{x}, r\vec{y}). \end{aligned}$$

Die Bilinearform a heißt

- a) symmetrisch** genau dann, wenn $a(\vec{x}, \vec{y}) = a(\vec{y}, \vec{x})$ für alle $\vec{x}, \vec{y} \in V$,
b) beschränkt genau dann, wenn ein $C \in \mathbb{R}$ existiert mit

$$|a(\vec{x}, \vec{y})| \leq C \|\vec{x}\|_V \|\vec{y}\|_V \text{ für alle } \vec{x}, \vec{y} \in V, \quad (5.43)$$

- c) V -elliptisch (koerziv)** genau dann, wenn ein $c > 0$ existiert mit

$$a(\vec{x}, \vec{x}) \geq c \|\vec{x}\|_V^2 \text{ für alle } \vec{x} \in V. \quad (5.44)$$

Eine beschränkte elliptische Bilinearform a wird zu einem Skalarprodukt, wenn a zusätzlich symmetrisch ist. Die Symmetrie ist aber häufig nicht gegeben, wenn die Bilinearform aus einer partiellen Differenzialgleichung entsteht. Die Bezeichnung V -elliptisch ist nicht zufällig gewählt, da solche Bilinearformen in der Praxis aus elliptischen Differenzialgleichungen (siehe Seite 791) hervorgehen. Eine zugehörige Randwertaufgabe kann z. B. durch partielle Integration in das folgende Variationsproblem überführt werden, das auch schwaches Problem genannt wird:

Gegeben ist eine beschränkte und V -elliptische Bilinearform a und ein lineares beschränktes Funktional f^* auf V , d. h., f^* ist Element des Dualraums von V , also $f^* \in V^*$. Gesucht ist ein $\vec{x} = \vec{x}_{f^*} \in V$, so dass

$$a(\vec{x}, \vec{y}) = f^*(\vec{y}) \text{ für alle } \vec{y} \in V. \quad (5.45)$$

Im Gegensatz zur Differenzialgleichung ist aber hier die Frage nach der Lösbarkeit leicht zu beantworten. Für Skalarprodukte haben wir die Existenz und Eindeutigkeit der Lösung bereits mit dem Riesz'schen Darstellungssatz erhalten. Allgemeiner gilt (Alt, 1992, S.118):

Satz 5.11 (Lemma von Lax und Milgram)

Sei a eine beschränkte V -elliptische Bilinearform. Dann existiert zu jedem $f^* \in V^*$ genau eine Lösung $\vec{x}$ von (5.45).

Das Lemma ist tatsächlich eine Folgerung aus dem Riesz-Darstellungssatz für Hilbert-Räume. Die Lösung von (5.45) ist zugleich auch die eindeutige Lösung der Minimierungsaufgabe, bei der ein $\vec{x} \in V$ mit $J(\vec{x}) = \inf_{\vec{x} \in V} J(\vec{x})$ gesucht ist. Dabei ist $J(\vec{x}) := a(\vec{x}, \vec{x})/2 - f^*(\vec{x})$, siehe z. B. (Ciarlet, 1990, S. 25).

Im Allgemeinen wird V unendlich-dimensional sein, so dass es schwierig ist, die Lösung von (5.45) explizit zu berechnen. Zur näherungsweisen Lösung von (5.45) betrachten wir wie im vorangehenden eindimensionalen Beispiel einen endlich-dimensionalen Untervektorraum $S_h \subset V$ und lösen die folgende Diskretisierung von (5.45):

Gegeben seien die beschränkte und V -elliptische Bilinearform a und die Inhomogenität $f^* \in V^*$ der Aufgabe (5.45). Gesucht ist ein $\vec{x}_h = \vec{x}_{h,f^*} \in S_h$, so dass

$$a(\vec{x}_h, \vec{y}) = f^*(\vec{y}) \text{ für alle } \vec{y} \in S_h. \quad (5.46)$$

Der Untervektorraum $S_h \subset V$ ist selbst wieder ein Hilbert-Raum mit dem Skalarprodukt $(\cdot, \cdot)_V$, so dass a auch S_h -elliptisch ist. Da außerdem Funktionale auf V auch Funktionale auf S_h sind (d. h., für die Dualräume gilt $V^* \subset V_h^*$), ist $f^* \in S_h^*$, d. h., f^* ist insbesondere ein beschränktes lineares Funktional auf S_h . Ersetzen wir also in Satz 5.11 den Raum V durch V_h , so folgt, dass auch das diskrete Variationsproblem (5.46) eine eindeutige Lösung besitzt.

Wir ändern nun unseren Blickwinkel. Jedes $\vec{x} \in V$ ist trivialerweise Lösung einer Aufgabe (5.45) mit der Inhomogenität $f^*(\cdot) := a(\vec{x}, \cdot)$. Über die zugehörige diskrete Lösung $\vec{x}_h$ von (5.46) definieren wir die **Ritz-Projektion** $P_h : V \rightarrow V_h$ vermöge $P_h \vec{x} = \vec{x}_h$, also als eindeutige Lösung $P_h \vec{x} \in S_h$ von

$$a(P_h \vec{x}, \vec{y}) = a(\vec{x}, \vec{y}), \text{ d. h. } a(\vec{x} - P_h \vec{x}, \vec{y}) = 0 \text{ für alle } \vec{y} \in S_h. \quad (5.47)$$

P_h ist offensichtlich eine lineare Abbildung, die wegen

$$c \|P_h \vec{x}\|_V^2 \stackrel{(5.44)}{\leq} a(P_h \vec{x}, P_h \vec{x}) \stackrel{(5.47)}{=} a(\vec{x}, P_h \vec{x}) \stackrel{(5.43)}{\leq} C \|\vec{x}\|_V \|P_h \vec{x}\|_V,$$

d. h. $\|P_h \vec{x}\|_V \leq \frac{C}{c} \|\vec{x}\|_V$, beschränkt ist ($P_h \in [V, S_h]$, vgl. Kapitel 3.7.4). P_h wird auch als **elliptische Projektion** bezeichnet – Projektion, da nach (5.47) für alle $\vec{x} \in S_h$ gilt: $P_h \vec{x} = \vec{x}$.

Werden schwache Probleme betrachtet, die aus Randwertproblemen von Differenzialgleichungen durch partielle Integration entstehen, so haben die Inhomogenitäten f^* wie oben die spezielle Gestalt $(\vec{f}, \cdot)_H$, wobei $H = L^2$ ist:

Lemma 5.7 (Funktional der rechten Seite)

Sei $V \subset H$ stetig eingebettet, d. h. $\|\vec{x}\|_H \leq C \|\vec{x}\|_V$ für alle $\vec{x} \in V$. Dann wird durch jedes $\vec{f} \in H$ ein Funktional $f^* := (\vec{f}, \cdot)_H \in V^* \subset S_h^*$ mit $\|f^*\|_{V^*} \leq C \|\vec{f}\|_H$ erzeugt.

Beweis: $|(\vec{f}, \vec{x})_H| \leq \|\vec{f}\|_H \|\vec{x}\|_H \leq C \|\vec{f}\|_H \|\vec{x}\|_V$ für alle $\vec{x} \in V$. □

Der Index h bei der Bezeichnung des endlich-dimensionalen Untervektorraums S_h deutet an, dass in den Anwendungen eine Familie von Untervektorräumen S_h betrachtet wird, so dass $\lim_{h \rightarrow 0+} \|\vec{x} - P_h \vec{x}\|_V = 0$ möglich werden kann. Bei der Finite-Elemente-Methode könnte z. B. h der größte Durchmesser der Zerlegungsintervalle oder Dreiecke sein, über die die Ansatzfunktionen definiert sind. Wenn $\vec{x}$ die exakte Lösung des Problems (5.45) und damit ggf. einer Differenzialgleichung ist, dann ist $P_h \vec{x}$ die zugehörige

Näherungslösung, die über (5.46) berechnet wird. Für den Fehler $\|\vec{x} - P_h \vec{x}\|_V$ gilt der folgende Satz:

Satz 5.12 (Lemma von Céa)

Sei a eine beschränkte V -elliptische Bilinearform mit Konstanten C und c wie in (5.43) und (5.44). Dann gilt für jedes $\vec{x} \in V$ die Fehlerabschätzung:

$$\|\vec{x} - P_h \vec{x}\|_V \leq \frac{C}{c} \inf_{\vec{y} \in S_h} \|\vec{x} - \vec{y}\|_V.$$

Der Fehler ist damit bis auf eine feste Konstante durch den **Fehler der besten Approximation** von $\vec{x}$ durch Elemente aus S_h beschränkt. Abgesehen von der Konstanten $\frac{C}{c}$ geht es nicht besser. Um den Fehler in der Praxis konkret anzugeben, müsste man aber die in der Regel unbekannte exakte Lösung $\vec{x}$ kennen. Allerdings sind oft durch die Problemstellung Eigenschaften der Lösung gegeben, mit denen man Aussagen über den Fehler der besten Approximation treffen kann.

Beweis: Für beliebiges $\vec{y} \in S_h$ gilt

$$a(\vec{x} - P_h \vec{x}, P_h \vec{x}) \stackrel{(5.47)}{=} 0 \stackrel{(5.47)}{=} a(\vec{x} - P_h \vec{x}, \vec{y}), \quad (5.48)$$

$$\begin{aligned} c\|\vec{x} - P_h \vec{x}\|_V^2 &\stackrel{(5.44)}{\leq} a(\vec{x} - P_h \vec{x}, \vec{x} - P_h \vec{x}) = a(\vec{x} - P_h \vec{x}, \vec{x}) - a(\vec{x} - P_h \vec{x}, P_h \vec{x}) \\ &\stackrel{(5.48)}{=} a(\vec{x} - P_h \vec{x}, \vec{x}) - a(\vec{x} - P_h \vec{x}, \vec{y}) = a(\vec{x} - P_h \vec{x}, \vec{x} - \vec{y}) \\ &\stackrel{(5.43)}{\leq} C\|\vec{x} - P_h \vec{x}\|_V \|\vec{x} - \vec{y}\|_V. \end{aligned}$$

Der Übergang zum Infimum bezüglich $\vec{y} \in S_h$ liefert die Behauptung. $\square$

Ein wichtiger Vorteil von Finite-Elemente-Methoden gegenüber klassischen Differenzenverfahren ist die Stabilität (Stetigkeit, siehe Kapitel 3.7.4.3). Seien $f_1, f_2 \in H$ zwei Inhomogenitäten, über die die Funktionale $f_1^*(\cdot) = (f_1, \cdot)_H$ und $f_2^*(\cdot) = (f_2, \cdot)_H$ definiert sind. Außerdem sei V stetig in H eingebettet mit $\|\vec{x}\|_H \leq C\|\vec{x}\|_V$ wie in Lemma 5.7. Dann gilt für die zugehörigen Lösungen $\vec{x}_{1,h}$ und $\vec{x}_{2,h}$ des diskreten Problems (5.46)

$$\begin{aligned} c\|\vec{x}_{1,h} - \vec{x}_{2,h}\|_V^2 &\stackrel{(5.44)}{\leq} a(\vec{x}_{1,h} - \vec{x}_{2,h}, \vec{x}_{1,h} - \vec{x}_{2,h}) \stackrel{(5.46)}{=} (f_1 - f_2, \vec{x}_{1,h} - \vec{x}_{2,h})_H \\ &\leq \|f_1 - f_2\|_H \|\vec{x}_{1,h} - \vec{x}_{2,h}\|_H \leq C\|f_1 - f_2\|_H \|\vec{x}_{1,h} - \vec{x}_{2,h}\|_V, \end{aligned}$$

so dass

$$\|\vec{x}_{1,h} - \vec{x}_{2,h}\|_V \leq \frac{C}{c} \|f_1 - f_2\|_H.$$

Kleine Änderungen der Eingabe (der Inhomogenität) führen also auch nur zu kleinen Änderungen der Finite-Elemente-Lösung. Das Verfahren ist stabil.

5.5.4 Beispiel für die Finite-Elemente-Methode in 2-D

Im letzten Abschnitt des Kapitels wenden wir die Finite-Elemente-Methode am Beispiel des zweidimensionalen Dirichlet-Randwertproblems für die Poisson-Gleichung mit homogenen Randbedingungen an.

5.5.4.1 Ein Dirichlet-Randwertproblem

Unter Verwendung des Laplace-Operators $\Delta u := \frac{\partial^2}{\partial x^2} u + \frac{\partial^2}{\partial y^2} u$ wird eine Lösung u des Randwertproblems

$$\Delta u(x, y) = f(x, y) \text{ für alle } (x, y) \in \Omega \text{ und } u(x, y) = 0 \text{ für alle } (x, y) \in \partial\Omega \quad (5.49)$$

gesucht. Hier ist $\Omega \subset \mathbb{R}^2$ ein Gebiet mit Rand $\partial\Omega$. In diesem Abschnitt liegt der Fokus auf der praktischen Umsetzung der Methode. Daher setzen wir das Gebiet, den Rand, die Inhomogenität f und damit auch die exakte Lösung u als hinreichend gutmütig für die Rechnungen voraus.

Die Poisson-Gleichung und den Laplace-Operator haben wir bereits auf den Seiten 680, 685 und im Kasten auf Seite 696 kennen gelernt.

Ausgangspunkt für die Finite-Elemente-Methode ist die schwache Form der Differentialgleichung. Wir betrachten den Sobolev-Raum der einmal schwach differenzierbaren Testfunktionen $V := W_0^{1,2}(\Omega)$, die die Randbedingung erfüllen, vgl. Kapitel 4.4.5.2 ab Seite 669. Für die Herleitung der schwachen Form von (5.49) multiplizieren wir beide Seiten der Differentialgleichung mit einer Testfunktion $v \in V$ und integrieren analog zur eindimensionalen Rechnung (5.42) über Ω . Eine Lösung u erfüllt

$$\int_{\Omega} v(x, y) \cdot \Delta u(x, y) d(x, y) = \int_{\Omega} f(x, y) \cdot v(x, y) d(x, y).$$

Mit der Produktregel für die Divergenz (Lemma 4.4 auf Seite 679)

$$\operatorname{div}(v \cdot \operatorname{grad} u) = \operatorname{grad} v \cdot \operatorname{grad} u + v \operatorname{div} \operatorname{grad} u = \operatorname{grad} v \cdot \operatorname{grad} u + v \Delta u$$

und dem Satz von Green (Satz 4.19 auf Seite 694 ist auch für Lebesgue-Integrale anwendbar) erhalten wir

$$\begin{aligned} \int_{\Omega} v \cdot \Delta u d(x, y) &= \int_{\Omega} \operatorname{div}(v \cdot \operatorname{grad} u) d(x, y) - \int_{\Omega} \operatorname{grad} v \cdot \operatorname{grad} u d(x, y) \\ &= \int_{\Omega} \frac{\partial}{\partial x} \left[v(x, y) \frac{\partial}{\partial x} u(x, y) \right] - \frac{\partial}{\partial y} \left[-v(x, y) \frac{\partial}{\partial y} u(x, y) \right] d(x, y) \\ &\quad - \int_{\Omega} \operatorname{grad} v \cdot \operatorname{grad} u d(x, y) \\ &= \underbrace{\oint_{\partial\Omega} v(x, y) \left[-\frac{\partial}{\partial y} u(x, y) + \frac{\partial}{\partial x} u(x, y) \right] d(x, y)}_{=0 \text{ da } v|_{\partial\Omega}=0} - \int_{\Omega} \operatorname{grad} v \cdot \operatorname{grad} u d(x, y) \end{aligned}$$

und gelangen zur schwachen Formulierung des Randwertproblems

$$\int_{\Omega} \operatorname{grad} v \cdot \operatorname{grad} u \, d(x, y) = - \int_{\Omega} v \cdot f \, d(x, y) \quad \text{für alle } v \in V. \quad (5.50)$$

Bei drei (und mehr) Variablen gelingt diese mehrdimensionale partielle Integration mit dem Satz von Gauß (siehe Satz 4.20 auf Seite 694).

Mit der Bilinearform $a(.,.)$ und dem Funktional $l(.)$

$$a(v, u) := \int_{\Omega} \operatorname{grad} v \cdot \operatorname{grad} u \, d(x, y), \quad l(v) := - \int_{\Omega} v \cdot f \, d(x, y) \quad (5.51)$$

erhalten wir die **schwache Form** der Poisson-Gleichung (vgl. (5.45), Kasten auf Seite 696): Wir suchen $u \in V$ mit

$$a(v, u) = l(v) \quad \text{für alle } v \in V. \quad (5.52)$$

Tatsächlich ist a in unserem Beispiel eine beschränkte V -elliptische Bilinearform, so dass das schwache Problem eindeutig lösbar ist (siehe Satz 5.11).

Statt der Differenzialgleichung lösen wir nun dieses Problem wieder näherungsweise mit dem Ritz-Galerkin-Verfahren, indem wir (5.52) auf einem endlich-dimensionalen Ansatzfunktionenraum $V_N \subset V$ statt auf V betrachten. Hier liefert Satz 5.11 nun die eindeutige Existenz einer Näherungslösung. Diese wird in der Praxis über ein Gleichungssystem mit der Steifigkeitsmatrix berechnet, wie wir es im eindimensionalen Fall bereits angedeutet haben. Das führen wir nun genauer aus.

Wir betrachten endlich-dimensionale Untervektorräume $V_N \subset V$ mit Basen $B_N := \{\phi_1(x, y), \dots, \phi_N(x, y)\}$ und suchen Näherungslösungen $\hat{u} \in V_N$ mit

$$a(v, \hat{u}) = l(v) \quad \text{für alle } v \in V_N. \quad (5.53)$$

Die Existenz der Finite-Elemente-Lösung $\hat{u} \in V_N$ von (5.53) und deren Eindeutigkeit folgen wieder aus dem Satz von Riesz bzw. dem Lemma von Lax und Milgram aus dem Kasten ab Seite 796.

Die Wahl von V_N und der Basisfunktionen ϕ_k hängt vom Gebiet Ω ab, darauf gehen wir in den folgenden Abschnitten genauer ein. Im Kasten auf Seite 795 haben wir die endlich-dimensionalen Vektorräume mit V_h bezeichnet. Der Parameter h bezeichnet typischerweise eine Abschätzung des Durchmessers der finiten Elemente, siehe Seite 807. Mit wachsendem N wird dann h kleiner.

Die Lösung $\hat{u}$ und die Testfunktionen v aus V_N stellen wir als Linearkombinationen der Basisfunktionen dar:

$$\hat{u}(x, y) = \sum_{i=1}^N \hat{u}_i \phi_i(x, y), \quad v(x, y) = \sum_{k=1}^N v_k \phi_k(x, y), \quad \hat{u}_i, v_k \in \mathbb{R}.$$

Einsetzen in (5.53) liefert:

$$\begin{aligned} a(v, \hat{u}) = l(v) &\iff a\left(\sum_{k=1}^N v_k \phi_k, \sum_{i=1}^N \hat{u}_i \phi_i\right) = l\left(\sum_{k=1}^N v_k \phi_k\right) \\ &\iff \sum_{k=1}^N \sum_{i=1}^N v_k \hat{u}_i a(\phi_k, \phi_i) = \sum_{k=1}^N v_k l(\phi_k) \iff \vec{v}^\top \cdot \mathbf{A} \cdot \hat{\vec{u}} = \vec{v}^\top \cdot \vec{l} \end{aligned}$$

mit $\vec{v} := (v_1, \dots, v_N)^\top$, $\hat{\vec{u}} := (\hat{u}_1, \dots, \hat{u}_N)^\top$, $\vec{l} := (l(\phi_1), \dots, l(\phi_N))^\top$ und der **Steifigkeitsmatrix**

$$\mathbf{A} = \begin{bmatrix} a(\phi_1, \phi_1) & \dots & a(\phi_1, \phi_N) \\ & \ddots & \\ a(\phi_N, \phi_1) & \dots & a(\phi_N, \phi_N) \end{bmatrix}. \quad (5.54)$$

Aufgrund der Eigenschaften der Bilinearform a ist die Matrix $\mathbf{A}$ symmetrisch und positiv definit (und damit insbesondere invertierbar). Der Vektor $\vec{l}$ heißt **Lastvektor**. Da die Gleichung für beliebige Vektoren $\vec{v} \in \mathbb{R}^N$ gilt, ist $\hat{\vec{u}}$ (eindeutige) Lösung des linearen Gleichungssystems

$$\mathbf{A} \cdot \hat{\vec{u}} = \vec{l}. \quad (5.55)$$

Der Ritz-Galerkin-Ansatz hat das schwache Randwertproblem in das lineare Gleichungssystem (5.55) überführt. Mit der Lösung $\hat{\vec{u}}$ von (5.55) erhalten wir die Finite-Elemente-Lösung

$$\hat{u}(x, y) = \sum_{i=1}^N \hat{u}_i \phi_i(x, y). \quad (5.56)$$

5.5.4.2 Finite-Elemente-Diskretisierung, nodale Basis

Im vorigen Abschnitt haben wir (5.53) auf dem endlich-dimensionalen Teilraum $V_N \subset V$ mit Basis $B_N = \{\phi_1(x, y), \dots, \phi_N(x, y)\}$ gelöst. Nun gehen wir auf die Konstruktion von konkreten Ansatzfunktionenräumen V_N mit Basen B_N für $\Omega \subset \mathbb{R}^2$ ein. Dazu zerlegen wir $\Omega \cup \partial\Omega$ in einfach strukturierte Teilmengen, die finiten Elemente. Bei zwei Variablen sind das in der Regel (und hier) Dreiecke. Damit eine solche Triangulierung gelingt, muss der Rand $\partial\Omega$ ein Polygonzug sein. Sonst wird $\Omega \cup \partial\Omega$ ersetzt durch eine polygonal berandete Menge $\hat{\Omega} \cup \partial\hat{\Omega}$, die als Vereinigung der finiten Elemente (hier Dreiecke) $E_r \subset \mathbb{R}^2$, $r = 1, \dots, N_{\text{El}}$, geschrieben werden kann, vgl. Abbildung 5.16. Wir werden das diskretisierte schwache Problem nicht auf Ω , sondern lediglich auf der offenen Menge $\hat{\Omega}$ lösen. Die Dreiecke besitzen **Knoten** (Eckpunkte), die entweder in $\hat{\Omega}$ oder in $\partial\hat{\Omega}$ liegen. Seien $\vec{x}_i \in \hat{\Omega}$, $i = 1, \dots, N$ die in $\hat{\Omega}$ enthaltenen Knoten. Diese heißen **innere Knoten**. Knoten in $\partial\hat{\Omega}$ heißen dagegen **Randknoten**. Wir verlangen, dass die Randknoten auch auf dem ursprünglichen Rand $\partial\Omega$ liegen. Aufgrund der Randbedingung muss die gesuchte Finite-Elemente-Lösung hier null werden.

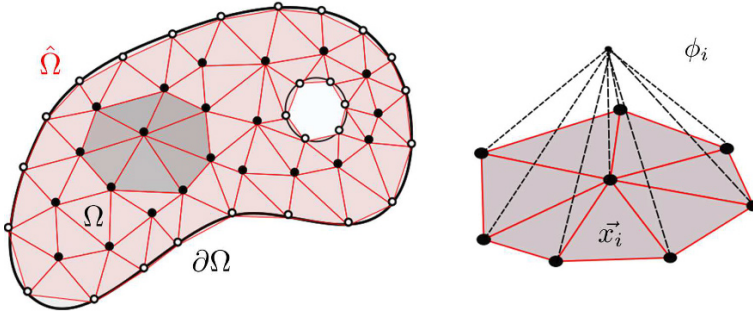


Abb. 5.16: Gebiet $\Omega \subset \mathbb{R}^2$ und Gebiet $\hat{\Omega}$ mit Knoten $\vec{x}_i$, $i = 1, \dots, N$. Die finiten Elemente sind Dreiecke mit $\vec{x}_i$ als (innere) Ecken. Für die Knotenfunktion $\phi_i(x, y)$ gilt $\phi_i(\vec{x}_r) = \delta_{i,r}$. Der Träger von ϕ_i besteht aus den Dreiecken, die $\vec{x}_i$ als Ecke haben.

Für einen inneren Knoten $\vec{x}_i$ erklären wir die **Knotenfunktion** (Basisfunktion) $\phi_i(x, y)$ wie folgt:

- Es gilt $\phi_i(\vec{x}_i) = 1$ und $\phi_i(\vec{x}_k) = 0$ für $k \neq i$, d. h. unter Verwendung des Kronecker-Delta (vgl. Seite 490)

$$\phi_i(\vec{x}_k) = \delta_{i,k}. \quad (5.57)$$

Darüber hinaus gilt $\phi_i(\vec{y}) = 0$ für alle Randknoten $\vec{y}$.

- ϕ_i „lebt“ auf denjenigen Elementen E_k , die $\vec{x}_i$ als Ecke haben. Ist $\vec{x}_i$ nicht Ecke von E_l , so gilt $\phi_i(x, y) = 0$ für alle $(x, y) \in E_l$. Der **Träger**

$$\text{supp}(\phi_i) := \bigcup_{k \in \{1, \dots, N_{\text{El}}\} : \vec{x}_i \text{ ist Ecke von } E_k} E_k \quad (5.58)$$

von ϕ_i besteht genau aus den finiten Elementen E_k , die $\vec{x}_i$ als Ecke besitzen, vgl. Abbildung 5.16.

- $\phi_i(x, y)$ ist stückweise linear, also ein stückweise lineares Polynom in zwei Variablen: Auf jedem Element E ist $\phi_i(x, y) = c_{0,0} + c_{1,0}x + c_{0,1}y$ mit Koeffizienten $c_{0,0}, c_{1,0}, c_{0,1} \in \mathbb{R}$ (die durch die vorangehenden Bedingungen eindeutig bestimmt sind).

Jede stückweise lineare Ansatzfunktion $\phi_i(x, y)$ ist stetig und im Inneren jedes finiten Elements partiell nach x und y differenzierbar. Die fehlende Differenzierbarkeit an den Elementrändern des Trägers ändert nichts an der Zugehörigkeit der Funktion zu $W_0^{1,2}(\hat{\Omega})$.

Die Funktionenmenge $B_N = \{\phi_i(x, y) : i = 1, \dots, N\}$ heißt **nodale Basis**. Der von B_N erzeugte Funktionenraum $V_N \subset W_0^{1,2}(\hat{\Omega})$ hat die Dimension N .

Für die Elemente der Steifigkeitsmatrix **A** bezüglich V_N erhalten wir mit (5.51), (5.54) und (5.58)

$$a(\phi_i, \phi_r) = \int_{\hat{\Omega}} \text{grad } \phi_i \cdot \text{grad } \phi_r \, d(x, y) = \int_{\text{supp}(\phi_i) \cap \text{supp}(\phi_r)} \text{grad } \phi_i \cdot \text{grad } \phi_r \, d(x, y). \quad (5.59)$$

Es gilt $a(\phi_i, \phi_r) = 0$, wenn die Knoten $\vec{x}_i$ und $\vec{x}_r$ nicht beide Ecken desselben Dreiecks E_k sind. Der Integrationsbereich bei (5.59) besteht aus den finiten Elementen E_k , die $\vec{x}_i$ und $\vec{x}_r$ als Ecke besitzen:

$$a(\phi_i, \phi_r) = \sum_{k \in \{1, \dots, N_{\text{El}}\}: \vec{x}_i \text{ und } \vec{x}_r \text{ sind Ecken von } E_k} a(\phi_i, \phi_r)|_{E_k}$$

mit

$$a(\phi_i, \phi_r)|_{E_k} := \int_{E_k} \text{grad } \phi_i \cdot \text{grad } \phi_r \, d(x, y). \quad (5.60)$$

Entsprechend berechnen wir den Lastvektor $\vec{l} = (l(\phi_1), \dots, l(\phi_N))^T$ mit

$$\begin{aligned} l(\phi_i) &:= - \int_{\hat{\Omega}} \phi_i \cdot f \, d(x, y) = - \int_{\text{supp}(\phi_i)} \phi_i \cdot f \, d(x, y) \\ &= - \sum_{k \in \{1, \dots, N_{\text{El}}\}: \vec{x}_i \text{ ist Ecke von } E_k} \int_{E_k} \phi_i \cdot f \, d(x, y). \end{aligned} \quad (5.61)$$

Die Steifigkeitsmatrix $\mathbf{A}$ und der Lastvektor $\vec{l}$ werden iterativ über alle Dreiecke E_n , $n = 1, \dots, N_{\text{El}}$, aufgebaut. Die Komponente $a_{i,r} = a(\phi_i, \phi_r)$ der Matrix ist die Summe der Anteile $a(\phi_i, \phi_r)|_{E_n}$. Dabei liefern nur Dreiecke E_n einen Beitrag zu $a_{i,r}$, die als Ecken sowohl den Knoten $\vec{x}_i$ als auch $\vec{x}_r$ besitzen. Falls $\vec{x}_i$ und $\vec{x}_r$ nicht beide zu einem Dreieck gehören, ist $a(\phi_i, \phi_r) = 0$. Daher ist $\mathbf{A}$ dünn besetzt. Ein Eintrag $l_i = l(\phi_i)$, $i \in \{1, \dots, N\}$, des Lastvektors entsteht durch Summation der Integrale $-\int_{E_n} \phi_i \cdot f \, d(x, y)$. Diese Integrale über E_n können nur von null verschieden sein, falls $\vec{x}_i$ eine Ecke von E_n ist.

5.5.4.3 Berechnung der Integrale durch Transformation auf ein Referenzdreieck

Für die Berechnung der $a(\phi_i, \phi_r)$ nach (5.60) und der $l(\phi_i)$ nach (5.61) sind Integrale über Dreieckselemente E zu berechnen. Hierfür transformieren wir die Integrale auf ein Referenzdreieck R . Da die nodalen Basisfunktionen ϕ_k als stückweise linear angenommen werden, ist die Berechnung der Integrale dann einfach durchführbar.

Ein Dreieck E habe als Ecken die inneren oder äußeren Knoten $\vec{x}_r = (x_r, y_r)^T$, $\vec{x}_k = (x_k, y_k)^T$ und $\vec{x}_s = (x_s, y_s)^T$. Die Transformation

$$\vec{X}(\xi_1, \xi_2) := \vec{x}_r + (\vec{x}_k - \vec{x}_r)\xi_1 + (\vec{x}_s - \vec{x}_r)\xi_2, \quad 0 \leq \xi_1 \leq 1, \quad 0 \leq \xi_2 \leq 1 - \xi_1,$$

bildet das Referenzdreieck

$$R := \{(\xi_1, \xi_2) : 0 \leq \xi_1 \leq 1, \quad 0 \leq \xi_2 \leq 1 - \xi_1\}$$

in der (ξ_1, ξ_2) -Ebene mit den Ecken $(0, 0)^T$, $(1, 0)^T$ und $(0, 1)^T$ auf das Dreieck E in der (x, y) -Ebene ab, vgl. Abbildung 5.17. Es gilt

$$\vec{X}(0, 0) = \vec{x}_r, \quad \vec{X}(1, 0) = \vec{x}_k, \quad \vec{X}(0, 1) = \vec{x}_s.$$

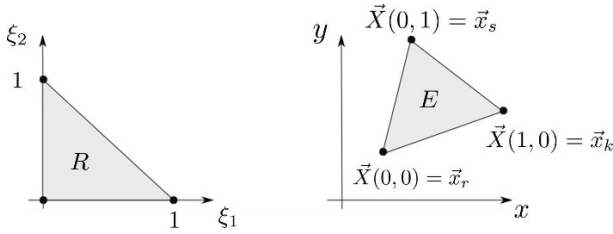


Abb. 5.17: Transformation des Referenzdreiecks R auf das Dreieck E

Wir betrachten für eine geeignete, auf E definierte reellwertige Funktion g die Substitution (vgl. Satz 4.16 auf Seite 661)

$$\int_E g(x, y) d(x, y) = \int_R g(\vec{X}(\xi_1, \xi_2)) |\det \vec{X}'(\xi_1, \xi_2)| d(\xi_1, \xi_2).$$

Für die Jacobi-Matrix der Transformation $\vec{X}$ erhalten wir

$$\vec{X}'(\xi_1, \xi_2) = \left[\frac{\partial}{\partial \xi_1} \vec{X}(\xi_1, \xi_2), \frac{\partial}{\partial \xi_2} \vec{X}(\xi_1, \xi_2) \right] = [\vec{x}_k - \vec{x}_r, \vec{x}_s - \vec{x}_r]. \quad (5.62)$$

Da die Jacobi-Matrix konstant ist und nur vom Element E abhängt, schreiben wir im Folgenden $\mathbf{J}_E$ anstatt $\vec{X}'(\xi_1, \xi_2)$. Mit der Funktionaldeterminante

$$\det \mathbf{J}_E = \begin{vmatrix} x_k - x_r & x_s - x_r \\ y_k - y_r & y_s - y_r \end{vmatrix} = (x_k - x_r) \cdot (y_s - y_r) - (x_s - x_r) \cdot (y_k - y_r)$$

folgt

$$\int_E g(x, y) d(x, y) = |\det \mathbf{J}_E| \int_R g(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2).$$

Jetzt wenden wir die Substitution zur näherungsweisen Berechnung des Lastvektors (5.61) an. Auf E integrieren wir dazu die Funktion $g(x, y) := f(x, y)\phi_m(x, y)$:

$$\int_E f(x, y)\phi_m(x, y) d(x, y) = |\det \mathbf{J}_E| \int_R f(\vec{X}(\xi_1, \xi_2))\phi_m(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2).$$

Diese Integrale lassen sich mittels eines Quadraturverfahrens numerisch berechnen. Dazu approximieren wir die Funktion f auf E mit der nodalen Basis, die wir nur zu diesem Zweck um Basisfunktionen für die Randknoten erweitern. Denn die Inhomogenität muss nicht die Nullrandbedingung erfüllen. Dann ist

$$f(x, y) \approx \hat{f}(x, y) := f(\vec{x}_r)\phi_r(x, y) + f(\vec{x}_k)\phi_k(x, y) + f(\vec{x}_s)\phi_s(x, y),$$

und wir setzen in (5.62) ein:

$$\int_E \hat{f}(x, y)\phi_m(x, y) d(x, y)$$

$$\begin{aligned}
&= \int_E [f(\vec{x}_r)\phi_r(x, y) + f(\vec{x}_k)\phi_k(x, y) + f(\vec{x}_s)\phi_s(x, y)] \cdot \phi_m(x, y) d(x, y) \\
&= |\det \mathbf{J}_E| (f(\vec{x}_r), f(\vec{x}_k), f(\vec{x}_s)) \cdot \begin{pmatrix} \int_R \phi_r(\vec{X}(\xi_1, \xi_2))\phi_m(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) \\ \int_R \phi_k(\vec{X}(\xi_1, \xi_2))\phi_m(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) \\ \int_R \phi_s(\vec{X}(\xi_1, \xi_2))\phi_m(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) \end{pmatrix}.
\end{aligned}$$

Wegen $\phi_r(\vec{X}(\xi_1, \xi_2)) = 1 - \xi_1 - \xi_2$, $\phi_k(\vec{X}(\xi_1, \xi_2)) = \xi_1$, $\phi_s(\vec{X}(\xi_1, \xi_2)) = \xi_2$ berechnen sich die Integrale (mittels Integration über das Referenzdreieck als Normalbereich, siehe Satz 4.15 auf Seite 659) zu

$$\begin{aligned}
\int_R \phi_k^2(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) &= \int_0^1 \int_0^{1-\xi_1} \xi_1^2 d\xi_2 d\xi_1 = \int_0^1 \xi_1^2 - \xi_1^3 d\xi_1 = \frac{1}{12} \\
\int_R \phi_r^2(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) &= \int_0^1 \int_0^{1-\xi_1} (1 - \xi_1 - \xi_2)^2 d\xi_2 d\xi_1 = \frac{1}{12} \\
\int_R \phi_s^2(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) &= \int_0^1 \int_0^{1-\xi_1} \xi_2^2 d\xi_2 d\xi_1 = \frac{1}{12} \\
\int_R \phi_r(\vec{X}(\xi_1, \xi_2))\phi_k(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) &= \int_0^1 \int_0^{1-\xi_1} \xi_1 - \xi_1^2 - \xi_1\xi_2 d\xi_2 d\xi_1 = \frac{1}{24} \\
\int_R \phi_r(\vec{X}(\xi_1, \xi_2))\phi_s(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) &= \int_0^1 \int_0^{1-\xi_1} \xi_2 - \xi_1\xi_2 - \xi_2^2 d\xi_2 d\xi_1 = \frac{1}{24} \\
\int_R \phi_k(\vec{X}(\xi_1, \xi_2))\phi_s(\vec{X}(\xi_1, \xi_2)) d(\xi_1, \xi_2) &= \int_0^1 \int_0^{1-\xi_1} \xi_1\xi_2 d\xi_2 d\xi_1 = \frac{1}{24}.
\end{aligned}$$

Damit erhalten wir für $m \in \{r, k, s\}$:

$$\begin{aligned}
\int_E \hat{f}(x, y)\phi_r(x, y) d(x, y) &= |\det \mathbf{J}_E| [f(\vec{x}_r), f(\vec{x}_k), f(\vec{x}_s)] \cdot \frac{1}{24} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \\
\int_E \hat{f}(x, y)\phi_k(x, y) d(x, y) &= |\det \mathbf{J}_E| [f(\vec{x}_r), f(\vec{x}_k), f(\vec{x}_s)] \cdot \frac{1}{24} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \\
\int_E \hat{f}(x, y)\phi_s(x, y) d(x, y) &= |\det \mathbf{J}_E| [f(\vec{x}_r), f(\vec{x}_k), f(\vec{x}_s)] \cdot \frac{1}{24} \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}.
\end{aligned}$$

Weiter gilt

$$\int_E \hat{f}(x, y)\phi_m(x, y) d(x, y) = 0 \quad \text{für } m \notin \{r, k, s\}.$$

Für die Berechnung der Steifigkeitsmatrix benötigen wir als Summanden die Zahlen

$$a(\phi_n, \phi_m)|_E = \int_E \text{grad } \phi_n(x, y) \cdot \text{grad } \phi_m(x, y)^\top d(x, y) \quad \text{falls } n, m \in \{r, k, s\}.$$

Die mehrdimensionale Kettenregel (Satz 4.4, Seite 607) liefert

$$\text{grad}[\phi_n(\vec{X}(\xi_1, \xi_2))] = [\text{grad } \phi_n](\vec{X}(\xi_1, \xi_2)) \cdot [\vec{x}_k - \vec{x}_r, \vec{x}_s - \vec{x}_r] = [\text{grad } \phi_n](\vec{X}(\xi_1, \xi_2)) \cdot \mathbf{J}_E,$$

wobei $[\text{grad } \phi_n](\vec{X}(\xi_1, \xi_2))$ den an der Stelle $\vec{X}(\xi_1, \xi_2)$ ausgewerteten Gradient von ϕ_n bezeichnet und wir rechts mit der (invertierbaren) Jacobi-Matrix von $\vec{X}$ multiplizieren. Insbesondere ist

$$[\text{grad } \phi_n](\vec{X}(\xi_1, \xi_2)) = \text{grad}[\phi_n(\vec{X}(\xi_1, \xi_2))] \cdot \mathbf{J}_E^{-1}.$$

Damit folgt als Beitrag des Integrals über das Element E zu $a(\phi_n, \phi_m)$

$$\begin{aligned} a(\phi_n, \phi_m)|_{E_i} &= \int_E \text{grad } \phi_n(x, y) \cdot \text{grad } \phi_m(x, y)^\top d(x, y) \\ &= |\det \mathbf{J}_E| \int_R [\text{grad } \phi_n](\vec{X}(\xi_1, \xi_2)) \cdot ([\text{grad } \phi_m](\vec{X}(\xi_1, \xi_2)))^\top d(\xi_1, \xi_2) \\ &= |\det \mathbf{J}_E| \int_R \text{grad}[\phi_n(\vec{X}(\xi_1, \xi_2))] \cdot \mathbf{J}_E^{-1} \cdot (\text{grad}[\phi_m(\vec{X}(\xi_1, \xi_2))] \cdot \mathbf{J}_E^{-1})^\top d(\xi_1, \xi_2) \\ &= |\det \mathbf{J}_E| \int_R \text{grad}[\phi_n(\vec{X}(\xi_1, \xi_2))] \cdot \mathbf{Q}_E \cdot (\text{grad}[\phi_m(\vec{X}(\xi_1, \xi_2))])^\top d(\xi_1, \xi_2) \end{aligned}$$

mit der symmetrischen Matrix (vgl. Satz 1.29 auf Seite 172)

$$\mathbf{Q}_E = \begin{bmatrix} Q_{1,1} & Q_{1,2} \\ Q_{2,1} & Q_{2,2} \end{bmatrix} := \mathbf{J}_E^{-1} \cdot [\mathbf{J}_E^{-1}]^\top = \mathbf{J}_E^{-1} \cdot [\mathbf{J}_E^\top]^{-1} = [\mathbf{J}_E^\top \cdot \mathbf{J}_E]^{-1}.$$

Die Gradienten der Knotenfunktionen sind konstant: $\text{grad}[\phi_r(\vec{X}(\xi_1, \xi_2))] = (-1, -1)$, $\text{grad}[\phi_k(\vec{X}(\xi_1, \xi_2))] = (1, 0)$ und $\text{grad}[\phi_s(\vec{X}(\xi_1, \xi_2))] = (0, 1)$. Damit folgt, falls $n, m \in \{r, k, s\}$:

$$a(\phi_n, \phi_m)|_E = \frac{|\det \mathbf{J}_E|}{2} \text{grad}[\phi_n(\vec{X}(\xi_1, \xi_2))] \cdot \mathbf{Q}_E \cdot \text{grad}[\phi_m(\vec{X}(\xi_1, \xi_2))]^\top.$$

Im Einzelnen gilt:

$$\begin{aligned} a(\phi_r, \phi_r)|_E &= \frac{|\det \mathbf{J}_E|}{2} (-1, -1) \cdot \mathbf{Q}_E \cdot \begin{pmatrix} -1 \\ -1 \end{pmatrix} \\ a(\phi_k, \phi_k)|_E &= \frac{|\det \mathbf{J}_E|}{2} (1, 0) \cdot \mathbf{Q}_E \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{|\det \mathbf{J}_E| Q_{1,1}}{2} \\ a(\phi_s, \phi_s)|_E &= \frac{|\det \mathbf{J}_E|}{2} (0, 1) \cdot \mathbf{Q}_E \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{|\det \mathbf{J}_E| Q_{2,2}}{2} \\ a(\phi_r, \phi_k)|_E &= \frac{|\det \mathbf{J}_E|}{2} (-1, -1) \cdot \mathbf{Q}_E \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -\frac{|\det \mathbf{J}_E|}{2} (Q_{1,1} + Q_{2,1}) \end{aligned}$$

$$a(\phi_r, \phi_s)|_E = \frac{|\det \mathbf{J}_E|}{2} (-1, -1) \cdot \mathbf{Q}_E \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = -\frac{|\det \mathbf{J}_E|}{2} (Q_{1,2} + Q_{2,2})$$

$$a(\phi_k, \phi_s)|_E = \frac{|\det \mathbf{J}_E|}{2} (1, 0) \cdot \mathbf{Q}_E \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{|\det \mathbf{J}_E|}{2} Q_{1,2}.$$

Aufgrund der Symmetrie ist $a(\phi_n, \phi_m)|_E = a(\phi_m, \phi_n)|_E$. Für $n \notin \{r, k, s\}$ oder $m \notin \{r, k, s\}$ ist $a(\phi_n, \phi_m)|_E = 0$.

5.5.4.4 Lösung des linearen Gleichungssystems und Fehlerverhalten

Wir können nun alle Komponenten der Steifigkeitsmatrix $\mathbf{A}$ und des Lastvektors $\vec{l}$ für das lineare Gleichungssystem (5.55) berechnen. Die Randbedingung $u(x, y) = 0$ für $(x, y) \in \partial\Omega$ ist bereits dadurch berücksichtigt, dass wir zu Randknoten keine Elemente der nodalen Basis gebildet haben.

Die Steifigkeitsmatrix $\mathbf{A}$ ist symmetrisch, positiv definit und dünn besetzt. Für derartige lineare Gleichungssysteme verwendet man standardmäßig CG-Verfahren, siehe Beispiel 4.23 auf Seite 623. Darüber hinaus wird eine Vorkonditionierung empfohlen (vgl. Seite 574), beispielsweise durch Verwendung einer unvollständigen Cholesky-Zerlegung, siehe z. B. (Jung, 2001, S. 493ff), vgl. Seite 627.

Mit Lemma 5.12 auf Seite 798 können wir (zumindest im Fall eines polygonalen Gebiets $\Omega = \hat{\Omega}$ und exakter Inhomogenität $\hat{f} = f$) eine Fehlerschranke in einer Sobolev-Norm (siehe Kapitel 4.4.5.2) angeben:

$$\|u - \hat{u}\|_{1,2,\Omega} \leq \frac{C}{c} \inf_{v \in V_N} \|u - v\|_{1,2,\Omega}.$$

Das Infimum gibt den geringstmöglichen Fehler der Approximation von u durch Ansatzfunktionen $v \in V_N$ an. Ist sogar $u \in W_0^{1,2}(\Omega) \cap W^{2,2}(\Omega)$, so kann bei den von uns betrachteten stückweise linearen Ansatzfunktionen weiter gegen eine von u abhängende Konstante C_u mal h abgeschätzt werden, wobei h der maximale Durchmesser der finiten Elemente ist: $\|u - \hat{u}\|_{1,2,\Omega} \leq C_u h$. Konstruiert man die Ansatzfunktionenräume über Polynome höheren Grades r , so lässt sich der Exponent von h von 1 zu r verbessern. Der Fehler wird hier in der Norm des Sobolev-Raums $W_0^{1,2}(\Omega)$ gemessen, bei der auch die ersten schwachen Ableitungen berücksichtigt werden. Mit dem sogenannten **Nitsche-Trick**, siehe (Ciarlet, 1990, S. 141), lässt sich daraus eine Abschätzung des Fehlers in der $L^2(\Omega)$ -Norm gewinnen: $\|u - \hat{u}\|_{L^2(\Omega)} \leq C_u h^2$. Da auf der linken Seite keine Ableitungen mehr verglichen werden, erhöht sich für stückweise lineare Ansatzfunktionen die Konvergenzordnung von h zu h^2 . In der L^2 -Norm wird durch die Integration der Fehler gemittelt. Interessanter ist aber häufig der punktweise Fehler an vorgegebenen Stellen. Bei stückweise linearen Ansatzfunktionen kann der punktweise Fehler außerhalb der Knoten aber um einen $|\ln(h)|$ -Faktor größer sein, siehe Schatz

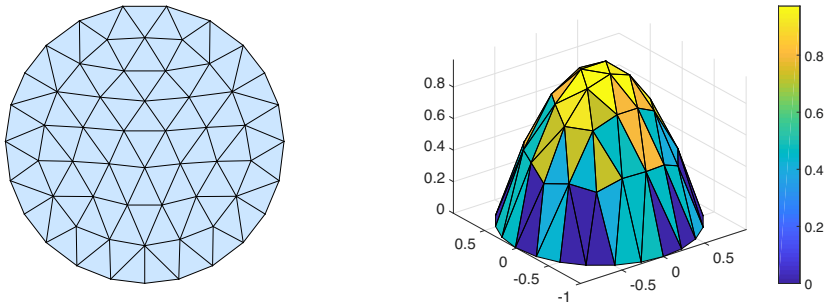


Abb. 5.18: Finite-Elemente-Lösung des homogenen Randwertproblems (5.63) (rechts) für den Einheitskreis Ω und der Triangulierung von Ω mit $h = 0,25$ (links)

(1980). Dieser Faktor tritt bei Verwendung stückweiser Polynome höherer Ordnung aber nicht auf.

5.5.4.5 Triangulierung des Gebiets und Beispielrechnungen

Wesentlich für die Finite-Elemente-Methode ist die **Triangulierung** von Ω bzw. die **Gittergenerierung**, d. h. die Wahl der inneren Knoten und Randknoten sowie der zugehörigen finiten Elemente. Hierfür gibt es verschiedene Strategien und Algorithmen, auf die wir im Rahmen dieser Einführung nicht eingehen können. Für die Triangulation von Ω haben wir das Paket „DistMesh“ von Per-Olof Persson und Gilbert Strang verwendet, siehe Persson und Strang (2004), das als MATLAB-Code verfügbar ist. Der Aufbau der Steifigkeitsmatrix $\mathbf{A}$ und des Lastvektors $\vec{l}$ sowie die Lösung des linearen Gleichungssystems werden mit dem MATLAB-Scriptfile „poissonv2.m“ durchgeführt, das im Rahmen der „Math 692 one-page MATLAB FEM code challenge“ unter der Leitung von Ed Bueler in 2005 erstellt wurde. Das DistMesh-Paket und das MATLAB-Scriptfile poissonv2.m unterliegen der GNU GPL.

Beispiel 5.53

Als Test für die Genauigkeit und Konvergenzordnung h^2 bzw. $|\ln(h)h^2|$ betrachten wir das Randwertproblem

$$\Delta u(x, y) = -4 \text{ für alle } (x, y) \in \Omega \text{ und } u(x, y) = 0 \text{ für alle } (x, y) \in \partial\Omega, \quad (5.63)$$

wobei $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ mit Rand $\partial\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ die Einheitskreisscheibe sei. Die Lösung ist $u(x, y) = 1 - x^2 - y^2$, $(x, y) \in \Omega$. Mit dem Paket DistMesh mit $h = 0,25$ und poissonv2.m erhalten wir die Finite-Elemente-Lösung, die in Abbildung 5.18 dargestellt ist. Für den maximalen Fehler $|\hat{u}(\vec{x}_k) - u(\vec{x}_k)|$ in den Knoten erhalten wir für verschiedene (maximale) Elementdurchmesser h die Werte in Tabelle 5.2. Die Tabelle bestätigt die Fehlerabschätzung gegen h^2 . Bei

Tab. 5.2: Knotenfehler im Beispiel 5.53

h	$\max_k \hat{u}(\vec{x}_k) - u(\vec{x}_k) $
0.5	0.039
0.25	0.0115
0.125	0.0026

Halbierung von h reduziert sich der maximale Fehler auf etwa ein Viertel. ■

Beispiel 5.54

Im zweiten Beispiel betrachten wir das Randwertproblem für die Poisson-Gleichung

$$\Delta u(x, y) = -e^{-4((x+3)^2 + (y-1)^2)} \text{ für alle } (x, y) \in \Omega, \quad u(x, y) = 0 \text{ für alle } (x, y) \in \partial\Omega,$$

auf dem Gebiet $\Omega = \{(x, y) \in \mathbb{R}^2 : 4 < x < 2 \wedge -2 < y < 2 \wedge x^2 + y^2 > 1\}$. Mit dem Paket DistMesh mit $h = 0,15$ und poissonv2.m erhalten wir die Finite-Elemente-Lösung, die in Abbildung 5.19 dargestellt ist. ■

5.6 Aufgaben

Aufgabe 5.1

Man berechne die (eindeutigen) Lösungen der homogenen linearen Differenzialgleichungen

$$y'(x) + y(x)e^x = 0, \quad y'(x) + y(x) \sin x = 0,$$

die die Anfangsbedingung $y(0) = 1$ erfüllen.

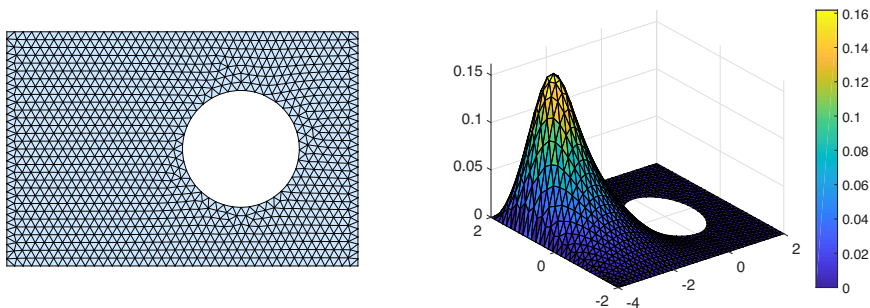


Abb. 5.19: Finite-Elemente-Lösung des homogenen Randwertproblems aus Beispiel 5.54 (rechts) für ein etwas komplizierteres Gebiet Ω mit der links gezeigten Triangulierung zu $h = 0,15$

Aufgabe 5.2

Man berechne die allgemeine Lösung der inhomogenen linearen Differenzialgleichung erster Ordnung:

$$y'(x) = \cos(3x)y(x) + \cos(3x).$$

Hinweis: Zur Berechnung des Integrals für die partikuläre Lösung substituiere man das auftretende Argument der Exponentialfunktion.

Aufgabe 5.3

Berechnen Sie die allgemeine Lösung der folgenden Differenzialgleichungen

- a) $x'(t) = x(t) + t + 4t^2$, b) $x'(t) = -tx(t) - 3t$, c) $tx'(t) = -5x(t) + \frac{e^t}{t^3}$,
 d) $x'(t) = -12t^2x(t) + t^2 + 2t^5$, e) $x'(t) = \frac{x(t)}{t^2} + \frac{4}{t^2}$.

Aufgabe 5.4

Berechnen Sie die Lösung der folgenden Anfangswertprobleme

- a) $\frac{d}{dt}x(t) = -3x(t) + 1 + t$, $x(0) = \frac{1}{3}$; b) $\frac{d}{dt}x(t) = 2t(2x(t) - 2)$, $x(0) = 0$;
 c) $\frac{d}{dt}x(t) = \frac{x(t)}{t} + t^2 - 1$, $x(1) = 1$; d) $\frac{d}{dt}x(t) = x(t) \ln t$, $x(1) = 1$.

Aufgabe 5.5

Berechnen Sie die allgemeine Lösung $x(t)$ der folgenden Differenzialgleichungen

- a) $\frac{d}{dt}x(t) = 17 \cdot x(t)$, b) $\frac{d}{dt}x(t) = 7t^2 \cdot x(t)$, c) $t^2 \frac{d}{dt}x(t) = \sqrt{x(t)}$,
 d) $\frac{d}{dt}x(t) = \cot x(t) \cdot (1 + t^2)$, e) $x(t) \cdot \frac{d}{dt}x(t) = \exp(t)$, f) $\frac{d}{dt}x(t) = a \cdot x(t)(x(t) - 2)$.

Aufgabe 5.6

Berechnen Sie die Lösung der folgenden Anfangswertprobleme:

- a) $\frac{d}{dt}x(t) = e^{x(t)+t}$, $x(0) = 0$; b) $\frac{d}{dt}x(t) = \frac{\sin t}{x^2(t)}$, $x(0) = 1$;
 c) $t^2 \cdot \frac{d}{dt}x(t) = x(t)(x(t) + 1)$, $x(1) = 1$; d) $\frac{d}{dt}x(t) = (x^2(t) - 1) \sin t$, $x(0) = 2$.

Aufgabe 5.7

Lösen Sie folgende Differenzialgleichungen erster Ordnung mit Hilfe einer geeigneten Substitution:

- a) $\frac{d}{dx}y(x) = (x + y(x) + 1)^2$, b) $x \frac{d}{dx}y(x) = y(x) + 5x$,
 c) $x^2 \frac{d}{dx}y(x) = \frac{5}{4}x^2 + y^2(x)$, d) $\frac{d}{dx}y(x) = 2 \sin\left(\frac{y(x)}{x}\right) + \frac{y(x)}{x}$.

Hinweis: Es ist $\int \frac{1}{\sin u} du = \ln \left| \tan \frac{u}{2} \right| + c$.

Aufgabe 5.8

Lösen Sie die folgenden linearen Anfangswertaufgaben:

- a) $\frac{d}{dx}y(x) = -\tan(x)y(x) + 5\sin(2x)$, $y(\pi) = 0$;
 b) $x \frac{d}{dx}y(x) = y(x) - x^2 \sin x$, $y(2\pi) = 6\pi$.

Aufgabe 5.9

Man bestimme die allgemeine Lösung des linearen, homogenen Differenzialgleichungssystems

$$\vec{y}'(t) = \begin{bmatrix} 8 & 21 \\ 1 & 4 \end{bmatrix} \vec{y}(t).$$

Aufgabe 5.10

Man berechne eine Wronski-Matrix für das lineare Differenzialgleichungssystem

$$\vec{y}' = \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix} \vec{y}.$$

Aufgabe 5.11

Man berechne ein reelles Fundamentalsystem des homogenen Systems $\frac{d}{dx}\vec{y}(x) = \mathbf{A}\vec{y}(x)$ mit

$$\mathbf{A} = \begin{bmatrix} 0 & 3 & -2 \\ -3 & 0 & 1 \\ 2 & -1 & 0 \end{bmatrix}.$$

Aufgabe 5.12

Gegeben sei das inhomogene lineare System

$$\vec{y}'(t) = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix} \vec{y}(t) + \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

- a) Berechnen Sie die allgemeine Lösung des zugehörigen homogenen Systems über Eigenwerte und Eigenvektoren.
 b) Bestimmen Sie damit eine partikuläre Lösung des inhomogenen Systems mit einem Ansatz vom Typ der rechten Seite.
 c) Wie lautet die allgemeine Lösung?
 d) Wie lautet die Lösung des Anfangswertproblems mit $y_1(0) = y_2(0) = 0$?

Aufgabe 5.13

Man bestimme die allgemeine Lösung des linearen, homogenen Systems

$$\vec{y}'(t) = \begin{bmatrix} 5 & 8 \\ 1 & 3 \end{bmatrix} \vec{y}(t).$$

Aufgabe 5.14

Man bestimme die allgemeine Lösung der Differenzialgleichung vierter Ordnung

$$y^{(4)}(t) - 7y^{(2)}(t) + 12y(t) = 0,$$

sowie die allgemeine Lösung von $y^{(2)}(t) - 4y(t) = 0$.

Aufgabe 5.15

Lösen Sie das Anfangswertproblem für das homogene System erster Ordnung

$$\vec{y}'(t) = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix} \vec{y}(t), \quad \vec{y}(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

a) durch Elimination einer Variable über eine Differenzialgleichung zweiter Ordnung,

b) durch Berechnung der Eigenwerte und Eigenvektoren.

Aufgabe 5.16

Man bestimme die Lösung der Differenzialgleichung zweiter Ordnung

$$y''(t) - 2y'(t) - 3y(t) = 0,$$

indem man die Hilfsfunktionen $y_1(t) := y(t)$ und $y_2(t) := y'(t)$ einführt und die Differenzialgleichung als System erster Ordnung schreibt und löst. Wie lautet die Lösung mit Anfangswerten $y(0) = 1$ und $y'(0) = 0$?

Aufgabe 5.17

Man löse $y''(x) + y'(x) - 2y(x) = 10e^x$ mit dem Ansatz vom Typ der rechten Seite.



6 Fourier-Reihen und Integraltransformationen

Übersicht

6.1	Fourier-Reihen	814
6.2	Fourier-Transformation	861
6.3	Laplace-Transformation	878
6.4	Diskrete Fourier-Transformation	903
6.5	Wavelets und schnelle Wavelet-Transformation *	954
6.6	Aufgaben	965

In Kapitel 1.6.5 haben wir für Spannungen der Form $u(t) = \hat{u} \cos(\omega t + \varphi_u)$ gesehen, dass das Ohm'sche Gesetz auch für Spulen und Kondensatoren gilt. Dabei rechnet man mit komplexen Widerständen (Impedanzen). Diese Widerstände hängen von der Kreisfrequenz ω ab und lassen sich nur für Spannungen angeben, die genau die angegebene Form haben. Wie kann man nun die Ströme berechnen, wenn statt einer kosinusförmigen Spannung ein anderer periodischer Spannungsverlauf vorliegt (z. B. eine Sägezahnspannung)? Uns wäre damit geholfen, wenn man diese periodische Spannung schreiben könnte als Überlagerung (Summe) von kosinusförmigen Spannungen, so dass wir für jede einzelne Spannung die Ströme bestimmen und diese dann anschließend überlagern können. Damit beschäftigen wir uns in diesem Kapitel. Die generelle Idee dabei ist, ein kompliziertes Problem in ein einfacheres zu transformieren, es in der einfacheren Form zu lösen und schließlich diese Lösung in die Lösung des Ursprungsproblems zurückzutransformieren. Möchte man Eisen verformen, so ist es auch einfacher, zunächst das Eisen zu erwärmen (Transformation), es im erwärmten Zustand zu verformen (Lösung des einfacheren Problems) und es anschließend wieder abzukühlen (Rücktransformation).

Indem man Funktionen mittels Sinus- und Kosinusausdrücken schreibt (transformiert), werden tatsächlich viele Aufgabenstellungen einfacher. Nicht nur die komplexe

Wechselstromrechnung wird auf beliebige periodische Spannungen anwendbar, auch werden Differenzialgleichungen zu algebraischen Gleichungen, in denen keine Ableitungen mehr vorkommen – und damit leicht zu lösen sind. Kompressionsverfahren für Bild- und Tondaten nutzen die Eigenschaften der Sinus- und Kosinusdarstellungen. Außerdem lassen sich Zusammenhänge in der Regelungstechnik mit Übertragungsfunktionen und Blockschaltbildern beschreiben.

Ein Ton, der z. B. von einer schwingenden Saite erzeugt werden kann, setzt sich wie in (5.41) aus Sinus- und Kosinus-Schwingungen unterschiedlicher Frequenzen mit zugehörigen Amplituden zusammen. Der Ausdruck

$$f(t) = \tilde{a}_0 + \sum_{k=1}^{\infty} [\tilde{a}_k \cos(kt) + \tilde{b}_k \sin(kt)]$$

beschreibt eine Überlagerung von Sinus- und Kosinus-Funktionen mit Amplituden $|\tilde{a}_k|$ bzw. $|\tilde{b}_k|$ zu den Kreisfrequenzen (Winkelgeschwindigkeiten) k bzw. zu den Frequenzen $\frac{k}{2\pi}$. Die Frequenzen geben die Anzahl der Schwingungen auf dem Intervall $[0, 1]$ an und werden in Hertz (Hz) gemessen. Dabei steht das Intervall für die Zeitspanne einer Sekunde.

Diese Reihendarstellung von f heißt eine **Fourier-Reihe**. Hier hat man abzählbar viele (diskrete) Kreisfrequenzen $k \in \mathbb{N}$. Außerdem hat man für $k = 0$ den Term $\tilde{a}_0 = \tilde{a}_0 \cos(0 \cdot t) + \tilde{b}_0 \sin(0 \cdot t)$. Man spricht von einem **diskreten Spektrum**. Bei einer schwingenden Saite sorgt die Randbedingung, dass sie an den Enden eingespannt ist, dafür, dass genau diese und keine anderen Frequenzen auftreten (siehe Kapitel 5.5.1).

Hat man dagegen einen nicht-periodischen Signalverlauf, der für eine unendlich große Zeitspanne betrachtet werden muss (z. B. Spannungsverlauf aus Kondensatorentladung), so kann man auch diesen unter Umständen als Überlagerung von Sinus- und Kosinus-Schwingungen auffassen. Allerdings geht man hier zu einem kontinuierlichen Spektrum über (statt $k \in \mathbb{N}$ bzw. $\mathbb{N}_0$ nun $\omega \in \mathbb{R}$). Dieser Übergang wird durch die Fourier-Transformation berechnet, die wir nach den Fourier-Reihen behandeln.

Mittels Fourier-Transformation werden wir auch Differenzialgleichungen lösen. Allerdings muss man dabei erhebliche Einschränkungen hinnehmen. Um die Einschränkungen abzumildern, betrachten wir schließlich als Weiterentwicklung der Fourier-Transformation die Laplace-Transformation. Diese eignet sich insbesondere zur Lösung von Differenzialgleichungen, die als Anfangswertproblem gestellt sind.

6.1 Fourier-Reihen

Wir wollen periodische Funktionen als Summe von Sinus- und Kosinus-Funktionen schreiben. Mit den Potenzreihen haben wir zuvor bereits etwas sehr Ähnliches getan. Wir haben dabei beliebig oft differenzierbare Funktionen als Summe von algebraischen

Monomen $(t - t_0)^k$ innerhalb eines Konvergenzradius geschrieben: $\sum_{k=0}^{\infty} \frac{f^{(k)}(t_0)}{k!} (t - t_0)^k$. Betrachtet man periodische Funktionen, so kann man diese in vergleichbarer Weise mit trigonometrischen Monomen $\cos(kt)$ und $\sin(kt)$, $k \in \mathbb{N}_0$ entwickeln. Dabei benötigt man erheblich weniger „Glattheit“ der Funktionen als bei Potenzreihen. Allerdings sind Konvergenzaussagen etwas komplizierter.

Die Einschränkung, dass es sich um periodische Funktionen handeln soll, ist in der Praxis nicht hinderlich. Hat man eine Funktion, die auf einem Intervall $[0, T]$ definiert ist, so kann man sie mittels

$$f(t + T) := f(t)$$

T -periodisch auf ganz $\mathbb{R}$ fortsetzen.

Aus Bequemlichkeit beginnen wir unsere Betrachtung mit $T = 2\pi$. Da dies die Periode von Sinus und Kosinus ist, werden die Formeln etwas einfacher. Bei einer allgemeineren Periode $T = 2p$ kommen Umrechnungsfaktoren hinzu, aber an den Aussagen ändert sich nichts (siehe Kapitel 6.1.7).

6.1.1 Fourier-Koeffizienten und Definition der Fourier-Reihe

Wir suchen zu einer integrierbaren, 2π -periodischen Funktion f die Koeffizienten a_0 , a_k und b_k , $1 \leq k \leq n$, so dass die 2π -periodische Summe

$$a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt))$$

möglichst gut mit $f(t)$ übereinstimmt.

In Abbildung 6.1 sieht man beispielsweise eine Summe $2 \sin(t) + \sin(2t) + 3 \sin(3t)$, die sich aus drei Kreisfrequenzen $k \in \{1, 2, 3\}$ mit Amplituden $b_1 = 2$, $b_2 = 1$ und $b_3 = 3$ zusammensetzt. Man beachte, dass alle Summanden bereits 2π -periodisch sind. Addiert man Sinus- und Kosinusfunktionen zu beliebigen reellen Winkelgeschwindigkeiten $\omega_1 > 0$ und $\omega_2 > 0$, so kann man nicht davon ausgehen, dass wieder eine periodische Funktion entsteht. $\sin(\omega_1 t) + \sin(\omega_2 t)$ ist periodisch, wenn beide Summanden eine gemeinsame Periode besitzen. Der erste Summand hat die Perioden $\frac{2\pi}{\omega_1} i$ für $i \in \mathbb{N}$, der zweite Summand die Perioden $\frac{2\pi}{\omega_2} m$ für $m \in \mathbb{N}$. Eine gemeinsame Periode entsteht, wenn es natürlichen Zahlen i und m gibt mit $\frac{2\pi}{\omega_1} i = \frac{2\pi}{\omega_2} m \iff \frac{\omega_1}{\omega_2} = \frac{i}{m} \in \mathbb{Q}$.

Den „Abstand“ einer Summe von Sinus- und Kosinustermen zu f kann man auf viele Weisen messen. Eingehen sollte die Differenz

$$\left| f(t) - \left(a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt)) \right) \right|$$

für jeden Wert t des Periodenintervalls $[-\pi, \pi]$. Da der Betrag über eine Fallunterscheidung definiert ist, erweist er sich beim Rechnen als sperrig. Einfacher wird es,

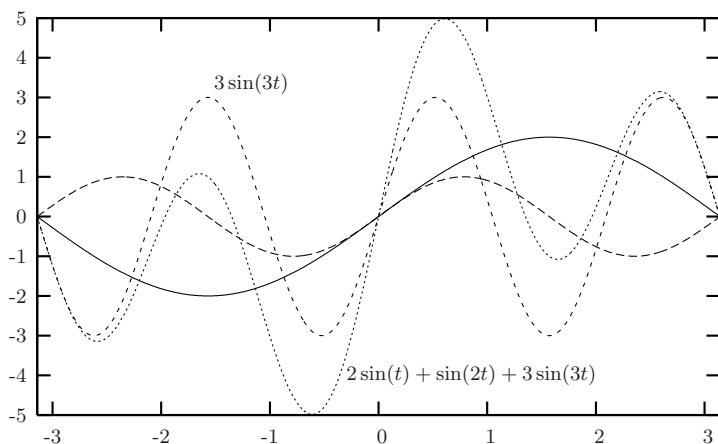


Abb. 6.1: Die Harmonischen $2 \sin(t)$, $\sin(2t)$, $3 \sin(3t)$ und ihre Überlagerung zu $2 \sin(t) + \sin(2t) + 3 \sin(3t)$

wenn wir die Differenz quadrieren. So bekommen wir für jedes t einen quadrierten Fehler. Wir messen nun die Summe der Fehler für alle Werte von t im Periodenintervall $[-\pi, \pi]$. Da es hier überabzählbar viele Werte gibt, können wir aber keine Summe $\sum_{t \in [-\pi, \pi]}$ verwenden, sondern müssen integrieren. Unsere Aufgabe besteht nun darin, die Koeffizienten so zu bestimmen, dass die Fehlerfunktion (vgl. Beispiel 3.34 auf Seite 487)

$$g(a_0, \dots, a_n, b_1, \dots, b_n) := \int_{-\pi}^{\pi} \left| f(t) - \left(a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt)) \right) \right|^2 dt \quad (6.1)$$

minimal wird. Nach Satz 4.7 ist $\text{grad } g = \vec{0}$ eine notwendige Bedingung für ein Minimum. Diese Bedingung ist für genau einen Punkt $(a_0, a_1, \dots, a_n, b_1, \dots, b_n)$ erfüllt. Man kann zeigen, dass hier tatsächlich ein lokales und sogar globales Minimum der nicht-negativwertigen Funktion g vorliegt. Wir haben bereits auf Seite 620 als Beispiel zur Bestimmung lokaler Extrema aus der Bedingung

$$\text{grad } g(a_0, a_1, \dots, a_n, b_1, b_2, \dots, b_n) = \vec{0}$$

berechnet, dass $a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) dt$ ist (siehe (4.8)). Werte für die anderen Koeffizienten a_k und b_k erhält man mit einem entsprechenden Ansatz, den wir in Kapitel 6.1.3 durchführen. Dabei stellt sich – wie für a_0 – heraus, dass die Werte unabhängig von n sind. Das ist sehr wichtig, da wir bei einer Vergrößerung von n bereits berechnete Ko-

effizienten nicht neu berechnen müssen und insbesondere den Fall $n \rightarrow \infty$ betrachten können. Für $k \in \mathbb{N}$ erhält man:

$$a_k := \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(kt) dt, \quad k \in \mathbb{N}, \quad b_k := \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(kt) dt, \quad k \in \mathbb{N}.$$

Da sowohl f als auch $\cos(kt)$ und $\sin(kt)$ die Periode 2π besitzen, kann man statt über $[-\pi, \pi]$ zu integrieren auch über jedes andere Intervall der Länge 2π integrieren (z. B. über $[0, 2\pi]$) und erhält den gleichen Wert.

Definition 6.1 (Fourier-Koeffizienten)

Sei $f : \mathbb{R} \rightarrow \mathbb{R}$ eine 2π -periodische Funktion, d. h. $f(t+k2\pi) = f(t)$ für alle $t \in \mathbb{R}, k \in \mathbb{Z}$. Weiterhin sei f auf dem Intervall $[-\pi, \pi]$ Riemann-integrierbar. Die Zahlen

$$\begin{aligned} a_0 &:= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) dt, \\ a_k &:= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(kt) dt, \quad k \in \mathbb{N}, \\ b_k &:= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(kt) dt, \quad k \in \mathbb{N}, \end{aligned}$$

heißen **Fourier-Koeffizienten** der Funktion f . Die darüber erklärte Reihe

$$a_0 + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt))$$

heißt **Fourier-Reihe** von f . Die Partialsummen bezeichnen wir mit

$$S_n(f, t) := a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt)), \quad n \in \mathbb{N}_0.$$

Man beachte, dass in der Definition von a_0 der Faktor $\frac{1}{2\pi}$ auftritt, während bei den anderen Koeffizienten $\frac{1}{\pi}$ steht. Dieser Unterschied wird später verschwinden, wenn wir zu komplexen Fourier-Koeffizienten übergehen.

Achtung: In einigen Büchern, wie im Klassiker Zygmund (2002), wird a_0 abweichend mit dem Faktor $\frac{1}{\pi}$ statt $\frac{1}{2\pi}$ definiert, dann aber im Gegensatz zu diesem Text stets in Fourier-Reihen als $\frac{a_0}{2}$ verwendet. Inhaltlich hat dies keine Konsequenzen.

Die Berechnung der Fourier-Koeffizienten und damit die Zerlegung einer Funktion (eines Signals) in Frequenzanteile nennt man **Fourier-Analyse**. Das Zusammensetzen einzelner Frequenzbestandteile zu einer Funktion über die Fourier-Reihe heißt **Fourier-Synthese**. Das Teilgebiet der Mathematik, das sich mit diesen Themen beschäftigt, heißt die **Fourier-Analysis**.

Eine lokale Definitionsänderung der Funktion f kann zu vollständig anderen Fourier-Koeffizienten führen. Die **Wavelet-Transformation** vermeidet das, indem ihre entsprechenden Koeffizienten nicht nur von der Kreisfrequenz k sondern auch von der betrachteten Stelle t abhängen, siehe Kapitel 6.5.

Da die Fourier-Koeffizienten so gewählt sind, dass die Fourier-Partialsummen möglichst gut die Ursprungsfunktion wiedergeben, kann man hoffen, dass die Fourier-Reihe gegen die Ursprungsfunktion konvergiert, dass also für jedes $t \in \mathbb{R}$ die Zahlenfolge $s_n := S_n(f, t)$,

$$(s_n)_{n=1}^\infty = \left(a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt)) \right)_{n=1}^\infty,$$

den Grenzwert $\lim_{n \rightarrow \infty} s_n = f(t)$ hat. In vielen Fällen ist das so, aber ganz ohne Voraussetzungen an f kann man die Konvergenz nicht erwarten. Um hier weiterzukommen, müssen wir erst die Fourier-Reihen näher kennenlernen. Wir greifen den Aspekt in Kapitel 6.1.5 wieder auf, indem wir dort eine hinreichende allgemeine Bedingung für die Konvergenz formulieren. Einfacher ist die Situation im folgenden Beispiel.

Beispiel 6.1

Wir geben die Fourier-Reihe eines **trigonometrischen Polynoms**

$$f(t) := \alpha_0 + \sum_{k=1}^n (\alpha_k \cos(kt) + \beta_k \sin(kt))$$

an. Dazu können wir die Koeffizienten über die in der Definition angegebenen Integrale ausrechnen. Einfacher wird es aber, wenn wir die Eigenschaft ausnutzen, dass die Koeffizienten den quadratischen Fehler (6.1) der Fourier-Partialsummen mit f minimieren. Der Fehler wird sogar null und damit sicher minimal, wenn wir

$$a_k = \begin{cases} \alpha_k & \text{für } 0 \leq k \leq n, \\ 0 & \text{für } k > n, \end{cases} \quad b_k = \begin{cases} \beta_k & \text{für } 1 \leq k \leq n, \\ 0 & \text{für } k > n. \end{cases}$$

setzen. Die Fourier-Reihe ist hier eine endliche Summe und gleich dem trigonometrischen Polynom f . Damit ist in diesem einfachen, aber wichtigen Fall die Konvergenz der Reihe offensichtlich. Diese Erkenntnis passt dazu, dass die Potenzreihe eines (algebraischen) Polynoms mit diesem übereinstimmt (siehe Seite 384). Entsprechend spricht man auch in diesem Zusammenhang von einer **Projektion**. ■

Bei der Berechnung von Fourier-Koeffizienten kann man sich auch dann viel Arbeit ersparen, wenn kein trigonometrisches Polynom gegeben ist. Dabei nutzt man aus, dass $\sin(kt)$ eine ungerade und $\cos(kt)$ eine gerade Funktion ist (vgl. Definition 1.26). Das Integral von $-\pi$ bis π über eine ungerade Funktion ist 0, da sich positive und negative Flächen aufheben.

- Ist f eine gerade Funktion, so ist $f(t) \sin(kt)$ eine ungerade Funktion (siehe Lemma 1.6, Seite 82) und damit $b_k = 0$. Die Fourier-Reihe ist eine **Kosinus-Reihe**.
- Ist f ungerade, so ist $f(t) \cos(kt)$ ebenfalls eine ungerade Funktion (siehe Lemma 1.6) und somit $a_k = 0$. Die Fourier-Reihe ist eine **Sinus-Reihe**.

Beispiel 6.2

Wir berechnen die Fourier-Reihe der 2π -periodische Funktion f mit

$$f(t) := \begin{cases} t, & -\pi < t < \pi, \\ 0, & t = \pi. \end{cases} \quad (6.2)$$

Die Funktion f ist ungerade und daher als Sinusreihe darstellbar, also $a_k = 0$. Wir berechnen die Koeffizienten b_k mittels partieller Integration:

$$\begin{aligned} b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} t \sin(kt) dt = \frac{1}{\pi} \left[t \left(-\frac{\cos(kt)}{k} \right) \right]_{-\pi}^{\pi} - \frac{1}{\pi} \int_{-\pi}^{\pi} 1 \cdot \left(-\frac{\cos(kt)}{k} \right) dt \\ &= -\frac{\pi}{\pi k} \cos(k\pi) - \frac{-\pi}{\pi} \left(-\frac{\cos(k(-\pi))}{k} \right) + \frac{1}{k\pi} \int_{-\pi}^{\pi} \cos(kt) dt \\ &= \frac{2(-1)^{k+1}}{k} + \left[\frac{\sin(kt)}{\pi k^2} \right]_{-\pi}^{\pi} = \frac{2(-1)^{k+1}}{k}. \end{aligned}$$

Damit erhalten wir die Fourier-Reihe

$$\sum_{k=1}^{\infty} \frac{2(-1)^{k+1}}{k} \sin(kt). \quad (6.3)$$

In Abbildung 6.2 sehen wir, dass die Fourier-Partialsummen mit wachsendem n die Ursprungsfunktion immer besser annähern. Es gibt lediglich eine Ausnahme: Die Unstetigkeitsstellen bei $(2k+1)\pi$. Hier streben die Partialsummen gegen das arithmetische Mittel der einseitigen Grenzwerte von f . Diese Beobachtung wird später mit dem Satz 6.4 präzisiert.

Wir erkennen aber schon jetzt, dass hohe Frequenzen wichtig sind, um die Sprungstellen gut wiederzugeben. Sprungstellen führen zu einem unendlich großen Spektrum. Denn eine endliche Summe von Sinus- und Kosinus-Termen wäre stetig, man benötigt also eine unendliche Reihe, um Sprünge wiederzugeben. Einige Autofokus-Systeme von Fotoapparaten basieren auf dieser Beobachtung. Sie berechnen die Frequenzverteilung im Bild, indem sie die Bildpunkte als Funktionswerte auffassen und die Fourier-Koeffizienten bestimmen. Das Bild wird schärfer, wenn der Anteil der hohen Frequenzen größer wird. Denn dann werden Kanten (Sprungstellen) klarer. ■

Gelegentlich kann man eine Funktion, die zunächst weder gerade noch ungerade ist, durch Verschiebung in eine solche verwandeln. Dann kann man dafür die Fourier-Reihe ausrechnen und diese anschließend wieder zurückverschieben:

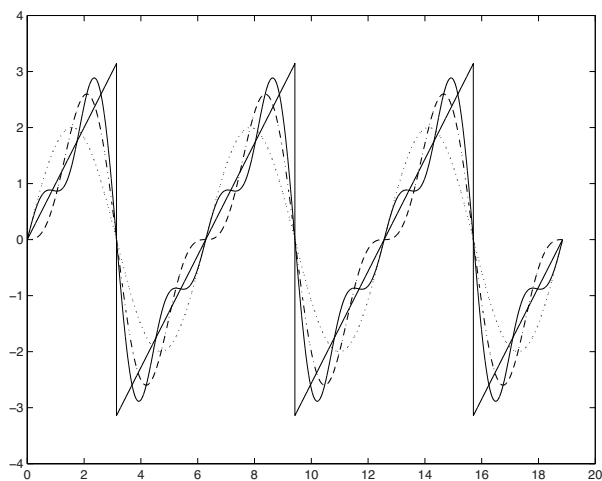


Abb. 6.2: Fourier-Summen $a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt))$ für $n = 1, 2, 3$ zur Funktion (6.2)

- Hat man eine Fourier-Summe $a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt))$ für die Funktion $f(t) + c$, so ergibt sich daraus die Fourier-Summe $(a_0 - c) + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt))$ für $f(t)$.
- Hat man eine Fourier-Summe $a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt))$ für die Funktion $f(t+c)$, so erhält man durch Einsetzen von $t-c$ mit den Additionstheoremen (1.27) und (1.28) für $f(t)$ eine Fourier-Summe

$$\begin{aligned}
 & a_0 + \sum_{k=1}^n (a_k \cos(k(t-c)) + b_k \sin(k(t-c))) \\
 &= a_0 + \sum_{k=1}^n (a_k [\cos(kt) \cos(kc) + \sin(kt) \sin(kc)] \\
 &\quad + b_k [\sin(kt) \cos(kc) - \cos(kt) \sin(kc)]) \\
 &= a_0 + \sum_{k=1}^n ([a_k \cos(kc) - b_k \sin(kc)] \cos(kt) + [a_k \sin(kc) + b_k \cos(kc)] \sin(kt)).
 \end{aligned} \tag{6.4}$$

Die komplexe Schreibweise der Fourier-Reihen, die in Kapitel 6.1.3 eingeführt wird, vereinfacht solche Rechnungen, siehe Lemma 6.1 b) auf Seite 831.

Beispiel 6.3

Bei der Kompression von Musik in das MP3-Format beruht ein wichtiger Schritt auf dem sogenannten psychoakustischen Modell. Dabei überführt man u. a. das zu komprimierende Musiksignal in eine Fourier-Reihe. Dies geschieht mit einer diskreten Fourier-Transformation, die später noch besprochen wird und mit der man Fourier-Koeffizienten näherungsweise aus endlich vielen abgetasteten Funktionswerten berechnen kann. Der dabei auszuführende numerische Algorithmus heißt schnelle Fourier-Transformation (Fast Fourier Transformation, FFT, siehe Kapitel 6.4.4). Nun kennt

man die einzelnen Frequenzen $\frac{k}{2\pi}$ (bzw. $k/(\text{Länge des Periodenintervalls})$, siehe Kapitel 6.1.7) und zugehörigen Amplituden a_k, b_k , aus denen sich das Signal zusammensetzt. Für das menschliche Ohr sind gewisse (leise) Frequenzen nicht hörbar und können ohne Klangeinbußen (in der Fourier-Reihe) weggelassen werden. Andere Frequenzen verhindern, dass man benachbarte Frequenzen bis zu einer bestimmten Lautstärke gut hören kann. Auch diese können weggelassen oder ihre Amplituden grob gerundet werden. Schließlich ist das Ohr träge und kann kurze Zeit nach lauten Frequenzen ebenfalls leisere nicht hören. Reduziert man das Ursprungssignal um all diese Komponenten, entsteht eine ausgedünnte Fourier-Reihe. Man hat weniger Koeffizienten bis zu einer vorgegebenen Grenzfrequenz und kann daher ein Musikfragment mit wenigen Daten darstellen.

Bei MP3 werden aber nicht die so reduzierten Fourier-Koeffizienten gespeichert. Vielmehr wird das Signal unter Verwendung der Ergebnisse des psychoakustischen Modells mit einer (modifizierten) diskreten Kosinus-Transformation gepackt. Dabei werden mit der diskreten Fourier-Transformation Koeffizienten einer Kosinus-Reihe berechnet (vgl. (6.16)). Die dabei vorgenommene periodische Fortsetzung der Signalstücke verhindert ungewünschte Sprünge. Die diskrete Kosinus-Transformation wird auch bei der JPEG-Bildkompression zur Transformation von 8×8 -Pixelblöcken eingesetzt. Bei der MP3- und der JPEG-Kompression werden die berechneten Koeffizienten quantisiert (in Klassen eingeteilt, vgl. Kapitel 7.1.2), so dass man weniger Speicherplatz benötigt. ■

6.1.2 Sinus- und Kosinus-Form der Fourier-Reihe

Für ein festes $k \in \mathbb{N}$ haben die Funktionen $a_k \cos(kt)$ und $b_k \sin(kt)$ die gleiche Kreisfrequenz k bzw. Frequenz $\frac{k}{2\pi}$. Damit hat auch die Summe $a_k \cos(kt) + b_k \sin(kt)$ diese Frequenz. Man bezeichnet sie auch als die k -te **Harmonische**, sie ist der Anteil der Funktion, der dem k -Fachen der Grundfrequenz entspricht. Die Summe $a_k \cos(kt) + b_k \sin(kt)$ kann durch Addition der Zeiger in einem Zeigerdiagramm konstruiert werden (siehe Kapitel 1.5.9):

$$a_k \cos(kt) + b_k \sin(kt) = \sqrt{a_k^2 + b_k^2} \sin(kt + \varphi_k), \quad (6.5)$$

$$a_k \cos(kt) + b_k \sin(kt) = \sqrt{a_k^2 + b_k^2} \cos(kt - \psi_k), \quad (6.6)$$

wobei der Nullphasenwinkel φ_k der Winkel mit Gegenkathete a_k und Ankathete b_k im rechtwinkligen Dreieck mit Hypotenusenlänge $\sqrt{a_k^2 + b_k^2}$ ist. Die Darstellung

$$a_0 + \sum_{k=1}^n \sqrt{a_k^2 + b_k^2} \sin(kt + \varphi_k)$$

heißt **Sinus-Form**, und mit $\psi_k = \frac{\pi}{2} - \varphi_k$ heißt

$$a_0 + \sum_{k=1}^n \sqrt{a_k^2 + b_k^2} \cos(kt - \psi_k)$$

Kosinus-Form der Fourier-Summe. Die Zahlen $\sqrt{a_k^2 + b_k^2}$ bilden das **Amplitudenspektrum**, die Winkel φ_k das **Phasenspektrum**. $\sqrt{a_k^2 + b_k^2}$ ist die Amplitude der Harmonischen zur Kreisfrequenz k . Der Winkel φ_k verschiebt den Graphen von $\sin(kt)$ um φ_k/k nach links zum Graphen von $\sin(k(t + \varphi_k/k)) = \sin(kt + \varphi_k)$. Entsprechend verschiebt ψ_k den Graphen von $\cos(kt)$ um ψ_k/k nach rechts.

Beweis: Im Fall $a_k = b_k = 0$ sind beide Seiten in (6.5) und (6.6) gleich null und stimmen überein. Sei also $a_k \neq 0$ oder $b_k \neq 0$. Wir setzen $\cos \varphi_k = \frac{b_k}{\sqrt{a_k^2 + b_k^2}}$ und $\sin \varphi_k = \frac{a_k}{\sqrt{a_k^2 + b_k^2}}$ und erhalten:

$$\begin{aligned} & \sqrt{a_k^2 + b_k^2} \sin(kt + \varphi_k) \stackrel{(1.23)}{=} \sqrt{a_k^2 + b_k^2} [\cos(kt) \sin \varphi_k + \sin(kt) \cos \varphi_k] \\ &= \sqrt{a_k^2 + b_k^2} \left[\frac{a_k}{\sqrt{a_k^2 + b_k^2}} \cos(kt) + \frac{b_k}{\sqrt{a_k^2 + b_k^2}} \sin(kt) \right] = a_k \cos(kt) + b_k \sin(kt). \end{aligned}$$

Wegen $\cos(kt - \psi_k) = \cos(kt - \frac{\pi}{2} + \varphi_k) = \sin(kt + \varphi_k)$ gilt insbesondere auch (6.6). $\square$

Beispiel 6.4

Wir kommen nun zur Anfangsbemerkung des Kapitels zurück und betrachten wieder die komplexe Wechselstromrechnung aus Kapitel 1.6.5, die dort für Spannungen der Form $u(t) = \hat{u} \cos(\omega t + \varphi_u)$ beschrieben ist, wobei die komplexen Widerstände von ω abhängen. Hat man mehrere solcher Spannungen (Harmonischen), die sich überlagern, kann man mit diesen einzeln rechnen und erhält das Gesamtergebnis ebenfalls als Überlagerung. Lässt sich nun eine beliebige periodische Funktion als Fourier-Reihe schreiben, so ist sie eine Überlagerung von ggf. unendlich vielen dieser Harmonischen. Damit kann man ein Netzwerk ebenfalls wieder für die einzelnen Harmonischen berechnen und erhält das Ergebnis als Überlagerung (unendliche Summation) der Einzelergebnisse. Wir betrachten eine Sägezahnspannung, für die wir bereits mit (6.3) eine Fourier-Reihe ausgerechnet haben:

$$u(t) := \sum_{k=1}^{\infty} \frac{2(-1)^{k+1}}{k} \sin(kt) = \sum_{k=1}^{\infty} u_k(t),$$

wobei $u_k(t) = \frac{2(-1)^{k+1}}{k} \sin(kt) = \frac{2(-1)^{k+1}}{k} \cos\left(kt - \frac{\pi}{2}\right)$ die k -te Harmonische ist. Komplex erweitert ist diese Teilspannung

$$\underline{u}_k(t) := \frac{2(-1)^{k+1}}{k} \left[\cos\left(kt - \frac{\pi}{2}\right) + j \sin\left(kt - \frac{\pi}{2}\right) \right] = \frac{2(-1)^{k+1}}{k} \exp\left(j \left[kt - \frac{\pi}{2}\right]\right).$$

Legen wir die Sägezahnspannung an eine Reihenschaltung aus Ohm'schem Widerstand R und Kondensator mit Kapazität C an (komplexer Widerstand ist $R - \frac{j}{\omega C}$ abhängig von der Kreisfrequenz $\omega = k$), so haben wir den Strom

$$\begin{aligned} i(t) &= \sum_{k=1}^{\infty} \frac{u_k(t)}{R - \frac{j}{kC}} = \sum_{k=1}^{\infty} \frac{2(-1)^{k+1}}{k} \frac{\cos\left(kt - \frac{\pi}{2}\right) + j \sin\left(kt - \frac{\pi}{2}\right)}{R - \frac{j}{kC}} \\ &= \sum_{k=1}^{\infty} \frac{2(-1)^{k+1}}{k \left[R^2 + \frac{1}{k^2 C^2}\right]} \left[R \cos\left(kt - \frac{\pi}{2}\right) - \frac{1}{kC} \sin\left(kt - \frac{\pi}{2}\right) \right. \\ &\quad \left. + jR \sin\left(kt - \frac{\pi}{2}\right) + \frac{j}{kC} \cos\left(kt - \frac{\pi}{2}\right) \right]. \end{aligned}$$

Die Aufteilung der Summe in Real- und Imaginärteil liefert die reelle Stromstärke als Fourier-Reihe

$$\begin{aligned} i(t) &= \sum_{k=1}^{\infty} \left[\frac{2R(-1)^{k+1}}{kR^2 + \frac{1}{kC^2}} \cos\left(kt - \frac{\pi}{2}\right) - \frac{2(-1)^{k+1}}{k^2 R^2 C + \frac{1}{C}} \sin\left(kt - \frac{\pi}{2}\right) \right] \\ &= \sum_{k=1}^{\infty} \left[\frac{2R(-1)^{k+1}}{kR^2 + \frac{1}{kC^2}} \sin(kt) + \frac{2(-1)^{k+1}}{k^2 R^2 C + \frac{1}{C}} \cos(kt) \right]. \end{aligned}$$

Die Konvergenz der Summe lässt sich mit den Konvergenzkriterien für Zahlenreihen nachrechnen. Man beachte, dass der komplexe Widerstand von der jeweiligen Kreisfrequenz $\omega = k$ abhängig ist. Damit haben wir für jeden der unendlich vielen Summanden einen anderen komplexen Widerstand. Das Ohm'sche Gesetz lässt sich somit nur auf die einzelnen Summanden, nicht aber auf die vollständige Summe und damit die Ursprungsfunktion anwenden. Ohne die Darstellung als Fourier-Reihe geht es hier also nicht. Durch Berechnung der Fourier-Koeffizienten haben wir die Ursprungsspannung so transformiert, dass wir Stromstärken berechnen können. Dass Transformation Aufgabenstellungen einfacher werden lässt, werden wir auch bei der Fourier-Transformation und der Laplace-Transformation sehen. ■

6.1.3 Komplexwertige Funktionen und Fourier-Koeffizienten

6.1.3.1 Komplexwertige Fourier-Koeffizienten

Wir haben schon beim Umgang mit Wechselströmen gesehen, dass es sich mit $e^{j\varphi}$ einfacher rechnen lässt als mit $\cos(\varphi)$ und $\sin(\varphi)$, da man die Regeln der Potenzrechnung statt die komplizierten Additionstheoreme verwenden kann.

Sinus- und Kosinus-Terme zur gleichen Kreisfrequenz lassen sich wie bei den komplexen Spannungen im vorangehenden Beispiel mittels der komplexen Exponentialfunktion zusammenfassen. Dadurch entstehen komplexe Fourier-Koeffizienten, deren Betrag die halbe Amplitude ist und deren Winkel sich aus der Phasenverschiebung der entsprechenden Harmonischen ergibt.

Mit $\exp(jt) = \cos(t) + j \sin(t)$, $t \in \mathbb{R}$ können wir die Fourier-Reihen $a_0 + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt)$ von 2π -periodischen Funktionen etwas anders darstellen als

$$\sum_{k=-\infty}^{\infty} c_k \exp(jkt) := \lim_{n \rightarrow \infty} \sum_{k=-n}^n c_k \exp(jkt)$$

mit ($k \in \mathbb{N}$)

$$\begin{aligned} c_0 &= a_0, & c_{-k} &= \frac{1}{2}[a_k + jb_k], & c_k &= \frac{1}{2}[a_k - jb_k], \\ a_k &= c_{-k} + c_k = 2 \operatorname{Re}(c_k), & b_k &= j[c_k - c_{-k}] = -2 \operatorname{Im}(c_k). \end{aligned}$$

Denn fasst man die Summanden zu k und $-k$, $k \in \mathbb{N}$, zusammen, erhält man die bekannte Darstellung:

$$\begin{aligned} & c_{-k} \exp(-jkt) + c_k \exp(jkt) \\ &= \frac{1}{2}[a_k + jb_k][\cos(kt) - j \sin(kt)] + \frac{1}{2}[a_k - jb_k][\cos(kt) + j \sin(kt)] \\ &= \frac{1}{2}[a_k \cos(kt) + b_k \sin(kt) + a_k \cos(kt) + b_k \sin(kt)] \\ &\quad + \frac{j}{2}[-a_k \sin(kt) + b_k \cos(kt) + a_k \sin(kt) - b_k \cos(kt)] \\ &= a_k \cos(kt) + b_k \sin(kt). \end{aligned}$$

Trotz der komplexen Schreibweise sind (bei einer reellwertigen Funktion f) damit die hier verwendeten symmetrischen Summen rein reell, die Koeffizienten c_{-k} und c_k sind konjugiert komplex. Die komplexen Summanden ergänzen sich aufgrund der symmetrischen Summe von $-n$ bis n paarweise zu reellen. Im Gegensatz zu Integralen $\int_{-\infty}^{\infty} f(t) dt$, die über zwei separate Grenzwerte $\int_{-\infty}^{x_0} f(t) dt$ und $\int_{x_0}^{\infty} f(t) dt$ berechnet werden, muss man daher hier einen gekoppelten Grenzwert verwenden, bei dem die untere und die obere Summationsgrenze gemeinsam gegen $\pm\infty$ streben.

Beispiel 6.5

Die Fourier-Reihe der „Sägezahn“-Funktion (6.3) lautet in komplexer Schreibweise:

$$\sum_{k=1}^{\infty} \frac{2(-1)^{k+1}}{k} \sin(kt) = \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} j \frac{(-1)^k}{k} e^{jkt}, \quad (6.7)$$

denn $a_0 = 0$, und für $k > 0$ ist $a_k = j \frac{(-1)^{-k}}{-k} + j \frac{(-1)^k}{k} = 0$ sowie

$$b_k = j \left[j \frac{(-1)^k}{k} - j \frac{(-1)^{-k}}{-k} \right] = - \left[\frac{(-1)^k}{k} + \frac{(-1)^k}{k} \right] = \frac{2(-1)^{k+1}}{k}. \quad \blacksquare$$

Es ist üblich, für die (komplexen) Fourier-Koeffizienten c_k von f die Bezeichnung

$$f^\wedge(k) := c_k, \quad k \in \mathbb{Z},$$

(sprich: „ f -Dach von k “) zu verwenden. So ist auch direkt ersichtlich, zu welcher Funktion der Fourier-Koeffizient gebildet wurde. Man nennt die Gesamtheit der Fourier-Koeffizienten (also $(f^\wedge(k))_{k=-\infty}^\infty$) auch das (diskrete) **Spektrum** von f .

Die Amplitude der k -ten Harmonischen (siehe Kapitel 6.1.2) einer reellen, 2π -periodischen Funktion f berechnet sich aus $f^\wedge(k)$ für $k \in \mathbb{N}$ nun zu

$$\sqrt{a_k^2 + b_k^2} = \sqrt{[2 \operatorname{Re}(f^\wedge(k))]^2 + [2 \operatorname{Im}(f^\wedge(k))]^2} = 2|f^\wedge(k)| = |f^\wedge(k)| + |f^\wedge(-k)|. \quad (6.8)$$

Für $k = 0$ ist die Amplitude $|a_0| = |f^\wedge(0)|$. Auch die Winkel der Phasenverschiebung φ_k aus (6.5) und ψ_k aus (6.6) entdecken wir in $f^\wedge(k)$ für $k \in \mathbb{N}$ wieder: φ_k ist der Winkel des rechtwinkligen Dreiecks mit Gegenkathete a_k und Ankathete b_k . Die komplexe Zahl $a_k - jb_k$ hat damit die Polarform $a_k - jb_k = |a_k - jb_k| \exp(-j[\pi/2 - \varphi_k])$, so dass

$$\begin{aligned} f^\wedge(k) &= \frac{1}{2}[a_k - jb_k] = |f^\wedge(k)| \exp\left(j\left[\varphi_k - \frac{\pi}{2}\right]\right) = |f^\wedge(k)| \exp(-j\psi_k) \\ &= \frac{\sqrt{a_k^2 + b_k^2}}{2} \exp(-j\psi_k) = \frac{\sqrt{a_k^2 + b_k^2}}{2} \exp\left(-j\frac{\pi}{2} + j\varphi_k\right) = -j \frac{\sqrt{a_k^2 + b_k^2}}{2} \exp(j\varphi_k). \end{aligned}$$

Die komplexen Fourier-Koeffizienten setzen sich also direkt aus Amplitude und Winkel der Phasenverschiebung zur Kreisfrequenz k zusammen.

In den nächsten beiden Abschnitten begründen wir die Definition der Fourier-Koeffizienten mit mehreren Ansätzen. Dann kommen wir in Abschnitt 6.1.3.4 auf die Berechnung der komplexen Fourier-Koeffizienten $f^\wedge(k)$ zurück.

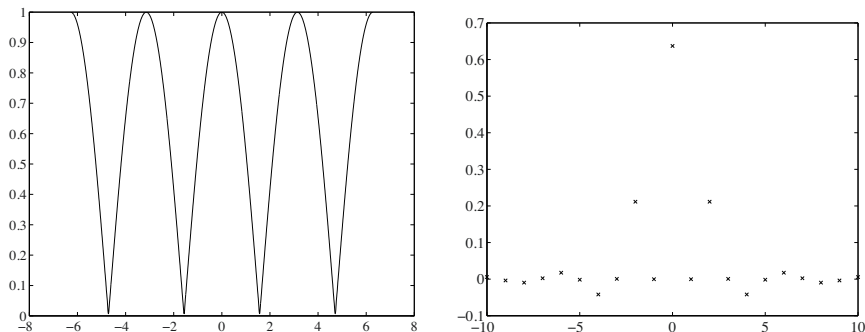


Abb. 6.3: Die 2π -periodische Funktion $|\cos t|$ (links) und die in diesem Fall reellen Fourier-Koeffizienten $f^\wedge(k)$, also das Spektrum (rechts)

6.1.3.2 Fourier-Koeffizienten als Lösung eines Minimierungsproblems

Die Fourier-Koeffizienten sind so definiert, dass der Fehler (6.1) minimal wird. Für den Koeffizienten a_0 haben wir dies schon auf Seite 620 nachgerechnet. Jetzt holen wir dies für die anderen Koeffizienten nach. Dabei nutzen wir die sogenannte **Orthogonalität** (vgl. Kapitel 6.1.3.3) der komplexwertigen Funktionen e^{jkt} aus:

$$\int_{-\pi}^{\pi} e^{jkt} e^{jlt} dt = \begin{cases} 2\pi & k+l=0 \\ 0 & k+l \neq 0. \end{cases}$$

Ist $k+l=0$, so erhalten wir

$$\int_{-\pi}^{\pi} e^{jkt} e^{jlt} dt = \int_{-\pi}^{\pi} e^0 dt = \int_{-\pi}^{\pi} 1 dt = 2\pi.$$

Für $k+l \neq 0$ ist

$$\begin{aligned} \int_{-\pi}^{\pi} e^{jkt} e^{jlt} dt &= \int_{-\pi}^{\pi} e^{j(k+l)t} dt = \int_{-\pi}^{\pi} \operatorname{Re} \left(e^{j(k+l)t} \right) dt + j \int_{-\pi}^{\pi} \operatorname{Im} \left(e^{j(k+l)t} \right) dt \\ &= \int_{-\pi}^{\pi} \cos((k+l)t) dt + j \int_{-\pi}^{\pi} \sin((k+l)t) dt \\ &= \left[\frac{1}{k+l} \sin((k+l)t) \right]_{-\pi}^{\pi} - j \left[\frac{1}{k+l} \cos((k+l)t) \right]_{-\pi}^{\pi} \\ &= 0 - 0 - \frac{j}{k+l} \cos((k+l)\pi) + \frac{j}{k+l} \cos((k+l)\pi) = 0. \end{aligned}$$

Bei dieser Rechnung haben wir Integrale von komplexwertigen Funktionen einer reellen Variable aufgeschrieben, ohne dass wir diese zuvor explizit definiert haben. Bei solchen Integralen wird der Real- und der Imaginärteil der komplexwertigen Funktion separat integriert und das Ergebnis wieder zu einer komplexen Zahl zusammengefasst (vgl. Kapitel 5.3.5.1). Da die Variable reell ist, handelt es sich um ganz normale Integrale (und nicht um Kurvenintegrale, siehe Kasten auf Seite 685). Kürzer darf man auch wie folgt rechnen (vgl. (5.21) auf Seite 763):

$$\int_{-\pi}^{\pi} e^{jkt} e^{jlt} dt = \int_{-\pi}^{\pi} e^{j(k+l)t} dt = \left[\frac{1}{j(k+l)} e^{j(k+l)t} \right]_{-\pi}^{\pi} = 0.$$

Diese Identität ist generell bei der Berechnung von Fourier-Koeffizienten sehr hilfreich. In reeller Schreibweise sieht sie so aus (vgl. (2.52) auf Seite 346):

$$\begin{aligned} \int_{-\pi}^{\pi} \cos(kt) \cdot \cos(lt) dt &= \int_{-\pi}^{\pi} \frac{e^{jkt} + e^{-jkt}}{2} \cdot \frac{e^{jlt} + e^{-jlt}}{2} dt \\ &= \frac{1}{4} \int_{-\pi}^{\pi} e^{j(k+l)t} + e^{j(k-l)t} + e^{j(-k+l)t} + e^{j(-k-l)t} dt \end{aligned}$$

$$= \begin{cases} \frac{1}{4}[2\pi + 0 + 0 + 2\pi] = \pi, & k = -l \\ \frac{1}{4}[0 + 2\pi + 2\pi + 0] = \pi, & k = l \\ 0, & k \neq \pm l. \end{cases} \quad (6.9)$$

Völlig analog gilt für den Sinus (vgl. (2.51) auf Seite 346):

$$\int_{-\pi}^{\pi} \sin(kt) \cdot \sin(lt) dt = \begin{cases} \pi, & k = l \\ -\pi, & k = -l \\ 0, & k \neq \pm l. \end{cases}$$

Daneben ist für $k, l \in \mathbb{N}_0$

$$\int_{-\pi}^{\pi} \sin(kt) \cdot \cos(lt) dt = 0, \quad (6.10)$$

da über eine ungerade Funktion integriert wird.

Nun berechnen wir die Fourier-Koeffizienten als Minimum von (6.1) mittels der eben gezeigten Orthogonalität über die notwendige Bedingung $\text{grad } g = \vec{0}$ für $1 \leq l \leq n$:

$$\begin{aligned} 0 &= \frac{\partial g}{\partial a_l}(a_0, \dots, a_n, b_1, \dots, b_n) \\ &= \frac{\partial}{\partial a_l} \int_{-\pi}^{\pi} \left[f(t) - \left(a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt)) \right) \right]^2 dt \\ &= \frac{\partial}{\partial a_l} \left[\int_{-\pi}^{\pi} \left(f(t) - a_0 - \sum_{k=1}^n b_k \sin(kt) \right)^2 dt \right. \\ &\quad \left. - 2 \int_{-\pi}^{\pi} \left(f(t) - a_0 - \sum_{k=1}^n b_k \sin(kt) \right) \left(\sum_{i=1}^n a_i \cos(it) \right) dt \right. \\ &\quad \left. + \int_{-\pi}^{\pi} \left(\sum_{k=1}^n a_k \cos(kt) \right) \left(\sum_{i=1}^n a_i \cos(it) \right) dt \right] \\ &= \frac{\partial}{\partial a_l} \left[\int_{-\pi}^{\pi} \left(f(t) - a_0 - \sum_{k=1}^n b_k \sin(kt) \right)^2 dt \right. \\ &\quad \left. - 2 \sum_{i=1}^n a_i \int_{-\pi}^{\pi} \cos(it) \left(f(t) - a_0 - \sum_{k=1}^n b_k \sin(kt) \right) dt \right. \\ &\quad \left. + \sum_{k=1}^n \sum_{i=1}^n a_k a_i \int_{-\pi}^{\pi} \cos(kt) \cos(it) dt \right]. \end{aligned}$$

Hier haben wir den quadratischen Term mit der Binomischen Formel ausmultipliziert. Das erste Integral ist unabhängig von a_l ($l \neq 0$), so dass dessen Ableitung 0 ist. Das gilt auch für die zweite Zeile für $i \neq l$, so dass deren Ableitung gleich

$$-2 \int_{-\pi}^{\pi} \cos(lt) \left(f(t) - a_0 - \sum_{k=1}^n b_k \sin(kt) \right) dt$$

$$\begin{aligned}
&= -2 \int_{-\pi}^{\pi} f(t) \cos(lt) dt + 2a_0 \int_{-\pi}^{\pi} \cos(lt) dt + 2 \sum_{k=1}^n b_k \int_{-\pi}^{\pi} \cos(lt) \sin(kt) dt \\
&\stackrel{(6.10)}{=} -2 \int_{-\pi}^{\pi} f(t) \cos(lt) dt
\end{aligned}$$

ist. Dabei haben wir die Orthogonalität von Sinus und Kosinus ausgenutzt. Sie hilft auch bei der Ableitung des verbleibenden Summanden:

$$\frac{\partial}{\partial a_l} \sum_{k=1}^n \sum_{i=1}^n a_k a_i \int_{-\pi}^{\pi} \cos(kt) \cos(it) dt \stackrel{(6.9)}{=} \frac{\partial}{\partial a_l} \sum_{k=1}^n a_k^2 \pi = 2a_l \pi.$$

Insgesamt haben wir damit die notwendige Bedingung $0 = -2 \int_{-\pi}^{\pi} f(t) \cos(lt) dt + 2a_l \pi$, so dass $a_l = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(lt) dt$ sein muss. Die Koeffizienten b_k findet man analog.

Nachdem wir jetzt 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 Ursprungsfunktion f streben, dass wir ein Integral über f schreiben dürfen als Grenzwert über Integrale der Partialsummen. Das ist nach Satz 2.58 auf Seite 409 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 $l \in \mathbb{N}_0$:

$$\begin{aligned}
\int_{-\pi}^{\pi} f(t) \cdot \cos(lt) dt &= \int_{-\pi}^{\pi} \left[a_0 + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt)) \right] \cdot \cos(lt) dt \\
&= \int_{-\pi}^{\pi} \lim_{n \rightarrow \infty} \left[a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt)) \right] \cdot \cos(lt) dt \\
&= \lim_{n \rightarrow \infty} \int_{-\pi}^{\pi} \left[a_0 + \sum_{k=1}^n (a_k \cos(kt) + b_k \sin(kt)) \right] \cdot \cos(lt) dt \\
&= \int_{-\pi}^{\pi} a_0 \cos(lt) dt + \lim_{n \rightarrow \infty} \sum_{k=1}^n \left[a_k \int_{-\pi}^{\pi} \cos(kt) \cos(lt) dt + b_k \int_{-\pi}^{\pi} \sin(kt) \cos(lt) dt \right] \\
&\stackrel{(6.10)}{=} \int_{-\pi}^{\pi} a_0 \cos(lt) dt + \lim_{n \rightarrow \infty} \sum_{k=1}^n a_k \int_{-\pi}^{\pi} \cos(kt) \cos(lt) dt.
\end{aligned}$$

Beim dritten 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 (6.9) 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 Ursprungsfunktion ergeben soll, dann müssen die Fourier-Koeffizienten genauso aussehen wie von uns definiert.

6.1.3.3 Fourier-Partialsummen als Orthogonalprojektion

Wir haben zuvor die Fourier-Koeffizienten über die Orthogonalität der Funktionen e^{jkt} bzw. der Funktionen 1 , $\cos(kt)$ und $\sin(kt)$ berechnet. Den Begriff Orthogonalität haben wir in einem anderen Zusammenhang bereits im Kapitel zur Linearen Algebra kennengelernt. Tatsächlich ist aber gar nichts anderes gemeint. Fourier-Reihen sind ein Spezialfall der allgemeinen Theorie der Linearen Algebra.

Genauer ist die Menge $\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos(kt), \frac{1}{\sqrt{\pi}} \sin(kt) : 1 \leq k \leq n \right\}$ eine Orthonormalbasis des von diesen Funktionen aufgespannten Unterraums des Vektorraums der 2π -periodischen (integrierbaren) Funktionen, wenn wir das Skalarprodukt

$$f \bullet g := \int_{-\pi}^{\pi} f(t)g(t) dt$$

verwenden (siehe Beispiel 3.21 auf Seite 485). Dass die Funktionen zueinander senkrecht stehen und normiert sind, haben wir mit (6.9)–(6.10) berechnet. Die orthogonale Projektion einer 2π -periodischen Funktion f auf den Unterraum lautet nach Satz 3.20 von Seite 496:

$$\begin{aligned} & \left(f \bullet \frac{1}{\sqrt{2\pi}} \right) \frac{1}{\sqrt{2\pi}} + \sum_{k=1}^n \left[\left(f \bullet \frac{\cos(kt)}{\sqrt{\pi}} \right) \frac{\cos(kt)}{\sqrt{\pi}} + \left(f \bullet \frac{\sin(kt)}{\sqrt{\pi}} \right) \frac{\sin(kt)}{\sqrt{\pi}} \right] \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) dt + \sum_{k=1}^n \left[\frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(kt) dt \cdot \cos(kt) + \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(kt) dt \cdot \sin(kt) \right] \\ &= a_0 + \sum_{k=1}^n [a_k \cos(kt) + b_k \sin(kt)]. \end{aligned}$$

Damit sind die Fourier-Partialsummen orthogonale Projektionen, und die Fourier-Koeffizienten bilden nach Definition 3.19 auf Seite 495 eine globale Minimalstelle der Fehlerfunktion g in (6.1). Mit wachsendem n wird die Dimension des Unterraums, auf den man projiziert, immer größer, so dass die Projektion immer besser die Ursprungsfunktion darstellt.

Damit haben wir jetzt drei (auf Orthogonalität basierende) Zugänge zu den Formeln der Fourier-Koeffizienten gefunden:

- Die Koeffizienten sind die Lösung eines Minimierungsproblems.
- Man erhält die Darstellung der Koeffizienten, falls man eine Funktion als ihre Fourier-Reihe schreiben kann und dann mit $\cos(kt)$ bzw. $\sin(kt)$ multipliziert und über $[-\pi, \pi]$ integriert.
- Die Koeffizienten entstehen durch Orthogonalprojektion.

Aus der Theorie der Orthogonalprojektionen geben wir ohne Beweis ein weiteres Ergebnis an, mit dem man z. B. prüfen kann, ob berechnete Fourier-Koeffizienten korrekt

sein können. Für 2π -periodische, integrierbare Funktionen f gilt die **Parseval'sche Gleichung**

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t)|^2 dt = \sum_{k=-\infty}^{\infty} |f^{\wedge}(k)|^2 = a_0^2 + \frac{1}{2} \sum_{k=1}^{\infty} (a_k^2 + b_k^2) \quad (6.11)$$

und damit die **Bessel'sche Ungleichung** $\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t)|^2 dt \geq \sum_{k=-n}^n |f^{\wedge}(k)|^2$.

6.1.3.4 Fourier-Koeffizienten komplexwertiger Funktionen und ihre Rechenregeln

Mit den Formeln für die reellen Fourier-Koeffizienten können wir auch die komplexen über ein Integral ausrechnen. Dabei beachte man, dass nun einheitlich der Faktor $\frac{1}{2\pi}$ auftritt (vgl. Seite 824, $k \in \mathbb{N}$):

$$\begin{aligned} c_k &= \frac{1}{2} [a_k - j b_k] = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \cos(kt) dt - j \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sin(kt) dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \exp(-jkt) dt, \\ c_{-k} &= \frac{1}{2} [a_k + j b_k] = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \cos(kt) dt + j \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sin(kt) dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \exp(jkt) dt. \end{aligned}$$

Beim Berechnen der Fourier-Koeffizienten wird mit $\exp(-jkt)$ gearbeitet, während in der Fourier-Reihe $\exp(jkt)$ verwendet wird. Das lässt sich so merken: Wir haben die Vorstellung, dass die Fourier-Reihe aus den Koeffizienten wieder die Ursprungsfunktion zurückgewinnt. Das passt zu $\exp(-jkt) \cdot \exp(jkt) = 1$.

Wir werden im Folgenden generell komplexwertige Funktionen $f : \mathbb{R} \rightarrow \mathbb{C}$, d. h. $f(t) = f_1(t) + j f_2(t)$ mit $f_1, f_2 : \mathbb{R} \rightarrow \mathbb{R}$, betrachten. So wie wir zuvor schon Real- und Imaginärteil separat integriert haben, beziehen sich Begriffe wie Stetigkeit und Differenzierbarkeit auf die Komponenten f_1 und f_2 . Insbesondere ist also für $k \in \mathbb{Z}$:

$$\begin{aligned} f^{\wedge}(k) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-jkt} dt := \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) [\cos(kt) - j \sin(kt)] dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{Re}(f(t)) \cos(kt) + \operatorname{Im}(f(t)) \sin(kt) dt \\ &\quad + j \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{Im}(f(t)) \cos(kt) - \operatorname{Re}(f(t)) \sin(kt) dt. \end{aligned}$$

Die Theorie der Fourier-Reihen gilt auch für komplexwertige, 2π -periodische Funktionen ohne Änderungen. Letztlich stellen hier Real- und Imaginärteil zwei reellwertige 2π -periodische Funktionen dar. Ist f reellwertig, so sind a_k und b_k reell. Damit sind $c_{-k} = f^\wedge(-k) = \frac{a_k}{2} + j\frac{b_k}{2}$ und $c_k = f^\wedge(k) = \frac{a_k}{2} - j\frac{b_k}{2}$ konjugiert komplex. Nimmt f jedoch auch komplexe Werte an, so sind die Koeffizienten $f^\wedge(-k)$ und $f^\wedge(k)$ in der Regel komplexe Zahlen, die nicht konjugiert komplex sind.

Lemma 6.1 (Rechenregeln für die Fourier-Koeffizienten)

Seien $f, g : \mathbb{R} \rightarrow \mathbb{C}$ auf dem Periodenintervall integrierbare, 2π -periodische Funktionen. Dann gelten die folgenden Rechenregeln für alle $k \in \mathbb{Z}$:

a) Linearität: Für Konstanten $a, b \in \mathbb{C}$ gilt: $[af(t) + bg(t)]^\wedge(k) = af^\wedge(k) + bg^\wedge(k)$.

b) Verschiebung (Translation):

i) $[f(t+h)]^\wedge(k) = e^{jh k} f^\wedge(k)$ für $h \in \mathbb{R}$.

ii) $[e^{-jht} f(t)]^\wedge(k) = f^\wedge(k+h)$ für $h \in \mathbb{Z}$.

c) Konjugation: $[\overline{f(t)}]^\wedge(k) = \overline{f^\wedge(-k)}$.

d) Ableitung: f möge zusätzlich stetig differenzierbar sein (separat für Real- und Imaginärteil):

$$(f')^\wedge(k) = jk f^\wedge(k).$$

Die Regel d) wird später im Rahmen der Fourier- und Laplace-Transformation sehr wichtig werden, da man mit ihr Ableitungen entfernen und so Differenzialgleichungen lösen kann.

Die Regel b) i) ist nicht verwunderlich, da der Winkel eines komplexen Fourier-Koeffizienten der Phasenverschiebung der Harmonischen entspricht. Mit dieser Regel schauen wir uns noch einmal die Formel (6.4) auf Seite 820 an. Dazu berechnen wir für die Funktion $f(t) := g(t-c)$ die Fourier-Koeffizienten. Dabei ist ein reelles trigonometrisches Polynom $g(t) = f(t+c) := a_0 + \sum_{k=0}^n (a_k \cos(kt) + b_k \sin(kt))$ gegeben, das die bekannten komplexen Fourier-Koeffizienten $g^\wedge(k) = \frac{1}{2}[a_k - jb_k]$ und $g^\wedge(-k) = \frac{1}{2}[a_k + jb_k]$, $0 \leq k \leq n$, besitzt. Mit b) erhalten wir

$$f^\wedge(k) = e^{-jck} g^\wedge(k).$$

Daraus ergeben sich für f die reellen Fourier-Koeffizienten

$$\begin{aligned} \tilde{a}_k &= f^\wedge(-k) + f^\wedge(k) = e^{jkc} g^\wedge(-k) + e^{-jkc} g^\wedge(k) \\ &= \frac{1}{2} \left[e^{jkc} [a_k + jb_k] + e^{-jkc} [a_k - jb_k] \right] \\ &= \frac{1}{2} [\cos(kc) + j \sin(kc)][a_k + jb_k] + [\cos(kc) - j \sin(kc)][a_k - jb_k] \\ &= a_k \cos(kc) - b_k \sin(kc) \end{aligned}$$

und analog $\tilde{b}_k = a_k \sin(kc) + b_k \cos(kc)$, so dass wir die Koeffizienten aus der letzten Zeile von (6.4) ohne Additionstheoreme erhalten.

Beweis: a) Die Linearität folgt sofort aus der Linearität des Integrals (siehe (2.43)).

b) Mit der Substitution $v = t + h$ erhalten wir (siehe (2.55) auf Seite 334):

$$\begin{aligned} (f(t+h))^\wedge(k) &= \int_{-\pi}^{\pi} f(t+h) e^{-jkt} dt = \int_{-\pi+h}^{\pi+h} f(v) e^{-jk(v-h)} dv \\ &= e^{jkh} \int_{-\pi+h}^{\pi+h} f(v) e^{-jkv} dv = e^{jkh} \int_{-\pi}^{\pi} f(v) e^{-jkv} dv = e^{jkh} f^\wedge(k), \end{aligned}$$

da der Integrand 2π -periodisch ist und damit die Integration über $[-\pi+h, \pi+h]$ das gleiche Ergebnis liefert wie über $[-\pi, \pi]$. Zur zweiten Gleichung:

$$(e^{-jht} f(t))^\wedge(k) = \int_{-\pi}^{\pi} f(t) e^{-j(k+h)t} dt = f^\wedge(k+h).$$

Hier ist zu beachten, dass h eine ganze Zahl sein muss, damit $k+h \in \mathbb{Z}$ sein kann.

c) Unter Verwendung der Regeln zur Konjugation (siehe (1.34) auf Seite 134) und der Definition des Integrals einer komplexwertigen Funktion über die Integration des Real- und Imaginärteils erhalten wir

$$\begin{aligned} [\overline{f(t)}]^\wedge(k) &= \int_{-\pi}^{\pi} \overline{f(t)} e^{-jkt} dt = \overline{\int_{-\pi}^{\pi} f(t) e^{-jkt} dt} \\ &= \overline{\int_{-\pi}^{\pi} f(t) e^{-jkt} dt} = \int_{-\pi}^{\pi} \overline{f(t) e^{-jkt}} dt = \int_{-\pi}^{\pi} \overline{f(t)} e^{jkt} dt = \overline{f^\wedge(-k)}. \end{aligned}$$

d) Mittels partieller Integration erhalten wir:

$$(f')^\wedge(k) = \int_{-\pi}^{\pi} f'(t) e^{-jkt} dt = \underbrace{\left[f(t) e^{-jkt} \right]_{-\pi}^{\pi}}_{=0} + jk \int_{-\pi}^{\pi} f(t) e^{-jkt} dt = jk f^\wedge(k).$$

Der erste Ausdruck, der bei der partiellen Integration entsteht, ist null, da die 2π -periodische Funktion $f(t) e^{-jkt}$ bei π und $-\pi$ den gleichen Funktionswert hat. $\square$

6.1.4 Faltung

Die Faltung ist eine Rechenoperation, mit der man das Verhalten von Filtern mathematisch beschreiben kann.

Beispiel 6.6

Eine (2π -periodische) Funktion f soll geglättet werden. Kleine Schwankungen im Funktionsgraphen sollen dabei mit einem Filter ausgeglichen werden. Eine Anwendung ist z. B. das Weichzeichnen eines Bildes. Eine Lösung besteht darin, jeden Funktionswert $f(t_0)$ durch eine geeignete Mittelung aller Funktionswerte in einer Umgebung von t_0 zu ersetzen. Geeignet heißt, dass die Funktionswerte gewichtet berücksichtigt werden sollen. Die Funktionswerte nahe bei t_0 sollen stärker eingehen als weiter entfernte. Wir definieren uns dazu z. B. eine (gerade) Gewichtsfunktion (Dreiecksfunktion, siehe g in Abbildung 6.5)

$$g(t) := 2\pi \begin{cases} 1 - |t| & \text{für } -1 \leq t \leq 1, \\ 0 & \text{für } 1 < |t| \leq \pi \end{cases}$$

und setzen diese 2π -periodisch fort. Der Funktionsgraph besteht aus Dreiecken mit Spitzen bei $k2\pi$, $k \in \mathbb{Z}$. $g(t_0 - t)$ ist eine Verschiebung dieser Funktion um t_0 , die Spitzen der Dreiecke liegen nun bei $t_0 + k2\pi$, $k \in \mathbb{Z}$. Wir benutzen diese Funktion, um den Funktionswert $f(t_0)$ zu mitteln, indem wir

$$\tilde{f}(t_0) := \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)g(t_0 - t) dt$$

berechnen. Wäre f konstant, so würde das Integral den Wert

$$\tilde{f}(t_0) = f(t_0) \frac{1}{2\pi} \int_{-\pi}^{\pi} g(t_0 - t) dt = f(t_0) \left[\frac{1}{2} + \frac{1}{2} \right] = f(t_0)$$

annehmen. Ist f nicht konstant, so ergibt sich eine neue, nun glattere Funktion. Diese ist das Ergebnis einer **Faltung** von f mit g . Betrachten wir konkret die 2π -periodisch fortgesetzt gedachte Funktion

$$f(t) := \begin{cases} 1 & \text{für } 0 \leq t \leq \pi, \\ 0 & \text{für } \pi < t < 2\pi. \end{cases}$$

Dann ist (siehe Abbildungen 6.4 und 6.5)

$$\tilde{f}(t_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)g(t_0 - t) dt = \begin{cases} \frac{1}{2}(1 + t_0)^2 & \text{für } -1 < t_0 \leq 0, \\ 1 - \frac{1}{2}(1 - t_0)^2 & \text{für } 0 < t_0 \leq 1, \\ 1 & \text{für } 1 < t_0 \leq \pi - 1, \\ 1 - \frac{1}{2}(\pi - 1 - t_0)^2 & \text{für } \pi - 1 < t_0 \leq \pi, \\ \frac{1}{2}(\pi + 1 - t_0)^2 & \text{für } \pi < t_0 \leq \pi + 1, \\ 0 & \text{für } \pi + 1 < t_0 \leq 2\pi - 1. \end{cases}$$

■

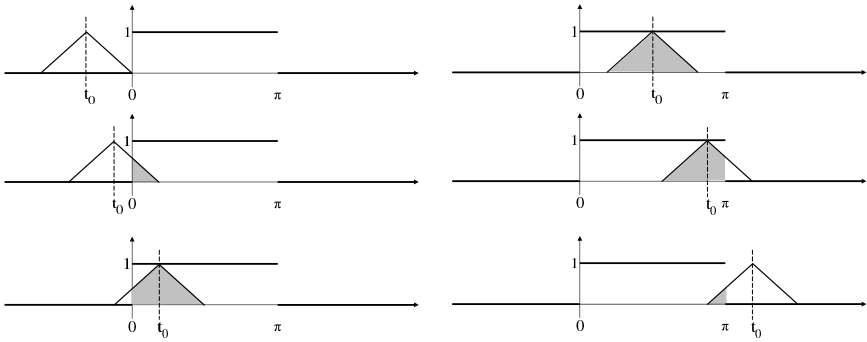


Abb. 6.4: Der Wert der Faltung eines Rechtecksignals mit einer Dreiecksfunktion an verschiedenen Stellen t_0 entspricht den eingezeichneten Flächeninhalten.

Definition 6.2 (Faltung periodischer Funktionen)
Die **Faltung** zweier 2π -periodischer, integrierbarer Funktionen f und g ist definiert als

$$[f * g](t) := \frac{1}{2\pi} \int_0^{2\pi} f(t-u)g(u) du.$$

Man beachte, dass für einen festen Wert t das Integral bezüglich der Variable u berechnet wird. Vor der Integration wird die Funktion f mittels $h(u) := f(t-u)$ an der y -Achse gespiegelt und dann um t nach rechts verschoben, bevor sie mit g multipliziert wird. Diese Spiegelung und Verschiebung wird durch den Namen Faltung (englisch: convolution) ausgedrückt.

Die Faltung ist kommutativ, d. h.

$$[f * g](t) = \frac{1}{2\pi} \int_0^{2\pi} f(t-u)g(u) du = \frac{1}{2\pi} \int_0^{2\pi} f(u)g(t-u) du = [g * f](t).$$

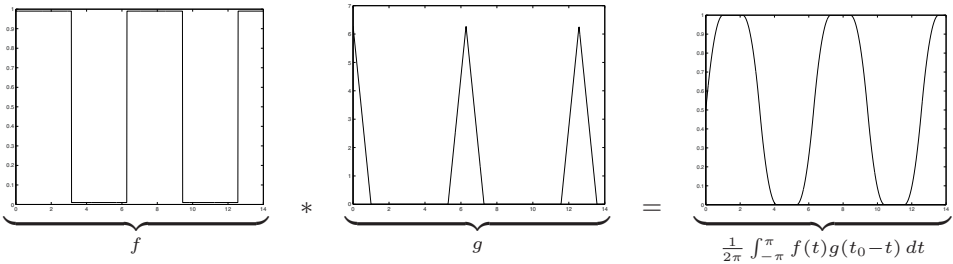


Abb. 6.5: Faltung eines Rechtecksignals mit einer Dreiecksfunktion als Kurve der Flächen aus Abbildung 6.4

Dass die beiden Integrale gleich sind, sieht man mittels Substitution $v = t - u$, $dv = -du$, unter Ausnutzung der Periodizität von f und g (siehe (2.55) auf Seite 349):

$$\begin{aligned} \int_0^{2\pi} f(t-u)g(u) du &= - \int_{t-0}^{t-2\pi} f(v)g(t-v) dv = \int_{t-2\pi}^t f(v)g(t-v) dv \\ &= \int_0^{2\pi} f(v)g(t-v) dv. \end{aligned}$$

Das Ergebnis der Faltung ist wieder eine 2π -periodische Funktion:

$$[f * g](t + 2\pi) := \frac{1}{2\pi} \int_0^{2\pi} f(t + 2\pi - u)g(u) du = \frac{1}{2\pi} \int_0^{2\pi} f(t - u)g(u) du = [f * g](t).$$

Die Funktion $f * g$ erbt die „Glattheit“ beider beteiligter Funktionen (vgl. Abbildung 6.5, in der eine unstetige und eine stetige Funktion zu einer stetigen Funktion gefaltet werden):

Lemma 6.2 (Glättungseigenschaft der Faltung)

Seien f und g 2π -periodische, auf $[0, 2\pi]$ integrierbare Funktionen.

- a) Ist f (oder alternativ g) stetig auf $\mathbb{R}$, so ist $f * g$ stetig auf $\mathbb{R}$.
 b) Ist f zusätzlich stetig differenzierbar, so ist $f * g$ stetig differenzierbar mit

$$\frac{d}{dt}[f * g](t) = [f' * g](t).$$

Beweis: a) Wir überprüfen die Stetigkeit mit der Definition. Sei dazu $\varepsilon > 0$. Wir zeigen, dass es ein $\delta > 0$ gibt, so dass $|[f * g](t + h) - [f * g](t)| < \varepsilon$ ist für alle $|h| < \delta$. Da g integrierbar ist, ist g mit einem $M > 0$ beschränkt. Da f stetig auf $[-2\pi, 2\pi]$ ist, ist f hier sogar gleichmäßig stetig (siehe Satz 2.30). Wir haben das Intervall so gewählt, dass es das Periodenintervall $[-\pi, \pi]$ überlappt. Damit ist die periodische Funktion f sogar auf $\mathbb{R}$ gleichmäßig stetig, d. h., zum vorgegebenen ε existiert ein δ , so dass $|f(t + h) - f(t)| < \frac{\varepsilon}{M}$ für alle $t \in \mathbb{R}$ und $|h| < \delta$. Damit erhalten wir für $|h| < \delta$:

$$\begin{aligned} |[f * g](t + h) - [f * g](t)| &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t + h - u) - f(t - u)| |g(u)| du \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\varepsilon}{M} M du = \varepsilon. \end{aligned}$$

b) Wir müssen hier den Grenzwert eines Differenzenquotienten berechnen, dessen Funktion über die Faltung gebildet ist. Um zum Ergebnis zu kommen, müssen wir das Integral der Faltung mit dem Grenzwert des Differenzenquotienten vertauschen. Das Vertauschen zweier Grenzwerte ist aber im Allgemeinen nicht erlaubt. Hier ist

es aufgrund eines Tricks analog zum Beweis des Satzes 4.14 doch möglich: Mit dem Satz von Fubini (Satz 4.13, Seite 654) können zwei Integrale (als zwei Grenzwerte) vertauscht werden. Diesen Satz wenden wir hier geschickt an, indem wir den Differenzenquotienten mit dem Hauptsatz (Satz 2.48, Seite 340) umschreiben.

$$\begin{aligned}
 \frac{d}{dt}[f * g](t) &= \lim_{h \rightarrow 0} \frac{[f * g](t+h) - [f * g](t)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{1}{2\pi h} \int_{-\pi}^{\pi} [f(t+h-u) - f(t-u)]g(u) du \\
 &\stackrel{\text{Satz 2.48}}{=} \lim_{h \rightarrow 0} \frac{1}{2\pi h} \int_{-\pi}^{\pi} \left[\int_0^{t+h} f'(v-u) dv - \int_0^t f'(v-u) dv \right] g(u) du \\
 &= \lim_{h \rightarrow 0} \frac{1}{2\pi h} \int_{-\pi}^{\pi} \left[\int_t^{t+h} f'(v-u) dv \right] g(u) du.
 \end{aligned}$$

$f'(v-u)g(u)$ ist nach u integrierbar, da nach Voraussetzung f' stetig (und damit integrierbar) und g integrierbar ist. Außerdem sind $\int_{-\pi}^{\pi} f'(v-u)g(u) du = [f' * g](v)$ und $\int_{-\pi}^{\pi} |f'(v-u)||g(u)| du = [|f'| * |g|](v)$ nach a) stetig und damit ebenfalls integrierbar. Damit ist Satz 4.13 anwendbar:

$$\begin{aligned}
 \frac{d}{dt}[f * g](t) &= \lim_{h \rightarrow 0} \frac{1}{2\pi h} \int_t^{t+h} \int_{-\pi}^{\pi} f'(v-u)g(u) du dv \\
 &= \lim_{h \rightarrow 0} \frac{1}{h} \int_t^{t+h} [f' * g](v) dv = [f' * g](t),
 \end{aligned}$$

wobei wir im letzten Schritt (2.49) von Seite 338 für die nach a) stetige Funktion $f' * g$ benutzt haben. $\square$

Ist f r_1 -mal und g r_2 -mal stetig differenzierbar, so ist die Faltung $f * g$ sogar $r_1 + r_2$ -mal stetig differenzierbar:

$$\begin{aligned}
 \frac{d^{r_1+r_2}}{dt^{r_1+r_2}}[f * g](t) &\stackrel{\text{Lemma 6.2b}}{=} \frac{d^{r_2}}{dt^{r_2}}[f^{(r_1)} * g](t) = \frac{d^{r_2}}{dt^{r_2}}[g * f^{(r_1)}](t) \\
 &\stackrel{\text{Lemma 6.2b}}{=} [g^{(r_2)} * f^{(r_1)}](t) = [f^{(r_1)} * g^{(r_2)}](t).
 \end{aligned}$$

Die Ableitungsregel für die Faltung ist damit viel einfacher als die entsprechende Regel für das Produkt zweier Funktionen f und g :

$$\frac{d}{dt}[f * g](t) = [f' * g](t), \quad \frac{d}{dt}[f \cdot g](t) = [f' \cdot g](t) + [f \cdot g'](t).$$

Durch Faltung mit einer glatten Hilfsfunktion g kann man die Ursprungsfunktion f wie im Eingangsbeispiel glätten. Dazu muss g so gewählt werden, dass $f * g$ näherungsweise mit f übereinstimmt. Entsprechende Funktionen g heißen **Glättungskerne**. Hier sehen wir uns dazu in Kürze den Dirichlet-Kern an. Es gibt übrigens keine Funktion g , die sich bezüglich der Faltung exakt als neutrales Element ($f * g = f$ für jede geeignete Funktion f) verhält. Dafür benötigt man das Konzept der Distributionen, das wir uns für die Fourier-Transformation im Kasten auf Seite 898 ansehen.

Die Bedeutung der Faltung für die Fourier-Koeffizienten zeigt sich im folgenden **Faltungssatz**.

Satz 6.1 (Faltungssatz)

Seien f und g integrierbare, 2π -periodische Funktionen. Nur für einen einfachen Beweis fordern wir zusätzlich, dass f (oder alternativ g) stetig auf $\mathbb{R}$ sei. Für $k \in \mathbb{Z}$ gilt:

$$(f * g)^\wedge(k) = f^\wedge(k) \cdot g^\wedge(k).$$

Der Faltungssatz ist ein ganz wesentlicher Grund, warum man sich mit Faltung beschäftigt! Die Fourier-Koeffizienten einer Faltung erhält man als Produkt der Fourier-Koeffizienten der beiden beteiligten Funktionen. Faltung wird bei der Berechnung der Fourier-Koeffizienten zur Multiplikation. Dem Produkt von Fourier-Koeffizienten entspricht die Faltung der zugehörigen Funktionen.

Beweis: Die Stetigkeit von f oder g sichert nach Lemma 6.2 a) Stetigkeit und Integrierbarkeit von $f * g$, so dass wir $(f * g)^\wedge(k)$ ausrechnen können. Dabei verwenden wir wieder den Satz 4.13 von Fubini, um die Integrationsreihenfolge zu vertauschen:

$$\begin{aligned} (f * g)^\wedge(k) &= \frac{1}{2\pi} \int_0^{2\pi} (f * g)(t) \exp(-jkt) dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left[\frac{1}{2\pi} \int_0^{2\pi} f(t-u)g(u) du \right] \exp(-jkt) dt \\ &= \left(\frac{1}{2\pi} \right)^2 \int_0^{2\pi} \int_0^{2\pi} f(t-u) \exp(-jk(t-u))g(u) \exp(-jku) du dt \\ &\stackrel{\text{Satz 4.13}}{=} \left(\frac{1}{2\pi} \right)^2 \int_0^{2\pi} \int_0^{2\pi} f(t-u) \exp(-jk(t-u))g(u) \exp(-jku) dt du \\ &= \left(\frac{1}{2\pi} \right)^2 \int_0^{2\pi} \underbrace{\left[\int_0^{2\pi} f(t-u) \exp(-jk(t-u)) dt \right]}_{= \int_0^{2\pi} f(t) \exp(-jkt) dt} g(u) \exp(-jku) du \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(t) \exp(-jkt) dt \cdot \frac{1}{2\pi} \int_0^{2\pi} g(u) \exp(-jku) du \\ &= f^\wedge(k) \cdot g^\wedge(k). \end{aligned}$$

Wir müssen noch überprüfen, dass die Voraussetzungen von Satz 4.13 tatsächlich erfüllt sind (vgl. Seite 654): $f(t-u) \exp(-jk(t-u))g(u) \exp(-jku)$ ist nach Voraussetzung ein Produkt integrierbarer Funktionen bezüglich u und damit integrierbar.

$$\begin{aligned} &\frac{1}{2\pi} \int_0^{2\pi} f(t-u) \exp(-jk(t-u))g(u) \exp(-jku) du \\ &= [f(t) \exp(-jkt)] * [g(t) \exp(-jkt)] \end{aligned}$$

und wegen $|\exp(-jk(t-u))| = |\exp(-jku)| = 1$

$$\frac{1}{2\pi} \int_0^{2\pi} |f(t-u) \exp(-jk(t-u))g(u) \exp(-jku)| du = [|f| * |g|](t)$$

sind nach Lemma 6.2 a) stetig und damit bezüglich t integrierbar. $\square$

Den Faltungssatz kann man für Filter nutzen. Möchte man gezielt Frequenzen verstärken oder ausblenden, so multipliziert man die zugehörigen Fourier-Koeffizienten mit den gewünschten Gewichten. Dieser Multiplikation im „**Frequenzbereich**“ entspricht eine Faltung im Zeitbereich, bei der man mit der Fourier-Reihe, deren Koeffizienten genau die Gewichte sind, faltet.

Wir wollen nun den Faltungssatz auf die Faltung mit einem wichtigen Glättungskern anwenden:

$$D_n(t) := \sum_{k=-n}^n e^{jkt}$$

heißt **Dirichlet-Kern**. D_n ist ein trigonometrisches Polynom, und wir können die Fourier-Koeffizienten von D_n direkt ablesen:

$$D_n^\wedge(k) = \begin{cases} 1, & \text{für } |k| \leq n, \\ 0, & \text{für } |k| > n. \end{cases} \quad (6.12)$$

Wir werden den Dirichlet-Kern benutzen, um die Konvergenz von Fourier-Reihen zu beweisen. Dazu benötigen wir eine geschlossene Darstellung:

$$D_n(t) = 1 + \sum_{k=1}^n 2 \cos(kt) = \begin{cases} \frac{\sin((n+\frac{1}{2})t)}{\sin(\frac{t}{2})} = \frac{\sin((2n+1)\frac{t}{2})}{\sin(\frac{t}{2})}, & t \neq l2\pi \\ 2n+1, & t = l2\pi \end{cases}, \quad l \in \mathbb{Z}.$$

Beweis: Zunächst erhalten wir die erste Darstellung über

$$D_n(t) = \sum_{k=-n}^n e^{jkt} = 1 + \sum_{k=1}^n [e^{jkt} + e^{-jkt}] = 1 + \sum_{k=1}^n 2 \cos(kt).$$

Die zweite ergibt sich für $t \neq l2\pi$ über die Formel zur geometrischen Summe (siehe Seite 60):

$$\begin{aligned} D_n(t) &= \sum_{k=0}^n e^{jkt} + \sum_{k=0}^n e^{-jkt} - 1 = \sum_{k=0}^n [e^{jt}]^k + \sum_{k=0}^n [e^{-jt}]^k - 1 \\ &\stackrel{(1.10)}{=} \frac{1 - [e^{jt}]^{n+1}}{1 - e^{jt}} + \underbrace{\frac{1 - [e^{-jt}]^{n+1}}{1 - e^{-jt}}}_{\frac{e^{jt} - e^{-jnt}}{e^{jt} - 1}} - 1 \end{aligned}$$

$$\begin{aligned}
&= \frac{1 - e^{j(n+1)t} - e^{jt} + e^{-jnt} - 1 + e^{jt}}{1 - e^{jt}} = \frac{-e^{j(n+1)t} + e^{-jnt}}{1 - e^{jt}} \\
&= \frac{-e^{j(n+\frac{1}{2})t} + e^{-j(n+\frac{1}{2})t}}{e^{-j\frac{t}{2}} - e^{j\frac{t}{2}}} = \frac{\frac{e^{j[-(n+\frac{1}{2})t]} - e^{-j[-(n+\frac{1}{2})t]}}{2j}}{\frac{e^{j[-\frac{t}{2}]} - e^{-j[-\frac{t}{2}]}{2j}} \\
&= \frac{\sin\left(-\left(n+\frac{1}{2}\right)t\right)}{\sin\left(-\frac{t}{2}\right)} = \frac{\sin\left(\left(n+\frac{1}{2}\right)t\right)}{\sin\left(\frac{t}{2}\right)}.
\end{aligned}$$

Für $k = l2\pi$ ist $D_n(l2\pi) = 1 + \sum_{k=1}^n 2\cos(lk2\pi) = 1 + 2n$, wobei wir anhand der Definition oder der ersten Darstellung sehen, dass der Dirichlet-Kern auch an diesen Stellen nicht nur stetig, sondern auf $\mathbb{R}$ beliebig oft stetig differenzierbar ist. $\square$

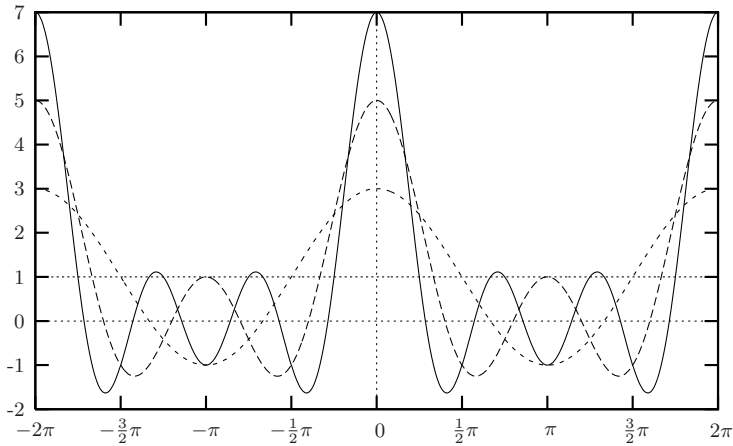


Abb. 6.6: Dirichlet-Kerne $D_0(t)$, $D_1(t)$, $D_2(t)$, $D_3(t)$ mit zunehmenden Frequenzen

Falten wir eine 2π -periodische, integrierbare Funktion f mit dem beliebig oft differenzierbaren Dirichlet-Kern D_n , so entsteht eine beliebig oft differenzierbare Funktion. Nach Satz 6.1 hat diese die Fourier-Koeffizienten

$$(f * D_n)^\wedge(k) = f^\wedge(k) \cdot D_n^\wedge(k) = \begin{cases} f^\wedge(k), & \text{für } |k| \leq n, \\ 0, & \text{für } |k| > n. \end{cases} \quad (6.13)$$

Bislang wissen wir noch nicht, unter welchen Voraussetzungen eine periodische Funktion als Grenzwert ihrer Fourier-Reihe darstellbar ist. Das ist aber bei der Funktion $f * D_n$ anders. Wir zeigen, dass $f * D_n$ exakt gleich der zugehörigen Fourier-Reihe ist, deren Fourier-Koeffizienten wir soeben berechnet haben und die mit einer Fourier-Partialsumme von f übereinstimmt:

$$[f * D_n](t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(u) \sum_{k=-n}^n e^{jk(t-u)} du$$

$$= \sum_{k=-n}^n \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} f(u) e^{-jku} du \right] e^{jkt} = \sum_{k=-n}^n \hat{f}(k) e^{jkt} = S_n(f, t).$$

An dieser Rechnung sehen wir auch, dass die Faltung mit einem trigonometrischen Polynom stets ein trigonometrisches Polynom ergibt. Hier glättet die Faltung also nicht nur, sondern führt sogar zu einer einfachen Darstellung. Wir haben eine sehr wichtige Sicht auf die Fourier-Partialsummen gewonnen, die wir bei der Untersuchung der Konvergenz von Fourier-Reihen benutzen werden:

Satz 6.2 (Partialsummen als Faltung mit Dirichlet-Kern)

$$S_n(f, t) = [f * D_n](t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t-u) \frac{\sin\left(\left[n + \frac{1}{2}\right]u\right)}{\sin\left(\frac{u}{2}\right)} du. \quad (6.14)$$

Die Fourier-Partialsummen lassen sich also als Faltung mit dem Dirichlet-Kern schreiben. Sie sind das Ergebnis einer Glättung oder Filterung mit dem Dirichlet-Kern. Alle Kreisfrequenzen $> n$ bzw. Frequenzen $> \frac{n}{2\pi}$ werden durch diese Faltung entfernt. Sie kann daher als (ideales) **Tiefpassfilter** verwendet werden. Was das bedeutet, sieht man in Abbildung 6.7: Wir fassen die Grundrisskurve einer Brücke als 2π -periodische, komplexwertige Funktion $f(t)$ auf (vgl. Kapitel 4.5.2). Während t das Intervall $[0, 2\pi]$ durchläuft, beschreibt $(x, y) = (\operatorname{Re}(f(t)), \operatorname{Im}(f(t)))$ den zugehörigen Punkt in der Ebene. Nachdem einmal der Grundriss umrundet ist, wiederholen sich die Punkte, daher ist die Funktion periodisch. Durch die Faltung mit $D_{10}(t)$ werden in der Fourier-Reihe alle Summanden zu Kreisfrequenzen $n > 10$ abgeschnitten. Es entsteht eine endliche Summe von Sinus- und Kosinus-Termen und damit eine beliebig oft differenzierbare Funktion. Entsprechend weist die dargestellte gefilterte Kurve keine Ecken mehr auf, der Grundriss wurde geglättet. Das Beispiel zeigt zudem, dass bereits in $21 = 2 \cdot 10 + 1$ Fourierkoeffizienten genügend Information steckt, um den Grundriss weitgehend zu rekonstruieren. Das nutzt man zur verlustbehafteten Datenkompression (z. B. bei JPEG) und auch zur Mustererkennung aus. Die numerische Berechnung der Fourier-Reihe wurde mit dem FFT-Algorithmus durchgeführt, siehe Kapitel 6.4.4.

6.1.5 Konvergenz von Fourier-Reihen*

Die Fourier-Koeffizienten minimieren den quadratisch über ein Integral gemittelten Fehler (6.1) auf Seite 816. Dass $\lim_{n \rightarrow \infty} \int_{-\pi}^{\pi} |f(t) - S_n(t)|^2 dt = 0$ gilt, ist damit nicht verwunderlich. Diese „Konvergenz im quadratischen Mittel“ lässt sich über die Interpretation der Fourier-Partialsummen als Orthogonalprojektion vergleichsweise einfach beweisen. Daraus kann man aber keinesfalls Rückschlüsse auf die Konvergenz der Partialsummen an einer festen Stelle $t = t_0$ ziehen.

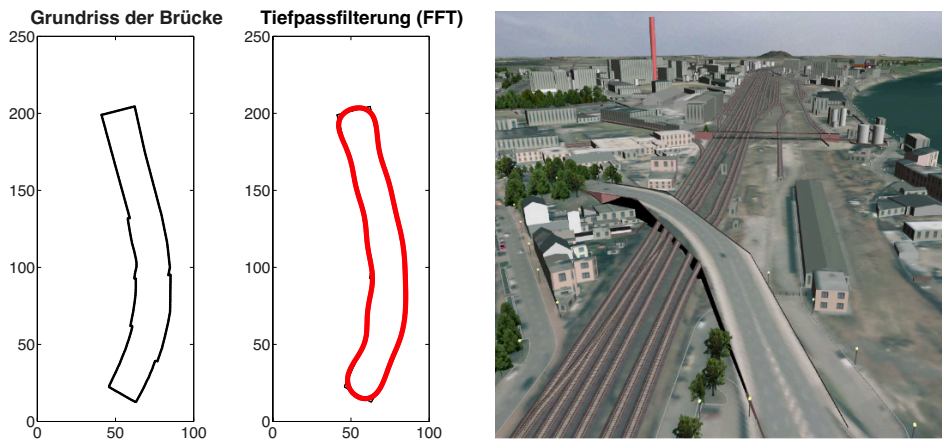


Abb. 6.7: Tiefpass-Filterung der Grundriss-Kurve einer Brücke. Im Vergleich zwischen gefilterter Kurve und Originalkurve können die relevanten vier Ecken der Brücke gefunden werden, die bei der Erstellung eines virtuellen Stadtmodells zum Einpassen der Brücke in die Landschaft benötigt werden. An diesen Ecken weichen die Kurven stark voneinander ab. Das rechte Bild zeigt das Ergebnis.

Satz 6.3 (Du Bois-Reymond 1831–1889, hier 1876)

Es existiert eine 2π -periodische, stetige (und damit insbesondere auf $[0, 2\pi]$ integrierbare) Funktion f_0 und eine streng monoton wachsende Folge $(n_k)_{k=1}^{\infty}$, $n_k \in \mathbb{N}$, so dass

$$\lim_{k \rightarrow \infty} S_{n_k}(f_0, 0) = \infty.$$

Insbesondere ist die Fourier-Reihe von f_0 im Punkt $t = 0$ nicht konvergent, obwohl f_0 sogar stetig ist.

Die Beweisidee des Satzes liefern wir im nachfolgenden Kasten. Die Stetigkeit reicht nicht aus, um auf die Konvergenz der Reihe für eine Stelle $t = t_0$ schließen zu können. Noch schwieriger wird es bei unstetigen Funktionen. Jede Fourier-Partialsumme ist als Summe beliebig oft differenzierbarer Funktionen ebenfalls beliebig oft differenzierbar, so dass man in jedem Fall eine unendliche Reihe benötigt, um der Ursprungsfunktion nahezukommen.

Entweder müssen wir die Fourier-Partialsummen so abändern, dass wir Konvergenz erzwingen (siehe Kasten auf Seite 845), oder wir benötigen eine Zusatzbedingung, die zur folgenden (etwas vereinfachten) Fassung einiger berühmter Konvergenzsätze führt. Unter der (abschwächbaren) Voraussetzung der Differenzierbarkeit in einer punktierten Umgebung des betrachteten Punktes können wir das **Riemann'sche Lokalisationsprinzip** und die darauf aufbauenden Kriterien von Dini (1845–1918) und Jordan (1838–1922) zusammenfassen.

Hintergrund: Divergenz von Fourier-Reihen

Die Aussage des Satzes von Du Bois-Reymond lässt sich mittels des Prinzips der gleichgradigen Beschränktheit (Satz 3.43 auf Seite 577) zeigen. Dazu fassen wir die Bildung der Fourier-Partialsummen als lineare Abbildung S_n ,

$$S_n(f) := \sum_{k=-n}^n f^\wedge(k) e^{ikt} = D_n * f$$

auf, wobei D_n der n -te Dirichlet-Kern ist. S_n ist eine beschränkte lineare Abbildung (siehe Kapitel 3.7) von $C_{2\pi}$ nach $C_{2\pi}$, wobei $C_{2\pi}$ der Vektorraum der 2π -periodischen stetigen Funktionen mit Norm $\|\cdot\|_{C_{2\pi}} := \|\cdot\|_{C[0,2\pi]}$ ist. Die Menge der Operatornormen $\{\|S_n\|_{[C_{2\pi}, C_{2\pi}]} : n \in \mathbb{N}\}$ ist nicht beschränkt. Das kann man über Eigenschaften des Dirichlet-Kerns zeigen. Faltungsintegrale können generell als lineare Abbildungen von $L_{2\pi}^p$ nach $L_{2\pi}^2$, $1 \leq p \leq \infty$, oder von $C_{2\pi}$ in sich aufgefasst werden. Hier bezeichnet $L_{2\pi}^p$ den Banach-Raum der (messbaren) 2π -periodischen Funktionen auf $\mathbb{R}$ als Unterraum von $L^p[0,2\pi]$ (siehe Kapitel 3.7.3). Sei $X_{2\pi} = L_{2\pi}^p$ oder $X_{2\pi} = C_{2\pi}$, dann gilt für $f \in X_{2\pi}$ und $g \in L_{2\pi}^1$, dass $f * g \in X_{2\pi}$ und

$$\|f * g\|_{X_{2\pi}} \leq \frac{1}{2\pi} \|f\|_{X_{2\pi}} \|g\|_{L^1[0,2\pi]}. \quad (6.15)$$

Diese Aussage lässt sich mit dem Satz von Fubini und der Hölder-Ungleichung zeigen, siehe z. B. (Butzer und Nessel, 1971, S. 10). Damit gilt insbesondere für die Fourier-Partialsummen:

$$\|S_n\|_{[C_{2\pi}, C_{2\pi}]} = \sup_{f \in C_{2\pi}, \|f\|_{C[0,2\pi]}=1} \|D_n * f\|_{C_{2\pi}} \leq \frac{1}{2\pi} \|D_n\|_{L^1[0,2\pi]}.$$

Tatsächlich lässt sich mit dem Riesz'schen Darstellungssatz (siehe Seite 584) zeigen, dass für stetige Kerne wie D_n und die verwendete $C_{2\pi}$ -Norm sogar die Gleichheit beider Seiten in (6.15) gilt, siehe (Butzer und Nessel, 1971, S. 54). Also ist insbesondere

$$\|S_n\|_{[C_{2\pi}, C_{2\pi}]} = \frac{1}{2\pi} \|D_n\|_{L^1[0,2\pi]},$$

und man kann die L^1 -Norm des Dirichlet-Kerns mit Konstanten $c, C \in \mathbb{R}$ abschätzen, so dass (Butzer und Nessel, 1971, S. 42)

$$\frac{4}{\pi^2} \ln(n) + c \leq \frac{1}{2\pi} \|D_n\|_{L^1[0,2\pi]} \leq \frac{4}{\pi^2} \ln(n) + C.$$

Für $n \rightarrow \infty$ streben die Operatornormen also gegen unendlich. Damit gibt es aber laut dem Prinzip der gleichgradigen Beschränktheit (Satz 3.43, Seite 577) eine stetige Funktion $f \in C_{2\pi}$, so dass $\{\|S_n(f)\|_{C_{2\pi}} : n \in \mathbb{N}\}$ nicht beschränkt ist. Es gibt also eine streng monoton wachsende Folge natürlicher Zahlen $(n_k)_{k=1}^\infty$, so dass $\lim_{k \rightarrow \infty} \|S_{n_k}(f)\|_{C_{2\pi}} = \infty$

gilt. Damit kann die Funktionenfolge $S_n(f)$ nicht gleichmäßig (also in der $C_{2\pi}$ -Norm) gegen f konvergieren:

$$\lim_{k \rightarrow \infty} \|S_{n_k}(f) - f\|_{C_{2\pi}} \geq \lim_{k \rightarrow \infty} \|S_{n_k}(f)\|_{C_{2\pi}} - \|f\|_{C_{2\pi}} = \infty.$$

Es gibt aber auch eine Funktion $f \in C_{2\pi}$, deren Fourier-Partialsummen an der festen Stelle $x = 0$ nicht gegen $f(0)$ konvergieren. Dazu benutzt man, dass die Menge der Operatornormen der Funktionele $F_n(f) := (S_n(f))(0)$ ebenfalls nicht beschränkt ist. Dann liefert der Satz sogar eine Funktion $f \in C_{2\pi}$ mit

$$\lim_{k \rightarrow \infty} \left| \sum_{l=-n_k}^{n_k} f^\wedge(l) \right| = \lim_{k \rightarrow \infty} \left| \sum_{l=-n_k}^{n_k} f^\wedge(l) e^{j l \cdot 0} \right| = \lim_{k \rightarrow \infty} |(S_{n_k}(f))(0)| = \infty.$$

Untersucht man die Konvergenz von Fourier-Reihen in den L^p -Normen, so stellt man fest, dass sie für $p = 1$ divergent sein können (da wie oben die Operatornormen nicht beschränkt sind), während sie für $1 < p < \infty$ stets in der Norm gegen die Funktion konvergieren. Hier sind die Operatornormen der Partialsummenoperatoren tatsächlich beschränkt, und es gilt nicht die Gleichheit in (6.15). Das Verhalten von beschränkten linearen Abbildungen kann also je nach Raum sehr unterschiedlich sein. Die Konvergenz in den L^p -Normen für $1 < p < \infty$ folgt aus dem Satz von Banach-Steinhaus (Folgerung 3.4 auf Seite 582): Die trigonometrischen Polynome sind Fourier-Summen, und es lässt sich (z. B. mit dem hier nicht behandelten Approximationssatz von Weierstraß für trigonometrische Polynome) zeigen, dass diese dicht in den L^p -Räumen liegen. Für trigonometrische Polynome konvergieren die Partialsummen.

Satz 6.4 (Konvergenzsatz für Fourier-Reihen)

Sei $f : \mathbb{R} \rightarrow \mathbb{C}$ eine 2π -periodische Funktion, die auf $[0, 2\pi]$ beschränkt und stetig bis auf endlich viele Unstetigkeitsstellen ist. Zu $t_0 \in \mathbb{R}$ möge eine punktierte Umgebung $U_\delta := [t_0 - \delta, t_0] \cup]t_0, t_0 + \delta]$ existieren, in der f differenzierbar (und damit insbesondere stetig) ist. Die Ableitung f' sei auf U_δ beschränkt. Dann gilt:

$$\lim_{n \rightarrow \infty} S_n(f, t_0) = \frac{1}{2} \left(\lim_{t \rightarrow t_0+} f(t) + \lim_{t \rightarrow t_0-} f(t) \right).$$

Für interessierte Leser geben wir einen Beweis des Satzes im Kasten ab Seite 851. Die Voraussetzung der Stetigkeit bis auf endlich viele Unstetigkeitsstellen lässt sich auf Integrierbarkeit abschwächen, auch die Differenzierbarkeit kann durch weniger einschränkende Eigenschaften aufgeweicht werden, die wir hier aber nicht einführen wollen (Stichwort: beschränkte Variation, siehe Definition 4.25 auf Seite 677 für $n = 1$).

Bemerkungen:

- Das Beispiel von du Bois-Reymond zeigt, dass Stetigkeit alleine für die Konvergenz der Fourier-Reihe an einer Stelle t_0 nicht ausreicht.
- Der Konvergenzsatz dagegen sagt, dass die Ursprungsfunktion nur in einer kleinen Umgebung der betrachteten Stelle t_0 „glatt“ sein soll, damit die Fourier-Reihe in t_0 konvergiert. Man muss also lokal Eigenschaften der Funktion überprüfen. Dabei handelt es sich aber in der Formulierung des Konvergenzsatzes lediglich um eine hinreichende Voraussetzung, die durch die nirgends differenzierbare Weierstraß-Funktion (siehe Seite 406) verletzt ist. Diese ist aber über eine konvergente Fourier-Reihe definiert. Das Riemann'sche Lokalisationsprinzip ist im Gegensatz zum hier angegebenen Konvergenzsatz eine notwendige und hinreichende (lokale) Bedingung für die Konvergenz einer Fourier-Reihe an einer Stelle t_0 . Allerdings ist die Bedingung sehr unhandlich und schwierig zu überprüfen. Wir werden sie als Zwischenergebnis (6.21) im Beweis des Konvergenzsatzes auf Seite 852 erhalten.
- In einer Umgebung der betrachteten Stelle soll die Funktion stetig und differenzierbar mit beschränkter Ableitung sein. Dabei darf die Stelle selbst ausgenommen werden. Aus der geforderten Beschränktheit der Ableitung kann man mit dem Mittelwertsatz (siehe Seite 315) zeigen, dass die einseitigen Grenzwerte $\lim_{t \rightarrow t_0+} f(t)$ und $\lim_{t \rightarrow t_0-} f(t)$ existieren. Damit ist f in t_0 stetig (ergänzbar) oder hat dort eine Sprungstelle.
 - Ist f stetig in t_0 , so konvergiert die Fourier-Reihe gegen den Funktionswert $f(t_0)$, der mit den einseitigen Grenzwerten übereinstimmt.
 - Hat f in t_0 eine Sprungstelle, so strebt hier die Fourier-Reihe gegen die „Mitte des Sprungs“ im Sinne des Mittelwerts der einseitigen Grenzwerte. Das liegt daran, dass die Fourier-Partialsummen stetig sind und ihr Funktionsgraph quasi einen Sprung der Funktion f überbrückt.
- Ist f eine stetig differenzierbare, 2π -periodische Funktion, so ist f' stetig auf $[0, 2\pi]$, und f' ist insbesondere beschränkt (siehe Satz 2.26). Damit konvergiert die Fourier-Reihe in jedem Punkt $t_0 \in \mathbb{R}$.

$$\lim_{n \rightarrow \infty} S_n(f, t_0) = f(t_0) \text{ für alle } t_0 \in \mathbb{R}.$$

- Die Voraussetzungen für Konvergenz gegen den Mittelwert der einseitigen Grenzwerte sind z. B. auch überall erfüllt, wenn eine periodische Funktion f stückweise zusammengesetzt ist aus endlich vielen stetig differenzierbaren Funktionen, die an den Nahtstellen ggf. nicht stetig (Sprungstellen) oder nicht differenzierbar (Knicke) zusammenpassen.
- Ohne die im Satz geforderte Differenzierbarkeit ist es für die Konvergenz gegen den Mittelwert der einseitigen Grenzwerte an jeder Stelle beispielsweise hinreichend,

wenn das Periodenintervall in endlich viele Teilintervalle zerlegt werden kann, auf denen f stetig und monoton ist. Zudem müssen an den Intervallgrenzen die einseitigen Grenzwerte existieren. Alle Unstetigkeitsstellen müssen also Sprungstellen sein.

- Ist eine 2π -periodische Funktion sogar unendlich oft differenzierbar, so können wir die Funktion sowohl als Fourier- als auch als Taylor- bzw. Potenzreihe um den Punkt $t_0 = 0$ mit Konvergenzradius ρ für $t \in]-\rho, \rho[$ darstellen:

$$f(t) = \sum_{k=0}^{\infty} \frac{f^{(k)}}{k!} x^k, \quad f(t) = \sum_{k=-\infty}^{\infty} f^{\wedge}(k) \left[e^{jt} \right]^k.$$

Wir können f also sowohl durch ein algebraisches Polynom als auch durch ein trigonometrisches Polynom annähern.

- Nach dem mathematisch sehr tiefgründigen Carleson'schen Satz von 1966 konvergiert die Fourier-Reihe einer Riemann-integrierbaren Funktion unabhängig von Voraussetzungen, wie wir sie gemacht haben, an fast jeder Stelle. Dass man nicht die Konvergenz an jeder Stelle erwarten kann, zeigt das stetige und damit integrierbare Beispiel von du Bois-Reymond. Die Formulierung „fast jede Stelle“ bedeutet wie in Satz 2.44 auf Seite 335, dass man die Menge U der Divergenzstellen durch eine Vereinigung von Intervallen mit jedem beliebig kleinen vorgegebenen Gesamthalt überdecken kann. Die Menge U kennt man aber nicht. Wegen des Beispiels von du Bois-Reymond kann sie verschieden sein von der Menge der Unstetigkeitsstellen. Das Konvergenzverhalten von Fourier-Reihen ist also ein schwieriges Thema.

Beispiel 6.7

Die 2π -periodisch fortgesetzte Sägezahn-Funktion f aus (6.2) auf Seite 819 erfüllt die Voraussetzungen von Satz 6.4, so dass die zugehörige Fourier-Reihe, die wir zuvor auf Seite 819 berechnet haben, in jedem Punkt $t \in \mathbb{R}$ gegen $f(t)$ konvergiert:

$$f(t) = \sum_{k=1}^{\infty} \frac{2(-1)^{k+1}}{k} \sin(kt).$$

Man beachte, dass die Funktionswerte an den Unstetigkeitsstellen $(2k+1)\pi$ als arithmetisches Mittel 0 der einseitigen Grenzwerte definiert sind: $\lim_{t \rightarrow 2(k+1)\pi-} f(t) = \pi$ und $\lim_{t \rightarrow 2(k+1)\pi+} f(t) = -\pi$. ■

Hintergrund: Fejér-Summen statt Fourier-Partialsummen

Nach dem Satz von Du Bois-Reymond muss die Fourier-Reihe einer stetigen Funktion nicht an jeder Stelle t_0 konvergieren. Die Fourier-Koeffizienten sind so definiert, dass die Fourier-Partialsummen einen über ein Integral definierten quadratischen Fehler mi-

nimieren. Sie minimieren keinen punktweisen Fehler $f(t_0) - S_n(f, t_0)$. Daher kommt es zu der Divergenzaussage.

Wenn wir die Partialsummen aber etwas modifizieren, können wir stetige, 2π -periodische Funktionen doch mittels trigonometrischer Polynome ohne die Zusatzvoraussetzungen des Konvergenzsatzes punktweise beliebig genau annähern, so dass wir Konvergenz für alle Stellen t_0 erhalten. Die Fourier-Reihe von f an der Stelle t_0 ist die Folge der Partialsummen

$$(S_n(f, t_0))_{n=0}^\infty = \left(\sum_{k=-n}^n f^\wedge(k) e^{jkt_0} \right)_{n=0}^\infty.$$

Stattdessen betrachten wir nun eine Folge arithmetischer Mittel der ersten n Partialsummen $\left(\frac{1}{n+1} \sum_{l=0}^n S_l(f, t_0) \right)_{n=0}^\infty$. Dabei ist

$$\frac{1}{n+1} \sum_{l=0}^n S_l(f, t_0) = \frac{1}{n+1} \sum_{l=0}^n \sum_{k=-l}^l f^\wedge(k) e^{jkt_0}.$$

Der Term $f^\wedge(0)e^{j \cdot 0 \cdot t_0}$ kommt hier in $n+1$ Summanden vor. $f^\wedge(1)e^{j \cdot 1 \cdot t_0}$ und $f^\wedge(-1)e^{j \cdot (-1) \cdot t_0}$ kommen in n Summanden vor, und allgemein gibt es $n+1-|k|$ Summanden $f^\wedge(k)e^{jkt_0}$. Daher ist

$$\frac{1}{n+1} \sum_{l=0}^n S_l(f, t_0) = \frac{1}{n+1} \sum_{k=-n}^n (n+1-|k|) f^\wedge(k) e^{jkt_0} = \sum_{k=-n}^n \left(1 - \frac{|k|}{n+1} \right) f^\wedge(k) e^{jkt_0}.$$

Man beachte, dass es sich bei dieser Folge nicht mehr um eine Reihendarstellung handelt, da für jedes n andere Glieder aufsummiert werden.

Man kann zeigen, dass durch die Mittelung sich tatsächlich eine Folge ergibt, die für stetiges f ohne Zusatzvoraussetzungen gegen $f(t_0)$ konvergiert, siehe (Natanson, 1955, S. 140).

Das Polynom $\sum_{k=-n}^n \left(1 - \frac{|k|}{n+1} \right) f^\wedge(k) e^{jkt}$ entsteht als Faltung $f * \chi_n$ von f mit dem **Fejér-Kern**, der das arithmetische Mittel der Dirichlet-Kerne ist.

$$\chi_n(t) := \frac{1}{n+1} \sum_{l=0}^n D_l(t) = \sum_{k=-n}^n \left(1 - \frac{|k|}{n+1} \right) e^{jkt}.$$

Außerdem lässt sich elementar nachrechnen, dass der Fejér-Kern die geschlossene Darstellung

$$\chi_n(t) = \frac{1}{n+1} \left(\frac{\sin\left(\frac{n+1}{2}t\right)}{\sin\left(\frac{t}{2}\right)} \right)^2$$

besitzt. Der Fejér-Kern ist ein weiteres klassisches Beispiel für einen Glättungskern. Insbesondere sehen wir die folgende Aussage:

Lemma 6.3 (Fourier-Reihen stetiger Funktionen)

Ist die Fourier-Reihe einer stetigen Funktion an einer Stelle t_0 konvergent, so ist der Grenzwert der Reihe der Funktionswert $f(t_0)$ und keine andere Zahl.

Beweis: Unter Verwendung des Fejér-Kerns ist

$$f(t) = \lim_{n \rightarrow \infty} \sum_{k=-n}^n \left(1 - \frac{|k|}{n+1}\right) f^\wedge(k) e^{jkt_0} = \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{l=0}^n S_l(f, t_0) = \lim_{l \rightarrow \infty} S_l(f, t_0),$$

da die Folge der Fourier-Partialsummen konvergent ist und nach Lemma 2.1 auf Seite 223 ihre Mittelwerte gegen den gleichen Grenzwert streben. Damit ist also $f(t_0)$ Grenzwert der Fourier-Reihe für t_0 . $\square$

Beispiel 6.8 (Kosinus-Transformation)

Sei f eine auf dem Intervall $[0, \pi]$ stetig differenzierbare Funktion (damit ist f' beschränkt). Dann können wir f auf $[0, \pi]$ als Kosinus-Reihe (ohne Nullphasenwinkel) schreiben: Wir erweitern f auf $[0, 2\pi]$ über die Spiegelung $f(\pi+t) = f(\pi-t)$, $0 < t < \pi$, und setzen dann f 2π -periodisch auf $\mathbb{R}$ fort. f ist nun so konstruiert, dass f eine gerade Funktion (ohne ungewünschte Sprungstellen, vgl. Kapitel 6.1.6) ist und die Fourier-Reihe nur Kosinus-Terme enthält. Da die fortgesetzte Funktion f stetig auf $\mathbb{R}$ und differenzierbar mit beschränkter Ableitung auf $\mathbb{R} \setminus \{k\pi, k \in \mathbb{Z}\}$ ist, konvergiert die Fourier-Reihe für jedes $t \in \mathbb{R}$ und insbesondere für $t \in [0, \pi]$:

$$f(t) = \sum_{k=0}^{\infty} a_k \cos(kt), \quad (6.16)$$

wobei

$$\begin{aligned} a_0 &= \frac{1}{2\pi} \int_0^{2\pi} f(t) dt = \frac{1}{\pi} \int_0^{\pi} f(t) dt, \\ a_k &= \frac{1}{\pi} \int_0^{2\pi} f(t) \cos(kt) dt = \frac{2}{\pi} \int_0^{\pi} f(t) \cos(kt) dt. \end{aligned}$$

Hat man nur endlich viele abgetastete Funktionswerte zur Verfügung, so erhält man die Koeffizienten a_k näherungsweise über eine diskrete Kosinus-Transformation, die sich durch Anwendung der weiter hinten beschriebenen diskreten Fourier-Transformation auf die durch Spiegelung fortgesetzte Funktion ergibt und für Verfahren der Datenkompression sehr wichtig ist. $\blacksquare$

Eine Fourier-Reihe kann nur konvergieren, wenn die Folge der Summanden hinreichend schnell gegen null konvergiert. Das hängt von der jeweiligen Funktion ab. Allerdings sorgt bereits die Bildung der Fourier-Koeffizienten dafür, dass die Folge gegen null strebt (wenn auch eventuell nicht schnell genug). Das ist die Aussage des Riemann-Lebesgue-Lemmas, das direkt aus der Konvergenz der Summe in (6.11) folgt, aber auch mit der Stetigkeit im Mittel (siehe Folgerung 6.1 im Kasten) bewiesen werden kann:

Satz 6.5 (Riemann-Lebesgue-Lemma)

Für jede integrierbare, 2π -periodische Funktion f gilt:

$$\lim_{k \rightarrow \pm\infty} f^\wedge(k) = 0. \quad (6.17)$$

Die Amplituden der hohen Kreisfrequenzen periodischer Funktionen werden zwangsläufig beliebig klein. Für numerische Zwecke kann man also bei einem genügend großen k mit der Berechnung der Fourier-Koeffizienten aufhören.

Hintergrund: Stetigkeit im Mittel

Ist f stetig an einer Stelle t , so ist $\lim_{h \rightarrow 0} |f(t) - f(t-h)| = 0$. Bei der **Stetigkeit im Mittel** (6.18) betrachtet man nicht eine Stelle t , sondern mittelt durch die Integration über alle Werte von t .

Lemma 6.4 (Stetigkeit im Mittel)

Sei $f : [a, b] \rightarrow \mathbb{C}$ auf $[a, b]$ beschränkt und stetig bis auf endliche viele Unstetigkeitsstellen $t_1, t_2, \dots, t_n$ von Real- oder Imaginärteil, dann ist

$$\lim_{h \rightarrow 0+} \int_{a+h}^b f(t) - f(t-h) dt = \lim_{h \rightarrow 0+} \int_{a+h}^b |f(t) - f(t-h)| dt = 0. \quad (6.18)$$

Eine stetige Funktion ist auch stetig im Mittel, die Umkehrung gilt aber nicht, da z. B. (6.18) auch bei Funktionen mit Sprungstellen erfüllt ist.

Beweis: Für den Nachweis des Grenzwerts sei $\varepsilon > 0$. Wir zeigen, dass es zu diesem ε ein $\delta > 0$ gibt, so dass für $0 < h < \delta$ das Integral kleiner als ε ist. Damit wir am Ende tatsächlich eine Abschätzung gegen ε ohne unschöne Faktoren bekommen, wählen wir zunächst einige Parameter recht kompliziert. Die Wahl wird verständlich, wenn man den Beweis rückwärts liest. So ist die Wahl auch entstanden.

Nach Voraussetzung ist f beschränkt auf $[a, b]$, also $|f(t)| \leq M$. Die Voraussetzung der Stetigkeit nutzen wir nur in einiger Entfernung zu den Unstetigkeitsstellen aus, da wir dort außer der Beschränktheit nichts von f wissen. Die Funktion f ist stetig auf den Intervallen $[a, b] \cap [a, t_1 - \frac{\varepsilon}{8Mn}]$, $[a, b] \cap [t_1 + \frac{\varepsilon}{8Mn}, t_2 - \frac{\varepsilon}{8Mn}]$, $\dots$, $[a, b] \cap [t_n + \frac{\varepsilon}{8Mn}, b]$ und damit auf jedem dieser Intervalle sogar gleichmäßig stetig (siehe Satz 2.30). Zum vorgegebenen $\varepsilon > 0$ gibt es damit ein $0 < \delta = \delta(\varepsilon)$ (das für alle Teilintervalle durch Bildung des Minimums gemeinsam gewählt werden kann), so dass $|f(t) - f(t-h)| < \frac{\varepsilon}{2[b-a+n2M]}$ für alle $0 < h < \delta$, sofern t und $t+h$ gemeinsam in einem der $n+1$ Teilintervalle liegen. Für $0 < h < \min \left\{ \delta, \frac{\varepsilon}{2[b-a+n2M]} \right\}$ erhalten wir:

$$\left| \int_{a+h}^b f(t) - f(t-h) dt \right| \leq \int_{a+h}^b |f(t) - f(t-h)| dt$$

$$\begin{aligned}
&\leq \int_{[a+h, b] \setminus [\bigcup_{k=1}^n [t_k - \frac{\varepsilon}{8Mn}, t_k + \frac{\varepsilon}{8Mn} + h]]} |f(t) - f(t-h)| dt + n \left[h + 2 \frac{\varepsilon}{8Mn} \right] 2M \\
&\leq [b - (a+h)] \frac{\varepsilon}{2[b-a+n2M]} + n2M \frac{\varepsilon}{2[b-a+n2M]} + \frac{\varepsilon}{2} \\
&\leq [b-a+n2M] \frac{\varepsilon}{2[b-a+n2M]} + \frac{\varepsilon}{2} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\end{aligned}$$

wobei jedes Integral von $|f(t) - f(t-h)| \leq 2M$ über jedem der n Intervalle $[a+h, b] \cap [t_k - \frac{\varepsilon}{8Mn}, t_k + \frac{\varepsilon}{8Mn} + h]$ abgeschätzt wurde durch $[h + 2 \frac{\varepsilon}{8Mn}] 2M$, so dass wir $n [h + 2 \frac{\varepsilon}{8Mn}] 2M$ addiert haben. Damit ist der Grenzwert 0 nachgerechnet. $\square$

Die Aussage des Lemmas bleibt gültig, wenn man statt der stückweisen Stetigkeit nur die Riemann-Integrierbarkeit der Funktion f fordert (vgl. Satz 2.44 auf Seite 335).

Mit der Stetigkeit im Mittel erhalten wir direkt eine Aussage für hohe Frequenzen, die ebenfalls auch für integrierbare Funktionen f gilt:

Folgerung 6.1 (Riemann-Lebesgue-Typ-Lemma)

Sei $f : [a, b] \rightarrow \mathbb{C}$ beschränkt und stetig bis auf endlich viele Unstetigkeitsstellen von Real- und Imaginärteil. Dann ist

$$\lim_{x \rightarrow \infty} \int_a^b f(t) \sin(xt) dt = 0. \quad (6.19)$$

Beweis: Wir erzeugen eine Differenz von Funktionswerten wie in (6.18), indem wir geschickt substituieren:

$$\int_a^b f(t) \sin(xt) dt = - \int_a^b f(t) \sin \left(x \left[t + \frac{\pi}{x} \right] \right) dt = - \int_{a+\frac{\pi}{x}}^{b+\frac{\pi}{x}} f \left(u - \frac{\pi}{x} \right) \sin(xu) du.$$

Hier haben wir die Substitution $u = t + \frac{\pi}{x}$ vorgenommen (siehe (2.55)). Damit erhalten wir die Differenz der Funktionswerte:

$$\begin{aligned}
&\left| \int_a^b f(t) \sin(xt) dt \right| = \frac{1}{2} \left| \int_a^b f(t) \sin(xt) dt - \int_{a+\frac{\pi}{x}}^{b+\frac{\pi}{x}} f \left(t - \frac{\pi}{x} \right) \sin(xt) dt \right| \\
&= \frac{1}{2} \left| \int_a^{a+\frac{\pi}{x}} f(t) \sin(xt) dt + \int_{a+\frac{\pi}{x}}^b [f(t) - f(t - \frac{\pi}{x})] \sin(xt) dt + \int_b^{b+\frac{\pi}{x}} f(t - \frac{\pi}{x}) \sin(xt) dt \right| \\
&\leq \frac{1}{2} \left[\int_a^{a+\frac{\pi}{x}} |f(t)| dt + \left| \int_{a+\frac{\pi}{x}}^b [f(t) - f(t - \frac{\pi}{x})] \sin(xt) dt \right| + \int_b^{b+\frac{\pi}{x}} |f(t - \frac{\pi}{x})| dt \right] \\
&\leq \frac{1}{2} \left[\frac{\pi}{x} M + \int_{a+\frac{\pi}{x}}^b |f(t) - f(t - \frac{\pi}{x})| dt + \frac{\pi}{x} M \right].
\end{aligned}$$

Dabei haben wir ausgenutzt, dass f auf $[a, b]$ mit einer Zahl $M > 0$ beschränkt ist. Für $x \rightarrow \infty$ konvergiert $\frac{\pi}{x}M$ gegen null. Für das verbleibende Integral nutzen wir die Stetigkeit im Mittel (6.18) mit $h = \frac{\pi}{x}$, so dass auch dieser Term gegen 0 strebt. $\square$

Auch diese Aussage gilt für Riemann-integrierbare Funktionen und auch, wenn man den Sinus durch einen Kosinus oder durch $\exp(-jxt)$ ersetzt. In dieser Form ergibt sich für das Intervall $[0, 2\pi]$ das Riemann-Lebesgue-Lemma (Satz 6.5).

Beispiel 6.9

Die Sägezahnfunktion f mit der Fourier-Reihe (6.7) auf Seite 824 hat für $k \neq 0$ die Fourier-Koeffizienten $f^\wedge(k) = j \frac{(-1)^k}{k}$. Offensichtlich ist sowohl $\lim_{k \rightarrow \infty} f^\wedge(k) = \lim_{k \rightarrow \infty} j \frac{(-1)^k}{k} = 0$ als auch $\lim_{k \rightarrow -\infty} f^\wedge(k) = 0$. $\blacksquare$

Man kann sich nun fragen, wie „schnell“ eine Fourier-Reihe konvergiert. Eine erste einfache Fehlerabschätzung ergibt sich direkt aus den nicht verwendeten Fourier-Koeffizienten. Lässt sich f für ein $t_0 \in \mathbb{R}$ als Grenzwert der Fourier-Reihe schreiben (z. B. bei Voraussetzungen wie in Satz 6.4), so gilt für den Fehler:

$$\begin{aligned} |f(t_0) - S_n(f, t_0)| &= \left| \sum_{k=-\infty}^{\infty} f^\wedge(k) e^{jkt_0} - \sum_{k=-n}^n f^\wedge(k) e^{jkt_0} \right| = \left| \sum_{|k|>n} f^\wedge(k) e^{jkt_0} \right| \\ &\leq \sum_{|k|>n} |f^\wedge(k)| \cdot |e^{jkt_0}| = \sum_{|k|>n} |f^\wedge(k)|. \end{aligned}$$

Für die Sägezahnfunktion im vorangehenden Beispiel ist die Schranke bereits zu grob, da rechts eine divergente harmonische Reihe entsteht. Wegen ihrer Unstetigkeitsstellen kann man aber auch keine Fehlerschranke unabhängig von der betrachteten Stelle t_0 erwarten (siehe Kapitel 6.1.6).

In Verschärfung zu (6.17) kann man zeigen, dass die Fourier-Koeffizienten glatter Funktionen schneller gegen null streben. Genauer gilt für eine r -mal stetig differenzierbare 2π -periodische Funktion f , dass

$$|f^\wedge(k)| \leq C_r \frac{1}{|k|^r}$$

ist, wobei die Konstante C_r nur von f und r aber nicht von k und t_0 abhängt. Damit lässt sich für $r > 1$ eine Fehlerschranke herleiten, wie wir sie vom Restglied der Taylor-Entwicklung kennen:

$$\begin{aligned} |f(t_0) - S_n(f, t_0)| &\leq C_r \sum_{|k|>n} \frac{1}{|k|^r} = 2C_r \sum_{k=n+1}^{\infty} \frac{1}{k^r} \leq 2C_r \sum_{k=n+1}^{\infty} \int_{k-1}^k \frac{1}{t^r} dt \\ &= 2C_r \int_n^{\infty} \frac{1}{t^r} dt = 2C_r \lim_{u \rightarrow \infty} \frac{1}{1-r} t^{1-r} \Big|_n^u = \frac{2C_r}{r-1} \frac{1}{n^{r-1}}. \end{aligned}$$

Der Fehler strebt also bis auf einen konstanten Faktor mindestens so schnell gegen null wie die Folge $(n^{-r+1})_{n=1}^{\infty}$. Diese Abschätzung lässt sich unter den gleichen Voraussetzungen für $n > 1$ noch weiter verschärfen zu (Natanson, 1955, S. 136)

$$|f(t_0) - S_n(f, t_0)| \leq C_r \frac{\ln n}{n^r} \quad (6.20)$$

für jedes $t_0 \in \mathbb{R}$. Hier ist die rechte Seite für große Werte von n kleiner als zuvor, der Fehler wird genauer abgeschätzt, die Folge $(\frac{\ln n}{n^r})_{n=2}^{\infty}$ strebt schneller gegen null als $(n^{-r+1})_{n=1}^{\infty}$.

Für Funktionen, die nur stetig sind, muss die Fourier-Reihe gar nicht konvergieren (Satz 6.20), so dass auch eine Abschätzung wie (6.20), die sogar gleichmäßige Konvergenz (Konvergenz unabhängig von der Stelle t_0 , siehe Definition 2.43 auf Seite 404) bedeutet, nicht gelten kann. Hat eine Funktion eine Sprungstelle, d. h., ist sie nicht einmal stetig, so kann man erst recht keine gleichmäßige Konvergenz erwarten. Damit beschäftigt sich das Gibbs-Phänomen, mit dem es nach dem Einschub des Beweises von Satz 6.4 weitergeht.

Hintergrund: Beweis des Konvergenzsatzes (Satz 6.4) für Fourier-Reihen

Der Beweis ist etwas technisch und nur für die Leser gedacht, die ein vertieftes Verständnis für das Funktionieren von Fourier-Reihen gewinnen möchten. Im Wesentlichen basiert er darauf, dass man unter geeigneten Voraussetzungen durch Faltung von f mit einem Dirichlet-Kern einen Funktionswert $f(t) = f(t + k2\pi)$ in guter Näherung erhalten kann. Durch die Faltung wird der Dirichlet-Kern so verschoben, dass er an den Stellen $t + k2\pi$ die größten Funktionswerte hat. Integriert man nun das Produkt des verschobenen Kerns und der Funktion f , d. h. berechnet man die Faltung, so gewinnt man durch die Gewichtung mit dem Dirichlet-Kern ungefähr den Funktionswert $f(t)$. Da die Faltung mit einem Dirichlet-Kern einer Fourier-Partialsumme entspricht, lässt sich so deren Konvergenz gegen $f(t)$ zeigen.

Beweis: Wir stellen die Fourier-Partialsummen als Faltung mit dem Dirichlet-Kern dar (siehe (6.14)) und vergleichen mit dem Wert $c := \frac{1}{2} (\lim_{t \rightarrow t_0+} f(t) + \lim_{t \rightarrow t_0-} f(t))$. Dies führt zu Integralen, die wegen der Stetigkeit im Mittel (siehe Folgerung 6.1) und aufgrund der Voraussetzungen gegen 0 konvergieren.

Da $1 = D_n^\wedge(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} D_n(t) dt$ (siehe (6.12)) und $S_n(f, t_0) = (f * D_n)(t_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t_0 - t) D_n(t) dt$ gilt (siehe (6.14)), ist

$$\begin{aligned} S_n(f, t_0) - c &= \frac{1}{2\pi} \int_{-\pi}^{\pi} [f(t_0 - t) - c] D_n(t) dt \\ &= \frac{1}{2\pi} \left[\int_{-\pi}^0 [f(t_0 - t) - c] D_n(t) dt + \int_0^{\pi} [f(t_0 - t) - c] D_n(t) dt \right] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2\pi} \left[\int_0^\pi [f(t_0 + t) - c] D_n(-t) dt + \int_0^\pi [f(t_0 - t) - c] D_n(t) dt \right] \\
&= \frac{1}{2\pi} \int_0^\pi \underbrace{[f(t_0 + t) + f(t_0 - t) - 2c]}_{=: g(t)} \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{\sin\left(\frac{t}{2}\right)} dt. \tag{6.21}
\end{aligned}$$

Im letzten Schritt haben wir ausgenutzt, dass $D_n(t) = \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{\sin\left(\frac{t}{2}\right)}$ eine gerade Funktion ist, also $D_n(t) = D_n(-t)$.

Die Darstellung (6.21) der Fourier-Reihe an der Stelle t_0 ist das eigentliche Riemann'sche Lokalisationsprinzip. Strebt (6.21) für $n \rightarrow \infty$ gegen 0, so konvergiert die Fourier-Reihe an der Stelle t_0 gegen c und umgekehrt.

Jetzt nutzen wir die weiteren Voraussetzungen aus, um damit zu zeigen, dass (6.21) tatsächlich gegen 0 konvergiert und um so ein hinreichendes Kriterium für die Konvergenz zu gewinnen. Um dabei die punktierte Umgebung $[t_0 - \delta, t_0] \cup [t_0, t_0 + \delta]$ ins Spiel zu bringen, zerlegen wir das Integrationsintervall $[0, \pi]$ in die Intervalle $[0, \delta]$ und $[\delta, \pi]$ und ergänzen geschickt $0 = g(t) \left[\frac{\sin\left(\frac{(2n+1)t}{2}\right)}{\frac{t}{2}} - \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{\frac{t}{2}} \right]$:

$$\begin{aligned}
&\frac{1}{2\pi} \int_0^\pi g(t) \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{\sin\left(\frac{t}{2}\right)} dt \\
&= \frac{1}{2\pi} \int_0^\delta g(t) \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{\frac{t}{2}} dt + \frac{1}{2\pi} \int_0^\delta g(t) \left[\frac{1}{\sin\left(\frac{t}{2}\right)} - \frac{1}{\frac{t}{2}} \right] \sin\left(\frac{(2n+1)t}{2}\right) dt \\
&\quad + \frac{1}{2\pi} \int_\delta^\pi g(t) \frac{1}{\sin\left(\frac{t}{2}\right)} \sin\left(\frac{(2n+1)t}{2}\right) dt =: I_1(n) + I_2(n) + I_3(n).
\end{aligned}$$

Mit Folgerung 6.1 zeigen wir jetzt, dass $\lim_{n \rightarrow \infty} I_2(n) = \lim_{n \rightarrow \infty} I_3(n) = 0$ ist. Die Funktion g ist nach Voraussetzung stetig auf $]0, \delta]$, und da f höchstens eine Sprungstelle bei t_0 besitzt, existiert der Grenzwert $\lim_{t \rightarrow 0+} g(t)$, und g ist in 0 stetig ergänzbar und auf $[0, \delta]$ beschränkt. Die Funktion $\frac{1}{\sin\left(\frac{t}{2}\right)} - \frac{1}{\frac{t}{2}}$ ist stetig auf $]0, \delta]$ und wegen

$$\begin{aligned}
\lim_{t \rightarrow 0+} \left[\frac{1}{\sin\left(\frac{t}{2}\right)} - \frac{1}{\frac{t}{2}} \right] &= \lim_{t \rightarrow 0+} \frac{\frac{t}{2} - \sin\left(\frac{t}{2}\right)}{\frac{t}{2} \sin\left(\frac{t}{2}\right)} = \lim_{t \rightarrow 0+} \frac{\frac{1}{2} - \frac{1}{2} \cos\left(\frac{t}{2}\right)}{\frac{1}{2} \sin\left(\frac{t}{2}\right) + \frac{t}{4} \cos\left(\frac{t}{2}\right)} \\
&= \lim_{t \rightarrow 0+} \frac{\frac{1}{4} \sin\left(\frac{t}{2}\right)}{\frac{1}{4} \cos\left(\frac{t}{2}\right) + \frac{1}{4} \cos\left(\frac{t}{2}\right) - \frac{t}{8} \sin\left(\frac{t}{2}\right)} = 0
\end{aligned}$$

kann sie ebenfalls stetig auf $[0, \delta]$ fortgesetzt werden und ist insbesondere beschränkt. Dabei haben wir zweimal den Satz von L'Hospital angewendet (siehe Sätze 2.40, 2.41). Für $x = \frac{2n+1}{2}$ folgt damit aus Folgerung 6.1 $\lim_{n \rightarrow \infty} I_2(n) = 0$.

Nun zu $I_3(n)$: Auf $[\delta, \pi]$ sind sowohl $g(t)$ beschränkt und stetig bis auf endlich viele Unstetigkeitsstellen (von $f(t_0 + t)$ und von $f(t_0 - t)$) und $\frac{1}{\sin\left(\frac{t}{2}\right)}$ stetig (und beschränkt), so dass $\lim_{n \rightarrow \infty} I_3(n) = 0$ nun ebenfalls mit Folgerung 6.1 für $x = \frac{2n+1}{2}$ folgt.

Wir wissen jetzt: $\lim_{n \rightarrow \infty} S_n(f, t_0) - c = 0 \iff \lim_{n \rightarrow \infty} I_1(n) = 0$, die Fourier-Reihe konvergiert an der Stelle t_0 also gegen c genau dann, wenn das erste Integral für $n \rightarrow \infty$ gegen 0 strebt, wobei hier nur Werte von f über die Definition der Funktion g in einer δ -Umgebung von t_0 berücksichtigt werden. Deshalb spricht man vom Lokalisationsprinzip. Um nun auch noch $\lim_{n \rightarrow \infty} I_1(n) = 0$ zu zeigen, verwenden wir die zusätzlichen Voraussetzungen des Satzes. Auf $]0, \delta]$ ist

$$\begin{aligned} g(t) &= f(t_0 + t) + f(t_0 - t) - 2c \\ &= f(t_0 + t) - \lim_{h \rightarrow 0+} f(t_0 + h) + f(t_0 - t) - \lim_{h \rightarrow 0-} f(t_0 + h) = tf'(\xi_1) - tf'(\xi_2), \end{aligned}$$

wobei es nach dem Mittelwertsatz (Seite 315) für die auf $[t_0, t_0 + t]$ bzw. $[t_0 - t, t_0]$ mit den einseitigen Grenzwerten stetig fortgesetzte Funktion f entsprechende Stellen $\xi_1 \in]t_0, t_0 + t[\subset]t_0, t_0 + \delta[$ und $\xi_2 \in]t_0 - t, t_0[\subset]t_0 - \delta, t_0[$ gibt. Nach Voraussetzung ist f' auf $]t_0 - \delta, t_0[\cup]t_0, t_0 + \delta[$ beschränkt: $|f'(t)| \leq M$. Damit ist $|g(t)| \leq 2Mt$, und $\left| \frac{g(t)}{\frac{t}{2}} \right|$ ist beschränkt auf $[0, \delta]$ sowie nach Voraussetzung stetig auf $]0, \delta]$. Mit Folgerung 6.1 erhalten wir schließlich wieder für $x = \frac{2n+1}{2}$:

$$\lim_{n \rightarrow \infty} I_1(n) = \lim_{x \rightarrow \infty} \frac{1}{2\pi} \int_0^\delta \frac{g(t)}{\frac{t}{2}} \sin(xt) dt = 0.$$

□

6.1.6 Gibbs-Phänomen

In diesem Abschnitt untersuchen wir das Verhalten einer Fourier-Reihe an einer Sprungstelle der Ursprungsfunktion genauer. In der praktischen Anwendung entstehen Sprungstellen, wenn man ein Signal periodisch fortsetzt. Selbst wenn man ein periodisches Signal hat, kennt man eventuell die Periode nicht und setzt es mit einer anderen Periode fort (siehe Kapitel 6.4.7). Sprungstellen führen zu einem unendlich großen Spektrum, das durch eine endliche Fourier-Partialsumme nicht vollständig erfasst werden kann. Als Konsequenz beobachtet man ein Überschwingen der Fourier-Partialsummen in der Nähe der Unstetigkeitsstellen, das wir an einem Beispiel analysieren wollen.

Beispiel 6.10

Die 2π -periodische Funktion f sei im Periodenintervall $[-\pi, \pi[$ erklärt über:

$$f(t) = \begin{cases} 0, & t = -\pi \text{ und } t = 0, \\ 1, & t \in]-\pi, 0[, \\ -1, & t \in]0, \pi[. \end{cases}$$

Hat man irgendeine 2π -periodische Funktion g mit Sprungstelle bei t_0 und einseitigen Grenzwerten $L := \lim_{t \rightarrow t_0-} g(t)$ und $R := \lim_{t \rightarrow t_0+} g(t)$, so ist

$$g(t) = \underbrace{g(t) - h(t)}_{\text{stetig ergänzbar in } t_0 \text{ mit } 0} + h(t), \quad h(t) = \frac{L - R}{2} f(t - t_0) + \frac{L + R}{2}. \quad (6.22)$$

Daher ist für das besondere Verhalten an der Sprungstelle auch hier das „Rechtecksignal“ f entscheidend, sofern sich die Fourier-Reihe von $g(t) - h(t)$ bei t_0 gutmütig verhält. Die Funktion f erfüllt die Voraussetzungen des Konvergenzsatzes (Satz 6.4): Bis auf die Sprungstellen $k\pi$ ist f als stückweise konstante Funktion stetig mit $f'(t) = 0$. Außerdem gilt für die Sprungstellen

$$\frac{\lim_{t \rightarrow k\pi+} f(t) + \lim_{t \rightarrow k\pi-} f(t)}{2} = 0 = f(k\pi).$$

Damit konvergiert die Fourier-Reihe in jedem Punkt. Bei der Berechnung der Fourier-Koeffizienten nutzen wir aus, dass f ungerade ist. Damit sind alle Koeffizienten $a_k = 0$, und wir erhalten eine Sinus-Reihe. Bei der Berechnung der Koeffizienten b_k stellt man fest, dass sich aufgrund der Funktionswerte des Kosinus die Koeffizienten b_{2k} zu geraden Indizes $2k$ (also $b_2, b_4, b_6, \dots$) anders verhalten als die Koeffizienten b_{2k-1} zu ungeraden Indizes $2k-1$ (also $b_1, b_3, b_5, \dots$): $\cos(2k\pi) = 1$, $\cos((2k-1)\pi) = -1$.

$$\begin{aligned} b_{2k} &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(2kt) dt = \frac{1}{\pi} \int_{-\pi}^0 \sin(2kt) dt - \frac{1}{\pi} \int_0^{\pi} \sin(2kt) dt \\ &= -\frac{1}{2k\pi} \cos(2kt) \Big|_{-\pi}^0 + \frac{1}{2k\pi} \cos(2kt) \Big|_0^{\pi} = -\frac{1}{2k\pi} - \left[-\frac{1}{2k\pi} \right] + \frac{1}{2k\pi} - \frac{1}{2k\pi} = 0, \\ b_{2k-1} &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin((2k-1)t) dt = \frac{1}{\pi} \int_{-\pi}^0 \sin((2k-1)t) dt - \frac{1}{\pi} \int_0^{\pi} \sin((2k-1)t) dt \\ &= -\frac{1}{(2k-1)\pi} \cos((2k-1)t) \Big|_{-\pi}^0 + \frac{1}{(2k-1)\pi} \cos((2k-1)t) \Big|_0^{\pi} \\ &= -\frac{1}{(2k-1)\pi} + \frac{1}{(2k-1)\pi}(-1) + \frac{1}{(2k-1)\pi}(-1) - \frac{1}{(2k-1)\pi} = -\frac{4}{(2k-1)\pi}. \end{aligned}$$

Damit ist

$$f(t) = -\sum_{k=1}^{\infty} \frac{4}{(2k-1)\pi} \sin((2k-1)t). \quad (6.23)$$

In diesem Beispiel haben wir die unstetige Funktion f als Grenzwert von Fourier-Partialsummen geschrieben. Für jedes (einzelne und fest gewählte) $t \in \mathbb{R}$ konvergieren die Fourier-Partialsummen als Folge gegen den Funktionswert $f(t)$. Dies ist eine punktweise Konvergenz der Funktionenfolge im Sinne von Definition 2.42. Man kann aber nicht erwarten, dass es eine „Mindestgeschwindigkeit“ gibt, mit der die Partialsummen gleichmäßig in jedem Punkt $t \in \mathbb{R}$ gegen $f(t)$ streben in dem Sinne, dass

$$\lim_{n \rightarrow \infty} \left(\sup_{t \in \mathbb{R}} \left| f(t) - \left[-\sum_{k=1}^n \frac{4}{(2k-1)\pi} \sin((2k-1)t) \right] \right| \right) = 0. \quad (6.24)$$

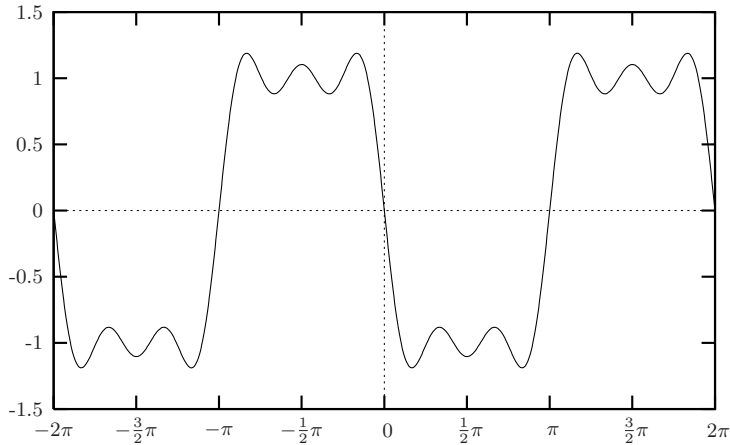


Abb. 6.8: $-\sum_{k=1}^3 \frac{4}{(2k-1)\pi} \sin(2k-1)t$ auf $[-2\pi, 2\pi]$

Man kann also keine gleichmäßige Konvergenz der Funktionenfolge im Sinne der Definition auf Seite 404 erwarten. Das liegt daran, dass die Fourier-Partialsummen als Summe der stetigen Sinus- und Kosinus-Funktionen stetig sind. Auch an den Unstetigkeitsstellen von f kann man die Funktionsgraphen der Partialsummen im Gegensatz zum Graphen von f durchzeichnen. Dabei macht man zwangsläufig einen Approximationsfehler. Hätten wir gleichmäßige Konvergenz, dann müsste wegen Satz 2.56 auch die Grenzfunktion stetig sein, was sie hier nicht ist.

Um zu verdeutlichen, was (6.24) bedeutet, sehen wir uns statt der Fourier-Partialsummen die stückweise linearen und stetigen Funktionen f_n an, die 2π -periodisch fortgesetzt seien (siehe Abbildung 6.9):

$$f_n(t) := \begin{cases} -nt, & t \in \left[-\frac{1}{n}, \frac{1}{n}\right], \\ n(t - \pi), & t \in \left[\pi - \frac{1}{n}, \pi + \frac{1}{n}\right], \\ 1, & t \in \left]-\pi + \frac{1}{n}, -\frac{1}{n}\right[, \\ -1, & t \in \left]\frac{1}{n}, \pi - \frac{1}{n}\right[. \end{cases}$$

Für jedes $t \in \mathbb{R}$ gilt $\lim_{t \rightarrow \infty} f_n(t) = f(t)$, aber es liegt keine gleichmäßige Konvergenz unabhängig von einem einzelnen $t \in \mathbb{R}$ vor (siehe Abbildung 6.9):

$$\sup_{t \in \mathbb{R}} |f(t) - f_n(t)| = 1.$$

Wir können diesen Effekt auch an einzelnen Funktionswerten festmachen. Dazu betrachten wir zu jedem f_n die (von n abhängige) Stelle $\frac{1}{2n}$. Hier ist $f_n\left(\frac{1}{2n}\right) = -\frac{1}{2}$, und wir haben für jedes $n \in \mathbb{N}$ den Fehler $\left|f\left(\frac{1}{2n}\right) - f_n\left(\frac{1}{2n}\right)\right| = \frac{1}{2}$. Bei der Approximation

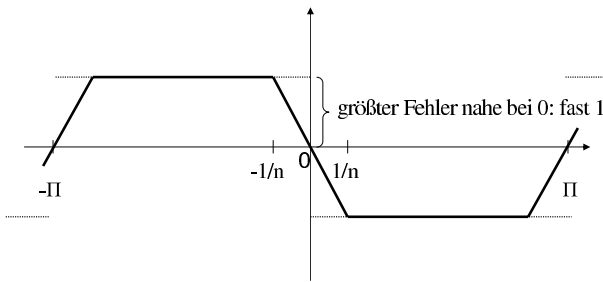


Abb. 6.9: Approximation von f durch eine stückweise lineare Funktion f_n

einer Funktion durch Fourier-Partialsummen kommt aber noch ein weiterer Effekt hinzu. In der Nähe der Unstetigkeitsstellen gibt es zusätzlich zum Fehler, den man macht, weil man die stetigen Partialsummen durchzeichnet, noch Überschwinger. Dies sind die überhohen „Höcker“ in Abbildung 6.8 und 6.11. Die Fourier-Koeffizienten werden so berechnet, dass ein mittlerer quadrierter Fehler minimal wird, siehe (6.1) auf Seite 816. Durch die Überschwinger werden die Flanken der Fourier-Partialsummen an der Sprungstelle steiler, so dass im Mittel die gegebene Funktion besser angenähert wird. Ein ähnlicher Effekt lässt sich bei der Annäherung von Höhendaten $h(x, y)$ durch eine aus Dreiecken fester Größe zusammengesetzte Oberfläche mit Höhen $\Delta(x, y)$ (Triangulierung) auf einem Gebiet G beobachten, wenn die Parameter der Dreiecke so gewählt werden, dass $\iint_G (h(x, y) - \Delta(x, y))^2 d(x, y)$ minimal wird. In Abbildung 6.10 ragen die gezeichneten Höhen $\Delta(x, y)$ am Rand einer Brücke durch die Gleise. An dieser Stelle fällt der Boden im verwendeten digitalen Höhenmodell von der Höhe des Bahndamms auf die Höhe der Straße unter der Brücke, d. h., $h(x, y)$ ist hier unstetig, und $\Delta(x, y)$ hat einen Überschwinger.

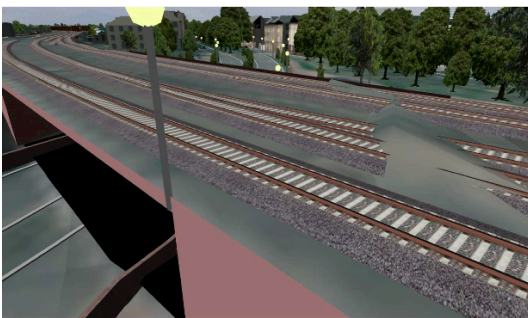


Abb. 6.10: Bei der Annäherung von Höhendaten durch Dreiecksflächen entsteht ein Überschwinger, der die Bahngleise durchbricht. Er liegt an der Sprungstelle von Bahndamm zur Straße unter der Brücke.

Wir zeigen jetzt, dass die Überschwinger von Fourier-Partialsummen nicht verschwinden, wenn man die Summen für sehr große n berechnet und diese auch nicht kleiner werden. Das ist der **Gibbs-Effekt**. Dazu berechnen wir für die Folge $t_n = \frac{\pi}{2n}$

(das sind gerade die Stellen, an denen der Höcker bei der entsprechenden Partialsumme auftritt) den Grenzwert

$$\lim_{n \rightarrow \infty} - \sum_{k=1}^n \frac{4}{(2k-1)\pi} \sin((2k-1)t_n).$$

Zur Berechnung des Grenzwerts der „Höckerfunktionswerte“ interpretieren wir die Summe als summierte Mittelpunktsregel (siehe Seite 360) und damit als Riemann-Zwischensumme der stetigen und damit integrierbaren Funktion $\operatorname{sinc} t = \frac{\sin t}{t}$ (siehe (2.18) auf Seite 271):

$$\begin{aligned} - \sum_{k=1}^n \frac{4}{(2k-1)\pi} \sin((2k-1)t_n) &= - \sum_{k=1}^n \frac{4}{(2k-1)\pi} \sin\left(\frac{(2k-1)\pi}{2n}\right) \\ &= - \frac{2}{n} \sum_{k=1}^n \frac{2n}{(2k-1)\pi} \sin\left(\frac{(2k-1)\pi}{2n}\right) = - \frac{2}{\pi} \sum_{k=1}^n \frac{\pi \sin\left(\frac{(2k-1)\pi}{2n}\right)}{\frac{(2k-1)\pi}{2n}} \\ &= - \frac{2}{\pi} \sum_{k=1}^n \frac{\pi}{n} \operatorname{sinc}\left(\frac{(2k-1)\pi}{2n}\right). \end{aligned}$$

Unterteilt man das Intervall $[0, \pi]$ in n Teilintervalle $[(k-1)\frac{\pi}{n}, k\frac{\pi}{n}]$, $1 \leq k \leq n$, der Länge $\frac{\pi}{n}$, so haben diese jeweils den Mittelpunkt $\frac{1}{2}[(k-1)\frac{\pi}{n} + k\frac{\pi}{n}] = \frac{(2k-1)\pi}{2n}$, so dass hier genau die Mittelpunktsregel für das Integral $\int_0^\pi \operatorname{sinc}(t) dt$ steht. Es gilt:

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{\pi}{n} \operatorname{sinc}\left(\frac{(2k-1)\pi}{2n}\right) = \int_0^\pi \operatorname{sinc} t dt = \int_0^\pi \underbrace{\frac{\sin t}{t}}_{>0, 0 < t < \pi} dt =: C > 0.$$

Damit ist

$$\lim_{n \rightarrow \infty} - \sum_{k=1}^n \frac{4}{(2k-1)\pi} \sin((2k-1)t_n) = -\frac{2}{\pi} C \approx -1,17898.$$

Der Höcker wandert mit wachsendem n immer näher an die Unstetigkeitsstelle und wird dabei auch immer schmaler. Allerdings hat die Funktion f für t_n jeweils den Wert -1 , d. h., der Höcker verschwindet auch für große n nicht, und im Grenzwert überragt er f um $0,17898$ und damit um etwa 18 % der halben Sprunghöhe. Wegen (6.22) gilt das sogar generell für Sprungstellen unabhängig vom konkreten Beispiel f .

Dieser Effekt steht nicht im Widerspruch zur (punktweisen) Konvergenz der Fourier-Reihe für jedes t . An den Unstetigkeitsstellen $t = k\pi$ sind die Partialsummen gleich dem Funktionswert $f(k\pi) = 0$. Für jede andere Stelle t gilt, dass ein gegebenenfalls störender Höcker die Stelle t auf seinem Weg zur Unstetigkeitsstelle mit wachsendem n irgendwann passiert hat.

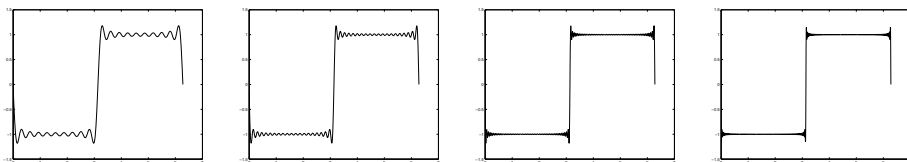


Abb. 6.11: $-\sum_{k=1}^n \frac{4}{(2k-1)\pi} \sin((2k-1)t)$ für $n = 10, 20, 50, 100$ auf $[0, 2\pi]$

Ist eine 2π -periodische Funktion f stetig auf $[0, 2\pi]$ bis auf endliche viele Sprungstellen $x_1, \dots, x_m$ und ist f auf jedem Intervall $[0, x_1], [x_1, x_2], \dots, [x_m, 2\pi]$ stetig differenzierbar, so erfüllt sie an jeder Stelle die Voraussetzungen des Konvergenzsatzes. Darüber hinaus kann man zeigen, dass die Fourier-Reihe sogar auf jedem Intervall $[a, b]$, das keine der Sprungstellen $x_1, \dots, x_m$ enthält, gleichmäßig gegen f konvergiert (Heuser, 2004, S. 148). Effekte nicht-gleichmäßiger Konvergenz können also ausschließlich an Sprungstellen auftreten.

Durch Mittelwertbildung kann man auch dort die Überschwinger verhindern. Das lässt sich z. B. durch Faltung mit dem Fejér-Kern (siehe Kasten auf Seite 845) bewerkstelligen: Statt f über eine Fourier-Reihe wie in (6.23) darzustellen, lässt sich f so durch zusätzliche Fejér-Mittelung annähern:

$$f(t) = - \lim_{n \rightarrow \infty} \sum_{k=1}^n \left(1 - \frac{2k-1}{2n}\right) \frac{4}{(2k-1)\pi} \sin((2k-1)t).$$

Für die Stellen t_n sehen wir (analog zur vorangehenden Berechnung)

$$\begin{aligned} & - \lim_{n \rightarrow \infty} \sum_{k=1}^n \left(1 - \frac{2k-1}{2n}\right) \frac{4}{(2k-1)\pi} \sin((2k-1)t_n) \\ &= - \lim_{n \rightarrow \infty} \frac{2}{\pi} \sum_{k=1}^n \left(1 - \frac{2k-1}{2n}\right) \frac{\pi}{n} \operatorname{sinc}\left(\frac{(2k-1)\pi}{2n}\right) \\ &= - \lim_{n \rightarrow \infty} \frac{2}{\pi^2} \sum_{k=1}^n \frac{\pi}{n} \left(\pi - \frac{(2k-1)\pi}{2n}\right) \operatorname{sinc}\left(\frac{(2k-1)\pi}{2n}\right) \\ &= - \frac{2}{\pi^2} \int_0^\pi (\pi - t) \operatorname{sinc}(t) dt = - \frac{2}{\pi} C + \frac{2}{\pi^2} \int_0^\pi \sin(t) dt = - \frac{2}{\pi} C + \frac{4}{\pi^2} \approx -0,7737. \end{aligned}$$

Damit gibt es an den Stellen t_n asymptotisch kein Überspringen mehr über -1 hinaus, allerdings auch weiterhin keine gleichmäßige Konvergenz in einer Umgebung der Sprungstelle.

Gängiger als Mittelwertbildung ist der Einsatz von Fensterfunktionen (vgl. Kapitel 6.4.7 und 6.4.10).

Den Gibbs-Effekt kann man auch bei im JPEG-Format vorliegenden Bildern beobachten, die mit Mitteln der Fourier-Analyse komprimiert wurden. An kontrastreichen Kanten treten hier Geisterlinien auf, die durch Überschwinger verursacht werden. Man kann diesen Effekt nutzen, um Kanten in einem Bild noch schärfer erscheinen zu lassen.

6.1.7 Entwicklung $2p$ -periodischer Funktionen

Bislang haben wir nur 2π -periodische Funktionen betrachtet, da dafür die Schreibweise (etwas) einfacher als bei einer beliebigen Periode $2p$ ist, die man in den Anwendungen vorfindet. Sei jetzt aber f eine integrierbare Funktion mit der allgemeinen Periode $2p > 0$, d. h.

$$f(t + 2p) = f(t) \text{ für alle } t \in \mathbb{R}.$$

Die daraus mittels Substitution gewonnene Funktion $g(t) := f\left(\frac{p}{\pi}t\right)$ ist wieder 2π -periodisch, denn

$$g(t + 2\pi) = f\left(\frac{p}{\pi}(t + 2\pi)\right) = f\left(\frac{p}{\pi}t + 2p\right) = f\left(\frac{p}{\pi}t\right) = g(t).$$

Damit können wir $g(t)$ als Fourier-Reihe entwickeln. Wir berechnen die zugehörigen Fourier-Koeffizienten mittels Substitution $u = \frac{p}{\pi}t$, $du = \frac{p}{\pi}dt$:

$$\begin{aligned} a_0 &:= \frac{1}{2\pi} \int_{-\pi}^{\pi} f\left(\frac{p}{\pi}t\right) dt = \frac{1}{2p} \int_{-p}^p f(u) du, \\ a_k &:= \frac{1}{\pi} \int_{-\pi}^{\pi} f\left(\frac{p}{\pi}t\right) \cos(kt) dt = \frac{1}{p} \int_{-p}^p f(u) \cos\left(\frac{\pi}{p}ku\right) du, \quad k \in \mathbb{N}, \\ b_k &:= \frac{1}{\pi} \int_{-\pi}^{\pi} f\left(\frac{p}{\pi}t\right) \sin(kt) dt = \frac{1}{p} \int_{-p}^p f(u) \sin\left(\frac{\pi}{p}ku\right) du, \quad k \in \mathbb{N}. \end{aligned}$$

Ist die $2p$ -periodische Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ auf $[-p, p]$ stetig bis auf endlich viele Sprungstellen und existiert f' außerhalb der Sprungstellen und ist beschränkt, so sind für $g(t) = f\left(\frac{p}{\pi}t\right)$ die Voraussetzungen des Konvergenzsatzes (Satz 6.4) erfüllt. Mit den soeben berechneten Koeffizienten gilt in jedem Stetigkeitspunkt $\frac{p}{\pi}t$ von f :

$$g(t) = f\left(\frac{p}{\pi}t\right) = a_0 + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt)),$$

also

$$f(t) = g\left(\frac{\pi}{p}t\right) = a_0 + \sum_{k=1}^{\infty} \left(a_k \cos\left(\frac{\pi}{p}kt\right) + b_k \sin\left(\frac{\pi}{p}kt\right) \right).$$

In komplexer Schreibweise erhalten wir (sogar für komplexwertige Funktionen $f : \mathbb{R} \rightarrow \mathbb{C}$)

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \exp\left(j\frac{\pi}{p}kt\right) := \lim_{n \rightarrow \infty} \sum_{k=-n}^n c_k \exp\left(j\frac{\pi}{p}kt\right) \quad (6.25)$$

mit

$$c_k = \frac{1}{2p} \int_{-p}^p f(t) \exp\left(-j \frac{\pi}{p} kt\right) dt = \frac{1}{2p} \int_{-p}^p f(t) \exp\left(-j \frac{k}{2p} 2\pi t\right) dt. \quad (6.26)$$

In der Signalverarbeitung bezeichnet man den Quotienten $\omega_0 := \frac{\pi}{p} = \frac{2\pi}{2p}$ als **Grundfrequenz**. Unter Verwendung von ω_0 werden (6.25) und (6.26) zu

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \exp(jk\omega_0 t), \quad c_k = \frac{\omega_0}{2\pi} \int_{-\pi/\omega_0}^{\pi/\omega_0} f(t) \exp(-jk\omega_0 t) dt.$$

Die Fourier-Koeffizienten c_{-k} und c_k , $k \in \mathbb{N}_0$, gehören zur in Hertz gemessenen Frequenz

$$\frac{\omega_0}{2\pi} \cdot k = \frac{k}{2p}.$$

Wir diskutieren also jetzt ein Spektrum mit den Frequenzen $\frac{k}{2p}$ Hz, $k \in \mathbb{N}_0$, benachbarte Frequenzen haben den Abstand $\frac{1}{2p}$ Hz.

Alles, was wir zuvor für 2π -periodische Funktionen gemacht haben, lässt sich direkt auf $2p$ -periodische Funktionen übertragen, indem man π durch p ersetzt. Beispielsweise berechnet sich die Amplitude der Frequenz $\frac{k}{2p}$ analog zu Kapitel 6.1.2 und (6.8) aus den Fourier-Koeffizienten als

$$\sqrt{a_k^2 + b_k^2} = \sqrt{[2 \operatorname{Re}(c_k)]^2 + [2 \operatorname{Im}(c_k)]^2} = 2|c_k| = |c_k| + |c_{-k}|.$$

Außerdem ist die Faltung zweier $2p$ -periodischer Funktionen definiert als $\frac{1}{2p} \int_{-p}^p f(t-u)g(u) du$, es gilt der Faltungssatz, und Fourier-Reihen sind unter den Voraussetzungen des Konvergenzsatzes für Fourier-Reihen konvergent.

Beispiel 6.11 (Eins-periodische Funktionen)

Wählen wir als Periodenintervall $[0, 1]$, also $2p = 1$ bzw. $p = 1/2$, so lautet die Fourier-Reihe einer 1-periodischen, reellen Funktion $f(t)$

$$a_0 + \sum_{k=1}^{\infty} (a_k \cos(2\pi kt) + b_k \sin(2\pi kt)) = \sum_{k=-\infty}^{\infty} c_k \exp(j2\pi kt),$$

wobei sich die reellen Fourier-Koeffizienten a_k , b_k und die komplexen Fourier-Koeffizienten c_k berechnen zu

$$a_0 = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(t) dt = \int_0^1 f(t) dt, \quad a_k = 2 \int_0^1 f(t) \cos(2\pi kt) dt,$$

$$b_k = 2 \int_0^1 f(t) \sin(2\pi kt) dt, \quad k \in \mathbb{N}, \quad c_k = \int_0^1 f(t) \exp(-j2\pi kt) dt, \quad k \in \mathbb{Z}.$$

Hier ist die Grundfrequenz $\omega_0 = 2\pi$, und die Variable $k = \frac{\omega_0}{2\pi} \cdot k$ beschreibt in der Fourier-Reihe die Frequenz des jeweiligen Summanden direkt in Hertz. ■

6.2 Fourier-Transformation

6.2.1 Fourier-Integral

Die Darstellung (6.25), (6.26) dient als Ausgangspunkt für die Fourier-Transformation nicht-periodischer Funktionen. Wir betrachten dazu eine nicht-periodische Funktion $f : \mathbb{R} \rightarrow \mathbb{C}$ zunächst auf einem Intervall $[-p, p[$. Auch wenn die Ursprungsfunktion nicht die Periode $2p$ hat, können wir sie vom Intervall $[-p, p[$ mit der Periode $2p$ (zu einer anderen Funktion) fortsetzen und eine Fourier-Reihe angeben. Falls f auf $[-p, p[$ genügend glatt ist, konvergiert diese Fourier-Reihe, und wir können die Werte der Ursprungsfunktion f zwar nicht auf ganz $\mathbb{R}$ aber immerhin auf $[-p, p[$ über die Reihe berechnen. Mit der Hilfsgröße $h_k := \frac{\pi}{p}k$ (also $h_{k+1} - h_k = \frac{\pi}{p}$) erhalten wir hier die Darstellung

$$\begin{aligned} f(x) &= \sum_{k=-\infty}^{\infty} \left[\frac{1}{2p} \int_{-p}^p f(t) \exp\left(-j \frac{\pi}{p} kt\right) dt \right] \exp\left(j \frac{\pi}{p} kx\right) \\ &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} [h_{k+1} - h_k] \left[\int_{-\pi/(h_{k+1}-h_k)}^{\pi/(h_{k+1}-h_k)} f(t) \exp(-jh_k t) dt \right] \exp(jh_k x). \end{aligned}$$

Wir lassen nun formal p gegen ∞ streben, so dass wir uns von der Periodizität verabschieden und f auf immer größeren Intervallen und schließlich auf $]-\infty, \infty[$ betrachten. Mit $p \rightarrow \infty$ gilt insbesondere $h_{k+1} - h_k \rightarrow 0$. Das Integral $\int_{-\pi/(h_{k+1}-h_k)}^{\pi/(h_{k+1}-h_k)} f(t) \exp(-jh_k t) dt$ wird dadurch zu $\int_{-\infty}^{\infty}$. Die Summe (die wir so eigentlich nicht hinschreiben dürfen, da wir so zwei separate und nicht einen über h_k im Inneren des Integrals gekoppelten Grenzwert betrachten)

$$\frac{1}{2\pi} \sum_{k=-\infty}^{\infty} [h_{k+1} - h_k] \left[\int_{-\infty}^{\infty} f(t) \exp(-jh_k t) dt \right] \exp(jh_k x)$$

ist eine Riemann-Zwischensumme zum unbeschränkten Intervall $]-\infty, \infty[$, das durch Intervalle $[h_k, h_{k+1}] = \left[\frac{\pi}{p}k, \frac{\pi}{p}(k+1) \right]$ zerlegt wird. Dies suggeriert, dass wie bei einer Quadraturformel (vgl. Seite 360) für $p \rightarrow \infty$ die Summe gegen das Integral

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(t) \exp(-j\omega t) dt \right] \exp(j\omega x) d\omega \quad (6.27)$$

strebt. Diese Überlegung ist rein heuristisch und kein Beweis! Allerdings ist der Ausdruck nun für nicht-periodische Funktionen f sinnvoll. Damit das Integral wohldefiniert ist, benötigt man Zusatzbedingungen wie z.B. die Existenz der Majorante $\int_{-\infty}^{\infty} |f(t)| dt$, so dass man hauptsächlich Funktionen mit $\lim_{t \rightarrow \pm\infty} f(t) = 0$ betrachtet. Wir nennen nun das Ergebnis des inneren Integrals die Fourier-Transformierte der Funktion f . Das äußere Integral überführt diese Fourier-Transformierte zurück in die Ursprungsfunktion.

Definition 6.3 (Fourier-Transformation)

Sei $f : \mathbb{R} \rightarrow \mathbb{C}$ (d. h. $f(t) = f_1(t) + j f_2(t)$) mit reellwertigen Funktionen f_1 und f_2) und $f(t)$ auf jedem endlichen Intervall integrierbar, so dass

$$\int_{-\infty}^{\infty} |f(t)| dt := \lim_{u \rightarrow \infty} \int_{-u}^0 |f(t)| dt + \lim_{u \rightarrow \infty} \int_0^u |f(t)| dt$$

existiert. Dann heißt f **Fourier-transformierbar**, und die **Fourier-Transformation** von f ist definiert durch

$$[\mathcal{F}(f)](\omega) := \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt.$$

Die Funktion $f^\wedge(\omega) := [\mathcal{F}(f)](\omega) : \mathbb{R} \rightarrow \mathbb{C}$ heißt die **Fourier-Transformierte** der Funktion f .

Bemerkungen:

- Die Fourier-Transformation $\mathcal{F}$ überführt eine Funktion f in eine andere Funktion $\mathcal{F}(f)$, die Fourier-Transformierte $f^\wedge$. Möchte man den Funktionswert $f^\wedge(\omega)$ der Transformierten an einer Stelle ω berechnen, so muss man dazu das Integral $\int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt$ lösen, wobei ω für die Integration konstant ist und eine Funktion der Variable t integriert wird. Daher wird die Fourier-Transformation als **Integraltransformation** bezeichnet.
- Für die Ursprungsfunktion f wird (neben x) häufig die Variable t verwendet, da f in vielen Anwendungen von der Zeit abhängt (ein zeitabhängiges Signal beschreibt). Man spricht daher auch von einer Funktion im **Zeitbereich**. Die Zielfunktion wird üblicherweise mit der Variable ω geschrieben, um die Abhängigkeit von Frequenzen anzudeuten. Hier spricht man von einer Funktion im **Frequenzbereich**. Der Wert der Variablen ω ist aber eine Kreisfrequenz. Durch Division mit 2π wird daraus eine Frequenz $\frac{\omega}{2\pi}$, die in Hertz angegeben werden kann. Denn $f^\wedge(\omega)$ ist in (6.27) der Vorfaktor der 2π -periodischen Funktion $\exp(j\omega x)$ mit der Frequenz $\frac{\omega}{2\pi}$.
- Die Fourier-Transformierte $[\mathcal{F}(f)](\omega)$ ist eine Funktion von $\mathbb{R}$ nach $\mathbb{C}$. Da sie für jeden Wert $\omega \in \mathbb{R}$ erklärt ist, spricht man von einem **kontinuierlichen Spektrum**. Dagegen bilden die Fourier-Koeffizienten

$$\left(\frac{1}{2p} \int_{-p}^p f(t) \exp\left(-j \frac{k\pi t}{p}\right) dt \right)_{k=-\infty}^{\infty}$$

eine Folge und damit eine Funktion von $\mathbb{Z}$ nach $\mathbb{C}$. Da als Frequenzen nur abzählbar viele Werte $\frac{k}{2p}$ auftreten, spricht man von einem **diskreten Spektrum**. Benachbarte Frequenzen haben dabei den Abstand $\frac{1}{2p}$ Hz. Durch die formale Rechnung für $p \rightarrow \infty$ geht der Abstand gegen null, so dass das kontinuierliche Spektrum der Fourier-Transformation entsteht.

Mit $f^\wedge(k)$ bzw. $f^\wedge(\omega)$ bezeichnen wir sowohl Fourier-Koeffizienten als auch die Fourier-Transformierte (Doppelbezeichnung). Aufgrund der Eigenschaften der Funktion f (2π -periodisch oder $|f|$ auf $\mathbb{R}$ integrierbar und damit bei Stetigkeit mit Grenzwert null im Unendlichen) ist jedoch stets klar, welche Definition gemeint ist.

- $f^\wedge(\omega)$ ist eine komplexe Zahl, $f^\wedge$ eine komplexwertige Funktion. Die zugehörige reellwertige Funktion $|f^\wedge|$ heißt **Amplitudenspektrum**. Sie wird häufig verwendet, um die Transformierte als Funktionsgraph darzustellen.
- Man kann das Integral der Fourier-Transformation ausrechnen, da uns die Voraussetzung der Definition eine integrierbare Majorante $|f(t)|$ liefert. Zunächst schreiben wir die Transformierte über Integrale von reellwertigen Funktionen:

$$\begin{aligned} \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt &= \int_{-\infty}^{\infty} [\operatorname{Re}(f(t)) + j \operatorname{Im}(f(t))] \cdot [\cos(\omega t) - j \sin(\omega t)] dt \\ &= \int_{-\infty}^{\infty} \operatorname{Re}(f(t)) \cos(\omega t) + \operatorname{Im}(f(t)) \sin(\omega t) dt \\ &\quad + j \int_{-\infty}^{\infty} -\operatorname{Re}(f(t)) \sin(\omega t) + \operatorname{Im}(f(t)) \cos(\omega t) dt. \end{aligned}$$

Da $|\operatorname{Re}(f(t))| \leq |f(t)|$, $|\operatorname{Im}(f(t))| \leq |f(t)|$ und $|\cos(\omega t)| \leq 1$, $|\sin(\omega t)| \leq 1$, ist die Funktion $2|f(t)|$ eine Majorante für beide Integrale. Damit existieren nach Voraussetzung an f und Satz 2.51 (siehe Seite 366) beide Integrale, und für jedes $\omega \in \mathbb{R}$ ist der Funktionswert $f^\wedge(\omega) \in \mathbb{C}$ wohldefiniert. Damit ist auch die Bezeichnung „Fourier-transformierbar“ gerechtfertigt.

- Man beachte, dass durch die Voraussetzung $\int_{-\infty}^{\infty} |f(t)| dt < \infty$ eine sogenannte absolute Konvergenz des Integrals $\int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt$ vorliegt. In einem solchen Fall darf man (wie bei der Umordnung von Reihen, vgl. Seite 244) statt der beiden Grenzwerte

$$\int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt = \lim_{u \rightarrow -\infty} \int_u^0 f(t) e^{-j\omega t} dt + \lim_{u \rightarrow \infty} \int_0^u f(t) e^{-j\omega t} dt$$

auch einen symmetrischen Grenzwert ausrechnen:

$$\int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt = \lim_{u \rightarrow \infty} \int_{-u}^u f(t) e^{-j\omega t} dt.$$

Beispiel 6.12

Wir betrachten den **Rechteckimpuls**

$$f(t) := \begin{cases} 1, & -1 \leq t \leq 1, \\ 0, & |t| > 1. \end{cases} \quad (6.28)$$

Dieser hat die reelle Fourier-Transformierte

$$[\mathcal{F}f(t)](\omega) = \int_{-1}^1 e^{-j\omega t} dt = \int_{-1}^1 \cos(\omega t) - j \sin(\omega t) dt$$

$$\begin{aligned}
&= \begin{cases} \frac{1}{\omega} [\sin(\omega t) + j \cos(\omega t)]_{-1}^1 = 2 \frac{\sin \omega}{\omega}, & \omega \neq 0, \\ 2, & \omega = 0. \end{cases} \\
&= 2 \cdot \text{sinc } \omega.
\end{aligned} \tag{6.29}$$

Den Sinus Cardinalis (siehe Seite 271) haben wir ebenfalls beim Gibbs-Phänomen (siehe Seite 853) vorgefunden. Dort haben wir die Fourier-Reihe eines periodischen Rechtecksignals analysiert. ■

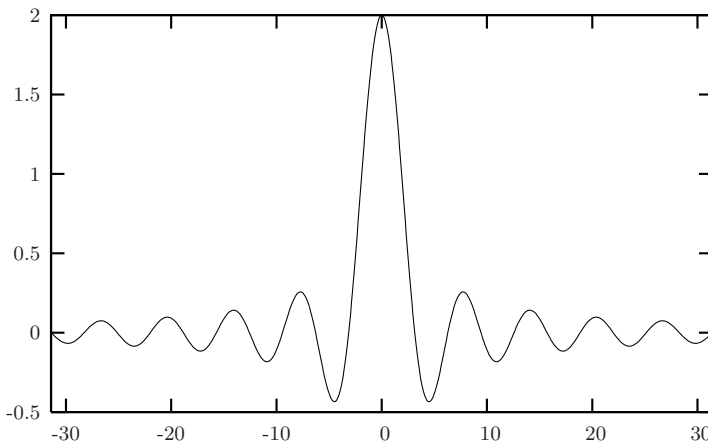


Abb. 6.12: $2 \cdot \text{sinc } \omega$

Bevor wir später auf die Eigenschaften der Fourier-Transformierten eingehen, möchten wir an dieser Stelle schon auf ein sehr schönes Verhalten aufmerksam machen:

Lemma 6.5 (Riemann-Lebesgue-Lemma)

Sei $f : \mathbb{R} \rightarrow \mathbb{C}$ Fourier-transformierbar. Bezogen auf Real- und Imaginärteil ist

- $f^\wedge(\omega)$ stetig und
- $\lim_{\omega \rightarrow \pm\infty} f^\wedge(\omega) = 0$ (**Riemann-Lebesgue-Lemma**, vgl. mit (6.17)).

Wie beim entsprechend lautenden Riemann-Lebesgue-Lemma für Fourier-Koeffizienten (Satz 6.5 auf Seite 848) steckt hinter dieser Aussage die Stetigkeit im Mittel (vgl. Folgerung 6.1), für einen jetzt unendlich großen Integrationsbereich.

Der Graph des Amplitudenspektrums $|f^\wedge|$ kann also durchgezeichnet werden, und er strebt für $\omega \rightarrow \pm\infty$ gegen null. Dies ermöglicht in der Praxis, nur mit Werten aus einem genügend großen Intervall $[-R, R]$ zu rechnen, das so gewählt ist, dass $f^\wedge$ außerhalb sehr klein und damit zu vernachlässigen ist.

Wie für die Fourier-Koeffizienten gilt auch für die Fourier-Transformierte eine Parseval'sche Gleichung (siehe (6.11) auf Seite 830), die jetzt **Satz von Plancherel** heißt. Falls beide Integrale existieren (was nicht selbstverständlich ist), gilt:

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |f^\wedge(\omega)|^2 d\omega.$$

6.2.2 Fourier-Umkehrtransformation

Ausgehend von den Fourier-Reihen haben wir als Kandidaten für eine Rücktransformation vom Frequenzbereich in den Zeitbereich mit (6.27) die Umkehrformel

$$f(t) = [\mathcal{F}^{-1} f^\wedge](t) \stackrel{?}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} f^\wedge(\omega) \exp(j\omega t) d\omega \quad (6.30)$$

erhalten. Durch Integration für jeden festen Wert von t wird so aus der Fourier-Transformierten wieder die Ursprungsfunktion. Die **Fourier-Umkehrtransformation** $\mathcal{F}^{-1}$ überführt $f^\wedge$ in f .

Die Summe von periodischen Funktionen, deren Periodenlängen keine rationalen Vielfachen voneinander sind, muss nicht wieder eine periodische Funktion sein (siehe Seite 815). Das wird hier ausgenutzt, um nicht-periodische Funktionen über die Umkehrtransformation mit Sinus und Kosinus darzustellen. Bei einer periodischen Funktion werden die Fourier-Koeffizienten mit der Fourier-Reihe zurück transformiert. Allerdings wird jetzt nicht summiert, sondern integriert.

Das Problem dabei ist jedoch, dass im Allgemeinen $\int_{-\infty}^{\infty} |f^\wedge(\omega)| d\omega$ nicht existiert, obwohl für f gilt: $\int_{-\infty}^{\infty} |f(t)| dt < \infty$. Man kann also das Integral in (6.30) dann in der Regel gar nicht ausrechnen. Das ist vergleichbar damit, dass die Fourier-Reihe einer stetigen, periodischen Funktion ebenfalls (an einer Stelle t_0) nicht konvergieren muss. Gelingt hier aber die Berechnung des Integrals, und ist f stetig, so erhält man tatsächlich f zurück:

Die Umkehrtransformation kann dann über die Fourier-Transformation ausgedrückt werden:

$$[\mathcal{F}^{-1} f^\wedge](t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f^\wedge(\omega) \exp(j\omega t) d\omega = \frac{1}{2\pi} [f^\wedge]^\wedge(-t),$$

also erhalten wir mit $g = f^\wedge$ und wegen $f = \mathcal{F}^{-1} f^\wedge$:

$$[\mathcal{F}^{-1} g](t) = \frac{1}{2\pi} g^\wedge(-t), \quad g^\wedge(t) = 2\pi [\mathcal{F}^{-1} g](t) \text{ und } f(t) = \frac{1}{2\pi} f^\wedge^\wedge(-t). \quad (6.31)$$

Fourier-Transformation und Fourier-Umkehrtransformation funktionieren problemlos für alle Funktionen des **Schwartz-Raums**, das sind Funktionen f , die beliebig

oft differenzierbar auf $\mathbb{R}$ sind und für die $|x^p f^{(k)}(x)|$ für jede Potenz $p \in \mathbb{N}_0$ und jede Ableitungsordnung $k \in \mathbb{N}_0$ auf ganz $\mathbb{R}$ beschränkt ist, also

$$|x^p f^{(k)}(x)| \leq C_{p,k} \text{ für alle } x \in \mathbb{R}. \quad (6.32)$$

Ist dagegen die Ursprungsfunktion nicht so schön glatt oder streben die Ableitungen für $x \rightarrow \pm\infty$ nicht schnell genug gegen null, kann man die Umkehrtransformation über sogenannte Summationsverfahren berechnen, die an die Stelle von (6.30) treten.

Für die konkrete Anwendung ist es am einfachsten, wenn man die Umkehrtransformation mittels Eigenschaften der Fourier-Transformation (siehe nächster Abschnitt) und Tabellenwerken bestimmt. Diesen Weg werden wir später auch für die Laplace-Transformation beschreiten.

Bei der Umkehrtransformation findet sich hier der Faktor $\frac{1}{2\pi}$. Alternativ wird dieser Faktor in der Literatur auch bei der Fourier-Transformation (und dann nicht bei der Umkehrtransformation) gesetzt. Auch findet man jeweils bei Fourier- und Umkehrtransformation den Faktor $\frac{1}{\sqrt{2\pi}}$. Wendet man beide Transformationen hintereinander an, benötigt man in jedem Fall zusammen den Faktor $\frac{1}{2\pi}$ – unabhängig davon, welche Definition man verwendet.

Wir betrachten den Rechteckimpuls im Frequenzbereich (so wie wir ihn zuvor im Zeitbereich betrachtet haben):

$$g(\omega) := \begin{cases} 1, & -1 \leq \omega \leq 1, \\ 0, & |\omega| > 1. \end{cases}$$

Auf g wenden wir die Umkehrtransformation an (vgl. Abbildung 6.12, beachte, dass $\text{sinc } t$ gerade ist, also $\text{sinc}(-t) = \text{sinc } t$):

$$[\mathcal{F}^{-1}g(\omega)](t) = \frac{1}{2\pi} g^\wedge(-t) \stackrel{(6.29)}{=} \frac{1}{2\pi} 2 \text{sinc}(-t) = \frac{1}{\pi} \text{sinc } t.$$

Man könnte also glauben, dass g die Fourier-Transformierte von $\frac{1}{\pi} \text{sinc } t$ ist. Dem ist aber nicht so, $\int_{-\infty}^{\infty} |\text{sinc } t| dt$ existiert nicht, und die Funktion $\text{sinc } t$ ist nicht Fourier-transformierbar. Des Rätsels Lösung: g kann nicht Fourier-Transformierte sein, da g nicht stetig ist.

Die Umkehrtransformation (6.30) produziert analog zur Fourier-Transformation stetige Funktionen, deren Grenzwerte für $t \rightarrow \pm\infty$ null sind (vgl. Lemma 6.5). Die einfache Formel (6.30) kann also insbesondere auch nur dann funktionieren, wenn die Ursprungsfunktion stetig ist.

Mit der stetigen (Dreiecks-)Funktion im Frequenzbereich haben wir mehr Glück:

$$g(\omega) := \begin{cases} \pi(1 - |\omega|), & -1 \leq \omega \leq 1, \\ 0, & |\omega| > 1. \end{cases}$$

Man erhält mittels Umkehrtransformation die stetige Funktion im Zeitbereich, deren Fourier-Transformierte g ist:

$$[\mathcal{F}^{-1}g](t) = \begin{cases} \frac{1-\cos t}{t^2} \stackrel{(1.25)}{=} \frac{1-\cos^2(\frac{t}{2})+\sin^2(\frac{t}{2})}{t^2} = \frac{2\sin^2(\frac{t}{2})}{t^2}, & t \neq 0, \\ \frac{1}{2}, & t = 0. \end{cases} \quad (6.33)$$

Insbesondere sind beide Funktionen reellwertig (da eine gerade Funktion transformiert wird, siehe Lemma 6.6 unten). In der Regel erhält man durch die Transformation eine komplexwertige Funktion. Für später berechnen wir noch

$$\left(\frac{1}{\pi}g\right)^\wedge(\omega) = \frac{2\pi}{\pi}[\mathcal{F}^{-1}g](-\omega) = \begin{cases} \frac{\sin^2(\frac{\omega}{2})}{(\frac{\omega}{2})^2}, & \omega \neq 0, \\ 1, & \omega = 0. \end{cases} \quad (6.34)$$

6.2.3 Fourier-Koeffizienten und Fourier-Transformation

Ein Vergleich der Begriffe Fourier-Koeffizienten und Fourier-Transformation ist in Tabelle 6.1 dargestellt. Wir können die Fourier-Koeffizienten als Funktionswerte einer Fourier-Transformierten auffassen:

Sei f eine stetige, 2π -periodische Funktion. Dann ist

$$g(t) := \begin{cases} f(t), & t \in [-\pi, \pi] \\ 0, & |t| > \pi. \end{cases}$$

eine Funktion, die Fourier-transformiert werden kann:

$$g^\wedge(\omega) = \int_{-\infty}^{\infty} g(t)e^{-j\omega t} dt = \int_{-\pi}^{\pi} f(t)e^{-j\omega t} dt.$$

Ist speziell $\omega \in \mathbb{Z}$, so nimmt die Fourier-Transformierte als Wert bis auf einen Faktor 2π den Fourier-Koeffizienten $f^\wedge(\omega)$ der Fourier-Reihe von f an:

$$g^\wedge(\omega) = 2\pi f^\wedge(\omega).$$

Hintergrund: Poisson-Summationsformel

Der allgemeine Zusammenhang zwischen Fourier-Koeffizienten und der Fourier-Transformierten ist durch die **Poisson-Summationsformel** gegeben. Sei g stetig auf $\mathbb{R}$ und $\int_{-\infty}^{\infty} |g(t)| dt < \infty$ und $\sum_{k=-\infty}^{\infty} |g^\wedge(k)| < \infty$, wobei hier $g^\wedge(k)$ ein Funktionswert der Fourier-Transformierten von g und kein Fourier-Koeffizient ist. Außerdem möge $\sum_{k=-\infty}^{\infty} g(t+k2\pi)$ gegen eine Funktion $f(t)$ gleichmäßig konvergieren (siehe Definition 2.43 auf Seite 404), d. h.

$$f(t) = \sum_{k=-\infty}^{\infty} g(t+k2\pi)$$

punktweise für jedes $t \in \mathbb{R}$, und zu jedem noch so kleinen $\varepsilon > 0$ existiert ein $n_0 = n_0(\varepsilon)$ (unabhängig von t), so dass für alle $n > n_0$ unabhängig vom konkreten $t \in \mathbb{R}$ gilt: $|\sum_{k=-n}^n g(t + k2\pi) - f(t)| < \varepsilon$.

Aus der nicht-periodischen Ursprungsfunktion g wird durch die Summation eine 2π -periodische Funktion f gemacht, so dass wir jetzt Werte $g^\wedge(k)$ der Fourier-Transformierten von g als Fourier-Koeffizienten $f^\wedge(k)$ von f auffassen können, d. h., f ist in eine konvergente Fourier-Reihe entwickelbar, und es gilt die **Poisson-Summationsformel**:

$$f(t) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} g^\wedge(k) e^{jkt}.$$

Wir wenden nun diese Formel für eine Funktion g an, die ihrerseits Fourier-Transformierte einer stetigen Funktion h ist: $g(\omega) := h^\wedge(\omega)$. Damit erhalten wir einen Anhaltspunkt, wie man die Funktionswerte einer Fourier-Transformierten numerisch mit dem Computer berechnen kann. Sei also $h: \mathbb{R} \rightarrow \mathbb{C}$ 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, dann durch Änderung der Summationsreihenfolge und schließlich über die Fourier-Umkehrtransformation mit (6.31):

$$\begin{aligned} \sum_{k=-\infty}^{\infty} h^\wedge(\omega + 2k\pi) &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} (h^\wedge)^\wedge(k) e^{jk\omega} \\ &= \sum_{k=-\infty}^{\infty} \frac{1}{2\pi} (h^\wedge)^\wedge(-k) e^{-jk\omega} \stackrel{(6.31)}{=} \sum_{k=-\infty}^{\infty} h(k) e^{-jk\omega}. \end{aligned} \quad (6.35)$$

Auf der linken Seite steht eine Summe von Werten der Fourier-Transformierten und auf der rechten eine Formel, wie man diese aus einzelnen Funktionswerten der Ursprungsfunktion berechnen kann. Wenn die Voraussetzungen so sind, dass die Summe auf der linken Seite nur aus einem Summanden besteht, dann hat man eine Rechenvorschrift zur Berechnung von Werten der Fourier-Transformierten ohne Integration. Das ist eine Aussage des Abtastsatzes von Shannon und Nyquist, den wir später auf Seite 941 behandeln.

Ein anderer Aspekt ist, dass man auch 2π -periodische Funktionen Fourier-transformieren kann. Da periodische Funktionen im Allgemeinen auf $\mathbb{R}$ nicht absolut integrierbar sind, kann man dazu aber nicht die klassische Fourier-Transformation verwenden. Vielmehr muss man dann 2π -periodische Funktionen zu sogenannten Dis-

Tab. 6.1: Vergleich von Fourier-Koeffizienten und Fourier-Transformation

Fourier-Koeffizienten	Fourier-Transformation
$f : \mathbb{R} \rightarrow \mathbb{C}$ ist 2π -periodisch (oder p -periodisch).	$f : \mathbb{R} \rightarrow \mathbb{C}$ ist absolut integrierbar: $\int_{-\infty}^{\infty} f(t) dt < \infty$. In der Regel: $\lim_{t \rightarrow \pm\infty} f(t) = 0$
Fourier-Koeffizienten für $k \in \mathbb{Z}$: $f^\wedge(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-jkt} dt$	Fourier-Transformierte für $\omega \in \mathbb{R}$: $f^\wedge(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt$
Unter den Voraussetzungen des Konvergenzsatzes (Seite 843) kann f aus den Fourier-Koeffizienten mittels einer Fourier-Reihe rekon- struiert werden: $f(t) = \sum_{k=-\infty}^{\infty} f^\wedge(k) e^{jkt}$.	Sollte das Integral der Fourier- Umkehrtransformation existieren (siehe Kapitel 6.2.2), kann f aus der Fourier-Transformierten rekon- struiert werden: $f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f^\wedge(\omega) e^{j\omega t} d\omega$.

tributionen verallgemeinern. Auf Distributionen und diesen Zusammenhang gehen wir
knapp im Kasten auf Seite 898 ein.

6.2.4 Eigenschaften der Fourier-Transformation

Lemma 6.6 (Symmetrie)
Sei $f : \mathbb{R} \rightarrow \mathbb{R}$ eine reellwertige, Fourier-transformierbare Funktion. Dann gilt:

- a) Der Realteil von $f^\wedge$ ist gerade.
- b) Der Imaginärteil von $f^\wedge$ ist ungerade.
- c) Ist f gerade, so ist $f^\wedge$ reellwertig und gerade.
- d) Ist f ungerade, so ist $f^\wedge$ rein imaginär und ungerade.

Ein Beispiel zu c) ist der Rechteckimpuls auf Seite 864.

Beweis:

$$\begin{aligned} f^\wedge(\omega) &= \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt = \lim_{u \rightarrow \infty} \left[\int_{-u}^u f(t) \cos(\omega t) dt + j \int_{-u}^u f(t) \sin(\omega t) dt \right] \\ &= \lim_{u \rightarrow \infty} \left[\int_{-u}^0 f(t) \cos(\omega t) dt + \int_0^u f(t) \cos(\omega t) dt \right. \\ &\quad \left. + j \int_{-u}^0 f(t) \sin(\omega t) dt + j \int_0^u f(t) \sin(\omega t) dt \right] \end{aligned}$$

$$\begin{aligned}
&= \lim_{u \rightarrow \infty} \left[\int_0^u f(-t) \cos(\omega t) dt + \int_0^u f(t) \cos(\omega t) dt \right. \\
&\quad \left. - j \int_0^u f(-t) \sin(\omega t) dt + j \int_0^u f(t) \sin(\omega t) dt \right], \text{ also} \\
f^\wedge(\omega) &= \lim_{u \rightarrow \infty} \left[\int_0^u [f(-t) + f(t)] \cos(\omega t) dt + j \int_0^u [f(t) - f(-t)] \sin(\omega t) dt \right] \\
&= \int_0^\infty [f(-t) + f(t)] \cos(\omega t) dt + j \int_0^\infty [f(t) - f(-t)] \sin(\omega t) dt.
\end{aligned}$$

Die Aussagen lassen sich nun an der Darstellung der Fourier-Transformierten direkt ablesen. $\square$

Die Aussagen c) und d) lassen sich z. B. für stetige Funktionen f auch umkehren.

Ein beliebter Lösungsansatz in der Mathematik ist die Überführung (Transformation) eines Problems in ein gleichwertiges, das aber einfacher zu lösen ist. Es ist z. B. recht aufwändig, zwei große Zahlen a und b zu multiplizieren. Früher wurden Logarithmen mittels Tabellen bestimmt. So konnte man das Problem durch Anwenden des Logarithmus transformieren, es im transformierten Zustand lösen und die Lösung dann zurücktransformieren:

- Ursprungsproblem: Es ist $a \cdot b$ für große positive Zahlen auszurechnen.
- Transformatiertes Problem: Berechne $\ln(a \cdot b)$. Dieses Problem lässt sich nun durch Addition einfach lösen: $\ln(a \cdot b) = \ln(a) + \ln(b)$.
- Rücktransformation führt zur Lösung des Ursprungsproblems: $a \cdot b = \exp(\ln(a) + \ln(b))$.

Mit diesem Trick kann man leicht die Ableitung des Produkts von n positiv-wertigen Funktionen $f_1, \dots, f_n$ berechnen. Während die Ableitung einer Summe die Summe der Ableitungen ist, ist das beim Produkt nicht so einfach. Hier muss man wiederholt die Produktregel anwenden (siehe Aufgabe 2.23). Mit der Transformation über den Logarithmus erhält man aber direkt:

$$\begin{aligned}
\frac{d}{dx} \prod_{k=1}^n f_k(x) &= \frac{d}{dx} \exp \left(\ln \left(\prod_{k=1}^n f_k(x) \right) \right) = \frac{d}{dx} \exp \left(\sum_{k=1}^n \ln(f_k(x)) \right) \\
&= \exp \left(\sum_{k=1}^n \ln(f_k(x)) \right) \sum_{k=1}^n \frac{d}{dx} \ln(f_k(x)) = \left(\prod_{k=1}^n f_k(x) \right) \sum_{k=1}^n \frac{f'_k(x)}{f_k(x)}.
\end{aligned}$$

Das folgende Lemma zeigt, dass z. B. die Ableitung unter der Fourier-Transformation zu einer Multiplikation wird, so dass man leichter ableiten kann.

Lemma 6.7 (Rechenregeln für die Fourier-Transformation)

Seien $f, g : \mathbb{R} \rightarrow \mathbb{C}$ Fourier-transformierbar. Dann gelten analog zu den Fourier-Koeffizienten, siehe Lemma 6.1 auf Seite 831, die folgenden Rechenregeln für alle $\omega \in \mathbb{R}$:

a) **Linearität:** Für Konstanten $a, b \in \mathbb{C}$ gilt: $[af(t) + bg(t)]^\wedge(\omega) = af^\wedge(\omega) + bg^\wedge(\omega)$.

b) **Verschiebung (Translation):**

i) $[f(t+h)]^\wedge(\omega) = e^{jh\omega} f^\wedge(\omega)$ für $h \in \mathbb{R}$.

ii) $[e^{-jht} f(t)]^\wedge(\omega) = f^\wedge(\omega + h)$ für $h \in \mathbb{R}$.

c) **Streckung:** $[f(at)]^\wedge(\omega) = \frac{1}{a} f^\wedge\left(\frac{\omega}{a}\right)$ für $a \in \mathbb{R}$, $a > 0$.

d) **Konjugation:** $[f(t)]^\wedge(\omega) = f^\wedge(-\omega)$.

e) **Ableitung im Zeitbereich:** f möge zusätzlich stetig differenzierbar sein (separat für Real- und Imaginärteil) mit $\lim_{t \rightarrow \pm\infty} f(t) = 0$, f' sei Fourier-transformierbar:

$$(f')^\wedge(\omega) = j\omega f^\wedge(\omega).$$

f) **Ableitung im Frequenzbereich:** Ist $t \cdot f(t)$ Fourier-transformierbar, so ist $f^\wedge$ differenzierbar mit

$$(-jt \cdot f(t))^\wedge(\omega) = \frac{d}{d\omega} f^\wedge(\omega).$$

Bemerkungen:

- Bei der Regel a) darf man nur konstante Faktoren wie a oder b aus der Transformation herausziehen. Dies gilt nicht für von t abhängige Funktionen. Falsch ist $[tf(t)]^\wedge(\omega) = t f^\wedge(\omega)$.
- Konzentriert man eine Funktion auf eine Umgebung der Null, indem man die Argumente mit $a > 1$ multipliziert, so besagt Regel c), dass die zugehörige Transformierte dazu gegenläufig ausgedehnt wird (nicht zu vermeidender Trade-off).
- Die Regel e) macht die Fourier-Transformation für das Lösen von Differenzialgleichungen interessant: Die Transformierte einer Ableitung ist gleich der Transformierten der Ursprungsfunktion, multipliziert mit j und der Variable. Das motiviert die später diskutierte Modifikation der Fourier-Transformation zur Laplace-Transformation. In den Wirtschaftswissenschaften werden zum Beispiel die Black-Scholes-Differenzialgleichungen, mit denen man Optionspreise berechnen kann, mittels der Fourier-Transformation gelöst.
- Die Anwendung mehrerer Regeln nacheinander macht in Übungsaufgaben immer wieder Probleme. Man darf nur dann eine Regel anwenden, wenn man exakt die beschriebene Situation vorfindet. Beispielsweise ist

$$[f(3t+4)]^\wedge(\omega) = \frac{1}{3} [f(t+4)]^\wedge\left(\frac{\omega}{3}\right) = \frac{1}{3} e^{j4\frac{\omega}{3}} f^\wedge\left(\frac{\omega}{3}\right).$$

Im ersten Schritt haben wir die Regel c) auf $g(3t)$ mit der Funktion $g(x) := f(x+4)$ angewendet. Im zweiten Schritt haben wir dann Regel b) i) verwendet, wobei allerdings die Transformierte hier nicht an der Stelle ω , sondern an der Stelle $\frac{\omega}{3}$

betrachtet wird. Wendet man die Regeln in umgekehrter Reihenfolge an, dann ergibt sich ebenfalls:

$$[f(3t+4)]^\wedge(\omega) = \left[f \left(3 \left(t + \frac{4}{3} \right) \right) \right]^\wedge(\omega) = e^{j\frac{4}{3}\omega} [f(3t)]^\wedge(\omega) = \frac{1}{3} e^{j\frac{4}{3}\omega} f^\wedge \left(\frac{\omega}{3} \right).$$

Beispiel 6.13

Für den Rechteckimpuls f aus (6.28) mit $f^\wedge(\omega) = 2 \operatorname{sinc} \omega$ gilt:

$$\frac{d}{d\omega} 2 \operatorname{sinc}(\omega) = (-j t f(t))^\wedge(\omega) = -j \int_{-1}^1 t e^{-j\omega t} dt.$$

■

Wir kommen zum Beweis von Lemma 6.7.

Beweis: Unter Berücksichtigung des unbeschränkten Integrationsbereichs lassen sich die Aussagen **a)**, **b)**, **d)** und **e)** analog zum Beweis von Lemma 6.1 (siehe Seite 831) zeigen, wobei wir das wegen der besonderen Bedeutung nur für **e)** tun.

c) Mit der Substitution $v = at$, $dv = a dt$, erhalten wir (siehe (2.55)):

$$(f(at))^\wedge(\omega) = \lim_{u \rightarrow \infty} \int_{-u}^u f(at) e^{-j\omega t} dt = \frac{1}{a} \lim_{u \rightarrow \infty} \int_{-au}^{au} f(v) e^{-j\frac{\omega}{a}v} dv = \frac{1}{a} f^\wedge \left(\frac{\omega}{a} \right).$$

e) Mittels partieller Integration erhält man wegen $\lim_{t \rightarrow \pm\infty} f(t) = 0$:

$$\begin{aligned} (f')^\wedge(\omega) &= \lim_{u \rightarrow \infty} \int_{-u}^u f'(t) e^{-j\omega t} dt \\ &= \lim_{u \rightarrow \infty} \left[\underbrace{\left[f(t) e^{-j\omega t} \right]_{-u}^u}_{\rightarrow 0, u \rightarrow \infty} + j\omega \int_{-u}^u f(t) e^{-j\omega t} dt \right] = j\omega f^\wedge(\omega). \end{aligned}$$

f) Wir haben unter strengerem als den hier gegebenen Voraussetzungen den Satz 4.14 auf Seite 656 zur Ableitung von Parameterintegralen formuliert. Da der Satz von Fubini aber auch für unbeschränkte Integrationsbereiche formuliert werden kann und $[tf(t)]^\wedge(\omega)$ nach Voraussetzung existiert und als Fourier-Transformierte stetig ist (siehe Seite 864), kann der Beweis des Satzes 4.14 an die hier gegebenen Voraussetzungen angepasst werden. Lange Rede kurzer Sinn: Wir dürfen Ableitung und Integral vertauschen:

$$\begin{aligned} \frac{d}{d\omega} f^\wedge(\omega) &= \frac{d}{d\omega} \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt = \int_{-\infty}^{\infty} f(t) \frac{d}{d\omega} e^{-j\omega t} dt \\ &= \int_{-\infty}^{\infty} f(t) \left[-j t e^{-j\omega t} \right] dt = (-j t f(t))^\wedge(\omega). \end{aligned}$$

□

Beispiel 6.14

Wir berechnen für $g(t) := 4f(t) + e^{jt}f(8t)$ mit den soeben bewiesenen Rechenregeln eine Darstellung der Fourier-Transformierten, in der nur Funktionswerte von $f^\wedge$ verwendet werden:

$$\begin{aligned} [\mathcal{F}g(t)](\omega) &= 4f^\wedge(\omega) + \mathcal{F}[e^{jt}f(8t)](\omega) = 4f^\wedge(\omega) + \mathcal{F}f(8t)(\omega - 1) \\ &= 4f^\wedge(\omega) + \frac{1}{8}f^\wedge\left(\frac{\omega - 1}{8}\right). \end{aligned} \quad \blacksquare$$

Mit der Linearität a) und der Ableitungsregel e) können wir Differenzialgleichungen lösen, bei denen eine Funktion $y(t)$ gesucht wird. Dazu werden beide Seiten der Gleichung transformiert, wobei die Ableitungen entfallen. Betrachten wir beispielsweise die inhomogene lineare Differenzialgleichung

$$y'(t) + 3y(t) = f(t).$$

Unter der oft zu starken Voraussetzung, dass die auftretenden Funktionen Fourier-transformierbar sind, wird daraus mit a) und e):

$$j\omega y^\wedge(\omega) + 3y^\wedge(\omega) = f^\wedge(\omega) \iff y^\wedge(\omega) = \frac{1}{3 + j\omega} \cdot f^\wedge(\omega).$$

Hier sehen wir, warum man Integraltransformationen zur Lösung von Differenzialgleichungen verwendet. Durch das Integrieren lassen sich mit der Regel zur partiellen Integration (siehe Beweis zur Regel e)) die Ableitungen entfernen. Nach der Transformation findet sich in der Gleichung nur noch die unbekannte Funktion $y^\wedge$, so dass wir danach auflösen können. Es handelt sich um eine einfache algebraische Gleichung ohne Ableitungen.

Hat man so $y^\wedge$ berechnet, dann muss man lediglich $y^\wedge$ zurücktransformieren, um die ursprünglich gesuchte Lösung y zu erhalten. Dabei gibt es im Beispiel aber ein kleines Problem, da wir das Produkt von $\frac{1}{3 + j\omega}$ und $f^\wedge(\omega)$ zurücktransformieren müssen. Mit der Funktion

$$g(t) := \begin{cases} e^{-3t}, & t \geq 0, \\ 0, & t < 0 \end{cases} \quad (6.36)$$

stellen wir dieses Produkt anders dar. Sie hat die Fourier-Transformierte

$$\begin{aligned} g^\wedge(\omega) &= \int_{-\infty}^{\infty} g(t)e^{-j\omega t} dt = \int_0^{\infty} e^{-3t}e^{-j\omega t} dt \\ &= \lim_{u \rightarrow \infty} \left[-\frac{1}{3 + j\omega} e^{(-3-j\omega)t} \right]_0^u = -\frac{1}{3 + j\omega} [0 - 1] = \frac{1}{3 + j\omega}. \end{aligned}$$

Damit haben wir für $y^\wedge$ gezeigt:

$$y^\wedge(\omega) = g^\wedge(\omega) \cdot f^\wedge(\omega).$$

Unsere verbleibende Aufgabe besteht also darin, eine Funktion y zu finden, deren Transformierte das Produkt von $\hat{g}$ und $\hat{f}$ ist. Bei den Fourier-Koeffizienten haben wir solche Produkte im Rahmen des Faltungssatzes (siehe Seite 837) gesehen. Sie entsprachen dort den Fourier-Koeffizienten einer Faltung. Das gilt auch für die Fourier-Transformation, wie wir im nächsten Abschnitt sehen. Wenn wir die Faltung zur Verfügung haben, rechnen wir hier weiter und lösen noch eine ähnliche Differenzialgleichung aus der Elektrotechnik (siehe (6.38)).

6.2.5 Faltung

Die Ableitung $(f \cdot g)'$ ist im Allgemeinen **nicht** das Produkt von f' und g' . Hier muss man die Produktregel anwenden. Analog gilt in der Regel **nicht** $(f \cdot g)^\wedge(\omega) = \hat{f}(\omega) \cdot \hat{g}(\omega)$. Dies ist ein gerne gemachter Fehler. Man benötigt etwas Analoges zur Produktregel, und das sind Faltungssätze, die wir schon von Fourier-Koeffizienten her kennen.

Definition 6.4 (Faltung nicht-periodischer Funktionen)

Die Faltung zweier Funktionen $f, g : \mathbb{R} \rightarrow \mathbb{C}$ ist definiert über (vgl. Seite 834)

$$(f * g)(t) := \int_{-\infty}^{\infty} f(t-u)g(u) du,$$

sofern dieses Integral (bezogen auf Real- und Imaginärteil) existiert.

Mittels Substitution (siehe (2.55)) zeigt man auch hier die Kommutativität

$$[f * g](t) = [g * f](t) = \int_{-\infty}^{\infty} g(t-u)f(u) du.$$

Ein Produkt im Zeitbereich führt zu einer Faltung im Frequenzbereich, und eine Faltung im Zeitbereich führt zu einem Produkt im Frequenzbereich. Damit beginnen wir:

Satz 6.6 (Faltungssatz für die Faltung im Zeitbereich)

Seien $f, g : \mathbb{R} \rightarrow \mathbb{C}$ für die $\int_{-\infty}^{\infty} |f(t)| dt < \infty$ und $\int_{-\infty}^{\infty} |g(t)| dt < \infty$. Dann ist $f * g$ Fourier-transformierbar, und es gilt:

$$(f * g)^\wedge(\omega) = \hat{f}(\omega) \cdot \hat{g}(\omega). \quad (6.37)$$

Nach den Rechenregeln für die Ableitung ist der Faltungssatz ein weiteres Beispiel dafür, wie die Fourier-Transformation Rechnungen vereinfacht: Die schwierige Berechnung der Faltung wird unter Anwendung der Fourier-Transformation zu einer einfachen Multiplikation (der Transformierten der Ursprungsfunktionen).

Den Satz beweist man analog zum Faltungssatz für Fourier-Koeffizienten (Satz 6.1). Allerdings muss man dabei den unbeschränkten Integrationsbereich berücksichtigen.

Die eingangs gestellte Frage, wie man das Produkt zweier Funktionen Fourier-transformiert, beantwortet jetzt diese Variante des Faltungssatzes:

Folgerung 6.2 (Faltungssatz für die Faltung im Frequenzbereich)

Seien die Funktionen f und g Elemente des Schwartz-Raums (siehe (6.32)), d. h. so gewählt, dass sowohl die Fourier-Transformation als auch die Fourier-Umkehrtransformation in der Form (6.30) anwendbar sind. Dann gilt:

$$(f \cdot g)^\wedge(\omega) = \frac{1}{2\pi} [f^\wedge * g^\wedge](\omega).$$

Beweis: Wir fassen $f \cdot g$ als Produkt im Frequenzbereich auf. Nach (6.37) ist $f \cdot g = \mathcal{F}([\mathcal{F}^{-1}f] * [\mathcal{F}^{-1}g])$, so dass

$$\begin{aligned} (f \cdot g)^\wedge(\omega) &\stackrel{(6.31)}{=} 2\pi \mathcal{F}^{-1}(f \cdot g)(-\omega) \\ &= 2\pi [\mathcal{F}^{-1}f] * [\mathcal{F}^{-1} \cdot g](-\omega) \stackrel{(6.31)}{=} 2\pi \left[\frac{1}{2\pi} f^\wedge(-t) \right] * \left[\frac{1}{2\pi} g^\wedge(-t) \right](-\omega) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f^\wedge(\omega + u) g^\wedge(-u) du = \frac{1}{2\pi} \int_{-\infty}^{\infty} f^\wedge(\omega - u) g^\wedge(u) du = \frac{1}{2\pi} [f^\wedge * g^\wedge](\omega). \end{aligned}$$

□

Die Voraussetzungen in Folgerung 6.2 können noch abgeschwächt werden. Man muss lediglich dafür sorgen, dass alle im Beweis benutzten Transformationen und Umkehrtransformationen erklärt sind. Der Satz gilt z. B. auch, wenn nur eine der beiden Funktionen f aus dem Schwartz-Raum ist und die andere Funktion g integrierbar ist mit $\int_{-\infty}^{\infty} |g(t)| < \infty$.

Man benötigt den Satz in der Praxis, wenn man nur ein endliches Zeitfenster eines Signals zur Verfügung hat. Dann multipliziert man das Signal im Zeitbereich mit einer sogenannten Fensterfunktion, um durch die Faltung im Frequenzbereich störende Artefakte zu verhindern (siehe Seite 950).

Beispiel 6.15 (Theoretischer Tiefpassfilter)

Wir falten mit einer Funktion, die im Frequenzbereich außerhalb eines Intervalls $[-R, R]$ verschwindet (z. B. mit $\frac{1-\cos t}{t^2}$, deren Transformierte auf $\mathbb{R} \setminus [-1, 1]$ null ist, siehe (6.33)). Die Fourier-Transformierte der Faltung ist dann außerhalb dieses Intervalls ebenfalls 0, d. h., höhere Frequenzen werden unterdrückt. ■

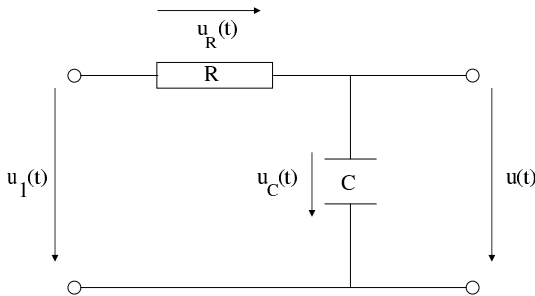


Abb. 6.13: R/C-Glied

Wir haben bereits auf Seite 873 gesehen, dass bei der Überführung der Differenzialgleichung $y'(t) + 3y(t) = f(t)$ in den Frequenzbereich ein Produkt $g^\wedge \cdot f^\wedge$ mit der Transformaten der Inhomogenität auftritt. Hier hilft jetzt der Faltungssatz:

$$y(t) = \mathcal{F}^{-1}(g^\wedge \cdot f^\wedge)(t) = \mathcal{F}^{-1}((g * f)^\wedge)(t) = (g * f)(t) = (f * g)(t) = \int_0^\infty f(t-u)e^{-3u} du.$$

Wir betrachten ein ähnliches Beispiel aus der Elektrotechnik:

Beispiel 6.16 (R/C-Glied als Tiefpassfilter)

Ein einfacher analoger passiver Tiefpassfilter für Spannungen ist das **R/C-Glied** (vgl. mit dem R/L-Kreis ab Seite 733). Dabei handelt es sich um einen Spannungsteiler, der aus einem Ohm'schen Widerstand R und einem Kondensator mit Kapazität C besteht. Der Wechselstrom-Widerstand des Kondensators ist frequenzabhängig. Bei niedrigen Frequenzen ist der Widerstand groß, bei großen Frequenzen wird er klein. Dann fällt nahezu die gesamte Spannung am Ohm'schen Widerstand ab. Greifen wir also die Spannung am Kondensator ab, erhalten wir hauptsächlich tiefe Frequenzen der Ursprungsspannung. Sei u_1 die Eingangsspannung, u_R die Spannung am Widerstand und u_C die Spannung am Kondensator. Die am Spannungsteiler abgegriffene Spannung sei $u = u_C$. Dann gilt: $u_R(t) = RQ'(t)$, wobei $Q(t)$ die Ladung des Kondensators ist und $Q'(t)$ der daraus resultierende Strom $i(t)$. Mit $Q(t) = Cu_C(t)$ ist $u_R(t) = RCu_C'(t)$. Nach Maschenregel ist $u_R(t) + u_C(t) = u_1(t)$, so dass wir insgesamt erhalten:

$$0 = -u_1(t) + RCu_C'(t) + u_C(t) = -u_1(t) + RCu'(t) + u(t).$$

Diese inhomogene Differenzialgleichung

$$RCu'(t) + u(t) = u_1(t) \quad (6.38)$$

lösen wir nun mittels Fourier-Transformation. Dazu nehmen wir an, dass die gegebene Spannungsfunktion $u_1(t)$ Fourier-transformierbar ist. Wir transformieren beide Seiten und erhalten wegen Linearität und Ableitungsregel ($RC \neq 0$)

$$RCj\omega u^\wedge(\omega) + u^\wedge(\omega) = u_1^\wedge(\omega), \quad u^\wedge(\omega) = \frac{1}{1 + j\omega RC} \cdot u_1^\wedge(\omega).$$

Der Multiplikation auf der rechten Seite entspricht im Zeitbereich eine Faltung mit der Inhomogenität:

$$u(t) = \left[\mathcal{F}^{-1} \left(\frac{1}{1 + j\omega RC} \right) \right] (t) * u_1(t).$$

Analog zu (6.36) ist $[\mathcal{F}f](\omega) = \frac{1}{1+j\omega RC}$ mit

$$f(t) := \left[\mathcal{F}^{-1} \left(\frac{1}{1 + j\omega RC} \right) \right] (t) = \frac{1}{RC} \begin{cases} e^{-\frac{1}{RC}t}, & t \geq 0, \\ 0, & t < 0 \end{cases}. \quad (6.39)$$

Man erhält eine Lösung der Differenzialgleichung als Faltung der Inhomogenität (der Eingangsspannung $u_1(t)$) mit der Funktion (6.39). Das RC -Glied wirkt so als Tiefpassfilter auf die Eingangsspannung. Das Filter ist allerdings nicht besonders gut, da

$$\left| \frac{1}{1 + j\omega RC} \right| = \frac{1}{\sqrt{1 + \omega^2 (RC)^2}}$$

für $|\omega| \rightarrow \infty$ nur langsam gegen 0 strebt und so höhere Frequenzen nur wenig gedämpft werden. ■

Hintergrund: Mehrdimensionale Fourier-Reihen und Fourier-Transformation

Die Theorie der Fourier-Reihen lässt sich direkt auf Funktionen mit mehreren Variablen übertragen. Ist $f : \mathbb{R}^n \rightarrow \mathbb{C}$ 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 $k_1 t_1 + \dots + k_n t_n$. Die **Fourier-Koeffizienten** sind hier

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

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

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

oder quadratische Partialsummen (Summation über Punkte aus $\mathbb{Z}^n$, die in einem „ n -dimensionalen Quadrat“ liegen)

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

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 $f : \mathbb{R}^n \rightarrow \mathbb{C}$ eine Funktion, für die

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

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. $f(t_1, t_2, \dots, t_n) = f_1(t_1) \cdot f_2(t_2) \cdot \dots \cdot f_n(t_n)$, so wird aus der mehrdimensionalen Fourier-Transformation das Produkt eindimensionaler Fourier-Transformationen:

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

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 Laplace-Transformation

6.3.1 Von der Fourier- zur Laplace-Transformation

Das Problem bei der Fourier-Transformation ist, dass das Integral zu ihrer Berechnung nur unter zusätzlichen Voraussetzungen wie

$$\int_{-\infty}^{\infty} |f(t)| dt < \infty$$

existiert. Wenn der Grenzwert $\lim_{t \rightarrow \pm\infty} f(t)$ erklärt ist, dann muss er gleich null sein.

Lösungen von Differenzialgleichungen setzen sich häufig aus Exponentialfunktionen zusammen. Bereits $f(t) = e^t$ erfüllt die Konvergenzbedingung nicht: $\int_{-\infty}^{\infty} |f(t)| dt =$

$\int_{-\infty}^{\infty} e^t dt = \infty$. Mit der Fourier-Transformation findet man daher keine entsprechenden Lösungen, sie ist trotz der schönen Rechenregel für Ableitungen nicht gut zum Lösen von Differenzialgleichungen geeignet. Der Trick der Integraltransformation, durch partielle Integration eine Ableitung zu eliminieren, ist aber zu schön, um an dieser Stelle aufzugeben. Daher wurde die Fourier-Transformation zur Laplace-Transformation modifiziert. Dabei nutzt man aus, dass häufig Differenzialgleichungen als Anfangswertproblem definiert sind und man nur an Lösungen interessiert ist, die erst ab einem Startzeitpunkt $t = 0$ (Anfangsbedingung) erklärt sind.

Für solche Funktionen f , die für $t < 0$ identisch 0 sind, kann die Konvergenzbedingung häufig erzwungen werden, indem man statt f die Funktion $f(t) \exp(-\delta t)$ mit $\delta > 0$ betrachtet und transformiert. Der Faktor $\exp(-\delta t)$ strebt dabei für $t \rightarrow \infty$ schnell gegen null und kann damit für die Existenz des Integrals sorgen. Dagegen ist $\lim_{t \rightarrow -\infty} \exp(-\delta t) = \infty$. Daher ist es gut, dass wir $f(t) = 0$ für $t < 0$ verlangen, so dass dieses divergente Verhalten keine Rolle spielt.

Wählen wir beispielsweise für $f(t) = \begin{cases} e^t, & \text{für } t \geq 0 \\ 0, & \text{für } t < 0 \end{cases}$ einen Wert $\delta > 1$, so ist

$$\int_{-\infty}^{\infty} |f(t) \exp(-\delta t)| dt = \int_0^{\infty} \exp(-[\delta - 1]t) dt = \frac{1}{\delta - 1} < \infty.$$

Man modifiziert durch dieses Vorgehen die Fourier-Transformation durch Einfügen des Faktors $e^{-\delta t}$ zu

$$\int_{-\infty}^{\infty} f(t) e^{-\delta t} e^{-j\omega t} dt = \int_0^{\infty} f(t) e^{-\delta t} e^{-j\omega t} dt = \int_0^{\infty} f(t) \exp(-\underbrace{(\delta + j\omega)t}_{=:s}) dt$$

und erhält damit eine Funktion mit einer komplexen Variable, die **Laplace-Transformierte** (nach dem französischen Mathematiker Laplace, der von 1749–1827 und damit während der französischen Revolution lebte und unter Napoleon Minister war). Dabei verwenden wir wieder die komplexe Exponentialfunktion (siehe Seite 140):

$$e^z := e^{\operatorname{Re}(z)} \cdot e^{j \operatorname{Im}(z)} = e^{\operatorname{Re}(z)} \cdot [\cos(\operatorname{Im}(z)) + j \sin(\operatorname{Im}(z))].$$

Definition 6.5 (Laplace-Transformation)

Eine Funktion $f : [0, \infty[\rightarrow \mathbb{C}$ heißt genau dann **Laplace-transformierbar**, wenn das Integral

$$F(s) := [\mathcal{L}f](s) := \int_0^{\infty} f(t) \exp(-st) dt = \lim_{u \rightarrow \infty} \int_0^u f(t) \exp(-st) dt$$

für ein $s \in \mathbb{C}$ erklärt ist und konvergiert.

Ist eine Funktion f Laplace-transformierbar, dann ist über $F := \mathcal{L}f$ wieder eine Funktion definiert, die **Laplace-Transformierte** F von f .

Die Laplace-Transformation ist nach der Fourier-Transformation die zweite **Integraltransformation**, die wir kennenlernen. Die Laplace-Transformierte wird zur Abgrenzung gegen die Ausgangsfunktion mit dem entsprechenden Großbuchstaben gekennzeichnet. Man schreibt auch $f(t) \circ \bullet F(s)$ bzw. $F(s) \bullet \circ f(t)$, um den Zusammenhang zwischen f und F auszudrücken.

Beispiel 6.17

a) Für die Funktion $f(t) = t^n$ konvergiert das Laplace-Integral

$$[\mathcal{L}(t^n)](s) = \int_0^\infty t^n e^{-st} dt$$

für alle $s \in \mathbb{C}$ mit $\operatorname{Re}(s) > 0$, da e^{-st} für $t \rightarrow \infty$ schneller abnimmt, als jede Potenz von t wächst. Für $n = 1$ folgt für $\operatorname{Re}(s) > 0$ durch partielle Integration

$$\begin{aligned} [\mathcal{L}(t)](s) &= \int_0^\infty t e^{-st} dt = \lim_{r \rightarrow \infty} \left[-\frac{t}{s} e^{-st} \right]_{t=0}^{t=r} + \int_0^\infty \frac{e^{-st}}{s} dt \\ &= 0 + \lim_{r \rightarrow \infty} \left[-\frac{e^{-st}}{s^2} \right]_{t=0}^{t=r} = \frac{1}{s^2}, \end{aligned}$$

und mit Vollständiger Induktion zeigen wir

$$[\mathcal{L}(t^n)](s) = \frac{n!}{s^{n+1}}.$$

Für den Induktionsschluss folgern wir aus der Gültigkeit der Transformationsformel für n wieder mit partieller Integration

$$\begin{aligned} [\mathcal{L}(t^{n+1})](s) &= \int_0^\infty t^{n+1} e^{-st} dt = \lim_{r \rightarrow \infty} \left[-\frac{t^{n+1}}{s} e^{-st} \right]_{t=0}^{t=r} + \int_0^\infty (n+1)t^n \frac{e^{-st}}{s} dt \\ &= \frac{n+1}{s} \int_0^\infty t^n e^{-st} dt = \frac{(n+1)!}{s^{n+2}}. \end{aligned}$$

b) Für $f(t) = e^{at}$ mit $a \in \mathbb{R}$ erhalten wir

$$[\mathcal{L}(e^{at})](s) = \int_0^\infty e^{at} e^{-st} dt = \int_0^\infty e^{-(s-a)t} dt = \lim_{r \rightarrow \infty} \left[-\frac{1}{s-a} e^{-(s-a)t} \right]_{t=0}^{t=r}.$$

Für $\operatorname{Re}(s) > a$ gilt $\lim_{r \rightarrow \infty} e^{-(s-a)t} = 0$, und man erhält

$$[\mathcal{L}(e^{at})](s) = \frac{1}{s-a}.$$

Für $\operatorname{Re}(s) \leq a$ existiert das Laplace-Integral nicht. ■

Die Laplace-Transformierte ist eine Funktion einer komplexen Variable. Wir werden aber im Folgenden nicht bezüglich dieser komplexen Variable differenzieren (vgl. Seite 419) oder integrieren (vgl. Seite 685). Darauf sind wir nur am Rande eingegangen, und wir müssen auch gar nicht mit der komplexen Variable rechnen.

Wir haben also für Funktionen f mit $f(t) = 0$ für $t < 0$ die Laplace-Transformation als die Fourier-Transformation der mit dem Faktor $\exp(-\operatorname{Re}(s)t)$ gewichteten Funktion f definiert:

$$\begin{aligned} F(s) &= [\mathcal{L}f](s) = \int_0^\infty f(t) \exp(-st) dt = \int_0^\infty f(t) \exp(-\operatorname{Re}(s)t) \exp(-j \operatorname{Im}(s)t) dt \\ &= [\mathcal{F}(f(t) \exp(-\operatorname{Re}(s)t))](\operatorname{Im}(s)) = [f(t) \exp(-\operatorname{Re}(s)t)]^\wedge(\operatorname{Im}(s)). \end{aligned} \quad (6.40)$$

Der Imaginärteil von s hat die Bedeutung der Kreisfrequenz ω bei der Fourier-Transformation. Trotz des zusätzlichen Realteils von s spricht man auch bei Anwendung der Laplace-Transformation vom Übergang in den Frequenzbereich.

Satz 6.7 (Existenz- und Eindeutigkeitsatz der Laplace-Transformation)

Sei $f : [0, \infty[\rightarrow \mathbb{C}$ auf jedem endlichen Intervall $[0, b]$ integrierbar und von höchstens exponentiellem Wachstum, d. h., es existieren reelle Zahlen $M, s_0 \geq 0$ mit

$$|f(t)| \leq M \exp(s_0 t)$$

für alle $t \geq 0$ (Majorantenbedingung), dann gilt:

- a) $F(s) := [\mathcal{L}f](s)$ existiert für (mindestens) alle $s \in \mathbb{C}$ mit $\operatorname{Re}(s) > s_0$.
- b) f ist bis auf die Funktionswerte in Unstetigkeitsstellen durch F eindeutig bestimmt.
- c) $\lim_{\operatorname{Re}(s) \rightarrow \infty} F(s) = 0$.

Vielfach sind Lösungen von Differenzialgleichungen über die Exponentialfunktion gebildet und erfüllen damit das geforderte exponentielle Wachstum. Eine ähnliche Wachstumsbedingung wie diese für Funktionen haben wir bereits für Folgen in Beispiel 2.29 (Z-Transformation) auf Seite 249 benutzt.

Beweis: (Skizze) b) kann man mit der entsprechenden Aussage für die Fourier-Transformation beweisen. Insbesondere kann man analog zur Herleitung der Laplace-Transformation aus der Fourier-Transformation auch eine Laplace-Umkehrtransformation aus der Umkehrformel (6.30) für das Fourier-Integral herleiten. Dabei handelt es sich dann aber um ein komplexes Kurvenintegral, siehe Kasten auf Seite 685 und z. B. (Arens et al., 2012, S. 1151). Eine genauere Betrachtung würde hier den Rahmen sprengen.

Die anderen Aussagen folgen aus dem Majorantenkriterium Satz 2.51 für uneigentliche Integrale:

$$\begin{aligned} |f(t) \exp(-st)| &= |f(t) \exp(-\operatorname{Re}(s)t) \exp(-j \operatorname{Im}(s)t)| = |f(t) \exp(-\operatorname{Re}(s)t)| \\ &\leq M \exp(s_0 t) \exp(-\operatorname{Re}(s)t) = M \exp([s_0 - \operatorname{Re}(s)]t), \end{aligned}$$

für $\operatorname{Re}(s) > s_0$ ist $s_0 - \operatorname{Re}(s) < 0$ und

$$\begin{aligned} \int_0^\infty \exp([s_0 - \operatorname{Re}(s)]t) dt &= \lim_{r \rightarrow \infty} \int_0^r \exp([s_0 - \operatorname{Re}(s)]t) dt \\ &= \lim_{r \rightarrow \infty} \frac{1}{s_0 - \operatorname{Re}(s)} \exp([s_0 - \operatorname{Re}(s)]t) \Big|_0^r = -\frac{1}{s_0 - \operatorname{Re}(s)}. \end{aligned}$$

Damit ist $\exp([s_0 - \operatorname{Re}(s)]t)$ eine integrierbare Majorante, und $F(s)$ existiert mit

$$|F(s)| = \left| \int_0^\infty f(t) \exp(-st) dt \right| \leq \frac{M}{\operatorname{Re}(s) - s_0},$$

so dass wegen $\lim_{\operatorname{Re}(s) \rightarrow \infty} \frac{M}{\operatorname{Re}(s) - s_0} = 0$ auch die Grenzwertaussage c) folgt. $\square$

Beispiel 6.18

a) Laut Beispiel 6.17 sind $f(t) = t^n$ und $f(t) = e^{at}$ Laplace-transformierbar, beide Funktionen sind von exponentiellem Wachstum, z. B. mit der gemeinsamen Majorante $\exp(|a| \cdot t)$.

b) Für $f(t) = \frac{1}{t}$ und $f(t) = \exp(t^2)$ ist die Voraussetzung jedoch nicht erfüllt, und es gibt auch keine Laplace-Transformierte. Die Funktion $f(t) = \exp(t^2)$ erfüllt die Majoranten-Bedingung für große t nicht.

$f(t) = \frac{1}{t}$ ist nicht beschränkt in einer Umgebung des Nullpunktes, damit kann es hier keine Majorante $M \exp(s_0 t)$ geben.

In der Ingenieurliteratur wird bisweilen nur $|f(t)| \leq M \exp(s_0 t)$ für alle $t > T$ zu einer Konstante $T \geq 0$ gefordert. Da wir in diesem Buch die Integrierbarkeit von f auf endlichen Intervallen voraussetzen, ist f insbesondere auf jedem Intervall $[0, T]$ beschränkt, so dass wir $T = 0$ wählen können. Wir verzichten hier auf die Untersuchung von Funktionen f , die z. B. wie $f(t) = \frac{1}{\sqrt{t}}$ auf $[0, T]$ zwar nicht beschränkt (und damit nicht integrierbar), aber uneigentlich integrierbar sind und für die die Laplace-Transformation sinnvoll sein könnte. Die Funktion $f(t) = \frac{1}{t}$ hingegen ist nicht einmal uneigentlich integrierbar auf $[0, T]$. $\blacksquare$

6.3.2 Rechnen mit der Laplace-Transformation

Mit der Laplace-Transformation kann man Differenzialgleichungen lösen. Das liegt einerseits daran, dass im Gegensatz zur Fourier-Transformation die typischen Lösungen

Tab. 6.2: Einige Laplace-Transformierte ($f(t) = 0$ für $t < 0$, $\operatorname{Re}(s) > 0$)

Zeitfunktion $f(t)$	Transformierte $F(s)$	Zeitfunktion $f(t)$	Transformierte $F(s)$
1	$\frac{1}{s}$	$\sinh^2(at)$	$\frac{2a^2}{s(s^2-4a^2)}, \operatorname{Re}(s) > 2 a $
$\exp(at)$	$\frac{1}{s-a}, \operatorname{Re}(s) > a$	$\cosh^2(at)$	$\frac{s^2-2a^2}{s(s^2-4a^2)}, \operatorname{Re}(s) > 2 a $
$\frac{t^n}{n!} \exp(at)$	$\frac{1}{(s-a)^{n+1}}, \operatorname{Re}(s) > a$	$t \sin(\omega t)$	$2\omega \frac{s}{(s^2+\omega^2)^2}$
$\cos(\omega t)$	$\frac{s}{s^2+\omega^2}$	$t \cos(\omega t)$	$\frac{s^2-\omega^2}{(s^2+\omega^2)^2}$
$\sin(\omega t)$	$\frac{\omega}{s^2+\omega^2}$	$\exp(-\delta t) \cos(\omega t)$	$\frac{s+\delta}{(s+\delta)^2+\omega^2}, \operatorname{Re}(s) + \delta > 0$
$\sin^2(\omega t)$	$\frac{2\omega^2}{s(s^2+4\omega^2)}$	$\exp(-\delta t) \sin(\omega t)$	$\frac{\omega}{(s+\delta)^2+\omega^2}, \operatorname{Re}(s) + \delta > 0$
$\cos^2(\omega t)$	$\frac{s^2+2\omega^2}{s(s^2+4\omega^2)}$	$\frac{\delta \sin(\omega t) - \omega \sin(\delta t)}{\omega \delta (\delta^2 - \omega^2)}$	$\frac{1}{(s^2+\omega^2)(s^2+\delta^2)}, \omega \neq \delta,$ $\omega, \delta \neq 0$
$\sin(\omega t + \varphi)$	$\frac{s \sin(\varphi) + \omega \cos(\varphi)}{s^2+\omega^2}$	$\frac{\cos(\omega t) - \cos(\delta t)}{\delta^2 - \omega^2}$	$\frac{s}{(s^2+\omega^2)(s^2+\delta^2)}, \omega \neq \delta$
$\cos(\omega t + \varphi)$	$\frac{s \cos(\varphi) - \omega \sin(\varphi)}{s^2+\omega^2}$	$\frac{\sin(\omega t) - \omega t \cos(\omega t)}{2\omega^3}$	$\frac{1}{(s^2+\omega^2)^2}, \omega \neq 0$
t^n	$\frac{n!}{s^{n+1}}$	$\frac{\sin(\omega t) + \omega t \cos(\omega t)}{2\omega}$	$\frac{s^2}{(s^2+\omega^2)^2}, \omega \neq 0$
$\cosh(at)$	$\frac{s}{s^2-a^2}, \operatorname{Re}(s) > a $	$\delta(t)$ (siehe	1
$\sinh(at)$	$\frac{a}{s^2-a^2}, \operatorname{Re}(s) > a $	Seite 898)	

Laplace-transformierbar sind. Andererseits gelten Rechenregeln wie bei der Fourier-Transformation, die aus einer Differenzialgleichung eine algebraische Gleichung ohne Ableitungen machen. Diese Rechenregeln sehen wir uns jetzt an.

6.3.2.1 Rechenregeln

Im Folgenden mögen die zu transformierenden Funktionen auf jedem endlichen Intervall $[0, b]$ integrierbar sein und die Bedingung $|f(t)| \leq M \exp(s_0 t)$ erfüllen, so dass damit die Laplace-Transformierte für $\operatorname{Re}(s) > s_0 \geq 0$ existiert.

a) Linearität:

$$[\mathcal{L}(af(t) + bg(t))](s) = a[\mathcal{L}f](s) + b[\mathcal{L}g](s), a, b \in \mathbb{R}.$$

Beispiel 6.19

Für $\operatorname{Re}(s) > 4$ ist (siehe Tabelle 6.2)

$$\begin{aligned} & [\mathcal{L}(3 \exp(4t) + 7 \exp(-2t))](s) \\ &= 3[\mathcal{L}(\exp(4t))](s) + 7[\mathcal{L}(\exp(-2t))](s) = \frac{3}{s-4} + \frac{7}{s+2}. \end{aligned}$$

■

b) **Streckung:**

$$[\mathcal{L}(f(ct))](s) = \frac{1}{c} [\mathcal{L}f] \left(\frac{s}{c} \right), \quad c > 0.$$

Beispiel 6.20

$$[\mathcal{L}(\cos(3t))](s) = \frac{1}{3} [\mathcal{L}(\cos t)] \left(\frac{s}{3} \right) = \frac{1}{3} \frac{\frac{s}{3}}{\left(\frac{s}{3}\right)^2 + 1} = \frac{s}{s^2 + 9}. \quad \blacksquare$$

c) **Dämpfung:**

$$[\mathcal{L}(\exp(-at)f(t))](s) = [\mathcal{L}f](s + a), \quad a > 0.$$

Beispiel 6.21

$$[\mathcal{L}(\exp(-t)\exp(t))](s) = [\mathcal{L}(\exp(t))](s + 1) = \frac{1}{(s + 1) - 1} = \frac{1}{s} = [\mathcal{L}(1)](s). \quad \blacksquare$$

d) **Ableitung:** Sei f zusätzlich stetig differenzierbar auf $[0, \infty[$ (oder zumindest differenzierbar, wobei f' auf jedem Intervall nur endlich viele Sprungstellen besitzt):

$$[\mathcal{L}(f')](s) = s[\mathcal{L}f](s) - f(0). \quad (6.41)$$

Die Regel sieht der Ableitungsregel für die Fourier-Transformation aus Lemma 6.7 sehr ähnlich, so dass sich die Laplace-Transformation tatsächlich zum Lösen von Differenzialgleichungen eignet. Statt $j\omega$ hat man hier nur die (jetzt komplexe) Variable s als Vorfaktor. Da man aber ein Integral von 0 bis ∞ berechnet, bekommt man zusätzlich den Einfluss der Stelle 0 in Form des Funktionswerts $f(0)$.

Beispiel 6.22

$$[\mathcal{L}(\sin' t)](s) = s[\mathcal{L}(\sin t)](s) - \sin(0) = s \frac{1}{s^2 + 1} = [\mathcal{L}(\cos t)](s). \quad \blacksquare$$

Für höhere Ableitungen gilt für n -mal auf $[0, \infty[$ stetig differenzierbares f iterativ:

$$\begin{aligned} [\mathcal{L}(f^{(n)})](s) &= s^n [\mathcal{L}f](s) - f^{(n-1)}(0) - s f^{(n-2)}(0) - \dots - s^{n-1} f(0) \\ &= s^n [\mathcal{L}f](s) - \sum_{k=0}^{n-1} s^k f^{(n-1-k)}(0). \end{aligned} \quad (6.42)$$

Die Funktionswerte $f(0), f'(0), \dots, f^{(n-1)}(0)$ gehen so in die Formel ein, dass bei wachsenden Potenzen von s der Grad der Ableitung fällt. In Anwendungen sind diese Werte meist als Anfangswerte vorgegeben.

e) **Stammfunktion** für stetiges f :

$$\left[\mathcal{L} \left(\int_0^t f(u) du \right) \right] (s) = \frac{1}{s} [\mathcal{L}f](s).$$

Beispiel 6.23

$$\left[\mathcal{L} \left(\int_0^t \cos(u) du \right) \right] (s) = \frac{1}{s} [\mathcal{L}(\cos t)](s) = \frac{1}{s} \frac{s}{s^2 + 1} = [\mathcal{L} \sin t](s). \quad \blacksquare$$

f) Faltung:

$$[\mathcal{L}(f * g)](s) = [\mathcal{L}f](s) \cdot [\mathcal{L}g](s),$$

wobei hier die **Faltung** für Funktionen f und g mit $f(t) = g(t) = 0$, $t < 0$, genauso definiert ist wie bei der Fourier-Transformation (siehe Definition 6.4). Allerdings kann man das Integrationsintervall auf den Bereich beschränken, auf dem die Funktionen von null verschieden sind:

$$(f * g)(t) := \int_0^t f(t-u)g(u) du = \int_0^\infty f(t-u)g(u) du = \int_{-\infty}^\infty f(t-u)g(u) du.$$

Man beachte, dass die Faltung periodischer Funktionen analog definiert ist, allerdings wird dabei nur über ein Periodenintervall integriert und mit einem zusätzlichen Faktor normiert.

Beispiel 6.24

$$[\mathcal{L}(e^t * e^t)](s) = [\mathcal{L}(e^t)](s)^2 = \frac{1}{(s-1)^2}. \text{ Andererseits ist}$$

$$e^t * e^t = \int_0^t e^{t-u} e^u du = te^t, \quad [\mathcal{L}(te^t)](s) = \frac{1}{(s-1)^2}. \quad \blacksquare$$

Beweis: (Skizze) Die Linearität a) folgt sofort aus der Linearität des Integrals. Exemplarisch zeigen wir d) mittels partieller Integration.

$$\begin{aligned} [\mathcal{L}(f')](s) &= \int_0^\infty f'(t) \exp(-st) dt = \lim_{r \rightarrow \infty} \int_0^r f'(t) \exp(-st) dt \\ &= \lim_{r \rightarrow \infty} \left[f(r) \exp(-sr) - f(0) \exp(0) + s \int_0^r f(t) \exp(-st) dt \right] \\ &= s[\mathcal{L}(f)](s) - f(0), \end{aligned}$$

denn für $\operatorname{Re}(s) > s_0$:

$$0 \leq \lim_{r \rightarrow \infty} |f(r) \exp(-sr)| = \lim_{r \rightarrow \infty} |f(r)| \exp(-\operatorname{Re}(s)r) \leq \lim_{r \rightarrow \infty} M \exp([s_0 - \operatorname{Re}(s)]r) = 0.$$

Mit der Regel für Ableitungen ergibt sich direkt auch die Regel für Stammfunktionen mit dem Hauptsatz der Differenzial- und Integralrechnung:

$$[\mathcal{L}(f)](s) = \left[\mathcal{L} \left(\frac{d}{dt} \int_0^t f(u) du \right) \right] (s) = s \left[\mathcal{L} \left(\int_0^t f(u) du \right) \right] (s) - \int_0^0 f(u) du,$$

also

$$\left[\mathcal{L} \left(\int_0^t f(u) du \right) \right] (s) = \frac{1}{s} [\mathcal{L}(f)](s).$$

□

6.3.2.2 Lösen von Differenzialgleichungen

Analog zur Fourier-Transformation können wir nun eine Differenzialgleichung Laplace-transformieren und dabei die Linearität a) und Ableitungsregel d) aus dem letzten Abschnitt anwenden. Dann hat man eine Gleichung ohne Ableitungen und kann diese (im günstigsten Fall) im Bildraum lösen. Anschließend muss man die Lösung zurücktransformieren. Dazu lesen wir die Tabelle der Laplace-Transformationen von Standard-Funktionen (siehe Tabelle 6.2) von rechts nach links und verwenden die Rechenregeln rückwärts. Man kann mit diesem Ansatz natürlich nur Lösungen finden, die auch Laplace-transformierbar sind. Dabei ist die Ausbeute aber besser als bei der Fourier-Transformation.

Beispiel 6.25

Wir betrachten das allgemeine lineare Anfangswertproblem zweiter Ordnung mit konstanten Koeffizienten (das wir auch mit Satz 5.10 auf Seite 774 lösen können):

$$y''(t) + ay'(t) + by(t) = f(t), \quad y(0) = y_0, \quad y'(0) = y_1.$$

Wir transformieren beide Seiten und erhalten wegen der Linearität und der Ableitungsregel ($s^2 + as + b \neq 0$):

$$\begin{aligned} (s^2 Y(s) - sy_0 - y_1) + a(sY(s) - y_0) + bY(s) &= F(s) \\ \iff (s^2 + as + b)Y(s) &= F(s) + (s + a)y_0 + y_1 \\ \iff Y(s) &= \frac{F(s)}{(s^2 + as + b)} + \frac{(s + a)y_0 + y_1}{(s^2 + as + b)}. \end{aligned} \quad (6.43)$$

Der Nenner ist hier das charakteristische Polynom der Differenzialgleichung. Es entsteht durch Anwendung der Ableitungsregel. Jetzt erhält man eine Lösung des Ursprungsproblems durch Rücktransformation:

$$\begin{aligned} y(t) &= [\mathcal{L}^{-1}Y](t) = \left[\mathcal{L}^{-1} \left(F(s) \cdot \frac{1}{(s^2 + as + b)} \right) \right] (t) + \left[\mathcal{L}^{-1} \left(\frac{(s + a)y_0 + y_1}{(s^2 + as + b)} \right) \right] (t) \\ &= [\mathcal{L}^{-1}F(s)](t) * \left[\mathcal{L}^{-1} \left(\frac{1}{(s^2 + as + b)} \right) \right] (t) + \left[\mathcal{L}^{-1} \left(\frac{(s + a)y_0 + y_1}{(s^2 + as + b)} \right) \right] (t) \end{aligned}$$

$$= f(t) * \left[\mathcal{L}^{-1} \left(\frac{1}{(s^2 + as + b)} \right) \right] (t) + \left[\mathcal{L}^{-1} \left(\frac{(s+a)y_0 + y_1}{(s^2 + as + b)} \right) \right] (t).$$

Dabei haben wir den Faltungssatz f) aus dem vorangehenden Abschnitt eingesetzt. Für lineare Differenzialgleichungen höherer Ordnung kann man analog vorgehen, das sehen wir uns am Ende des Abschnitts an. ■

Hintergrund: Zusammenhang zwischen Laplace-Transformation und Lösungsformel Satz 5.10 für lineare Differenzialgleichungen höherer Ordnung

Mit Satz 5.10 (Seite 774) haben wir zunächst unabhängig von der Laplace-Transformation ein Verfahren kennengelernt, um eine partikuläre Lösung der Differenzialgleichung

$$y^{(2)}(t) + ay'(t) + by(t) = f(t)$$

basierend auf einer homogenen Lösung $y_0(t)$ zu erhalten. Wir zeigen jetzt, dass dieses Verfahren mit der Laplace-Transformation erklärt werden kann: Mit $x_0 = 0$ und Inhomogenität $q(t) = f(t)$ lautet die dort angegebene Formel für $n = 2$

$$y_p(x) = \int_0^x y_0(x-t)f(t) dt = [y_0 * f](x),$$

wobei y_0 eine spezielle homogene Lösung mit $y_0(0) = 0$ und $y'_0(0) = 1$ ist. y_p wurde also bereits in Satz 5.10 über ein Faltungsintegral angegeben, nur dass wir die Faltung damals noch nicht kannten.

Nach Wahl von $y_0(t)$ ist $0 = y''_0(t) + ay'_0(t) + by_0(t)$ und damit Laplace-transformiert

$$\begin{aligned} 0 &= s^2 Y_0(s) - y'_0(0) - s y_0(0) + a(s Y_0(s) + y_0(0)) + b Y_0(s) \\ &= s^2 Y_0(s) - 1 + a s Y_0(s) + b Y_0(s) = [s^2 + as + b] Y_0(s) - 1. \end{aligned}$$

Demnach ist $Y_0(s) = \frac{1}{s^2 + as + b}$. Wir können diese homogene Lösung durch Rücktransformation ausrechnen: Das charakteristische Polynom $s^2 + as + b$ hat die Nullstellen $s_0 = -\frac{a}{2} + \sqrt{\frac{a^2}{4} - b}$ und $s_1 = -\frac{a}{2} - \sqrt{\frac{a^2}{4} - b}$. Sind diese verschieden, erhalten wir eine Partialbruchzerlegung

$$Y_0(s) = \frac{1}{s_0 - s_1} \left[\frac{1}{s - s_0} - \frac{1}{s - s_1} \right],$$

und eine Rücktransformation liefert

$$y_0(t) = \frac{1}{s_0 - s_1} [e^{s_0 t} - e^{s_1 t}].$$

Sind die Nullstellen $s_1 = s_2$ gleich, so lautet die Partialbruchzerlegung

$$Y_0(s) = \frac{1}{(s - s_0)^2} + \frac{0}{s - s_0},$$

und wir erhalten über die Rücktransformation $y_0(t) = te^{s_0 t}$.

Dieses Ergebnis entspricht dem in Satz 5.9 auf Seite 773 dargestellten Vorgehen für homogene Gleichungen, bei dem man linear unabhängige Lösungen direkt aus den Nullstellen des charakteristischen Polynoms gewinnt. Wählt man in der dort berechneten allgemeinen homogenen Lösung die Konstanten so, dass die Anfangsbedingung $y_0(0) = 0$ und $y'_0(0) = 1$ erfüllt ist, ergibt sich das hier berechnete $y_0(t)$.

Nun aber zurück zur inhomogenen Lösung $y_p = y_0 * f$ aus Satz 5.10:

$$[\mathcal{L}y_p](s) = [\mathcal{L}(y_0 * f)](s) = [\mathcal{L}y_0](s) \cdot [\mathcal{L}f](s) = Y_0(s) \cdot F(s) = \frac{F(s)}{s^2 + as + b}.$$

Damit ist aber $\mathcal{L}y_p$ die Transformierte Y einer Lösung, die wir mit (6.43) für die Anfangswerte $y_0 = y_1 = 0$ berechnet haben. y_p muss daher mit dieser Lösung übereinstimmen. Damit haben wir für Differenzialgleichungen zweiter Ordnung nachgerechnet, dass die in Satz 5.10 angegebene Funktion y_p tatsächlich eine partikuläre Lösung der Differenzialgleichung ist.

Ein Faltungsintegral tritt sowohl auf bei Anwendung von Satz 5.10 als auch bei Verwendung der Laplace-Transformation im Rahmen der Rücktransformation von (6.43). Allerdings kann man hier bei bestimmten Inhomogenitäten auf das Ausrechnen der Faltung verzichten, wenn man die Umkehrtransformation des Produkts durch Ablesen aus einer Tabelle ermitteln kann.

Beispiel 6.26

Wir setzen im vorangehenden Beispiel 6.25 konkrete Daten ein:

$$y''(t) + 4y(t) = \sin(\omega t), \quad \omega > 0.$$

Als allgemeinen Anfangswert wählen wir $y(0) = c_0$, $y'(0) = c_1$ und erhalten mit der obigen Rechnung für $a = 0$, $b = 4$ und $f(t) = \sin(\omega t)$:

$$Y(s) = \frac{[\mathcal{L}(\sin(\omega t))](s)}{(s^2 + 4)} + \frac{sc_0 + c_1}{(s^2 + 4)}.$$

Setzen wir die bekannte Laplace-Transformierte von $f(t)$ ein, ergibt sich für $s > 0$:

$$Y(s) = \frac{\frac{\omega}{s^2 + \omega^2}}{(s^2 + 4)} + \frac{sc_0 + c_1}{(s^2 + 4)} = \frac{\omega}{(s^2 + \omega^2)(s^2 + 4)} + \frac{sc_0 + c_1}{(s^2 + 4)}.$$

Die Rücktransformation geschieht nun durch Ablesen von bekannten Urbildern in Tabelle 6.2 unter Berücksichtigung der Rechenregeln (ohne Ausrechnen eines Faltungsintegrals).

$$\begin{aligned} \left[\mathcal{L}^{-1} \left(\frac{sc_0 + c_1}{(s^2 + 4)} \right) \right] (t) &= c_0 \left[\mathcal{L}^{-1} \left(\frac{s}{(s^2 + 2^2)} \right) \right] (t) + \frac{c_1}{2} \left[\mathcal{L}^{-1} \left(\frac{2}{(s^2 + 2^2)} \right) \right] (t) \\ &= c_0 \cos(2t) + \frac{c_1}{2} \sin(2t). \end{aligned}$$

$$\left[\mathcal{L}^{-1} \left(\frac{\omega}{(s^2 + \omega^2)(s^2 + 4)} \right) \right] (t) = \omega \frac{2 \sin(\omega t) - \omega \sin(2t)}{2\omega(4 - \omega^2)} = \frac{2 \sin(\omega t) - \omega \sin(2t)}{2(4 - \omega^2)},$$

falls $\omega \neq 2$. Im Fall $\omega = 2$ ist

$$\begin{aligned} \left[\mathcal{L}^{-1} \left(\frac{\omega}{(s^2 + \omega^2)(s^2 + 4)} \right) \right] (t) &= 2 \left[\mathcal{L}^{-1} \left(\frac{1}{(s^2 + 2^2)^2} \right) \right] (t) = 2 \frac{\sin(2t) - 2t \cos(2t)}{2 \cdot 2^3} \\ &= \frac{1}{8} (\sin(2t) - 2t \cos(2t)). \end{aligned}$$

Insgesamt erhalten wir damit

$$y(t) = c_0 \cos(2t) + \frac{c_1}{2} \sin(2t) + \begin{cases} \frac{2 \sin(\omega t) - \omega \sin(2t)}{2(4 - \omega^2)}, & \omega \neq 2, \\ \frac{1}{8} (\sin(2t) - 2t \cos(2t)), & \omega = 2. \end{cases} \quad \blacksquare$$

Ist die Laplace-transformierte Lösung eine gebrochen rationale Funktion, so hilft wie oben eine Partialbruchzerlegung, um die Funktion mit der Tabelle 6.2 zurückzutransformieren. Im folgenden Beispiel hat die rechte Seite bereits direkt die Gestalt einer Partialbruchzerlegung:

Beispiel 6.27

Wir bestimmen die Lösung y des folgenden Anfangswertproblems mittels Laplace-Transformation:

$$y'(t) + 2y(t) = \exp(-2t), \quad y(0) = 1.$$

Dabei verwenden wir aus Tabelle 6.2 die Regel

$$\left[\mathcal{L} \left(\frac{t^n}{n!} \exp(at) \right) \right] (s) = \frac{1}{(s - a)^{n+1}}, \quad s > a, \quad n \in \mathbb{N}_0.$$

$$[\mathcal{L}(y'(t) + 2y(t))](s) = [\mathcal{L}(\exp(-2t))](s) = \frac{1}{s + 2},$$

wegen $[\mathcal{L}(y'(t) + 2y(t))](s) = sY(s) - y(0) + 2Y(s) = sY(s) + 2Y(s) - 1$ ist

$$Y(s) = \frac{1}{(s + 2)^2} + \frac{1}{s + 2}.$$

Rücktransformation liefert $y(t) = t \exp(-2t) + \exp(-2t)$. ■

Beispiel 6.28 (Differenzialgleichungen höherer Ordnung)

Wir haben zuvor lineare Differenzialgleichungen erster und zweiter Ordnung gelöst. Mit der Laplace-Transformation können wir genauso lineare Anfangswertprobleme beliebiger n -ter Ordnung

$$y^{(n)}(t) + a_1 y^{(n-1)}(t) + a_2 y^{(n-2)}(t) + \cdots + a_n y(t) = f(t),$$

$$y^{(k)}(0) = y_k, \quad 0 \leq k < n,$$

mit konstanten reellen Koeffizienten $a_1, \dots, a_n$, Laplace-transformierbarer Inhomogenität f und den Anfangswerten $y_0, \dots, y_{n-1}$ lösen. Dabei gelangt man zur Gleichung

$$s^n Y(s) - \sum_{k=0}^{n-1} s^k y_{n-1-k} + a_1 \left[s^{n-1} Y(s) - \sum_{k=0}^{n-2} s^k y_{n-2-k} \right] + \dots + a_n Y(s) = F(s).$$

Löst man diese nach $Y(s)$ auf, erhält man

$$Y(s) = \frac{1}{s^n + a_1 s^{n-1} + \dots + a_n s^0} \left[F(s) + \sum_{k=0}^{n-1} s^k y_{n-1-k} + a_1 \sum_{k=0}^{n-2} s^k y_{n-2-k} + \dots \right],$$

wobei der Nenner wieder das charakteristische Polynom der Differenzialgleichung ist. Damit erhalten wir unter Verwendung des Faltungssatzes

$$\begin{aligned} y &= \left[\mathcal{L}^{-1} \left(\frac{1}{s^n + a_1 s^{n-1} + \dots + a_n s^0} \right) \right] * f \\ &+ \mathcal{L}^{-1} \left(\frac{\sum_{k=0}^{n-1} s^k y_{n-1-k} + a_1 \sum_{k=0}^{n-2} s^k y_{n-2-k} + \dots}{s^n + a_1 s^{n-1} + \dots + a_n s^0} \right). \end{aligned} \quad (6.44)$$

Nun besteht die Aufgabe lediglich noch darin, die Umkehrtransformation der beiden gebrochen-rationalen Funktionen zu bestimmen. Dazu berechnet man jeweils eine Partialbruchzerlegung. In den Zerlegungen treten bis auf Faktoren nur Summanden der Form

$$\frac{1}{(s-a)^k}, \quad \frac{s+c}{[(s-a)^2+b^2]^k}$$

mit $a, b, c \in \mathbb{R}$ auf. Die Umkehrfunktion des ersten Terms sehen wir in Tabelle 6.2 nach:

$$\mathcal{L}^{-1} \left(\frac{1}{(s-a)^k} \right) (t) = \frac{t^{k-1}}{(k-1)!} \exp(at).$$

Für den zweiten Term (komplexe Nullstellen $a \pm jb$) betrachten wir zunächst den

- Fall $k=1$: $\mathcal{L}^{-1} \left(\frac{s+c}{(s-a)^2+b^2} \right) (t) = e^{at} \cos(bt) + \frac{c+a}{b} e^{at} \sin(bt)$,
denn laut Tabelle 6.2 ist

$$\mathcal{L}(e^{at} \cos(bt)) = \frac{s-a}{(s-a)^2+b^2}, \quad \mathcal{L}(e^{at} \sin(bt)) = \frac{b}{(s-a)^2+b^2},$$

$$\begin{aligned} \mathcal{L} \left(e^{at} \cos(bt) + \frac{c+a}{b} e^{at} \sin(bt) \right) &= \frac{s-a}{(s-a)^2+b^2} + \frac{c+a}{b} \frac{b}{(s-a)^2+b^2} \\ &= \frac{s+c}{(s-a)^2+b^2}. \end{aligned}$$

Damit ist umgekehrt

$$\mathcal{L}^{-1} \left(\frac{s+c}{(s-a)^2+b^2} \right) (t) = e^{at} \cos(bt) + \frac{c+a}{b} e^{at} \sin(bt).$$

- Ist $k > 1$, so ist die Umkehrtransformierte eines Produkts von $\frac{s+c}{[(s-a)^2+b^2]}$ mit $k-1$ Termen $\frac{1}{(s-a)^2+b^2}$ zu berechnen, das sich wieder über die Faltung der einzelnen Umkehrtransformierten ergibt. Dazu benötigen wir nur noch zusätzlich, dass

$$\mathcal{L}\left(\frac{1}{b}e^{at}\sin(bt)\right)(s) = \frac{1}{(s-a)^2+b^2} :$$

$$\begin{aligned} \mathcal{L}^{-1}\left(\frac{s+c}{[(s-a)^2+b^2]^k}\right)(t) \\ = \left(e^{at}\cos(bt) + \frac{c+a}{b}e^{at}\sin(bt)\right) * \underbrace{\left(\frac{1}{b}e^{at}\sin(bt)\right) * \dots * \left(\frac{1}{b}e^{at}\sin(bt)\right)}_{(k-1)\text{-mal}}. \end{aligned}$$

Mit einer Partialbruchzerlegung und diesen Rücktransformationen kann so die Lösung (6.44) vollständig ausgerechnet werden. ■

Wir vergleichen abschließend die Laplace-Transformation und die Fourier-Transformation als Lösungsmethoden für Differenzialgleichungen:

- Im Gegensatz zur Laplace-Transformation führt die Ableitungsregel der Fourier-Transformation zu keiner Anfangsbedingung. Dies liegt daran, dass wir hier von $-\infty$ bis ∞ integrieren und nicht wie bei Laplace das Integral erst bei null starten.
- Funktionen mit exponentiellem Wachstum sind Laplace- aber nicht Fourier-transformierbar. Man findet solche Lösungen nicht mit dem Fourier-Ansatz. Da die Exponentialfunktion ein Grundbaustein von Lösungen von Differenzialgleichungen ist, ist dies der entscheidende Vorteil der Laplace-Transformation.

6.3.2.3 Grenzwertsätze

Den Aufwand für die Rücktransformation einer Laplace-Transformierten $F(s)$ kann man sich unter Umständen ersparen, wenn man nur an den Werten $f(0)$ und $\lim_{t \rightarrow \infty} f(t)$ interessiert ist. In der Regelungstechnik ist gerade dieser Grenzwert einer mittels Laplace-Transformation berechneten Lösung einer Differenzialgleichung wichtig. Existiert er, so nennt man die Lösung **stabil**. Sonst heißt die Lösung **instabil**.

Die Werte $f(0)$ und $\lim_{t \rightarrow \infty} f(t)$ können direkt aus der Transformierten $F(s)$ über Grenzwerte berechnet werden. Den Anfangswertsatz, der $f(0)$ liefert, erhält man unter Ausnutzung der Ableitungsregel (6.41), wobei wir jetzt in die Variable s nur reelle Zahlen einsetzen (da wir in diesem Buch $s \rightarrow \infty$ für komplexe Zahlen s nicht definiert haben und wir uns hier nicht mit dem Übergang zum Betrag oder Realteil von s belasten möchten):

$$\lim_{s \rightarrow \infty} sF(s) = \lim_{s \rightarrow \infty} s[\mathcal{L}(f)](s) \stackrel{(6.41)}{=} \lim_{s \rightarrow \infty} [\mathcal{L}(f')](s) + f(0) = f(0).$$

Im letzten Schritt der Rechnung haben wir die Grenzwertaussage $\lim_{s \rightarrow \infty} [\mathcal{L}g](s) = 0$ für eine Laplace-transformierbare Funktion g aus Satz 6.7 (Seite 881) verwendet. Der Grenzwert $\lim_{s \rightarrow \infty} sF(s)$ ist damit auch insbesondere vom Typ $[\infty \cdot 0]$ und kann mit dem Satz von L'Hospital berechnet werden.

Wir haben in der Rechnung die Ableitung f' verwendet. Der Anfangswertsatz lässt sich aber auch ohne eine Voraussetzung an die Differenzierbarkeit von f zeigen:

Satz 6.8 (Anfangswertsatz)

Sei die Funktion $f : [0, \infty[\rightarrow \mathbb{C}$ auf $[0, \infty[$ stetig und von höchstens exponentiellem Wachstum. Dann gilt ($s \in \mathbb{R}$):

$$\lim_{s \rightarrow \infty} sF(s) = f(0).$$

Der Beweis unter diesen Voraussetzungen geschieht in der Übungsaufgabe 6.12.

Eine ähnliche Aussage gilt für den Endwert $f(\infty) := \lim_{t \rightarrow \infty} f(t)$, falls dieser existiert. Die Existenz ist alles andere als selbstverständlich, da eine Funktion von exponentiellem Wachstum recht schnell gegen ∞ streben darf.

Satz 6.9 (Endwertsatz)

Sei die Funktion $f : [0, \infty[\rightarrow \mathbb{C}$ stetig auf $[0, \infty[$, so dass der Grenzwert $\lim_{t \rightarrow \infty} f(t) = c$ als (komplexe) Zahl existiert. Dann gilt ($s \in \mathbb{R}$):

$$\lim_{s \rightarrow 0+} sF(s) = \lim_{t \rightarrow \infty} f(t).$$

Da der Grenzwert $\lim_{t \rightarrow \infty} f(t)$ existiert, ist die stetige Funktion f auf $[0, \infty[$ beschränkt. Damit ist sie insbesondere von exponentiellem Wachstum, so dass die Laplace-Transformation erklärt ist. $F(s)$ existiert für jedes $s \in \mathbb{R}$ mit $s > s_0 := 0$.

Wie der Anfangswertsatz lässt sich auch der Endwertsatz über die Ableitungsregel (6.41) herleiten. Allerdings ist hier der Umgang mit den Grenzwerten etwas schwieriger:

$$\begin{aligned} \lim_{s \rightarrow 0+} sF(s) &\stackrel{(6.41)}{=} \lim_{s \rightarrow 0+} [\mathcal{L}(f')](s) + f(0) = \lim_{s \rightarrow 0+} \int_0^{\infty} f'(t) e^{-st} dt + f(0) \\ &= \int_0^{\infty} \left[\lim_{s \rightarrow 0+} f'(t) e^{-st} \right] dt + f(0) = \int_0^{\infty} f'(t) dt + f(0) \\ &= \lim_{u \rightarrow \infty} [f(t)]_0^u + f(0) = \lim_{u \rightarrow \infty} f(u). \end{aligned}$$

Bei dieser Rechnung haben wir Integral und Grenzwert vertauscht. Das ist leider nicht generell erlaubt. Außerdem haben wir die Ableitung benötigt. Im elementaren Beweis des Satzes kommen wir ohne diese Einschränkungen aus (vgl. Aufgabe 6.12):

Beweis: Nach Definition der Laplace-Transformation ist

$$sF(s) = s \int_0^\infty f(t)e^{-st} dt = s \int_0^{\frac{1}{\sqrt{s}}} f(t)e^{-st} dt + s \int_{\frac{1}{\sqrt{s}}}^\infty f(t)e^{-st} dt.$$

Hier haben wir das Laplace-Integral in zwei Teile zerlegt, die wir für kleine $s > 0$ näher untersuchen. Da die Funktion f für $t \rightarrow \infty$ konvergiert, wissen wir etwas über die Funktionswerte für große t . Das werden wir zur Berechnung des zweiten Integrals einsetzen: Die Grenze $\frac{1}{\sqrt{s}}$, an der wir das Integral aufspalten, ist so gewählt, dass sie für $s \rightarrow 0+$ gegen unendlich geht und wir damit das Wissen über f verwenden können. Andererseits ist die Grenze so gewählt, dass das erste Integral gerade noch keinen signifikanten Beitrag liefert.

Wie zuvor bemerkt, ist f wegen der Stetigkeit und der Konvergenz auf $[0, \infty[$ beschränkt, z. B. mit $M > 0$. Damit erhalten wir für das erste Integral

$$\left| s \int_0^{\frac{1}{\sqrt{s}}} f(t)e^{-st} dt \right| \leq s \int_0^{\frac{1}{\sqrt{s}}} \underbrace{|f(t)|}_{\leq M} e^{-st} dt \leq M \left[s \left(-\frac{1}{s} \right) e^{-st} \right]_{t=0}^{t=\frac{1}{\sqrt{s}}} = -M [e^{-\sqrt{s}} - e^0].$$

Für $s \rightarrow 0+$ strebt die rechte Seite und damit auch die linke Seite gegen 0. Für das zweite Integral nutzen wir aus, dass $\lim_{t \rightarrow \infty} f(t) = c$ ist. Zu einem beliebig gewählten $\varepsilon > 0$ gibt es damit ein $t_0 = t_0(\varepsilon) > 0$, so dass $|f(t) - c| < \varepsilon$ für alle Werte $t > t_0$. Insbesondere ist also dafür $|\operatorname{Re}(f(t)) - \operatorname{Re}(c)| < \varepsilon$ und $|\operatorname{Im}(f(t)) - \operatorname{Im}(c)| < \varepsilon$. Sei nun $0 < s < \frac{1}{t_0^2}$, d. h. $\frac{1}{\sqrt{s}} > t_0$. Für diese s gilt:

$$\begin{aligned} \operatorname{Re} \left(s \int_{\frac{1}{\sqrt{s}}}^\infty f(t)e^{-st} dt \right) &= s \int_{\frac{1}{\sqrt{s}}}^\infty \operatorname{Re}(f(t))e^{-st} dt \leq s \int_{\frac{1}{\sqrt{s}}}^\infty [\operatorname{Re}(c) + \varepsilon]e^{-st} dt \\ &= [\operatorname{Re}(c) + \varepsilon] \lim_{u \rightarrow \infty} \left[s \left(-\frac{1}{s} \right) e^{-st} \right]_{t=\frac{1}{\sqrt{s}}}^{t=u} = [\operatorname{Re}(c) + \varepsilon]e^{-\sqrt{s}}. \end{aligned}$$

Analog zeigt man $\operatorname{Re} \left(s \int_{\frac{1}{\sqrt{s}}}^\infty f(t)e^{-st} dt \right) \geq [\operatorname{Re}(c) - \varepsilon]e^{-\sqrt{s}}$.

Da wieder $\lim_{s \rightarrow 0+} e^{-\sqrt{s}} = 1$ ist und außerdem $\varepsilon > 0$ beliebig gewählt wurde, ist

$$\lim_{s \rightarrow 0+} \operatorname{Re} \left(s \int_{\frac{1}{\sqrt{s}}}^\infty f(t)e^{-st} dt \right) = \operatorname{Re}(c).$$

Entsprechendes erhält man für den Imaginärteil, so dass insgesamt gilt:

$$\lim_{s \rightarrow 0+} s \int_{\frac{1}{\sqrt{s}}}^\infty f(t)e^{-st} dt = c.$$

Zusammen erhalten wir $\lim_{s \rightarrow 0+} sF(s) = 0 + c = \lim_{t \rightarrow \infty} f(t)$. □

Im Beweis wird die Existenz des Grenzwerts $\lim_{t \rightarrow \infty} f(t)$ zweimal benötigt. Ist diese nicht gegeben, so gilt der Satz nicht. Für die Funktion $f(t) = \sin t$ ist $F(s) = \frac{1}{s^2+1}$, so dass der Grenzwert $\lim_{s \rightarrow 0+} sF(s) = \lim_{s \rightarrow 0+} \frac{s}{s^2+1} = 0$ existiert. Aber $\lim_{t \rightarrow \infty} f(t)$ existiert nicht.

Beispiel 6.29

- a) Für $f(t) = 1 + 3e^{-t}$ ist offensichtlich $f(0) = 4$ und $\lim_{t \rightarrow \infty} f(t) = 1$. Die gleichen Werte erhalten wir über den Anfangs- und Endwertsatz: Nach Tabelle 6.2 ist $F(s) = \frac{1}{s} + \frac{3}{s+1}$. $\lim_{s \rightarrow \infty} sF(s) = \lim_{s \rightarrow \infty} \left[1 + \frac{3s}{s+1}\right] = 4$ und $\lim_{s \rightarrow 0+} sF(s) = \lim_{s \rightarrow 0+} \left[1 + \frac{3s}{s+1}\right] = 1$.
- b) Für die Funktion $f(t) = \exp(-t) \sin t$ ist $f(0) = \lim_{t \rightarrow \infty} f(t) = 0$. Auch diese Werte erhalten wir über den Anfangs- und den Endwertsatz. Nach Tabelle 6.2 ist $F(s) = \frac{1}{(s+1)^2+1}$. Damit ist $\lim_{s \rightarrow \infty} sF(s) = \lim_{s \rightarrow \infty} \frac{s}{s^2+2s+2} = 0 = f(0)$ sowie $\lim_{s \rightarrow 0+} sF(s) = \frac{s}{s^2+2s+2} \Big|_{s=0} = 0 = \lim_{t \rightarrow \infty} f(t)$. ■

6.3.3 Laplace-Transformation in der Systemtheorie *

Die Laplace-Transformation findet überwiegend in der Systemtheorie und der Regelungstechnik ihre Anwendung. Darauf wollen wir hier kurz eingehen.

6.3.3.1 Lineare zeitinvariante Übertragungssysteme

Ein großer Teil der **linearen zeitinvarianten Übertragungssysteme** (sogenannte LTI-Systeme, wobei LTI für „Linear-Time-Invariant“ steht) wird durch gewöhnliche lineare Differenzialgleichungen n -ter Ordnung

$$\begin{aligned} a_n \frac{d^n}{dt^n} y(t) + a_{n-1} \frac{d^{n-1}}{dt^{n-1}} y(t) + \cdots + a_1 \frac{d}{dt} y(t) + a_0 y(t) \\ = b_m \frac{d^m}{dt^m} x(t) + b_{m-1} \frac{d^{m-1}}{dt^{m-1}} x(t) + \cdots + b_1 \frac{d}{dt} x(t) + b_0 x(t) \end{aligned}$$

mit Koeffizienten $a_k, b_k \in \mathbb{R}$, $a_n \neq 0$ und $b_m \neq 0$ beschrieben. Aus dem Eingangssignal $x(t)$, $t \geq 0$, erzeugt das System ein Ausgangssignal $y(t)$, $t \geq 0$. Die Inhomogenität auf der rechten Seite der Differenzialgleichung berechnet sich aus dem gegebenen Eingangssignal $x(t)$. Gesucht ist ein Ausgangssignal $y(t)$, das zu einer Inhomogenität berechnet werden soll, die sich aus den Ableitungen des Eingangssignals zusammensetzt.

Man spricht von Zeitinvarianz, da die Koeffizienten a_k und b_k nicht von t abhängig sind. Bei einem zeitlich verschobenen Eingangssignal $x(t + t_0)$ ist ein entsprechend verschobenes Ausgangssignal Lösung der Gleichung. Ist $y(t)$ eine zu $x(t)$ gehörende Lösung, so ist $y(t + t_0)$ eine zu $x(t + t_0)$ gehörende Lösung. Das lässt sich wegen $\frac{d}{dt}[y(t + t_0)] = \left[\frac{d}{dt}y\right](t + t_0)$ durch Einsetzen in die Differenzialgleichung verifizieren.

Hat man die Anfangsbedingung $y(0) = y'(0) = \dots = y^{(n-1)}(0) = x(0) = \dots = x^{(m-1)}(0) = 0$, so erhält man mit der Ableitungsregel (6.42)

$$[a_n s^n + a_{n-1} s^{n-1} + \dots + a_1 s + a_0][\mathcal{L}y](s) = [b_m s^m + b_{m-1} s^{m-1} + \dots + b_1 s + b_0][\mathcal{L}x](s),$$

indem man beide Seiten der Differenzialgleichung Laplace-transformiert. Die Funktion

$$H(s) = \frac{Y(s)}{X(s)} = \frac{(\mathcal{L}y)(s)}{(\mathcal{L}x)(s)} = \frac{b_m s^m + b_{m-1} s^{m-1} + \dots + b_1 s + b_0}{a_n s^n + a_{n-1} s^{n-1} + \dots + a_1 s + a_0}$$

heißt **Übertragungsfunktion** des Systems.

Die Übertragungsfunktion lässt sich direkt aus der Differenzialgleichung ablesen, ist unabhängig vom Eingangssignal $x(t)$ und damit wohldefiniert. Mit der Übertragungsfunktion $H(s)$ wird die Dynamik eines Übertragungssystems vollständig beschrieben: Hat man ein Eingangssignal $x(t)$ mit der Anfangsbedingung $x(0) = \dots = x^{(m-1)}(0) = 0$, so erhält man im Bildbereich (transformierten Zustand) sofort das (wegen der Anfangsbedingung $y(0) = y'(0) = \dots = y^{(n-1)}(0) = 0$ eindeutige) Ausgangssignal über die Multiplikation mit der Übertragungsfunktion:

$$(\mathcal{L}y)(s) = H(s)(\mathcal{L}x)(s). \quad (6.45)$$

Beispiel 6.30

Wir betrachten einfache elektrische Bauteile als Regler, die ein Eingangssignal in ein Ausgangssignal überführen.

a) An einem Kondensator gilt $i(t) = Cu'(t)$.

Wir betrachten $u(t)$ als Eingangssignal mit $u(0) = 0$ und $i(t)$ als Ausgangssignal.

Dann ist $H(s) = \frac{Cs}{1} = Cs$ und $(\mathcal{L}i)(s) = Cs(\mathcal{L}u)(s)$.

b) Ist am Kondensator $i(t)$ Eingangs- und $u(t)$ Ausgangssignal, so gilt $H(s) = \frac{1}{Cs}$.

c) Für einen Ohm'schen Widerstand R gilt die „Differenzialgleichung“ $u(t) = Ri(t)$.

Für das Eingangssignal $i(t)$ und das Ausgangssignal $u(t)$ erhalten wir die Übertragungsfunktion $H(s) = R$.

d) Ist $u(t)$ das Eingangssignal am Widerstand, so ist $H(s) = \frac{1}{R}$. ■

Der Nenner der Übertragungsfunktion ist das charakteristische Polynom der Differenzialgleichung. Über die Nullstellen dieses Polynoms findet man mit Satz 5.9 (Seite 773) die homogenen Lösungen des Übertragungssystems, also die Lösungen für $x(t) = 0$.

Hat eine Nullstelle s_0 einen positiven Realteil, so gibt es mit $e^{s_0 t}$ eine homogene Lösung, deren Betrag für $t \rightarrow \infty$ gegen Unendlich strebt, die also instabil ist (vgl. Kapitel 6.3.2.3).

Eine spezielle Lösung der inhomogenen Differenzialgleichung erhält man durch Rücktransformation von (6.45) in den Zeitbereich als Faltungsintegral

$$y(t) = (h * x)(t) = \int_0^t h(t-u)x(u) du \quad (6.46)$$

mit der **Gewichtsfunktion** $h(t) = [\mathcal{L}^{-1}H](t)$ als Laplace-Umkehrtransformierte der Übertragungsfunktion $H(s)$. Hier sieht man sehr schön, dass in $y(t)$ nur Werte $x(u)$ mit $u \leq t$ eingehen, der Systemzustand zum Zeitpunkt t also durch die Vergangenheit $u < t$ bestimmt ist.

Ist die Übertragungsfunktion unbekannt, so kann sie experimentell bestimmt werden, indem ein geeignetes Eingangssignal angelegt wird. Optimal wäre eine Funktion $x(t)$ mit $\mathcal{L}x(s) = 1$. Dann hätten wir $H(s) = (\mathcal{L}y)(s)$, wobei $y(t)$ als Antwort gemessen werden kann. Leider gibt es aber z. B. wegen Satz 6.7 c) auf Seite 881 keine Funktion, deren Laplace-Transformierte die Konstante eins ist. Man verwendet daher verallgemeinerte Funktionen und hier speziell die **Dirac'sche δ -Distribution**. Man kann sich δ als eine „Funktion“ vorstellen, die bis auf die Stelle 0 überall den Wert null annimmt und an der Stelle 0 den Wert ∞ hat, so dass das Integral über die Funktion den Wert eins ergibt. Der Wert ∞ verletzt die oben geforderte Anfangsbedingung $x(0) = 0$ für $m > 0$. Das ist aber wegen Eigenschaften der Distribution kein Problem (siehe Hintergrundinformationen ab Seite 898, insbesondere gilt wegen (6.50): $\left[\mathcal{L} \left(b_m \frac{d^m}{dt^m} \delta + b_{m-1} \frac{d^{m-1}}{dt^{m-1}} \delta + \dots + b_1 \frac{d}{dt} \delta + b_0 \delta \right) \right] (s) = b_m s^m + b_{m-1} s^{m-1} + \dots + b_0$). Als Systemantwort auf diesen kurzen, aber heftigen Impuls erhält man die Gewichtsfunktion, und ihre Laplace-Transformierte ist die Übertragungsfunktion.

Neben der Impulsantwort kann man die Übertragungsfunktion aus der Reaktion $y(t)$ auf eine Sprungfunktion $x(t)$ berechnen. Sei $x(t) := 1$ für $t \geq 0$ und $x(t) = 0$ für $t < 0$. Für die Laplace-Transformation verhält sich $x(t)$ wie die konstante Funktion 1 mit $(\mathcal{L}x)(s) = \frac{1}{s}$. Allerdings fassen wir $x(t)$ wieder als verallgemeinerte Funktion auf, deren Ableitung die δ -Distribution (und nicht die Nullfunktion, siehe (6.49)) ist. In diesem Sinne kann man zeigen, dass die Laplace-Transformierte der k -ten Ableitung $\frac{1}{s} s^k$ ist. Damit folgt $Y(s) = H(s) \cdot \frac{1}{s}$, also $H(s) = sY(s)$. Die Antwort $y(t) = \left[\mathcal{L}^{-1} \left(\frac{H(s)}{s} \right) \right] (t)$ auf die Sprungfunktion nennt man **Übergangsfunktion** oder **Sprungantwort**.

In der Systemtheorie sind einige Begriffe sehr wichtig, die auf der Übertragungsfunktion basieren. Setzt man $s = j\omega$ (also $\operatorname{Re}(s) = 0$) in die Übertragungsfunktion ein, so erhält man eine Funktion mit einer reellen Variable ω , die **Frequenzgang** $H(j\omega)$ heißt. Verwendet man statt der Laplace- die Fourier-Transformation, um zu einer Eingangsfunktion die Ausgangsfunktion eines LTI-Systems zu berechnen, dann erhält man statt der Übertragungsfunktion den Frequenzgang. Schreibt man den Frequenzgang in Polarkoordinatenform $H(j\omega) = A(\omega)e^{j\varphi(\omega)}$, so heißt $A(\omega)$ der **Amplitudengang** und $\varphi(\omega)$ der **Phasengang**. Trägt man die komplexen Werte für $\omega \in [0, \infty[$ als Punkte in der komplexen Ebene auf, so entsteht eine Kurve, die man **Ortskurve** nennt. Analog zu Definition 4.22 (siehe Seite 674) kann man sie mit einem unbeschränkten Parameterintervall über eine Parametrisierung $((\operatorname{Re} H(j\omega), \operatorname{Im} H(j\omega)), [0, \infty[)$ mit der Parametervariable ω darstellen.

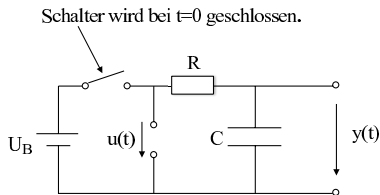


Abb. 6.14: Laden eines Kondensators

Beispiel 6.31

Ein Kondensator wird über eine Gleichspannungsquelle geladen. Zum Zeitpunkt $t = 0$ wird dazu eine konstante Spannung $u(t) = U_B$ an eine Reihenschaltung aus Kondensator mit Kapazität C und Ohm'schem Widerstand R angelegt, siehe Abbildung 6.14. Für $t < 0$ ist $u(t) = 0$ (Sprungfunktion), so dass $[\mathcal{L}u](s) = U(s) = \frac{U_B}{s}$. Die Spannung am Kondensator wird mit $y(t)$ bezeichnet. Gesucht ist $\lim_{t \rightarrow \infty} y(t)$, also die Spannung, die der Kondensator nach sehr langer Ladezeit hat. Diese ermitteln wir über den Endwertsatz der Laplace-Transformation.

Für die Spannungen gilt: $u(t) - y(t) - Ri(t) = 0$ und $i(t) = Cy'(t)$, also $Cy'(t) + \frac{1}{R}y(t) = \frac{1}{R}u(t)$ mit der Anfangsbedingung $y(0) = 0$. Die Übertragungsfunktion kann damit sofort abgelesen werden: $H(s) = \frac{\frac{1}{R}}{Cs + \frac{1}{R}} = \frac{1}{RCs + 1}$.

Damit erhalten wir über den Endwertsatz (siehe Satz 6.9)

$$\lim_{t \rightarrow \infty} y(t) = \lim_{s \rightarrow 0+} sY(s) = \lim_{s \rightarrow 0+} sH(s)U(s) = \lim_{s \rightarrow 0+} s \frac{1}{RCs + 1} \frac{U_B}{s} = U_B.$$

Asymptotisch stellt sich also die Ladespannung ein.

Den vollständigen Spannungsverlauf erhalten wir durch Rücktransformation von

$$Y(s) = H(s)U(s) = \frac{1}{RCs + 1} \frac{U_B}{s} = \frac{\frac{U_B}{RC}}{\left(s + \frac{1}{RC}\right)s} = -\frac{U_B}{s + \frac{1}{RC}} + \frac{U_B}{s}$$

$$\text{zu } y(t) = U_B \left[-\exp\left(-\frac{t}{RC}\right) + 1 \right].$$

Die Gewichtsfunktion ist in diesem Beispiel $h(t) = [\mathcal{L}^{-1}H](t) = \frac{1}{RC}e^{-\frac{t}{RC}}$. Damit können wir $y(t)$ alternativ (als Sprungantwort) auch so gewinnen:

$$\begin{aligned} y(t) &= h(t) * u(t) = \int_0^t \frac{1}{RC} e^{-\frac{t-w}{RC}} u(w) dw = \frac{U_B}{RC} \int_0^t e^{-\frac{t-w}{RC}} dw \\ &= \frac{U_B}{RC} e^{-\frac{t}{RC}} \int_0^t e^{\frac{w}{RC}} dw = U_B e^{-\frac{t}{RC}} \left[e^{\frac{w}{RC}} \right]_{w=0}^{w=t} = U_B \left(1 - e^{-\frac{t}{RC}} \right). \end{aligned}$$

Wir nutzen das Beispiel auch noch zur Illustration des Anfangswertsatzes. Durch Messen von Spannungswerten $y(t)$ kann man den Wert für RC ermitteln. Da $y'(t) = U_B \frac{1}{RC} \exp\left(-\frac{t}{RC}\right)$ ist, gilt $RC = \frac{U_B}{y'(0)}$. Das gleiche Ergebnis erhält man auch mit den

Sätzen zur Laplace-Transformation. Zunächst gilt wegen der Ableitungsregel (siehe (6.41) und beachte, dass die Spannung $y(0)$ am Kondensator zunächst 0 ist), dass $[\mathcal{L}y'](s) = sY(s) - y(0) = sY(s) = \frac{U_B}{RCs+1}$. Mit dem Anfangswertsatz (Satz 6.8) erhalten wir damit ebenfalls $RC = \frac{U_B}{y'(0)}$, da

$$y'(0) = \lim_{s \rightarrow \infty} s[\mathcal{L}y'](s) = \lim_{s \rightarrow \infty} s \frac{U_B}{RCs+1} = \frac{U_B}{RC}. \quad \blacksquare$$

Betrachtet man Ein- und Ausgangssignale nicht als Funktionen, sondern als Folgen einzelner Messwerte, so kann man mit der Z-Transformation (siehe Beispiel 2.29 auf Seite 249) ähnlich arbeiten, wie wir es hier mit der Laplace-Transformation tun, siehe z. B. Goebbels (2014). Das liegt daran, dass sich die Z-Transformation für Differenzen (bzw. Differenzenquotienten) analog zur Laplace-Transformation für Ableitungen verhält.

Hintergrund: Distributionen

Die Reaktion $y(t)$ eines linearen zeitinvarianten Übertragungssystems mit $y(0) = 0$ auf ein Eingangssignal $x(t), t \geq 0$, ist gegeben durch das Faltungsintegral (6.46) aus der Gewichtsfunktion $h(t)$ und dem Signal $x(t)$: $y(t) = \int_0^t h(t-u)x(u) du$.

In zahlreichen Anwendungen interessiert die Reaktion des technischen Systems auf eine kurzzeitig wirkende Störung $x(t)$, eine sogenannte **Impulsfunktion**, wobei sich der Gesamtimpuls $\int_{t_0}^{t_1} x(t) dt$ auf ein kleines Zeitfenster $[t_0, t_1]$ konzentriert. Man denke etwa an die Reaktion eines mechanischen Systems auf einen Hammerschlag oder die Reaktion einer elektrischen Schaltung auf einen Stromstoß.

Eine Impulsfunktion ist für kleines $\varepsilon > 0$ gegeben durch

$$\delta_\varepsilon(t) := \begin{cases} 0, & t < 0, \\ \frac{1}{\varepsilon}, & 0 < t < \varepsilon, \\ 0, & \varepsilon < t < \infty, \end{cases} \quad \text{mit} \quad \int_{-\infty}^{\infty} \delta_\varepsilon(t) dt = \int_0^\infty \delta_\varepsilon(t) dt = 1.$$

Ist der Gesamtimpuls auf den Zeitpunkt $t = 0$ konzentriert, dann müsste durch

$$\delta(t) := \lim_{\varepsilon \rightarrow 0+} \delta_\varepsilon(t) = \begin{cases} 0, & t \neq 0, \\ \infty, & t = 0, \end{cases}$$

eine „Funktion“ erklärt sein, für die $\int_{-\infty}^{\infty} \delta(t) = 1$ gilt. Eine solche Funktion gibt es nicht, da weder ∞ ein Funktionswert ist noch das Riemann'sche Integral existiert. Dennoch ist es sinnvoll, den Grenzwert der Systemreaktion auf die Eingangssignale $\delta_\varepsilon(t)$ für $\varepsilon \rightarrow 0+$ zu betrachten, denn die Gewichtsfunktion δ_ε ist ja eine „vernünftige“ Funktion:

$$y(t) = \lim_{\varepsilon \rightarrow 0+} \int_{-\infty}^{\infty} h(t-u)\delta_\varepsilon(u) du = \lim_{\varepsilon \rightarrow 0+} \int_0^t h(t-u)\delta_\varepsilon(u) du = h(t). \quad (6.47)$$

Um ohne den Umweg eines Grenzwerts mit δ rechnen zu können, erweitert man den Funktionenbegriff und kommt dabei zu verallgemeinerten Funktionen, auch **Distribu-**

tionen genannt. In der Menge der Distributionen finden sich die bekannten Funktionen wieder. Es gibt aber darüber hinaus neue Elemente wie δ . Zu einer (geeigneten) Funktion $g : \mathbb{R} \rightarrow \mathbb{C}$ ist über

$$T_g(f) := \int_{-\infty}^{\infty} g(t) \cdot f(t) dt \quad (6.48)$$

eine Distribution definiert. Dabei muss das Integral existieren. Das ist z. B. der Fall, wenn f aus dem Schwartz-Raum stammt (siehe (6.32), Seite 866) und die Funktion g langsam wachsend im Sinne von $|g(x)| \leq C(1 + |x|)^n$ für ein $C > 0$, ein $n \in \mathbb{N}_0$ und alle $x \in \mathbb{R}$ ist. Langsam wachsende Funktionen wachsen für $x \rightarrow \pm\infty$ nicht schneller als Polynome und damit erheblich langsamer als beispielsweise die Exponentialfunktion.

Die Distribution T_g bildet Funktionen (sogenannte Testfunktionen, die z. B. aus dem Schwartz-Raum stammen) auf komplexe Zahlen ab. Eine solche Abbildung nennt man auch **Funktional** (siehe Definition auf Seite 570). Sie ist also ein völlig anderes Objekt als die Ursprungsfunktion g . Es gibt aber eine Korrespondenz zwischen g und T_g . Man sagt, dass g diese Distribution erzeugt, T_g heißt **reguläre Distribution**, da sie von g erzeugt wird.

δ wird nun als Distribution definiert, die von keiner Funktion erzeugt wird, also keine reguläre Distribution ist. δ ist eine Abbildung, die den Funktionswert der Testfunktion an der Stelle 0 liefert: $\delta(f) := f(0)$. Das passt dazu, dass für eine in 0 stetige Testfunktion f gilt: $\lim_{\varepsilon \rightarrow 0+} T_{\delta_\varepsilon}(f) = \lim_{\varepsilon \rightarrow 0+} \frac{1}{\varepsilon} \int_0^\varepsilon f(t) dt = f(0)$ (siehe (2.49) auf Seite 338).

Begriffe und Eigenschaften von Funktionen kann man konsistent auf Distributionen ausdehnen, so dass man mit ihnen (fast) wie mit Funktionen rechnen kann. Die Grundidee dabei ist folgende: Führt man mit einer Funktion g Rechenoperationen aus, so soll das Ergebnis die gleiche Distribution erzeugen wie die, die bei Anwendung der analogen Rechenoperationen für Distributionen auf der von g erzeugten Distribution entsteht. So kann man beim Rechnen zwischen Funktionen und Distributionen hin- und herwechseln.

Sind beispielsweise g und g' langsam wachsende stetige Funktionen, so gilt für jede Funktion f aus dem Schwartz-Raum mittels partieller Integration ($\lim_{t \rightarrow \pm\infty} f(t) = 0$):

$$T_{g'}(f) = \int_{-\infty}^{\infty} g'(t)f(t) dt = - \int_{-\infty}^{\infty} g(t)f'(t) dt = -T_g(f'),$$

wobei mit f auch f' im Schwartz-Raum ist, also in T_g eingesetzt werden kann. Die distributionentheoretische Ableitung ist daher (vgl. mit schwachen Ableitungen in Kapitel 3.7.6)

$$T'(f) := -T(f'), \quad T^{(n)}(f) = (-1)^n T(f^{(n)}).$$

Als Beispiel leiten wir die von der Sprungfunktion $g(t) := 1$ für $t \geq 0$ und $g(t) = 0$ für $t < 0$ erzeugte reguläre Distribution ab. Dabei erhalten wir die δ -Distribution:

$$T'_g(f) = -T_g(f') = - \int_0^\infty f'(t) dt = f(0) = \delta(f). \quad (6.49)$$

Ist für eine (geeignete) Funktion g die Fourier-Transformation $\mathcal{F}g$ erklärt, so sollte also für die Transformation der zugehörigen Distribution T_g die Distribution $T_{\mathcal{F}g}$ entstehen. Unter Verwendung des Satzes von Fubini (vgl. Satz 4.13) erhält man daraus:

$$\begin{aligned} T_{g^\wedge}(f) &= \int_{-\infty}^{\infty} g^\wedge(t) \cdot f(t) dt = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u) e^{-jut} du \cdot f(t) dt \\ &= \int_{-\infty}^{\infty} g(u) \cdot \int_{-\infty}^{\infty} f(t) e^{-jut} dt du = \int_{-\infty}^{\infty} g(u) \cdot f^\wedge(u) du = T_g(f^\wedge). \end{aligned}$$

Die rechte Seite macht auch hier für Distributionen Sinn, die nicht von einer (Fourier-transformierbaren) Funktion erzeugt werden. Daher wird so die Fourier-Transformation für Distributionen definiert. Man beachte, dass für Testfunktionen f aus dem Schwartz-Raum auch $f^\wedge$ im Schwartz-Raum ist, so dass T_g auf $f^\wedge$ anwendbar ist. Damit:

$$\delta^\wedge(f) = \delta(f^\wedge) = f^\wedge(0) = \int_{-\infty}^{\infty} 1 \cdot f(t) dt = T_1(f).$$

Die Fourier-Transformierte der δ -Distribution ist also regulär und wird von der konstanten Funktion 1 erzeugt. Wir starten mit einem Funktional δ , das keine klassische Funktion ist, und erhalten als Fourier-Transformierte eine Distribution, die mit einer (konstanten) klassischen Funktion korrespondiert. Oft identifiziert man eine reguläre Distribution mit ihrer erzeugenden Funktion, also hier $\delta^\wedge(\omega) = 1$. Mit δ lässt sich wie mit einer Funktion rechnen.

Die Faltung einer Distribution mit einer (geeigneten) Funktion lässt sich so als Funktion definieren, dass der Faltungssatz gilt. Insbesondere erhält man für die δ -Distribution mit $\delta^\wedge = 1$: $(f * \delta)^\wedge = f^\wedge \cdot \delta^\wedge = f^\wedge \cdot 1$, so dass $f * \delta = f$ ist. Passend zu (6.47) ist δ neutrales Element der Faltung.

Bei der Laplace-Transformation verfährt man analog zur Fourier-Transformation. Die Laplace-Transformation einer Funktion f (mit $f(t) = 0$, $t < 0$) ergibt sich durch Anwendung einer Distribution auf die Funktion e^{-st} für $t \geq 0$, die sich für $t < 0$ zu einer Funktion aus dem Schwartz-Raum fortsetzen lässt:

$$[\mathcal{L}f](s) = \int_0^\infty f(t) e^{-st} dt = \int_{-\infty}^\infty f(t) e^{-st} dt = T_f(e^{-st}).$$

So lässt sich als Verallgemeinerung auch die δ -Distribution Laplace-transformieren:

$$\begin{aligned} [\mathcal{L}\delta](s) &= \delta(e^{-st}) = e^{-s \cdot 0} = 1, \\ [\mathcal{L}\delta^{(n)}](s) &= \delta^{(n)}(e^{-st}) = (-1)^n \delta \left(\frac{d^n}{dt^n} e^{-st} \right) = (-1)^n (-s)^n e^{-s \cdot 0} = s^n. \end{aligned} \quad (6.50)$$

Distributionen erlauben eine neue Sicht der Fourier-Reihen als Fourier-Transformierte. Dazu berechnen wir zu einem $k \in \mathbb{Z}$ die Fourier-Transformation der von e^{jku} als Funktion von u erzeugten regulären Distribution $T_{e^{jku}}$:

$$[T_{e^{jku}}]^\wedge(f) = T_{e^{jku}}(f^\wedge) = \int_{-\infty}^{\infty} e^{jku} f^\wedge(u) du = 2\pi f(k),$$

sofern im letzten Schritt die Fourier-Umkehrformel (6.30) anwendbar ist. Damit können wir nun auch (2π) -periodische Funktionen g Fourier-transformieren, indem wir sie als Distributionen auffassen. Ist g darstellbar als Fourier-Reihe $g(t) = \sum_{k=-\infty}^{\infty} g^{\wedge}(k)e^{jkt}$, so gilt, wenn wir die Konvergenz der Summe voraussetzen und die Vertauschung von Summe und Integral erlaubt ist:

$$\begin{aligned} [T_g]^{\wedge}(f) &= \int_{-\infty}^{\infty} \left[\sum_{k=-\infty}^{\infty} g^{\wedge}(k)e^{jku} \right] f^{\wedge}(u) du = \sum_{k=-\infty}^{\infty} g^{\wedge}(k) \int_{-\infty}^{\infty} e^{jku} f^{\wedge}(u) du \\ &= \sum_{k=-\infty}^{\infty} g^{\wedge}(k) [T_{e^{jku}}]^{\wedge}(f) = \sum_{k=-\infty}^{\infty} g^{\wedge}(k) 2\pi f(k) = \sum_{k=-\infty}^{\infty} g^{\wedge}(k) 2\pi \delta_k(f), \end{aligned}$$

wobei $\delta_k(f) := f(k)$, also $\delta_0(f) := f(0)$ und damit $\delta_0 = \delta$. Obwohl δ keine reelle Variable ω hat, schreibt man in Anlehnung an Funktionen auch $\delta(\omega - k) := \delta_k$. Die Fourier-Transformierte von g ist damit die Distribution $2\pi \sum_{k=-\infty}^{\infty} g^{\wedge}(k) \delta_k = 2\pi \sum_{k=-\infty}^{\infty} g^{\wedge}(k) \delta(\omega - k)$. Statt der Darstellung einer periodischen Funktion als Fourier-Reihe findet man bisweilen diese distributionentheoretische Schreibweise.

6.3.3.2 Verknüpfung von Übertragungssystemen

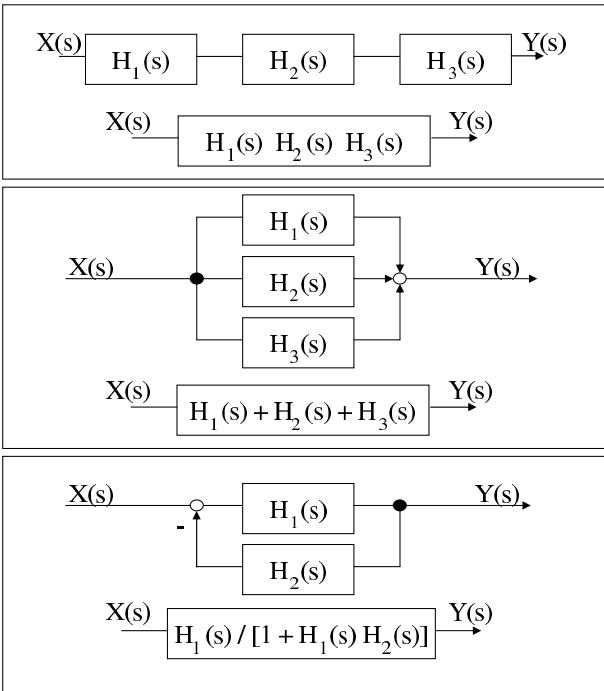


Abb. 6.15: Blockschaltbilder: Reihenschaltung, Parallelschaltung und Rückkopplung jeweils mit Ersatzschaltbild

Reihenschaltung

Schaltet man mehrere Übertragungssysteme hintereinander (das Ausgangssignal wird bei passenden Anfangsbedingungen zum Eingangssignal des nächsten Übertragungssystems, siehe erstes Blockschaltbild in Abbildung 6.15), so erhält man ein neues Übertragungssystem, dessen Übertragungsfunktion das Produkt der Einzelnen ist.

Beispiel 6.32

Wir schalten zwei Widerstände R_1 und R_2 parallel. Für R_1 ist das Eingangssignal der Strom $i(t)$. Das führt zu einer Spannung $u(t) = R_1 i(t)$ als Ausgangssignal, das als Eingangssignal an R_2 anliegt. Daraus ergibt sich ein Strom $\frac{1}{R_2} u(t) = R_1 \frac{1}{R_2} i(t)$ als Ausgangssignal des zweiten Widerstands. Die Übertragungsfunktion ist also das Produkt der beiden Einzelnen: $H(s) = \frac{R_1}{R_2}$. Man beachte, dass zwar die Widerstände parallel geschaltet, die Übertragungssysteme aber in Reihe angeordnet sind. Das liegt daran, dass das Eingangssignal des Ersten ein Strom ist, während das Eingangssignal des Zweiten kein Strom, sondern hier eine Spannung ist. ■

Parallelschaltung

Schaltet man zwei Übertragungssysteme mit Übertragungsfunktionen H_1 und H_2 parallel, so gilt für die Transformierte $Y(s)$ der Ausgangsgröße $y(t)$:

$$Y(s) = H_1(s)X(s) + H_2(s)X(s) = [H_1 + H_2](s)X(s).$$

Entsprechend ergibt sich die neue Übertragungsfunktion bei einer Parallelschaltung von n Übertragungssystemen als die Summe der einzelnen (siehe zweites Blockschaltbild in Abbildung 6.15).

Rückkopplung

Hat man eine Rückkopplung (siehe drittes Blockschaltbild in Abbildung 6.15) ohne zeitlichen Verzug, so wird vom transformierten Eingangssignal $X(s)$ das aus der Rückkopplung entstehende Signal $Y_R(s)$ abgezogen:

$$Y(s) = H_1(s)[X(s) - Y_R(s)] = H_1(s)[X(s) - H_2(s)Y(s)]$$

$$\iff [1 + H_1(s)H_2(s)]Y(s) = H_1(s)X(s).$$

Damit lautet die Übertragungsfunktion

$$\frac{Y(s)}{X(s)} = \frac{H_1(s)}{1 + H_1(s)H_2(s)}.$$

Unter Verwendung der vorangehenden Regeln für die Verknüpfung von Übertragungsfunktionen kann man eine Schaltung in eine einfach zu lösende algebraische Gleichung übertragen. Dabei wird quasi die Schaltung direkt in den Frequenzbereich transformiert, während wir zuvor erst aus der Schaltung eine Differenzialgleichung gewonnen und diese dann zur Lösung transformiert haben.

6.4 Diskrete Fourier-Transformation

Wir beschäftigen uns nun damit, wie man Fourier-Koeffizienten und die Fourier-Transformation automatisch berechnen kann. Dies geschieht heute natürlich mit dem Computer. Vor dem Computerzeitalter halfen mechanische Apparaturen wie der Harmonic Analyzer, der von Olaus Henrici (1840–1918) erfunden und von Gottlieb Coradi in Zürich gebaut wurde. Dabei hat man ausgenutzt, dass man das Integral eines Fourier-Koeffizienten (nach einer partiellen Integration) durch mechanische Längenbestimmung einer Kurve auf einer Kugeloberfläche ausrechnen kann, wobei sich die Kurve durch Überlagerung von zwei Bewegungen beim Abfahren eines Funktionsgraphen ergibt.

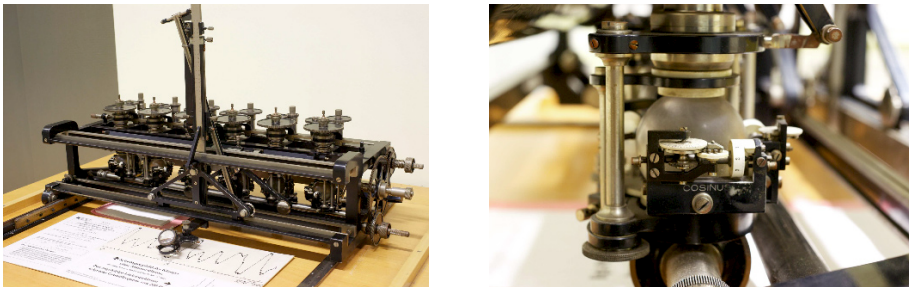


Abb. 6.16: Coradi Harmonic Analyzer an der Hochschule Niederrhein

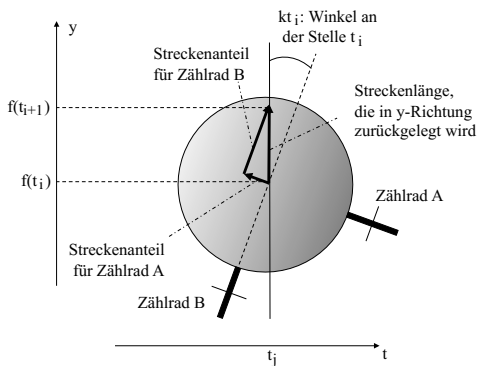


Abb. 6.17: Integration mittels Glaskugel

Hintergrund: Funktionsweise des historischen Coradi Harmonic Analyzers

Ausgangspunkt ist der gezeichnete Funktionsgraph einer vollen Periode einer (zur Vereinfachung wieder 2π -) periodischen, stetig differenzierbaren Funktion f . Hat man eine nicht-periodische Funktion, so kann man sich diese als periodisch fortgesetzt vorstellen (siehe Kapitel 6.4.7 hinsichtlich der Konsequenzen).

Aufgrund des Konvergenzsatzes für Fourier-Reihen (Satz 6.4 auf Seite 843) lässt sich f als Fourier-Reihe mit reellen Koeffizienten schreiben:

$$f(t) = a_0 + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt)).$$

Der entscheidende Trick für die Konstruktion der Maschine besteht nun darin, dass man die Fourier-Koeffizienten $a_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos(kt) dt$ und $b_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin(kt) dt$ partiell integriert und so für $k > 0$ Integrale über f' erhält:

$$\begin{aligned} \pi a_k &= \int_0^{2\pi} f(t) \cos(kt) dt = \left[\frac{1}{k} f(t) \sin(kt) \right]_0^{2\pi} - \frac{1}{k} \int_0^{2\pi} f'(t) \sin(kt) dt \\ &= -\frac{1}{k} \int_0^{2\pi} f'(t) \sin(kt) dt, \\ \pi b_k &= \frac{1}{k} \int_0^{2\pi} f'(t) \cos(kt) dt. \end{aligned}$$

Dabei haben wir ausgenutzt, dass f und auch $f(t) \sin(kt)$ sowie $f(t) \cos(kt)$ 2π -periodisch sind und damit den gleichen Funktionswert bei 2π und 0 haben.

Fährt man den Funktionsgraphen von f über ein Intervall $[a, b]$ der t -Achse von links nach rechts mit der an der Apparatur vorne angebrachten Linse ab, so bewegt sich die gesamte Vorrichtung um die parallel zur y -Achse zurückgelegten Entfernungen vor und zurück. Diese Bewegung wird über ein Rad auf eine Glaskugel mit sehr gleichmäßiger Oberfläche übertragen, die sich mit gleicher Geschwindigkeit wie die Apparatur parallel zur $y(= f(t))$ -Achse dreht. Die Kugel läuft nicht auf der Grundfläche und hat keinen Bewegungsanteil in t -Richtung.

Um die Kugel herum dreht sich ein Träger in der t - y -Ebene beim Abfahren des Graphen in Abhängigkeit von der t -Position. Die Bewegung der Linse in t -Richtung wird nicht auf die Kugel, sondern ausschließlich auf den Träger übertragen. Nach einem kompletten Periodenintervall hat er sich zur Berechnung der ersten reellen Fourier-Koeffizienten a_1 und b_1 einmal komplett gedreht. Möchte man statt der ersten die k -ten Fourier-Koeffizienten berechnen, so muss sich der Träger k -mal drehen. Dies geschieht über ein Getriebe mit auswechselbaren Zahnrädern, die an der Oberseite der Maschine angeordnet sind. Durch Austausch der Zahnräder kann man unterschiedliche Werte für k einstellen. Da die abgebildete Maschine über sechs Kugeln verfügt, können damit sechs Paare (a_k, b_k) zu unterschiedlichen k gleichzeitig bestimmt werden.

Am Träger sind in einem Winkel von 90 Grad zwei Messrädchen A und B angebracht, die auf der Kugel rollen. Im Startzustand ($t = 0$) befindet sich zunächst das Rädchen A parallel zur t -Achse und B parallel zur y -Achse. Die Rädchen summieren die von ihnen

auf der Kugeloberfläche zurückgelegten Entfernungen auf. Dabei muss die Überlagerung der Drehung der Kugel in y -Richtung und die Drehung des Trägers beachtet werden. Aus der relativen Sicht des Trägers dreht sich die Kugel mit der Geschwindigkeit, die die Maschine in y -Richtung zurücklegt, um eine Achse in der t - y -Ebene, die insgesamt k -mal gedreht wird (siehe Abbildung 6.17). An den beiden Rädchen können die Koeffizienten a_k und b_k am Ende des Periodenintervalls abgelesen werden.

Wir berechnen die Kurvenlänge, die das Rädchen A zurücklegt, indem wir diese durch kleine Teilstrecken annähern. Deren Länge können wir anschließend gegen null gehen lassen (siehe Bemerkung zu Definition 4.24 auf Seite 676). Dazu sei $0 = t_0 < t_1 < t_2 < \dots < t_n = 2\pi$ eine Zerlegung des Intervalls $[0, 2\pi]$. Zwischen zwei nahe zusammenliegenden Stellen t_i und t_{i+1} legt die Kugeloberfläche in Richtung der y -Achse die Strecke $f(t_{i+1}) - f(t_i)$ zurück. An der Stelle t_i hat die Drehbewegung der Kugel aber den Winkel $k \cdot t_i$ im Bogenmaß relativ zur Achse des Rads A des Trägers (beachte: f ist 2π -periodisch, bei anderen Winkeln kommt hier ein Umrechnungsfaktor hinzu), da sich dieser entsprechend weitergedreht hat. Der Winkel vergrößert sich bis zur Stelle t_{i+1} zu $k \cdot t_{i+1}$. Die Stellen mögen so nahe zusammenliegen, dass wir nur mit dem Winkel $k \cdot t_i$ rechnen müssen. Die Strecke, die Rad A auf der Kugeloberfläche zurücklegt, ist dann $[f(t_{i+1}) - f(t_i)] \sin(k \cdot t_i)$, die Strecke, die sich Rad B bewegt, ist $[f(t_{i+1}) - f(t_i)] \cos(k \cdot t_i)$. Insgesamt misst das Rad A ungefähr den Wert

$$\begin{aligned} \sum_{i=0}^{n-1} [f(t_{i+1}) - f(t_i)] \sin(kt_i) &= \sum_{k=0}^{n-1} \underbrace{\frac{f(t_{i+1}) - f(t_i)}{t_{i+1} - t_i}}_{=f'(\xi_i)} \sin(kt_i) (t_{i+1} - t_i) \\ &\approx \sum_{i=0}^{n-1} f'(\xi_i) \sin(k\xi_i) (t_{i+1} - t_i), \end{aligned}$$

wobei wir den Mittelwertsatz der Differenzialrechnung (Satz 2.36) verwendet haben, der die Zwischenstellen ξ_i liefert. Das ist aber eine Riemann-Zwischensumme, die für immer feinere Zerlegungen gegen das Integral $\int_0^{2\pi} f'(t) \sin(kt) dt$ konvergiert. Das Rad B zeigt den Wert $\int_0^{2\pi} f'(t) \cos(kt) dt$ nach Abfahren des Periodenintervalls. Die Fourier-Koeffizienten sind damit über die eingangs beschriebene partielle Integration bestimmt.

6.4.1 Ausgangspunkt: Koeffizienten einer Fourier-Reihe

In der digitalen Realität (wie z. B. bei der MP3-Codierung oder in Messgeräten) haben wir für die Berechnung einer Fourier-Reihe nur endlich viele abgetastete Funktionswerte einer 2π -periodischen Funktion f zur Verfügung. Den Vorgang des Ab tastens von endlich vielen Funktionswerten nennt man **Sampling**, man spricht bei einem Übergang zu endlich (oder höchstens abzählbar) vielen Werten von **Diskretisierung**.

Auf Basis von n Abtastungen von f im Intervall $[0, 2\pi]$ zu den Zeitpunkten $t_l = l \cdot \frac{2\pi}{n}$, $l = 0, 1, \dots, n-1$, müssen die Integrale zur Berechnung der Fourier-Koeffizienten

numerisch approximiert werden. Dazu benutzen wir eine einfache Quadraturformel (vgl. Seite 360), für die wir $[0, 2\pi]$ in n Teilintervalle der Länge $\frac{2\pi}{n}$ unterteilen und pro Teilintervall den Funktionswert am linken Intervallrand berechnen, d. h. beim l -ten Intervall die Funktion f an der Stelle $t_l = l \cdot \frac{2\pi}{n}$ auswerten:

$$\begin{aligned} f^\wedge(k) &= \frac{1}{2\pi} \int_0^{2\pi} f(t) \exp(-jkt) dt \approx \frac{1}{2\pi} \sum_{l=0}^{n-1} \frac{2\pi}{n} f\left(l \frac{2\pi}{n}\right) \exp\left(-jkl \frac{2\pi}{n}\right) \\ &= \frac{1}{n} \sum_{l=0}^{n-1} f\left(l \frac{2\pi}{n}\right) \left[\exp\left(-j \frac{2\pi}{n}\right) \right]^{kl} = \frac{1}{n} \sum_{l=0}^{n-1} f_l v^{kl} =: d_k, \end{aligned} \quad (6.51)$$

wobei $f_l := f\left(l \frac{2\pi}{n}\right)$ und $v := \exp\left(-j \frac{2\pi}{n}\right)$. Mit n Funktionswerten f_l erhalten wir näherungsweise Fourier-Koeffizienten d_k (und unter gewissen Voraussetzungen sogar exakte – siehe Abschnitt 6.4.6). Die d_k nennt man **Fourier-Lagrange-Koeffizienten** oder **gestörte Fourier-Koeffizienten**. In Matrix-Schreibweise wird aus (6.51)

$$\begin{pmatrix} d_0 \\ d_1 \\ \dots \\ d_{n-1} \end{pmatrix} = \frac{1}{n} \underbrace{\begin{bmatrix} v^{0 \cdot 0} & v^{0 \cdot 1} & \dots & v^{0 \cdot (n-1)} \\ v^{1 \cdot 0} & v^{1 \cdot 1} & \dots & v^{1 \cdot (n-1)} \\ \vdots & \vdots & & \vdots \\ v^{(n-1) \cdot 0} & v^{(n-1) \cdot 1} & \dots & v^{(n-1) \cdot (n-1)} \end{bmatrix}}_{\mathbf{F}} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix}.$$

Man nennt das Produkt dieser speziellen Matrix $\mathbf{F}$ (ohne den Faktor $\frac{1}{n}$) mit dem Vektor der Funktionswerte eine **diskrete Fourier-Transformation**.

Beispiel 6.33

Für $n = 4$ ist $v = \exp\left(-j \frac{2\pi}{4}\right) = \cos\left(\frac{\pi}{2}\right) - j \sin\left(\frac{\pi}{2}\right) = -j$ mit $v^0 = 1$, $v^1 = -j$, $v^2 = -1$ und $v^3 = j$. Danach wiederholen sich die Werte periodisch, so dass

$$\begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \end{pmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & +j \\ 1 & -1 & 1 & -1 \\ 1 & +j & -1 & -j \end{bmatrix} \begin{pmatrix} f(0) \\ f\left(\frac{2\pi}{4}\right) \\ f\left(2 \cdot \frac{2\pi}{4}\right) \\ f\left(3 \cdot \frac{2\pi}{4}\right) \end{pmatrix}. \quad (6.52) \quad \blacksquare$$

Die Zahl v ist eine n -te komplexe Wurzel der 1: $v^n = \exp(-j2\pi) = \cos(2\pi) - j \sin(2\pi) = 1$. Man bezeichnet v als eine n -te **Einheitswurzel**. Insbesondere ist $v^{nl} = 1^l = 1$, und wir erhalten eine Periode n für die näherungsweisen Fourier-Koeffizienten d_k im folgenden Sinne:

$$d_k = \frac{1}{n} \sum_{l=0}^{n-1} f_l v^{kl} = \frac{1}{n} \sum_{l=0}^{n-1} f_l v^{kl} \underbrace{v^{nl}}_{=1} = \frac{1}{n} \sum_{l=0}^{n-1} f_l v^{n l + k l} = \frac{1}{n} \sum_{l=0}^{n-1} f_l v^{(n+k)l} = d_{n+k}. \quad (6.53)$$

Bei n vorliegenden Funktionswerten kann man also nur maximal n verschiedene Fourier-Lagrange-Koeffizienten berechnen. Deshalb hat die Fourier-Matrix $\mathbf{F}$ auch nicht mehr als n Zeilen. Man benötigt für eine Fourier-Partialsomme $\sum_{k=-m}^m f^\wedge(k) e^{jkx}$ daher mindestens $2m+1$ Funktionswerte zur näherungsweisen Angabe der Fourier-Koeffizienten und der Partialsomme. Wollen wir die Partialsomme durch $\sum_{k=-m}^m d_k e^{jkx}$ annähern, so benötigen wir auch Fourier-Lagrange-Koeffizienten zu negativen Indizes k . Mit (6.53) gilt für einen geraden Wert für n :

$$\begin{aligned} \left(d_0, d_1, \dots, d_{\frac{n}{2}-1}, d_{-\frac{n}{2}}, \dots, d_{-2}, d_{-1} \right)^\top &= \left(d_0, d_1, \dots, d_{\frac{n}{2}-1}, d_{\frac{n}{2}}, \dots, d_{n-2}, d_{n-1} \right)^\top \\ &= \frac{1}{n} \mathbf{F} \cdot (f_0, f_1, \dots, f_{n-1})^\top. \end{aligned} \quad (6.54)$$

An dieser Stelle ist noch nicht verständlich, dass die n so berechneten Fourier-Lagrange-Koeffizienten tatsächlich auch die Fourier-Koeffizienten $f^\wedge(-\frac{n}{2}), \dots, f^\wedge(\frac{n}{2}-1)$ gut wiedergeben. Es könnte auch sein, dass die Werte $d_{\frac{n}{2}}, \dots, d_{n-1}$ nicht den Koeffizienten $f^\wedge(-\frac{n}{2}), \dots, f^\wedge(-1)$, sondern $f^\wedge(\frac{n}{2}), \dots, f^\wedge(n-1)$ entsprechen. Wir werden aber in Kapitel 6.4.6 sehen, dass wir hier die richtige Wahl getroffen haben.

Nun befinden sich die Fourier-Lagrange-Koeffizienten zu negativen Indizes in der zweiten Vektorhälfte. Um das zu ändern, vertauschen wir in einem ersten Schritt die ersten $\frac{n}{2}$ mit den zweiten $\frac{n}{2}$ Zeilen von $\mathbf{F}$.

$$\begin{aligned} \begin{pmatrix} d_{-\frac{n}{2}} \\ \vdots \\ d_{-1} \\ d_0 \\ d_1 \\ \vdots \\ d_{\frac{n}{2}-1} \end{pmatrix} &= \frac{1}{n} \begin{bmatrix} v^{\frac{n}{2} \cdot 0} & v^{\frac{n}{2} \cdot 1} & \dots & v^{\frac{n}{2} \cdot (n-1)} \\ \vdots & \vdots & & \vdots \\ v^{(n-1) \cdot 0} & v^{(n-1) \cdot 1} & \dots & v^{(n-1) \cdot (n-1)} \\ v^{0 \cdot 0} & v^{0 \cdot 1} & \dots & v^{0 \cdot (n-1)} \\ \vdots & \vdots & & \vdots \\ v^{\left(\frac{n}{2}-1\right) \cdot 0} & v^{\left(\frac{n}{2}-1\right) \cdot 1} & \dots & v^{\left(\frac{n}{2}-1\right) \cdot (n-1)} \end{bmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} \\ &= \frac{1}{n} \begin{bmatrix} v^{\frac{n}{2} \cdot \frac{n}{2}} & \dots & v^{\frac{n}{2} \cdot (n-1)} & v^{\frac{n}{2} \cdot 0} & \dots & v^{\frac{n}{2} \cdot \left(\frac{n}{2}-1\right)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v^{(n-1) \cdot \frac{n}{2}} & \dots & v^{(n-1) \cdot (n-1)} & v^{(n-1) \cdot 0} & \dots & v^{(n-1) \cdot \left(\frac{n}{2}-1\right)} \\ v^{0 \cdot \frac{n}{2}} & \dots & v^{0 \cdot (n-1)} & v^{0 \cdot 0} & \dots & v^{0 \cdot \left(\frac{n}{2}-1\right)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v^{\left(\frac{n}{2}-1\right) \cdot \frac{n}{2}} & \dots & v^{\left(\frac{n}{2}-1\right) \cdot (n-1)} & v^{\left(\frac{n}{2}-1\right) \cdot 0} & \dots & v^{\left(\frac{n}{2}-1\right) \cdot \left(\frac{n}{2}-1\right)} \end{bmatrix} \begin{pmatrix} f_{-\frac{n}{2}} \\ f_{-\frac{n}{2}+1} \\ \vdots \\ f_0 \\ f_1 \\ \vdots \\ f_{\frac{n}{2}-1} \end{pmatrix}. \end{aligned} \quad (6.55)$$

Bei der zweiten Gleichung haben wir die ersten $\frac{n}{2}$ Spalten mit den zweiten vertauscht. Dabei verwenden wir, dass f 2π -periodisch ist mit $f_{-\frac{n}{2}} = f\left(-\frac{n}{2} \cdot \frac{2\pi}{n}\right) = f(-\pi) =$

$f(\pi) = f_{\frac{n}{2}}, f_{-\frac{n}{2}+1} = f_{\frac{n}{2}+1}, \dots, f_{-1} = f_{n-1}$. So stehen in den Vektoren auf der linken und rechten Seite die gleichen Indizes.

Für $n = 4$ wird daraus

$$\begin{pmatrix} d_{-2} \\ d_{-1} \\ d_0 \\ d_1 \end{pmatrix} = \frac{1}{4} \begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & -j & 1 & +j \\ 1 & 1 & 1 & 1 \\ -1 & +j & 1 & -j \end{bmatrix} \begin{pmatrix} f(-\pi) \\ f(-\frac{\pi}{2}) \\ f(0) \\ f(\frac{\pi}{2}) \end{pmatrix}.$$

Tatsächlich koppelt man die Anzahl der zu berechnenden Koeffizienten an die Anzahl ausgewerteter Funktionswerte. Für gerades n kann man so mit n Funktionswerten symmetrisch $n - 1$ Koeffizienten $f^{\wedge}(-\frac{n}{2} + 1) \approx d_{-\frac{n}{2}+1} = d_{\frac{n}{2}+1}, \dots, f^{\wedge}(\frac{n}{2} - 1) \approx d_{\frac{n}{2}-1}$ approximieren, und für ungerades n erhält man analog n Koeffizienten $f^{\wedge}(-\frac{n-1}{2}) \approx d_{-\frac{n-1}{2}} = d_{\frac{n+1}{2}}, \dots, f^{\wedge}(\frac{n-1}{2}) \approx d_{\frac{n-1}{2}}$ (siehe Kapitel 6.4.5).

Zur Berechnung der Fourier-Lagrange-Koeffizienten d_k mit einem Computer verwendet man effiziente Implementierungen der diskreten Fourier-Transformation. Auf die diskrete Fourier-Transformation gehen wir in den folgenden Abschnitten ein und nehmen anschließend in Kapitel 6.4.5 wieder Bezug zur numerischen Berechnung einer Fourier-Reihe, indem wir die vorangehenden Überlegungen mittels der diskreten Fourier-Transformation schreiben. Anschließend sehen wir uns auch an, wie die Fourier-Transformation (und damit wegen (6.40) auch die Laplace-Transformation) numerisch berechnet werden kann.

6.4.2 Diskrete Fourier-Transformation

In diesem Abschnitt diskutieren wir in einem allgemeineren Rahmen den Aufbau und die Eigenschaften der $(n \times n)$ -Matrizen $\mathbf{F}$, die wir im vorangehenden Kapitel kennengelernt haben. Möchte man die Fourier-Koeffizienten von sehr hochfrequenten Funktionen bestimmen, so benötigt man sehr viele abgetastete Funktionswerte. Die Matrix $\mathbf{F}$ wird sehr groß. Selbst mit schnellen Computern ist die Multiplikation einer großen Matrix mit einem Vektor eine aufwändige Operation, bei der eine Größenordnung von n^2 Multiplikationen zu bewerkstelligen ist. (Bei n Elementen des Ergebnisvektors wird jedes mittels n Multiplikationen berechnet.) Das geht deutlich besser. Um zu verstehen, wie man die Berechnung optimieren kann, verwenden wir in diesem Abschnitt neben $v = \exp(-j\frac{2\pi}{n})$ weitere Zahlen mit ähnlichen Eigenschaften, die zu allgemeineren Matrizen $\mathbf{F}$ führen. Das dabei gewonnene Wissen führt dann anschließend zum FFT-Algorithmus.

Das Polynom $z^n - 1$ hat nach dem Fundamentalsatz der Algebra (siehe Seite 147) n komplexe Nullstellen unter Berücksichtigung ihrer Vielfachheit. Hier gibt es tatsächlich

n verschiedene einfache Nullstellen. Dies sind die n -ten Wurzeln der Eins (siehe Kapitel 1.6), die n -ten **Einheitswurzeln**

$$\exp\left(jl\frac{2\pi}{n}\right) = \exp\left(j\frac{2\pi}{n}\right)^l = \cos\left(l\frac{2\pi}{n}\right) + j\sin\left(l\frac{2\pi}{n}\right), \quad 0 \leq l \leq n-1. \quad (6.56)$$

Die zuvor für die numerische Berechnung von Fourier-Koeffizienten verwendete Zahl

$$v := \exp\left(-j\frac{2\pi}{n}\right) = \exp\left(j\left[2\pi - \frac{2\pi}{n}\right]\right) = \exp\left(j(n-1)\frac{2\pi}{n}\right)$$

ist eine dieser Wurzeln.

Neben den über Potenzen von v gebildeten Matrizen $\mathbf{F}$ werden wir entsprechende Matrizen auch für andere n -te Einheitswurzeln w aufstellen. Es wird sich zeigen, dass man diese sogar benötigt, wenn man ursprünglich nur mit v arbeiten möchte. Damit die Matrizen invertierbar werden, verwendet man dabei aber nur solche w , die eine zusätzliche Bedingung erfüllen:

Definition 6.6 (Primitive Einheitswurzel)

Eine n -te Einheitswurzel $w \in \mathbb{C}$, d. h., w ist eine Zahl aus (6.56) mit $w^n = 1$, heißt **primitiv** genau dann, wenn

$$w^k \neq 1, \quad 1 \leq k \leq n-1. \quad (6.57)$$

Durch Bildung der Potenzen w^k , $k \in \mathbb{N}$, einer primitiven n -ten Einheitswurzel erhält man n verschiedene komplexe Zahlen. So ist z. B. $j = \exp\left(j\frac{2\pi}{4}\right) = e^{j\frac{\pi}{2}}$ eine primitive vierte Einheitswurzel: $j^1 = j$, $j^2 = -1$, $j^3 = -j$, $j^4 = 1$, $j^5 = j$, $j^6 = -1$, ... Bei der Bildung der Potenzen wiederholen sich die Zahlen, wenn man ein Vielfaches von $n = 4$ und damit den Wert 1 erreicht.

Lemma 6.8 (Potenzen einer primitiven n -ten Einheitswurzel)

Sei w eine primitive n -te Einheitswurzel. Dann sind die Potenzen $w^1, w^2, \dots, w^n$ alle voneinander verschieden.

Beweis: Wäre $w^k = w^m$ für zwei Zahlen k und m mit $0 < k < m \leq n$, so wäre $w^{m-k} = \frac{w^m}{w^k} = 1$. Allerdings ist $1 \leq m-k \leq n-1$. Das ist ein Widerspruch dazu, dass w primitiv ist. $\square$

Wegen $(w^k)^n = (w^n)^k = 1^k = 1$ (bzw. mit der Darstellung $w = \exp\left(jl\frac{2\pi}{n}\right)$ ist $[\exp\left(jl\frac{2\pi}{n}\right)]^n = \exp(jl2\pi) = 1$) sind alle Potenzen einer n -ten Einheitswurzel w ebenfalls n -te Einheitswurzeln. Ist zudem w primitiv, so erhält man durch Bildung der Potenzen nach der vorangehenden Überlegung n verschiedene und damit alle n -ten Einheitswurzeln. Mit einer primitiven Einheitswurzel erhält man also alle Einheitswurzeln durch Potenzieren.

Lemma 6.9 (Charakterisierung primitiver Wurzeln)

Primitive n -te Einheitswurzeln sind genau die Zahlen $\exp\left(jl\frac{2\pi}{n}\right)$, $1 \leq l < n$, bei denen l und n teilerfremd sind.

Insbesondere ist $v = \exp\left(j(n-1)\frac{2\pi}{n}\right)$ eine primitive n -te Wurzel, da $n-1$ und n teilerfremd sind.

Beweis: Die angegebenen Zahlen sind n -te Wurzeln, so dass wir zeigen müssen, dass sie a) primitiv sind und es b) keine weiteren primitiven Wurzeln gibt.

a) Wir untersuchen, ob für teilerfremde Zahlen l und n die Gleichung

$$\left[\exp\left(jl\frac{2\pi}{n}\right)\right]^k = \exp\left(jlk\frac{2\pi}{n}\right) = 1$$

eine Lösung $1 \leq k < n$ besitzen kann (dann wäre diese n -te Wurzel nicht primitiv). Die Gleichung ist genau dann erfüllt, wenn lk ein Vielfaches von n ist, also $lk = mn$, d. h. $k = \frac{mn}{l}$. Da n und l teilerfremd sind, ist dann m durch l teilbar, $\frac{m}{l}$ ist eine natürliche Zahl und damit k ein Vielfaches von n . Es gibt daher keine Lösung $1 \leq k < n$. Die angegebenen n -ten Wurzeln sind primitiv.

b) Es gibt keine weiteren primitiven n -ten Wurzeln. Denn falls l ($1 \leq l < n$) und n nicht teilerfremd sind, ist $l = mr$ und $n = ms$ für Zahlen $m, r, s \in \mathbb{N}$ mit $1 \leq r, s < n$. Damit ist aber

$$\left[\exp\left(jl\frac{2\pi}{n}\right)\right]^s = \exp\left(j\underbrace{mr}_=l s\frac{2\pi}{n}\right) = \exp\left(j\underbrace{ms}_=n r\frac{2\pi}{n}\right) = 1^r = 1.$$

Die Zahl $\exp\left(jl\frac{2\pi}{n}\right)$ ist also keine primitive n -te Wurzel. □

Lemma 6.10 (Eigenschaften primitiver n -ter Einheitswurzeln)

Für eine primitive n -te Einheitswurzel w gilt:

- a) w^{-1} ist ebenfalls eine n -te primitive Einheitswurzel, erfüllt also (6.57). Außerdem ist $w^{-1} = \bar{w}$.
 b) Ist n gerade, so ist die Zahl w^2 eine $\frac{n}{2}$ -te primitive Einheitswurzel von 1. Außerdem ist

$$w^{\frac{n}{2}} = -1. \quad (6.58)$$

Beweis: a) Zunächst ist der Kehrwert jeder n -ten Einheitswurzel wieder eine n -te Einheitswurzel, da

$$(w^{-1})^n = \frac{1}{w^n} = \frac{1}{1} = 1. \quad (6.59)$$

Mit w ist auch w^{-1} primitiv. Denn angenommen, es gibt ein $1 \leq k \leq n-1$ mit $1 = (w^{-1})^k$, dann ist $1 = w^0 = w^{k-k} = w^k (w^{-1})^k = w^k$. Dies ist ein Widerspruch dazu, dass w primitiv ist, siehe (6.57).

Wegen $1 = |w^n| = |w|^n$ ist $|w| = 1$, so dass außerdem $w^{-1} = \frac{1}{w} = \frac{\bar{w}}{w\bar{w}} = \frac{\bar{w}}{|w|^2} = \bar{w}$.

b) Die Zahl w^2 ist eine $\frac{n}{2}$ -te Wurzel von 1, da

$$(w^2)^{\frac{n}{2}} = w^n = 1. \quad (6.60)$$

w^2 ist primitiv, d. h. $(w^2)^k \neq 1$, $1 \leq k < \frac{n}{2}$. Das ergibt sich für die Exponenten $2k$ direkt aus der Voraussetzung, dass w primitiv ist, vgl. (6.57).

Da $(w^{\frac{n}{2}})^2 = 1$ gilt, ist entweder $w^{\frac{n}{2}} = 1$ oder $w^{\frac{n}{2}} = -1$. 1 kommt aber nicht in Frage, da w primitiv ist. Damit gilt (6.58). $\square$

Wir abstrahieren nun von der numerischen Berechnung von Fourier-Koeffizienten. Das Ergebnis ist die diskrete Fourier-Transformation, die unabhängig vom Bezug zu den Fourier-Koeffizienten eine interessante lineare Abbildung (vgl. Kapitel 3.4) ist:

Definition 6.7 (Diskrete Fourier-Transformation)

Sei w eine primitive n -te Einheitswurzel. Die **diskrete Fourier-Transformation** $\text{DFT}_w : \mathbb{C}^n \rightarrow \mathbb{C}^n$ bildet (Zeilen- oder Spalten-)Vektoren $\vec{y} = (y_0, y_1, \dots, y_{n-1})$ ab auf (Zeilen- oder Spalten-)Vektoren $\vec{c} = (c_0, c_1, \dots, c_{n-1})$ mit

$$\text{DFT}_w(\vec{y}) = \vec{c}, \text{ wobei } c_k = \sum_{l=0}^{n-1} y_l w^{kl}. \quad (6.61)$$

Üblicherweise beginnt man hier den Index bei 0. Über eine (symmetrische) **Fourier-Matrix** ausgedrückt lautet die (lineare) Abbildung:

$$\begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} = \begin{bmatrix} w^{0 \cdot 0} & w^{0 \cdot 1} & \dots & w^{0 \cdot (n-1)} \\ w^{1 \cdot 0} & w^{1 \cdot 1} & \dots & w^{1 \cdot (n-1)} \\ \vdots & \vdots & & \vdots \\ w^{(n-1) \cdot 0} & w^{(n-1) \cdot 1} & \dots & w^{(n-1) \cdot (n-1)} \end{bmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{pmatrix}.$$

Beispiel 6.34

a) In (6.52) haben wir die Fourier-Matrix zur primitiven vierten Wurzel $w = v = j$ verwendet.

b) Die Fourier-Matrix für die diskrete Fourier-Transformation in $\mathbb{C}^6$ zur primitiven Wurzel $w := e^{j\frac{\pi}{3}}$ lautet

$$\begin{aligned}
 \begin{bmatrix} w^{0\cdot0} & w^{0\cdot1} & w^{0\cdot2} & w^{0\cdot3} & w^{0\cdot4} & w^{0\cdot5} \\ w^{1\cdot0} & w^{1\cdot1} & w^{1\cdot2} & w^{1\cdot3} & w^{1\cdot4} & w^{1\cdot5} \\ w^{2\cdot0} & w^{2\cdot1} & w^{2\cdot2} & w^{2\cdot3} & w^{2\cdot4} & w^{2\cdot5} \\ w^{3\cdot0} & w^{3\cdot1} & w^{3\cdot2} & w^{3\cdot3} & w^{3\cdot4} & w^{3\cdot5} \\ w^{4\cdot0} & w^{4\cdot1} & w^{4\cdot2} & w^{4\cdot3} & w^{4\cdot4} & w^{4\cdot5} \\ w^{5\cdot0} & w^{5\cdot1} & w^{5\cdot2} & w^{5\cdot3} & w^{5\cdot4} & w^{5\cdot5} \end{bmatrix} &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 & w^4 & w^5 \\ 1 & w^2 & w^4 & 1 & w^2 & w^4 \\ 1 & w^3 & 1 & w^3 & 1 & w^3 \\ 1 & w^4 & w^2 & 1 & w^4 & w^2 \\ 1 & w^5 & w^4 & w^3 & w^2 & w \end{bmatrix} \\
 &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & e^{j\pi/3} & e^{j2\pi/3} & -1 & e^{j4\pi/3} & e^{j5\pi/3} \\ 1 & e^{j2\pi/3} & e^{j4\pi/3} & 1 & e^{j2\pi/3} & e^{j4\pi/3} \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & e^{j4\pi/3} & e^{j2\pi/3} & 1 & e^{j4\pi/3} & e^{j2\pi/3} \\ 1 & e^{j5\pi/3} & e^{j4\pi/3} & -1 & e^{j2\pi/3} & e^{j\pi/3} \end{bmatrix}. \quad (6.62)
 \end{aligned}$$

Die diskrete Fourier-Transformation hat eine Umkehrtransformation (**inverse diskrete Fourier-Transformation, IDFT**), die wir erhalten, indem wir die Fourier-Matrix invertieren (vgl. Kapitel 3.4.4). Wir werden sehen, dass die Matrix gerade deshalb invertierbar ist, weil wir primitive Einheitswurzeln verwenden. Die Berechnung der Inversen der Matrix gestaltet sich hier besonders einfach. Man muss nicht den Gauß-Algorithmus bemühen oder die Cramer'sche Regel anwenden. Vielmehr genügt es, die Adjungierte, die sich hier aus den Kehrwerten der Elemente von $\mathbf{F}$ zusammensetzt, mit dem Faktor $\frac{1}{n}$ zu multiplizieren:

Lemma 6.11 (Inverse der Fourier-Matrix)

Die Fourier-Matrix $\mathbf{F}$ zu einer n -ten primitiven Wurzel w der Eins hat die inverse Matrix $\mathbf{F}^{-1} = \frac{1}{n}\mathbf{F}^*$, wobei $\mathbf{F}^*$ die Adjungierte von $\mathbf{F}$ ist (siehe Seite 173), d. h.

$$\mathbf{F}^{-1} = \frac{1}{n} \begin{bmatrix} w^{-0\cdot0} & w^{-0\cdot1} & \dots & w^{-0\cdot(n-1)} \\ w^{-1\cdot0} & w^{-1\cdot1} & \dots & w^{-1\cdot(n-1)} \\ \vdots & \vdots & & \vdots \\ w^{-(n-1)\cdot0} & w^{-(n-1)\cdot1} & \dots & w^{-(n-1)\cdot(n-1)} \end{bmatrix}.$$

Beweis: Wir prüfen, dass $\mathbf{F}^{-1}$ tatsächlich die inverse Matrix der Fourier-Matrix $\mathbf{F}$ ist, indem wir $\mathbf{F}^{-1}\mathbf{F}$ berechnen. Das Ergebnis muss die Einheitsmatrix sein, da wir

$\vec{y} = \mathbf{F}^{-1}\mathbf{F}\vec{y}$ für jeden Vektor $\vec{y}$ erwarten. Dazu verwenden wir die Formel für eine endliche geometrische Summe (siehe (1.10) auf Seite 60):

$$\sum_{r=0}^{n-1} x^r = \frac{1-x^n}{1-x}.$$

Dabei muss $x \neq 1$ sein. Es ist

$$\begin{aligned} [\mathbf{F}^{-1}\mathbf{F}]_{k,l} &= \sum_{r=0}^{n-1} \frac{1}{n} w^{-kr} w^{rl} = \frac{1}{n} \sum_{r=0}^{n-1} [w^{l-k}]^r \\ &= \begin{cases} \frac{1}{n} \sum_{r=0}^{n-1} 1 = 1, & l = k, \\ \frac{1}{n} \frac{1-[w^{l-k}]^n}{1-w^{l-k}} = \frac{1}{n} \frac{1-1}{1-w^{l-k}} = 0, & l \neq k. \end{cases} \end{aligned}$$

Dabei ist $x = w^{l-k}$ im Nenner nur deshalb von 1 verschieden, da w primitiv ist. An genau dieser Stelle benötigt man die Einschränkung auf primitive Einheitswurzeln. Wir erhalten die Einheitsmatrix und haben die explizite Darstellung von $\mathbf{F}^{-1}$ verifiziert. Da $\mathbf{F}$ symmetrisch ist und $w^{-1} = \bar{w}$ die zu w konjugiert komplexe Zahl ist, entspricht die Inverse der Adjungierten bis auf den Faktor $\frac{1}{n}$. $\square$

Die Fourier-Umkehrtransformation wird bezeichnet mit $\text{IDFT}_w(\vec{c}) := \mathbf{F}^{-1}\vec{c}$, bezogen auf eine Komponente y_l des Vektors $\vec{y} := \text{IDFT}_w(\vec{c})$ ist also

$$y_l = \frac{1}{n} \sum_{k=0}^{n-1} c_k w^{-kl}.$$

Insbesondere sehen wir im Vergleich mit (6.61), dass wir die IDFT über die DFT berechnen können:

Folgerung 6.3 (Berechnung der Umkehrtransformation)

Seien w eine n -te primitive Einheitswurzel, DFT_w und IDFT_w die zugehörige diskrete Fourier-Transformation und Fourier-Umkehrtransformation. Dann gilt:

$$\text{IDFT}_w(\vec{c}) = \frac{1}{n} \text{DFT}_{w^{-1}}(\vec{c}), \quad (6.63)$$

Man beachte dabei, dass mit w auch w^{-1} eine primitive n -te Einheitswurzel ist (siehe (6.59) und die zugehörigen Bemerkungen).

Beispiel 6.35

Für $n = 4$ und $w = v$ ist zu $\mathbf{F}$ aus (6.52) die inverse Matrix

$$\mathbf{F}^{-1} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \end{bmatrix}.$$

■

Beispiel 6.36

Wir berechnen $\mathbf{F}^{-1}$ in $\mathbb{C}^6$ für die primitive Wurzel $w := e^{j\frac{\pi}{3}}$ (siehe (6.62)). Da $e^{-jk\pi/3} = e^{j[2\pi-k\pi/3]} = e^{j[6-k]\pi/3}$ ist

$$\begin{aligned} \mathbf{F}^{-1} &= \frac{1}{6} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & e^{-j\pi/3} & e^{-j2\pi/3} & -1 & e^{-j4\pi/3} & e^{-j5\pi/3} \\ 1 & e^{-j2\pi/3} & e^{-j4\pi/3} & 1 & e^{-j2\pi/3} & e^{-j4\pi/3} \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & e^{-j4\pi/3} & e^{-j2\pi/3} & 1 & e^{-j4\pi/3} & e^{-j2\pi/3} \\ 1 & e^{-j5\pi/3} & e^{-j4\pi/3} & -1 & e^{-j2\pi/3} & e^{-j\pi/3} \end{bmatrix} \\ &= \frac{1}{6} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & e^{j5\pi/3} & e^{j4\pi/3} & -1 & e^{j2\pi/3} & e^{j\pi/3} \\ 1 & e^{j4\pi/3} & e^{j2\pi/3} & 1 & e^{j4\pi/3} & e^{j2\pi/3} \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & e^{j2\pi/3} & e^{j4\pi/3} & 1 & e^{j2\pi/3} & e^{j4\pi/3} \\ 1 & e^{j\pi/3} & e^{j2\pi/3} & -1 & e^{j4\pi/3} & e^{j5\pi/3} \end{bmatrix}. \end{aligned}$$

■

Damit ist die DFT eine bijektive Abbildung von $\mathbb{C}^n$ auf $\mathbb{C}^n$, d. h., jeder Vektor hat ein eindeutiges Bild, und jedes $\vec{c} \in \mathbb{C}^n$ wird auch als Wert der Abbildung angenommen.

Hintergrund: Diskrete Fourier-Transformation in reeller Notation

Wir können die diskrete Fourier-Transformation analog zu den Fourier-Koeffizienten auch rein reell schreiben. Das ist sinnvoll, wenn Vektoren aus $\mathbb{R}^n$ transformiert werden sollen. Die diskrete Fourier-Transformation bildet diese auf Vektoren eines n -dimensionalen reellen Untervektorraums des $\mathbb{C}^n$ ab. Damit lässt sich die Transformierte mit n reellen Zahlen statt mit $2n$ reellen Werten für die n Real- und die n Imaginärteile ausdrücken. Sei $w = \exp\left(j\frac{r}{n}2\pi\right)$ eine primitive n -te Einheitswurzel, d. h., r und n sind

teilerfremd. Für $k, l \in \mathbb{N}_0$ ist $w^{kl} = \cos\left(kl \frac{r}{n} 2\pi\right) + j \sin\left(kl \frac{r}{n} 2\pi\right)$, so dass die Fourier-Matrix $\mathbf{F} = [w^{k \cdot l}]_{0 \leq k, l < n}$ als Summe $\mathbf{F}_1 + j\mathbf{F}_2$ mit den beiden reellen Matrizen

$$\mathbf{F}_1 = \begin{bmatrix} \cos\left(0 \cdot 0 \frac{r}{n} 2\pi\right) & \cos\left(0 \cdot 1 \frac{r}{n} 2\pi\right) & \dots & \cos\left(0 \cdot (n-1) \frac{r}{n} 2\pi\right) \\ \cos\left(1 \cdot 0 \frac{r}{n} 2\pi\right) & \cos\left(1 \cdot 1 \frac{r}{n} 2\pi\right) & \dots & \cos\left(1 \cdot (n-1) \frac{r}{n} 2\pi\right) \\ \vdots & \vdots & & \vdots \\ \cos\left((n-1) \cdot 0 \frac{r}{n} 2\pi\right) & \cos\left((n-1) \cdot 1 \frac{r}{n} 2\pi\right) & \dots & \cos\left((n-1) \cdot (n-1) \frac{r}{n} 2\pi\right) \end{bmatrix},$$

$$\mathbf{F}_2 = \begin{bmatrix} \sin\left(0 \cdot 0 \frac{r}{n} 2\pi\right) & \sin\left(0 \cdot 1 \frac{r}{n} 2\pi\right) & \dots & \sin\left(0 \cdot (n-1) \frac{r}{n} 2\pi\right) \\ \sin\left(1 \cdot 0 \frac{r}{n} 2\pi\right) & \sin\left(1 \cdot 1 \frac{r}{n} 2\pi\right) & \dots & \sin\left(1 \cdot (n-1) \frac{r}{n} 2\pi\right) \\ \vdots & \vdots & & \vdots \\ \sin\left((n-1) \cdot 0 \frac{r}{n} 2\pi\right) & \sin\left((n-1) \cdot 1 \frac{r}{n} 2\pi\right) & \dots & \sin\left((n-1) \cdot (n-1) \frac{r}{n} 2\pi\right) \end{bmatrix}$$

geschrieben werden kann. Wir erkennen, dass die $(n-k)$ -te Zeile von $\mathbf{F}_1$ mit der k -ten übereinstimmt für $1 \leq k \leq \frac{n}{2}$, da der Kosinus eine gerade Funktion ist und $n \cdot \frac{r}{n} 2\pi$ wegen der 2π -Periodizität im Argument weggelassen werden kann. Bei $\mathbf{F}_2$ haben die k -te und $(n-k)$ -te Zeile genau entgegengesetzte Vorzeichen für $1 \leq k \leq \frac{n}{2}$. Ist n gerade, so ist die Zeile zu $\frac{n}{2}$ in $\mathbf{F}_2$ daher gleich $\vec{0}$. Außerdem ist die nullte Zeile von $\mathbf{F}_2$ der Nullvektor.

Wir transformieren nun $\vec{x} \in \mathbb{R}^n$ in das Paar $(\vec{a}, \vec{b})$ reeller Vektoren mit $\vec{a} = \mathbf{F}_1 \cdot \vec{x}$ und $\vec{b} = \mathbf{F}_2 \cdot \vec{x}$, wobei $\mathbf{F}\vec{x} = \vec{a} + j\vec{b}$. Aufgrund der soeben beschriebenen redundanten Struktur der Matrizen $\mathbf{F}_1$ und $\mathbf{F}_2$ müssen aber nur die Komponenten $a_0, a_1, \dots, a_{m_1}$ und $b_1, b_2, \dots, b_{m_2}$ berechnet werden, wobei $m_1 = \frac{n}{2}$ und $m_2 = \frac{n}{2} - 1$ für gerades n und $m_1 = m_2 = \frac{n-1}{2}$ für ungerades n gilt. Lösen wir die Matrixmultiplikation auf, dann entstehen diese relevanten Komponenten von $\vec{a}$ und $\vec{b}$ so:

$$a_k = \sum_{l=0}^{n-1} x_l \cos\left(kl \frac{r}{n} 2\pi\right) \text{ für } 0 \leq k \leq m_1, \quad b_k = \sum_{l=1}^{n-1} x_l \sin\left(kl \frac{r}{n} 2\pi\right) \text{ für } 1 \leq k \leq m_2.$$

Für die verbleibenden Komponenten von $\vec{a}$ und $\vec{b}$ gilt: $b_0 = 0$, $a_k = a_{n-k}$ und $b_k = -b_{n-k}$ für $m_1 < k < n$ bzw. $m_2 < k < n$.

Aus $\vec{a} + j\vec{b}$ erhalten wir $\vec{x}$ über die inverse diskrete Fourier-Transformation zurück: $\vec{x} = \mathbf{F}^{-1}[\vec{a} + j\vec{b}]$. Dabei entsteht $\mathbf{F}^{-1}$ aus $\mathbf{F}$, indem alle Komponenten durch ihre Kehrwerte ersetzt werden und die so entstehende Matrix mit $\frac{1}{n}$ multipliziert wird. Da $\vec{x}$ hier reellwertig ist, müssen wir bei der Umkehrtransformation nur die Realteile ausrechnen ($0 \leq k < n$):

$$\begin{aligned} x_k &= \frac{1}{n} \sum_{l=0}^{n-1} \left[\cos\left(-kl \frac{r}{n} 2\pi\right) + j \sin\left(-kl \frac{r}{n} 2\pi\right) \right] (a_l + jb_l) \\ &= \frac{1}{n} \sum_{l=0}^{n-1} \left[\cos\left(kl \frac{r}{n} 2\pi\right) - j \sin\left(kl \frac{r}{n} 2\pi\right) \right] (a_l + jb_l) \\ &= \frac{1}{n} \sum_{l=0}^{n-1} \left[a_l \cos\left(kl \frac{r}{n} 2\pi\right) + b_l \sin\left(kl \frac{r}{n} 2\pi\right) \right] = \frac{1}{n} \sum_{l=0}^{n-1} (a_l, b_l) \begin{pmatrix} \cos\left(kl \frac{r}{n} 2\pi\right) \\ \sin\left(kl \frac{r}{n} 2\pi\right) \end{pmatrix}. \end{aligned}$$

Insbesondere sehen wir, dass $\mathbf{F}^{-1} = \frac{1}{n}[\mathbf{F}_1 - j\mathbf{F}_2]$ ist. Wenn wir die redundanten Koeffizienten a_k und b_k in dieser Formel ersetzen, dann erhalten wir bei ungeradem n :

$$\begin{aligned}
 x_k &= \frac{a_0}{n} + \frac{1}{n} \sum_{l=1}^{n-1} \left[a_l \cos\left(kl \frac{r}{n} 2\pi\right) + b_l \sin\left(kl \frac{r}{n} 2\pi\right) \right] \\
 &= \frac{a_0}{n} + \frac{1}{n} \sum_{l=1}^{\frac{n-1}{2}} \left[a_l \cos\left(kl \frac{r}{n} 2\pi\right) + b_l \sin\left(kl \frac{r}{n} 2\pi\right) \right] + \\
 &\quad \frac{1}{n} \sum_{l=\frac{n-1}{2}+1}^{n-1} \left[a_{n-l} \cos\left(k[n-l] \frac{r}{n} 2\pi\right) + b_{n-l} \sin\left(k[n-l] \frac{r}{n} 2\pi\right) \right] \\
 &= \frac{a_0}{n} + \frac{1}{n} \sum_{l=1}^{\frac{n-1}{2}} \left[a_l \cos\left(kl \frac{r}{n} 2\pi\right) + b_l \sin\left(kl \frac{r}{n} 2\pi\right) \right] + \\
 &\quad \frac{1}{n} \sum_{i=1}^{\frac{n-1}{2}} \left[a_i \cos\left(ki \frac{r}{n} 2\pi\right) + b_i \sin\left(ki \frac{r}{n} 2\pi\right) \right] \quad (\text{Substitution } i = n - l) \\
 &= \frac{a_0}{n} + \frac{2}{n} \sum_{l=1}^{\frac{n-1}{2}} \left[a_l \cos\left(kl \frac{r}{n} 2\pi\right) + b_l \sin\left(kl \frac{r}{n} 2\pi\right) \right].
 \end{aligned}$$

Wir haben damit eine Darstellung, die analog zu der der Fourier-Partialsummen für reelle periodische Funktionen ist. Das gilt mit $b_{\frac{n}{2}} = 0$ auch, wenn n gerade ist:

$$x_k = \frac{a_0}{n} + \frac{2}{n} \sum_{l=1}^{\frac{n}{2}} \left[a_l \cos\left(kl \frac{r}{n} 2\pi\right) + b_l \sin\left(kl \frac{r}{n} 2\pi\right) \right].$$

Wir betrachten im Folgenden Vektoren $\vec{y} \in \mathbb{C}^n$ als n -periodisch fortgesetzt, d. h. $y_{k \pm n} := y_k$. So können wir auch mit einem Index außerhalb $\{0, 1, \dots, n-1\}$ zugreifen. Für $\vec{y} = (1, 2, 3)$ ist beispielsweise $y_0 = y_3 = y_{-3} = y_6 = \dots = 1$, $y_1 = y_4 = y_{-2} = 2$.

Es gelten analog zu den Eigenschaften der Fourier-Koeffizienten und -Transformation (siehe Seiten 831 und 870) die folgenden Rechenregeln ($\vec{y}, \vec{z} \in \mathbb{C}^n$, $a, b \in \mathbb{C}$):

a) Linearität:

$$\text{DFT}_w(a \cdot \vec{y} + b \cdot \vec{z}) = a \cdot \text{DFT}_w(\vec{y}) + b \cdot \text{DFT}_w(\vec{z}).$$

b) Verschiebung (Translation) im Frequenzbereich:

Aus $\text{DFT}_w(\vec{y}) = \vec{c}$ folgt für $m \in \mathbb{Z}$

$$\text{DFT}_w((w^{lm} y_l)_{l=0}^{n-1}) = (c_{k+m})_{k=0}^{n-1}. \quad (6.64)$$

Beispiel 6.37

Mit $n = 4$, $w = v = -j$, $\mathbf{F}$ aus (6.52) und $m = 1$ erhalten wir

$$\text{DFT}_v((1, 2, 3, 4)) = (10, -2 + 2j, -2, -2 - 2j),$$

$$\begin{aligned} \text{DFT}_v((v^0 \cdot 1, v^1 \cdot 2, v^2 \cdot 3, v^3 \cdot 4)) &= \text{DFT}_v((1, -2j, -3, 4j)) \\ &= (-2 + 2j, -2, -2 - 2j, 10). \quad \blacksquare \end{aligned}$$

c) Verschiebung im Zeitbereich:

Aus $\text{DFT}_w(\vec{y}) = \vec{c}$ folgt für $m \in \mathbb{Z}$

$$\text{DFT}_w((y_{l+m})_{l=0}^{n-1}) = \left(w^{-km} c_k \right)_{k=0}^{n-1}. \quad (6.65)$$

In vielen Anwendungen ist bereits der Betrag der Komponenten c_k aussagekräftig. Da $|w^{-km}| = 1$ ist, sind die Beträge invariant gegen Verschiebungen im Zeitbereich. Das lässt sich z. B. zur Erkennung von Mustern ausnutzen.

Beispiel 6.38

Für $n = 4$, $w = v$, also $w^{-1} = j$, $\mathbf{F}$ aus (6.52) und $m = 1$ gilt einerseits

$$\text{DFT}_v((2, 3, 4, 1)) = (10, -2 - 2j, 2, -2 + 2j),$$

andererseits ist

$$\begin{aligned} (v^0 \cdot 10, v^{-1} \cdot (-2 + 2j), v^{-2} \cdot (-2), v^{-3} \cdot (-2 - 2j)) \\ = (1 \cdot 10, j \cdot (-2 + 2j), (-1) \cdot (-2), (-j) \cdot (-2 - 2j)) = (10, -2 - 2j, 2, -2 + 2j). \end{aligned}$$

■

Beweis: Aus der Linearität der Matrixmultiplikation folgt a).

Die Regel b) resultiert daraus, dass sich die Einträge einer Zeile der Fourier-Matrix aus den der vorangehenden ergeben, indem man den Eintrag der Spalte l der vorangehenden Zeile mit w^l multipliziert. Wegen $w^n = 1$ ergibt sich außerdem die erste Zeile durch Multiplikation der letzten mit w^l pro Spalte l . Bei einer zyklischen Rotation aller Zeilen um m Stellen (Multiplikation aller Spalten l mit w^{lm}) nach oben und anschließender Multiplikation mit $\vec{y}$ ergibt sich ein um m Stellen zyklisch nach oben verschobener Ergebnisvektor $(c_{k+m})_{k=0}^{n-1}$:

$$\text{DFT}_w((w^{lm} y_l)_{l=0}^{n-1}) = \mathbf{F} \cdot \begin{bmatrix} w^{0 \cdot m} & 0 & \dots & 0 \\ 0 & w^{1 \cdot m} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & w^{(n-1) \cdot m} \end{bmatrix} \cdot \vec{y} = (c_{k+m})_{k=0}^{n-1}.$$

Bei der Regel c) können wir eine Verschiebung der Komponenten des Vektors $\vec{y}$ nach oben durch eine Verschiebung der Spalten der Matrix nach rechts darstellen:

$$\begin{aligned} \text{DFT}_w((y_{l+m})_{l=0}^{n-1}) &= \mathbf{F} \cdot (y_{l+m})_{l=0}^{n-1} = \begin{bmatrix} w^{-0 \cdot m} & 0 & \dots & 0 \\ 0 & w^{-1 \cdot m} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & w^{-(n-1) \cdot m} \end{bmatrix} \mathbf{F} \cdot \vec{y} \\ &= \begin{bmatrix} w^{-0 \cdot m} & 0 & \dots & 0 \\ 0 & w^{-1 \cdot m} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & w^{-(n-1) \cdot m} \end{bmatrix} \vec{c} = \left(w^{-km} c_k \right)_{k=0}^{n-1}. \quad \square \end{aligned}$$

Ist n gerade und $m = \frac{n}{2}$, so ergibt sich aus c) wegen $w^{\frac{n}{2}} = -1$ der folgende Spezialfall für die Verschiebung im Zeitbereich, der später für die Anwendung der diskreten Fourier-Transformation bei der Berechnung von Fourier-Koeffizienten verwendet werden kann:

$$\text{DFT}_w((y_{l+n/2})_{l=0}^{n-1}) = \left((-1)^k c_k \right)_{k=0}^{n-1}. \quad (6.66)$$

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 877). 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 $\text{DFT}_v((a_{r,0}, a_{r,1}, \dots, a_{r,n-1}))$. Mit diesen Ergebnissen wird dann eine weitere eindimensionale Fourier-Transformation zur Wurzel w durchgeführt:

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

Mit $m + n$ eindimensionalen diskreten Fourier-Transformationen ist die zweidimensionale diskrete Fourier-Transformation berechnet. Denn man kann zunächst für jedes $r \in \{0, 1, \dots, m-1\}$ die innere Transformation ausrechnen und sich die Ergebnisse merken. Für jedes $k \in \{0, 1, \dots, n-1\}$ wird nun aus den zuvor berechneten Transformatierten 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.

6.4.3 Diskrete Faltung*

Bislang haben wir bei Fourier-Reihen nur die Faltung im Zeitbereich betrachtet. Haben wir dagegen ein Produkt im Zeitbereich, so erhalten wir unter geeigneten Voraussetzungen wie bei der Fourier-Transformation eine Faltung im Frequenzbereich. Dazu betrachten wir die Fourier-Koeffizienten des Produkts zweier 2π -periodischer Funktionen f und g . Falls die zugehörigen Fourier-Reihen konvergieren und wir Summe und Integral vertauschen dürfen, erhalten wir für die Fourier-Koeffizienten des Produkts der beiden Funktionen:

$$\begin{aligned} (f \cdot g)^\wedge(k) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \cdot g(t) \cdot e^{-jkt} dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \cdot \left[\sum_{r=-\infty}^{\infty} g^\wedge(r) e^{jrt} \right] \cdot e^{-jkt} dt = \sum_{r=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \cdot g^\wedge(r) e^{jrt} e^{-jkt} dt \\ &= \sum_{r=-\infty}^{\infty} g^\wedge(r) \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-j[k-r]t} dt = \sum_{r=-\infty}^{\infty} f^\wedge(k-r) g^\wedge(r). \end{aligned} \quad (6.67)$$

Die Summe ist eine Faltung der Folgen der Fourier-Koeffizienten. Im Hinblick auf die diskrete Fourier-Transformation wollen wir aber hier nicht unendliche Folgen, sondern Vektoren mit endlich vielen Komponenten falten. Diese Faltung ist analog definiert:

Definition 6.8 (Faltung von (periodisch fortgesetzten) Vektoren)

Die (diskrete) **Faltung** zweier Vektoren $\vec{y}, \vec{z} \in \mathbb{C}^n$ ist erklärt über

$$\vec{y} * \vec{z} := \left(\sum_{r=0}^{n-1} y_{k-r} z_r \right)_{k=0}^{n-1},$$

wobei wir das Ergebnis in Abhängigkeit von $\vec{y}$ und $\vec{z}$ als Zeilen- oder als Spaltenvektor verstehen.

Zu dieser Definition gelangen wir auch, wenn wir ein Faltungsintegral über eine Riemann-Zwischensumme annähern und die dabei ausgewerteten Funktionswerte mit $f_k := f\left(k\frac{2\pi}{n}\right)$ und $g_k := g\left(k\frac{2\pi}{n}\right)$ bezeichnen:

$$\begin{aligned}(f * g)\left(k\frac{2\pi}{n}\right) &= \frac{1}{2\pi} \int_0^{2\pi} f\left(k\frac{2\pi}{n} - u\right) g(u) du \\ &\approx \frac{1}{2\pi} \sum_{r=0}^{n-1} \frac{2\pi}{n} f\left(k\frac{2\pi}{n} - r\frac{2\pi}{n}\right) g\left(r\frac{2\pi}{n}\right) = \frac{1}{n} \sum_{r=0}^{n-1} f\left([k-r]\frac{2\pi}{n}\right) g\left(r\frac{2\pi}{n}\right) \\ &= \frac{1}{n} \sum_{r=0}^{n-1} f_{k-r} g_r = \frac{1}{n} (f_0, \dots, f_{n-1}) * (g_0, \dots, g_{n-1}).\end{aligned}$$

Die Definition der Faltung von Vektoren entspricht also der Faltung $\frac{1}{2\pi} \int_{-\pi}^{\pi} f(t-u)g(u) du$ zweier 2π -periodischer Funktionen f und g . Statt zu integrieren, wird nun summiert, die Integrationsvariable u wird durch die Summationsvariable r ersetzt. Der Vorfaktor $\frac{1}{2\pi}$ würde einem Faktor $\frac{1}{n}$ bei der diskreten Faltung entsprechen. Der wurde aber in der Definition weggelassen, denn die diskrete Fourier-Transformation haben wir im Gegensatz zu den Fourier-Koeffizienten ohne Vorfaktor definiert. Der Vorfaktor wird dafür bei der diskreten Umkehrtransformation angegeben.

Beispiel 6.39

$$(1, 2, 3) * (4, 5, 6) = (1 \cdot 4 + 3 \cdot 5 + 2 \cdot 6, 2 \cdot 4 + 1 \cdot 5 + 3 \cdot 6, 3 \cdot 4 + 2 \cdot 5 + 1 \cdot 6) \quad \blacksquare$$

Auch die diskrete Faltung ist kommutativ, d. h. $\vec{y} * \vec{z} = \vec{z} * \vec{y}$, denn mit $l = k - r$ und der Periodizität der Vektoren ist

$$\vec{z} * \vec{y} := \left(\sum_{r=0}^{n-1} z_{k-r} y_r \right)_{k=0}^{n-1} = \left(\sum_{l=k-[n-1]}^k z_l y_{k-l} \right)_{k=0}^{n-1} = \left(\sum_{l=0}^{n-1} y_{k-l} z_l \right)_{k=0}^{n-1} = \vec{y} * \vec{z}.$$

Bemerkung: Die Koeffizienten des Polynomprodukts

$$(y_0 x^0 + y_1 x^1 + \dots + y_{n-1} x^{n-1}) \cdot (z_0 x^0 + z_1 x^1 + \dots + z_{n-1} x^{n-1})$$

erhält man über die Faltung

$$(y_0, y_1, \dots, y_{n-1}, \underbrace{0, 0, \dots, 0}_{n-1}) * (z_0, z_1, \dots, z_{n-1}, \underbrace{0, 0, \dots, 0}_{n-1}).$$

Das passt zur Definition des Cauchy-Produkts (siehe Seite 244), und über die Ergebniskoeffizienten einer Polynommultiplikation ist bisweilen die Faltung zweier Vektoren $(y_0, y_1, \dots, y_{n-1})$ und $(z_0, z_1, \dots, z_{n-1})$ alternativ definiert. Da das Produktpolynom einen höheren Grad als die Eingangspolynome hat, hat das Ergebnis aber mehr Komponenten als die Eingangsvektoren. Daher verwenden wir diese Definition hier nicht.

Beispiel 6.40 (Multiplikation natürlicher Zahlen mittels Faltung)

Wir multiplizieren $123 = 3 \cdot 10^0 + 2 \cdot 10^1 + 1 \cdot 10^2$ mit $45 = 5 \cdot 10^0 + 4 \cdot 10^1 + 0 \cdot 10^2$, indem wir wie zuvor das Polynomprodukt für $x = 10$ mittels Faltung berechnen:

$$(3, 2, 1, 0, 0) * (5, 4, 0, 0, 0) = (15, 22, 13, 4, 0).$$

Damit ist $123 \cdot 45 = 15 \cdot 10^0 + 22 \cdot 10^1 + 13 \cdot 10^2 + 4 \cdot 10^3 + 0 \cdot 10^4 = 5\,535$. ■

Die Übertragungsfunktion der Regelungstechnik kann man als Laplace-Transformierte einer Impulsantwort finden (siehe Seite 899). Ähnlich ist folgendes Beispiel:

Beispiel 6.41 (Faltung mit der Impulsantwort)

Die folgende Matrix $\mathbf{A}$ bildet durch Multiplikation einen Vektor $\vec{y}$ auf einen Vektor $\vec{c}$ ab, d. h. $\vec{c} = \mathbf{A}\vec{y}$, indem jede Komponente mit den benachbarten verknüpft wird:

$$\mathbf{A} := \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}.$$

Es gilt also $c_k = 2y_k + y_{k-1} + y_{k+1}$. Die Gewichtung mit Nachbarkomponenten ist eine Situation, wie sie z. B. häufig in der Bildverarbeitung (Weichzeichnen) oder in der Signalverarbeitung auftritt. Verschiebt man das Eingangssignal (Translation des (periodisch fortgesetzten) Vektors $\vec{y}$), so entsteht bei dieser Form der Abbildung ein entsprechend zu $\vec{c}$ verschobenes Ausgangssignal. Man spricht daher von einer **zeitinvarianten** (oder **verschiebeinvarianten**) Abbildung. Bei einer Verschiebung um m Stellen lautet das bei einer $(n \times n)$ -Matrix $\mathbf{A}$ als Formel:

$$\mathbf{A}(y_{k+m})_{k=0 \dots n-1} = (c_{k+m})_{k=0 \dots n-1}. \quad (6.68)$$

Insbesondere erhält man in unserem Beispiel alle Einheitsvektoren durch Verschieben des „Impulses“ $(1, 0, 0, 0)^\top$, die zugehörigen Ausgangssignale sind $\mathbf{A}(1, 0, 0, 0)^\top = (2, 1, 0, 1)^\top$ (Impulsantwort), $\mathbf{A}(0, 1, 0, 0)^\top = (1, 2, 1, 0)^\top$ (um eine Stelle verschobene Impulsantwort), $\mathbf{A}(0, 0, 1, 0)^\top = (0, 1, 2, 1)^\top$ (um zwei Stellen verschobene Impulsantwort) und $\mathbf{A}(0, 0, 0, 1)^\top = (1, 0, 1, 2)^\top$ (um drei Stellen verschobene Impulsantwort). Damit lässt sich die Multiplikation mit einem beliebigen Vektor (die Verarbeitung eines beliebigen Eingangssignals) auf eine Faltung mit der Impulsantwort zurückführen:

$$\mathbf{A} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = 1 \cdot \mathbf{A} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 2 \cdot \mathbf{A} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 3 \cdot \mathbf{A} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + 4 \cdot \mathbf{A} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2 \cdot 1 + 1 \cdot 2 + 0 \cdot 3 + 1 \cdot 4 \\ 1 \cdot 1 + 2 \cdot 2 + 1 \cdot 3 + 0 \cdot 4 \\ 0 \cdot 1 + 1 \cdot 2 + 2 \cdot 3 + 1 \cdot 4 \\ 1 \cdot 1 + 0 \cdot 2 + 1 \cdot 3 + 2 \cdot 4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} * \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

So kann man generell bei linearen, verschiebeinvarianten Abbildungen vorgehen. ■

Ist $\mathbf{A} = \mathbf{E}_n$ die Einheitsmatrix im $\mathbb{C}^{n \times n}$, so erfüllt offensichtlich auch diese das Translationsverhalten (6.68). Der Impuls $\vec{e}_0$ wird auf $\vec{e}_0$ abgebildet. Demnach ist $\vec{e}_0 * \vec{y} = \vec{y}$ für jeden Vektor $\vec{y} \in \mathbb{C}^n$. Das lässt sich mit der Definition der Faltung nachrechnen:

$$\vec{e}_0 * \vec{y} = \left(\sum_{l=0}^{n-1} e_{0,k-l} y_l \right)_{k=0}^{n-1} = \left(\sum_{l=k}^k e_{0,k-l} y_l \right)_{k=0}^{n-1} = (y_k)_{k=0}^{n-1} = \vec{y}.$$

Ähnlich verhält sich auch die Faltung 2π -periodischer Funktionen mit einem Testimpuls, den man dann aber als Distribution definieren muss (vgl. Seite 899).

Auch im diskreten Fall gilt ein **Faltungssatz** (vgl. Seite 837 und Seite 874):

Satz 6.10 (Faltungssatz)

Ist $\text{DFT}_w(\vec{y}) = \vec{c}$ und $\text{DFT}_w(\vec{z}) = \vec{d}$, so gilt:

- a) $\text{DFT}_w(\vec{y} * \vec{z}) = (c_k \cdot d_k)_{k=0}^{n-1}$,
b) $\text{DFT}_w((y_k z_k)_{k=0}^{n-1}) = \frac{1}{n} [\vec{c} * \vec{d}] = \frac{1}{n} [\text{DFT}_w(\vec{y}) * \text{DFT}_w(\vec{z})].$

Dieser Satz kann z. B. genutzt werden, um die Faltung mit einer Impulsantwort zu einer Multiplikation im Frequenzbereich zu vereinfachen. Während die Faltungssätze für Fourier-Koeffizienten und für die Fourier-Transformation über das Vertauschen der Integrationsreihenfolge mit dem Satz von Fubini bewiesen wurden, vertauschen wir hier die (endlichen) Summen. Dafür braucht man keinen Satz:

Beweis: a) Wir zeigen, dass die k -te Komponente des Vektors auf der linken gleich der k -ten Komponente des Vektors auf der rechten Seite ist.

$$\begin{aligned} [\text{DFT}_w(\vec{y} * \vec{z})]_k &= \sum_{l=0}^{n-1} w^{kl} [\vec{y} * \vec{z}]_l = \sum_{l=0}^{n-1} \left[w^{kl} \sum_{r=0}^{n-1} y_{l-r} z_r \right] = \sum_{r=0}^{n-1} \sum_{l=0}^{n-1} w^{kl} y_{l-r} z_r \\ &= \sum_{r=0}^{n-1} w^{kr} z_r \sum_{l=0}^{n-1} w^{k[l-r]} y_{l-r} = \sum_{r=0}^{n-1} w^{kr} z_r \left[\sum_{l=0}^{n-1-r} w^{kl} y_l + \sum_{l=-r}^{-1} w^{kl} y_l \right] \\ &= \sum_{r=0}^{n-1} w^{kr} z_r \left[\sum_{l=0}^{n-1-r} w^{kl} y_l + \sum_{l=-r+n}^{-1+n} w^{k(l-n)} y_{l-n} \right] \end{aligned}$$

$$\begin{aligned}
&= \sum_{r=0}^{n-1} w^{kr} z_r \left[\sum_{l=0}^{n-1-r} w^{kl} y_l + \sum_{l=n-r}^{n-1} w^{kl} y_l \right] = \sum_{r=0}^{n-1} w^{kr} z_r \sum_{l=0}^{n-1} w^{kl} y_l \\
&= \left[\sum_{r=0}^{n-1} w^{kr} z_r \right] \cdot \left[\sum_{l=0}^{n-1} w^{kl} y_l \right] = [\text{DFT}_w(\vec{z})]_k \cdot [\text{DFT}_w(\vec{y})]_k = d_k \cdot c_k.
\end{aligned}$$

b) Da die Fourier-Transformation eine bijektive Abbildung ist und $\frac{1}{n} \text{DFT}_{w^{-1}}$ die Umkehrtransformation zu DFT_w ist, genügt es,

$$\frac{1}{n} \text{DFT}_{w^{-1}} \text{DFT}_w((y_k z_k)_{k=0}^{n-1}) = \frac{1}{n^2} \text{DFT}_{w^{-1}}[\text{DFT}_w(\vec{y}) * \text{DFT}_w(\vec{z})]$$

zu zeigen. Wegen a) für $\text{DFT}_{w^{-1}}$ erhalten wir aber genau diese Gleichung:

$$\begin{aligned}
\frac{1}{n^2} \text{DFT}_{w^{-1}}[\text{DFT}_w(\vec{y}) * \text{DFT}_w(\vec{z})] &= \frac{1}{n^2} (ny_k \cdot nz_k)_{k=0}^{n-1} = (y_k \cdot z_k)_{k=0}^{n-1} \\
&= \frac{1}{n} \text{DFT}_{w^{-1}} \text{DFT}_w((y_k z_k)_{k=0}^{n-1}). \quad \square
\end{aligned}$$

6.4.4 FFT-Algorithmus

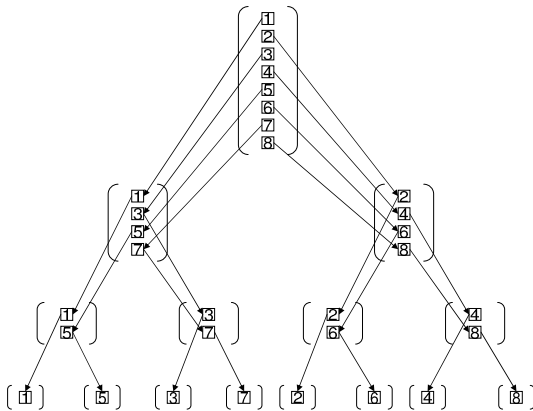


Abb. 6.18: So werden die Komponenten des Ergebnisvektors beim FFT-Algorithmus berechnet.

Zur Berechnung einer diskreten Fourier-Transformation multipliziert man eine $(n \times n)$ -Matrix $\mathbf{F}$ mit einem Vektor. Dabei benötigt man n^2 Multiplikationen. Multiplikationen benötigen auf Computern relativ viel Rechenzeit. Die Matrix $\mathbf{F}$ besitzt zwar n^2 Komponenten, aber alle sind Potenzen einer primitiven Einheitswurzel w , so dass es nur genau n verschiedene Komponenten gibt. Diese spezielle Form der Matrix ermöglicht ein schnelleres Vorgehen: Der berühmte Algorithmus von Cooley und Tukey zur schnellen (diskreten) Fourier-Transformation (**Fast Fourier Transform, FFT**) reduziert die Anzahl der Multiplikationen auf eine Größenordnung $n \log_2 n$, die

für große n erheblich besser als n^2 ist. Denn mit dem Satz von L'Hospital für einen Grenzwert $[\infty]$ (siehe Seite 323) gilt:

$$\lim_{n \rightarrow \infty} \frac{n \log_2 n}{n^2} = \lim_{n \rightarrow \infty} \frac{\frac{\ln n}{\ln 2}}{n} = \frac{1}{\ln 2} \lim_{n \rightarrow \infty} \frac{\frac{1}{n}}{1} = 0.$$

Der Trick dabei ist das schon aus der Antike bekannte Prinzip „**Teile und herrsche!**“. Dazu muss $n = 2^m$ eine Zweierpotenz sein, so dass man n fortlaufend durch 2 teilen kann. Der Algorithmus bildet aus dem zu multiplizierenden Vektor zwei Vektoren mit halb so vielen Komponenten, die einzeln mit $(\frac{n}{2} \times \frac{n}{2})$ -Matrizen multipliziert werden. Dies sind auch wieder Fourier-Matrizen. Aus dem Ergebnis kann man den ursprünglich gesuchten Ergebnisvektor zusammensetzen, vgl. Abbildung 6.18. Die beiden Matrixmultiplikationen benötigen zusammen nur noch $2 \frac{n^2}{4} = \frac{n^2}{2}$ Multiplikationen. Hinzu kommen aber noch einige Multiplikationen, die man für die Berechnung der Vektoren mit halb so vielen Komponenten benötigt. Jetzt führt man nicht die Multiplikationen mit den $(\frac{n}{2} \times \frac{n}{2})$ -Fourier-Matrizen durch, sondern führt diese auf $(\frac{n}{4} \times \frac{n}{4})$ -Fourier-Matrizen zurück usw. Schließlich gelangt man zur (1×1) -Matrix [1], deren Multiplikation trivial ist. Relevant sind damit nur die Multiplikationen, die für den sukzessiven Aufbau der Vektoren halber Länge benötigt werden.

Wir sehen uns jetzt den Algorithmus von Cooley und Tukey für $n = 2^m$ genauer an. Zu berechnen ist das Produkt

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{bmatrix} w^{0 \cdot 0} & w^{0 \cdot 1} & \dots & w^{0 \cdot (n-1)} \\ w^{1 \cdot 0} & w^{1 \cdot 1} & \dots & w^{1 \cdot (n-1)} \\ \vdots & \vdots & & \vdots \\ w^{(n-1) \cdot 0} & w^{(n-1) \cdot 1} & \dots & w^{(n-1) \cdot (n-1)} \end{bmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{pmatrix}.$$

Die Aufgabe wird wie angekündigt in zwei gleichwertige Aufgaben der halben Größe aufgeteilt, die dann ihrerseits wieder genauso aufgeteilt werden usw. Die erste Aufgabe besteht in der Berechnung der Komponenten a_{2k} , $0 \leq k < \frac{n}{2}$ von $\vec{a}$ mit geradem Index. Die zweite Aufgabe ist dann später die Bestimmung der Einträge an ungeradzahligem Positionen.

$$\begin{aligned} a_{2k} &= \sum_{l=0}^{n-1} w^{2kl} y_l = \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} y_l + \sum_{l=\frac{n}{2}}^{n-1} w^{2kl} y_l \quad (n = 2^m \text{ ist durch 2 teilbar}) \\ &= \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} y_l + \sum_{l=0}^{\frac{n}{2}-1} w^{2k(l+\frac{n}{2})} y_{l+\frac{n}{2}} = \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} y_l + \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} \underbrace{w^{nk}}_{=1} y_{l+\frac{n}{2}} \\ &= \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} (y_l + y_{l+\frac{n}{2}}) = \sum_{l=0}^{\frac{n}{2}-1} w^{kl} (y_l + y_{l+\frac{n}{2}}), \end{aligned}$$

wobei $v := w^2$ eine primitive $\frac{n}{2}$ -te Einheitswurzel ist (siehe Lemma 6.10 b)). Damit ist a_{2k} die k -te Komponente von $\text{DFT}_v((y_l + y_{l+\frac{n}{2}})_{l=0}^{\frac{n}{2}-1})$, berechnet mit v statt mit w . Den Vektor halber Größe $(y_l + y_{l+\frac{n}{2}})_{l=0}^{\frac{n}{2}-1}$ kann man gänzlich ohne eine Multiplikation aus den Eingangsdaten erstellen.

Ähnlich können die Komponenten a_{2k+1} , $0 \leq k < \frac{n}{2}$ an ungradzahligen Positionen berechnet werden:

$$\begin{aligned}
 a_{2k+1} &= \sum_{l=0}^{n-1} w^{(2k+1)l} y_l = \sum_{l=0}^{\frac{n}{2}-1} w^{(2k+1)l} y_l + \sum_{l=\frac{n}{2}}^{n-1} w^{(2k+1)l} y_l \\
 &= \sum_{l=0}^{\frac{n}{2}-1} w^{(2k+1)l} y_l + \sum_{l=0}^{\frac{n}{2}-1} w^{(2k+1)(l+\frac{n}{2})} y_{l+\frac{n}{2}} \\
 &\stackrel{(6.58)}{=} \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} w^l y_l + \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} \underbrace{w^{nk}}_{=1} w^l \underbrace{w^{\frac{n}{2}}}_{=-1} y_{l+\frac{n}{2}} \\
 &= \sum_{l=0}^{\frac{n}{2}-1} w^{2kl} w^l (y_l - y_{l+\frac{n}{2}}) = \sum_{l=0}^{\frac{n}{2}-1} v^{kl} w^l (y_l - y_{l+\frac{n}{2}}),
 \end{aligned}$$

wobei wie oben $v = w^2$. Damit ist a_{2k+1} die k -te Komponente von $\text{DFT}_v((w^l[y_l - y_{l+\frac{n}{2}}])_{l=0}^{\frac{n}{2}-1})$ berechnet mit v statt mit w . Den Vektor $(w^l[y_l - y_{l+\frac{n}{2}}])_{l=0}^{\frac{n}{2}-1}$ kann man mit maximal $\frac{n}{2}$ Multiplikationen aus den Ursprungsdaten erstellen. Berücksichtigt man $w^0 = 1$, so sind sogar nur $\frac{n}{2} - 1$ Multiplikationen nötig.

Beim Berechnen von Polynomwerten konnten wir in Kapitel 1.5.6.4 mittels des Horner-Schemas geschickt Faktoren ausklammern und so die Anzahl der Multiplikationen reduzieren. Hier wird ein ähnlicher Trick angewendet, indem in den beiden vorangehenden Rechnungen die Summen geteilt und anschließend (nach einer Indextransformation) so zusammengefasst werden, dass sich gemeinsamen Faktoren ausklammern lassen.

Wir haben das Ursprungsproblem überführt in zwei neue Probleme für $v = w^2$:

$$\begin{aligned}
 (a_0, a_2, a_4, a_6, \dots, a_{n-2}) &= \text{DFT}_v((y_l + y_{l+\frac{n}{2}})_{l=0}^{\frac{n}{2}-1}), \\
 (a_1, a_3, a_5, a_7, \dots, a_{n-1}) &= \text{DFT}_v((w^l[y_l - y_{l+\frac{n}{2}}])_{l=0}^{\frac{n}{2}-1}).
 \end{aligned}$$

$\text{DFT}_v((y_l + y_{l+\frac{n}{2}})_{l=0}^{\frac{n}{2}-1})$ und $\text{DFT}_v((w^l[y_l - y_{l+\frac{n}{2}}])_{l=0}^{\frac{n}{2}-1})$ kann man, da $\frac{n}{2} = 2^{m-1}$ wieder durch zwei teilbar ist, entsprechend aufteilen in jeweils zwei weitere diskrete Fourier-Transformationen für Vektoren der Länge $\frac{n}{4}$.

Man zerlegt nun sukzessive weiter. Dabei hat man stets mit Vektorlängen zu tun, die durch 2 teilbar sind. Schließlich erhält man Vektoren, die nur noch eine Komponente haben. Die Matrix-Multiplikation wird dafür zur Multiplikation mit der Konstanten $v^0 = 1$.

Beispiel 6.42

Zur achten primitiven Wurzel der Eins $w = e^{j\frac{2\pi}{8}} = e^{j\frac{\pi}{4}} = \frac{1}{\sqrt{2}}(1+j)$ berechnen wir $(a_0, a_1, \dots, a_{n-1}) = \text{DFT}_w((0, 1, 2, 3, 4, 5, 6, 7))$. Zunächst ist $w^2 = j$, $w^3 = j\frac{1}{\sqrt{2}}(1+j)$, $j = \frac{1}{\sqrt{2}}(j-1)$, $w^4 = -1$, $w^5 = -\frac{1}{\sqrt{2}}(1+j)$.

Im ersten Schritt wird das Ursprungsproblem in zwei Probleme der halben Größe zerlegt:

$$\begin{aligned}(a_0, a_2, a_4, a_6) &= \text{DFT}_{w^2}((0+4, 1+5, 2+6, 3+7)) \\ (a_1, a_3, a_5, a_7) &= \text{DFT}_{w^2}((w^0[0-4], w^1[1-5], w^2[2-6], w^3[3-7])).\end{aligned}$$

Im zweiten Schritt werden auch diese halbiert:

$$\begin{aligned}(a_0, a_4) &= \text{DFT}_{w^4}((0+4+2+6, 1+5+3+7)) \\ (a_2, a_6) &= \text{DFT}_{w^4}((w^2)^0[0+4-2-6], [w^2]^1[1+5-3-7])) \\ (a_1, a_5) &= \text{DFT}_{w^4}((w^0[0-4] + w^2[2-6], w^1[1-5] + w^3[3-7])) \\ (a_3, a_7) &= \text{DFT}_{w^4}((w^2)^0[w^0[0-4] - w^2[2-6]], [w^2]^1[w^1[1-5] - w^3[3-7]])) \\ &= \text{DFT}_{w^4}((w^0[0-4] - w^2[2-6], w^1[1-5] - w^3[3-7])).\end{aligned}$$

Nochmalige Teilung liefert das Ergebnis:

$$\begin{aligned}(a_0) &= \text{DFT}_{w^8}((0+4+2+6+1+5+3+7)) \\ &= (0+4+2+6+1+5+3+7) = (28) \\ (a_4) &= \text{DFT}_{w^8}([w^4]^0(0+4+2+6-1-5-3-7)) \\ &= (0+4+2+6-1-5-3-7) = (-4) \\ (a_2) &= \text{DFT}_{w^8}((w^2)^0[0+4-2-6] + [w^2]^1[1+5-3-7])) = (-4-4j) \\ (a_6) &= \text{DFT}_{w^8}((w^4)^0([w^2]^0[0+4-2-6] - [w^2]^1[1+5-3-7])) = (-4+4j) \\ (a_1) &= \text{DFT}_{w^8}((w^0[0-4] + w^2[2-6] + w^1[1-5] + w^3[3-7])) \\ &= \left(-4-4j - \frac{4}{\sqrt{2}}(1+j) - \frac{4}{\sqrt{2}}(j-1)\right) = (-4-4[1+\sqrt{2}]j) \\ (a_5) &= \text{DFT}_{w^8}((w^4)^0[w^0[0-4] + w^2[2-6] - w^1[1-5] - w^3[3-7]])) \\ &= \left(-4-4j + \frac{4}{\sqrt{2}}(1+j) + \frac{4}{\sqrt{2}}(j-1)\right) = (-4-4[1-\sqrt{2}]j) \\ (a_3) &= \text{DFT}_{w^8}((w^2)^0[0-4] - w^2[2-6] + w^1[1-5] - w^3[3-7])) \\ &= \left(-4+4j - \frac{4}{\sqrt{2}}(j-1) - \frac{4}{\sqrt{2}}(1+j)\right) = (-4+4(1-\sqrt{2})j) \\ (a_7) &= \text{DFT}_{w^8}((w^4)^0([w^2]^0[0-4] - w^2[2-6] - w^1[1-5] + w^3[3-7])) \\ &= \left(-4+4j + \frac{4}{\sqrt{2}}(j-1) + \frac{4}{\sqrt{2}}(1+j)\right) = (-4+4(1+\sqrt{2})j). \quad \blacksquare\end{aligned}$$

Wir betrachten den Aufwand der schnellen Fourier-Transformation:

- Beim ersten Schritt zerlegt man den Eingangsvektor in zwei Vektoren mit je $n/2 = 2^m/2 = 2^{m-1}$ Komponenten.
- Beim zweiten Schritt zerlegt man diese zwei Vektoren wieder in zwei Vektoren mit nun $2^{m-1}/2 = 2^{m-2}$ Komponenten und erhält $4 = 2^2$ Vektoren.
- Beim k -ten Schritt zerlegt man 2^{k-1} Vektoren in 2^k Vektoren mit 2^{m-k} Komponenten.
- Im m -ten Schritt zerlegt man schließlich 2^{m-1} Vektoren in 2^m Vektoren mit einer Komponente.

Wegen $n = 2^m$ benötigt der Algorithmus also $m = \log_2 n$ Schritte. In jedem Schritt muss man die Vektoren für den nächsten Schritt vorbereiten. Hier fasst man jeweils $n = 2^m$ Komponenten an (Produkt der Anzahl der Vektoren mit der Anzahl ihrer Komponenten). Dabei wird maximal die Hälfte der Komponenten mit einem Faktor w^l multipliziert, so dass wir insgesamt nur eine Größenordnung von $\frac{n}{2} \log_2 n$ Multiplikationen erhalten (siehe Tabelle 6.3).

Tab. 6.3: In der Tabelle ist der Zeitbedarf für n^2 komplexe Multiplikationen im Vergleich zu den ca. $\frac{n}{2} \log_2 n$ Multiplikationen der schnellen Fourier-Transformation dargestellt, wobei wir von 10^8 Multiplikationen pro Sekunde ausgehen. Der Zeitaufwand aller anderen Operationen ist nicht berücksichtigt.

n	Zeitbedarf Multiplikation Matrix mit Vektor (n^2)	Zeitbedarf FFT ($\frac{n}{2} \log_2 n$)
10^3	0,01 s	ca. $5 \cdot 10^{-5}$ s
10^4	1,0 s	ca. $6,6 \cdot 10^{-4}$ s
10^5	100 s	ca. $8 \cdot 10^{-3}$ s
10^6	ca. 2,8 h	ca. 0,1 s

Wir erhalten unmittelbar weitere Anwendungen des Algorithmus:

- Den Algorithmus kann man wegen (6.63) mit w ersetzt durch w^{-1} direkt auch für die inverse diskrete Fourier-Transformation verwenden. Dabei muss das Ergebnis noch mit $\frac{1}{n}$ multipliziert werden.
- Wir können die diskrete Faltung $\vec{y} * \vec{z}$ beschleunigen. Bislang haben wir für jede der n Komponenten $(\vec{y} * \vec{z})_k$ des Ergebnisvektors n Produkte $y_{k-r} z_r$ ($0 \leq r < n$) ausgerechnet und addiert. Das macht n^2 Multiplikationen. Wenn wir aber zunächst beide Vektoren transformieren zu $\vec{c} = \text{DFT}_w(\vec{y})$ und $\vec{d} = \text{DFT}_w(\vec{z})$, so können wir die Faltung mit dem Faltungssatz Satz 6.10 a) auch so schreiben:

$$\vec{y} * \vec{z} = \text{IDFT}_w(c_k \cdot d_k)_{k=0}^{n-1} = \frac{1}{n} \text{DFT}_{w^{-1}}(c_k \cdot d_k)_{k=0}^{n-1}.$$

Dabei wenden wir dreimal die schnelle Fourier-Transformation an (zur Berechnung von $\vec{c}$, $\vec{d}$ und $\text{DFT}_{w^{-1}}(c_k \cdot d_k)_{k=0}^{n-1}$ und müssen zudem ein Produkt pro Komponente bilden. Das ist eine Größenordnung von $n + 3\frac{n}{2}\log_2 n$. Für große n sind damit wesentlich weniger als n^2 Multiplikationen auszuführen (**schnelle Faltung**).

- Die Multiplikation sehr großer ganzer Zahlen lässt sich effizient mit der schnellen Faltung durchführen, vgl. Beispiel 6.40. Das ist die Idee hinter dem Schönhage-Strassen-Algorithmus für die effiziente Multiplikation sehr sehr großer ganzer Zahlen.

Wir haben vorausgesetzt, dass n eine Zweierpotenz ist. Für die Anwendungen ist das in der Regel keine Einschränkung, da man lediglich n groß genug wählen muss und dabei Zweierpotenzen verwenden kann. Es gibt aber auch Varianten des Algorithmus, die ohne diese Voraussetzung auskommen, aber dann nicht ganz so schnell sind. Lässt sich beispielsweise n in zwei ganzzahlige Faktoren p und q zerlegen, d. h. $n = pq$, so kann man auf eine Größenordnung von $n(p + q)$ Multiplikationen gelangen.

6.4.5 Numerische Berechnung von Fourier-Koeffizienten

Nach dem Exkurs zur schnellen Fourier-Transformation kommen wir nun wieder zu unserem Problem aus Kapitel 6.4 zurück. Wir wollen Fourier-Koeffizienten aus abgetasteten Funktionswerten berechnen.

Will man eine Funktion näherungsweise als Fourier-Reihe schreiben, so benötigt man Näherungswerte für die Fourier-Koeffizienten $f^\wedge(k)$ und $f^\wedge(-k)$ paarweise. Denn so kann man eine symmetrische Partialsumme berechnen:

$$\sum_{k=-m}^m f^\wedge(k) e^{jkt} = f^\wedge(0) + \sum_{k=1}^m \left[f^\wedge(-k) e^{-jkt} + f^\wedge(k) e^{jkt} \right].$$

Näherungsweise ergeben sich Fourier-Koeffizienten als Fourier-Lagrange-Koeffizienten $d_k = \frac{1}{n} \sum_{l=0}^{n-1} f_l v^{kl} \approx f^\wedge(k)$ aus (6.51) aus der diskreten Fourier-Transformation $\vec{a} := \text{DFT}_v[(f_0, f_1, \dots, f_{n-1})]$ mit $v = \exp(-j\frac{2\pi}{n})$. Wir müssen lediglich den Vektor $\vec{a}$ mit $\frac{1}{n}$ multiplizieren und seine Komponenten in die richtige Reihenfolge bringen.

- Für eine gerade Anzahl n von Abtastwerten haben wir bereits auf Seite 907 mit (6.54) gesehen, dass

$$d_k = \frac{1}{n} \begin{cases} a_k, & 0 \leq k \leq \frac{n}{2} - 1, \\ a_{n+k}, & -\frac{n}{2} + 1 \leq k < 0. \end{cases}$$

Damit kann man die symmetrische Partialsumme für $m = \frac{n}{2} - 1$ numerisch bilden.

- Falls n ungerade ist, erhält man dazu analog mit (6.53):

$$d_k = \frac{1}{n} \begin{cases} a_k, & 0 \leq k \leq \frac{n-1}{2}, \\ a_{n+k}, & -\frac{n-1}{2} \leq k < 0. \end{cases}$$

So lassen sich die Koeffizienten einer Partialsumme für $m = \frac{n-1}{2}$ annähern.

Insbesondere können wir die Berechnung der $d_k \approx f^\wedge(k)$ z. B. für gerades n mit der Verschiebung in Frequenzbereich (6.64) auch so berechnen:

$$\vec{d} = \frac{1}{n} \text{DFT}_v[(v^{\frac{n}{2}0} f_0, v^{\frac{n}{2}1} f_1, \dots, v^{\frac{n}{2}(n-1)} f_{n-1})].$$

Hier vertauschen die Faktoren die ersten $\frac{n}{2}$ Einträge des Ergebnisvektors mit den zweiten $\frac{n}{2}$ Einträgen (siehe auch (6.55)). Für ungerades n werden die ersten $\frac{n+1}{2}$ Komponenten mit den dann folgenden $\frac{n-1}{2}$ Einträgen wie folgt getauscht:

$$\vec{d} = \frac{1}{n} \text{DFT}_v[(v^{\frac{n+1}{2}0} f_0, v^{\frac{n+1}{2}1} f_1, \dots, v^{\frac{n+1}{2}(n-1)} f_{n-1})].$$

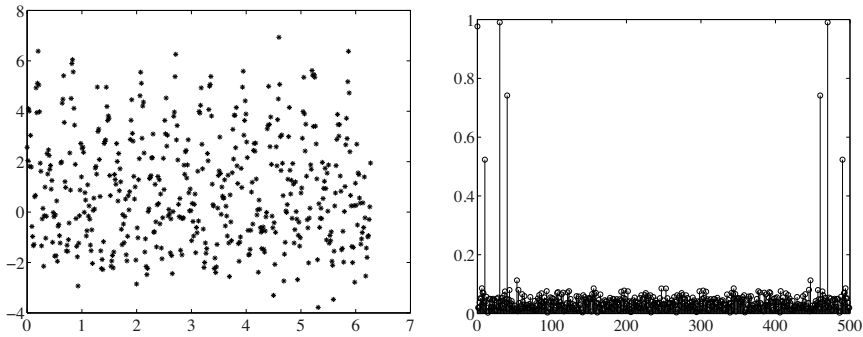


Abb. 6.19: Links: verrauschte Abtastwerte einer Periode, rechts: Beträge der aus den Abtastwerten berechneten Fourier-Koeffizienten

Beispiel 6.43

Wir betrachten ein 2π -periodisches Signal $f(t) = 1 + \sin(10 \cdot t) + 2 \cos(30 \cdot t) + \frac{3}{2} \sin(40 \cdot t)$, das durch Rauschen gestört ist. Mit der diskreten Fourier-Transformation lassen sich die ursprünglichen Kreisfrequenzen über abgetastete Werte rekonstruieren. Zur Simulation des Rauschens wählen wir 500 Abtastwerte von f im Intervall $[0, 2\pi[$ und addieren je eine Zufallszahl (mit Erwartungswert 0 und Varianz 1, vgl. Kapitel 7.2.6). Das Ergebnis entspricht den Abtastwerten eines verrauschten Signals und ist links in Abbildung 6.19 dargestellt. Mit diesen Abtastwerten führen wir eine diskrete Fourier-Transformation durch, um Annäherungen an die Fourier-Koeffizienten zu erhalten. Geplottet sind rechts in Abbildung 6.19 die Beträge der so näherungsweise berechneten komplexen Fourier-Koeffizienten $f^\wedge(k)$, wobei die rechte Bildhälfte (ab 250) Fourier-Koeffizienten zu negativen Indizes darstellt. Wir haben also noch nicht, wie oben beschrieben, die beiden Vektorhälften vertauscht. Ein Strich bei 450 ist der Betrag des Fourier-Koeffizienten $f^\wedge(450 - 500)$. Man erkennt deutlich die Kreisfrequenzen 0, 10, 30 und 40.

Filtern wir die durch das Rauschen entstehenden kleinen Fourier-Koeffizienten weg, so kann man nun durch Bildung der Fourier-Summe zu den verbliebenen Koeffizienten das Ursprungssignal ohne Rauschen sehr gut rekonstruieren. Wir haben ein digitales Filter zur Rauschunterdrückung. ■

Der nächste Abschnitt beantwortet die Frage, welchen Fehler man macht, wenn man in einer Fourier-Reihe die Fourier-Koeffizienten durch die Näherungswerte d_k ersetzt.

6.4.6 Trigonometrische Interpolation, Abtastsatz für trigonometrische Polynome, Aliasing

Die Frage nach der Güte einer näherungsweise berechneten Fourier-Reihe klärt sich überraschenderweise aus einer völlig anderen Fragestellung heraus:

Gegeben ist eine 2π -periodische, stetige Funktion f . Wir suchen ein trigonometrisches Polynom

$$\sum_{k=-n}^n d_k e^{jkt} = \sum_{k=0}^{2n} d_{k-n} e^{j(k-n)t},$$

das an den Stützstellen $t_l = l \frac{2\pi}{2n+1}$, $0 \leq l \leq 2n$, die gleichen Funktionswerte hat wie f (vgl. Abbildung 6.20), also

$$f(t_l) = \sum_{k=0}^{2n} d_{k-n} e^{j(k-n)l \frac{2\pi}{2n+1}}. \quad (6.69)$$

Diese Aufgabenstellung nennt man **trigonometrische Interpolation**. Sie entspricht der algebraischen Interpolation, die wir auf Seite 92 behandelt haben.

Eine Fourier-Partialsumme und ein Interpolationspolynom sind (zunächst) zwei verschiedene Dinge. Bei der Fourier-Partialsumme wird ein trigonometrisches Polynom so berechnet, dass die quadrierte Differenz zur Ursprungsfunktion eine möglichst kleine Fläche unter dem Funktionsgraphen hat. Ein Interpolationspolynom wird so gebildet, dass an vorgegebenen Stellen der Fehler null ist, aber alle anderen Stellen nicht betrachtet werden.

Die $2n+1$ Koeffizienten $d_{-n}, d_{-n+1}, \dots, d_n$ sind als Lösungen des Gleichungssystems (6.69) mit Gleichungen für $0 \leq l \leq 2n$ zu berechnen. Mit der $2n+1$ -ten primitiven Einheitswurzel $w = \exp(j \frac{2\pi}{2n+1})$ ist

$$e^{j(k-n)l \frac{2\pi}{2n+1}} = w^{(k-n)l} = w^{(k-n+2n+1)l} = w^{(k+n+1)l},$$

$$f(t_l) = \sum_{k=0}^{2n} d_{k-n} w^{(k-n)l} = \sum_{k=0}^{n-1} d_{-n+k} w^{(n+1+k)l} + \sum_{k=n}^{2n} d_{k-n} w^{(k-n)l}$$

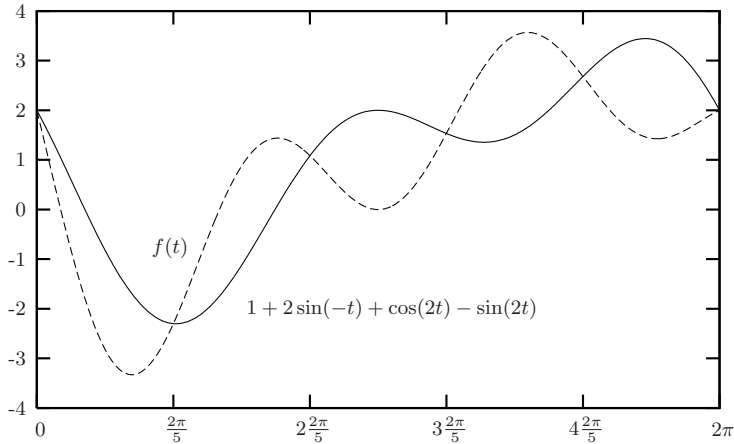


Abb. 6.20: Interpolation einer Funktion f an den Stellen $k \frac{2\pi}{5}$, $k \in \{0, 1, 2, 3, 4\}$ mit dem trigonometrischen Polynom $1 + 2 \sin(-t) + \cos(2t) - \sin(2t)$

$$= \sum_{k=0}^n d_k w^{kl} + \sum_{k=0}^{n-1} d_{-n+k} w^{(n+1+k)l} = (\text{DFT}_w((d_0, d_1, \dots, d_n, d_{-n}, \dots, d_{-1})))_l.$$

Damit erhält man aber die Koeffizienten des Interpolationspolynoms über die inverse diskrete Fourier-Transformation (siehe Seite 913) zu

$$\begin{aligned} (d_0, d_1, \dots, d_n, d_{-n}, \dots, d_{-1}) &= \text{IDFT}_w(f(t_0), \dots, f(t_{2n})) \\ &\stackrel{(6.63)}{=} \frac{1}{2n+1} \text{DFT}_{w^{-1}}(f(t_0), \dots, f(t_{2n})). \end{aligned}$$

Daraus ergeben sich die einige Schlussfolgerungen.

- Mit den $2n+1$ Funktionswerten sind die $2n+1$ Koeffizienten des trigonometrischen Interpolationspolynoms eindeutig bestimmt. Wenn wir für die Ursprungsfunktion f selbst schon ein trigonometrisches Polynom mit Kreisfrequenzen von $-n$ bis n wählen, muss das Interpolationspolynom mit diesem übereinstimmen (Projektion).
- Vergleichen wir die hier bestimmten Koeffizienten mit den bei der numerischen Berechnung der Fourier-Koeffizienten für $2n+1$ mit der Einheitswurzel $v = \exp(-j \frac{2\pi}{2n+1})$ erhaltenen, so sind diese wegen $w^{-1} = v$ identisch! Damit liefert die numerische Berechnung der Fourier-Reihe genau das trigonometrische (Interpolations-)Polynom, das an den Stellen t_l die durch f vorgegebenen Werte hat. Da man diese Form der Interpolation Lagrange-Interpolation nennt, heißen die näherungsweise berechneten Fourier-Koeffizienten auch Fourier-Lagrange-Koeffizienten.

- Hat man also ein trigonometrisches Polynom f mit Kreisfrequenzen von $-n$ bis n , so ist die numerisch approximierte Fourier-Reihe gleich dem Interpolationspolynom gleich der Ursprungsfunktion f . Die Ursprungsfunktion stimmt als trigonometrisches Polynom außerdem mit ihrer Fourier-Reihe überein (siehe Seite 818). Damit berechnen wir die Fourier-Koeffizienten und die Fourier-Reihe exakt (= statt $\approx$):

Satz 6.11 (Abtastsatz für trigonometrische Polynome)

Ist $f(t)$ ein trigonometrisches Polynom

$$f(t) = \sum_{k=-n}^n d_k e^{jkt} = a_0 + \sum_{k=1}^n a_k \cos(kt) + b_k \sin(kt)$$

mit Koeffizienten d_k bzw. a_k und b_k , so sind a_k und b_k genau die reellen und $d_k = f^\wedge(k)$ die komplexen Fourier-Koeffizienten von f , und diese lassen sich exakt berechnen über $2n+1$ Funktionswerte:

$$f^\wedge(k) = \begin{cases} \frac{1}{2n+1} \sum_{l=0}^{2n} f\left(l \frac{2\pi}{2n+1}\right) \left[\exp\left(-j \frac{2\pi}{2n+1}\right)\right]^{kl}, & -n \leq k \leq n, \\ 0, & |k| > n. \end{cases} \quad (6.70)$$

$$\begin{aligned} & (f^\wedge(0), f^\wedge(1), \dots, f^\wedge(n), f^\wedge(-n), f^\wedge(-n+1), \dots, f^\wedge(-1)) \\ &= \frac{1}{2n+1} \text{DFT}_{\exp\left(-j \frac{2\pi}{2n+1}\right)} \left(f(0), f\left(\frac{2\pi}{2n+1}\right), f\left(\frac{2 \cdot 2\pi}{2n+1}\right), \dots, f\left(\frac{2n \cdot 2\pi}{2n+1}\right) \right). \end{aligned}$$

Diesen Zusammenhang werden wir später auch bei der Fourier-Transformation nicht-periodischer Funktionen vorfinden. Dort wird er als Abtastsatz von Shannon und Nyquist (siehe Satz 6.12) formuliert.

Bemerkung: Im Hinblick auf den FFT-Algorithmus ist die ungradzahlige Wahl von $2n+1$ Stützstellen ungünstig, aber hier vereinfacht sie die Darstellung. Modifiziert man die Aufgabe, so dass man ein Interpolationspolynom der Gestalt

$$\sum_{k=-(n-1)}^{n-1} d_k e^{jkt} + \frac{d_n}{2} [e^{jnx} + e^{-jnx}]$$

für $2n$ Stützstellen betrachtet, ergeben sich ähnliche Formeln.

Achtung: Ist n die größte Kreisfrequenz, so benötigt man mindestens $2n+1$ Abtastwerte, um die Funktion exakt zu rekonstruieren. Es genügen im Allgemeinen nicht $2n$ Werte. Häufig findet man hier irreführende Formulierungen. Tastet man z. B. $\sin(2t)$ an den $2 \cdot 2 = 4$ Stellen $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$ ab, so sind alle Abtastwerte 0, und man würde fälschlich alle Fourier-Koeffizienten zu null berechnen.

Verwendet man für das trigonometrische Polynom $f(t) = \sum_{k=-n}^n d_k e^{jkt}$ sogar mehr Funktionswerte, als im Satz gefordert, also z. B. $2m+1$ Werte mit $m > n$, so ergeben sich auch in diesem Fall die exakten Fourier-Koeffizienten (**Oversampling**). Denn das Polynom kann aufgefasst werden als ein Polynom mit Kreisfrequenzen bis m , wobei $d_k = 0, n < |k| \leq m$. Da die Koeffizienten aber schon für $2n+1$ Werte exakt berechnet werden, macht die Verwendung von mehr Funktionswerten hier keinen Sinn.

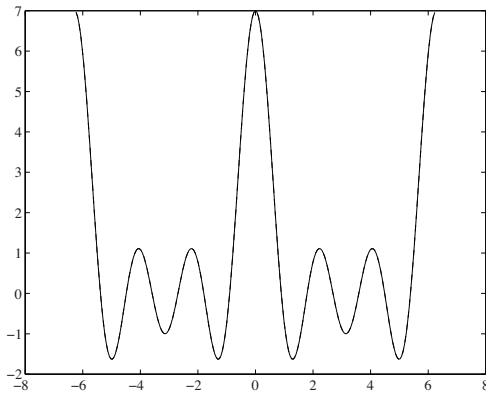


Abb. 6.21: Exakte Approximation von D_3 mittels 7 Abtastwerten

In Abbildung 6.21 sehen wir den Dirichlet-Kern D_3 mit Kreisfrequenzen von -3 bis 3 . Gezeichnet wurde nicht direkt der Graph von $D_3(t)$, sondern die mit den nach (6.70) berechneten Fourier-Koeffizienten gebildete Fourier-Reihe zu $n = 3$. Bei der Berechnung wurden $2n+1 = 7$ Funktionswerte herangezogen, so dass die Koeffizienten exakt bestimmt sind und sich tatsächlich der Graph von $D_3(t)$ zeigt.

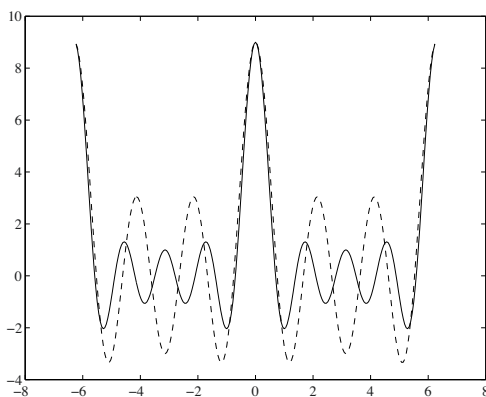


Abb. 6.22: Approximation (gestrichelt) von D_4 (durchgezogen) mittels 7 Abtastwerten

Man macht jedoch einen prinzipiellen Fehler, wenn es noch Kreisfrequenzen $|k| > n$ gibt und man daher mit $2n+1$ zu wenige Abtastwerte verwendet. In den Abbildungen 6.22 und 6.23 sind D_4 bzw. D_6 durchgezogen dargestellt, die näherungsweise mit 7

Abtastwerten (statt der mindestens erforderlichen 9 bzw. 13) berechneten Fourier-Partialsommen gestrichelt. Die Abweichungen sind offensichtlich.

Hier beeinflussen höhere Kreisfrequenzen von f (also k mit $|k| > n$) die approximierten Fourier-Koeffizienten. Dies heißt **Aliasing** oder **Undersampling**. Bei 7

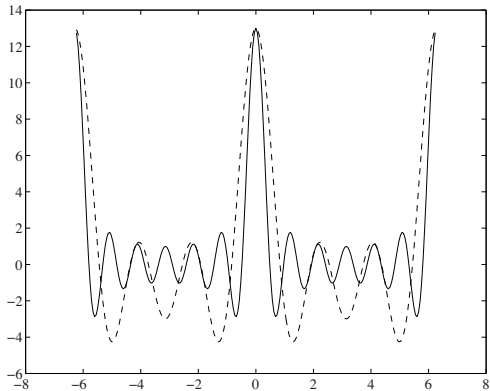


Abb. 6.23: Approximation (gestrichelt) von D_6 (durchgezogen) mittels 7 Abtastwerten

Abtastwerten erhalten wir für Dirichlet-Kerne höherer Ordnung statt der Fourier-Koeffizienten $(f^\wedge(-3), \dots, f^\wedge(3)) = (1, 1, 1, 1, 1, 1, 1)$ die folgenden numerischen Werte:

$$\begin{array}{l|l} D_4 & (2, 1, 1, 1, 1, 1, 2) \\ D_5 & (2, 2, 1, 1, 1, 2, 2) \\ D_6 & (2, 2, 2, 1, 2, 2, 2) \\ D_7 & (2, 2, 2, 3, 2, 2, 2) \\ D_8 & (2, 2, 3, 3, 3, 2, 2) \end{array}$$

Die Ursache ist in Abbildung 6.24 zu erkennen. D_4 setzt sich zusammen als Überlagerung von 1 und den Schwingungen $2\cos(kt)$, $1 \leq k \leq 4$. Wenn wir sieben äquidistante Abtastwerte verwenden, dann haben $2\cos(3t)$ (gestrichelt) und $2\cos(4t) = 2\cos(-4t) = 2\cos([3-7]t)$ (durchgezogen) an diesen äquidistanten Stellen die gleichen Funktionswerte (die Graphen schneiden sich). Wir erkennen damit fälschlich eine $4\cos(3t)$ -Funktion, die statt $f^\wedge(-3) = f^\wedge(3) = 1$ zu den fehlerhaft berechneten Fourier-Koeffizienten $f^\wedge(-3) = f^\wedge(3) \approx 2$ führt. Der $2\cos(4t)$ -Anteil wird quasi am Rand des betrachteten Frequenzbereichs nach innen gespiegelt und überlagert den $2\cos(3t)$ -Anteil. Entsprechend wird ein $2\cos(5t)$ -Anteil von D_5 nach innen gespiegelt und überlagert den $2\cos(2t)$ -Anteil usw.

Wir sehen uns die Ursache des Aliasing-Phänomens etwas genauer an. Bei $2n+1$ Abtastwerten werden nur Funktionswerte an den Stellen $t_l = l \frac{2\pi}{2n+1}$, $0 \leq l \leq 2n$, ausgewertet. Die Funktionen $f_m(t) := e^{j(k+m[2n+1])t}$, $m \in \mathbb{Z}$, haben an den Stellen t_l

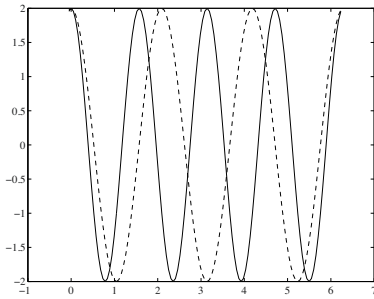


Abb. 6.24: Frequenzanteile $2 \cos(3t)$ (gestrichelt) und $2 \cos(4t)$ (durchgezogen) von D_4 : Wenn genau an den Schnittstellen abgetastet wird, kann nicht zwischen den beiden Funktionen unterschieden werden. Das führt zum Aliasing.

alle die gleichen Funktionswerte und können daher nicht unterschieden werden. Denn es gilt (vgl. (6.53)):

$$e^{j(k+m[2n+1])t_l} = e^{j(k+m[2n+1])l \frac{2\pi}{2n+1}} = e^{j\left(kl \frac{2\pi}{2n+1} + ml2\pi\right)} = e^{jkl \frac{2\pi}{2n+1}} = e^{jkt_l}. \quad (6.71)$$

Hat man nun eine Funktion $f(t) = \sum_{k=-\infty}^{\infty} f^{\wedge}(k)e^{jkt}$, wobei die Fourier-Reihe mindestens an den Stellen t_l absolut konvergieren möge, dann darf man die Summationsreihenfolge vertauschen (siehe Seite 244) und erhält:

$$\begin{aligned} f(t_l) &= \sum_{k=-\infty}^{\infty} f^{\wedge}(k)e^{jkt_l} = \sum_{k=-n}^n \sum_{m=-\infty}^{\infty} f^{\wedge}(k+m(2n+1))e^{j[k+m(2n+1)]t_l} \\ &\stackrel{(6.71)}{=} \sum_{k=-n}^n \left[\sum_{m=-\infty}^{\infty} f^{\wedge}(k+m(2n+1)) \right] e^{jkt_l}. \end{aligned}$$

Die Vertauschung der Summationsreihenfolge geschieht durch Einführung der Summationsvariable m . Wenn k von $-n$ bis n läuft und m alle ganzen Zahlen durchläuft, dann erhält man über $k+m(2n+1)$ ebenfalls jede ganze Zahl genau einmal.

Damit ist aber das Polynom

$$\sum_{k=-n}^n \left[\sum_{m=-\infty}^{\infty} f^{\wedge}(k+m(2n+1)) \right] e^{jkt}$$

das eindeutige Interpolationspolynom vom Grad höchstens n für die Stellen t_l . Die Koeffizienten dieses Interpolationspolynoms sind die diskret berechneten Fourier-Koeffizienten, bei denen damit $f^{\wedge}(k)$ durch höhere Koeffizienten überlagert wird. Statt $f^{\wedge}(k)$ berechnet man also mit der diskreten Fourier-Transformation

$$f^{\wedge}(k) + \sum_{m=1}^{\infty} f^{\wedge}(k-m[2n+1]) + f^{\wedge}(k+m[2n+1]).$$

Bei der Berechnung der Koeffizienten von D_4 erhalten wir so statt $D_4^{\wedge}(3)$ den Wert ($n=3$)

$$\sum_{m=-\infty}^{\infty} D_4^{\wedge}(3+m \cdot 7) = D_4^{\wedge}(3-7) + D_4^{\wedge}(3) = D_4^{\wedge}(-4) + D_4^{\wedge}(3).$$

Da $D_4^\wedge(-4) = D_4^\wedge(4)$ ist, sieht es so aus, als würden die Koeffizienten am Rand gespiegelt. Tatsächlich überlagert sich jeder siebte (allgemein: jeder $2n + 1$ -te) Koeffizient. Dieses Aliasing-Phänomen werden wir auch bei der diskretisierten Fourier-Transformation entdecken.

Beispiel 6.44

In einem Film, der mit 24 Bildern pro Sekunde gezeigt wird, scheinen sich ab und zu Speichenräder in die falsche Richtung zu drehen. Hier wird eine höhere Frequenz falsch abgetastet. Dieser Effekt entsteht auch bei stillstehenden Computerbildern, die aus einzelnen abgetasteten Pixeln zusammengesetzt werden. Ändert sich zwischen zwei Pixeln die Farbe, so geht diese Information verloren. Feine Strukturen werden dadurch verfälscht. Das sieht man in der Nähe des Horizontes am Schachbrettmuster in Abbildung 6.25.

Aufgrund der Auflösung der Wiedergabegeräte werden diagonale Linien häufig als Treppen dargestellt. Auch dieser Effekt wird Aliasing genannt, da zu grob abgetastet wird. ■

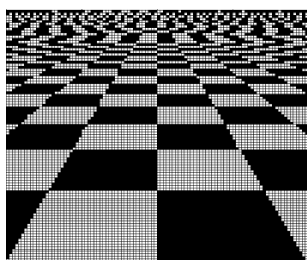


Abb. 6.25: Aliasing in Horizontnähe bei der Abtastung eines Bildes

Eine gängige Lösung zur Verhinderung von Aliasing besteht darin, vor der Fourier-Analyse ein **analoges** Tiefpassfilter auf eine Funktion f anzuwenden, das hohe Kreisfrequenzen $|k| > n$ wegfiltert. Die aus der Faltung von f mit dem Dirichlet-Kern D_n hervorgehende Funktion $g := f * D_n$ (vgl. (6.12) auf Seite 838) entspricht wegen des Faltungssatzes (Satz 6.1 auf Seite 837) der auf diese Weise gefilterten Funktion. Die Funktion g ist nun ein trigonometrisches Polynom mit $2n + 1$ Koeffizienten, das sich mit der diskreten Fourier-Transformation exakt über $2n + 1$ Funktionswerte berechnen lässt. Digitale Oszilloskope verwenden z. B. ein (nicht ganz so gutes) analoges Tiefpassfilter, bevor ein Signal an n Stellen abgetastet und für eine Darstellung der Frequenzen anschließend eine diskrete Fourier-Transformation durchgeführt wird. Ergänzend zu einem analogen Tiefpassfilter kann man auch zunächst noch für höhere Frequenzen abtasten (Oversampling) und dann anschließend ein digitales Tiefpassfilter einsetzen.

Eine mögliche andere Lösung nutzt das Riemann-Lebesgue-Lemma (6.17) aus, das besagt, dass $\lim_{|k| \rightarrow \infty} f^\wedge(k) = 0$ ist. Damit ist die Größe der Fourier-Koeffizienten ab einem genügend groß zu wählenden n von keinem großen Einfluss mehr. Dabei weiß man aber vor dem Sampling in der Regel nicht, wo diese Frequenz liegt. Hier könnte

man iterativ vorgehen und die Sampling-Rate erhöhen, bis keine signifikante Änderung mehr eintritt. Oder man startet mit einer sehr hohen Sampling-Rate und reduziert sie dann in Abhängigkeit der vorgefundenen Fourier-Koeffizienten. Das MP3-Format sieht unterschiedliche Sampling-Raten innerhalb einer Datei vor.

Die maximale Abweichung des Interpolationspolynoms von der Fourier-Partialsumme kann man wie folgt eingrenzen: Für eine differenzierbare, 2π -periodische Funktion f mit beschränkter Ableitung gilt mit Konstanten C_1 und C_2 unabhängig von $n > 1$ für den maximalen Fehler (vgl. (6.20) auf Seite 851):

$$\begin{aligned} \frac{1}{\ln(n)} \max_{t \in [-\pi, \pi]} \left| \sum_{k=-n}^n f^{\wedge}(k) e^{jkt} - f(t) \right| &\leq C_1 \max_{t \in [-\pi, \pi]} \left| \sum_{k=-n}^n d_k e^{jkt} - f(t) \right| \\ &\leq C_2 \ln(n) \max_{t \in [-\pi, \pi]} \left| \sum_{k=-n}^n f^{\wedge}(k) e^{jkt} - f(t) \right|. \end{aligned}$$

6.4.7 Abtastung $2p$ -periodischer Funktionen und Leck-Effekt (Leakage)*

Wir haben zuvor ausschließlich 2π -periodische Signale abgetastet. Völlig analog berechnet man die Fourier-Koeffizienten von $2p$ -periodischen Funktionen, indem man Funktionswerte auf $[-p, p]$ oder $[0, 2p]$ abtastet. Wie bereits bei der Berechnung von Fourier-Koeffizienten zu $2p$ -periodischen Funktionen auf Seite 859 betrachten wir zu einer $2p$ -periodischen Funktion f die 2π -periodische Hilfsfunktion $g(t) := f\left(\frac{2p}{2\pi}t\right) = f\left(\frac{t}{\omega_0}\right)$ mit der Grundfrequenz $\omega_0 = \frac{\pi}{p}$. Diese tasten wir äquidistant auf $[0, 2\pi]$ an Stellen $t_l := l \frac{2\pi}{2n+1}$, $0 \leq l < 2n+1$ ab: $g_l := g(t_l)$. Das entspricht der Abtastung der Funktion f auf $[0, 2p]$ an den Stellen $l \frac{2p}{2n+1}$: $f_l := f\left(l \frac{2p}{2n+1}\right) = g_l$. Die Funktion f wird somit mit einer **Abtastfrequenz** von

$$\frac{1}{\frac{2p}{2n+1}} \text{ Hz} = \frac{2n+1}{2p} \text{ Hz}$$

abgetastet. Mit der diskreten Fourier-Transformation des Vektors $(f_0, f_1, \dots, f_{2n})$ berechnen sich die Fourier-Lagrange-Koeffizienten von g , die eine numerische Näherung der Fourierkoeffizienten $g^{\wedge}$ sind:

$$\begin{aligned} (g^{\wedge}(0), g^{\wedge}(1), \dots, g^{\wedge}(n), g^{\wedge}(-n), \dots, g^{\wedge}(-1)) \\ \approx (d_0, d_1, \dots, d_n, d_{-n}, \dots, d_{-1}) = \frac{1}{2n+1} \text{DFT}_{\exp(-j2\pi/(2n+1))}(f_0, \dots, f_{2n}). \end{aligned}$$

Die Fourier-Koeffizienten c_k zu den Frequenzen $\frac{k}{2p} \text{ Hz} = \frac{\omega_0}{2\pi} \cdot k \text{ Hz}$ der $2p$ -periodischen Funktion f stimmen mit $g^{\wedge}(k)$ überein (siehe Kapitel 6.1.7) und sind damit ebenfalls näherungsweise berechnet. Sie sind wegen des Abtastsatzes sogar exakt bestimmt, falls die Kreisfrequenzen von g kleinergleich n sind, d. h., wenn $c_k = 0$ für $|k| > n$ ist. Das ist

der Fall, wenn f keine Frequenzen größer $\frac{n}{2p}$ Hz oder größer der halben Abtastfrequenz hat.

Beispiel 6.45 (Fourier-Koeffizienten einer 0,1-periodischen Funktion)

Wir betrachten eine reellwertige 0,1-periodische Funktion

$$f(t) := \sum_{k=-n}^n c_k \exp(jk20\pi t) = \sum_{k=-n}^n c_k \exp(jk\omega_0 t),$$

wobei $\omega_0 = \frac{\pi}{0,1/2} = 10 \cdot 2\pi$ die Grundfrequenz ist. Auf dem Intervall $[0, 0,1]$ tasten wir mit mindestens einer Abtastfrequenz $\frac{2n+1}{0,1}$ Hz = $20n + 10$ Hz ab und erhalten damit mindestens $2n + 1$ Funktionswerte $f_l = f(l \cdot 0,1/(2n + 1))$, $0 \leq l \leq 2n$. Damit sind die Fourier-Koeffizienten c_k zu den Frequenzen $\frac{|k|}{0,1} = 10|k|$ bis einschließlich zur Frequenz $\frac{n}{0,1} = 10n$ (und damit bis zur halben Abtastfrequenz $10n + 5$, da Frequenzen nur mit der Schrittweite $1/0,1$ Hz = 10 Hz > 5 Hz auftreten) exakt bestimmt:

$$(c_0, c_1, \dots, c_n, c_{-n}, \dots, c_{-1}) = \frac{1}{2n+1} \text{DFT}_{\exp(-j2\pi/(2n+1))}(f_0, \dots, f_{2n}).$$

Die Koeffizienten c_k und c_{-k} beschreiben die Harmonische $c_{-k} \exp(-jk\omega_0 t) + c_k \exp(jk\omega_0 t)$ zu $10|k|$ Hz, die die Amplitude $2|c_k| = |c_k| + |c_{-k}|$ besitzt. Die Amplitude des Gleichanteils (Frequenz $k = 0$) ist $|c_0|$. ■

Ein reales Problem besteht darin, dass man häufig im Voraus nicht weiß, welche Periode ein Signal hat. Wählt man ein falsches Periodenintervall, so werden gegebenenfalls Informationen des Signals abgeschnitten. Der daraus resultierende Fehler heißt **Leck-Effekt (Leakage)**. Genauer betrachten wir durch das Abschneiden eine neue Funktion, die auf dem angenommenen Periodenintervall identisch mit f ist, über das angenommene Periodenintervall hinaus aber mit der angenommenen Periode periodisch fortgesetzt ist. Solange wir also nur innerhalb des angenommenen Intervalls bleiben, sollte es also keine Probleme geben. Dem ist aber nicht so. Denn durch die Fortsetzung entsteht in der Regel eine Sprungstelle, die sich mit dem Gibbs-Phänomen bemerkbar macht und zu beliebig hohen Frequenzen führt. Bei der Diskretisierung ziehen diese dann zwangsläufig auch noch Aliasing nach sich. Man kann die Sprungstelle in praktischen Anwendungen durch Modifikation des Ursprungssignals verhindern:

- Ein Ausweg ist, die Funktion mit einer (möglichst) stetigen Hilfsfunktion zu multiplizieren, die an den Rändern des Abtastintervalls gleich null ist. Man nennt solche Funktionen **Fensterfunktionen**. In Verbindung mit der Fourier-Transformation werden wir uns Fensterfunktionen noch genauer ansehen (vgl. Kapitel 6.4.10).
- Statt die Länge des Abtastintervalls als volle Periodenlänge aufzufassen, interpretiert man sie als halbe Periodenlänge und setzt das Ursprungssignal stetig durch Spiegelung auf die volle Periodenlänge fort. Damit kann das Signal auf dem Abtastintervall als Kosinus-Reihe geschrieben werden (siehe Seite 847).

Sei f eine $2p$ -periodische Funktion mit Fourier-Reihe $f(t) = \sum_{k=-\infty}^{\infty} f^{\wedge}(k) e^{j \frac{\pi}{p} kt}$. Wir wollen uns ansehen, was passiert, wenn man statt $f^{\wedge}(k) = \frac{1}{2p} \int_{-p}^p f(t) e^{-j \frac{\pi}{p} kt} dt$ fälschlicherweise $\frac{1}{2q} \int_{-q}^q f(t) e^{-j \frac{\pi}{q} kt} dt$ für eine abweichende Periode $2q < 2p$ berechnet ohne die Funktion zu modifizieren. (Da f insbesondere $2pn$ -periodisch ist für $n \in \mathbb{N}$, kann der Fall $q > p$ zurückgeführt werden auf $q < pn$ für ein $n \in \mathbb{N}$.)

Mit der Funktion $1_{[-1,1]}$, die auf dem Intervall $[-1,1]$ gleich eins und sonst gleich null ist, erhalten wir über den Umweg der Fourier-Transformation

$$\begin{aligned}
 \frac{1}{2q} \int_{-q}^q f(t) e^{-j \frac{\pi}{q} kt} dt &= \frac{1}{2q} \int_{-q}^q \sum_{n=-\infty}^{\infty} f^{\wedge}(n) e^{j \frac{\pi}{p} nt} e^{-j \frac{\pi}{q} kt} dt \\
 &= \sum_{n=-\infty}^{\infty} f^{\wedge}(n) \frac{1}{2q} \int_{-q}^q e^{j \frac{\pi}{p} nt} e^{-j \frac{\pi}{q} kt} dt = \sum_{n=-\infty}^{\infty} f^{\wedge}(n) \frac{1}{2q} \int_{-\infty}^{\infty} 1_{[-1,1]} \left(\frac{t}{q} \right) e^{-j \left[\frac{\pi}{q} k - \frac{\pi}{p} n \right] t} dt \\
 &= \sum_{n=-\infty}^{\infty} f^{\wedge}(n) \frac{1}{2q} \mathcal{F} \left[1_{[-1,1]} \left(\frac{t}{q} \right) \right] \left(\frac{\pi}{q} k - \frac{\pi}{p} n \right) \\
 &= \sum_{n=-\infty}^{\infty} f^{\wedge}(n) \frac{1}{2q} q \mathcal{F}[1_{[-1,1]}] \left(q \frac{\pi}{q} k - q \frac{\pi}{p} n \right) \quad (\text{Lemma 6.7 c), Seite 870}) \\
 &= \sum_{n=-\infty}^{\infty} f^{\wedge}(n) \frac{1}{2} 2 \operatorname{sinc} \left(\pi k - \frac{\pi q}{p} n \right) = \sum_{n=-\infty}^{\infty} f^{\wedge}(n) \operatorname{sinc} \left(\frac{\pi p k - \pi q n}{p} \right), \tag{6.72}
 \end{aligned}$$

wobei wir im vorletzten Schritt (6.29) von Seite 864 verwendet haben. Wir erhalten also eine Summe, die einer diskreten Faltung sehr ähnelt (siehe (6.67)). Voraussetzung für diese Rechnung ist die Vertauschbarkeit von Summe und Integral, die z. B. bei gleichmäßiger Konvergenz der Fourier-Reihe nach Satz 2.58 gegeben ist. Gleichmäßige Konvergenz kann man z. B. für eine differenzierbare Funktion f mit beschränkter Ableitung mittels (6.20) zeigen.

Da $\operatorname{sinc}(k\pi) = \frac{\sin(k\pi)}{k\pi} = 0$ für $k \in \mathbb{Z} \setminus \{0\}$ ist, bleibt von der Summe im Fall $p = q$ erwartungsgemäß nur $f^{\wedge}(k)$ übrig. Bei einer Abweichung der Periode kommt es jedoch zu einer über den Sinus-Cardinalis gewichteten Verschmierung mit anderen Fourier-Koeffizienten. Die Koeffizienten (6.72) gehören im Gegensatz zu $f^{\wedge}(k)$ zu einer $2q$ -periodischen Funktion. Entsprechend würden wir damit über eine $2q$ -periodische Fourier-Reihe die Funktion f auf $[-q, q]$ korrekt rekonstruieren. Daher dürfen wir die Differenz von (6.72) und $f^{\wedge}(k)$ nicht als Fehler auffassen - beide Koeffizienten gehören lediglich zu unterschiedlichen Perioden.

6.4.8 Numerische Berechnung der Fourier-Transformation

Wir können nicht nur die Koeffizienten von Fourier-Reihen mittels der diskreten Fourier-Transformation berechnen, sondern auch Funktionswerte der Fourier-Transformation nicht-periodischer Funktionen f . Bei der Berechnung von Fourier-Koeffizien-

ten haben wir Abtastwerte aus genau einem Periodenintervall verwendet. Jetzt ist die Funktion f auf ganz $\mathbb{R}$ definiert und nicht-periodisch. Da wir nur mit endlich vielen Abtastwerten arbeiten können, müssen wir aber die Abtastpunkte zwangsläufig aus einem beschränkten Intervall $[-R, R]$ wählen. Da hier $\lim_{t \rightarrow \pm\infty} f(t) = 0$ sein muss, kann man ein Intervall $[-R, R]$ für ein $R > 0$ auswählen, so dass f außerhalb dieses Intervalls keinen signifikanten Beitrag mehr liefert. Dieses Intervall wird nun z. B. in $2n$ Teilintervalle zerlegt, f wird an $2n$ -Stellen abgetastet, mit denen das Fourier-Integral näherungsweise als Riemann-Zwischensumme (Quadraturformel, vgl. Seite 360) berechnet wird:

$$f^\wedge(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \approx \int_{-R}^R f(t) e^{-j\omega t} dt \approx \frac{R}{n} \sum_{l=-n}^{n-1} f\left(l \frac{R}{n}\right) \exp\left(-j\omega l \frac{R}{n}\right). \quad (6.73)$$

Um die rechte Seite als diskrete Fourier-Transformation zu interpretieren, müssen wir ω speziell wählen. Wir können so nicht alle Funktionswerte von $f^\wedge$ ausrechnen, sondern erhalten auch für die Transformierte nur Abtastwerte. Wir betrachten nun Werte von $f^\wedge$ an den Stellen $\omega_k := \frac{\pi}{R}k$, $-n \leq k < n$. Dafür treten in der Summe Potenzen einer $2n$ -ten primitiven Einheitswurzel auf:

$$\begin{aligned} f^\wedge(\omega_k) &\approx \frac{R}{n} \sum_{l=-n}^{n-1} f\left(l \frac{R}{n}\right) \exp\left(-j \frac{2\pi}{2R} k l \frac{R}{n}\right) \\ &= \frac{R}{n} \sum_{l=-n}^{n-1} f\left(l \frac{R}{n}\right) v^{lk} = \frac{R}{n} \sum_{l=0}^{2n-1} f\left([l-n] \frac{R}{n}\right) v^{[l-n]k} \end{aligned}$$

mit der $2n$ -ten primitiven Einheitswurzel $v = \exp\left(-j \frac{2\pi}{2n}\right)$. Der Exponent k kann jetzt noch negativ sein. Das ändern wir zu $0 \leq k < 2n$:

$$f^\wedge(\omega_{k-n}) = \frac{R}{n} \sum_{l=0}^{2n-1} f\left([l-n] \frac{R}{n}\right) v^{[l-n][k-n]},$$

wobei mit $v^{2n} = 1$ und $v^{\frac{2n}{2}} = v^n = -1$:

$$v^{[l-n][k-n]} = v^{lk} \underbrace{[v^{ln} v^{-2ln}]}_{=1} v^{-nk+n^2} = v^{lk} v^{ln} [v^n]^{n-k} = v^{lk} v^{ln} (-1)^{n-k}.$$

Damit haben wir eine Gestalt erreicht, die wir mit der diskreten Fourier-Transformation berechnen können. Der Faktor v^{ln} führt wegen (6.64) auf Seite 916 lediglich zu einer Vertauschung der beiden Vektorhälften der Transformierten (wie schon bei der Berechnung der Fourier-Reihen). Wir berechnen also

$$\text{DFT}_v \left(\left(f\left(-n \frac{R}{n}\right), f\left([-n+1] \frac{R}{n}\right), \dots, f\left((n-1) \frac{R}{n}\right) \right) \right).$$

Dann vertauschen wir die beiden Vektorhälften. Außerdem müssen wir noch die k -te Komponente des Ergebnisses mit $\frac{R}{n}(-1)^{n-k}$ multiplizieren, da wir diesen Faktor noch nicht berücksichtigt haben. Nach (6.66) entfällt $(-1)^{n-k} = [v^n]^{n-k}$, wenn man die beiden Vektorhälften des Vektors der Funktionswerte vertauscht.

Genau die gleiche Rechnung haben wir in (6.55) auf Seite 907 im Fall einer geraden Anzahl von Abtastwerten für die Berechnung der Fourier-Koeffizienten durchgeführt, wenn wir $R = \pi$ setzen und den Faktor $\frac{1}{2\pi}$ der Fourier-Koeffizienten berücksichtigen. Man kann also den Algorithmus zur Berechnung der Fourier-Transformation mit der Wahl $R = \pi$ direkt auch zur Berechnung von Fourier-Koeffizienten einer 2π -periodischen Funktion einsetzen. Ingenieure unterscheiden daher häufig nicht zwischen der Berechnung von Fourier-Koeffizienten und der Berechnung von Funktionswerten einer Fourier-Transformierten.

Es stellt sich auch hier die Frage, wie gut die so berechneten Werte der Transformierten sind und ob man vielleicht ähnlich wie bei der Diskretisierung der Fourier-Reihen (fast) exakte Werte erhalten kann. Die Antwort liefert der nächste Abschnitt.

6.4.9 Abtastsatz der Fourier-Transformation

Eine periodische Funktion, die keine höhere Kreisfrequenz als n besitzt und daher gleich einem trigonometrischen Polynom ist, konnten wir mittels (6.70) aus $2n + 1$ Funktionswerten rekonstruieren. Eine ähnliche Rekonstruktion einer nicht-periodischen Zeitfunktion $f(t)$ aus diskreten Funktionswerten ist ebenfalls möglich. Damit diese funktioniert, müssen aber auch hier die auftretenden Kreisfrequenzen beschränkt sein. Entsprechende Funktionen heißen bandbegrenzt. Ihre Fourier-Transformierte nimmt nur auf einem beschränkten Intervall $[-\Omega, \Omega]$ von null verschiedene Funktionswerte an. Der Frequenzbereich von f ist also entsprechend beschränkt.

Im Gegensatz zu periodischen Funktionen kommt man aber nicht mit endlich vielen abgetasteten Funktionswerten aus, wenn man eine exakte Rekonstruktion ohne Näherungsfehler haben möchte. Hier benötigt man Funktionswerte zu unendlich vielen diskreten Zeitpunkten $k\Delta t$, $k \in \mathbb{Z}$, um das komplette „Periodenintervall“ $]-\infty, \infty[$ abzutasten (zuvor haben wir $\Delta t := R/n$ verwendet):

Satz 6.12 (Abtastsatz von Shannon und Nyquist)

Sei f stetig auf $]-\infty, \infty[$ mit $\int_{-\infty}^{\infty} |f(t)| dt < \infty$ (also insbesondere Fourier-transformierbar). Außerdem sei für $\Delta t > 0$ die technische Bedingung

$$\sum_{k=-\infty}^{\infty} |f(k\Delta t)| < \infty \quad (6.74)$$

erfüllt. Weiter sei $f^\wedge(\omega) = 0$ für alle $|\omega| > \Omega$. Es dürfen also keine Kreisfrequenzen $|\omega|$ größer als Ω auftreten, f ist in diesem Sinne **bandbegrenzt**.

Unter der **Shannon-Nyquist-Bedingung** für den Abstand Δt von Abtastpunkten

$$\Delta t \leq \frac{\pi}{\Omega}$$

lässt sich die Fourier-Transformierte (analog zu den Fourier-Koeffizienten einer Fourier-Reihe, siehe (6.70)) für jedes $\omega \in [-\Omega, \Omega]$ exakt aus den Funktionswerten von f an den Stellen $k\Delta t$ berechnen (vgl. mit (6.73) für $\Delta t = \frac{R}{n}$):

$$f^\wedge(\omega) = \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) \exp(-j\omega k\Delta t). \quad (6.75)$$

Außerdem lässt sich dann die Funktion f selbst vollständig aus ihren Funktionswerten an den Stellen $k\Delta t$, $k \in \mathbb{Z}$, rekonstruieren. Dabei ist für $t \in \mathbb{R}$:

$$f(t) = \sum_{k=-\infty}^{\infty} f(k\Delta t) \operatorname{sinc}\left(\frac{\pi}{\Delta t}(t - k\Delta t)\right). \quad (6.76)$$

Diese Formel heißt **Whittaker'sche Rekonstruktion**.

Bemerkungen:

- Werden bei $2p$ -periodischen Funktionen $2n + 1$ Abtastwerte verwendet, so sprechen wir von einer Abtastfrequenz von $\frac{2n+1}{2p}$ Hz = $\frac{1}{\Delta t}$ Hz, wobei Δt der Abstand zwischen zwei Abtaststellen ist. Im Rahmen des Shannon-Nyquist-Abtastsatzes verwenden wir ebenfalls $\frac{1}{\Delta t}$ als **Abtastfrequenz** und zählen damit die Samples, also die abgetasteten Funktionswerte, pro $[0, 1]$ -Intervall. f ist aus den Abtastwerten rekonstruierbar, wenn $\frac{1}{\Delta t} \geq \frac{1}{\frac{\pi}{\Omega}} = 2\frac{\Omega}{\pi}$ ist, d. h., wenn mit **mindestens der doppelten Frequenz** (im Sinne von Kreisfrequenz dividiert durch Länge des „Periodenintervalls“ von Sinus und Kosinus) abgetastet wird, die im Fourier-transformierten Zustand auftritt.
- Beim Abtastsatz für 2π -periodische Funktionen (siehe Kapitel 6.4.6) haben wir zuvor die Bedingung kennen gelernt, dass bei einer höchsten Kreisfrequenz $\Omega = n$ mindestens $2n + 1$ Abtastwerte erforderlich sind. Da wir nun unendlich viele Abtastwerte verwenden müssen, wird die Shannon-Nyquist-Bedingung nicht für die Anzahl, sondern für den Abstand Δt zweier Abtastwerte formuliert. Wenn wir das auch im Fall der 2π -periodischen Funktionen tun, dann erhalten wir fast die gleiche Bedingung: Dort ist bei k Abtastwerten der Abstand $\Delta t = \frac{2\pi}{k}$, wobei $k \geq 2n + 1$ sein musste, d. h.

$$\frac{2\pi}{\Delta t} = \frac{2\pi}{\frac{2\pi}{k}} = k \geq 2n + 1 = 2\Omega + 1 > 2\Omega, \text{ d. h. } \Delta t < \frac{\pi}{\Omega}.$$

- Die Voraussetzung, dass f stetig sein soll, ist in dieser Situation keine Einschränkung, sondern natürlich. Einerseits wird es erst durch die Stetigkeit sinnvoll,

mit einzelnen abgetasteten Funktionswerten zu rechnen, da die Werte aussagekräftig für eine ganze Umgebung sind. Andererseits folgt aus der Bandbegrenztheit von f , dass die Fourier-Umkehrtransformation von $f^\wedge$ mit dem Fourier-Integral berechnet werden kann (siehe Kapitel 6.2.2). Man beachte, dass die Umkehrtransformierte wie die Fourier-Transformierte stetig ist. So erhält man als zusätzliche Eigenschaft von f auch noch $\lim_{t \rightarrow \pm\infty} f(t) = 0$ und die Beschränktheit von f .

- Der Abtastsatz (also die exakte Gleichheit in (6.75) und (6.76)) funktioniert tatsächlich nur für bandbegrenzte Funktionen. (6.76) zeigt, wie diese aussehen: Sie sind eine Überlagerung von gedämpften Sinus-Schwingungen (vgl. Seite 271).
- Bandbegrenztheit ist eine sehr einschränkende Bedingung. Wir haben bereits mit Lemma 6.7 c) auf Seite 870 gesehen, dass die Ausdehnung einer Funktion im Zeitbereich gegenläufig zur Ausdehnung im Frequenzbereich sein kann. Tatsächlich verhindert die Begrenzung im Frequenzbereich eine Begrenzung im Zeitbereich. 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 Satz 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 statt (6.76) für jedes $t \in \mathbb{R}$ Brown (1967)

$$f(t) = \lim_{\Delta t \rightarrow 0+} \sum_{k=-\infty}^{\infty} f(k\Delta t) \operatorname{sinc}\left(\frac{\pi}{\Delta t}(t - k\Delta t)\right). \quad (6.77)$$

Diese Konvergenz ist sogar gleichmäßig (siehe Definition 2.43 auf Seite 404).

Ohne Bandbegrenztheit gilt auch die Berechnung von Werten der Transformierten über Werte der Ursprungsfunktion in (6.75) nicht mehr. Den entstehenden Aliasing-Fehler diskutieren wir im Kasten auf Seite 949.

- Man beachte, dass sich die Summe in (6.76) für $t = n\Delta t$ auf den einen Summanden $f(n\Delta t) \operatorname{sinc}(0) = f(n\Delta t)$ reduziert, da $\operatorname{sinc}(k\pi) = 0$ für $k \in \mathbb{Z} \setminus \{0\}$. Hat man unendlich viele Abtastwerte $f(k\Delta t)$, so erhält man über die unendliche Summe eine Funktion, die an allen Abtaststellen $\{n\Delta t : n \in \mathbb{Z}\}$ mit der Ursprungsfunktion f übereinstimmt. Die Summenfunktion interpoliert f an diesen Stellen. Man beachte, dass wir auch in Kapitel 6.4.6 eine Interpolationsfunktion diskutiert haben, dort allerdings für trigonometrische Polynome unter Verwendung endlich vieler Abtaststellen.

- Für $t = n\Delta t$ kann (6.76) als diskrete Faltung der Folgen $(a_k)_{k=1}^\infty := (f(k\Delta t))_{k=-\infty}^\infty$ und $(b_k)_{k=1}^\infty := (\text{sinc}(\pi k))_{k=-\infty}^\infty$ verstanden werden (vgl. Seite 919). Es gilt:

$$\begin{aligned} f(n\Delta t) &= [(b_k)_{k=-\infty}^\infty * (a_k)_{k=-\infty}^\infty]_n := \sum_{k=-\infty}^\infty b_{n-k} a_k \\ &= \sum_{k=-\infty}^\infty \text{sinc}(\pi(n-k)) f(k\Delta t) = \sum_{k=-\infty}^\infty \text{sinc}\left(\frac{\pi}{\Delta t}(n\Delta t - k\Delta t)\right) f(k\Delta t). \end{aligned}$$

Dabei hat der Vektor $(b_k)_{k=-\infty}^\infty$, wie zuvor bemerkt, nur eine Komponente ungleich null. Diese Faltung bildet den Ausgangspunkt für allgemeinere Abtastreihen.

- Die technische Bedingung (6.74) sorgt für einen einfachen Beweis von (6.75). Häufig findet man (6.76) unter der Voraussetzung, dass (für eine stetige und bandbegrenzte Funktion) $\int_{-\infty}^\infty |f(t)|^2 dt < \infty$ ist, z. B. in (Arens et al., 2012, S. 1170). Dabei lässt sich die Definition der Fourier-Transformation auf solche quadrat-integrierbaren Funktionen ausdehnen. Darauf gehen wir hier nicht ein und sichern die Existenz der Fourier-Transformation über die restriktivere Bedingung $\int_{-\infty}^\infty |f(t)| dt < \infty$. Wegen der Beschränktheit von f mit $M > 0$ folgt daraus sofort $\int_{-\infty}^\infty |f(t)|^2 dt \leq M \int_{-\infty}^\infty |f(t)| dt < \infty$.

Die Gleichung (6.75) folgt direkt aus der Poisson-Summationsformel mit (6.35) auf Seite 868 für $h(t) := \Delta t f(t \cdot \Delta t)$ (vgl. Kasten auf Seite 949). Da die Summationsformel hier unbewiesen angegeben ist, wollen wir den Beweis, den Sie gerne überspringen können, explizit führen.

Beweis: Der Beweis basiert auf einer Kombination der Fourier-Umkehrtransformation mit Fourier-Reihen. Die Fourier-Umkehrtransformation (6.30) von $f^\wedge$ funktioniert, da $f^\wedge$ außerhalb von $[-\Omega', \Omega']$ verschwindet (siehe Kapitel 6.2.2).

Nach Voraussetzung ist $\Delta t \leq \frac{\pi}{\Omega}$. Wir setzen $\Omega' := \frac{\pi}{\Delta t}$. Dann ist $\Omega' \geq \Omega$ und $f^\wedge(\omega) = 0$ für alle $|\omega| \geq \Omega'$ (auch für $|\omega| = \Omega'$, da die Fourier-Transformierte stetig ist). Wir führen eine 2π -periodische Hilfsfunktion ein:

$$g(\omega) := f^\wedge\left(\frac{\Omega'}{\pi}\omega\right), \quad -\pi < \omega \leq \pi,$$

sei 2π -periodisch fortgesetzt. Jetzt können wir diese nach Lemma 6.5 stetige Funktion in eine Fourier-Reihe entwickeln. Zur Vereinfachung der Schreibweise benutzen wir dabei eine „anonyme“ Variable $\cdot$, wobei die Funktion $f\left(\frac{\pi}{\Omega'}\cdot\right)$ definiert ist über $f\left(\frac{\pi}{\Omega'}\cdot\right)(t) := f\left(\frac{\pi}{\Omega'}t\right)$.

$$\begin{aligned} g^\wedge(k) &= \frac{1}{2\pi} \int_{-\pi}^\pi f^\wedge\left(\frac{\Omega'}{\pi}v\right) e^{-jkv} dv = \frac{1}{2\pi} \int_{-\infty}^\infty f^\wedge\left(\frac{\Omega'}{\pi}v\right) e^{-jkv} dv \\ &= \frac{\pi}{\Omega'} \frac{1}{2\pi} \int_{-\infty}^\infty \left[f\left(\frac{\pi}{\Omega'}\cdot\right)\right]^\wedge(v) e^{-jkv} dv \quad (\text{Lemma 6.7, Streckung}) \end{aligned}$$

$$= \frac{\pi}{\Omega'} \left[\mathcal{F}^{-1} \left(f \left(\frac{\pi}{\Omega'} \cdot \right) \right) \right] (-k) = \frac{\pi}{\Omega'} f \left(-\frac{\pi}{\Omega'} k \right) = \Delta t f(-k\Delta t).$$

Damit ist die Fourier-Reihe der 2π -periodischen Funktion $g(\omega)$ gleich

$$\sum_{k=-\infty}^{\infty} g^{\wedge}(k) e^{jk\omega} = \sum_{k=-\infty}^{\infty} g^{\wedge}(-k) e^{-jk\omega} = \sum_{k=-\infty}^{\infty} \Delta t f(k\Delta t) e^{-jk\omega}.$$

Die Tatsache, dass g stetig ist, lässt bekanntlich noch nicht darauf schließen, dass diese Reihe auch konvergiert. Wegen (6.74) hat sie aber für jedes ω mit $\sum_{k=-\infty}^{\infty} |f(k\Delta t)|$ eine absolut konvergente Majorante und konvergiert damit nach Satz 2.15 auf Seite 246 für jedes $\omega \in \mathbb{R}$. Wenn die Fourier-Reihe einer stetigen Funktion konvergiert, dann gegen den Funktionswert der Ursprungsfunktion (siehe Lemma 6.3 auf Seite 846). Insbesondere ist nach Definition von g und Ω' für $\omega \in [-\Omega, \Omega] \subset [-\Omega', \Omega']$

$$f^{\wedge}(\omega) = g\left(\frac{\pi}{\Omega'}\omega\right) = g(\omega\Delta t) = \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) e^{-j\omega k\Delta t}.$$

Damit ist die Darstellung der Fourier-Transformierten im Satz gezeigt. Die Darstellung von f erhalten wir durch Anwendung der Fourier-Umkehrtransformation auf die Reihendarstellung von $f^{\wedge}$. Dabei dürfen wir Integration und Summation mit Satz 2.58 (Seite 409) vertauschen. Dies ist wegen der Voraussetzung (6.74) möglich, die nach Lemma 2.12 (Seite 405) die gleichmäßige Konvergenz der Funktionenreihe erzwingt.

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f^{\wedge}(\omega) e^{j\omega t} d\omega = \frac{1}{2\pi} \int_{-\Omega'}^{\Omega'} f^{\wedge}(\omega) e^{j\omega t} d\omega \\ &= \frac{1}{2\pi} \int_{-\Omega'}^{\Omega'} \Delta t \left[\sum_{k=-\infty}^{\infty} f(k\Delta t) e^{-j\omega k\Delta t} \right] e^{j\omega t} d\omega \\ &= \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) \frac{1}{2\pi} \int_{-\Omega'}^{\Omega'} e^{j\omega[t-k\Delta t]} d\omega = \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) \frac{\Omega'}{2\pi} \int_{-1}^1 e^{j\Omega' v[t-k\Delta t]} dv \end{aligned}$$

mit der Substitution $v = \frac{\omega}{\Omega'}$, $dv = \frac{1}{\Omega'} d\omega$.

Im Beispiel (6.29) auf Seite 864 haben wir diese Fourier-Transformation berechnet:

$$\int_{-1}^1 e^{j\Omega' v[t-k\Delta t]} dv = \int_{-1}^1 e^{-jv[-\Omega'[t-k\Delta t]]} dv = 2 \operatorname{sinc}(-\Omega'(t-k\Delta t)) = 2 \operatorname{sinc}(\Omega'(t-k\Delta t)),$$

da sinc eine gerade Funktion ist. Dies eingesetzt ergibt:

$$f(t) = \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) \frac{\Omega'}{\pi} \operatorname{sinc}(\Omega'(t-k\Delta t)) = \sum_{k=-\infty}^{\infty} f(k\Delta t) \operatorname{sinc}\left(\frac{\pi}{\Delta t}(t-k\Delta t)\right). \quad \square$$

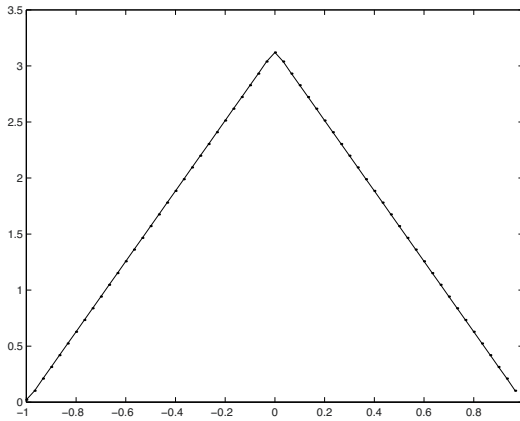


Abb. 6.26: Approximation von $\mathcal{F}\left(\frac{1-\cos t}{t^2}\right)$ mit $R = 30\pi$ und $n = 30$

Eine Abtastung einer Funktion f an unendlich vielen Stellen ist nur theoretisch denkbar, nicht jedoch technisch umsetzbar. Bei der im vorangehenden Abschnitt 6.4.8 hergeleiteten Diskretisierung der Fourier-Transformation haben wir f daher nur an endlich vielen Stellen (nämlich an $2n$) aus einem Intervall $[-R, R[$ (statt $]\infty, \infty[$) abgetastet und $\Delta t = \frac{R}{n}$ gewählt. (6.73) sieht also so aus:

$$f^\wedge(\omega) \approx \frac{R}{n} \sum_{l=-n}^{n-1} f\left(l \frac{R}{n}\right) \exp\left(-j\omega l \frac{R}{n}\right) = \Delta t \sum_{l=-n}^{n-1} f(l\Delta t) \exp(-j\omega l \Delta t).$$

Diese Formel stimmt mit (6.75) überein, falls die Funktion f außerhalb des Intervalls $[-R, R[$ null ist. Das ist aber nach den vorangehenden Bemerkungen zum Abtastsatz bei bandbegrenzten Funktionen nur für die Nullfunktion $f(t) = 0$ möglich. Daher sollte R zumindest so groß sein, dass die Funktionswerte von f außerhalb des Intervalls $[-R, R[$ keinen wesentlichen Beitrag mehr liefern. Das ist möglich, da $\lim_{u \rightarrow \infty} \left[\int_{-\infty}^{-u} |f(t)| dt + \int_u^{\infty} |f(t)| dt \right] = 0$ ist. Wählt man das Intervall $[-R, R[$ zu klein, so steht nicht genügend Information über die Ursprungsfunktion zur Verfügung, und man erhält einen Fehler wie in Abbildung 6.27, den man wie bei Fourier-Reihen auch hier **Leck-Effekt** (Leakage) nennt. In Abschnitt 6.4.10 wird dieser nicht komplett vermeidbare Fehler näher untersucht.

Laut Abtastsatz sollte n nun so gewählt werden, dass

$$\frac{R}{n} = \Delta t \leq \frac{\pi}{\Omega}, \text{ also } n \geq \frac{R\Omega}{\pi}.$$

Im Frequenzbereich erhält man so äquidistante Werte $\omega_k = \frac{\pi}{R}k$, $-n \leq k < n$, aus dem Intervall $[-\frac{\pi}{R}n, \frac{\pi}{R}(n-1)]$. Die Schrittweite $\frac{\pi}{R}$ im Frequenzbereich ist unabhängig von der Größe von Δt . Wählt man die Grenzfrequenz $n = \frac{R\Omega}{\pi}$, so ergibt sich genau eine Zerlegung von $[-\Omega, \Omega]$ (siehe Abbildung 6.26). Wählt man dagegen n größer, bekommt man auch Frequenzen außerhalb dieses Intervalls. Diese sind aber 0, so dass wir durch eine höhere Abtastfrequenz **keine** Verbesserung erzielen (siehe Abbildung 6.28).

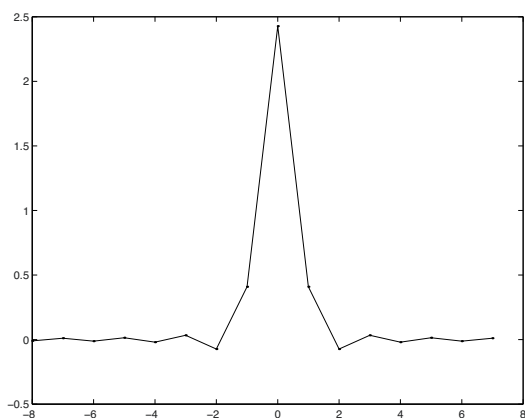


Abb. 6.27: Leck-Effekt bei der Approximation von $\mathcal{F}\left(\frac{1-\cos t}{t^2}\right)$ mit $R = \pi$ und $n = 8$

Wählt man dagegen n zu klein, so können wie bei den Fourier-Reihen Aliasing-Effekte auftreten. Für zu kleines n zeigt Abbildung 6.30, dass an den Rändern $\pm \frac{\pi}{R}n$ bei einer Funktion $f^\wedge$ Fehler durch Superposition mit höheren Frequenzen entstehen (vgl. Abschnitt 6.4.6 und den folgenden Kasten). Die Grafik 6.29 verdeutlicht die Entstehung dieses Effekts. Damit sollte n möglichst so gewählt sein, dass $n \geq \frac{R\Omega}{\pi}$, $n < \frac{R\Omega}{\pi} + 1$.

Beispiel 6.46

Häufig werden Signale mit einer vorgegebenen Abtastfrequenz von f_a Werten pro Sekunde abgetastet – also mit f_a Hz (vgl. Beispiel 6.45). Zwei Abtastwerte liegen $\Delta t = \frac{1}{f_a}$ Sekunden auseinander. Verwendet man nun n Abtastwerte auf dem Zeitintervall $[0, (n-1)\Delta t] = \left[0, \frac{(n-1)}{f_a}\right]$, so lassen sich mit der diskreten Fourier-Transformation

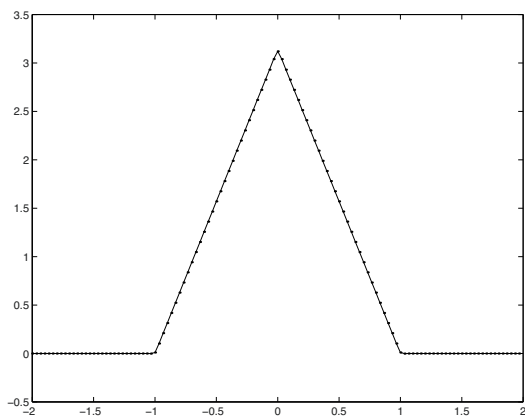


Abb. 6.28: Approximation von $\mathcal{F}\left(\frac{1-\cos t}{t^2}\right)$ mit $R = 30\pi$ und $n = 60$

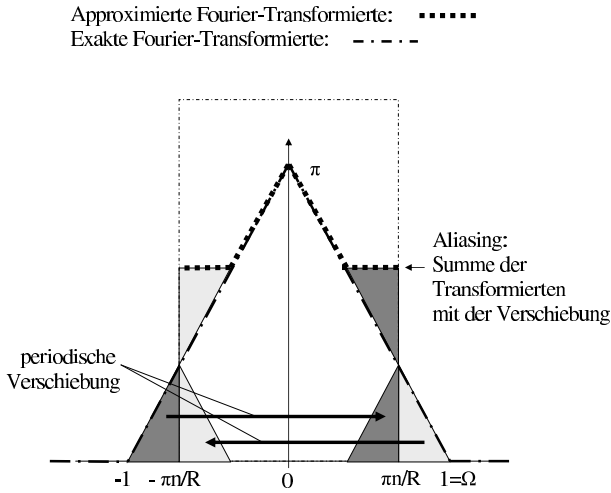


Abb. 6.29: Aliasing-Effekt bei zu grober Abtastrate mit $\frac{\pi}{R}n < \Omega$

Näherungswerte der Fourier-Transformierten an Stellen $\omega_k = \frac{2\pi}{n\Delta t}k = \frac{2\pi f_a}{n}k$ für $-\frac{n}{2} \leq k < \frac{n}{2}$ berechnen:

$$f^\wedge\left(\frac{2\pi f_a}{n}k\right) \approx \frac{1}{f_a} \sum_{l=0}^{n-1} f\left(\frac{l}{f_a}\right) \exp\left(-j2\pi k \frac{l}{n}\right).$$

Die Shannon-Nyquist-Bedingung lautet hier $f_a \geq \frac{\Omega}{\pi}$ für eine mit Ω bandbegrenzte Funktion f . ■

Für Funktionen, deren Transformierte nicht außerhalb eines Intervalls $[-\Omega, \Omega]$ verschwinden, die also nicht bandbegrenzt sind, ist der Abtastsatz nicht anwendbar. Hier könnte man ein (analoges) Tiefpassfilter vorschalten, das hohe Frequenzen entfernt.

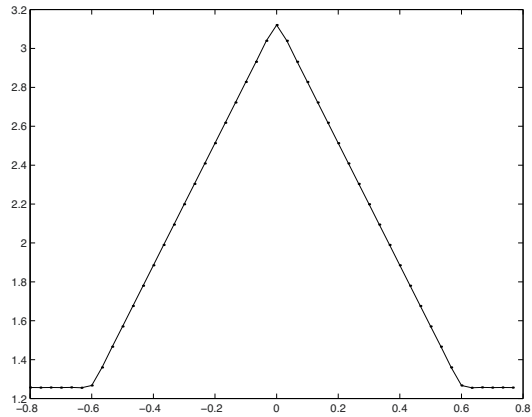


Abb. 6.30: Approximation von $\mathcal{F}\left(\frac{1-\cos t}{t^2}\right)$ mit $R = 30\pi$ und $n = 25$

Alternativ kann man aber auch das Riemann-Lebesgue-Lemma (Lemma 6.5) ausnutzen. Da $\lim_{|\omega| \rightarrow \infty} f^\wedge(\omega) = 0$ ist, kann man Ω so groß wählen, dass die Funktionswerte von $f^\wedge$ außerhalb von $[-\Omega, \Omega]$ sehr klein sind und nahezu keinen Beitrag mehr leisten. Dann ist die Funktion nahezu bandbegrenzt mit Ω . Die Shannon-Nyquist-Bedingung bewirkt, dass bei einer Vergrößerung von Ω die Abtastfrequenz größer und Δt kleiner werden muss. Aber auch wenn man Ω sehr groß wählt, bleibt ein kleiner prinzipieller Fehler. Dies ist der gleiche Aliasing-Fehler, der auch bei der Wahl einer zu kleinen Abtastfrequenz bei bandbegrenzten Funktionen auftritt (siehe Kasten).

Hintergrund: Aliasing

Der Aliasing-Effekt tritt auf, wenn eine Funktion nicht bandbegrenzt ist oder wenn die Abtastrate zu klein gewählt wird, so dass die Shannon-Nyquist-Bedingung verletzt ist. Was hier genau geschieht, erkennt man durch Anwendung der Poisson-Summationsformel. 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 (6.35) erfüllen, die sich aus der Poisson-Summationsformel ergibt (siehe Seite 868). Dann wird (6.35) 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

$$f^\wedge(\omega) - \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) e^{-j\omega k\Delta t} = - \underbrace{\left[\sum_{k=1}^{\infty} f^\wedge\left(\omega - k\frac{2\pi}{\Delta t}\right) + \sum_{k=1}^{\infty} f^\wedge\left(\omega + k\frac{2\pi}{\Delta t}\right) \right]}_{\text{Aliasing-Fehler}}. \quad (6.78)$$

Diese Darstellung entspricht genau der Situation, die wir auch schon bei der Berechnung von Fourier-Koeffizienten auf Seite 935 vorgefunden haben. Der Aliasing-Fehler verschwindet, wenn f bandbegrenzt und die Shannon-Nyquist-Bedingung erfüllt ist. Dann erhält man die Formel (6.75) aus dem Abtastsatz. Sonst führen die zusätzlichen Summen zu dem in Abbildung 6.29 dargestellten Effekt.

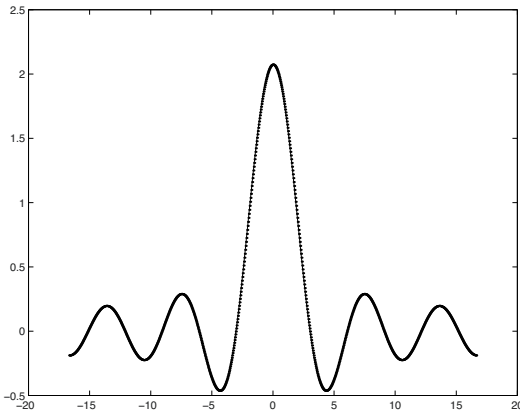


Abb. 6.31: Approximation der Fourier-Transformierten $2 \operatorname{sinc}(\omega)$ mittels $R = 30\pi$ und $n = 500$

Beispiel 6.47

Der Rechteckimpuls (6.28) hat die Fourier-Transformierte $2 \operatorname{sinc} \omega$, die in Abbildung 6.12 dargestellt ist. Diese ist nicht bandbegrenzt. Tastet man $2n = 1000$ Funktionswerte innerhalb des Intervalls $[-R, R] = [-30\pi, 30\pi]$ ab, so erhält man die in Abbildung 6.31 eingezeichnete Näherung für $2 \operatorname{sinc} \omega$, die dem Graphen in Abbildung 6.12 bereits sehr nahekommt. Das liegt daran, dass die Transformierte außerhalb des Intervalls $[-\frac{\pi}{R}n, \frac{\pi}{R}(n-1)] = [-\frac{500}{30}, \frac{499}{30}]$ bereits sehr klein ist. ■

6.4.10 Leck-Effekt und Fensterfunktionen *

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 (6.28) kennen wir bereits die Fourier-Transformierte:

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

Der Faltungssatz im Frequenzbereich (Folgerung 6.2, siehe insbesondere die Anmerkung zu den Voraussetzungen) besagt nun, dass

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

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 (siehe (6.72)).

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 (6.78) die Funktion $f \cdot g$ einsetzen. Ist die Ursprungsfunktion f bandbegrenzt und die Fensterfunktion g so gewählt, dass $(f \cdot g)^\wedge = f^\wedge * g^\wedge$ sich möglichst gut an $f^\wedge$ 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 (6.78) ebenfalls klein.

Die $2R \operatorname{sinc}(R\omega)$ -Funktion als Transformierte des Rechteckfensters $1_{[-R, R]}(t)$ hat ein absolutes Maximum bei $\omega = 0$, aber unendlich viele weitere positive Maxima und negative Minima, siehe erste Zeile in Abbildung 6.32. Aufgrund des Aussehens des Funktionsgraphen spricht man bei diesen Extrema von Keulen. 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 hohe Hauptmaximum (die hohe Hauptkeule) 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 $\operatorname{sinc}(R\omega)$ um den Nullpunkt zusammen, und weiter entfernte Werte werden weniger berücksichtigt, Leck- und Aliasing-Fehler werden kleiner.

Günstiger als Rechteckfenster sind Fensterfunktionen g , deren Fourier-Transformierte $g^\wedge$ der δ -Distribution ähneln, vgl. Kasten auf Seite 898. Denn die δ -Distribution ist das neutrale Element der Faltung. Leider ist aber die δ -Distribution keine Funktion im klassischen Sinne, und es gibt keine Fensterfunktion, die sie als Transformierte hat. Damit muss man einen Kompromiss eingehen. Günstig ist, wenn $g^\wedge$ betragsmäßig kleine Nebenmaxima (Nebenkeulen) hat. Fallen diese für $|\omega| \rightarrow \infty$ schnell ab, dann werden Leck- und Aliasing-Fehler bei der Verbreiterung des Abtastintervalls für $R \rightarrow \infty$ schnell klein. Allerdings erkaufte man sich das schnelle Abfallen in der Regel mit relativ großen Nebenmaxima in der Nähe der Hauptkeule bei Null. Da sich in der Praxis das Abtastintervall nicht beliebig vergrößern lässt, führen die großen Nebenmaxima

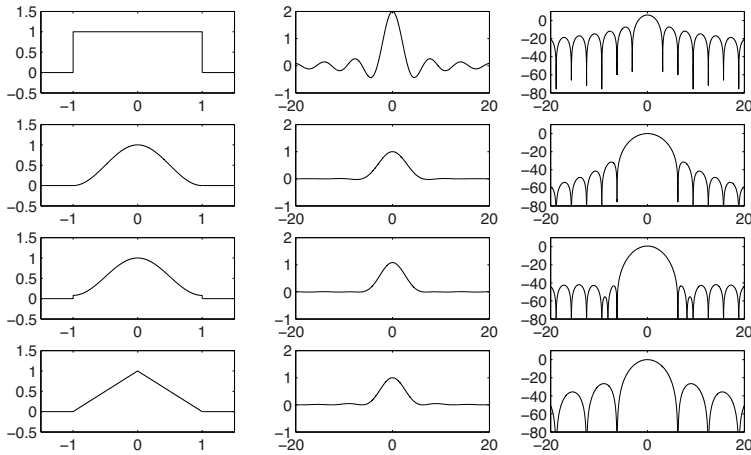


Abb. 6.32: Rechteck-, Hann-, Hamming- und Bartlett-Fenster: Links sind die Fensterfunktionen $g(t)$ zu $R = 1$ und in der Mitte die zugehörigen Transformaten $g^\wedge(\omega)$ eingezeichnet. Die Darstellung trägt: Keine der Fensterfunktionen ist bandbegrenzt. Das sieht man besser, wenn man eine logarithmische Darstellung der Graphen benutzt. Dazu ist rechts $10 \cdot \lg(|g^\wedge(\omega)|^2)$ eingezeichnet (Intensitätsverhältnisse in Dezibel).

bei Null durch die Faltung dazu, dass nah benachbarte Frequenzen nicht gut aufgelöst werden können. Es gilt also, einen Kompromiss zwischen schnellem Abfallen der Nebenkeulen und kleinen Nebenkeulen in der Nähe von Null zu finden.

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 Transformierte (vgl. (6.34), Seite 867)

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

Durch das Quadrat fällt diese Transformierte für $|\omega| \rightarrow \infty$ schneller ab als die sinc-Funktion. $g^\wedge$ entspricht dem Fejér-Kern bei 2π -periodischen Funktionen (siehe Kasten auf Seite 845), und die Faltung von $f^\wedge$ mit $g^\wedge$ strebt für $R \rightarrow \infty$ entsprechend gut gegen $f^\wedge$. Das ist ein Ergebnis der **Approximationstheorie**. In der Nachrichtentechnik werden andere gängige Fensterfunktionen eingesetzt, um den oben beschriebenen Kompromiss zu realisieren. Sie ergeben sich durch Wahl des Parameters $0 \leq a \leq 1$ aus

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

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

$$g(t) := \begin{cases} \frac{1}{2} + \frac{1}{2} \cos\left(\frac{\pi}{R}t\right) \stackrel{(1.25)}{=} \cos^2\left(\frac{\pi}{2R}t\right) & : -R \leq t \leq R \\ 0 & : |t| > R, \end{cases}$$

so dass man g auch $\cos^2$ -Fenster nennt. Um die Nebenkeulen in der Nähe der Hauptkeule bei Null zu reduzieren, 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 6.32):

$$g(t) := \begin{cases} 0,54 + 0,46 \cos\left(\frac{\pi}{R}t\right) \stackrel{(1.25)}{=} 0,08 + 0,92 \cos^2\left(\frac{\pi}{2R}t\right) & : -R \leq t \leq R \\ 0 & : |t| > R. \end{cases}$$

Die Fourier-Transformierte von g wird in Aufgabe 6.22 berechnet:

$$g^\wedge(\omega) = 2R \operatorname{sinc}(\omega R) \left[a + (1-a) \frac{1}{\left(\frac{\pi}{R\omega}\right)^2 - 1} \right].$$

Für $|\omega| \rightarrow \infty$ strebt der Term in der eckigen Klammer gegen $a - (1-a) = 2a - 1$. Insbesondere im Fall des Hann-Fensters ($a = \frac{1}{2}$) werden dadurch die Schwingungen der sinc-Funktion zusätzlich gedämpft, der Aliasing-Fehler ist (für großes R) günstiger als beim Rechteck- und Bartlett-Fenster. Beim Hamming-Fenster strebt der Term in der eckigen Klammer zwar nicht gegen null, dafür ist er aber, wie angekündigt, nahe beim Hauptmaximum kleiner als beim Hann-Fenster, siehe Abbildung 6.32. Das liegt daran, dass die Transformierte eine Linearkombination der Transformaten des Sockels, also des Rechteckfensters und der Transformaten des Hann-Fensters ist. Dabei nutzt man aus, dass die Hauptkeule der Transformaten $2 \operatorname{sinc}(\omega)$ des Rechteckfensters für $R = 1$ von den Nullstellen bei $-\pi$ bis π reicht. An diesen Stellen wechselt der Sinus-Cardinalis das Vorzeichen und ist auf $] - 2\pi, -\pi[$ und $] \pi, 2\pi[$ negativ. Dann wechselt das Vorzeichen, so dass die Funktion auf $] - 3\pi, -2\pi[$ und $] 2\pi, 3\pi[$ wieder positiv ist. Die Transformierte des von Hann-Fensters dagegen ist auf $] - 2\pi, 2\pi[$ positiv und auf $] - 3\pi, -2\pi[$ und $] 2\pi, 3\pi[$ negativ (siehe Abbildung 6.32). Aufgrund des unterschiedlichen Vorzeichens kommt es durch die Addition des Sinus-Cardinalis zu Auslöschungen auf $] - 3\pi, -\pi[$ und $] \pi, 3\pi[$.

Weitere Fenster findet man z. B. in (Butz, 2007, S. 87).

6.4.11 Zusammenfassung

Möchte man die Fourier-Koeffizienten einer Fourier-Reihe oder Werte einer Fourier-Transformierten praktisch ausrechnen, benutzt man dazu endlich viele Funktionswerte der Ursprungsfunktion. Diese werden in einem Vektor zusammengefasst, mit dem dann eine diskrete Fourier-Transformation (Matrix-Multiplikation mit einer speziellen Matrix) vorgenommen wird. Das Ergebnis sind näherungsweise berechnete Fourier-Koeffizienten oder Funktionswerte der Fourier-Transformierten. Unter gewissen Voraussetzungen ist das Ergebnis für Fourier-Koeffizienten sogar exakt, nämlich dann, wenn die Ursprungsfunktion keine zu hohen Frequenzen besitzt (Abtastsatz). Diese kann man aber vor dem Ausrechnen der Fourier-Koeffizienten mit einem Tiefpassfilter beseitigen. Außerdem muss man die richtige Periodenlänge kennen. Für die Fourier-Transformation gilt ein entsprechender Abtastsatz, wobei man aber einen zusätzlichen kleinen Fehler dadurch erhält, dass man im Gegensatz zur Berechnung der Fourier-Koeffizienten (hier betrachtet man als Abtastintervall eine volle Periode der Funktion) ein beschränktes Abtastintervall $[-R, R]$ auswählen muss und nicht $]-\infty, \infty[$ verwenden kann. Dabei tritt ein Aliasing-Fehler auf, der durch eine genügend hohe Abtastrate und geeignete Fensterfunktion klein gehalten werden kann.

6.5 Wavelets und schnelle Wavelet-Transformation *

In die Berechnung von Fourier-Koeffizienten und von Werten der Fourier-Transformierten gehen alle Werte der Eingangsfunktion ein. Schöner wäre es, wenn sich nur Funktionswerte aus kleinen Intervallen auf gewisse zugehörige Werte der Fourier-Transformierten auswirken würden. Denn in der Praxis lassen sich 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.

Das folgende Beispiel zeigt, wie sich bei einer Verschiebung des Abtastfensters tatsächlich alle Werte eines Fourier-Spektrums ändern.

Beispiel 6.48 (Fortlaufende diskrete Fourier-Transformation)

Wir möchten ein zeitlich fortlaufendes Signal Fourier-analysieren, so wie es beispielsweise für die Frequenzanzeige einer Stereo-Anlage erforderlich ist.

In einem Abtastintervall stehen uns zunächst n Funktionswerte $f_0, \dots, f_{n-1}$ zur Verfügung, mit denen wir über die diskrete Fourier-Transformation (mit dem FFT-

Algorithmus) ein Spektrum berechnen können. Das Ergebnis der diskreten Fourier-Transformation sei $(a_k)_{k=0}^{n-1} := DFT_w((f_0, \dots, f_{n-1}))$, also $a_k = \sum_{i=0}^{n-1} w^{ki} f_i$.

Dann werden alle Abtastwerte um eine Position nach links geschoben. f_0 fällt weg, dafür kommt rechts ein neuer Abtastwert f_n dazu. Eine erneute diskrete Fourier-Transformation führt in der Regel zu völlig anderen Werten $(b_k)_{k=0}^{n-1} := DFT_w((f_1, \dots, f_n))$. Statt diesen Vektor mit einem Aufwand der Größenordnung $n \lg(n)$ vollständig neu zu berechnen, können wir schneller mit dem bereits vorliegenden ersten Ergebnis $(a_k)_{k=0}^{n-1}$ weiterrechnen:

$$\begin{aligned} b_k &= \sum_{i=0}^{n-1} w^{ki} f_{i+1} = \sum_{i=1}^n w^{k(i-1)} f_i = w^{k(n-1)} f_n + w^{-k} \sum_{i=1}^{n-1} w^{ki} f_i \\ &= w^{k(n-1)} f_n - w^{-k} f_0 + w^{-k} \sum_{i=0}^{n-1} w^{ki} f_i = w^{k(n-1)} f_n - w^{-k} [f_0 + a_k]. \end{aligned}$$

Damit lässt sich die zweite diskrete Fourier-Transformation (und alle weiteren) unter Berücksichtigung der vorangehenden Transformatierten mit höchstens $2n$ Multiplikationen berechnen. Dieser Ansatz funktioniert aber nur, solange keine Fensterfunktion eingesetzt wird, die die Abtastwerte in Abhängigkeit ihrer Position verändern würde. ■

Die Ursache des globalen Verhaltens der Fourier-Transformation liegt darin, dass die verwendeten Sinus- und Kosinus-Funktionen nicht nur lokal Werte ungleich null annehmen. Während im Beispiel das geänderte Spektrum effizient berechnet werden kann, müssen bei anderen Änderungen alle Werte neu berechnet werden.

Bei der **Wavelet-Transformation** werden die Sinus- und Kosinus-Funktionen durch Funktionen ersetzt, die mit wachsenden Frequenzen auf immer kleineren Intervallen von null verschieden sind. Bei einer lokalen Änderung ändern sich dann nur die Koeffizienten bzw. Werte der Transformatierten, die einen lokalen Bezug haben, die anderen bleiben unverändert. Wie Fourier-Transformationen werden Wavelet-Transformationen z. B. zur verlustbehafteten Datenkompression eingesetzt, u. a. bei JPEG 2000.

6.5.1 Idee der Wavelet-Transformation

Eine gegebene Funktion wird zunächst auf eine feine Näherung abgebildet. 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 grösste 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.

Der Begriff **Multiskalenanalyse** drückt aus, dass bei der Wavelet-Transformation mit unterschiedlich genauen Auflösungen einer Funktion gearbeitet wird. Die grösste Skala erhält den Index 0, die feinste Skala, mit der die Berechnung begonnen wird, hat den grössten Index.

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

Wir beschreiben das Prinzip der Transformation zunächst anhand einer stückweise konstanten Funktion. In Abbildung 6.33 ist links diese Funktion (durchgezeichnet)

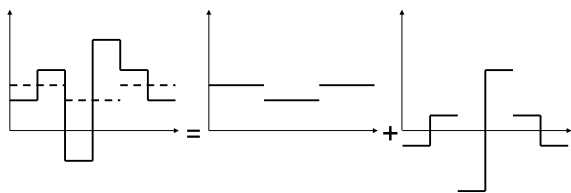


Abb. 6.33: Links: Approximation einer Funktion auf einer (feinen) Skala V_n mittels der Skalierungsfunktion. Mitte: Approximation der Funktion auf der nächst-gröberen Skala V_{n-1} . Rechts: Differenz zwischen beiden Darstellungen als Vielfache des Wavelets

dargestellt. Jetzt vergrößern wir die Darstellung, indem wir das arithmetische Mittel von Funktionswerten zu Paaren benachbarter konstanter Bereiche bilden. Die daraus entstehende gröbere Funktion ist in der Mitte dargestellt. Von dieser gröberen Funktion gelangen wir wieder zur feineren Startfunktion, indem wir die rechts dargestellten Funktionen („kleine Wellen“) addieren. Die mittlere Funktion kann nun durch Mittelwertbildung weiter vergrößert werden, wobei sich die Differenzen wieder mittels „kleinen Wellen“ beschreiben lassen. Die Wellen lassen sich alle als Faktor mal verschobener und skaliertes Standardfunktion, dem Wavelet, schreiben. Die Faktoren sind die bei der Transformation berechneten Werte.

Wir haben soeben die Wavelet-Transformation am Beispiel des (einfachen, aber leider unstetigen) Haar-Wavelets betrachtet. Um die gegebene Funktion zu beschreiben, benötigen wir eine **Skalierungsfunktion** Φ . Beim Haar-Wavelet ist

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

Durch Verschiebung (um k nach rechts) 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 oder deren konstante Abschnitte kleiner sind, näherungsweise als eine solche Linearkombination ausdrücken können, benötigen wir beliebig kleine konstante Stücke, die sich aneinander setzen lassen. Diese erhalten wir durch Skalierung von $\Phi_{0,k}$:

$$\Phi_{i,k}(x) := 2^{i/2} \Phi(2^i x - k) = \begin{cases} 2^{i/2} & \text{falls } \frac{k}{2^i} \leq x < \frac{k+1}{2^i}, \\ 0 & \text{sonst.} \end{cases}$$

In der Tat kann man sich über Linearkombinationen dieser 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 verkleinern die Länge der Intervalle, auf denen die Funktionen einen Beitrag liefern, auf 2^{-i} . Die Werte von i werden den Frequenzen der Sinus- und Kosinus-Terme bei Fourier-Reihen entsprechen. Mit wachsenden Frequenzen werden hier entsprechend die Periodenintervalle kleiner. Neu im Vergleich zur Fourier-Analyse ist die durch $-k$ verursachte Translation, die den lokalen Bezug der Funktionen herstellt. Der Vorfaktor $2^{i/2}$ dient lediglich der Normierung. Auf ihn gehen wir später ein, er ist zunächst nicht wichtig.

Den (unabhängig von den Vorfaktoren $2^{i/2}$) 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. 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, siehe Abbildung 6.33). Der Ansatz wird fortgesetzt, bis man eine Funktion aus V_0 plus n Differenzen erhält.

Wir können nun wie in der Abbildung 6.33 vorgehen. Dazu sei mit $d_{n,k} := a_k$

$$g(x) := \sum_{k=0}^{2^n-1} d_{n,k} \Phi(2^n x - k) = \sum_{k=0}^{2^n-1} a_k \Phi(2^n x - k) = 2^{-n/2} \sum_{k=0}^{2^n-1} a_k \Phi_{n,k}(x) \in V_n$$

eine Näherung an eine Funktion f auf einem Intervall $[0, 1]$, die aus 2^n stückweise konstanten Abschnitten besteht. Wir können nun $g(x)$ umschreiben in

$$g(x) = \sum_{k=0}^{2^{n-1}-1} d_{n-1,k} \Phi(2^{n-1} x - k) + h_{n-1,k} \Psi(2^{n-1} x - k), \quad (6.79)$$

wobei mit den Funktionen $\Phi(2^{n-1} x - k)$ die größeren konstanten Bereiche beschrieben werden und die Differenzen mit dem **Haar-Wavelet**

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

ausgedrückt werden. Die Koeffizienten $d_{n-1,k}$ entstehen dabei durch Mittelwertbildung:

$$d_{n-1,k} = \frac{d_{n,2k} + d_{n,2k+1}}{2}. \quad (6.80)$$

Die Koeffizienten der verschobenen und skalierten Wavelets sind

$$h_{n-1,k} = \frac{d_{n,2k} - d_{n,2k+1}}{2}. \quad (6.81)$$

Wir beobachten, dass $d_{n-1,k} + h_{n-1,k} = d_{n,2k}$ und $d_{n-1,k} - h_{n-1,k} = d_{n,2k+1}$ ist. Aufgrund der Definition von Ψ wird daher tatsächlich g in (6.79) berechnet, Ψ liefert genau die vor $h_{n-1,k}$ stehenden Vorzeichen $+$ und $-$. Wenn wir als nächstes $\sum_{k=0}^{2^{n-1}-1} d_{n-1,k} \Phi(2^{n-1}x - k) \in V_{n-1}$ wieder über eine gröbere Darstellung und Differenzen schreiben und das solange fortsetzen, bis wir eine konstante Funktion erhalten, dann haben wir fast die Haar-Wavelet-Transformation durchgeführt und diese Darstellung von g gewonnen:

$$g(x) = d_{0,0} + \sum_{l=0}^{n-1} \sum_{k=0}^{2^l-1} h_{l,k} \Psi(2^l x - k).$$

Wir erhalten also den Koeffizientenvektor

$$(d_{0,0}, h_{0,0}, h_{1,0}, h_{1,1}, h_{2,0}, h_{2,1}, h_{2,2}, h_{3,0}, \dots, h_{n-1, 2^{n-1}-1})$$

aus den gegebenen Funktionswerten $\vec{a} = (a_0, a_1, \dots, a_{2^n-1})$ über den Algorithmus 2, der sukzessive den in Algorithmus 1 beschriebenen Transformationsschritt verwendet. In den Algorithmen verwenden wir die Schreibweise $\vec{a}[i : k]$, um den Vektor mit den $k - i + 1$ Komponenten $a_i, a_{i+1}, \dots, a_k$ zu bezeichnen. Die Algorithmen sind mit „**discrete** Wavelet-Transformation“ beschriftet, da sie mit einem Vektor von endlich vielen Funktionswerten und nicht mit einer Funktion mit überabzählbar vielen Funktionswerten arbeiten.

Die einzige kleine Erweiterung zur üblicher Weise verwendeten Haar-Wavelet-Transformation besteht noch in der Verwendung von zur Normierung dienenden Faktoren, die wir bislang nur in der Definition von $\Phi_{i,k}$ berücksichtigt haben und auf die man in praktischen Anwendungen aber auch verzichten kann. Die Faktoren resultieren daraus, dass man gerne mit einem Orthonormalsystem von Funktionen arbeitet, so dass sich die beste Annäherung an eine Funktion leicht über eine Orthogonalprojektion berechnen lässt. Das wird dann wichtiger, wenn man mit komplizierteren Funktionen Φ und Ψ arbeitet, die schönere Eigenschaften (z.B. Stetigkeit) als die Haar-Skalierungsfunktion und das Haar-Wavelet haben. Zwar betrachten wir weiterhin konkret die Haar-Funktionen, aber wir stellen die Wavelet-Theorie jetzt so dar, dass die Formeln auch für andere orthogonale Wavelets hergeleitet werden.

Algorithmus 1 Ein Schritt der vereinfacht normierten diskreten Wavelet-Transformation

```

procedure DWTSCHRITT( $\vec{a}$ )
   $n := \text{ANZAHLKOMONENTEN}(\vec{a})$  (muss gerade sein)
  for  $j = 0 : n/2 - 1$  do
     $b_j := \frac{a_{2j} + a_{2j+1}}{2}$ 
     $b_{j+n/2} := \frac{a_{2j} - a_{2j+1}}{2}$ 
  return  $\vec{b}$ 

```

6.5.2 Eindimensionale Wavelet-Transformation mit orthogonalen Wavelets

Wir sind zuvor mit einer Funktion $g \in V_n$ gestartet. Wenn eine Funktion $f \notin V_n$ gegeben ist, dann müssen wir sie zunächst möglichst gut durch eine Funktion aus V_n ersetzen. Die beste Approximation (im quadratischen Mittel) aus V_n an eine gegebene Funktion erhält man (wie Fourier-Partialsummen) über eine Orthogonalprojektion. Der Vorfaktor $2^{i/2}$ der $\Phi_{i,k}(x)$ im Falle des Haar-Wavelets, den wir erst einmal nicht erklärt haben, sorgt dafür, dass die Funktionen zu festem i bezüglich des (reellen) Skalarprodukts $f \bullet g := \int_{-\infty}^{\infty} f(x)g(x) dx$ (vgl. mit Beispiel (3.21) auf Seite 485) nicht nur paarweise orthogonal, sondern auch zu 1 normiert sind. Man erhält als beste Approximation aus V_n an f die Funktion (vgl. mit Satz 3.20 auf Seite 496 für endlich-dimensionales $U := V_n$)

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

Die für die Orthogonalprojektion zu berechnenden Integrale können näherungsweise mittels Quadraturverfahren aus Abtastwerten berechnet werden. So kann man abgesehen vom Haar-Wavelet bei anderen Wavelet-Skalierungsfunktionen Φ vorgehen, wenn die skalierten und verschobenen Funktionen $\Phi_{n,k}$ eine Orthonormalbasis von V_n bilden.

Algorithmus 2 Vereinfacht normierte eindimensionale diskrete Wavelet-Transformation

```

procedure DWT( $\vec{a}$ )
   $n := \text{ANZAHLKOMONENTEN}(\vec{a})$  (muss Zweierpotenz sein)
  if  $n \leq 1$  then
    return  $\vec{a}$ 
   $\vec{b} := \text{DWTSCHRITT}(\vec{a})$ 
   $\vec{c}[0 : n/2 - 1] := \text{DWT}(\vec{b}[0 : n/2 - 1])$ 
   $\vec{c}[n/2 : n - 1] := \vec{b}[n/2 : n - 1]$ 
  return  $\vec{c}$ 

```

Die Haar-Skalierungsfunktion $\Phi \in V_0$ lässt sich in der nächst feineren Stufe V_1 ausdrücken:

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

Allgemein benötigt man für eine Multiskalenanalyse eine Skalierungsfunktion Φ , die sich als Linearkombination der $\Phi_{1,k}$ schreiben lässt, so dass $V_0 \subset V_1$ ist. Diese Linearkombination heißt **Verfeinerungsgleichung** oder **Skalierungsgleichung** (two-scale relation):

$$\Phi(x) = \sum_{k \in \mathbb{Z}} \frac{c_k}{\sqrt{2}} \Phi_{1,k}(x) = \sum_{k \in \mathbb{Z}} c_k \Phi(2^1 x - k). \quad (6.82)$$

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 $k \in \mathbb{Z} \setminus \{0, 1\}$. Bei den ebenfalls orthogonalen und zusätzlich stetigen Daubechies-Skalierungsfunktionen $D_4, D_6, D_8, \dots$ sind $4, 6, 8, \dots$ Koeffizienten ungleich null, siehe z. B. (Louis, Maaß und Rieder, 1998, S. 169ff). Wir betrachten nur Verfeinerungsgleichungen mit endlichen Summen.

Die für die ersten beiden Skalen formulierte Verfeinerungsgleichung gilt auch für den Übergang zwischen höheren Skalen:

$$\Phi_{i,k}(x) = \sum_{l \in \mathbb{Z}} \frac{c_l}{\sqrt{2}} \Phi_{i+1,2k+l}(x). \quad (6.83)$$

Diese Gleichung im Rahmen des Haar-Wavelets lautet

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

Mit den Koeffizienten der Verfeinerungsgleichung lassen sich ganz allgemein aus Skalierungsfunktionen Wavelets konstruieren, mit denen die Differenzen zwischen den Skalen ausgedrückt werden können. Zu jeder Skalierungsfunktion, bei der die Funktionen $\Phi_{i,k}(x)$ für festes i ein Orthonormalsystem bilden, erhält man ein zugehöriges Wavelet über den Ansatz (Louis, Maaß und Rieder, 1998, S. 122)

$$\Psi(x) = \sum_{k \in \mathbb{Z}} (-1)^k c_{1-k} \Phi(2x - k), \quad (6.84)$$

wobei $(c_k)_{k \in \mathbb{Z}}$ die Folge der Koeffizienten der Verfeinerungsgleichung (6.82) ist. Tatsächlich ist so auch das Haar-Wavelet ($c_0 = c_1 = 1$, $c_k = 0$ für $k \in \mathbb{Z} \setminus \{0, 1\}$) gebildet: Da für das Haar-Wavelet

$$\Phi(x) + \Psi(x) = 2\Phi(2x) \quad (6.85)$$

$$\Phi(x) - \Psi(x) = 2\Phi(2x - 1) \quad (6.86)$$

gilt, erhalten wir in diesem Fall

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

Abweichend von dieser Darstellung (6.84) 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 entgegengesetztes Vorzeichen hat. Das führt dann lediglich zu einem anderen Vorzeichen der über die Wavelet-Transformation berechneten Koeffizienten.

Wie zuvor für Φ verschieben und skalieren wir Ψ und schreiben normiert

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

Aus (6.84) erhalten wir für höhere Skalen

$$\begin{aligned} \Psi_{i,k}(x) &= 2^{i/2} \Psi(2^i x - k) = 2^{i/2} \sum_{l \in \mathbb{Z}} (-1)^l c_{1-l} \Phi(2[2^i x - k] - l) \\ &= 2^{i/2} \sum_{l \in \mathbb{Z}} (-1)^l c_{1-l} \Phi(2^{i+1} x - [2k + l]) = 2^{i/2} \sum_{l \in \mathbb{Z}} (-1)^l c_{1-l} \frac{1}{2^{\frac{i+1}{2}}} \Phi_{i+1, 2k+l}(x) \\ &= \sum_{l \in \mathbb{Z}} (-1)^l \frac{c_{1-l}}{\sqrt{2}} \Phi_{i+1, 2k+l}(x). \end{aligned} \quad (6.87)$$

Die **schnelle Wavelet-Transformation** basiert auf den Gleichungen (6.83) und (6.87). 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

$$\begin{aligned} d_{n,k} &\approx f(k2^{-n}) \int_{-\infty}^{\infty} \Phi_{n,k}(x) dx = f(k2^{-n}) \int_{-\infty}^{\infty} 2^{n/2} \Phi(2^n x - k) dx \\ &= f(k2^{-n}) 2^{-n/2} \int_{-\infty}^{\infty} \Phi(u) du = 2^{-n/2} f(k2^{-n}). \end{aligned}$$

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 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 (6.83):

$$\begin{aligned} d_{i,k} &= g_{i+1} \bullet \Phi_{i,k} = \sum_{m \in \mathbb{Z}} d_{i+1,m} \Phi_{i+1,m} \bullet \Phi_{i,k} \\ &\stackrel{(6.83)}{=} \sum_{m \in \mathbb{Z}} d_{i+1,m} \Phi_{i+1,m} \bullet \left(\sum_{l \in \mathbb{Z}} \frac{c_l}{\sqrt{2}} \Phi_{i+1,2k+l} \right) \\ &= \sum_{m \in \mathbb{Z}} \sum_{l \in \mathbb{Z}} d_{i+1,m} \frac{c_l}{\sqrt{2}} \Phi_{i+1,m} \bullet \Phi_{i+1,2k+l} = \sum_{l \in \mathbb{Z}} d_{i+1,2k+l} \frac{c_l}{\sqrt{2}}. \end{aligned}$$

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 (6.87):

$$\begin{aligned} h_{i,k} &= g_{i+1} \bullet \Psi_{i,k} = \sum_{m \in \mathbb{Z}} d_{i+1,m} \Phi_{i+1,m} \bullet \Psi_{i,k} \\ &\stackrel{(6.87)}{=} \sum_{m \in \mathbb{Z}} d_{i+1,m} \Phi_{i+1,m} \bullet \left(\sum_{l \in \mathbb{Z}} (-1)^l \frac{c_{1-l}}{\sqrt{2}} \Phi_{i+1,2k+l} \right) = \sum_{l \in \mathbb{Z}} d_{i+1,2k+l} (-1)^l \frac{c_{1-l}}{\sqrt{2}}. \end{aligned}$$

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, $0 \leq i < n$, $k \in \mathbb{Z}$. 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 halb so viele Werte $d_{i,k}$ der Stufe i berechnet. Aus m Werten der Stufe $i = n$ werden damit $m \left(1 + \frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^{n-1}}\right) \leq 2m$ Werte (vgl. (2.11) auf Seite 237). 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 Eingangsfunktion 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.

Speziell für das Haar-Wavelet sieht die schnelle Wavelet-Transformation so aus ($0 \leq i \leq n-1$):

$$\begin{aligned} d_{n,k} &\approx 2^{-n/2} f(k2^{-n}), & d_{i,k} &= \frac{d_{i+1,2k} + d_{i+1,2k+1}}{\sqrt{2}}, \\ h_{i,k} &= \frac{d_{i+1,2k} - d_{i+1,2k+1}}{\sqrt{2}}. \end{aligned}$$

Im Vergleich mit (6.80) und (6.81) sowie dem Algorithmus 1 fällt auf, dass durch $\sqrt{2}$ und nicht durch 2 dividiert wird. Das liegt ausschließlich an der Normierung für das Orthonormalsystem. Die Algorithmen 1 und 2 werden also zur üblichen diskreten Wavelet-Transformation, wenn die Komponenten von $\vec{a}$ vorab mit $2^{-n/2}$ multipliziert werden und dann in Algorithmus 1 statt durch 2 durch $\sqrt{2}$ dividiert wird.

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 Eingangsfunktion auch nur auf die entsprechenden lokalen Koeffizienten aus.

6.5.3 Zweidimensionale diskrete Wavelet-Transformation

Wir haben zuvor einen Vektor von Funktionswerten transformiert. In der Bildverarbeitung hat man es dagegen oft mit Matrizen von Grauwerten zu tun. Die in Algorithmus 3 beschriebene Standard-Wavelet-Transformation für Matrizen nutzt die eindimensionale diskrete Wavelet-Transformation DWT, indem zuerst alle Zeilen damit transformiert

werden. Die Ergebniskoeffizienten bilden eine neue Matrix, bei der nun die Spalten transformiert werden. Dadurch entsteht die Ergebnismatrix. Dieses Vorgehen lässt sich auch mittels Skalierungs- und Wavelet-Funktionen mit zwei Variablen erklären. Diese Funktionen sind zerfallend und entstehen als Produkt der Funktionen mit einer Variable und ihren Skalierungen.

Algorithmus 3 Diskrete zweidimensionale Standard-Wavelet-Transformation

```

procedure STANDARDWDT(A)
   $n \times n := \text{MATRIXFORMAT}(\mathbf{A})$ 
  if  $n \leq 1$  then
    return A
  for  $j = 0 : n - 1$  do
     $\mathbf{B}[j, 0 : n - 1] := \text{DWT}(\mathbf{A}[j, 0 : n - 1])$ 
  for  $j = 0 : n - 1$  do
     $\mathbf{C}[0 : n - 1, j] := \text{DWT}(\mathbf{B}[0 : n - 1, j])$ 
  return C

```

Bei der Standard-Transformation in Algorithmus 4 werden auch die bereits berechneten Wavelet-Faktoren durch die Spaltentransformationen weiter zerlegt. Das ist bei der Realisierung von Filtern, bei denen Koeffizienten klassifiziert oder weggelassen werden, aber gar nicht unbedingt erforderlich. Daher wird häufig in der Bildverarbeitung eine vereinfachte Transformation vorgenommen, bei der zunächst ein Schritt der Transformation auf alle Zeilen und dann auf alle Spalten angewendet wird. Danach geht es aber nur mit dem linken oberen Viertel der Matrix in dieser Form weiter, siehe Abbildung 6.34. Alle anderen Komponenten der Matrix werden nicht weiter modifiziert. Auf diese Weise benötigt man etwas weniger Rechenoperationen.

Algorithmus 4 Diskrete zweidimensionale Nichtstandard-Wavelet-Transformation

```

procedure NICHTSTANDARDWDT(A)
   $n \times n := \text{MATRIXFORMAT}(\mathbf{A})$ 
  if  $n \leq 1$  then
    return A
  for  $j = 0 : n - 1$  do
     $\mathbf{B}[j, 0 : n - 1] := \text{DWT}(\mathbf{A}[j, 0 : n - 1])$ 
  for  $j = 0 : n - 1$  do
     $\mathbf{C}[0 : n - 1, j] := \text{DWT}(\mathbf{B}[0 : n - 1, j])$ 
   $\mathbf{D}[0 : n/2 - 1, 0 : n/2 - 1] := \text{NICHTSTANDARDWDT}(\mathbf{C}[0 : n/2 - 1, 0 : n/2 - 1])$ 
   $\mathbf{D}[n/2 : n - 1, 0 : n - 1] := \mathbf{C}[n/2 : n - 1, 0 : n - 1]$ 
   $\mathbf{D}[0 : n/2 - 1, n/2 : n - 1] := \mathbf{C}[0 : n/2 - 1, n/2 : n - 1]$ 
  return D

```

In Goebbels und Pohle-Fröhlich (2015) wird beispielsweise die diskrete zweidimensionale Nichtstandard-Wavelet-Transformation genutzt, um damit Laserscan-Höhendaten der Stadt Krefeld stark zu komprimieren. Dabei werden vergleichsweise kleine Wavelet-Koeffizienten zu feinen Skalen (d. h. zu hohen Frequenzen) weggelassen (Tiefpass-Filterung), falls aufgrund der lokalen Bodennutzung Details für die Visualisierung unwichtig sind.

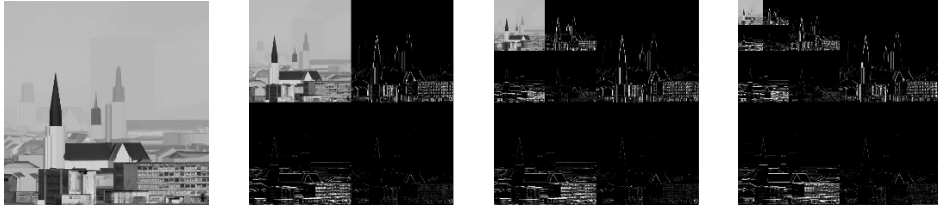


Abb. 6.34: Erste Schritte einer Nichtstandard-Wavelet-Transformation eines Grauwertbildes: Der Betrag der Wavelet-Koeffizienten wird ebenfalls als Grauwert dargestellt, null ist schwarz. Da die Koeffizienten klein sind, wurden sie zur besseren Sichtbarkeit mit zehn multipliziert.

6.6 Aufgaben

Aufgabe 6.1

Man entwickle die Funktionen $\sin t$ und $\cos(3t)$ als Fourier-Reihe und gebe diese sowohl in der Gestalt $a_0 + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt))$ als auch in der Darstellung $\sum_{k=-\infty}^{\infty} c_k \exp(jkt)$ an.

Aufgabe 6.2

Sei $f(t) = |\sin t|$.

- Man berechne die Fourier-Reihe $a_0 + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt))$ von f .
- Für welche $t \in \mathbb{R}$ konvergiert die Fourier-Reihe (als Zahlenreihe) gegen $f(t)$?

Aufgabe 6.3

Man berechne die Fourier-Transformierte der Funktion $f(t) := \begin{cases} e^{-t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$

Aufgabe 6.4

Man berechne mittels Integration die Fourier-Transformierte der Impulsfunktion

$$f(t) := \begin{cases} \frac{1}{T}, & 0 \leq t \leq T, \\ 0, & \text{sonst,} \end{cases} \quad \text{für } T > 0,$$

die unabhängig von T eine Fläche der Größe 1 einschließt: $\frac{1}{T} \int_0^T 1 \, dt = 1$. Was geschieht mit der Transformierten, wenn man $T \rightarrow 0+$ gehen lässt?

Aufgabe 6.5

Man berechne die Fourier-Transformierte von $g(t) := 3e^{-j4t}f(t) + 4f(2t)$, wobei

$$f(t) := \begin{cases} 1, & -1 \leq t \leq 1, \\ 0, & \text{sonst,} \end{cases} \quad \text{und } f^\wedge(\omega) = 2 \operatorname{sinc}(\omega) = 2 \cdot \begin{cases} \frac{\sin \omega}{\omega} & \omega \neq 0, \\ 1 & \omega = 0. \end{cases}$$

Aufgabe 6.6

Man berechne mittels der Rechenregeln für die Fourier-Transformation die aus Aufgabe 6.4 bekannte Transformierte der Impulsfunktion f für $T > 0$.

Aufgabe 6.7

Man berechne **a)** $[\mathcal{L}(3 \sin(5t) + 2 \cos t)](s)$ und **b)** $[\mathcal{L}((3t)^2)](s)$.

Aufgabe 6.8

Man zeige mittels der Regel $[\mathcal{L}(f')](s) = s[\mathcal{L}f](s) - f(0)$, dass $[\mathcal{L}e^t](s) = \frac{1}{s-1}$.

Aufgabe 6.9

Man löse das folgende Anfangswertproblem mittels Laplace-Transformation:

$$y'(t) + 3y(t) = t, \quad y(0) = 1.$$

Dabei verwende man $[\mathcal{L}(\frac{t^n}{n!} \exp(at))](s) = \frac{1}{(s-a)^{n+1}}, \quad s > a, \quad n \in \mathbb{N}_0$.

Hinweis: Man benötigt die Partialbruchzerlegung von $\frac{1}{(s+3)s^2}$.

Aufgabe 6.10

Man löse das folgende Anfangswertproblem mittels Laplace-Transformation:

$$y''(t) + 2y'(t) + y(t) = 0, \quad y(0) = -3, \quad y'(0) = 11.$$

Hinweis: Die Laplace-Transformierte von $\frac{t^n}{n!} \exp(at)$ ist $\frac{1}{(s-a)^{n+1}}$.

Aufgabe 6.11

Man verifiziere unter Verwendung der Faltungsregel, dass die Funktion, deren Laplace-Transformierte

$$F(s) = \frac{s}{s^2 + 1} \cdot \frac{1}{s^2 + 1}$$

ist, $f(t) = \frac{1}{2}t \sin t$ lautet. Dabei hilft $\cos(t-u) = \cos(t)\cos(u) + \sin(t)\sin(u)$.

Aufgabe 6.12

Man beweise den Anfangswertsatz (Satz 6.8): Ist f stetig auf $[0, \infty[$ und von höchstens exponentiellem Wachstum, dann gilt: $f(0) = \lim_{s \rightarrow \infty} sF(s)$.

Hinweis: Der Beweis kann analog zum Beweis des Endwertsatzes (Satz 6.9) aufgebaut werden. Man zerlege das Laplace-Integral in zwei Integrale und zeige, dass s -mal das erste Integral wegen der Stetigkeit von f an der Stelle 0 für $s \rightarrow \infty$ gegen $f(0)$ strebt. Anschließend zeige man, dass s -mal das zweite Integral für $s \rightarrow \infty$ gegen 0 konvergiert. Dazu kann man ausnutzen, dass f von höchstens exponentiellem Wachstum ist.

Aufgabe 6.13

Bestimmen Sie alle primitiven dritten Wurzeln der 1.

Aufgabe 6.14

- a) Bestimmen Sie die Matrix für die diskrete Fourier-Transformation in $\mathbb{C}^3$ zur primitiven Wurzel $w := e^{j\frac{2\pi}{3}}$.
- b) Wie sieht die Matrix der zugehörigen Umkehrtransformation aus?
- c) Berechnen Sie $\text{DFT}_w((1, 0, 1))$.

Aufgabe 6.15

- a) Bestimmen Sie alle primitiven (komplexen) fünften Wurzeln der Eins.
- b) Bestimmen Sie die Matrix für die diskrete Fourier-Transformation in $\mathbb{C}^5$ zur primitiven Wurzel $w := e^{j\frac{2\pi}{5}}$. Berechnen Sie $\text{DFT}_w((1, 0, 1, 0, 0))$.
- c) Geben Sie die Matrix der zugehörigen Umkehrtransformation so an, dass keine negativen Exponenten innerhalb der Matrix auftreten.

Aufgabe 6.16

Gegeben sei die Funktion $f(t) := 17 + 23 \cos(5t) + \sin(8t) + 11 \sin(3t) + \cos(2t)$.

- a) Wie viele äquidistant berechnete Funktionswerte benötigt man, um alle Fourier-Koeffizienten von f mittels diskreter Fourier-Transformation exakt zu bestimmen?
- b) Was geschieht, wenn mehr oder wenn weniger Funktionswerte verwendet werden?

Aufgabe 6.17

Was bedeutet „Teile und herrsche!“ in Bezug auf die schnelle Fourier-Transformation?

Aufgabe 6.18

- a) Man falte die Vektoren $\vec{y} := (3, 2)$ und $\vec{z} := (1, -1)$.
- b) Verifizieren Sie den Faltungssatz anhand der diskreten Fourier-Transformation

$$\text{DFT}_{-1} \text{ in } \mathbb{C}^2 \text{ zur Fourier-Matrix } F := \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Aufgabe 6.19

- a) Man falte die Vektoren $\vec{y} := (4, 3, 2, 1)$ und $\vec{z} := (5, 6, 7, 8)$.
- b) Verifizieren Sie den Faltungssatz anhand der diskreten Fourier-Transformation DFT_j , die über die folgende Fourier-Matrix gegeben ist:

$$F := \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \end{bmatrix}.$$

- c) Man invertiere die Fourier-Matrix F aus Teil b).

Aufgabe 6.20

Beantworten Sie die Fragen mit eigenen Worten:

- Was ist die diskrete Fourier-Transformation?
- Wie sehen Interpolationspolynom vom Grad n und Fourier-Reihe eines trigonometrischen Polynoms $p(x) = \sum_{k=-m}^m d_k e^{jkx}$ aus, wenn $m \leq n$ ist?
- Was ist der Unterschied zwischen einem trigonometrischen Interpolationspolynom vom Grad n und der mittels der diskreten Fourier-Transformation im $\mathbb{C}^{2n+1}$ näherungsweise berechneten n -ten Partialsumme einer Fourier-Reihe einer 2π -periodischen, stetigen Funktion f ?
- Wie kann die diskrete Fourier-Transformation genutzt werden, um
 - ein Interpolationspolynom vom Grad n zu berechnen,
 - Fourier-Koeffizienten $f^\wedge(-n), \dots, f^\wedge(n)$ zu berechnen?

Aufgabe 6.21

Die stetige, Fourier-transformierbare Funktion $f: \mathbb{R} \rightarrow \mathbb{C}$ mit $\int_{-\infty}^{\infty} |f(t)| dt < \infty$ erfülle die Bedingung, dass $f^\wedge$ außerhalb von $[-1, 1]$ null ist.

- Wie ist $\Delta t > 0$ zu wählen, damit man die Fourier-Transformierte exakt über die Reihe

$$f^\wedge(\omega) = \Delta t \sum_{k=-\infty}^{\infty} f(k\Delta t) \exp(-j\omega k\Delta t), \quad |\omega| \leq \Omega,$$

erhält, die absolut konvergieren möge?

- Um $f^\wedge$ mittels der diskreten Fourier-Transformation zu berechnen, werten wir f auf einem Intervall $[-R, R]$ mit einer Schrittweite R/n aus. Wie groß muss n gewählt werden, damit nur der Fehler des Abschneidens auf $[-R, R]$ und nicht zusätzlich der Fehler einer zu großen Schrittweite auftritt?
- Was passiert, wenn man n größer oder kleiner wählt?

Aufgabe 6.22

Da man eine auf ganz $\mathbb{R}$ definierte Funktion nur auf einem endlichen Intervall $[-R, R]$ abtasten kann, kann man praktisch nicht $f^\wedge$ berechnen, sondern man berechnet die Fourier-Transformierte des Produkts von f mit einer Fensterfunktion, die außerhalb des Intervalls $[-R, R]$ gleich null ist. Gängige Fensterfunktionen sind ($0 \leq a \leq 1$):

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

Ein Rechteckfenster liegt bei $a = 1$ vor, das Hann-Fenster bei $a = \frac{1}{2}$ und bei $a = 0,54$ das Hamming-Fenster. Berechnen Sie die Fourier-Transformierte $g^\wedge(\omega)$.



7 Wahrscheinlichkeitsrechnung und Statistik

Übersicht

7.1	Beschreibende Statistik	970
7.2	Wahrscheinlichkeitsrechnung	997
7.3	Schließende Statistik	1056
7.4	Aufgaben	1076

„Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.“ – Aaron Levenstein

Der Begriff „Statistik“ ist aus einer Vorlesung mit der Bezeichnung „collegium politico-statisticum“ von Martin Schmeitzel (1679–1747) entstanden. siehe (Menges, 1982, S. 4). Hier ging es um Staatenkunde, eine Disziplin, die sich mit (ziemlich vagen) Informationen über Staaten beschäftigte – ein entfernter Vorläufer der heutigen Statistik. Heute werden mittels Statistik große und unübersichtliche Datenmengen übersichtlich aufbereitet. Mit der Darstellung der Daten und der Berechnung von aussagekräftigen Kenngrößen beschäftigt sich die **beschreibende Statistik** (deskriptive Statistik, siehe Kapitel 7.1).

Beispiel 7.1

Wir sind an der mittleren Studienzeit (in Semestern) aller Absolventen der Hochschule Niederrhein im Jahr 2007 interessiert. Dazu muss die Studienzeit aller Absolventen von 2007 vorliegen. (Neben dem bekannteren arithmetischen Mittel ist hier der Median aussagekräftiger, den wir später kennenlernen). ■

Oft kann man sich nicht die Gesamtheit der Daten ansehen, sondern ist auf einen kleinen Ausschnitt (eine Stichprobe) angewiesen. Dies gilt insbesondere, wenn man Aussagen über die Zukunft machen möchte. Hier liegen nur Daten der Vergangenheit vor, die Daten der Zukunft sind natürlich nicht verfügbar. Die **schließende Statistik** versucht, aus Stichproben Aussagen für die Gesamtheit zu schließen (siehe Kapitel

7.3). Da man nur einen Teil der Informationen zur Verfügung hat, muss man hier mit Wahrscheinlichkeiten operieren (siehe Kapitel 7.2).

Beispiel 7.2

In einem Softwareprojekt wird die Anzahl der nach einer gewissen Zeit gefundenen Fehler pro Klasse als Qualitätsmaß eingesetzt. Zur Klasse A mit 1 000 Programmzeilen wurden 10 Fehler gefunden, zur Klasse B mit 2 000 Programmzeilen wurden 18 Fehler gefunden, d. h., Klasse A hat $\frac{10}{1\,000} = 0,01$ und Klasse B hat $\frac{18}{2\,000} = 0,009$ Fehler pro Programmzeile. A und B wurden durch unterschiedliche Teams erstellt. Kann man daraus schließen, dass Team A schlechter arbeitet als Team B? ■

Die schließende Statistik ermöglicht es, zu prüfen, ob man dieser Hypothese zustimmen sollte oder nicht. Eine Antwort finden Sie am Ende dieses Kapitels auf Seite 1073.

7.1 Beschreibende Statistik

In diesem Kapitel beschäftigen wir uns mit den wichtigsten Begriffen zur Aufbereitung und Darstellung großer Datenmengen, wobei wir uns in der Darstellung an Sachs (2003) orientieren. Das Buch eignet sich damit sehr gut zur Vertiefung und Ergänzung (z. B. hinsichtlich der hier nicht behandelten Begriffe für klassifizierte Merkmale).

7.1.1 Grundbegriffe

7.1.1.1 Modellbildung und Häufigkeit

Um Mathematik betreiben zu können, müssen wir zunächst den Untersuchungsgegenstand mit einem Modell beschreiben. Dazu führen wir halbformal einige Begriffe ein:

■ Wer wird untersucht?

Die **statistische Masse** (**Grundgesamtheit**, **Population**) ist die Menge von Objekten, deren Eigenschaften untersucht werden sollen. Die Elemente der statistischen Masse heißen **Merkmalsträger** (**statistische Elemente**, **statistische Einheiten**). Wir werden ausschließlich endliche statistische Massen betrachten. Dies ist sinnvoll, da wir die Eigenschaften der Merkmalsträger tatsächlich erfassen müssen (z. B. durch Messen, Befragung etc.).

Bei der Untersuchung der Studienzeiten bilden beispielsweise die Absolventen der Hochschule Niederrhein in 2007 die statistische Masse. Weitere Beispiele statistischer Massen sind

- Personen und Sachen sowie
- Ereignisse (z. B. Fehler, Geburten, Todesfälle).

■ Was wird untersucht?

Merkmalsträger besitzen (gemeinsame) **Merkmale**, die gewisse **Merkmalsausprägungen (Merkmalswerte)** annehmen können. So hat jeder Absolvent das Merkmal Studiendauer, das Merkmalsausprägungen aus $\mathbb{N}$ (mit der Einheit Semester) annehmen kann.

Wir sehen uns hauptsächlich die vorgefundenen Merkmalsausprägungen und ihre Häufigkeit an. Die beschreibende Statistik beschäftigt sich mit der Aufbereitung der Merkmalsausprägungen aller Elemente einer statistischen Masse. Man spricht hier von einer **Voll- oder Totalerhebung**. Arbeitet man dagegen nur mit einer echten Teilmenge der statistischen Masse, spricht man von einer **Teilerhebung** oder **Stichprobe**. Eine besondere Rolle spielen **repräsentative Stichproben**. Darin weisen die Merkmalsausprägungen gewisser Merkmale die gleiche relative Häufigkeit auf wie in der statistischen Masse. Die schließende Statistik beschäftigt sich mit Rückschlüssen von Stichproben auf die Grundgesamtheit. Überall dort, wo wir in der beschreibenden Statistik die statistische Masse betrachten, kann man diese durch eine Stichprobe austauschen.

Man unterscheidet zwischen verschiedenen Typen von Merkmalsausprägungen, wobei sich die Typen durchaus überlappen können:

- **Qualitative Merkmale:** Die Merkmalsausprägungen gehören einer endlichen Aufzählung an wie z. B. „grün“, „rot“, „blau“. In der Regel werden hier keine Zahlen verwendet.
- **Quantitative Merkmale:** Die Merkmalsausprägungen sind reelle Zahlen (ggf. ergänzt durch physikalische Einheiten). Hier unterscheidet man zwischen **diskreten** und **stetigen** Merkmalen. Bei diskreten Merkmalen gibt es „relativ wenige“ verschiedene Merkmalsausprägungen, so dass gleiche Ausprägungen i. Allg. für mehrere Merkmalsträger angenommen werden, insbesondere wenn die statistische Masse groß ist – wie z. B. die Studiendauer in Semestern. Hier ist es sinnvoll, zu zählen, wie oft eine Merkmalsausprägung angenommen wird. Dies gilt auch für qualitative Merkmale. Bei stetigen Merkmalen gibt es zu nahezu jedem Merkmalsträger eine eigene, von den anderen verschiedene Merkmalsausprägung, und zwar auch dann, wenn die statistische Masse sehr groß ist. Hier ist es daher nicht sinnvoll, das Auftreten einer einzelnen Merkmalsausprägung zu zählen. Betrachtet man beispielsweise das Gewicht von Personen in Gramm, so wird eine Merkmalsausprägung nur selten für mehrere Merkmalsträger (Personen) angenommen.

Ein Merkmal X kann als Abbildung (Funktion) verstanden werden, die einem Merkmalsträger i die zugehörige Merkmalsausprägung x_i zuordnet.

Achtung: Bei einem stetigen Merkmal kann man keineswegs auf eine irgendwie definierte Stetigkeit dieser Funktion schließen. Die Stetigkeit einer Funktion ist ein völlig anderer Begriff!

Definition 7.1 (Notationen, Häufigkeit)

Gegeben seien $n \in \mathbb{N}$ statistische Elemente (z. B. Absolventen), die zur Vereinfachung von 1 bis n durchnummeriert sind. Weiter sei X ein Merkmal (z. B. Studiendauer). Dann bezeichnen wir mit $x_i = X(i)$, $1 \leq i \leq n$, die Merkmalsausprägung des Merkmals X für das statistische Element i . Weiter sei $m \in \mathbb{N}$ ($m \leq n$) die Anzahl der **unterschiedlichen** Merkmalsausprägungen, die X auf der statistischen Masse annimmt. Wir nummerieren die unterschiedlichen Merkmalsausprägungen durch und bezeichnen sie mit a_k , $1 \leq k \leq m$. Zu jedem a_k interessiert uns, wie viele statistischen Elemente diese Merkmalsausprägung besitzen. Die absolute Häufigkeit $h_k > 0$ und die relative Häufigkeit $f_k > 0$ sind definiert über

$$\begin{aligned} h_k &:= \text{Anzahl der Merkmalsausprägungen } x_i, \text{ die gleich } a_k \text{ sind,} \\ f_k &:= \frac{h_k}{n}. \end{aligned}$$

Es gilt:

$$\sum_{k=1}^m h_k = n, \quad \sum_{k=1}^m f_k = \frac{1}{n} \sum_{k=1}^m h_k = 1.$$

Relative Häufigkeiten sind der Schlüssel zum Verständnis von Wahrscheinlichkeiten im nächsten Kapitel. Etwas unpräzise formuliert sind Wahrscheinlichkeiten relative Häufigkeiten für größere Anzahlen n , also für $n \rightarrow \infty$.

Beispiel 7.3

Merkmalsträger bzw. statistische Elemente sind die Studenten 1 (Simon), 2 (Sandra), 3 (Ralf), 4 (Schluffi) und 5 (Eva). Wir betrachten das Merkmal X Studiendauer und erhalten dazu pro Merkmalsträger (Student) einen Wert:

Student i	1	2	3	4	5
Studiendauer $x_i = X(i)$	7	8	7	20	8

Damit haben wir $m = 3$ verschiedene Merkmalsausprägungen $a_1 = 7$, $a_2 = 8$ und $a_3 = 20$. Die zugehörigen absoluten und relativen Häufigkeiten sind

	$a_1 = 7$	$a_2 = 8$	$a_3 = 20$
absolute Häufigkeit h_k	2	2	1
relative Häufigkeit f_k	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

■

Beispiel 7.4 (Bildererkennung)

Viele Verfahren zur Bildererkennung 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. ■

7.1.1.2 Darstellungen der Häufigkeit

Mit gängigen Darstellungen von absoluten und relativen Häufigkeiten werden wir jeden Tag in den Medien konfrontiert. Beim **Stabdiagramm** oder **Bar Chart** wird die Häufigkeit durch die Länge einer Line (Bar) dargestellt. Dabei gibt es zwei Varianten (siehe Abbildung 7.1):

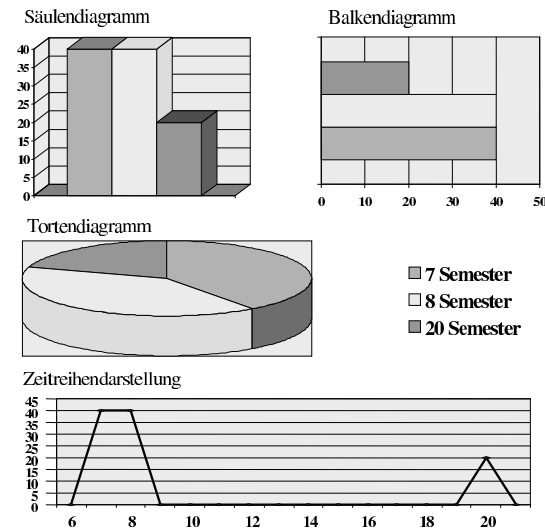


Abb. 7.1: Darstellung der Studiendauern in Prozent

- **Säulendiagramm:** Statt Linien werden Rechtecke gleicher Breite gezeichnet.
- **Balkendiagramm:** Dies ist ein Säulendiagramm mit vertauschten x - und y -Achsen.

Geht es um Anteile (also relative Häufigkeiten), so findet man oft **Tortendiagramme** vor. Die Häufigkeiten werden dabei über die Größe von Tortenstücken dargestellt. Als

Winkel des k -ten Tortenstücks wählt man $f_k \cdot 2\pi$, also den Anteil f_k von 360 Grad. Die zugehörige Fläche bei einem Kreis mit Radius r ist $f_k \cdot \pi r^2$.

Möchte man eine zeitliche Entwicklung von Häufigkeiten in den Vordergrund stellen, so verwendet man eine **Zeitreihendarstellung**. Hier verbindet man diskrete Häufigkeitspunkte mit Strecken, so dass der Graph einer stückweise linearen Funktion entsteht. Man interpoliert, obwohl keine Zwischenwerte vorliegen. Dadurch suggeriert man eine Tendenz.

In der Rheinischen Post vom 08.05.2010 ist auf Seite eins der Aufmacher eine Prognose zur Landtagswahl in Nordrhein-Westfalen. Als Zeitreihe werden Wahlergebnisse ab 2005 und zwei Prognosen vom 06.05.2010 (irrtümlich: 06.05.2009) chronologisch dargestellt. Die Abstände der Einzelwerte sind konstant und entsprechen nicht der Zeitspanne zwischen den jeweiligen Ergebnissen. Selbst die Prognosewerte mit gleichem Datum werden mit diesem Abstand dargestellt. Dadurch entsteht fälschlich der Eindruck einer Tendenz, was durchaus das Wahlergebnis beeinflussen konnte.

In die Irre führende Interpretationen sind ein großes Problem bei der Aufbereitung statistischer Daten, siehe Beck-Bornholdt et al. (2006):

- In einem Raum befinden sich 4 Eier und 4 Würstchen. Die relative Häufigkeit des Lebensmittels Eier beträgt 50 %. Dann wird ein Hund in den Raum gelassen. Er frisst 3 Würstchen. Nachdem der Hund im Raum war, beträgt die relative Häufigkeit der Eier 80 %. Man könnte nun schließen: Der Hund legt Eier.
- Wir möchten herausfinden, wie viele Züge bei der Bahn überfüllt sind und befragen „zufällig“ einen Tag lang Reisende am Krefelder Hauptbahnhof. 300 von 1 000 berichten von überfüllten Zügen. Sind dann 30 % aller Züge überfüllt? Nein, denn in einem überfüllten Zug sind mehr Fahrgäste, so dass wir auch entsprechend häufiger Reisende aus überfüllten Zügen treffen.
- Es werden Autounfälle bei Nebel und bei nebelfreiem Wetter gezählt. Die absolute Zahl der Unfälle bei Nebel ist viel geringer als die bei klarem Wetter. Kann man daraus schließen, dass bei Nebel vorsichtiger gefahren wird? Nein, denn Nebel ist sehr selten, so dass alleine deswegen weniger Unfälle bei Nebel geschehen.

Neben der fragwürdigen Interpretation korrekter Daten können aber auch bereits die Daten verfälscht sein. Bei Umfragen neigen Befragte dazu, mit Ja zu antworten. Durch entsprechende Formulierung der Fragen kann so das Ergebnis manipuliert werden. Auch ist nicht klar, wie ehrlich geantwortet wird. Die Westdeutsche Zeitung vom 10.12.2009 beschäftigt sich auf Seite 25 mit diesem Thema und nennt als Beispiel eine Statistik, nach der in Großbritannien mehr verheiratete Frauen als Männer leben.

Wie oben gesehen, kann auch die grafische Aufbereitung von Daten trügerisch sein:

- Es wird nur ein Ausschnitt der Daten gezeigt, der zu einer These passt.
- Koordinatenachsen beginnen nicht bei 0. Eine relativ kleine Veränderung kann dadurch riesig erscheinen. Diese Manipulation ist heute fast üblich.

- Es werden Flächen eingesetzt, deren Inhalt nicht proportional zur dargestellten Größe ist. Ein Beispiel dazu liefert wieder die Rheinische Post, die am 22.05.2010 mit den Wahrscheinlichkeiten für Koalitionen nach der Landtagswahl aufmacht. Diese sind jeweils als Kreisscheiben mit unterschiedlichen Durchmessern visualisiert:
 - Große Koalition aus SPD und CDU: 50 %, Durchmesser 10 cm
 - Neuwahlen: 35 %, Durchmesser 8 cm
 - Jamaika-Koalition aus CDU, Grünen und FDP: 10 %, Durchmesser 2 cm
 - Ampel-Koalition aus SPD, FDP und Grünen: 5 %, Durchmesser 1 cm.

Wir fragen uns lieber nicht, wie die Prozentangaben begründet sind (zumal es eine rot-grüne Minderheitsregierung geworden ist) und beschränken die Betrachtung auf die Darstellung der Prozentwerte. Bei der großen Koalition, der Ampel- und der Jamaika-Koalition sind die Prozente proportional zum Durchmesser. Das ist irreführend, da man die Flächen und nicht die Durchmesser vergleicht. Die Große Koalition wirkt also viel wahrscheinlicher, als sie ist. Der Kreis zu Neuwahlen müsste, führt man die Darstellung konsequent fort, einen Durchmesser von 7 cm haben, aber bei den verwendeten 8 cm sehen Neuwahlen wahrscheinlicher aus, und die Dramatik nimmt zu.

- Durch eine dreidimensionale Anordnung verzerrt die Perspektive die tatsächlichen Größenverhältnisse.

7.1.2 Empirische Verteilungsfunktionen

In diesem Abschnitt führen wir die empirische Verteilungsfunktion ein. Sie wird uns dabei helfen, wichtige Lageparameter und damit aussagekräftige Kenngrößen für unsere Daten zu bestimmen. Außerdem entspricht sie der Verteilungsfunktion in der Wahrscheinlichkeitsrechnung, die bei der Berechnung von Wahrscheinlichkeiten eine große Rolle spielt (siehe Kapitel 7.2.5.3).

Bei stetigen Merkmalsausprägungen ist in der Regel $h_k = 1$ und $f_k = \frac{1}{n}$, so dass diese Größen keine Aussagekraft besitzen. Statt mit einzelnen Merkmalsausprägungen zu arbeiten, rechnet man daher hier mit Klassen, mit denen Merkmalsausprägungen zusammengefasst werden. Insbesondere macht es nun Sinn, zu zählen, wie oft Merkmalsausprägungen einer Klasse angenommen werden. Dabei verliert man etwas von der Genauigkeit der einzelnen Merkmalsausprägungen, erhält dafür aber neue diskrete Merkmalsausprägungen. In der Literatur, siehe z. B. Sachs (2003), werden Begriffe für diskrete und stetige Merkmalsausprägungen etwas unterschiedlich definiert. Wir begnügen uns mit den Begriffen für diskrete Ausprägungen.

Im Folgenden betrachten wir durchgängig reellwertige Merkmalsausprägungen. Außerdem seien die verschiedenen Ausprägungen sortiert: $a_1 < a_2 < \dots < a_m$. Uns interessiert nun die relative Häufigkeit der Merkmalsausprägungen bis zu einer vorgegebenen Zahl. Damit bilden wir eine Funktion:

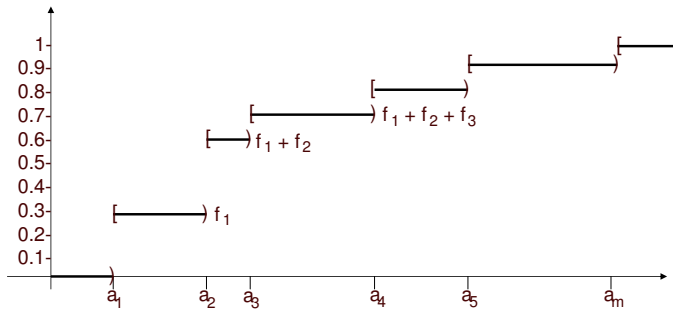


Abb. 7.2: Verteilungsfunktion eines Merkmals

Definition 7.2 (Empirische Verteilungsfunktion)

Die verschiedenen reellen Ausprägungen seien wie im restlichen Kapitel der Größe nach aufsteigend sortiert. Die **empirische Verteilungsfunktion** $F(x) : \mathbb{R} \rightarrow [0, 1]$ ist definiert über die Summe der relativen Häufigkeiten aller Ausprägungen, die kleiner oder gleich der Zahl x sind.

$$F(x) := \sum_{k \in \{1, \dots, n\} \text{ mit } a_k \leq x} f_k.$$

Dabei besagt $n \cdot F(x)$, wie viele Merkmalsausprägungen kleiner oder gleich x sind, und $F(x)$ gibt die relative Häufigkeit aller Merkmalsausprägungen kleiner oder gleich x an. Wir erhalten eine rechtsseitig stetige, monoton wachsende Treppenfunktion (siehe Abbildung 7.2).

Beispiel 7.5

Wir greifen das Beispiel der Studiendauern auf und erhalten für die verschiedenen Merkmalsausprägungen $a_1 = 7$, $a_2 = 8$ und $a_3 = 20$ die Verteilungsfunktion

$$F(x) = \begin{cases} 0, & x < 7 \\ \frac{2}{5}, & 7 \leq x < 8 \\ \frac{4}{5}, & 8 \leq x < 20 \\ 1, & x \geq 20. \end{cases}$$

Beachten Sie, dass **nicht** $F(9) = \frac{2}{5}$ gilt, obwohl die relative Häufigkeit aller Merkmalsausprägungen im Intervall $[8, 20[$ diesen Wert hat. Um $F(9)$ zu berechnen, muss man

die relative Häufigkeit aller Merkmalsausprägungen kleiner oder gleich 9 (also von 7 und 8) summieren: $F(9) = \frac{4}{5}$. ■

7.1.3 Lageparameter

Das Hauptziel der beschreibenden Statistik besteht darin, eine unübersichtliche Datenmenge durch wenige Kennzahlen (Lageparameter) zu charakterisieren. Ein ähnliches Vorgehen haben wir bei der Kurvendiskussion benutzt, um das Aussehen von Funktionsgraphen aus wenigen Parametern abzuleiten.

7.1.3.1 Arithmetisches Mittel

In der Bildverarbeitung benötigt man eine Information über die Helligkeit des Bildes. Dazu kann man die Helligkeit (den **Grauwert**) jedes einzelnen Bildpunktes bestimmen und anschließend daraus einen mittleren Grauwert über alle Bildpunkte berechnen. Dabei verwendet man das arithmetische Mittel:

Definition 7.3 (Arithmetisches Mittel)

Das **arithmetische Mittel** $\bar{x}$ eines Merkmals X wird berechnet, indem man die Summe aller (reellen) Merkmalsausprägungen x_i bildet und durch ihre Anzahl n dividiert:

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \\ &= \frac{a_1 h_1 + a_2 h_2 + \cdots + a_m h_m}{n} = \frac{1}{n} \sum_{k=1}^m a_k h_k = \sum_{k=1}^m a_k f_k.\end{aligned}\quad (7.1)$$

Achtung: In diesem Kapitel bezeichnet $\bar{x}$ nicht die konjugiert komplexe Zahl zu x , sondern das arithmetische Mittel.

Beispiel 7.6

Das arithmetische Mittel der Studiendauern ist $\frac{1}{5}(7 + 8 + 7 + 20 + 8) = 10$. ■

Wir können das arithmetische Mittel auch als Lösung der folgenden Optimierungsaufgabe verstehen: Gesucht ist ein Wert, für den die auf $\mathbb{R}$ erklärte Funktion

$$f(u) := \sum_{i=1}^n (x_i - u)^2 \quad (7.2)$$

ein Minimum annimmt, der also die Summe der quadrierten Differenzen zu den einzelnen Merkmalsausprägungen minimiert. Da f nicht-negativwertig und $\lim_{u \rightarrow \pm\infty} f(u) =$

∞ ist, muss die Funktion mindestens ein Minimum annehmen. Da sie differenzierbar ist, muss die Ableitung an der Stelle des Minimums null sein. Wir suchen also die Nullstellen von

$$f'(u) = -2 \sum_{i=1}^n (x_i - u) = 2 \left(nu - \sum_{i=1}^n x_i \right) = 2n(u - \bar{x}).$$

Nur $\bar{x}$ ist Nullstelle dieser linearen Funktion, und damit muss hier auch das Minimum angenommen werden.

Das arithmetische Mittel erfüllt wegen $f'(u) = -2 \sum_{i=1}^n (x_i - u)$

$$\sum_{i=1}^n (\bar{x} - x_i) = (-1)^n \sum_{i=1}^n (x_i - \bar{x}) = (-1)^{n+1} \frac{1}{2} f'(\bar{x}) = 0. \quad (7.3)$$

Das sieht man auch direkt ohne Verwendung der Ableitung:

$$\sum_{i=1}^n (\bar{x} - x_i) = n\bar{x} - \sum_{i=1}^n x_i = n\bar{x} - n\bar{x} = 0.$$

In diesem Sinne ist $\bar{x}$ der Schwerpunkt der Merkmalsausprägungen x_i : Die Addition aller Entfernungen der x_i zu $\bar{x}$ ergibt null. Stellt man sich die Strecken von $\bar{x}$ zu den Werten x_i als gleich dicke Rohre vor, dann ist $\bar{x}$ die Position des physikalischen Schwerpunkts der zusammengeklebten Rohre auf dem Zahlenstrahl.

Das arithmetische Mittel ist linear: Hat man Werte $x_1, \dots, x_n$ und $y_1, \dots, y_n$ und eine Konstante $c \in \mathbb{R}$, so gilt:

$$\overline{cx} = \frac{1}{n} \sum_{i=1}^n cx_i = c \frac{1}{n} \sum_{i=1}^n x_i = c\bar{x},$$

d. h., wenn man zunächst alle Merkmalsausprägungen mit c multipliziert und dann das arithmetische Mittel berechnet, kommt man zum gleichen Ergebnis wie bei der anschließenden Multiplikation von $\bar{x}$ mit c . Außerdem gilt

$$\overline{x+y} = \frac{1}{n} \sum_{i=1}^n (x_i + y_i) = \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{n} \sum_{i=1}^n y_i = \bar{x} + \bar{y}.$$

Hat man also zwei Merkmale X und Y und bildet man daraus ein neues Merkmal als Linearkombination (vgl. Seite 470) $c \cdot X + d \cdot Y$, so ist dessen arithmetisches Mittel gleich der entsprechenden Linearkombination der arithmetischen Mittel von X und Y .

7.1.3.2 Median

Einzelne Ausreißer, also besonders kleine und besonders große Werte, können das arithmetische Mittel stark beeinflussen. Im Beispiel der Studiendauern zieht ein Bummelstudent mit seinen 20 Semestern den ganzen Schnitt deutlich nach oben. Das kann nicht im Interesse einer Hochschule sein. Möchte man den Einfluss von Ausreißern klein halten, eignet sich der folgende Lageparameter besser:

Definition 7.4 (Median)

Der **Median (Zentralwert)** Z ist der mittlere Wert der nach Größe geordneten (reellen) Merkmalsausprägungen, d. h., für die aufsteigend geordneten Ausprägungen $x_1, x_2, \dots, x_n$ ist

$$Z := \begin{cases} x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade} \\ \frac{1}{2} \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), & \text{falls } n \text{ gerade.} \end{cases}$$

Achtung: $x_{\frac{n+1}{2}}$ ist die Merkmalsausprägung an der Stelle $\frac{n+1}{2}$ in der Liste aller Merkmalsausprägungen. $\frac{n+1}{2}$ ist ein Index und nicht der Wert des Medians.

Beispiel 7.7 (Wahre und scheinbare Armut)

In der Rheinischen Post vom 06.04.2013 erschien auf Seite B4 eine Kolumne, in der behauptet wurde, dass eine Einwanderung des Milliardärs Bill Gates die Zahl der nach einer offiziellen Statistik als arm geltenden Menschen in Deutschland erhöhen würde: „Das Problem dieser Armutsmessung ist die Orientierung am mittleren Einkommen, dem Median. Das führt zu dem paradoxen Ergebnis, dass die Zahl der Armen in einem Land steigt, sobald ein Milliardär einwandert“. Hier verwechselt die Autorin die Begriffe „Median“ und „arithmetisches Mittel“. Der Median wird gerade deshalb verwendet, um den Einfluss von Ausreißern gering zu halten. Als Beispiel betrachten wir die Einkommen 1 000, 2 000, 3 000, 3 000 und 11 000 Euro. Das arithmetische Mittel ist $\bar{x} = 20\,000/5 = 4\,000$ Euro, der Median liegt bei 3 000 Euro. Jetzt kommt ein Einkommensmilliardär hinzu. Das arithmetische Mittel liegt jetzt bei 166 670 000 Euro, der Median bleibt aber bei 3 000 Euro. ■

Beispiel 7.8

Üblicherweise wird die mittlere Studiendauer über den Median und nicht über das arithmetische Mittel angegeben. In unserem Beispiel sortieren wir die Werte von X der Größe nach: 7, 7, 8, 8, 20. Der mittlere Wert (und damit der Median) ist 8. Die lange Studiendauer von 20 Semestern spielt hier keine Rolle mehr. ■

So wie das arithmetische Mittel die Zahl mit der kleinsten Summe quadrierter Abstände zu den Merkmalsausprägungen ist, minimiert der Median die Summe der absoluten Differenzen

$$f(u) := \sum_{i=1}^n |x_i - u|.$$

Bei geradem n ist jeder Wert zwischen $x_{\frac{n}{2}}$ und $x_{\frac{n}{2}+1}$ ein Minimum.

7.1.3.3 p-Quantil

Der Begriff des Medians lässt sich verallgemeinern zum p -Quantil x_p (wobei der Name an Quantum erinnert): Die Werte $x_1, \dots, x_n$ seien aufsteigend sortiert. Links vom Median liegen höchstens 50 % der Merkmalsausprägungen, rechts vom Median liegen ebenfalls höchstens 50 % der Merkmalsausprägungen. Jetzt wählen wir statt 50 % eine beliebige Prozentzahl $(100 \cdot p) \%$ (beim Median ist $p = \frac{1}{2}$) und suchen einen Wert x_p , so dass links von x_p höchstens $(100 \cdot p) \%$ der Merkmalsausprägungen und rechts von x_p höchstens $[100 \cdot (1 - p)] \%$ der Merkmalsausprägungen liegen. Da x_p dadurch noch nicht eindeutig festgelegt ist, werden wir etwas präziser und verwenden sowohl bei der Definition als auch bei der Berechnung des p -Quantils die Verteilungsfunktion F :

Definition 7.5 (p -Quantil)

Ein p -Quantil ist die Stelle, an der die Verteilungsfunktion den Wert p überspringt. Falls diese Stelle nicht eindeutig ist, nimmt man die Mitte des Intervalls, auf dem die Verteilungsfunktion gleich p ist (siehe Abbildung 7.3).

Für $p = 0,5$ kann man so auch direkt den Median an der Verteilungsfunktion ablesen. In Formeln lautet die Definition: Zu $p \in]0, 1[$ ist das p -Quantil x_p definiert zu

- $x_p := a_k$, falls $F(a_k) > p$ und $(k = 1 \text{ oder } F(a_{k-1}) < p)$,
- $x_p := \frac{a_k + a_{k+1}}{2}$, falls $F(a_k) = p$.

Beispiel 7.9

Für die Studiendauern (7, 7, 8, 8, 20) berechnen wir einige Quantile:

- a) $x_{0,2} = 7$, $x_{0,00001} = 7$, denn $F(7) = \frac{2}{5} = 0,4$, $0,4 > 0,2 > 0,00001$ und $7 = a_1$,
- b) $x_{0,4} = 7,5$, denn $F(7) = 0,4$, also $x_{0,4} = \frac{7+8}{2}$.
- c) $x_{0,25} = 7$ (wie oben), $x_{0,75} = 8$, denn $F(8) = \frac{4}{5} > 0,75$ und $F(7) = 0,4 < 0,75$. ■

Eine besondere Rolle bei der Einschätzung der Streuung spielen die 0,25- und 0,75-Quantile, die auch als **Quartile** bezeichnet werden (siehe Kapitel 7.1.4). Der Name stammt aus dem Lateinischen und ist verwandt mit dem englischen „quarter“ (Viertel). $x_{0,25}$ gibt an, bis wo das erste Viertel der Werte reicht, $x_{0,75}$ besagt, wo das letzte Viertel beginnt.

7.1.3.4 Modalwert

Definition 7.6 (Modalwert)

*Ein häufigster Wert (**Modalwert**) ist eine Merkmalsausprägung a_k mit einer größten absoluten (oder – äquivalent – relativen) Häufigkeit h_k (oder f_k). Gibt es nur einen häufigen Wert (einen Modalwert, der sich bezüglich der Häufigkeiten deutlich von*

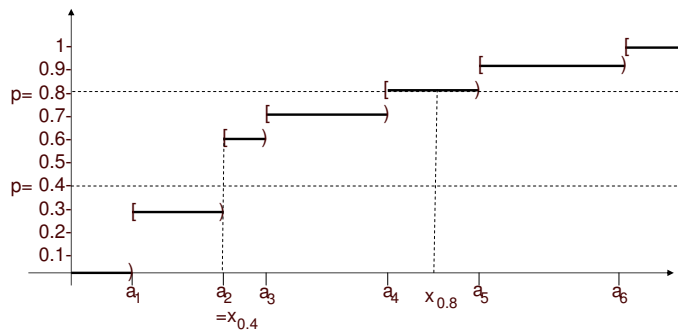


Abb. 7.3: Bestimmung der Quantile zu $p = 0,4$ und $p = 0,8$

den anderen Werten abhebt), spricht man von einer **unimodalen Verteilung**, bei zwei Werten, die im Vergleich mit den anderen Werten sehr häufig sind, von einer **bimodalen Verteilung**.

Der Modalwert eignet sich nicht bei stetigen Merkmalen, da ja hier in der Regel gilt: $h_k = 1$. Dagegen ist er gut anwendbar bei diskreten Merkmalen.

Beispiel 7.10

Bei unseren Studiendauern gibt es die beiden Modalwerte 7 und 8, es liegt damit eine bimodale Verteilung vor. ■

Bimodale Verteilungen finden sich oft auch bei den Ergebnissen von Mathe-Klausuren im ersten Semester, da die Vorkenntnisse sehr unterschiedlich sind und damit sowohl die Note 1 als auch die Note 5 häufig vorkommen.

Eine schwierige Aufgabe der digitalen Bildverarbeitung ist das Separieren von Objekten in Bildern. Ein Ansatz besteht darin, Helligkeitsinformationen der einzelnen Bildpunkte (**Grauwerte**) zu betrachten. Gibt es eine Häufung von nahe zusammenliegenden Grauwerten, so könnte sich dahinter ein Objekt mit entsprechender Farbinformation verbergen. Man sucht hier nicht nur nach Modalwerten, sondern generell nach häufigen Werten.

7.1.3.5 Geometrisches Mittel

Ein weiterer Lageparameter ist das geometrische Mittel, das wir schon kurz im Rahmen von Ungleichungen kennengelernt haben (siehe Seite 70). Dort haben wir es als Mittelwert für Wachstumsraten wie beispielsweise Zinssätze eingesetzt. Wenn die In-

flation in den letzten drei Jahren jeweils 5 %, 10 % und 15 % betragen hat, dann ist die durchschnittliche Inflation nicht das arithmetische Mittel 10 %. Denn ein Artikel, der zunächst x Euro gekostet hat, kostet nach den drei Jahren $x \cdot 1,05 \cdot 1,1 \cdot 1,15 = x \cdot 1,32825$ Euro. Das arithmetische Mittel führt aber zu einem Faktor $1,1 \cdot 1,1 \cdot 1,1 = 1,331$ und ist damit etwas zu groß. Richtig wäre eine Aufteilung des Faktors 1,32825 auf drei Jahre, indem wir die dritte Wurzel ziehen: 1,0992... Die durchschnittliche Inflation beträgt ca. 9,92 %.

Definition 7.7 (Geometrisches Mittel)

Seien die Merkmalsausprägungen x_i größer als null, $1 \leq i \leq n$. Dann ist das **geometrische Mittel** definiert als die n -te Wurzel des Produkts der Ausprägungen:

$$G := \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} = \sqrt[n]{\prod_{i=1}^n x_i}.$$

Betrachten wir statt der Ausprägungen x_i die Werte $\ln x_i$ und bilden deren arithmetisches Mittel L , so gilt:

$$L := \frac{1}{n} \sum_{i=1}^n \ln x_i = \frac{1}{n} \ln \left(\prod_{i=1}^n x_i \right) = \ln \left(\left[\prod_{i=1}^n x_i \right]^{\frac{1}{n}} \right).$$

Damit ist $\exp L = G$. Beim geometrischen Mittel betrachtet man in diesem Sinne die einzelnen Merkmalsausprägungen auf einer logarithmischen Skala.

7.1.4 Streuungsparameter

In diesem Abschnitt interessiert uns, ob die einzelnen Merkmalsausprägungen alle eng zusammenliegen, ob sie sich über einen breiten Bereich erstrecken oder ob es einzelne Ausreißer gibt. Durch Vergleich des arithmetischen Mittels $\bar{x}$ (Schwerpunkt) und des Medians Z (mittlerer Wert) erhält man bereits ein Gefühl zur Streuung:

- $\bar{x} \approx Z$: symmetrische Verteilung,
- $\bar{x} > Z$: rechts stärker als links streuende Verteilung,
- $\bar{x} < Z$: links stärker als rechts streuende Verteilung.

Allerdings lassen sich auch Ausnahmen zu diesen heuristischen Regeln angeben, siehe von Hippel (2005).

Definition 7.8 (Einfache Streuungsparameter)

Die **Spannweite** R (Range) der Merkmalsausprägungen $x_1, \dots, x_n$ ist definiert als die maximale Differenz zweier Ausprägungen, also als

$$R := \max\{x_k : k = 1, \dots, n\} - \min\{x_k : k = 1, \dots, n\}.$$

Der **Quartilsabstand** Q ist der Abstand zwischen dem 0,75- und dem 0,25-Quantil:

$$Q := x_{0,75} - x_{0,25}.$$

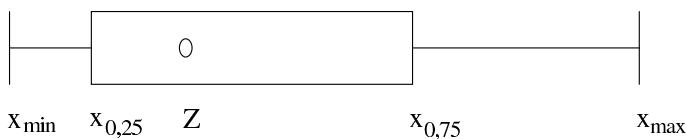
Bemerkungen:

- Man beachte, dass man hier wegen der Wahl von $p = 1/4$ und $p = 3/4$ von Quartilsabstand und nicht von Quantilsabstand spricht.
- Mit der Spannweite misst man insbesondere den Abstand zwischen einem kleinsten und einem größten Ausreißer.
- Diese Ausreißer gehen nicht in den Quartilsabstand ein.

Beispiel 7.11

Bei den Studienlängen 7, 7, 8, 8 und 20 ist $R = 20 - 7 = 13$ und $Q = 8 - 7 = 1$. ■

Zeichnet man die $x_{0,75}$ - und $x_{0,25}$ -Quartile zusammen mit dem Median, der kleinsten und der größten Merkmalsausprägung in eine Grafik (**Box-Plot**), so sieht man sofort, ob die Verteilung symmetrisch ist und ob es Ausreißer gibt.

**Abb. 7.4:** Box-Plot**Definition 7.9 (Empirische Varianz)**

Die **empirische Varianz** s^2 der Merkmalsausprägungen $x_1, \dots, x_n$ ist definiert über die Summe der quadrierten Differenzen durch $n - 1$:

$$s^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (7.4)$$

Beispiel 7.12

Bei den Studiendauern 7, 7, 8, 8 und 20 mit arithmetischem Mittel 10 ist die Varianz

$$s^2 = \frac{1}{4} [(7-10)^2 + (7-10)^2 + (8-10)^2 + (8-10)^2 + (20-10)^2] = \frac{126}{4} = 31,5. \quad \blacksquare$$

Bemerkungen:

- Wir haben bereits in (7.2) gesehen, dass das arithmetische Mittel die Funktion $f(u) := \sum_{i=1}^n (x_i - u)^2$ minimiert. Das Minimum ist nun $(n-1)s^2$.
- Die Bezeichnung s^2 rührt daher, dass wir hier quadrierte Abstände aufsummieren und daher das Ergebnis ≥ 0 ist.
- Die Summe wird durch $n-1$ geteilt. Naheliegender wäre bei n Summanden die Division durch n . Daher findet man auch $s^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ als abweichende Definition. Für große n sind beide Definitionen nahezu gleich. Wir werden jedoch (7.4) verwenden, da die Formel eine wichtige Eigenschaft besitzt, die wir im Rahmen der schließenden Statistik kennenlernen werden (Erwartungstreue, siehe Seite 1059).
- Statt der quadrierten Differenzen wäre es naheliegend, ein Streuungsmaß über Beträge der Differenzen einzuführen, wie wir es schon beim Median gesehen haben:

$$\tilde{s} := \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

Allerdings ist der Absolutbetrag wegen der eingebauten Fallunterscheidung schwieriger zu handhaben als die Quadratfunktion. Diese Größe findet sich auch nicht als Parameter in Ergebnissen der Wahrscheinlichkeitsrechnung wieder – im Gegensatz zur Varianz, die ein Parameter der Normalverteilung ist und damit quasi in der Natur vorkommt.

- Für die Regressionsrechnung hilft eine etwas andere Darstellung der Varianz:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 \right] - 2\frac{n}{n-1}\bar{x}^2 + \frac{n}{n-1}\bar{x}^2 = \frac{1}{n-1} \left(\left[\sum_{i=1}^n x_i^2 \right] - n\bar{x}^2 \right). \end{aligned} \quad (7.5)$$

Definition 7.10 (Empirische Standardabweichung)

Sei s^2 die empirische Varianz aus (7.4). Man nennt $s = \sqrt{s^2}$ die **empirische Standardabweichung** der Merkmalsausprägungen $x_1, \dots, x_n$.

Eigentlich wäre einer der beiden Begriffe Varianz und Standardabweichung ausreichend. Historisch bedingt werden jedoch beide eingesetzt.

7.1.5 Zweidimensionale Häufigkeitsverteilungen und Korrelation

Statt nur ein Merkmal X zu betrachten, sehen wir uns nun zwei Merkmale X und Y an, die für die gleiche statistische Masse erklärt sind. Jedem Merkmalsträger i wird dadurch ein Paar von Merkmalsausprägungen (x_i, y_i) zugeordnet. Jetzt mögen

$a_1, \dots, a_l$ die verschiedenen Merkmalsausprägungen von X und $b_1, \dots, b_m$ die verschiedenen Merkmalsausprägungen von Y bezeichnen. Die absolute Häufigkeit jeder Ausprägung a_i von X bezeichnen wir nun mit $h_{i,\bullet}$, die absolute Häufigkeit jeder Ausprägung b_k von Y mit $h_{\bullet,k}$. Nun interessiert auch die absolute Häufigkeit eines Werts (a_i, b_k) , die wir $h_{i,k}$ nennen. $h_{i,k}$ ist also genau die Anzahl der statistischen Elemente, für die gleichzeitig X den Wert a_i und Y den Wert b_k annimmt. Damit entsteht die folgende **zweidimensionale Häufigkeitstabelle (Kontingenztafel)**:

	$Y = b_1$	$Y = b_2$	$\dots$	$Y = b_m$	$\sum$
$X = a_1$	$h_{1,1}$	$h_{1,2}$	$\dots$	$h_{1,m}$	$h_{1,\bullet}$
$X = a_2$	$h_{2,1}$	$h_{2,2}$	$\dots$	$h_{2,m}$	$h_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$X = a_l$	$h_{l,1}$	$h_{l,2}$	$\dots$	$h_{l,m}$	$h_{l,\bullet}$
$\sum$	$h_{\bullet,1}$	$h_{\bullet,2}$	$\dots$	$h_{\bullet,m}$	n

- Die absoluten Häufigkeiten der Zeile i ergeben in Summe die absolute Häufigkeit $h_{i,\bullet}$ von a_i .
- Die absoluten Häufigkeiten der Spalte k ergeben in Summe die absolute Häufigkeit $h_{\bullet,k}$ von b_k .
- Diese Werte in der rechten Spalte und unteren Zeile werden als **Randhäufigkeiten** bezeichnet.
- Man beachte, dass hier $h_{i,k} = 0$ erlaubt ist. Für $X = Y$ ist beispielsweise nur die Hauptdiagonale gefüllt. Die Randhäufigkeiten sind dagegen stets > 0 .

Trägt man die Punkte (x_i, y_i) in ein Diagramm ein, spricht man von einem **Streudiagramm** oder auch von einer **Punktwolke**. Hier kann man ablesen, ob es einen funktionalen Zusammenhang zwischen den Werten von X und von Y gibt. Der folgende Begriff ist ein erster Schritt, einen solchen Zusammenhang zu erfassen:

Definition 7.11 (Empirische Kovarianz)

Die **empirische Kovarianz** der Punkte $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ist

$$\text{Cov}(X, Y) := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Analog zur Varianz verwenden wir auch die Schreibweise $s_{xy} := \text{Cov}(X, Y)$.

Bemerkungen:

- Die Formel erinnert stark an die Definition der Varianz. Man beachte, dass hier aber die Bezeichnung s_{xy} und nicht s_{xy}^2 ist. Das hat einen guten Grund: Die Kovarianz kann negativ sein.

- Die absolute Größe der Kovarianz sagt wenig aus, da wir mit den Werten von X und Y zwei völlig verschiedene Größen multiplizieren.
- Das Vorzeichen der Kovarianz hingegen erlaubt einen Rückschluss:
 - Ist die Kovarianz positiv, so überwiegen Punkte im ersten und dritten Quadranten eines zum Ursprung $(\bar{x}, \bar{y})$ verschobenen Koordinatensystems. Es liegt eine **positive Korrelation** vor.
 - Ist die Kovarianz negativ, so überwiegen Punkte im zweiten und vierten Quadranten. Es liegt eine **negative Korrelation** vor (siehe Abbildung 7.5).
- Multipliziert man das Produkt innerhalb der Summe aus, erhält man eine Darstellung, die wir noch bei der Regressionsrechnung verwenden:

$$\begin{aligned}
 s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\
 &= \frac{1}{n-1} \left(\left[\sum_{i=1}^n x_i y_i \right] + n \bar{x} \bar{y} \right) - \frac{1}{n-1} \sum_{i=1}^n x_i \bar{y} - \frac{1}{n-1} \sum_{i=1}^n \bar{x} y_i \\
 &= \frac{1}{n-1} \left(\left[\sum_{i=1}^n x_i y_i \right] + n \bar{x} \bar{y} \right) - \frac{2n}{n-1} \bar{x} \bar{y} = \frac{1}{n-1} \left(\left[\sum_{i=1}^n x_i y_i \right] - n \bar{x} \bar{y} \right). \quad (7.6)
 \end{aligned}$$

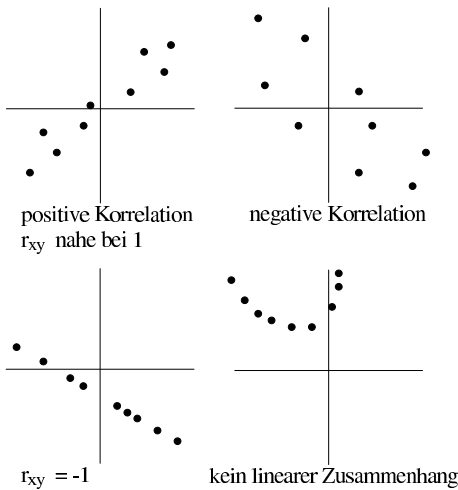


Abb. 7.5: Korrelation

Während die Kovarianz nur die Tendenz des Zusammenhangs im Sinne positiver oder negativer Korrelation angibt, wird der Wert erst durch Normierung aussagekräftig:

Definition 7.12 (Pearson'scher Korrelationskoeffizient)

Gegeben seien $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \in \mathbb{R}^2$, wobei mindestens ein $x_i \neq \bar{x}$ und ein $y_i \neq \bar{y}$ ist, d. h., nicht alle x_i sind gleich, und nicht alle y_i sind gleich. Der empirische Korrelationskoeffizient (Pearson'scher Korrelationskoeffizient)

$$\begin{aligned} r_{xy} &:= \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

ist der Quotient aus Kovarianz und dem Produkt der Standardabweichungen.

Satz 7.1 (Bedeutung des Pearson'schen Korrelationskoeffizienten)

Unter den Voraussetzungen von Definition 7.12 gilt:

- a) $-1 \leq r_{xy} \leq 1$,
- b) $r_{xy} = 1$ gilt genau dann, wenn die Punkte auf einer Geraden mit positiver Steigung liegen, also $y_i = c + dx_i$ mit $c, d \in \mathbb{R}$ und $d > 0$,
- c) $r_{xy} = -1$ gilt genau dann, wenn die Punkte auf einer Geraden mit negativer Steigung liegen, also $y_i = c + dx_i$ mit $c, d \in \mathbb{R}$ und $d < 0$ (siehe Abbildung 7.5).

Beweis: Für Vektoren $\vec{a} := (a_1, a_2, \dots, a_n)$ und $\vec{b} := (b_1, b_2, \dots, b_n) \in \mathbb{R}^n$ haben wir auf Seite 484 das Standardskalarprodukt $\vec{a} \cdot \vec{b} = \sum_{k=1}^n a_k b_k$ kennengelernt. Ist φ der Winkel zwischen $\vec{a}$ und $\vec{b}$, so gilt (siehe Seite 435):

$$|\vec{a} \cdot \vec{b}| = |\vec{a}| \cdot |\vec{b}| \cdot |\cos \varphi| \leq |\vec{a}| |\vec{b}| = \sqrt{\sum_{k=1}^n a_k^2} \sqrt{\sum_{k=1}^n b_k^2}.$$

Gleichheit liegt in dieser Beziehung genau dann vor, wenn φ ein Vielfaches von π ist, d. h., falls es ein $\lambda \in \mathbb{R}$ gibt mit $\vec{b} = \lambda \vec{a}$.

Wir wenden nun diese Aussage an, indem wir $\vec{a} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$ und $\vec{b} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$ setzen:

$$\begin{aligned} |s_{xy}| &= \frac{1}{n-1} \left| \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right| = \frac{1}{n-1} |\vec{a} \cdot \vec{b}| \\ &\leq \frac{1}{n-1} |\vec{a}| |\vec{b}| = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = s_x s_y, \end{aligned}$$

also $-1 \leq r_{xy} = \frac{s_{xy}}{s_x s_y} \leq 1$. Gleichheit $|r_{xy}| = 1$ gilt genau dann, wenn es ein $\lambda \in \mathbb{R}$ gibt mit $\vec{b} = \lambda \vec{a}$, d. h. $y_i - \bar{y} = \lambda(x_i - \bar{x})$ für alle Komponenten $1 \leq i \leq n$. Damit gilt aber $y_i = c + dx_i$ mit $d = \lambda$ und $c = -\lambda\bar{x} + \bar{y}$, $1 \leq i \leq n$.

Wir berechnen r_{xy} für $y_i = c + dx_i$ mit $c, d \in \mathbb{R}$: Zunächst ist $\bar{y} = c + d\bar{x}$, so dass

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(c + dx_i - (c + d\bar{x}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (c + dx_i - (c + d\bar{x}))^2}} = \frac{ds_x^2}{|d|s_x^2} = \text{sign}(d). \quad (7.7)$$

Damit ist die Richtung $\Leftarrow$ in b) und c) gezeigt. Umgekehrt haben wir für $\Rightarrow$ bereits bewiesen, dass $y_i = c + dx_i$ gilt. Wegen (7.7) folgt außerdem für b) $d > 0$ und für c) $d < 0$. $\square$

Bemerkungen:

- Ist der Wert $|r_{xy}|$ nahe bei 1, liegen die Punkte fast auf einer Geraden. Je nach Vorzeichen spricht man dann von einer **stark positiven** oder **stark negativen** Korrelation.
- Ist $r_{xy} = 0$, so nennt man X und Y **unkorreliert**. Die Vektoren $\vec{a}$ und $\vec{b}$ aus dem vorangehenden Beweis (Vektoren der Differenzen zum arithmetischen Mittel) stehen dann senkrecht zueinander.
- Eine starke Korrelation muss nicht automatisch bedeuten, dass zwischen den Merkmalswerten für X und Y ein ursächlicher Zusammenhang besteht. Wenn überhaupt, so kann dies auch nur Hinweis auf **lineare** Zusammenhänge sein. Es mag eine starke Korrelation zwischen der Anzahl von Störchen und der Geburtenrate geben, aber was kann man daraus schließen?

7.1.6 Kovarianzmatrix

Die Rolle der Kovarianz wird verständlicher, wenn wir Eigenschaften der **Kovarianzmatrix**

$$\mathbf{K} := \begin{bmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{bmatrix} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}$$

untersuchen. Hier liegen zwei Merkmale X und Y mit gleich vielen Merkmalswerten und den empirischen Varianzen s_x^2 und s_y^2 sowie der empirischen Kovarianz s_{xy} vor. Offensichtlich ist die Kovarianzmatrix symmetrisch, so dass wir den folgenden Hilfssatz anwenden können:

Satz 7.2 (Bedeutung der Eigenwerte symmetrischer Matrizen)

Sei $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ eine symmetrische Matrix. (Wegen der Symmetrie sind alle Eigenwerte reell, und Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal zueinander bezüglich des Standardskalarprodukts, siehe z. B. Satz 3.36 auf Seite 545.) Wir betrachten (Richtungs-) Vektoren der Länge eins in der Darstellung $(\cos(\alpha), \sin(\alpha))$ $\alpha \in [0, 2\pi[$. Dann wird

$$(\cos(\alpha), \sin(\alpha)) \cdot \mathbf{A} \cdot \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix}$$

genau dann maximal, falls $(\cos(\alpha), \sin(\alpha))$ ein Eigenvektor zum größten Eigenwert λ ist. Die Zahl wird genau dann minimal, falls $(\cos(\alpha), \sin(\alpha))$ ein Eigenvektor zum kleinsten Eigenwert μ ist.

Die Aussage gilt allgemein für beliebige symmetrische $n \times n$ -Matrizen, aber wir benötigen für die Kovarianzmatrix nur den 2×2 -Fall.

Beweis: Wegen Satz 3.39 auf Seite 549 und Satz 3.41 auf Seite 552 hat die symmetrische Matrix entweder einen doppelten reellen Eigenwert $\lambda = \mu$, zu dem es zwei linear unabhängige Eigenvektoren gibt, oder sie hat zwei verschiedene Eigenwerte $\lambda > \mu$ mit zugehörigen linear unabhängigen und sogar hinsichtlich des Standardskalarprodukts orthogonalen Eigenvektoren $\vec{d}_\lambda$ und $\vec{d}_\mu$, deren Länge wir durch Normierung zu eins wählen können. Im ersten Fall sind alle Vektoren $\vec{d} \in \mathbb{R}^2$, $\vec{d} \neq \vec{0}$, Eigenvektoren und

$$\begin{aligned} (\cos(\alpha), \sin(\alpha)) \cdot \mathbf{A} \cdot \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix} &= \lambda(\cos(\alpha), \sin(\alpha)) \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix} \\ &= \lambda[\cos^2(\alpha) + \sin^2(\alpha)] = \lambda = \mu. \end{aligned}$$

Damit ist in diesem Fall die Aussage trivial. Interessanter ist die Situation $\lambda > \mu$. Zusammenfassend haben wir in diesem Fall die Situation

$$\mathbf{A}\vec{d}_\lambda = \lambda\vec{d}_\lambda, \quad \mathbf{A}\vec{d}_\mu = \mu\vec{d}_\mu, \quad \vec{d}_\lambda \cdot \vec{d}_\mu = 0, \quad \vec{d}_\lambda \cdot \vec{d}_\lambda = 1, \quad \vec{d}_\mu \cdot \vec{d}_\mu = 1.$$

Wegen der linearen Unabhängigkeit der Eigenvektoren können wir die Richtungsvektoren als Linearkombination schreiben:

$$\begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix} = r\vec{d}_\lambda + s\vec{d}_\mu.$$

Wegen der normierten Länge gilt dann

$$1 = (r\vec{d}_\lambda + s\vec{d}_\mu) \cdot (r\vec{d}_\lambda + s\vec{d}_\mu) = r^2\vec{d}_\lambda \cdot \vec{d}_\lambda + 2rs\vec{d}_\lambda \cdot \vec{d}_\mu + s^2\vec{d}_\mu \cdot \vec{d}_\mu = r^2 + s^2,$$

so dass wir $r = \cos(\varphi)$ und $s = \sin(\varphi)$ schreiben können. Damit suchen wir Maxima und Minima der Funktion

$$\begin{aligned} f(\varphi) &:= (\cos(\varphi)\vec{d}_\lambda + \sin(\varphi)\vec{d}_\mu)^\top \cdot \mathbf{A} \cdot (\cos(\varphi)\vec{d}_\lambda + \sin(\varphi)\vec{d}_\mu) \\ &= \cos^2(\varphi)\lambda\vec{d}_\lambda \cdot \vec{d}_\lambda + \cos(\varphi)\sin(\varphi)[\lambda + \mu]\vec{d}_\lambda \cdot \vec{d}_\mu + \sin^2(\varphi)\mu\vec{d}_\mu \cdot \vec{d}_\mu \\ &= \cos^2(\varphi)\lambda + \sin^2(\varphi)\mu. \end{aligned}$$

Mittels eindimensionaler Extremwertrechnung bestimmen wir die Winkel φ , für die Maxima bzw. Minima angenommen werden. Aus der notwendigen Bedingung

$$0 = f'(\varphi) = 2\sin(\varphi)\cos(\varphi) \underbrace{[-\lambda + \mu]}_{<0}$$

folgt $\varphi = n\frac{\pi}{2}$, $n \in \mathbb{Z}$. Wegen

$$f''(\varphi) = 2[\cos^2(\varphi) - \sin^2(\varphi)][-\lambda + \mu]$$

und $f''(0) = f''(\pi) = 2[-\lambda + \mu] < 0$, $f''(\frac{\pi}{2}) = f''(\frac{3}{2}\pi) = -2[-\lambda + \mu] > 0$, liegen im Intervall $[0, 2\pi[$ bei 0 und π lokale (und wegen der Periodizität auch globale) Maxima mit Wert λ und bei $\frac{\pi}{2}$ und $\frac{3}{2}\pi$ lokale und globale Minima mit Wert μ . Genau in Richtung des Eigenvektors $\pm \vec{d}_\lambda$ wird also dessen größerer Eigenwert λ als Maximum und genau in Richtung des Eigenvektors $\pm \vec{d}_\mu$ wird dessen kleinerer Eigenwert μ als Minimum angenommen. $\square$

Wir zeigen jetzt, dass die Zahl

$$(\cos(\alpha), \sin(\alpha)) \cdot \mathbf{K} \cdot \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix}$$

die Varianz eines Merkmals $D = D(\alpha)$ beschreibt, dessen n Merkmalswerte d_i durch Projektion der n Punkte (x_i, y_i) auf die Ursprungsgerade mit Richtungsvektor $(\cos(\alpha), \sin(\alpha))$ und dazu senkrechter Normale $(-\sin(\alpha), \cos(\alpha))$ entstehen. Genauer ist d_i die (vorzeichenbehaftete) Entfernung des Lotfußpunktes (Lot von (x_i, y_i) auf die Gerade) zum Nullpunkt. Diese Entfernung entspricht der Entfernung der Geraden mit Richtungsvektor $(-\sin(\alpha), \cos(\alpha))$ durch (x_i, y_i) zum Nullpunkt, die sich gemäß der Hesse-Normalform (siehe Seite 463) zu

$$d_i = (\cos(\alpha), \sin(\alpha)) \cdot (x_i, y_i) = x_i \cos(\alpha) + y_i \sin(\alpha)$$

ergibt. Damit berechnen wir die Varianz des Merkmals $D = D(\alpha)$. Dessen arithmetisches Mittel ist $\bar{d} = \bar{x} \cos(\alpha) + \bar{y} \sin(\alpha)$, so dass

$$\begin{aligned} s_d^2 &= \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = \frac{1}{n-1} \sum_{i=1}^n ((x_i - \bar{x}) \cos(\alpha) + (y_i - \bar{y}) \sin(\alpha))^2 \\ &= s_x^2 \cos^2(\alpha) + 2s_{xy} \cos(\alpha) \sin(\alpha) + s_y^2 \sin^2(\alpha) = (\cos(\alpha), \sin(\alpha)) \cdot \mathbf{K} \cdot \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix}. \end{aligned}$$

Nach Lemma 7.2 zeigen Eigenvektoren zum größten Eigenwert von $\mathbf{K}$ in Richtung der größten und Eigenvektoren zum kleinsten Eigenwert in Richtung der kleinsten Varianz. Der Eigenvektor zum größten Eigenwert drückt daher die Richtung aus, in der die Punktwolke $\{(x_i, y_i) : 1 \leq i \leq n\}$ am größten streut. Bei der Erstellung eines virtuellen Stadtmodells kann so z. B. die Lage des Firstes einer Kirche aus einer Wolke von Grundrisspunkten abgeschätzt werden, siehe Abbildung 7.6.

Die Kovarianz s_{xy} ist genau dann null, wenn $(1, 0)$ und $(0, 1)$ Eigenvektoren der Kovarianzmatrix sind, wenn also die größte Streuung in Richtung einer Koordinatenachse liegt. Dreht man eine gegebene Punktwolke entsprechend z. B. um den Schwerpunkt $(\bar{x}, \bar{y})$, so erhält man Punkte mit Werten zweier unkorrelierter Merkmale. Dieses Vorgehen, mit dem korrelierte in unkorrelierte Merkmale überführt werden, heißt **Hauptachsentransformation (Principal Component Analysis)**.

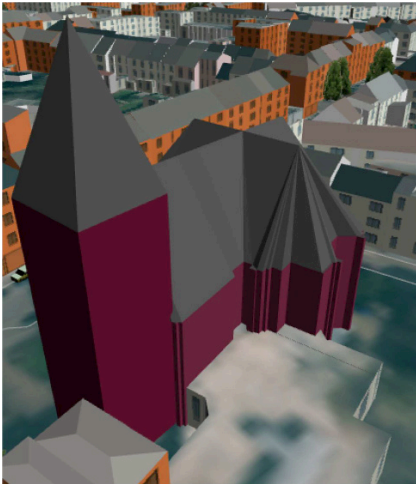


Abb. 7.6: Der Hauptfirst der Kirche wurde mit einer Hauptachsentransformation berechnet. Der First liegt auf einer Geraden durch den Schwerpunkt der (x, y) -Koordinaten des Grundrisses. Die Richtung der Geraden ist durch einen Eigenvektor zum größten Eigenwert der Kovarianzmatrix festgelegt.

Hat man die Punktwolke so gedreht, dass die größte Streuung in x -Richtung vorliegt, dann wird zugleich die Varianz der y -Koordinaten minimal, d. h., die quadrierten Abstände der y -Werte zu deren arithmetischem Mittel werden aufsummiert minimal. Betrachtet man eine horizontale Gerade, deren y -Achsenabschnitt dieses arithmetische Mittel ist, so ist daher die Summe der quadrierten Abstände der Punkte zur Geraden minimal. In dieser Lage geht die Gerade durch den Schwerpunkt der Punktwolke. Wenn wir nun die Drehung der Punktwolke rückgängig machen, so wird verständlich, dass die Gerade durch den Schwerpunkt $(\bar{x}, \bar{y})$ mit einem Eigenvektor zum größten Eigenwert der Kovarianzmatrix als Richtungsvektor den kleinsten aufsummierten quadrierten Abstand zu den Punkten aufweist.

Der kürzeste Abstand eines Punktes zur Geraden ist die Länge des Lots auf die Gerade. Häufig wird aber ein anderer Abstand minimiert, nämlich die Differenz der y -Koordinate eines Punktes zur y -Koordinate der Gerade bei gemeinsamer x -Koordinate. Das ist also der Abstand in Richtung der y -Achse. Wird die Gerade durch eine Funktion g beschrieben, dann ergibt sich für einen Punkt (x, y) der quadrierte Abstand $(g(x) - y)^2$. Die Gerade, die diese Abstände minimiert, zeigt in der Regel (falls die größte Streuung nicht in x -Richtung vorliegt) in eine andere Richtung als der Eigenvektor und heißt Regressionsgerade. Damit beschäftigt sich der nächste Abschnitt.

7.1.7 Lineare Regressionsrechnung

Ist der Korrelationskoeffizient nahe bei ± 1 , so stellt sich die Frage nach der Geraden, die einen möglichen Zusammenhang der Werte am besten beschreibt. Dies führt zum Begriff **Regressionsgerade**. Sie wird z. B. benötigt, wenn man über das Ohm'sche Gesetz $U = R \cdot I$ aus Messpunkten für Strom I und Spannung U den Widerstand R als Steigung der Geraden, die I auf U abbildet, ermitteln möchte. Verwendet man eine logarithmische Darstellung wie in Kapitel 1.5.8.4, so kann man mit einer Regressionsgerade auch exponentielle oder logarithmische Zusammenhänge analysieren.

Definition 7.13 (Empirische Regressionsgerade)

Gegeben seien die Punkte $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \in \mathbb{R}^2$, wobei mindestens ein $x_i \neq \bar{x}$ ist. Wir betrachten die quadrierten Abweichungen einer Geraden $u + vx$ von den vorliegenden Punkten im Sinne von

$$f(u, v) := \sum_{i=1}^n [y_i - (u + v \cdot x_i)]^2 = \sum_{i=1}^n [y_i - u - v \cdot x_i]^2.$$

Wird die Fehlerfunktion f für Parameter $u = u_0$ und $v = v_0$ minimal, so heißt die Gerade $y(x) := u_0 + v_0 x$, die damit den Punkten am nächsten liegt, eine **empirische Regressionsgerade**.

Die Forderung, dass ein x_i vom arithmetischen Mittel $\bar{x}$ verschieden ist, verhindert, dass man eine Parallele zur y -Achse erhält. Diese kann man nicht als Funktionsgraph schreiben. Die analoge Forderung, dass mindestens ein $y_i \neq \bar{y}$, ist nicht nötig. Dieser Fall führt zu einer Parallelen zur x -Achse.

Man spricht hier von der **Methode der kleinsten Quadrate**, da wir die Summe der quadrierten Differenzen minimieren. Etwas Vergleichbares haben wir auch bereits bei der Einführung der Fourier-Koeffizienten getan (siehe Seite 816).

Wir überlegen uns jetzt, wie die Regressionsgerade berechnet wird und sehen insbesondere, dass sie eindeutig ist. Dazu betrachten wir die „Straffunktion“ $f(u, v)$ mit den beiden Variablen u und v . Die Funktion ist stetig differenzierbar auf $\mathbb{R}^2$, und wir suchen eine Stelle $(u_0, v_0) \in \mathbb{R}^2$, an der die Funktion ein Minimum annimmt. Eine notwendige Bedingung dazu ist, dass beide partiellen Ableitungen an der Stelle (u_0, v_0) null sind (siehe Satz 4.7). Wir berechnen zunächst beide partiellen Ableitungen:

$$\begin{aligned} \frac{\partial}{\partial u} f(u, v) &= \sum_{i=1}^n \frac{\partial}{\partial u} (y_i - u - vx_i)^2 = \sum_{i=1}^n -2(y_i - u - vx_i) \\ &= -2 \left(\left[\sum_{i=1}^n y_i \right] - nu - v \sum_{i=1}^n x_i \right) = -2n(\bar{y} - u - v\bar{x}), \end{aligned}$$

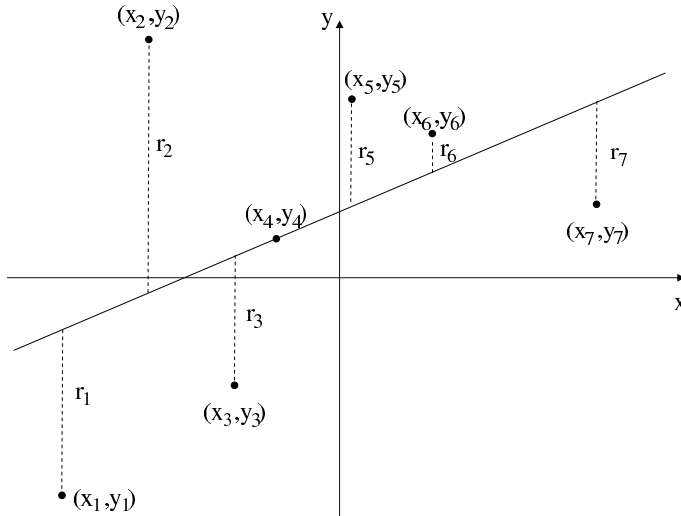


Abb. 7.7: empirische Regressionsgerade: $\sum_{i=1}^7 r_i^2$ ist minimal

$$\begin{aligned} \frac{\partial}{\partial v} f(u, v) &= \sum_{i=1}^n \frac{\partial}{\partial v} (y_i - u - vx_i)^2 = \sum_{i=1}^n -2x_i(y_i - u - vx_i) \\ &= -2 \left(\sum_{i=1}^n x_i y_i - u \sum_{i=1}^n x_i - v \sum_{i=1}^n x_i^2 \right) = -2n \left(\frac{1}{n} \left[\sum_{i=1}^n x_i y_i \right] - u\bar{x} - v \frac{1}{n} \sum_{i=1}^n x_i^2 \right). \end{aligned}$$

Aus der notwendigen Bedingung, dass die partiellen Ableitungen null sein müssen, ergibt sich das folgende lineare Gleichungssystem:

$$u + v\bar{x} = \bar{y} \quad \wedge \quad u\bar{x} + v \frac{1}{n} \sum_{i=1}^n x_i^2 = \frac{1}{n} \sum_{i=1}^n x_i y_i.$$

Ziehen wir von der zweiten Gleichung die erste multipliziert mit $\bar{x}$ ab, ergibt sich:

$$\begin{aligned} v \left(\frac{1}{n} \left[\sum_{i=1}^n x_i^2 \right] - \bar{x}^2 \right) &= \frac{1}{n} \left[\sum_{i=1}^n x_i y_i \right] - \bar{x}\bar{y} \\ \iff v \left(\left[\sum_{i=1}^n x_i^2 \right] - n\bar{x}^2 \right) &= \left[\sum_{i=1}^n x_i y_i \right] - n\bar{x}\bar{y} \\ \iff v \frac{1}{n-1} \left(\left[\sum_{i=1}^n x_i^2 \right] - n\bar{x}^2 \right) &= \frac{1}{n-1} \left(\left[\sum_{i=1}^n x_i y_i \right] - n\bar{x}\bar{y} \right) \\ \stackrel{(7.5), (7.6)}{\iff} v s_x^2 = s_{xy} \stackrel{s_x^2 > 0}{\iff} v &= \frac{s_{xy}}{s_x^2}. \end{aligned}$$

Man beachte, dass nach Voraussetzung mindestens ein x_i von $\bar{x}$ abweicht, so dass $s_x^2 > 0$ ist. Wir setzen den für v berechneten Wert in die erste Gleichung des Systems ein:

$$u = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}.$$

Das Gleichungssystem hat also genau diese eine Lösung $(u_0, v_0) := (\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}, \frac{s_{xy}}{s_x^2})$. Es gibt also höchstens ein (lokales) Maximum oder Minimum. Die Hesse-Matrix, in der die zweiten partiellen Ableitungen stehen (siehe Seite 612), ist hier die konstante

$$\text{Matrix} \begin{bmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^n x_i^2 \end{bmatrix}.$$

Die erste Hauptabschnittsdeterminante ist $2n > 0$, die zweite ist

$$\det \begin{bmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^n x_i^2 \end{bmatrix} = 4n \sum_{i=1}^n x_i^2 - 4n^2 \bar{x}^2 = 4 \left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right]. \quad (7.8)$$

Als Anwendung des Skalarprodukts ist (siehe Seite 435)

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

wobei φ der Winkel zwischen $(x_1, x_2, \dots, x_n)$ und $(1, 1, \dots, 1)$ ist. Damit erhalten wir $(\sum_{i=1}^n x_i)^2 = \cos^2(\varphi) \cdot n \sum_{i=1}^n x_i^2$ und insbesondere $(\sum_{i=1}^n x_i)^2 \leq n \sum_{i=1}^n x_i^2$, 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 $\bar{x}$ verschieden sein muss, ausgeschlossen haben. Damit ist aber die Determinante in (7.8) positiv, und die Hesse-Matrix ist nach Lemma 4.2 auf Seite 625 positiv definit. Es liegt also nach Satz 4.9 auf Seite 625 ein lokales Minimum vor. Man kann zeigen, dass es sich dabei auch um ein globales Minimum handelt.

Satz 7.3 (Regressionsgerade)

Die Regressionsgerade aus Definition 7.13 ist eindeutig bestimmt und lautet

$$y(x) = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right) + \frac{s_{xy}}{s_x^2} x,$$

d. h., wenn wir für s_{xy} und s_x^2 die Definitionen einsetzen und die arithmetischen Mittel $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ und $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ verwenden:

$$y(x) = \left(\bar{y} - \bar{x} \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \right) + \left(\frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \right) x.$$

- Man sieht sofort, dass das Vorzeichen der Steigung $\frac{s_{xy}}{s_x^2}$ dieser Geraden direkt dem Vorzeichen der Kovarianz entspricht.
- Setzt man $x = \bar{x}$ in die Geradengleichung ein, erhält man den Funktionswert $\bar{y}$. Der Schwerpunkt $(\bar{x}, \bar{y})$ der Punkte liegt damit auf der Regressionsgeraden.

Beispiel 7.13

Wir berechnen eine Regressionsgerade zu den Punkten (1, 1), (2, 3) und (3, 2): Die arithmetischen Mittel sind $\bar{x} = \frac{1}{3}(1 + 2 + 3) = 2$ und $\bar{y} = \frac{1}{3}(1 + 3 + 2) = 2$. Für die Kovarianz erhalten wir

$$s_{xy} = \frac{1}{2} \sum_{i=1}^3 (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{2}[-1 \cdot (-1) + 0 \cdot 1 + 1 \cdot 0] = \frac{1}{2}.$$

Die Varianz des ersten Merkmals ist $s_x^2 = \frac{1}{2} \sum_{i=1}^3 (x_i - \bar{x})^2 = \frac{1}{2}[1 + 0 + 1] = 1$, so dass $s_x = \sqrt{s_x^2} = 1$. Die Regressionsgerade ist damit

$$y(x) = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right) + \frac{s_{xy}}{s_x^2} x = \left(2 - \frac{1}{2} \cdot 2 \right) + \frac{1}{2} x = 1 + \frac{1}{2} x. \quad \blacksquare$$

Wir betrachten nun den Fehler, der entsteht, wenn wir die Werte y_i durch die Funktionswerte $\hat{y}_i := a + bx_i$ der Regressionsgeraden ersetzen. Die $\hat{y}_i$ heißen **theoretische y-Werte**. Diese haben das gleiche arithmetische Mittel wie die Werte y_i :

$$\bar{\hat{y}} := \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \bar{y},$$

denn wegen der Linearität des arithmetischen Mittels ist $\bar{\hat{y}} = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right) + \frac{s_{xy}}{s_x^2} \bar{x} = \bar{y}$. Falls tatsächlich zwischen den Merkmalen X und Y nicht nur ein vermuteter, sondern ein echter linearer Zusammenhang besteht, dann kann die Differenz $r_i := y_i - \hat{y}_i$ als durch zufällige Einflüsse entstanden verstanden werden. Die Differenz r_i heißt i -tes **Residuum**. Die Summe aller Residuen ist null:

$$\sum_{i=1}^n r_i = \sum_{i=1}^n (y_i - \hat{y}_i) = n \left[\frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n \hat{y}_i \right] = n(\bar{y} - \bar{\hat{y}}) = 0.$$

Für den durch die Regressionsgerade minimierten Fehler der Quadrate gilt:

$$\begin{aligned} \sum_{i=1}^n r_i^2 &= f \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}, \frac{s_{xy}}{s_x^2} \right) = \sum_{i=1}^n \left(y_i - \bar{y} + \frac{s_{xy}}{s_x^2} \bar{x} - \frac{s_{xy}}{s_x^2} x_i \right)^2 \\ &= \sum_{i=1}^n \left[(y_i - \bar{y}) - \frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2 \frac{s_{xy}}{s_x^2} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \left(\frac{s_{xy}}{s_x^2} \right)^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= (n-1) \left[s_y^2 - 2 \frac{s_{xy}}{s_x^2} s_{xy} + \frac{s_{xy}^2}{s_x^4} s_x^2 \right] = \frac{n-1}{s_x^2} [s_x^2 s_y^2 - 2s_{xy}^2 + s_{xy}^2] \\ &= \frac{n-1}{s_x^2} [s_x^2 s_y^2 - s_{xy}^2]. \end{aligned}$$

Der Fehler $\sum_{i=1}^n r_i^2$ ist also genau dann null, wenn $s_x^2 s_y^2 = s_{xy}^2$, d. h., wenn für den Korrelationskoeffizienten r_{xy} gilt: $r_{xy}^2 := \frac{s_{xy}^2}{s_x^2 s_y^2} = 1$. Diese Aussage haben wir bereits in Satz 7.1 kennengelernt, denn genau für $|r_{xy}| = 1$ liegen die Punkte (x_i, y_i) auf einer Geraden, die damit die Regressionsgerade sein muss.

Hintergrund: Regressionsrechnung mit Polynomen

Bei der linearen Regressionsrechnung haben wir eine Gerade, also ein Polynom ersten Grades, gesucht, die möglichst gut das Verhalten einer Punktwolke annähert. Statt eines Polynoms ersten Grades kann man analog auch Polynome höheren Grades verwenden, um nicht-lineare Zusammenhänge abzubilden. Möchte man beispielsweise das Weg-Zeit-Diagramm eines freien Falls basierend auf n Messwerten berechnen, so kann man keine Gerade durch die Punkte aus Zeit- und Wegkoordinate legen. Vielmehr wird der Zusammenhang zwischen Zeit t und Weg s bei Vernachlässigung des Luftwiderstands durch ein Polynom vom Grad zwei der Zeit beschrieben: $s(t) = a_2 \cdot t^2 + a_1 \cdot t + a_0$ mit $a_0, a_1, a_2 \in \mathbb{R}$. Diese Funktion kann man aus den Messdaten über **polynomiale Regression** bestimmen:

Es seien wieder die Punkte $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \in \mathbb{R}^2$ gegeben. Statt der Geraden $u + vx$ ist nun ein Polynom m -ten Grades $\sum_{k=0}^m a_k x^k$ gesucht, für das der Fehler

$$f(a_0, a_1, \dots, a_m) := \sum_{i=1}^n \left[y_i - \sum_{k=0}^m a_k x_i^k \right]^2$$

minimal wird. Aus der notwendigen Bedingung, dass bei einem Minimum die partiellen Ableitungen nach den Polynomkoeffizienten $a_0, \dots, a_m$ alle null sind (siehe wieder Satz 4.7 auf Seite 618), erhält man auch in diesem allgemeineren Fall Kandidaten für die Koeffizienten des Polynoms, $l = 0, \dots, m$:

$$\begin{aligned} 0 &= \frac{\partial f(a_0, a_1, \dots, a_m)}{\partial a_l} = 2 \sum_{i=1}^n \left[y_i - \sum_{k=0}^m a_k x_i^k \right] \cdot (-x_i^l) \\ \iff \sum_{i=1}^n \sum_{k=0}^m a_k x_i^k x_i^l &= \sum_{i=1}^n y_i x_i^l \iff \sum_{k=0}^m \left(\sum_{i=1}^n x_i^k x_i^l \right) a_k = \sum_{i=1}^n y_i x_i^l. \quad (7.9) \end{aligned}$$

Mit

$$\mathbf{A} := \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \quad \text{und} \quad \mathbf{A}^\top = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_1^m & x_2^m & \dots & x_n^m \end{bmatrix}$$

sowie der $((m+1) \times (m+1))$ -Matrix $\mathbf{A}^\top \mathbf{A} = [\sum_{i=1}^n x_i^l x_i^k]_{l=0, \dots, m, k=0, \dots, m}$ kann (7.9) geschrieben werden in der Form

$$\mathbf{A}^\top \mathbf{A} \vec{a} = \mathbf{A}^\top \vec{y}$$

mit $\vec{a} = (a_0, \dots, a_m)^\top$ und $\vec{y} = (y_1, \dots, y_n)^\top$.

Man nennt dieses lineare Gleichungssystem zur Bestimmung von $\vec{a}$ Gauß'sche Normalgleichungen. Für $m = 1$ sind die Koeffizienten der Regressionsgraden die Lösung.

7.2 Wahrscheinlichkeitsrechnung

Bei der beschreibenden Statistik liegen alle Daten vor. Hat man dagegen nur einen Ausschnitt und möchte man damit dennoch Aussagen über die Gesamtheit machen (z. B. bei Prognosen), benötigt man die Wahrscheinlichkeitsrechnung. In den folgenden Abschnitten werden die Grundzüge der Wahrscheinlichkeitstheorie dargestellt.

7.2.1 Zufallsexperimente und Ereignisse

Wir reden von einem **Zufallsexperiment**, wenn wir nicht alle Faktoren, die einen Einfluss auf das Ergebnis des Experiments haben, kennen oder bestimmen können. Der **Zufall** ist dabei der Einfluss dieser unbekannten Faktoren. Die Messung nahezu aller physikalischer Größen ist ein Zufallsexperiment. Die Quantentheorie basiert darüber hinaus auf der Existenz eines echten Zufalls, der nicht nur auf fehlendes Wissen zurückzuführen ist und sich bei der Messung von Quantenzuständen äußert (siehe Beispiel 7.30 auf Seite 1011).

Was ist nun die Wahrscheinlichkeit, dass ein gewisses Ergebnis eintritt? Die exakte Formulierung des vagen Begriffs Wahrscheinlichkeit, so dass man damit einerseits exakt rechnen kann und die Ergebnisse andererseits mit der Anschauung dieses Begriffs übereinstimmen, war eine schwierige Aufgabe für die Mathematik. Der Erfolg lag darin, aus der Anschauung ein Modell zu gewinnen und dann ohne weitere Anschauung rein mathematisch auf der Basis von Axiomen vorzugehen. Als Begründer der modernen Wahrscheinlichkeitsrechnung kann man den sowjetischen Mathematiker Andrey Nikolaevich Kolmogorov (1903–1987) sehen, der den axiomatischen Zugang in Kolmogorov (1933) (auf Deutsch) publizierte.

Betrachten wir eine Urne mit 20 weißen und 30 schwarzen Kugeln, die alle gleich seien. Anschaulich ist die Wahrscheinlichkeit, mit geschlossenen Augen eine weiße Kugel zu ziehen, gleich der relativen Häufigkeit der Merkmalsausprägung „weiß“, also $20/50 = 2/5$ bzw. 40 %. Die relative Häufigkeit dient als Vorbild für die Definition

der Wahrscheinlichkeit. Weiß man nicht, wie viele weiße und schwarze Kugeln in der Urne liegen, kann man n -mal mit Zurücklegen ziehen und dann die relative Häufigkeit der vorgefundenen Merkmalsausprägungen weiß und schwarz als Wahrscheinlichkeit für das Ziehen von „weiß“ oder „schwarz“ auffassen. Allerdings kann man bei einer erneuten Messung zu einem anderen Wert gelangen. Für eine Definition ist dies also ungeeignet. Allerdings vermuten wir, dass mit wachsendem n die relative Häufigkeit immer näher an der tatsächlichen Wahrscheinlichkeit ist. Eine vernünftige mathematische Definition der Wahrscheinlichkeit muss dies nachbilden (das führt zum Gesetz der großen Zahlen). Dazu benötigen wir ein mathematisches Modell für Zufallsexperimente.

Die Durchführung eines Zufallsexperiments heißt ein **Versuch**, sein Ergebnis ein **Elementarereignis**. Die Menge aller möglichen Versuchsausgänge, also aller Elementarereignisse, heißt **Grundgesamtheit** oder **Elementarereignisraum** und wird üblicherweise mit Ω bezeichnet. Ω entspricht dem Begriff der statistischen Masse in der beschreibenden Statistik. Ein **Ereignis** ist eine Teilmenge der Grundgesamtheit Ω . Mit einem Ereignis fasst man Elementarereignisse zusammen. Wir sagen, dass bei einem Experiment ein Ereignis E eingetreten ist, genau dann, wenn ein Elementarereignis beobachtet wird, das Element von E ist. Als Ereignis sind auch $\emptyset$ und Ω zugelassen. Da in $\emptyset$ keine Elementarereignisse liegen, kann $\emptyset$ auch niemals eintreten. Da in Ω alle Elementarereignisse sind, tritt das Ereignis Ω bei jedem Experiment ein. Man spricht hier einerseits vom **unmöglichen Ereignis** $\emptyset$ und andererseits vom **sicheren Ereignis** Ω .

Beispiel 7.14

Die Grundgesamtheit beim einmaligen Würfeln ist $\Omega := \{1, 2, 3, 4, 5, 6\}$. Das Ereignis eines geraden Wurfs ist dann $G := \{2, 4, 6\}$, das Ereignis eines ungeraden Wurfs ist $U := \{1, 3, 5\}$. ■

Bei einer endlichen Grundgesamtheit Ω mit N Elementen gibt es genau 2^N unterschiedliche Teilmengen und damit Ereignisse (siehe Kapitel 1.1).

Häufig möchte man nicht alle möglichen Ereignisse eines Zufallsexperiments betrachten, sondern nur eine Teilmenge $\mathcal{U}$ der Menge aller Ereignisse. Dies ist insbesondere dann der Fall, wenn die Grundgesamtheit unendlich viele Elemente hat. Diese Menge $\mathcal{U}$ bezeichnen wir als **Ereignis-Algebra**. Damit wir später den Ereignissen Wahrscheinlichkeiten zuordnen können, muss dann $\mathcal{U}$ eine gewisse Struktur haben, die bei den hier diskutierten Experimenten immer gegeben ist (siehe Kasten).

Hintergrund: Ereignis-Algebren

Die folgende Definition der Ereignis-Algebra ist für die mathematische Theorie vor allem dann wichtig, wenn die Grundgesamtheit überabzählbar viele Elemente hat. Damit man dann Wahrscheinlichkeiten sinnvoll definieren kann (mittels Maß- und Integrati-

onstheorie), muss man die Struktur der Ereignisalgebra wie in der folgenden Definition festlegen. Diese spielt für die praktische Anwendung der Wahrscheinlichkeitstheorie aber keine besondere Rolle.

Definition 7.14 (Ereignis-Algebra)

Sei die Grundgesamtheit Ω eine nicht-leere Menge und $\mathcal{U}$ eine Menge von Ereignissen, für die folgende Bedingungen erfüllt sind:

- $\mathcal{U} \neq \emptyset$, es gibt also Ereignisse.
- Für jedes $E \in \mathcal{U}$ gilt: $C_\Omega E \in \mathcal{U}$ (mit $CE := C_\Omega E := \{\omega \in \Omega : \omega \notin E\}$), d. h., mit jedem Ereignis E ist auch das Gegenereignis CE (d. h. E tritt nicht ein) vorhanden.
- Jede (höchstens abzählbare) Vereinigung von Mengen $E_k \in \mathcal{U}$, $k \in \mathbb{N}$, gehört selbst wieder zu $\mathcal{U}$. Man hat also auch das Ereignis zur Verfügung, dass mindestens eines der Ereignisse E_k eintritt.

Dann heißt $\mathcal{U}$ eine **Ereignis-Algebra** (**σ -Algebra**) über Ω .

Man beachte, dass für ein beliebiges Ereignis $E \in \mathcal{U}$ gilt:

$$\Omega = E \cup CE \in \mathcal{U}, \quad \emptyset = C\Omega \in \mathcal{U}.$$

Das unmögliche und das sichere Ereignis gehören also in jedem Fall zur Ereignis-Algebra.

Die Wahrscheinlichkeitsrechnung kann auch ohne Ereignis-Algebren verstanden werden:

- Die Potenzmenge von Ω , das ist die Menge aller Teilmengen von Ω (siehe Seite 2), erfüllt die Bedingungen der Definition 7.14. Hier können und werden wir also alle denkbaren Ereignisse betrachten.
- Bei einer unendlichen Grundgesamtheit Ω benötigt man bisweilen aber „kleinere“ Ereignisalgebren als die Potenzmenge. In diesen Fällen (z. B. wenn wir Folgen von untereinander unabhängigen Wahrscheinlichkeitsexperimenten betrachten), kann $\mathcal{U}$ geeignet konstruiert werden.

7.2.2 Wahrscheinlichkeit und Satz von Laplace

Der Ausgang eines einzelnen Zufallsexperiments ist völlig offen. Die Erfahrung zeigt aber, dass man durch n -fache Wiederholung eine Gesetzmäßigkeit erkennen kann. Man beobachtet, dass die relative Häufigkeit des Eintretens eines Ereignisses E für $n \rightarrow \infty$ gegen einen festen Wert $p \in [0, 1]$ konvergiert:

$$p = \lim_{n \rightarrow \infty} \frac{\text{„Anzahl des Eintretens des Ereignisses } E\text{“}}{n}. \quad (7.10)$$

Damit liegt es nahe, dem Ereignis E die Wahrscheinlichkeit p zuzuordnen: $P(E) = p$. Die folgende Definition gibt die Eigenschaften relativer Häufigkeiten wieder:

Definition 7.15 (Axiome des Wahrscheinlichkeitsmaßes)

Mit den zuvor eingeführten Bezeichnungen sei P eine Abbildung, die jedem Ereignis eine Zahl aus $[0, 1]$ zuordnet und die beide folgenden Axiome erfüllt:

- Die Wahrscheinlichkeit, dass bei einem Wahrscheinlichkeitsexperiment irgendwas herauskommt, ist auf eins normiert (also auf $\frac{100}{100} = 100\%$): $P(\Omega) = 1$.
- Für jede endliche oder unendliche Folge von Ereignissen E_k , die paarweise disjunkt (d. h. elementfremd, $E_i \cap E_k = \emptyset$ für $i \neq k$) sind, gilt die **Additivität**:

$$P(E_1 \cup E_2 \cup E_3 \cup \dots) = P(E_1) + P(E_2) + P(E_3) + \dots$$

P heißt ein **Wahrscheinlichkeitsmaß**.

Wenn wir ein Wahrscheinlichkeitsmaß verwenden, dann gehen wir ab jetzt stets davon aus, dass wir eine Grundgesamtheit Ω und eine zugehörige Menge $\mathcal{U}$ aller relevanten Ereignisse kennen, die eine Ereignis-Algebra ist (siehe Definition 7.14). Man sagt, dass diese Objekte einen **Wahrscheinlichkeitsraum** bilden. Der Vollständigkeit halber sei gesagt, dass die in Definition 7.15 aus Ereignissen gebildeten Mengen dann alle selbst wieder Ereignisse sind.

Vereinigt man zwei disjunkte Ereignisse, so tritt das neue Ereignis ein, wenn entweder das eine oder das andere Ereignis eintritt. Da sie disjunkt sind, können sie nicht zugleich eintreten. Die relative Häufigkeit des vereinigten Ereignisses ist damit gleich der Summe der relativen Häufigkeiten der beiden einzelnen Ereignisse. Dies ist mit dem Axiom zur Additivität nachempfunden. Ab jetzt lösen wir uns von der Anschauung und rechnen mit den Axiomen eines Wahrscheinlichkeitsmaßes, so dass wir für Wahrscheinlichkeiten praktisch anwendbare Sätze zeigen können.

Beispiel 7.15

Beim einmaligen Würfeln mit einem ungezinkten Würfel gilt:

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}.$$

Die Wahrscheinlichkeit eines geraden Wurfs ist $P(\{2, 4, 6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{3}{6} = \frac{1}{2}$, die eines ungeraden Wurfs ist ebenfalls $P(\{1, 3, 5\}) = \frac{1}{2}$. ■

Lemma 7.1 (Rechenregeln für ein Wahrscheinlichkeitsmaß)

Für ein Wahrscheinlichkeitsmaß P sowie Ereignisse A , B und das zu A entgegengesetzte Ereignis $CA := C_\Omega A := \{\omega \in \Omega : \omega \notin A\}$ gelten die Rechenregeln

$$P(CA) = 1 - P(A),$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (7.11)$$

Die Regel (7.11) kann von 2 auf $n \geq 2$ Mengen erweitert werden und heißt dann **Siebformel von Sylvester und Poincaré**.

Beweis: Es ist $1 = P(\Omega) = P(A \cup \mathcal{C}A) = P(A) + P(\mathcal{C}A)$, so dass die erste Aussage folgt. Weiter ist mit $A \setminus B = A \cap \mathcal{C}B = \{\omega \in A : \omega \notin B\}$ (siehe Kapitel 1.1):

$$\begin{aligned} P(A \cup B) &= P((A \setminus B) \cup (B \setminus A) \cup (A \cap B)) = P(A \setminus B) + P(B \setminus A) + P(A \cap B) \\ &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) + P(A \cap B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned} \quad \square$$

Die Wahrscheinlichkeit (Elementarwahrscheinlichkeit) von Elementarereignissen $\omega \in \Omega$ ist erklärt über $P(\{\omega\})$. Ist Ω endlich, so ist P vollständig erklärt über die Elementarwahrscheinlichkeiten $P(\{\omega\})$. Wegen $P(E) = \sum_{\omega \in E} P(\omega)$ nennt man die Abbildung $\omega \mapsto P(\{\omega\})$ **Zähldichte**. Besonders wichtig ist der Fall, bei dem (wie oben) alle Elementarwahrscheinlichkeiten gleich sind:

Beispiel 7.16 (Würfelexperiment)

Das folgende Würfelexperiment wird uns durch die nächsten Abschnitte der Wahrscheinlichkeitsrechnung begleiten. Es ist kein Beispiel aus dem Berufsalltag, dafür aber viel leichter zu durchschauen als „echte“ Probleme: Mit einem Würfel wird zweimal gewürfelt. Dieses Zufallsexperiment lässt sich beschreiben über $\Omega := \{(i, k) : 1 \leq i \leq 6, 1 \leq k \leq 6\}$. Ist der Würfel nicht gezinkt, ist jedes Ergebnis gleich wahrscheinlich. Wir lassen als Ereignisse alle Teilmengen von Ω zu, so dass die Menge der Ereignisse die Potenzmenge $\mathcal{P}(\Omega)$ ist und definieren P über $P(\{(i, k)\}) := \frac{1}{|\Omega|} = \frac{1}{36}$. $E := \{(i, 1) : 1 \leq i \leq 6\}$ ist das Ereignis, dass mit dem zweiten Wurf eine 1 gewürfelt wurde. $P(E) = \frac{6}{36} = \frac{1}{6}$. ■

Das Würfelbeispiel ist ein **Laplace-Experiment**:

Satz 7.4 (Laplace-Experiment)

Die Grundgesamtheit Ω sei endlich, und alle Elementarereignisse mögen die gleiche Wahrscheinlichkeit haben. Dann gilt für jedes Ereignis E :

$$P(E) = \frac{|E|}{|\Omega|} = \frac{\text{Anzahl der Elementarereignisse in } E}{\text{Anzahl aller möglichen Elementarereignisse}}.$$

Wir wissen quasi alles über dieses Experiment und sind nicht auf Beobachtungen angewiesen. Die Wahrscheinlichkeit ergibt sich aus der relativen Häufigkeit der Elementarereignisse innerhalb eines Ereignisses.

Beweis: Sei p die gemeinsame Wahrscheinlichkeit aller Elementarereignisse und $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$. Dann ist

$$1 = P(\Omega) = P\left(\bigcup_{k=1}^N \{\omega_k\}\right) = \sum_{k=1}^N P(\{\omega_k\}) = N \cdot p,$$

also ist $p = \frac{1}{N}$. Damit ist $P(E) = \sum_{\omega \in E} \frac{1}{N} = \frac{|E|}{N} = \frac{|E|}{|\Omega|}$. □

Um die Wahrscheinlichkeit eines Ereignisses bei Laplace-Experimenten zu berechnen, muss man die Mächtigkeit des Ereignisses bestimmen. Dieses Abzählen von Elementen kann sich bisweilen als recht schwierig erweisen – daher folgt im nächsten Kapitel ein kleiner Exkurs in die Kombinatorik.

Achtung: Ein häufig gemachter Fehler besteht darin, dass man bei einem Experiment alle möglichen Ergebnisse (Elementarereignisse) zählt und dann einfach annimmt, dass jedes Elementarereignis gleich wahrscheinlich ist. Das ist aber oft nicht so! Diese Situation stand auch am Anfang der Wahrscheinlichkeitstheorie. In dem Fall, dass ein Würfelspiel vorzeitig abgebrochen wird, sollen die Spieler einen Gewinnanteil erhalten, der ihrer Gewinnwahrscheinlichkeit zum Zeitpunkt des Abbruchs entspricht (von Randow, 2005, S. 24). Konkret würfeln zwei Spieler A und B in jeder Runde. Derjenige mit der höheren Augenzahl erhält einen Punkt (wenn die gleiche Augenzahl vorliegt, wird nochmal gewürfelt). Die Wahrscheinlichkeit, einen Punkt zu erhalten, ist pro Spieler und Runde $\frac{1}{2}$. Derjenige, der zuerst fünf Punkte hat, hat gewonnen. Das Spiel wird abgebrochen bei einem Spielstand von 4 (A) zu 3 (B). Welche Gewinnanteile erhalten A und B? 1654 löste Blaise Pascal (1623–1662) diese Frage in einem Briefwechsel an Pierre de Fermat (1601–1665), der als Geburtsstunde der Wahrscheinlichkeitsrechnung angesehen werden kann. Zuvor hatte man die möglichen Spielfortsetzungen gezählt:

- A würfelt eine höhere Zahl als B und gewinnt.
- B würfelt eine höhere Zahl als A und gleicht aus zu 4:4.
 - A gewinnt den nächsten Wurf und damit das Spiel.
 - B gewinnt den nächsten Wurf und das Spiel.

A hat also zwei Gewinnmöglichkeiten, B nur eine. Damit hatte man $\frac{2}{3}$ des Einsatzes an A und $\frac{1}{3}$ des Einsatzes an B gegeben. Dies entspricht aber nicht den Gewinnwahrscheinlichkeiten, da die drei Ausgänge unterschiedlich wahrscheinlich sind. Sie sind keine Elementarereignisse eines Laplace-Experiments. Dies hat Pascal festgestellt und die Wahrscheinlichkeiten berechnet. Wir machen das auf Seite 1013.

Beispiel 7.17

Wenn man die Grundgesamtheit des Würfelxperiments (zweimaliger Wurf) anders wählt, hat man ggf. auch kein Laplace-Experiment mehr. Statt

$$\Omega = \{(i, k) : 1 \leq i \leq 6, \quad 1 \leq k \leq 6\}$$

definieren wir nun

$$\Omega^* := \{(k_1, k_2, k_3, k_4, k_5, k_6) : k_1, k_2, \dots, k_6 \in \{0, 1, 2\} \text{ und } k_1 + k_2 + \dots + k_6 = 2\}.$$

Als Elementarereignis betrachtet man dabei nur noch die reinen Anzahlen pro Augenzahl, nicht mehr die Reihenfolge, in der die Augenzahlen gewürfelt werden. Dann ist $P(\{(1, 1, 0, 0, 0, 0)\}) = \frac{2}{36}$ die Wahrscheinlichkeit, dass eine 1 und eine 2 gewürfelt wird. Dagegen ist die Wahrscheinlichkeit $P(\{(2, 0, 0, 0, 0, 0)\}) = \frac{1}{36}$, dass zwei Einsen gewürfelt werden, eine andere. Es hängt also viel davon ab, wie man die Grundgesamtheit modelliert. ■

7.2.3 Kombinatorik

In Kapitel 1.3.2.4 haben wir bereits die Fakultät $n! = 1 \cdot 2 \cdot 3 \cdots n$, $0! := 1$, als die Anzahl der Permutationen von n verschiedenen Objekten kennengelernt, d. h., $n!$ gibt an, auf wie viele verschiedene Weisen man die Zahlen von 1 bis n in einer Liste anordnen kann. Weiterhin wurde in diesem Zusammenhang auch der Binomialkoeffizient

$$\binom{n}{m} := \frac{n!}{(n-m)!m!}$$

eingeführt. Binomialkoeffizient und Fakultät sind die wichtigsten Hilfsmittel, um die Elementarereignisse eines Zufallsexperiments zu zählen.

Wir betrachten vier verschiedene Szenarien, bei denen m Kugeln aus einer Urne mit n unterscheidbaren Kugeln (mit Nummern 1 bis n) gezogen werden:

- a) Die Reihenfolge der gezogenen Kugeln ist wichtig. Werden die gleichen Kugeln in einer anderen Reihenfolge gezogen, so handelt es sich um einen anderen Ausgang (ein anderes Elementarereignis) des Experiments. Man spricht hier von **Variationen**.
- i) Wird nach dem Ziehen die Kugel nicht wieder zurückgelegt, so spricht man von einem **Ziehen ohne Zurücklegen**. Jede Kugel kann im Ergebnis maximal einmal vorkommen. Hier handelt es sich um **Variationen ohne Wiederholung**. Es gibt n Möglichkeiten, die erste Kugel zu ziehen. Da diese nicht zurückgelegt wird, hat man anschließend für die zweite nur noch $n - 1$ Möglichkeiten, für die dritte $n - 2$ usw. Schließlich gibt es beim m -ten Ziehen noch $n - m + 1$ verbleibende Kugeln und damit Möglichkeiten. Insgesamt gibt es damit

$$\frac{n!}{(n-m)!}$$

verschiedene Elementarereignisse.

Beispiel 7.18

Wir haben 365 nummerierte Kugeln und ziehen ohne Zurücklegen m Kugeln. Unter Beachtung der Reihenfolge gibt es $\frac{365!}{(365-m)!}$ Elementarereignisse. ■

- ii) Wird nach dem Ziehen eine Kugel sofort wieder zurückgelegt, bevor die nächste gezogen wird, so spricht man von einem **Ziehen mit Zurücklegen**. Jede Kugel kann im Ergebnis bis zu m -mal vorkommen. Hier handelt es sich um **Variationen mit Wiederholung**.

Beispiel 7.19

Wir ziehen fünf Kugeln mit Zurücklegen aus einer Urne mit 10 Kugeln, die mit 0 bis 9 beschriftet sind und notieren das Ergebnis unter Berücksichtigung der Reihenfolge. Jedes Elementarereignis ist eine Ziffernfolge, die als natürliche Zahl zwischen 0 und 99 999 interpretiert werden kann. Es gibt also 10^5 verschiedene Elementarereignisse. ■

Da man m -mal die Wahl zwischen n Kugeln hat, gibt es allgemein

$$n^m$$

verschiedene Variationen mit Wiederholung.

Beispiel 7.20

m Personen schreiben nacheinander ihren Geburtstag (als Tag 1 bis 365 im Jahr) auf. Es gibt damit genau 365^m verschiedene Ausgänge des Experiments. ■

- b) Die Reihenfolge der gezogenen Kugeln ist unwichtig. Werden die gleichen Kugeln in einer anderen Reihenfolge gezogen, so handelt es sich um den gleichen Ausgang (das gleiche Elementarereignis) des Experiments. Man spricht hier von **Kombinationen**.
- i) Beim Ziehen ohne Zurücklegen handelt es sich um **Kombinationen ohne Wiederholung**.

Beispiel 7.21

Beim Lotto wird eine Kombination von 6 Zahlen aus 49 gezogen. Die Reihenfolge, mit der die Zahlen gezogen werden, spielt keine Rolle. Eine Kugel wird nicht zurückgelegt, jede Zahl kann also höchstens einmal vorkommen.

Die Anzahl der Elementarereignisse ergibt sich aus der Formel für Variationen ohne Wiederholung. Da nun jedoch die Reihenfolge unwichtig ist, müssen wir noch durch die Anzahl der Permutationen dividieren, mit der m Zahlen unterschiedlich angeordnet werden können – also durch $m!$. Wir erhalten wie in Kapitel 1.3.2.4 den Binomialkoeffizienten

$$\frac{n!}{(n-m)!m!} = \binom{n}{m}.$$

Beim Lotto gibt es demnach genau

$$\binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13\,983\,816$$

Elementarereignisse. Geht man davon aus, dass Lotto ein Laplace-Experiment ist, dann ist die Wahrscheinlichkeit, bei einer Ziehung mit einem Tipp sechs Richtige zu bekommen, gleich $\frac{1}{|\Omega|} = 1/13\,983\,816$. Hinsichtlich der Gewinnwahrscheinlichkeit ist es also völlig egal, welche Zahlen man ankreuzt. Im Falle eines Gewinns ist diese Wahl aber dennoch wichtig. Denn Muster oder Geburtstage werden häufiger angekreuzt. Mehrere Gewinner müssen sich den Gewinn teilen. Daher sollte man seltene Zahlenkombinationen wählen. Der erwartete Gewinn ist dann höher. Darauf basiert z. B. das Geschäftsmodell der Firma Faber. Gemäß einer Pressererklärung von Westlotto (9.12.2014), die auf einer Auswertung von 80 Millionen Lottoscheinen basiert, werden in NRW am häufigsten die folgenden Zahlen getippt (absteigende Reihenfolge): 9, 11, 7, 19, 17, 3. Am seltensten wird die 43 angekreuzt, wobei 9 ca. 40 % häufiger als 43 ist. Das Feld der 43 befindet sich in der linken unteren Ecke und wirkt durch eine Diagonale kleiner. ■

ii) Beim Ziehen mit Wiederholung spricht man von **Kombinationen mit Wiederholung**.

Wir ziehen m Kugeln mit Zurücklegen und zählen, wie oft jede der Zahlen 1 bis n im Ergebnis vorkommt. Die Liste dieser Anzahlen (die addiert m ergeben), sei ein Elementarereignis. Die Anzahl dieser Elementarereignisse entspricht der Anzahl der Kombinationen mit Wiederholung. Diese berechnet sich zu

$$\binom{n+m-1}{m} = |\{(x_1, \dots, x_m) \in \{1, 2, \dots, n\}^m : x_1 \leq x_2 \leq \dots \leq x_m\}|.$$

Diese Formel lässt sich so interpretieren: Zu den n Kugeln fügt man $m-1$ weitere von Kugeln $n+1, n+2, \dots, n+m-1$ hinzu. Die Anzahl der Kombinationen mit m Elementen ohne Wiederholung zu dieser erweiterten Menge ist $\binom{n+m-1}{m}$. Jetzt kann aber jede dieser neuen Kombinationen als eine eindeutige Kodierung einer Kombination mit Wiederholung verstanden werden und umgekehrt: Falls in einer Kombination mit Wiederholung das bei Sortierung k -te Element ($k > 1$) schon vorher (an der Stelle $k-1$) vorkommt, dann ersetze es durch $n+k-1$. Zum Beispiel wird bei $n=3$ die Kombination $\{1, 1, 2, 2, 2, 3\}$ kodiert durch $\{1, 4, 2, 6, 7, 3\} = \{1, 2, 3, 4, 6, 7\}$. Das Vorkommen von $n+k-1$ heißt also: An der Stelle k steht der gleiche Wert wie an der Stelle $k-1$. Umgekehrt kann man aus dieser Kodierung auch genau wieder die Kombination mit Wiederholungen zurückgewinnen.

Beispiel 7.22

Wir würfeln einmal mit drei Würfeln. Dann notieren wir, wie oft bei diesem Wurf die Zahlen 1, 2, 3, 4, 5 und 6 auftreten. Wie viele verschiedene Ausgänge (Elementarereignisse) hat dieses Experiment? Wir können dies als dreimaliges

($m = 3$) Ziehen mit Zurücklegen aus einer Menge von $n = 6$ Kugeln auffassen, wobei die Reihenfolge keine Rolle spielt. Die Anzahlformel liefert:

$$\binom{n+m-1}{m} = \binom{8}{3} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2} = 56.$$

Wir können die Anzahl aber auch anders herleiten: Es gibt $\binom{6}{3}$ Möglichkeiten, drei verschiedene Zahlen zu Würfeln. Die Anzahl der Möglichkeiten für zwei verschiedene Zahlen ergibt sich als Produkt von $\binom{6}{2}$ (Auswahl von zwei verschiedenen Zahlen) mit 2 (Welche der beiden Zahlen wird verdoppelt?). Dazu kommen die 6 Möglichkeiten, dass dreimal die gleiche Zahl gezogen wurde. ■

Wir verallgemeinern den Zählansatz aus dem Beispiel: Sei $W_{n,m}$ die Anzahl der Kombinationen mit Wiederholung von m aus n und $W_{n,0} := 1$, so gilt:

$$W_{1,m} = 1, \quad W_{n,m} = \sum_{k=1}^m \binom{n}{k} W_{k,m-k}. \quad (7.12)$$

Denn dabei ist $\binom{n}{k}$ die Anzahl der Kombinationen, k verschiedene Kugeln aus n zu ziehen. Für jede dieser Möglichkeiten können die verbleibenden $m - k$ Plätze mit Wiederholungen aufgefüllt werden. Dabei kann man jeweils aus k Kugeln ohne Berücksichtigung der Reihenfolge und mit Zurücklegen ziehen. Dies sind wieder Kombinationen mit Wiederholung, also $W_{k,m-k}$ Stück. So erhält man als Produkt damit die Zahl aller Möglichkeiten, genau k verschiedene Kugeln beim Ziehen von m Kugeln zu erhalten, und die Gesamtzahl entsteht durch Summation über k . Dieser Ansatz ist vielleicht etwas naheliegender als die Ergänzung um $m - 1$ Kugeln, die wir zuvor verwendet haben.

(7.12) liefert eine Vorschrift, mit der man rekursiv den Wert für $W_{n,m} = \binom{n+m-1}{m}$ aus der Anfangsbedingung $W_{1,m} = 1$ eindeutig berechnen kann.

Beispiel 7.23

Für die Anzahl der Ergebnisse beim dreimaligen Würfeln aus dem vorangehenden Beispiel erhalten wir damit so:

$$\begin{aligned} W_{6,3} &= \sum_{k=1}^3 \binom{6}{k} W_{k,3-k} = 6W_{1,2} + 15W_{2,1} + 20W_{3,0} \\ &= 6 \left[\binom{1}{1} W_{1,1} + \binom{1}{2} W_{2,1} \right] + 15 \left[\binom{2}{1} W_{1,0} \right] + 20W_{3,0} \\ &= 6[1 + 0] + 30 + 20 = 56. \end{aligned} \quad \blacksquare$$

Beispiel 7.24

Wie wahrscheinlich ist es, dass in einer Gruppe mit m Personen mindestens zwei am gleichen Tag Geburtstag haben, wenn die Wahrscheinlichkeit für jeden Geburtstag $\frac{1}{365}$ ist?

Reihenfolge	mit Wiederholung	ohne Wiederholung
unwichtig: Kombinationen	$\binom{n+m-1}{m}$	$\binom{n}{m} := \frac{n!}{(n-m)!m!}$
wichtig: Variationen	n^m	$\frac{n!}{(n-m)!}$

Abb. 7.8: Anzahlformeln für Kombinationen und Variationen

Wir bilden ein Modell mit $\Omega = \{(x_1, x_2, \dots, x_m) : x_k \in \{1, \dots, 365\}\}$. Damit können wir als Elementarereignis die Geburtstage der m Personen erfassen (Variationen mit Wiederholung). Wie bereits oben gesehen, gibt es 365^m dieser Elementarereignisse, so dass jeder Ausgang des Experimentes die Wahrscheinlichkeit $\frac{1}{365^m}$ besitzt. Jetzt überlegen wir uns, wie viele Elementarereignisse es gibt, bei denen kein Tag mehrfach vorkommt. Dies sind genau die oben betrachteten $\frac{365!}{(365-m)!}$ Variationen ohne Wiederholung. Die Wahrscheinlichkeit, dass alle Geburtstage verschieden sind, ist also $\frac{365!}{(365-m)!365^m}$. Die komplementäre Wahrscheinlichkeit, dass mindestens zwei den gleichen Geburtstag haben, ist $1 - \frac{365!}{(365-m)!365^m}$. Bereits bei $m = 23$ ist diese Wahrscheinlichkeit größer als 50 %.

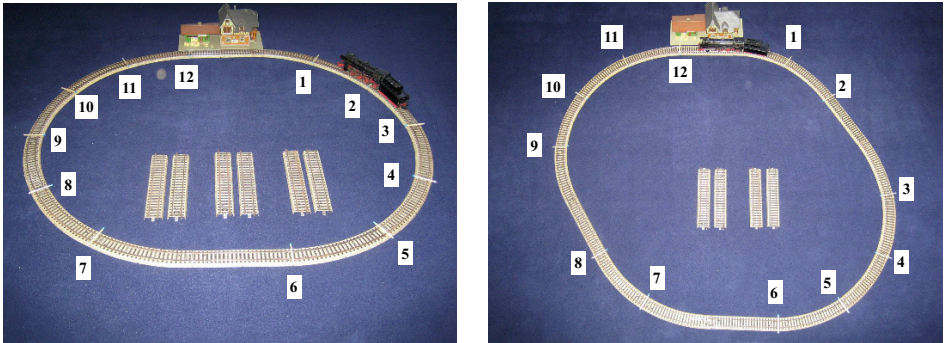


Abb. 7.9: Aufbau verschiedener Gleispläne

Beispiel 7.25

Zu Weihnachten gab es eine Modelleisenbahn mit 12 identischen gebogenen und vielen identischen geraden Gleisen. Mit jeweils 6 gebogenen Gleisen werden zwei Halbkreise gebaut, die mit zwei gegenüberliegenden geraden Gleisen zu einem Oval zusammengesteckt werden. An einem der beiden Geradenstücke wird ein Bahnhof positioniert (siehe Abbildung 7.9 links).

Jetzt soll die Eisenbahn mit den restlichen Geradenstücken erweitert werden. Dazu werden jeweils zwei Stücke an gegenüberliegenden Positionen eingebaut, damit die

Strecke geschlossen bleibt (siehe Abbildung 7.9 rechts). Mit dem Einbau des ersten Geradenpaares können 6 verschiedene Gleispläne gebaut werden: Wenn wir den Anfang jedes Kurvenstücks von 1 bis 12 wie bei einer Uhr durchnummerieren, dann können wir an jeder Position $1 \leq n \leq 6$ ein Geradenstück einfügen und müssen das zweite Stück an der Position $n + 6$ einpassen.

- Wie viele verschiedene Gleispläne können entstehen, wenn man nicht nur ein, sondern zwei Geradenpaare einbaut? Wie viele können bei drei Geradenpaaren entstehen und wie viele bei $n \in \mathbb{N}$?

Für jedes Geradenpaar lässt sich eine Position $1, 2, \dots, 6$ mit Wiederholung auswählen. Die Reihenfolge spielt dabei keine Rolle. Es handelt sich um Kombinationen mit Wiederholung. Bei n Geradenpaaren gibt es daher $\binom{6+n-1}{n}$ verschiedene Gleispläne, speziell für $n = 2$: $\binom{7}{2} = 21$ und für $n = 3$: $\binom{8}{3} = 56$.

Bemerkung: Durch den Bahnhof ist ein Gleis des Ovals so ausgezeichnet, dass durch Drehung und Spiegelung eine andere Modellbahn entsteht.

- Wenn man für das erste Geradenpaar zufällig eine der Positionen $1, 2, \dots, 6$ (alle gleich wahrscheinlich) und ebenso zufällig eine dieser Positionen für das zweite Geradenpaar auswählt, wie wahrscheinlich ist dann das Gleisbild, bei dem jeweils eine Gerade an Position 3 und 4 eingefügt wird?

Mit $\Omega := \{(x_1, x_2) : x_1, x_2 \in \{1, 2, 3, 4, 5, 6\}\}$ hat jedes Elementarereignis die gleiche Wahrscheinlichkeit: $P(\{(x_1, x_2)\}) = \frac{1}{|\Omega|} = \frac{1}{36}$. Für das gesuchte Ereignis gilt: $P(\{(3, 4), (4, 3)\}) = \frac{2}{36} = \frac{1}{18}$. Man beachte, dass nicht alle Gleisbilder gleich wahrscheinlich sind. So ist z. B. $P(\{(1, 1)\}) = \frac{1}{36}$. ■

7.2.4 Unabhängige Ereignisse und bedingte Wahrscheinlichkeiten

Beispiel 7.26 (Fortsetzung des Würfelexperiments)

Sei wie oben $E := \{(i, 1) : 1 \leq i \leq 6\}$ das Ereignis, dass mit dem zweiten Wurf eine 1 gewürfelt wird, und $F := \{(1, k) : 1 \leq k \leq 6\}$ das Ereignis, dass mit dem ersten Wurf eine 1 gewürfelt wird. Die Wahrscheinlichkeit, dass bei Eintreten von F auch E eintritt, ist

$$P(\text{„}E \text{ tritt ein, wenn } F \text{ eintritt“}) = \frac{|\{(1, 1)\}|}{|F|} = \frac{1}{6}.$$

Das Ergebnis des zweiten Wurfs hängt damit nicht vom ersten ab. ■

Allgemein betrachten wir ein Laplace-Experiment mit Ereignissen E und F . Wie wahrscheinlich ist E , wenn man weiß, dass in jedem Fall das Ereignis F eintritt (bzw. F schon eingetreten ist)?

Statt die Grundgesamtheit Ω zu betrachten, kann man hier zur neuen Grundgesamtheit F übergehen. Das Ereignis E wird dann zu $E \cap F$. Die Eintrittswahrscheinlichkeit ist damit

$$\frac{|E \cap F|}{|F|} = \frac{\frac{|E \cap F|}{|\Omega|}}{\frac{|F|}{|\Omega|}} = \frac{P(E \cap F)}{P(F)}.$$

Die rechte Seite ist für jedes Wahrscheinlichkeitsmaß sinnvoll (und wieder durch die Vorstellung von Wahrscheinlichkeiten als relative Häufigkeiten motiviert):

Definition 7.16 (Bedingte Wahrscheinlichkeit)

Für Ereignisse E und F mit $P(F) > 0$ heißt

$$P(E|F) := \frac{P(E \cap F)}{P(F)}$$

die **bedingte Wahrscheinlichkeit** von E unter der Bedingung von F .

Für $P(„E tritt ein, wenn F eintritt“)$ schreiben wir ab jetzt kurz $P(E|F)$. Dies ist die Wahrscheinlichkeit, dass E eintritt, falls bekannt ist, dass F eintritt.

Wir sagen, dass zwei Ereignisse stochastisch unabhängig sind, wenn die Eintrittswahrscheinlichkeit des einen nicht davon abhängt, ob das andere Ereignis eintritt, wenn also z. B. $P(E|F) = P(E)$ ist, d. h., wenn $\frac{P(E \cap F)}{P(F)} = P(E) \iff P(E \cap F) = P(E) \cdot P(F)$ gilt:

Definition 7.17 (Unabhängigkeit von Ereignissen)

Zwei Ereignisse E_1 und E_2 heißen **stochastisch unabhängig** bezüglich eines Wahrscheinlichkeitsmaßes P genau dann, wenn

$$P(E_1 \cap E_2) = P(E_1)P(E_2).$$

Allgemeiner heißen n Ereignisse $E_1, \dots, E_n$ **stochastisch unabhängig** unter P genau dann, wenn die Wahrscheinlichkeit des gemeinsamen Eintretens von 2 bis n dieser Ereignisse gleich dem Produkt der jeweiligen Einzelwahrscheinlichkeiten ist, d. h., wenn

$$P\left(\bigcap_{k \in J} E_k\right) = \prod_{k \in J} P(E_k)$$

für jede Teilmenge $J \subset \{1, \dots, n\}$ mit mindestens zwei Elementen.

Man erhält eine andere Definition, wenn man bei $n > 2$ Ereignissen lediglich die paarweise stochastische Unabhängigkeit fordert.

Für die praktische Anwendung haben wir den wichtigen Multiplikationssatz kennengelernt: Die Wahrscheinlichkeit, dass zwei unabhängige Ereignisse gleichzeitig eintreten, ist gleich dem Produkt der beiden Einzelwahrscheinlichkeiten.

Beispiel 7.27 (Fortsetzung des Würfelexperiments)

Da

$$P(E \cap F) = P(\{(1, 1)\}) = \frac{1}{36} = P(E)P(F),$$

sind die Ereignisse E und F stochastisch unabhängig. ■

Hintergrund: Modellierung unabhängiger Ereignisse

Im Würfelexperiment wird das zweimalige Würfeln durch eine Grundgesamtheit $\Omega = \Omega_1 \times \Omega_2$ als kartesisches Produkt modelliert, wobei $\Omega_1 = \Omega_2 = \{1, 2, 3, 4, 5, 6\}$. Die Ereignisse sind $E = \{1, 2, 3, 4, 5, 6\} \times \{1\}$ und $F = \{1\} \times \{1, 2, 3, 4, 5, 6\}$.

Generell kann man über das kartesische Produkt das Ergebnis unabhängiger Versuche so modellieren, dass unabhängige Ereignisse entstehen. Seien dazu Ω_1 und Ω_2 endliche Grundgesamtheiten, wobei wir jeweils alle Teilmengen als Ereignisse betrachten. Weiter seien P_1 ein Wahrscheinlichkeitsmaß auf Ω_1 und P_2 ein Wahrscheinlichkeitsmaß auf Ω_2 .

$$\Omega := \Omega_1 \times \Omega_2 := \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

Dann ist über die Zähldichte

$$P(\{(\omega_1, \omega_2)\}) := P_1(\{\omega_1\}) \cdot P_2(\{\omega_2\}) \quad (7.13)$$

ein Wahrscheinlichkeitsmaß auf Ω gegeben. Die Additivität folgt sofort aus der Definition über die Zähldichte. Es bleibt zu zeigen: $P(\Omega) = 1$. Dazu:

$$\begin{aligned} P(\Omega) &= \sum_{\omega_1 \in \Omega_1} \sum_{\omega_2 \in \Omega_2} P(\{(\omega_1, \omega_2)\}) = \sum_{\omega_1 \in \Omega_1} \sum_{\omega_2 \in \Omega_2} P_1(\{\omega_1\}) \cdot P_2(\{\omega_2\}) \\ &= \sum_{\omega_1 \in \Omega_1} P_1(\{\omega_1\}) \cdot \sum_{\omega_2 \in \Omega_2} P_2(\{\omega_2\}) = P(\Omega_1) \cdot P(\Omega_2) = 1. \end{aligned}$$

P und Ω sind so konstruiert, dass Ereignisse $E \times \Omega_2$ und $\Omega_1 \times F$ für jedes $E \subset \Omega_1$ und $F \subset \Omega_2$ stochastisch unabhängig sind:

$$\begin{aligned} P([E \times \Omega_2] \cap [\Omega_1 \times F]) &= P(E \times F) = \sum_{\omega_1 \in E} \sum_{\omega_2 \in F} P_1(\{\omega_1\}) \cdot P_2(\{\omega_2\}) \\ &= P_1(E) \cdot P_2(F) = P(E \times \Omega_2) \cdot P(\Omega_1 \times F). \end{aligned}$$

Beispiel 7.28

Die Ausfallwahrscheinlichkeit einer Glühbirne sei 0,01. Schaltet man zwei Glühbirnen parallel, ist die Wahrscheinlichkeit, dass beide ausfallen, $0,01 \cdot 0,01 = 0,0001$. (Wir gehen dabei davon aus, dass die Ausfälle unabhängig voneinander sind – auch wenn es durch den Ausfall eine Wechselwirkung geben sollte.) Mittels Redundanz kann man also die Fehleranfälligkeit signifikant senken. Schaltet man aber beide Birnen hintereinander, führt der Ausfall einer Lampe zum Ausfall beider Lampen. Die Ereignisse sind nicht mehr stochastisch unabhängig. Die Wahrscheinlichkeit ergibt sich nun als Summe der Wahrscheinlichkeiten, dass die erste kaputtgeht, die zweite aber nicht, dass umgekehrt die zweite kaputtgeht, die erste aber nicht, und dass beide kaputtgehen: $2 \cdot 0,01 \cdot 0,99 + 0,01 \cdot 0,01 = 0,0199$. Die Fehlerwahrscheinlichkeit hat sich gegenüber einer einzelnen Lampe fast verdoppelt. ■

Beispiel 7.29

Gegeben sei eine Dezimalzahl mit unendlich vielen Nachkommastellen, die unabhängig voneinander zufällig aus den Ziffern 0 bis 9 ausgewählt seien. Wie wahrscheinlich ist es, dass Ihr Geburtstag (in der Form TTMMJJJJ) irgendwo am Stück in den Nachkommastellen auftritt?

Die Wahrscheinlichkeit, dass sich Ihr Geburtstag (acht Ziffern) aus zehn zufälligen Ziffern ergibt, ist $\frac{1}{10^8}$ (Variationen mit Wiederholung). Die Wahrscheinlichkeit, dass an keiner $8k$ -ten Nachkommastelle, $1 \leq k \leq n$, der Geburtstag steht, ist aufgrund der Unabhängigkeit der n Ereignisse $(1 - \frac{1}{10^8})^n$ und wird damit umso kleiner, je größer n wird: $\lim_{n \rightarrow \infty} (1 - \frac{1}{10^8})^n = 0$. Damit ist die Wahrscheinlichkeit, bereits an einer solchen ($8k$ -ten) Stelle den Geburtstag zu finden, bereits 100 % und damit fast sicher. Dies gilt dann erst recht für beliebige Startstellen. Es ist sogar fast sicher, dass der Geburtstag unendlich oft auftritt. Da es sich um Zufall handelt, beachte man, dass man durchaus beobachten könnte, dass der Geburtstag gar nicht auftritt, z. B. bei der Zahl 0,1. Nur gibt es eben extrem viel mehr Zahlen, bei denen die Ziffernkombination zu finden ist. ■

Beispiel 7.30 (Quantencomputer)

In der klassischen Informatik ist die kleinste Informationseinheit das Bit, das entweder eins oder null ist. Bei Quantencomputern, die die Quantentheorie der Physik ausnutzen, wird in einem Bit eine komplexe Linearkombination (Überlagerung) der „Zustände“ $|0\rangle$ und $|1\rangle$ gespeichert. Hier verwenden wir die aus der Quantenmechanik übliche Klamberschreibweise. $\{|0\rangle, |1\rangle\}$ ist Basis eines zweidimensionalen komplexen Vektorraums. Nur solche Linearkombinationen $\alpha \cdot |0\rangle + \beta \cdot |1\rangle$ sind als Wert eines Bits erlaubt, für die $(\alpha, \beta) \in \mathbb{C}^2$ mit $|\alpha|^2 + |\beta|^2 = 1$ gilt. Die Koeffizienten α und β sind aber leider nicht auslesbar. Ein Bit lässt sich nur „messen“. Beim Messen wird allerdings die Linearkombination aufgelöst und der echte Zufall schlägt zu. Auf der Grundgesamtheit $\Omega = \{|0\rangle, |1\rangle\}$ führt die Messung mit der Wahrscheinlichkeit $P(\{|0\rangle\}) = |\alpha|^2$ zum Wert $|0\rangle$, und $|1\rangle$ wird mit Wahrscheinlichkeit $P(\{|1\rangle\}) = |\beta|^2$ gemessen. Mehr er-

fahren wir nicht über das Bit. Technisch ist es auch nicht möglich mehrfach zu messen. Die Wahrscheinlichkeiten $|\alpha|^2$ und $|\beta|^2$ lassen sich also nicht bestimmen, erst recht nicht die komplexen Zahlen α und β .

Zum Glück lässt sich aber vor einer Messung mit der kompletten Linearkombination rechnen. Technisch lässt sich die Multiplikation von unitären Matrizen $\mathbf{A}$ (siehe Kasten auf Seite 176) mit den Vektoren $(\alpha, \beta)^\top$ realisieren, so dass der Wert eines Bits von $\alpha \cdot |0\rangle + \beta \cdot |1\rangle$ in $\alpha' \cdot |0\rangle + \beta' \cdot |1\rangle$ mit $(\alpha', \beta')^\top = \mathbf{A} \cdot (\alpha, \beta)^\top$ übergeht. Da die Matrix unitär ist, gilt auch $|\alpha'|^2 + |\beta'|^2 = 1$ (vgl. Folgerung 3.2 auf Seite 497), so dass ein neuer wohldefinierter Wert des Bits entsteht.

Durch formale Multiplikation lassen sich die Werte von zwei einzelnen Bits in den Wert eines 2 Bit-Registers überführen:

$$\begin{aligned} & (\alpha \cdot |0\rangle + \beta \cdot |1\rangle) \cdot (\alpha' \cdot |0\rangle + \beta' \cdot |1\rangle) \\ & \text{,, = “ } \alpha\alpha'|0\rangle|0\rangle + \alpha\beta'|0\rangle|1\rangle + \beta\alpha'|1\rangle|0\rangle + \beta\beta'|1\rangle|1\rangle. \end{aligned}$$

Das ist hier eine rein formale Rechnung, dahinter steckt eigentlich das Tensorprodukt von Vektorräumen (ähnlich dem Kreuzprodukt von Mengen). Wir beobachten, dass

$$|\alpha\alpha'|^2 + |\alpha\beta'|^2 + |\beta\alpha'|^2 + |\beta\beta'|^2 = (|\alpha|^2 + |\beta|^2) \cdot (|\alpha'|^2 + |\beta'|^2) = 1 \cdot 1 = 1$$

ist. Damit können wir die vier Faktoren $|\alpha\alpha'|^2$, $|\alpha\beta'|^2$, $|\beta\alpha'|^2$ und $|\beta\beta'|^2$ als Wahrscheinlichkeiten für die neuen Basisvektoren $|00\rangle$, $|01\rangle$, $|10\rangle$ und $|11\rangle$, die den klassischen Werten eines 2 Bit-Wortes entsprechen, verwenden. Sind die Werte der beiden verknüpften Bits stochastisch unabhängig, dann ergibt sich die Wahrscheinlichkeit eines Werts des 2-Bit-Registers zu dieser Rechnung passend als Produkt der Einzelwahrscheinlichkeiten, z. B. ist für $|00\rangle$ die Wahrscheinlichkeit $|\alpha\alpha'|^2 = |\alpha|^2|\alpha'|^2$.

Der Wert des Quanten-2-Bit-Registers ist

$$\alpha\alpha'|00\rangle + \alpha\beta'|01\rangle + \beta\alpha'|10\rangle + \beta\beta'|11\rangle.$$

Wird er gemessen, dann wird auf der Grundgesamtheit $\Omega = \{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$ ein Zufallsexperiment ausgeführt, und mit Wahrscheinlichkeit $P(\{|00\rangle\}) = |\alpha\alpha'|^2$ wird $|00\rangle$ beobachtet usw. Auf die gleiche Weise werden durch formale Multiplikation Register mit n Bits beschrieben. Diese können dann analog zu einem Bit über unitäre $2^n \times 2^n$ -Matrizen manipuliert werden. Register können aber auch Werte annehmen, die sich nicht mit der stochastischen Unabhängigkeit der einzelnen Bits erklären lassen. Hier sind die Werte der Bits miteinander verknüpft, man spricht von Verschränkung. Der Wert eines Bits beeinflusst den Wert eines anderen. Beispielsweise gibt es zu $\frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$ keine Lösung des Gleichungssystems

$$\alpha\alpha' = 0 \wedge \alpha\beta' = \frac{1}{\sqrt{2}} \wedge \beta\alpha' = \frac{1}{\sqrt{2}} \wedge \beta\beta' = 0.$$

So können n -Bit Quantenregister viel mehr Information aufnehmen als n stochastisch unabhängige Quantenbits oder n klassische Bits. Zudem können mit verschränkten Bits Informationen teleportiert werden, d.h. ohne die Beschränkung durch die Lichtgeschwindigkeit übertragen werden.

Trotz der Umkehrbarkeit jeder Matrixmultiplikation (unitäre Matrizen sind invertierbar) erhält man überraschender Weise über die Matrixmultiplikationen ein Maschinenmodell, das genau so universell ist, wie die klassischen Computer. Einige Probleme lassen sich mit Quantencomputern sogar effizienter realisieren, z. B. die Primfaktorzerlegung, die zum Brechen von gängigen Verschlüsselungen benötigt wird. Geheimdienste finden Quantencomputer daher sicher sehr interessant. ■

Satz 7.5 (Totale Wahrscheinlichkeit)

Seien $E_1, E_2, \dots, E_n$ paarweise disjunkte Ereignisse mit $P(E_k) > 0$, $1 \leq k \leq n$, und $\Omega = \bigcup_{k=1}^n E_k$. Die Ereignisse bilden also eine Zerlegung von Ω (wie beim Würfeln die Ereignisse „gerader Wurf“ und „ungerader Wurf“). Dann gilt für jedes Ereignis A :

$$P(A) = \sum_{k=1}^n P(A|E_k)P(E_k). \quad (7.14)$$

Die **totale Wahrscheinlichkeit** (Wahrscheinlichkeit ohne Nebenbedingungen) von A kann also aus den bedingten Wahrscheinlichkeiten berechnet werden, indem man für jedes Ereignis E_k die bedingte Wahrscheinlichkeit $P(A|E_k)$ (Wahrscheinlichkeit, dass A eintritt, wenn E_k eintritt) mit der Wahrscheinlichkeit des Eintretens von E_k multipliziert (gewichtet) und dann alle Werte aufsummiert. Das hilft, wenn man die Wahrscheinlichkeit von A nur unter gewissen Voraussetzungen kennt und daraus die Gesamtwahrscheinlichkeit von A berechnen möchte.

Beweis: In (7.14) wird A lediglich in elementfremde Mengen $A \cap E_k$ zerlegt, die vereinigt A ergeben. Deren Einzelwahrscheinlichkeiten addieren sich zu $P(A)$:

$$\begin{aligned} \sum_{k=1}^n P(A|E_k)P(E_k) &\stackrel{\text{Def. 7.16}}{=} \sum_{k=1}^n \frac{P(A \cap E_k)}{P(E_k)} P(E_k) = \sum_{k=1}^n P(A \cap E_k) \\ &\stackrel{\text{Def. 7.15}}{=} P\left(\bigcup_{k=1}^n (A \cap E_k)\right) = P\left(A \cap \bigcup_{k=1}^n E_k\right) = P(A \cap \Omega) = P(A). \quad \square \end{aligned}$$

Beispiel 7.31

Wir müssen noch die Gewinnwahrscheinlichkeiten bestimmen, die Pascal für das abgebrochene Würfelspiel (siehe Seite 1002) ermittelt hat. Hier liefert die achte Runde die Ereignisse:

- Ereignis E_1 : A würfelt eine höhere Zahl als B und gewinnt, $P(E_1) = \frac{1}{2}$.

- Ereignis E_2 : B würfelt eine höhere Zahl als A und gleicht aus zu 4:4, $P(E_2) = \frac{1}{2}$.

E_1 und E_2 sind disjunkte Ereignisse. Außerdem betrachten wir

- Ereignis F_1 : A gewinnt den neunten Wurf und damit das Spiel: $P(F_1|E_1) = 0$ und $P(F_1|E_2) = \frac{1}{2}$.
- Ereignis F_2 : B gewinnt den neunten Wurf und das Spiel: $P(F_2|E_1) = 0$ und $P(F_2|E_2) = \frac{1}{2}$.

Nach dem Satz über die totale Wahrscheinlichkeit ist

$$P(F_1) = P(F_1|E_1)P(E_1) + P(F_1|E_2)P(E_2) = 0 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

$$P(F_2) = P(F_2|E_1)P(E_1) + P(F_2|E_2)P(E_2) = 0 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Spieler A gewinnt, falls eines der disjunkten Ereignisse E_1 oder F_1 eintritt und damit mit Wahrscheinlichkeit $\frac{3}{4}$. Spieler B gewinnt nur, falls F_2 eintritt, also mit Wahrscheinlichkeit $\frac{1}{4}$. Diese Wahrscheinlichkeiten stimmen nicht mit dem Teilungsverhältnis $\frac{2}{3} : \frac{1}{3}$ überein, das vor Pascal verwendet wurde. ■

Beispiel 7.32 (Test auf eine Krankheit)

In einem geeigneten Wahrscheinlichkeitsraum (Ω sei z. B. die Menge der Menschen in Krefeld) hat man die folgenden vier Ereignisse:

- K sei die Menge der kranken Menschen, $\bar{K}$ die der gesunden,
- T sei die Menge der Menschen mit einem positiven Testergebnis, $\bar{T}$ die Menge mit einem negativen Ergebnis.

Die Qualität eines Krankheitstests ist bestimmt durch die Werte von

- $P(T|K)$, also der Wahrscheinlichkeit, dass der Test bei einem Kranken die Krankheit auch erkennt (**Sensitivität**) und
- $P(\bar{T}|\bar{K})$, also der Wahrscheinlichkeit, dass ein Gesunder auch als gesund erkannt wird (**Spezifität**).

Oft kann man nicht Sensitivität und Spezifität gemeinsam optimieren. Daher sind Testergebnisse generell mit Vorsicht zu beurteilen. Wichtig ist auch die Prävalenz, d. h. die Wahrscheinlichkeit, überhaupt an einer Krankheit zu erkranken, also $P(K)$. Die Wahrscheinlichkeit eines positiven Tests ist

$$\begin{aligned} P(T) &= P(T|K)P(K) + P(T|\bar{K})P(\bar{K}) \\ &= [1 - P(\bar{T}|\bar{K})]P(K) + P(T|\bar{K})P(\bar{K}) \\ &= \underbrace{[1 - P(\bar{T}|\bar{K})]}_{\text{Spezifität}} \cdot \underbrace{(1 - P(K))}_{\text{Prävalenz}} + \underbrace{P(T|K)}_{\text{Sensitivität}} \cdot \underbrace{P(K)}_{\text{Prävalenz}}. \end{aligned}$$

Sei z. B. $P(K) = 0,01$ (1 % ist krank), die Sensitivität $P(T|K) = 0,99$ sei hoch und die Spezifität $P(CT|CK) = 0,8$. Dann ist $P(T) = (1 - 0,8)(1 - 0,01) + 0,99 \cdot 0,01 = 0,21 \cdot 0,99 = 0,2079$, also mehr als 20 %. Die Krankheit tritt aber nur bei 1 % auf. ■

Zweifelt man die Hypothese an, dass ein Mensch gesund ist, so liefert der Test in Abhängigkeit von den beiden Wahrscheinlichkeiten also ein Ergebnis, das fehlerhaft sein kann.

- Trifft die Hypothese zu, d. h., der Mensch ist tatsächlich gesund, aber der Test diagnostiziert eine Krankheit (und ist in diesem Sinne falsch positiv), so spricht man von einem **Fehler erster Art** (vgl. Seite 1072). Die Wahrscheinlichkeit dieses Fehlers ist $P(T|CK) = 1 - P(CT|CK)$. Je niedriger die Spezifität ist, desto größer ist die Wahrscheinlichkeit eines Fehlers erster Art.
- Trifft die Hypothese nicht zu, d. h., der Mensch ist tatsächlich krank, aber der Test diagnostiziert keine Krankheit (und ist in diesem Sinne falsch negativ), so spricht man von einem **Fehler zweiter Art**. Die Wahrscheinlichkeit dieses Fehlers ist $P(CT|K) = 1 - P(T|K)$. Je niedriger die Sensitivität ist, desto größer ist die Wahrscheinlichkeit eines Fehlers zweiter Art.

Die Begriffe „Fehler erster Art“ und „Fehler zweiter Art“ sind verwirrend, da sie sich auf eine zu formulierende Hypothese beziehen. Negiert man die Hypothese, so wird aus dem Fehler erster Art ein Fehler zweiter Art und umgekehrt.

Beispiel 7.33

Drei Maschinen produzieren einen Artikel mit unterschiedlichen Fehlerraten. Die relative Häufigkeit von fehlerhaften Teilen beträgt bei Maschine 1: 1 %, bei Maschine 2: 2 % und bei Maschine 3: 3 %. Insgesamt werden 10 %, 40 % und 50 % der Teile mit Maschine 1, 2 bzw. 3 produziert. Wie wahrscheinlich ist es, dass ein zufällig ausgewähltes Teil der Tagesproduktion defekt ist (Ereignis D)? Sei M_1 das Ereignis, dass ein Teil aus Maschine 1 stammt, entsprechend seien M_2 und M_3 die Ereignisse, dass der Ursprung Maschine 2 und 3 ist:

$$\begin{aligned} P(D) &= P(D|M_1)P(M_1) + P(D|M_2)P(M_2) + P(D|M_3)P(M_3) \\ &= 0,01 \cdot 0,1 + 0,02 \cdot 0,4 + 0,03 \cdot 0,5 = 0,024. \end{aligned}$$

In der Regel sind also 2,4 % aller Teile defekt. ■

Beispiel 7.34

In einer Urne befinden sich 99 weiße und eine schwarze Kugel. Zwei Spieler ziehen abwechselnd mit verbundenen Augen Kugeln. Der Spieler, der die schwarze Kugel zieht, hat gewonnen. Naheliegend ist die Frage, ob der Starter einen Vorteil hat,

- falls mit Zurücklegen gezogen wird.

■ falls ohne Zurücklegen gezogen wird.

Er hat einen Vorteil, wenn die Kugeln zurückgelegt werden. Hier hat jeder Zug die gleiche Gewinnwahrscheinlichkeit 1 %. Der Starter hat die Gewinnwahrscheinlichkeit

$$\begin{aligned} & 0,01 + [0,99]^2 \cdot 0,01 + [0,99]^4 \cdot 0,01 + [0,99]^6 \cdot 0,01 \dots \\ &= 0,01 \sum_{k=0}^{\infty} ([0,99]^2)^k = \frac{0,01}{1 - [0,99]^2} = 0,5025 \dots > \frac{1}{2}, \end{aligned}$$

die wir mittels der geometrischen Reihe berechnet haben. Er hat aber keinen Vorteil, wenn die Kugeln nicht zurückgelegt werden. Denn hier wächst mit der Spiellänge die Trefferwahrscheinlichkeit. Dass hier $\frac{1}{2}$ herauskommt, sieht man, da zu jedem Zeitpunkt beide Spieler die gleichen Chancen haben, z. B. beim ersten Zug: Spieler A: $\frac{1}{100}$, Spieler B: $\frac{99}{100} \cdot \frac{1}{99} = \frac{1}{100}$. ■

Kombiniert man den Satz über die totale Wahrscheinlichkeit und die Definition der bedingten Wahrscheinlichkeit, so erhält man:

Satz 7.6 (Satz von Bayes (1702–1761))

Seien $E_1, E_2, \dots, E_n$ paarweise disjunkte Ereignisse mit $P(E_k) > 0$, $1 \leq k \leq n$, und $\Omega = \bigcup_{k=1}^n E_k$. Die Ereignisse E_k bilden also wie zuvor eine Zerlegung von Ω . Dann gilt für jedes Ereignis A mit $P(A) > 0$ und $1 \leq k \leq n$:

$$P(E_k|A) = \frac{P(E_k \cap A)}{P(A)} = \frac{\frac{P(A \cap E_k)P(E_k)}{P(E_k)}}{P(A)} \stackrel{(7.14)}{=} \frac{P(A|E_k)P(E_k)}{\sum_{i=1}^n P(A|E_i)P(E_i)}.$$

Die bedingte Wahrscheinlichkeit, dass E_k eintritt, wenn A eingetreten ist, lässt sich umgekehrt ermitteln aus den bedingten Wahrscheinlichkeiten, dass A eintritt, wenn die einzelnen E_k eingetreten sind.

Beispiel 7.35 (Das Ziegenproblem, siehe von Randow (2005))

Bei einer Quiz-Show sind zwei Ziegen und ein Auto hinter je einer Tür versteckt. Der Kandidat soll erraten, wo sich das Auto befindet. Dazu darf er eine Tür raten. Dann öffnet der Moderator eine der beiden verbleibenden Türen, wobei er die Tür so wählt, dass sich dahinter eine Ziege befindet. (Das kann er immer tun, da sich hinter den beiden verbleibenden Türen mindestens eine Ziege verbergen muss). Der Kandidat darf nun erneut zwischen den beiden noch geschlossenen Türen wählen. Lohnt es sich für ihn, sich umzuentcheiden?

In der Kolumne „Ask Marilyn“ des amerikanischen Wochenmagazins „Parade“ erklärte die Journalistin Marilyn vos Savant 1991: „Es ist besser, zu wechseln, da sich die Gewinnchancen dann verdoppeln.“. Und sie hat recht. Beim ersten Raten hat er das Auto mit einer Wahrscheinlichkeit von $\frac{1}{3}$ getroffen. Die Wahrscheinlichkeit, dass

es sich hinter den beiden anderen Türen befindet, ist $\frac{2}{3}$. Daran ändert auch das Öffnen einer dieser Türen durch den Moderator nichts. Denn er kann immer eine Tür öffnen. Nun konzentriert sich aber die $\frac{2}{3}$ -Wahrscheinlichkeit auf die nicht geöffnete Tür. Der Kandidat verdoppelt seine Gewinnchancen, wenn er die Tür wechselt (vgl. Abbildung 7.10). Das wird klarer, wenn man 1 000 Türen betrachtet und der Moderator 998 davon öffnet.

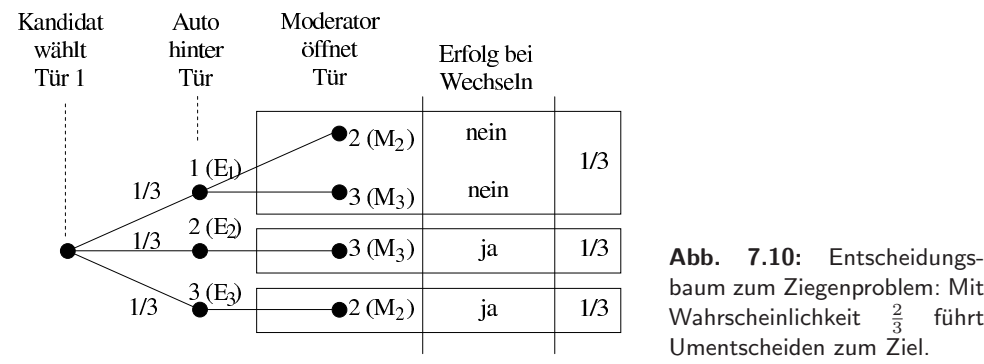


Abb. 7.10: Entscheidungsbaum zum Ziegenproblem: Mit Wahrscheinlichkeit $\frac{2}{3}$ führt Umentscheiden zum Ziel.

Wir können mit dem Satz von Bayes berechnen, wie wahrscheinlich ein Erfolg mit Wechseln in Abhängigkeit von der Reaktion des Moderators ist:

Der Kandidat entscheidet sich für Tür 1 (alle andere Fälle sind analog). E_k sei das Ereignis, dass sich das Auto hinter Tür k verbirgt. E_1 , E_2 und E_3 sind disjunkt mit $E_1 \cup E_2 \cup E_3 = \Omega$. Es ist $P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$. Entscheidet sich der Kandidat unabhängig von der gezeigten Tür um, so ist seine Trefferwahrscheinlichkeit $P(E_2 \cup E_3) = \frac{2}{3}$. Nun sehen wir uns mit dem Satz von Bayes an, ob man noch weitere Informationen aus der vom Moderator gewählten Tür ableiten kann. M_k sei das Ereignis, dass der Moderator Tür k öffnet. Da der Kandidat sich für Tür 1 entschieden hat, ist $P(M_1) = 0$. $P(M_3|E_2) = 1$, da der Moderator in diesem Fall nur Tür 3 öffnen darf. $P(M_3|E_3) = 0$, da der Moderator nicht das Auto zeigen darf. $P(M_2|E_1) + P(M_3|E_1) = 1$, da der Moderator zwischen Tür 2 und Tür 3 wählen kann.

$$\begin{aligned}
 P(E_2|M_3) &= \frac{P(M_3|E_2)P(E_2)}{P(M_3|E_1)P(E_1) + P(M_3|E_2)P(E_2) + P(M_3|E_3)P(E_3)} \\
 &= \frac{1 \cdot \frac{1}{3}}{P(M_3|E_1)\frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1}{1 + P(M_3|E_1)}, \\
 P(E_3|M_2) &= \frac{P(M_2|E_3)P(E_3)}{P(M_2|E_1)P(E_1) + P(M_2|E_2)P(E_2) + P(M_2|E_3)P(E_3)} \\
 &= \frac{1 \cdot \frac{1}{3}}{P(M_2|E_1)\frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{1 + P(M_2|E_1)}.
 \end{aligned}$$

Entscheidet sich der Moderator mit gleicher Wahrscheinlichkeit zwischen den Türen 2 und 3, also $P(M_2|E_1) = P(M_3|E_1) = \frac{1}{2}$, so erhalten wir wie zuvor überlegt die Wahr-

scheinlichkeit $P(E_2|M_3) = P(E_3|M_2) = \frac{2}{3}$. Vielleicht gibt es aber eine Absprache zwischen Moderator und Kandidat. Er könnte sich z. B. immer für Tür 3 entscheiden, falls dies möglich ist. Dann ist $P(M_2|E_1) = 0$ und $P(M_3|E_1) = 1$, so dass $P(E_2|M_3) = \frac{1}{2}$ und $P(E_3|M_2) = 1$. Wählt der Moderator nun die zweite Tür und kennt der Kandidat dessen Vorgehen, so weiß der Kandidat, dass das Auto nur hinter der dritten Tür sein kann. Allerdings ist dann $P(M_2) = P(E_3) = \frac{1}{3}$ und $P(M_3) = 1 - P(M_2) = \frac{2}{3}$. Der Kandidat weiß also nur in einem Drittel der Fälle Bescheid. In $\frac{2}{3}$ der Fälle hat er dagegen nur eine Chance von 50 %. Im Mittel sind die Chancen des Kandidaten bei einer Umentscheidung selbst dann, wenn er die Strategie des Moderators genau kennt, $\frac{2}{3}$. Aus den bedingten Wahrscheinlichkeiten ergibt sich natürlich auch in diesem Fall die bekannte Wahrscheinlichkeit von $P(E_2 \cup E_3) = \frac{2}{3}$:

$$\begin{aligned} P(E_2 \cup E_3) &= P(E_2 \cup E_3|M_2)P(M_2) + P(E_2 \cup E_3|M_3)P(M_3) \\ &= P(E_3|M_2)P(M_2) + P(E_2|M_3)P(M_3) = 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} = \frac{2}{3}. \quad \blacksquare \end{aligned}$$

7.2.5 Zufallsvariablen

7.2.5.1 Diskrete Zufallsvariablen und ihre Verteilung

Statt mit unhandlichen Elementarereignissen oder Ereignissen zu hantieren, möchte man mit Zahlenwerten arbeiten. Daher führt man Zufallsvariablen ein. Eine Zufallsvariable ist eine Funktion (der Name „Variable“ ist irreführend, deckt sich aber später mit der Anschauung), die Elemente der Grundgesamtheit auf Zahlen abbildet. Man betrachtet in der Praxis nur noch die Werte von Zufallsvariablen und nicht mehr die ursprünglichen Ereignisse. Damit braucht man sich in vielen Fällen nicht mit dem Erstellen eines Wahrscheinlichkeitsraums aufzuhalten (den wir daher auch kaum behandelt haben), sondern kann direkt mit bekannten Eigenschaften der Zufallsvariablen rechnen. Die Idee ist ähnlich wie bei der Laplace- oder Fourier-Transformation, bei denen man ein anderes mathematisches Modell wählt, um leichter rechnen zu können. In Abbildung 7.11 ist beispielsweise eine Zufallsvariable dargestellt, die die Augensumme beim zweimaligen Würfeln durch Addition der Einzelergebnisse berechnet. Statt mit Ω kann man nun mit der Menge $\{2, 3, \dots, 12\}$ arbeiten, wobei sich die Wahrscheinlichkeit für jede dieser Zahlen aus der Wahrscheinlichkeit der Elementarereignisse in Ω addiert, die zu dieser Augensumme führen.

Man braucht noch rein technische Zusatzeigenschaften, damit Zufallsvariablen in das mathematische Modell passen und Wahrscheinlichkeiten für ihre Funktionswerte erklärt sind. Für die Anwendung sind diese aber nicht entscheidend, so dass man die folgende Definition auch überspringen kann.

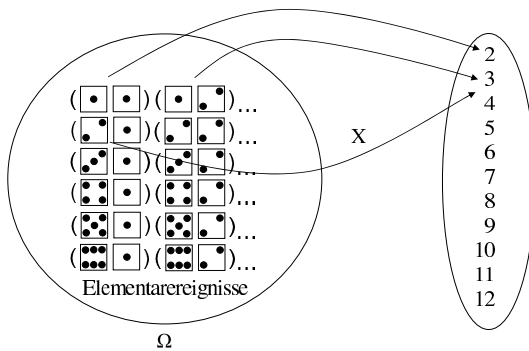


Abb. 7.11: Zufallsvariable X für die Augensumme, die mit der Wahrscheinlichkeit $\frac{1}{36}$ den Wert 2 und mit der Wahrscheinlichkeit $\frac{2}{36}$ den Wert 3 annimmt

Definition 7.18 (Zufallsvariable)

Eine Abbildung $X : \Omega \rightarrow \mathbb{R}$ heißt eine **diskrete Zufallsvariable** (für ein Wahrscheinlichkeitsmaß P) genau dann, wenn

- sie nur endlich oder abzählbar unendlich viele Werte annimmt, d. h., wenn $X(\Omega) := \{X(\omega) : \omega \in \Omega\}$ eine endliche oder abzählbare Menge reeller Zahlen ist. Die Werte einer diskreten Zufallsvariable können durchnummeriert werden (vgl. Kapitel 1.3.2.3), und wir können über alle Werte der Zufallsvariable summieren.
- für jeden Wert x der Abbildung die Wahrscheinlichkeit $P(X = x)$ erklärt ist, dass X den Wert x annimmt. Dies ist dann der Fall, wenn für jedes x aus dem Wertebereich von X die Menge der zugehörigen Elementarereignisse $\{\omega \in \Omega : X(\omega) = x\} = X^{-1}(\{x\})$ (vgl. Definition 1.4 auf Seite 8) ein Ereignis ist, so dass dessen Eintrittswahrscheinlichkeit mit P bekannt ist. Man spricht hier von der **Messbarkeit** von X .

Bemerkungen:

- Ohne die Bedingung, dass die Wertemenge von X höchstens abzählbar ist, spricht man von einer Zufallsvariable (ohne das Attribut „diskret“). Diesen allgemeinen Fall werden wir nur erwähnen, ohne in die Tiefe zu gehen, siehe Kapitel 7.2.9. Prinzipiell ist er aber analog zum diskreten Fall, man muss „nur“ die Summen, die wir noch verwenden werden, durch ein geeignetes Integral ersetzen.
- Die „Messbarkeitsbedingung“ ist nur erforderlich, wenn man nicht alle Teilmengen von Ω als Ereignisse betrachten möchte oder kann. In diesem Fall benötigt man die Definition der Ereignis-Algebra (siehe Seite 999). Sie stellt sicher, dass man auch in dieser Situation die Wahrscheinlichkeit angeben kann, mit der die Zufallsvariable X einen Wert x annimmt.

Statt mit Ereignissen aus Ω können wir nun mit reellen Zahlen und Mengen von reellen Zahlen rechnen. Mit

- $P(X = x)$ bezeichnen wir die Wahrscheinlichkeit des Ereignisses, dass die reelle Zahl x von der Zufallsvariablen X angenommen wird.
- $P(X \leq x)$ ist die Wahrscheinlichkeit, dass ein reeller Wert kleiner oder gleich x angenommen wird.
- $P(X \in A)$ ist die Wahrscheinlichkeit, dass der Wert von X in der Menge $A \subset \mathbb{R}$ liegt.

Entsprechend sind analoge Schreibweisen zu deuten. Unter $X \leq x$ verstehen wir das Ereignis mit den Elementen $\omega \in \Omega$, für die $X(\omega) \leq x$ ist. Die wie logische Ausdrücke aussehenden Terme $X = x$, $X \leq x$ und $X \in A$ beschreiben also Mengen. Soll daher z. B. die Wahrscheinlichkeit des gemeinsamen Eintretens der Ereignisse $X \leq x$ und $X \in A$ angegeben werden, so schreiben wir $P((X \leq x) \cap (X \in A))$. Man findet in der Literatur dagegen auch die gleichbedeutende Notation $P((X \leq x) \wedge (X \in A))$.

Die mit den Kurzschreibweisen angegebenen Wahrscheinlichkeiten können sinnvoll berechnet werden:

Satz 7.7 (Induziertes Wahrscheinlichkeitsmaß)

Mit einer diskreten Zufallsvariable X aus Definition 7.18 ist ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ gegeben, das jeder Teilmenge von $\mathbb{R}$ eine Wahrscheinlichkeit zuordnet. Da es sich aus dem auf Ω gegebenen Wahrscheinlichkeitsmaß P über die Zufallsvariable X berechnet, wird es mit P^X bezeichnet. Für jede Teilmenge $B \subset \mathbb{R}$ berechnet sich $P^X(B)$, indem man die Wahrscheinlichkeit der Menge aller Elementarereignisse aus Ω , die von X in B abgebildet werden, bestimmt. Damit gibt P^X die Wahrscheinlichkeit an, dass X einen Wert aus B annimmt:

$$P^X(B) := P(X \in B) = P(X^{-1}(B)) = P(\{\omega \in \Omega : X(\omega) \in B\}).$$

In der Schreibweise P^X ist das X lediglich als Zusatz und nicht als Exponent zu verstehen.

Beweis: Wir müssen zeigen, dass P^X sinnvoll definiert und tatsächlich ein Wahrscheinlichkeitsmaß ist.

Damit $P(X^{-1}(B))$ wohldefiniert ist, muss $X^{-1}(B)$ ein Ereignis sein. Hier können wir ausnutzen, dass X höchstens abzählbar viele Werte annehmen kann. Denn damit kann X auch höchstens abzählbar viele Werte aus B annehmen. Diese seien $\{x_k \in B : k \in J\}$ für eine Indexmenge $J \subset \mathbb{N}$. Dann ist

$$X^{-1}(B) = X^{-1}(\{x_k \in B : k \in J\}) = \bigcup_{k \in J} X^{-1}(\{x_k\}).$$

Das Urbild von B ist also die Vereinigung der elementfremden Urbilder der abzählbar vielen Elemente von B , die tatsächlich von X angenommen werden. Wegen der

Messbarkeit der Zufallsvariable X ist jedes Urbild $X^{-1}(\{x_k\})$ ein Ereignis. Wegen der Definition der Ereignis-Algebra (Seite 999) ist auch die abzählbare Vereinigung dieser Mengen wieder ein Ereignis, und $X^{-1}(B)$ ist als Argument von P erklärt.

Wir müssen die beiden Eigenschaften des Wahrscheinlichkeitsmaßes aus Definition 7.15 nachrechnen. Dazu benutzen wir, dass P diese Eigenschaften hat.

- Die Wahrscheinlichkeit, dass die Zufallsvariable irgendeine reelle Zahl annimmt, ist eins: $P^X(\mathbb{R}) = P(X^{-1}(\mathbb{R})) = P(\Omega) = 1$.
- Seien $E_1, E_2, \dots$ paarweise disjunkte Teilmengen von $\mathbb{R}$. Dann sind auch die verschiedenen Mengen $X^{-1}(E_1), X^{-1}(E_2), \dots$ der Elementarereignisse, die durch X auf diese Teilmengen abgebildet werden, elementfremd, so dass die Wahrscheinlichkeit ihrer Vereinigung gleich der Summe ihrer Einzelwahrscheinlichkeiten ist:

$$\begin{aligned} P^X(E_1 \cup E_2 \cup \dots) &= P(X^{-1}(E_1 \cup E_2 \cup \dots)) = P(X^{-1}(E_1) \cup X^{-1}(E_2) \cup \dots) \\ &= P(X^{-1}(E_1)) + P(X^{-1}(E_2)) + \dots \\ &= P^X(E_1) + P^X(E_2) + \dots \end{aligned} \quad \square$$

$P^X(B) = P(X \in B)$ gibt also die Wahrscheinlichkeit an, dass beim Experiment ein $\omega \in \Omega$ beobachtet wird, das durch die Zufallsvariable X auf eine reelle Zahl $X(\omega)$ aus B abbildet wird, dass wir also nach Anwendung der Zufallsvariable das Ereignis $B \subset \mathbb{R}$ beobachten. Wir haben damit Wahrscheinlichkeiten auf der „neuen Grundgesamtheit“ $\mathbb{R}$ und können hier als Ereignisse Teilmengen wie z.B. Intervalle betrachten. Jetzt müssen wir uns in den Anwendungen nicht mehr um das ursprüngliche Ω kümmern.

Definition 7.19 (Verteilung einer Zufallsvariable)

Das von der Zufallsvariable X erzeugte Wahrscheinlichkeitsmaß P^X aus Satz 7.7 heißt die **Verteilung** von X . Es gibt an, mit welcher Wahrscheinlichkeit die Zufallsvariable X ihre Werte x annimmt.

Die eingangs eingeführten Schreibweisen $P(X = x)$, $P(X \leq x)$ usw. drücken damit tatsächlich wohldefinierte Wahrscheinlichkeiten aus. Es ist

- $P(X = x) := P(\{\omega \in \Omega : X(\omega) = x\}) = P^X(\{x\}) = P(X^{-1}(\{x\})),$
- $P(X \leq x) := P(\{\omega \in \Omega : X(\omega) \leq x\}) = P^X([-\infty, x]) = P(X^{-1}([-\infty, x])),$
- $P(X \in A) := P(\{\omega \in \Omega : X(\omega) \in A\}) = P^X(A) = P(X^{-1}(A)).$

Nun kann man auch verstehen, warum die Abbildung $X(\omega)$ eine Zufallsvariable genannt wird. In der Schreibweise $P(X = x)$ ist X eine Variable, die zufällig Werte annimmt. Die Wahrscheinlichkeit, dass in X zufällig der Wert x steht, ist $P(X = x)$.

Beispiel 7.36 (Fortsetzung des Würfelexperiments)

Wir definieren eine Zufallsvariable $X_1 : \Omega \rightarrow \{1, 2, \dots, 6\}$ über $X_1((i, k)) := i$. Analog sei $X_2 : \Omega \rightarrow \{1, 2, \dots, 6\}$ mit $X_2((i, k)) := k$. Die Zufallsvariablen liefern also den

Wert des ersten bzw. zweiten Wurfs. $X_1^{-1}(\{i\}) = \{(i, k) : 1 \leq k \leq 6\}$, $X_2^{-1}(\{k\}) = \{(i, k) : 1 \leq i \leq 6\}$, $P(X_1 = i) = P(\{(i, k) : 1 \leq k \leq 6\}) = \frac{1}{6} = P(X_2 = k)$. ■

Beispiel 7.37 (Geometrische Verteilung)

Bei einem Experiment möge ein Ereignis A mit Wahrscheinlichkeit p eintreten. Das Experiment wird nun immer wieder (und unabhängig von den vorausgehenden Versuchen) durchgeführt. Auf einem geeigneten Wahrscheinlichkeitsraum definieren wir eine Zufallsvariable X , die die Nummer des ersten Experiments angibt, bei dem A eintritt:

$$P(X = n) = (1 - p)^{n-1} \cdot p, \quad n \in \mathbb{N},$$

denn in den ersten $n - 1$ Experimenten tritt A mit Wahrscheinlichkeit $1 - p$ nicht ein, dagegen im n -ten Experiment mit Wahrscheinlichkeit p ein. Die Wahrscheinlichkeiten dürfen wegen der Unabhängigkeit der Experimente multipliziert werden. Der Name **geometrische Verteilung** und die Bezeichnung $X \sim G(p)$ leiten sich von der geometrischen Summe ab, denn die Wahrscheinlichkeit, dass A in mindestens einem der ersten n Experimente eintritt, ist

$$\begin{aligned} P(X \leq n) &= \sum_{k=1}^n P(X = k) = \sum_{k=1}^n (1 - p)^{k-1} \cdot p \\ &= p \cdot \sum_{k=0}^{n-1} (1 - p)^k \stackrel{(1.10)}{=} p \cdot \frac{1 - (1 - p)^n}{1 - (1 - p)} = 1 - (1 - p)^n. \end{aligned}$$

Für $n \rightarrow \infty$ strebt die Wahrscheinlichkeit erwartungsgemäß gegen 1. ■

7.2.5.2 Hypergeometrische Verteilung

Wir betrachten folgende Aufgabenstellung: In einer Kiste befinden sich N Teile, von denen $M \leq N$ defekt sind. Nun werden zufällig n Teile der Kiste entnommen. Wie wahrscheinlich ist es, dass davon genau m defekt sind? Die Zufallsvariable X möge die Anzahl der bei dem Experiment gezogenen defekten Teile angeben. Es handelt sich hier um ein „Ziehen ohne Zurücklegen“. Beim Ziehen des ersten Teils ist die Wahrscheinlichkeit, ein defektes Teil zu treffen, genau $p := \frac{M}{N}$. Die Wahrscheinlichkeiten für jeden weiteren Zug hängen jetzt aber von den Ergebnissen der vorangehenden ab.

Wir suchen jetzt $P(X = m)$. Dazu müssen wir die Anzahl der Kombinationen zählen, bei denen genau m Teile defekt sind, und mit der Elementarwahrscheinlichkeit $1/\binom{N}{n}$ multiplizieren. Es gibt genau $\binom{M}{m}$ Kombinationen von m unterschiedlichen defekten Teilen ohne Beachtung der Reihenfolge. Aus den $N - M$ funktionierenden Teilen können $\binom{N-M}{n-m}$ verschiedene Kombinationen für die verbleibenden $n - m$ Teile der Auswahl gebildet werden, so dass es insgesamt $\binom{M}{m} \binom{N-M}{n-m}$ Elementarereignisse mit genau m defekten Teilen gibt. Die Multiplikation ist erlaubt, da wir die Reihenfolge nicht beachten. Insgesamt erhalten wir:

Definition 7.20 (Hypergeometrische Verteilung)

Eine Zufallsvariable X heißt genau dann **hypergeometrisch verteilt** bzw. hat eine **hypergeometrische Verteilung**, wenn

$$P(X = m) = \frac{\binom{M}{m} \binom{N-M}{n-m}}{\binom{N}{n}}.$$

Bezeichnung: $X \sim H(N; M; n)$.

Beispiel 7.38

Die Wahrscheinlichkeit beim Lotto ($n = 6$ aus $N = 49$), genau m Richtige zu haben, kann über die hypergeometrische Verteilung bestimmt werden. Die $M = 6$ gezogenen Kugeln stellen dabei die defekten Teile dar, also

$$P(X = m) = \frac{\binom{6}{m} \binom{43}{6-m}}{\binom{49}{6}}.$$

Damit erhalten wir folgende Wahrscheinlichkeiten:

$$\begin{aligned} P(X = 0) &= \frac{6\,096\,454}{13\,983\,816} \approx 0,436 & P(X = 1) &= \frac{5\,775\,588}{13\,983\,816} \approx 0,413 \\ P(X = 2) &= \frac{1\,851\,150}{13\,983\,816} \approx 0,132 & P(X = 3) &= \frac{246\,820}{13\,983\,816} \approx 0,018 \\ P(X = 4) &= \frac{13\,545}{13\,983\,816} \approx 0,001 & P(X = 5) &= \frac{258}{13\,983\,816} \approx 0 \\ P(X = 6) &= \frac{1}{13\,983\,816} \approx 0. \end{aligned}$$

■

7.2.5.3 Verteilungsfunktion und Dichte**Definition 7.21 (Verteilungsfunktion einer Zufallsvariable)**

Sei X eine diskrete Zufallsvariable. Die Funktion $F^X : \mathbb{R} \rightarrow [0, 1]$ mit

$$F^X(x) := P(X \leq x) = P^X([-\infty, x])$$

heißt die zu P^X gehörende **Verteilungsfunktion**. Zwei Zufallsvariablen heißen genau dann **identisch verteilt**, wenn sie die gleiche Verteilungsfunktion besitzen.

Der Funktionswert $F^X(x)$ der Verteilungsfunktion ist also die Wahrscheinlichkeit, dass die Zufallsvariable einen Wert $\leq x$ annimmt. Die Funktion $F^X(x)$ ist zwangsläufig monoton steigend mit $\lim_{x \rightarrow \infty} F^X(x) = 1$ und $\lim_{x \rightarrow -\infty} F^X(x) = 0$.

Wir haben bereits die empirische Verteilungsfunktion mit den gleichen Eigenschaften kennengelernt. Hier war für reelle Merkmalsausprägungen $x_1, \dots, x_n$

$$F(x) := \frac{\text{Anzahl der Merkmalsausprägungen } x_k \leq x}{n}.$$

Für großes n kann diese relative Häufigkeit genau als die Wahrscheinlichkeit verstanden werden, dass eine Merkmalsausprägung im Intervall $] -\infty, x]$ liegt. Genau wie bei der empirischen Verteilungsfunktion lassen sich auch hier **Quantile** definieren, also Stellen, an denen die Verteilungsfunktion angegebene Wahrscheinlichkeitsmarken übersteigt, siehe Definition 7.5 auf Seite 980.

Beispiel 7.39 (Fortsetzung des Würfelexperiments)

Die Zufallsvariablen X_1 und X_2 besitzen die gleiche Verteilungsfunktion (Treppenfunktion)

$$F^{X_1}(x) = F^{X_2}(x) = \begin{cases} 0, & x < 1, \\ \frac{k}{6}, & k \leq x < k+1 \text{ für } k \in \{1, 2, 3, 4, 5\}, \\ 1 & x \geq 6. \end{cases} \quad \blacksquare$$

Mit P^X hat man ein Wahrscheinlichkeitsmaß, das man auf Intervalle als spezielle Ereignisse anwenden kann. Die Verteilungsfunktion gibt die Wahrscheinlichkeit der speziellen Intervalle $] -\infty, x]$ an. Möchte man unabhängig von einer Zufallsvariable ein Wahrscheinlichkeitsmaß P für reelle Intervalle definieren, kann man dies auch mittels einer nicht-negativwertigen, reellen Funktion f tun, die die Bedingung

$$\int_{-\infty}^{\infty} f(t) dt = 1 \quad (7.15)$$

erfüllt. Eine solche Funktion heißt **(Wahrscheinlichkeits-)Dichte**. Mit ihr führen wir ein Wahrscheinlichkeitsmaß ein vermöge

$$P([a, b]) := \int_a^b f(t) dt.$$

Genauso sind die Wahrscheinlichkeiten der Intervalle $]a, b[$, $]a, b]$ und $[a, b[$ über das Integral $\int_a^b f(t) dt$ festgelegt. Wegen (7.15) ist $P(\mathbb{R}) = \int_{-\infty}^{\infty} f(t) dt = 1$, und die Linearität des Integrals (siehe (2.43)) führt dazu, dass auch die Additivität in der Definition 7.15 des Wahrscheinlichkeitsmaßes erfüllt ist. Insbesondere ist die zugehörige (unabhängig von einer Zufallsvariable definierte) Verteilungsfunktion

$$F(x) := P(]-\infty, x]) = \int_{-\infty}^x f(t) dt \quad (7.16)$$

eine Stammfunktion der Dichte. Die Dichte ist die Ableitung der Verteilungsfunktion.

Beispiel 7.40

An einem Bahnhof fährt ein Zug im Stundentakt. Wenn man zu einem rein zufälligen Zeitpunkt zum Bahnhof gelangt, dann sind alle exakten Wartezeiten zwischen null Minuten und einer Stunde gleich wahrscheinlich. Gibt X die Wartezeit an, so ist bei einer Rechnung in Minuten mit

$$f(x) := \begin{cases} \frac{1}{60} & : x \in [0, 60[\\ 0 & : x < 0 \vee x \geq 60 \end{cases}$$

die Wahrscheinlichkeit, einen Zug nach höchstens $x \geq 0$ Minuten zu bekommen, gleich

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt = \int_0^{\min\{x, 60\}} \frac{1}{60} dt = \frac{\min\{x, 60\}}{60}.$$

Nach einer halben Stunde hat man also mit 50 % Wahrscheinlichkeit einen Zug bekommen. Man nennt X eine **gleichverteilte Zufallsvariable**. ■

Wir haben zuvor mit Zähldichten gearbeitet und die Wahrscheinlichkeit eines Ereignisses durch Summation der Zähldichte erhalten. Hier wird die Wahrscheinlichkeit durch eine „überabzählbar-unendliche Summation“, nämlich durch Integration der Dichte, gewonnen. Hat man (nicht-diskrete) Zufallsvariablen, die wie eine gleichverteilte Zufallsvariable jede reelle Zahl aus einem Intervall als Wert haben können, so gibt man deren Verteilungsfunktionen wie in (7.16) über eine Dichte an.

Achtung: Über eine Dichte ist ein Wahrscheinlichkeitsmaß P definiert, das für alle einelementigen Ereignisse $\{x\}$ den Wert 0 hat (Integration über ein Intervall der Länge 0). Das entspricht ungefähr den relativen Häufigkeiten von Merkmalsausprägungen eines stetigen Merkmals in der beschreibenden Statistik (siehe Seite 971). Die Wahrscheinlichkeit, einen Zug genau nach einer vorgegebenen Zeit mit unendlich vielen Nachkommastellen zu bekommen, ist null. Trotzdem tritt bei einem Experiment zwangsläufig auch ein einelementiges Ereignis ein. Eine Wahrscheinlichkeit 0 bedeutet hier nicht, dass etwas gar nicht passieren kann. Das Ereignis ist lediglich **fast** unmöglich. Ebenso bedeutet eine Wahrscheinlichkeit 1 im Allgemeinen nur, dass ein Ereignis **fast** sicher eintritt. Andererseits kann man bei den Ereignissen $\emptyset$ und Ω auf den Zusatz „fast“ verzichten, da sie nie bzw. in jedem Fall eintreten.

Die neben der Dichte der Gleichverteilung vielleicht wichtigste Dichte ist die **Dichte der Standardnormalverteilung**

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

deren Graph die Gauß'sche Glockenkurve (siehe Abbildung 7.15 auf Seite 1048) ist. Tatsächlich kann man ausrechnen, dass die Fläche unter dem Funktionsgraphen gleich eins ist. Auf die Standardnormalverteilung kommen wir zurück, wenn wir uns den Zentralen Grenzwertsatz ansehen.

7.2.5.4 Stochastische Unabhängigkeit von Zufallsvariablen

Wenn man ein Zufallsexperiment n -fach wiederholt, ist es wichtig, dass die einzelnen Experimente sich nicht gegenseitig beeinflussen. Beschreibt man jedes Einzelexperiment über eine Zufallsvariable und den Ausgang über den Wert, den die Zufallsvariable annimmt (wie im Würfelexperiment), so sind die Zufallsvariablen unabhängig:

Definition 7.22 (Unabhängigkeit von Zufallsvariablen)

Seien $X_1, \dots, X_n$, $n \geq 2$, diskrete Zufallsvariablen auf Ω .

X_1 und X_2 heißen genau dann **stochastisch unabhängig** (bezüglich eines Wahrscheinlichkeitsmaßes P), falls für beliebige Funktionswerte x_1 von X_1 und x_2 von X_2 die Ereignisse $X_1 = x_1$ und $X_2 = x_2$ unter P stochastisch unabhängig sind, d. h., die Funktionswerte beeinflussen sich gegenseitig nicht:

$$P((X_1 = x_1) \cap (X_2 = x_2)) = P(X_1 = x_1) \cdot P(X_2 = x_2).$$

Die n Zufallsvariablen $X_1, \dots, X_n$ heißen genau dann **stochastisch unabhängig**, wenn die Kenntnis der Werte eines Teils der Variablen keinen Einfluss auf die Wahrscheinlichkeiten der Werte der anderen Variablen hat, d. h., wenn für jede Indexmenge $J \subset \{1, \dots, n\}$, $|J| \geq 2$, mit der mindestens zwei Zufallsvariablen ausgewählt werden, und für jeden Wert $x_i \in X_i(\Omega)$, $i \in J$, der mittels J ausgewählten Variablen gilt:

$$P\left(\bigcap_{i \in J} (X_i = x_i)\right) = \prod_{i \in J} P(X_i = x_i),$$

d. h., die Ereignisse $X_i = x_i$ sind stochastisch unabhängig im Sinne der Definition 7.17 auf Seite 1009.

Beispiel 7.41 (Fortsetzung des Würfel experiments)

Die Zufallsvariablen X_1 und X_2 sind stochastisch unabhängig, da für jedes $1 \leq i \leq 6$ und $1 \leq k \leq 6$ gilt:

$$P((X_1 = i) \cap (X_2 = k)) = P(\{(i, k)\}) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = P(X_1 = i) \cdot P(X_2 = k). \quad \blacksquare$$

Das Beispiel lässt sich verallgemeinern. Im folgenden Kasten wird gezeigt, dass man die Ergebnisse unabhängiger Experimente generell über die Werte unabhängiger Zufallsvariablen schreiben kann.

Hintergrund: Darstellung unabhängiger Experimente über unabhängige Zufallsvariablen

Sei $\Omega := \Omega_1 \times \Omega_2$ mit $\Omega_1 = \Omega_2$ sowie P , P_1 und P_2 wie in (7.13) auf Seite 1010. Damit können wir wieder die Ergebnisse zweier unabhängiger Experimente abbilden. Weiter sei X eine Zufallsvariable auf Ω_1 . Dann sind $X_1 : \Omega \rightarrow \mathbb{R}$, $X_1((\omega_1, \omega_2)) := X(\omega_1)$ und $X_2 : \Omega \rightarrow \mathbb{R}$, $X_2((\omega_1, \omega_2)) := X(\omega_2)$ Zufallsvariablen auf Ω (Projektion einer Komponente). Diese sind stochastisch unabhängig:

$$\begin{aligned} P((X_1 = x_1) \cap (X_2 = x_2)) &= P([X^{-1}(\{x_1\}) \times \Omega_2] \cap [\Omega_1 \times X^{-1}(\{x_2\})]) \\ &= P(X^{-1}(\{x_1\}) \times X^{-1}(\{x_2\})) = P_1(X^{-1}(\{x_1\})) \cdot P_2(X^{-1}(\{x_2\})) \end{aligned}$$

$$= P(X_1 = x_1) \cdot P(X_2 = x_2).$$

Wiederholt man ein Experiment völlig unabhängig von der ersten Durchführung, so kann man mittels X_1 das Ergebnis der ersten und mit X_2 das Ergebnis der zweiten Durchführung darstellen. Dies lässt sich auf n -malige Wiederholung (Zufallsvariablen $X_1, \dots, X_n$) und sogar auf abzählbar-unendliche Wiederholung eines Experiments ausdehnen.

7.2.5.5 Binomialverteilung

Wir führen ein Wahrscheinlichkeitsexperiment durch, bei dem zwei Ereignisse A („ A tritt ein“) und $\bar{A}$ („ A tritt nicht ein“) interessieren. Das Ereignis A trete mit der Wahrscheinlichkeit p und $\bar{A}$ mit der Wahrscheinlichkeit $1 - p$ ein. Das Experiment wird n -mal wie schon zuvor im Beispiel zur geometrischen Verteilung (siehe Seite 1022) wiederholt, wobei sich die Experimente wechselseitig nicht beeinflussen mögen. Es stellt sich die Frage, mit welcher Wahrscheinlichkeit das Ereignis A dabei genau k -mal eintritt. Hierbei handelt es sich um ein „Ziehen mit Zurücklegen“. Die Zufallsvariable X möge angeben, wie oft A eingetreten ist. Wir suchen $P(X = k)$.

Wegen der Unabhängigkeit der Einzelexperimente hat jeder mögliche Ausgang der n Experimente, bei dem genau k -mal A beobachtet wird, die Elementarwahrscheinlichkeit $p^k(1 - p)^{n-k}$. Wir müssen also nur noch zählen, wie viele Ausgänge des Experiments die Bedingung $X(\omega) = k$ erfüllen: Auf wie viele Arten kann man k von n Stellen markieren, d. h., wie viele Kombinationen ohne Wiederholung von k Elementen aus der Menge der n Stellen gibt es? Dies sind $\binom{n}{k}$. Damit ist

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Dies ist die **Binomialverteilung**. Man sagt, X ist **binomialverteilt** und $X \sim \text{Bi}(n; p)$. Der Binomische Satz (1.11) auf Seite 64 kann nun als Eigenschaft eines Wahrscheinlichkeitsmaßes interpretiert werden:

$$1 = P(\Omega) = \sum_{k=0}^n P(X = k) = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k}.$$

Der Unterschied zur hypergeometrischen Verteilung besteht im Zurücklegen. Eine Lieferung von N Teilen enthalte M defekte Teile. Bei einer Qualitätskontrolle entnehmen wir n Teile der Lieferung. Die Zufallsvariable X_i , $i = 1, \dots, n$, möge das Ergebnis der i -ten Entnahme ohne Zurücklegen bezeichnen. Dabei bedeutet $X_i = 1$, dass bei der i -ten Entnahme ein defektes Teil gefunden wird, sonst ist $X_i = 0$. Dann sind die X_i

nicht stochastisch unabhängig, da z. B. für die Wahrscheinlichkeit von defekten Teilen in den ersten beiden Zügen gilt (hypergeometrische Verteilung mit $m = n = 2$):

$$\begin{aligned} P((X_1 = 1) \cap (X_2 = 1)) &= P(X_1 + X_2 = 2) = \frac{\binom{M}{2} \binom{N-M}{2-2}}{\binom{N}{2}} \\ &= \frac{M!(N-2)!}{(M-2)!N!} = \frac{M(M-1)}{N(N-1)}. \end{aligned}$$

Andererseits ist für den ersten Zug $P(X_1 = 1) = \frac{M}{N}$, $P(X_1 = 0) = 1 - \frac{M}{N} = \frac{N-M}{N}$, und nach Formel zur totalen Wahrscheinlichkeit gilt für den zweiten Zug

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) + P(X_2 = 1|X_1 = 0)P(X_1 = 0) \\ &= \frac{M-1}{N-1} \frac{M}{N} + \frac{M}{N-1} \frac{N-M}{N} = \frac{M(N-1)}{N(N-1)} = \frac{M}{N}, \end{aligned}$$

so dass

$$P(X_1 = 1) \cdot P(X_2 = 1) = \frac{M}{N} \frac{M}{N} = \frac{M^2}{N^2} \neq \frac{M(M-1)}{N(N-1)} = P(X_1 + X_2 = 2).$$

Beispiel 7.42

Bei der Einführung der Folgen haben wir auf Seite 216 bereits eine binomialverteilte Zufallsvariable kennengelernt. Hier wurde mit einer Monte-Carlo-Methode eine Kreisfläche durch zufälliges Auswählen von Punkten geschätzt. Beschreibt X die Anzahl der Treffer innerhalb des Kreises bei n unabhängigen Versuchen, so ist $\frac{1}{n}X = a_n$. $X \sim \text{Bi}(n; p)$ mit $p = \frac{\pi}{4}$. ■

Beispiel 7.43

Bei einem Multiple-Choice-Test werden zehn Fragen mit jeweils vier Antworten vorgegeben. Davon ist jeweils genau eine Antwort richtig. Zum Bestehen muss man mindestens m Fragen richtig beantwortet haben. Wie wahrscheinlich ist das Bestehen bei rein zufälligem Ankreuzen je einer Lösung?

Die Zufallsvariable X gebe die Anzahl der richtigen Treffer an. Für jede Frage ist die Trefferwahrscheinlichkeit $p = \frac{1}{4}$. Damit ist

$$P(X = k) = \binom{10}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{10-k}.$$

Man besteht also mit Wahrscheinlichkeit $P(X \geq m) = \sum_{k=m}^{10} \binom{10}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{10-k}$. ■

Beispiel 7.44

Wird ein binärer Datenstrom über einen gestörten Kanal übertragen, so interessiert den Nachrichtentechniker die Wahrscheinlichkeit, dass innerhalb eines Blocks von n übertragenen Bits k Bits falsch sind, d.h. ihren Wert von 0 nach 1 oder umgekehrt geändert haben. Kennt man diese Wahrscheinlichkeit, so kann man geeignete Sicherungsmaßnahmen ergreifen, die eine Erkennung oder sogar eine Korrektur von Übertragungsfehlern ermöglichen. Ist p die Wahrscheinlichkeit für eine Störung eines Bits und treten Störungen unabhängig voneinander auf, so ist wieder $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, wobei die Zufallsvariable X die Anzahl der falschen Bits angibt. Auf einem ISDN-Kanal ist z. B. $p \approx 10^{-7}$ und damit sehr klein. Dagegen kann p im Mobilfunk erheblich größer sein. ■

7.2.5.6 Poisson-Verteilung**Beispiel 7.45**

Ein Hersteller von Computerchips sagt, dass durchschnittlich pro Quadratcentimeter Oberfläche $\lambda > 0$ punktförmige Materialfehler vorliegen. Man weiß, dass hier generell nur endlich viele Fehler auftreten, die sich damit an keiner Stelle häufen. Die punktförmigen Fehler sind also selten.

Die Zufallsvariable X möge die tatsächliche endliche Anzahl der Fehler auf einem Quadratcentimeter angeben. Uns interessiert die Wahrscheinlichkeit $P(X = k)$, dass k Fehler vorliegen. Dazu unterteilen wir das Quadrat in n gleich große Abschnitte (z. B. $n = m^2$ Quadrate). Da nur endlich viele Fehler vorliegen, können wir davon ausgehen, dass bei genügend großem $n > \lambda$ in jedem Abschnitt höchstens ein Fehler liegt. Die Wahrscheinlichkeit für einen Fehler in einem Abschnitt ist $p(n) := \frac{\lambda}{n}$. Nun ist $X \approx X_n$, wobei die Zufallsvariable X_n die Anzahl der Abschnitte mit einem Fehler zählt und damit $\text{Bi}(n; p)$ -verteilt ist. Um nun die Unschärfe „genügend großes n “, das von der zufälligen Fehleranordnung abhängt, zu beseitigen, lassen wir nun die Anzahl der Abschnitte gegen Unendlich gehen und erhalten:

$$\begin{aligned} \lim_{n \rightarrow \infty} P(X_n = k) &= \lim_{n \rightarrow \infty} \binom{n}{k} p(n)^k (1-p(n))^{n-k} = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \cdot \underbrace{\lim_{n \rightarrow \infty} \frac{n!}{n^k (n-k)!}}_{=1} \cdot \underbrace{\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n}_{=\exp(-\lambda)} \cdot \underbrace{\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{-k}}_{=1} = \frac{\lambda^k}{k!} \exp(-\lambda). \end{aligned}$$

Hier haben wir verwendet:

$$\lim_{n \rightarrow \infty} \frac{n!}{n^k (n-k)!} = \lim_{n \rightarrow \infty} \frac{n(n-1) \cdots (n-k+1)}{n^k} = \lim_{n \rightarrow \infty} 1 \cdot \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) = 1.$$

Die Wahrscheinlichkeit, dass ein solch seltener, im Durchschnitt λ -fach auftretender Fehler tatsächlich k -mal beobachtet wird, ist also $P(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda)$. ■

Definition 7.23 (Poisson-Verteilung)

Eine Zufallsvariable X , welche die Werte $0, 1, 2, \dots$ annimmt, heißt genau dann **Poisson-verteilt**, d. h. $X \sim \text{Po}(\lambda)$, wenn ihre Verteilung gegeben ist durch ($\lambda > 0$)

$$P(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda).$$

Aufgrund der Taylor-Entwicklung der Exponentialfunktion ist darüber auch tatsächlich eine Verteilung definiert:

$$P^X(\mathbb{R}) = \sum_{k=0}^{\infty} P(X = k) = \exp(-\lambda) \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \exp(-\lambda) \exp(\lambda) = 1.$$

Hintergrund: Generieren von Wahrscheinlichkeitsverteilungen

Benötigt man für eine Simulation Zufallswerte, die einer gewissen Wahrscheinlichkeitsverteilung genügen, so muss man bisweilen Werte eines Zufallszahlengenerators an die gewünschte Verteilung anpassen. Wir nehmen an, dass wir Werte einer Zufallsvariablen X mit einer Verteilungsfunktion F^X haben, aber Zufallswerte einer Zufallsvariable Y mit einer (anderen) Verteilungsfunktion F^Y benötigen. Falls es eine streng monotone Funktion g gibt, so dass Y und $g(X)$ die gleiche Verteilung haben, dann können wir die verfügbaren Zufallswerte mittels g umrechnen. Falls es eine entsprechende Funktion g gibt, dann gilt

$$F^Y(x) = P(Y \leq x) = P(g(X) \leq x) = P(X \leq g^{-1}(x)) = F^X(g^{-1}(x)).$$

Über diese Gleichung können wir die unbekannte Funktion g finden, indem wir „nur“ noch nach g^{-1} auflösen und g als Umkehrfunktion von g^{-1} berechnen. Diese Methode zur Anpassung von Zufallswerten heißt **Inversionsmethode**.

Beispiel 7.46 (Gleichverteilte Punkte)

In eine Kreisscheibe mit Radius $R > 0$ um den Nullpunkt sollen zufällig Punkte eingetragen werden. Das soll gleichverteilt geschehen, d. h., auf jedem Quadratzentimeter der Fläche sollen ungefähr gleich viele Punkte gezeichnet werden. Es stehen dazu zwei stochastisch unabhängige gleichverteilte Zufallsvariablen X_1 und X_2 mit Werten in $[0, 1]$ zur Verfügung mit Verteilungsfunktion

$$F^{X_{1,2}}(x) = P(X_{1,2} \leq x) = \int_0^x 1 \, dt = x \text{ für } x \in [0, 1].$$

Werte von $X_{1,2}$ erhält man üblicherweise mit einem Zufallszahlengenerator, der in vielen Programmiersprachen zur Verfügung steht.

Wie können wir nun mit $X_{1,2}$ sukzessive Punkte in den Kreis einzeichnen? Die erste Lösung besteht darin, einen Punkt

$$(x, y) = (-R + 2RX_1(\omega), -R + 2RX_2(\omega))$$

nur dann einzuzeichnen, wenn er nicht weiter als R vom Nullpunkt entfernt ist. Punkte im Quadrat mit den Kantenlängen $2R$, die nicht im Kreis liegen, werden verworfen. Diese **Rückweisungsmethode** ist nicht schön, da Zufallswerte weggeworfen werden müssen. Eleganter wäre eine Berechnung über Polarkoordinaten:

$$(x, y) = RX_1(\omega) \cdot (\cos(2\pi X_2(\omega)), \sin(2\pi X_2(\omega))).$$

Hier sind die Werte für den Radius und für den Winkel gleichverteilt. Das wollen wir aber gar nicht. Wenn wir damit viele Punkte zeichnen, dann liegen nahe beim Nullpunkt pro Quadratzentimeter viel mehr Punkte als am Rand des Kreises. Wir dürfen zwar den Winkel, aber nicht den Radius gleichverteilt wählen, da der Umfang eines Kreises mit kleinem Radius kleiner als der Umfang eines Kreises mit größerem Radius ist. Damit müssen kleine Radien seltener als große vorkommen. Damit auf jedem Kreis mit Radius $0 \leq r \leq R$ die Dichte der Punkte etwa gleich ist, muss die entsprechende Wahrscheinlichkeit proportional zum Umfang $2\pi r$ und damit zu r sein. Genauer benötigen wir statt RX_1 eine Zufallsvariable X_3 mit der Dichte $f(r) = cr$, wobei $c = \frac{2}{R^2}$ gewählt ist, damit $\int_0^R f(r) dr = \frac{c}{2} R^2 = 1$, also

$$F^{X_3}(x) = P(X_3 \leq x) = \int_0^x \frac{2}{R^2} r dr = \frac{x^2}{R^2} \text{ für } x \in [0, R].$$

Wir stellen nun mit der Inversionsmethode X_3 als $g(X_1)$ mit einer streng monotonen Funktion g dar.

$$\frac{x^2}{R^2} = P(X_3 \leq x) = P(g(X_1) \leq x) = P(X_1 \leq g^{-1}(x)) = \int_0^{g^{-1}(x)} 1 dt = g^{-1}(x).$$

Damit ist also $g(x) = R\sqrt{x}$, und wir erhalten die gewünschten Zufallspunkte über

$$(x, y) = R\sqrt{X_1(\omega)} \cdot (\cos(2\pi X_2(\omega)), \sin(2\pi X_2(\omega))).$$

Was wir soeben für eine Kreisscheibe gemacht haben, funktioniert auch für eine Kugel mit Radius $R > 0$. Dabei ergibt sich die Wahrscheinlichkeitsverteilung für die einzelnen Variablen r , φ und θ der geometrischen Kugelkoordinatendarstellung über das Verzerrungsverhältnis der Kugelkoordinatentransformation, also über die Funktionaldeterminante $r^2 \cos(\vartheta)$ (welche für die oben betrachteten Polarkoordinaten r ist, siehe Kapitel 4.4.4). Die Funktionaldeterminante beschreibt, wie sich das Volumen eines (infinitesimalen) Raumelements beim Übergang von kartesischen Koordinaten zu Kugelkoordinaten ändert.

Wir halten nun jeweils zwei Variablen fest und arbeiten mit der Verteilung für die verbleibende:

- Seien φ und r fest. Bezüglich ϑ ist die Dichte $\cos(\vartheta)/2$ auf $[-\frac{\pi}{2}, \frac{\pi}{2}]$ proportional zum Faktor $\cos(\vartheta)$ der Funktionaldeterminante, wobei durch 2 geteilt wird, um auf 1 zu normieren: $\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos(\vartheta)/2 d\vartheta = 1$. Wir möchten eine Zufallsvariable mit dieser Dichte über die gleichverteilte Zufallsvariable X_1 , jetzt mit Werten in $[-\frac{\pi}{2}, \frac{\pi}{2}]$, generieren.

$$\frac{1}{2} \int_{-\frac{\pi}{2}}^x \cos \vartheta d\vartheta = P(f_1(X_1) \leq x) = P(X_1 \leq f_1^{-1}(x)) = \frac{1}{\pi} \int_{-\frac{\pi}{2}}^{f_1^{-1}(x)} 1 d\vartheta = \frac{1}{\pi} f_1^{-1}(x) + \frac{1}{2}.$$

Damit ist $f_1^{-1}(x) = \frac{\pi}{2} \sin(x)$, also $f_1(x) = \arcsin(\frac{2}{\pi}x)$. Wenn man statt der geometrischen die astronomischen Kugelkoordinaten verwendet, ist die Funktionaldeterminante $r^2 \sin(\vartheta)$, X_1 muss gleichverteilt auf $[0, \pi]$ sein und

$$\frac{1}{2} \int_0^x \sin \vartheta d\vartheta = P(f_1(X_1) \leq x) = P(X_1 \leq f_1^{-1}(x)) = \frac{1}{\pi} \int_0^{f_1^{-1}(x)} 1 d\vartheta = \frac{1}{\pi} f_1^{-1}(x).$$

Damit ist in diesem Fall $f_1^{-1}(x) = \frac{\pi}{2}(-\cos(x) + 1)$, $f(x) = \arccos(1 - \frac{2}{\pi}x)$, also wird ϑ über $\arccos(1 - \frac{2}{\pi}X_1)$ berechnet.

- Zur Berechnung des Radius halten wir φ und ϑ fest. Wir konstruieren eine Zufallsvariable mit Werten in $[0, R]$ und Dichte $\frac{3}{R^3}r^2$, die proportional zu r^2 ist. Mit der gleichverteilten Zufallsvariable X_2 , die Werte in $[0, R]$ annimmt und hier die Dichte $1/R$ hat, suchen wir eine Funktion f_2 , so dass

$$\frac{3}{R^3} \int_0^x r^2 dr = \frac{1}{R} \int_0^{f_2^{-1}(x)} 1 dr = \frac{1}{R} f_2^{-1}(x)$$

gilt. Damit ist $f_2^{-1}(x) = \frac{3}{R^2} \int_0^x r^2 dr = \frac{1}{R^2} x^3$ und $f_2(x) = R^{2/3} x^{1/3}$. So können wir für den Radius r Werte der Zufallsvariable $R^{2/3} X_2^{1/3}$ verwenden.

- Schließlich kann $\varphi \in [0, 2\pi]$ direkt gleichverteilt gewählt werden.

Für höhere Dimensionen funktioniert der Ansatz exakt in dieser Weise. Allerdings wird das Rechnen sehr aufwändig, da in der Funktionaldeterminante für die zusätzlichen Winkel $\cos^k$ - bzw. $\sin^k$ -Faktoren auftreten, die durch die Integration auf schwierig zu invertierende Funktionen führen. ■

7.2.6 Lage- und Streuungsparameter von Zufallsvariablen

Da wir mit Zufallsvariablen die Ergebnisse eines Experiments als Zahlen ausdrücken können, können wir nun auch einen Mittelwert und ein Maß für die Streuung angeben.

Definition 7.24 (Erwartungswert einer Zufallsvariable)

Sei X eine diskrete Zufallsvariable auf Ω mit

$$\sum_{x \in X(\Omega)} |x| \cdot P(X = x) < \infty. \quad (7.17)$$

Der **Erwartungswert** (oder der **Mittelwert**) von X unter P ist

$$E(X) := \sum_{x \in X(\Omega)} x \cdot P(X = x).$$

Das Symbol $\sum_{x \in X(\Omega)}$ bedeutet, dass über alle (höchstens abzählbar vielen) Werte x summiert wird, die die Zufallsvariable X annehmen kann. Gibt X die Augenzahl eines Wurfs mit einem Würfel an, so wird über die Werte 1, 2, 3, 4, 5 und 6 summiert. Jeder dieser Werte x wird bei der Summation mit der Wahrscheinlichkeit seines Eintreffens $P(X = x)$ multipliziert. Beim Würfeln wäre das immer $\frac{1}{6}$:

Beispiel 7.47 (Fortsetzung des Würfel experiments)

Die identisch verteilten Zufallsvariablen X_1 und X_2 besitzen den gleichen Erwartungswert: $E(X_1) = E(X_2) = \sum_{k=1}^6 k \cdot \frac{1}{6} = \frac{21}{6} = 3,5$. Im Durchschnitt erwarten wir also eine Augenzahl von 3,5. ■

Die rein technische Voraussetzung (7.17) wird nur benötigt, wenn X unendlich viele Werte annimmt. Dann sichert sie die Konvergenz der Summe und erlaubt Änderungen der Summationsreihenfolge (siehe Bemerkungen zu Satz 2.13 auf Seite 243).

Der Erwartungswert entspricht dem arithmetischen Mittel (7.1) $\bar{x} = \sum_{k=1}^m a_k f_k$ in der beschreibenden Statistik. Dabei sind a_k die verschiedenen Merkmalsausprägungen mit ihrer relativen Häufigkeit f_k . Wenn man die Wahrscheinlichkeit $P(X = x)$ als die relative Häufigkeit f_k der Merkmalsausprägung $a_k = x$ versteht, gehen beide Begriffe ineinander über.

Der Name „Erwartungswert“ drückt aus, dass man als Wert von X bei der Durchführung vieler Experimente im Mittel die Zahl $E(X)$ erwartet. Damit beschäftigen wir uns in Kapitel 7.2.7.

Statt über die Werte der Zufallsvariablen zu summieren, kann es einfacher sein, die Summe über die Elementarereignisse des Experiments aufzubauen. Denn mehrere Elementarereignisse können zum gleichen Wert der Zufallsvariable führen, dessen Wahrscheinlichkeit dann nicht sofort ablesbar ist. Allerdings bekommen wir es so wieder mit den ursprünglichen Elementarereignissen zu tun, von denen wir uns mit der Zufallsvariable eigentlich befreien möchten.

Satz 7.8 (Erwartungswert diskreter Zufallsvariablen *)

Sei X eine diskrete Zufallsvariable auf höchstens abzählbarer Grundgesamtheit Ω .

a) Die Bedingung

$$\sum_{x \in X(\Omega)} |x| P(X = x) < \infty$$

ist äquivalent mit

$$\sum_{\omega \in \Omega} |X(\omega)| P(\{\omega\}) < \infty.$$

b) Besitzt X (gemäß der Bedingung unter a)) einen Erwartungswert, so ist

$$E(X) = \sum_{x \in X(\Omega)} x P(X = x) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}). \quad (7.18)$$

Gibt X z. B. beim Würfeln mit zwei Würfeln die Augenzahl an und ist auf $\Omega := \{(l, k) : 1 \leq l, k \leq 6\}$ definiert, so gilt:

$$\begin{aligned} E(X) &= \sum_{x \in \{2, 3, \dots, 12\}} x \cdot P(X = x) = \sum_{(l, k) \in \Omega} X((l, k)) \cdot P(\{(l, k)\}) \\ &= \sum_{l=1}^6 \sum_{k=1}^6 (l + k) \frac{1}{36} = \frac{1}{36} \sum_{l=1}^6 \left[6l + \frac{6 \cdot 7}{2} \right] = \frac{12}{36} \frac{6 \cdot 7}{2} = 7. \end{aligned} \quad (7.19)$$

Beweis: Wir zeigen nur b). Die Aussage unter a) ergibt sich völlig analog. Wegen der absoluten Konvergenz der Summe dürfen wir beliebig umsortieren:

$$\begin{aligned} \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) &= \sum_{x \in X(\Omega)} \sum_{\omega \in \Omega: X(\omega)=x} X(\omega) P(\{\omega\}) = \sum_{x \in X(\Omega)} x \sum_{\omega \in \Omega: X(\omega)=x} P(\{\omega\}) \\ &= \sum_{x \in X(\Omega)} x P(X^{-1}(\{x\})) = \sum_{x \in X(\Omega)} x P(X = x). \end{aligned}$$

Dabei haben wir die Additivität von P ausgenutzt. □

Falls die Verteilung P^X einer (nicht-diskreten) Zufallsvariable über eine Dichte f (siehe Seite 1024) gegeben ist, so verwendet man bei der Berechnung des Erwartungswerts ein Integral statt der Summe. Da die Wahrscheinlichkeit $P(X = x)$ für jede Zahl $x \in \mathbb{R}$ dann gleich null ist, fassen wir Werte in kleinen Intervallen der Länge $\frac{1}{n}$ zusammen und ordnen einem Wert $\frac{k}{n}$ pro Intervall $\left[\frac{k}{n}, \frac{k+1}{n}\right]$ die Wahrscheinlichkeit des Intervalls zu. Wählen wir n hinreichend groß und damit die Intervalle sehr klein, so erhalten wir

$$E(X) \approx \sum_{k \in \mathbb{Z}} \frac{k}{n} P\left(X \in \left[\frac{k}{n}, \frac{k+1}{n}\right]\right) = \sum_{k \in \mathbb{Z}} \frac{k}{n} P^X\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right)$$

$$= \sum_{k \in \mathbb{Z}} \frac{k}{n} \int_{\frac{k}{n}}^{\frac{k+1}{n}} f(t) dt \approx \sum_{k \in \mathbb{Z}} \int_{\frac{k}{n}}^{\frac{k+1}{n}} t \cdot f(t) dt = \int_{-\infty}^{\infty} t \cdot f(t) dt.$$

Dieses Integral findet man häufig als Definition des Erwartungswerts. Das liegt daran, dass man mit einem etwas allgemeineren Integrationsbegriff als dem von Riemann auch unendliche Summen als Integrale schreiben kann.

Den Erwartungswert einer standardnormalverteilten Zufallsvariable X mit der Dichte $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ (siehe ebenfalls Seite 1024) erhalten wir damit über eine Substitution $u = \frac{t^2}{2}$, $du = t dt$:

$$E(X) = \int_{-\infty}^{\infty} t \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \left[\int_{\infty}^0 e^{-u} du + \int_0^{\infty} e^{-u} du \right] = 0. \quad (7.20)$$

Satz 7.9 (Linearität des Erwartungswerts)

a) Seien X eine diskrete Zufallsvariable mit Erwartungswert $E(X)$ und $a, b \in \mathbb{R}$. Dann gilt analog zum arithmetischen Mittel für die Zufallsvariable $a \cdot X + b$:

$$E(a \cdot X + b) = a \cdot E(X) + b.$$

b) Seien X und Y zwei diskrete Zufallsvariablen auf Ω , für die der Erwartungswert erklärt ist. Dann existiert $E(X + Y)$ und

$$E(X + Y) = E(X) + E(Y).$$

Beweis: (Skizze) a) Da X eine diskrete Zufallsvariable ist, ist $X(\Omega)$ höchstens abzählbar, und wir können über den Wertebereich summieren:

$$\begin{aligned} \sum_{y \in (aX+b)(\Omega)} y P(aX+b=y) &= \sum_{x \in X(\Omega)} (ax+b) P(X=x) \\ &= a \sum_{x \in X(\Omega)} x P(X=x) + \sum_{x \in X(\Omega)} b P(X=x) = a E(X) + b P(\Omega) = a E(X) + b. \end{aligned}$$

b) Wir zeigen die Aussage nur für höchstens abzählbares Ω mittels (7.18):

$$\begin{aligned} \sum_{\omega \in \Omega} |(X+Y)(\omega)| P(\{\omega\}) &\leq \sum_{\omega \in \Omega} [|X(\omega)| + |Y(\omega)|] P(\{\omega\}) \\ &= \sum_{\omega \in \Omega} |X(\omega)| P(\{\omega\}) + \sum_{\omega \in \Omega} |Y(\omega)| P(\{\omega\}) < \infty. \end{aligned}$$

$$\begin{aligned} E(X+Y) &= \sum_{\omega \in \Omega} (X+Y)(\omega) P(\{\omega\}) \\ &= \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) + \sum_{\omega \in \Omega} Y(\omega) P(\{\omega\}) = E(X) + E(Y). \end{aligned}$$

□

Wir berechnen den Erwartungswert für die diskreten Verteilungen, mit denen wir uns zuvor beschäftigt haben:

- Für eine geometrisch verteilte Zufallsvariable $X \sim G(p)$ (siehe Beispiel 7.37, $k \in \mathbb{N}$) mit $P(X = k) = (1 - p)^{k-1}p$ erhalten wir den Erwartungswert mittels gliedweiser Ableitung (Satz 2.61 auf Seite 416) der Potenzreihe $\sum_{k=0}^{\infty} x^k$, die den Konvergenzradius 1 hat, siehe Seite 414, und für $|x| < 1$ gegen $\frac{1}{1-x}$ konvergiert (Seite 237). Für $|x| < 1$ ist

$$\frac{1}{(1-x)^2} = \frac{d}{dx} \frac{1}{1-x} = \frac{d}{dx} \sum_{k=0}^{\infty} x^k = \sum_{k=0}^{\infty} \frac{d}{dx} x^k = \sum_{k=1}^{\infty} kx^{k-1}.$$

Damit erhalten wir für den Erwartungswert für $x = 1 - p$:

$$E(X) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p \sum_{k=1}^{\infty} k(1-p)^{k-1} = p \cdot \frac{1}{(1-(1-p))^2} = \frac{1}{p}.$$

- Sei $X \sim H(N; M; n)$ hypergeometrisch verteilt. Der Trick beim Ausrechnen des Erwartungswerts ist die Beziehung

$$\binom{i}{k} = \frac{i}{k} \frac{(i-1)!}{(i-1-(k-1))!(k-1)!} = \frac{i}{k} \binom{i-1}{k-1}.$$

$$\begin{aligned} E(X) &= \sum_{k=0}^n kP(X=k) = \sum_{k=0}^n k \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} = \sum_{k=1}^n k \frac{\frac{M}{k} \binom{M-1}{k-1} \binom{N-M}{n-k}}{\frac{N}{n} \binom{N-1}{n-1}} \\ &= n \frac{M}{N} \sum_{k=1}^n \frac{\binom{M-1}{k-1} \binom{N-M}{n-k}}{\binom{N-1}{n-1}} = n \frac{M}{N} \sum_{k=0}^{n-1} \frac{\binom{M-1}{k} \binom{N-1-(M-1)}{n-1-k}}{\binom{N-1}{n-1}} = n \frac{M}{N}, \end{aligned}$$

denn hier werden die Wahrscheinlichkeiten der hypergeometrischen Verteilung $H(N-1; M-1; n-1)$ aufsummiert. $\frac{M}{N}$ ist das Verhältnis der defekten zu allen Teilen. Wir „erwarten“, dass dieses Verhältnis auch für die Stichprobe vorliegt, d. h., wir erwarten Stichprobenumfang n multipliziert mit $\frac{M}{N}$ fehlerhafte Teile in der Stichprobe – genau den Erwartungswert.

- Für die Binomialverteilung $Bi(n; p)$ gilt wegen des Binomischen Lehrsatzes (1.11):

$$\begin{aligned} E(X) &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n k \frac{n(n-1)!}{k(k-1)![n-1-(k-1)]!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n k \frac{n}{k} \binom{n-1}{k-1} p^k (1-p)^{n-1-(k-1)} \\ &= n \sum_{k=0}^{n-1} \binom{n-1}{k} p^{k+1} (1-p)^{n-1-k} = np(p+1-p)^{n-1} = np. \end{aligned}$$

Das passt zur Anschauung: Wir erwarten im Mittel np -maliges Eintreffen.

■ Schließlich erhalten wir für die Poisson-Verteilung $Po(\lambda)$:

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k P(X = k) = \sum_{k=0}^{\infty} k \exp(-\lambda) \frac{\lambda^k}{k!} = \exp(-\lambda) \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} \\ &= \lambda \exp(-\lambda) \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \exp(-\lambda) \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda \exp(-\lambda) \exp(\lambda) = \lambda. \end{aligned}$$

Beispiel 7.48

Eine Lotterie hat im September 2009 ein Gewinnspiel 4 aus 48 an Haushalte verschickt. Dabei wird ein Gewinn einer Rente von 1 000 Euro für 120 Monate (also 120 000 Euro ohne Zinsen) versprochen, wenn vier verschiedene rein zufällig aufgedruckte Zahlen (zwischen 1 und 48) mit vier verschiedenen bereits gezogenen Zahlen der gleichen Menge übereinstimmen. Wir nehmen an, dass die Wahrscheinlichkeit, mit der ein Adressat am Gewinnspiel teilnimmt, 0,1 ist und dass insgesamt eine Million Briefe verschickt wurden. Welche auszuschüttende Gewinnsumme muss die Lotterie erwarten?

Ist X die Anzahl der Gewinner, so ist X binomialverteilt. Jeder Adressat hat die gleiche Chance. Stochastisch unabhängig davon, ob seine Zahlen gewinnen, nimmt er mit der Wahrscheinlichkeit $\frac{1}{10}$ teil. Die Anzahl der Kombinationen von 4 aus 48 ohne Wiederholung ist $\binom{48}{4}$. Die Wahrscheinlichkeit, dass ein Adressat gewinnt, ist also

$$p = \frac{1}{10} \cdot \frac{1}{\binom{48}{4}} = \frac{2 \cdot 3 \cdot 4}{10 \cdot 48 \cdot 47 \cdot 46 \cdot 45} = \frac{1}{5 \cdot 12 \cdot 47 \cdot 46 \cdot 15} = \frac{1}{1\,945\,800}.$$

Damit sind $E(X) = 1\,000\,000 \cdot p = \frac{10\,000}{19\,458}$ Gewinne zu erwarten, also muss mit einer Gewinnsumme von $E(X) \cdot 120\,000 \approx 61\,671$ Euro gerechnet werden. Wenn die Lotterie das Gewinnspiel oft durchführt, wird sie im Mittel diesen Betrag zahlen müssen. Bei einmaliger Durchführung ist 120 000 Euro etwas wahrscheinlicher als 0 Euro. Wenn das Spiel nur an 500 000 Personen verschickt wird, ist $E(X) \approx \frac{1}{4}$, und damit ist es recht wahrscheinlich, dass die Lotterie gar nichts zahlen muss. ■

Ist X eine diskrete Zufallsvariable mit Erwartungswert $E(X)$ und $f : \mathbb{R} \rightarrow \mathbb{R}$ eine Abbildung, so dass auch der Erwartungswert der Verkettung $f \circ X : \Omega \rightarrow \mathbb{R}, \omega \mapsto f(X(\omega))$ existiert, dann können wir diesen auch mit dem Wissen über P^X (wie beim Nachrechnen der Linearität des Erwartungswerts) berechnen. $P(f(X) = y)$ ist die Wahrscheinlichkeit, dass X einen Wert annimmt, der von f auf y abgebildet wird. Man erhält diese Wahrscheinlichkeit, indem man die Einzelwahrscheinlichkeiten aller Werte von X aufsummiert, die von f auf y abgebildet werden.

$$\begin{aligned} E(f \circ X) &= \sum_{y \in f(X(\Omega))} y P(f(X) = y) = \sum_{y \in f(X(\Omega))} \sum_{\{x \in X(\Omega) : f(x) = y\}} \underbrace{y P(X = x)}_{= f(x) P(X = x)} \\ &= \sum_{x \in X(\Omega)} f(x) P(X = x). \end{aligned} \tag{7.21}$$

Man darf hier die Summationsreihenfolge beliebig wählen und auch beim Zusammenfassen der Summen ändern, da in der Definition des Erwartungswerts auf Seite 1034 die absolute Konvergenz der Summen gefordert ist.

Wie in der beschreibenden Statistik interessieren wir uns nicht nur für den Mittel- bzw. Erwartungswert, sondern auch für die Streuung:

Definition 7.25 (Streuung einer Zufallsvariable)

Sei X eine diskrete Zufallsvariable auf Ω , für die der Erwartungswert erklärt ist.

$$\text{Var}(X) := E([X - E(X)]^2)$$

*heißt die **Varianz** von X unter P . Die **Standardabweichung***

$$\sigma := \sqrt{\text{Var}(X)}$$

ist (analog zur beschreibenden Statistik) die Wurzel aus der Varianz. Daher wird die Varianz mit σ^2 und die Standardabweichung mit σ bezeichnet.

Man misst hier, wie viel die Werte der Zufallsvariable vom Erwartungswert abweichen, wobei man den Abstand quadriert. Dabei muss berücksichtigt werden, dass die Werte von X mit unterschiedlicher Wahrscheinlichkeit angenommen werden. Eine große Abweichung $[X(\omega) - E(X)]^2$ spielt keine große Rolle, wenn $P(\{\omega\})$ klein ist, diese Abweichung also selten beobachtet wird. Daher definiert man die Varianz über den Erwartungswert von $[X - E(X)]^2$, so dass die Wahrscheinlichkeiten der Abweichungen berücksichtigt sind. Die Varianz ist die erwartete quadrierte Abweichung von $E(X)$.

In (7.4) haben wir die empirische Varianz als $s^2 := \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$ definiert. s^2 und σ^2 sind fast analog definiert. Der Erwartungswert $E(X)$ entspricht dem arithmetischen Mittel $\bar{x}$, ebenso entspricht $E([X - E(X)]^2)$ dem arithmetischen Mittel

$$\overline{(x - \bar{x})^2} = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{n-1}{n} \cdot \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{n-1}{n} s^2.$$

Der Vorfaktor ist für große Anzahlen n nahe bei 1. Er entsteht, weil man bei der empirischen Varianz durch $n-1$ statt durch n teilt. Bis auf diesen Faktor ergibt sich die Varianz aus der empirischen Varianz, indem man relative Häufigkeiten als Wahrscheinlichkeiten interpretiert. Die Bedeutung des Vorfaktors wird später in Kapitel 7.3.1 verständlich.

Die Varianz ist ein Erwartungswert, der für die Zufallsvariable $f \circ X$ mit $f(x) := [x - E(X)]^2$ zu berechnen ist. Dabei hilft uns (7.21):

$$\text{Var}(X) = \sum_{x \in X(\Omega)} [x - E(X)]^2 P(X = x).$$

Beispiel 7.49

Wir berechnen den Erwartungswert und die Varianz einer Zufallsvariable X mit $P(X = 1) = 0,2$, $P(X = 2) = 0,4$ und $P(X = 5) = 0,4$.

$$\begin{aligned} E(X) &= 1 \cdot 0,2 + 2 \cdot 0,4 + 5 \cdot 0,4 = 3. \\ \text{Var}(X) &= (1 - 3)^2 \cdot 0,2 + (2 - 3)^2 \cdot 0,4 + (5 - 3)^2 \cdot 0,4 = 2,8. \end{aligned}$$

Beispiel 7.50 (Fortsetzung des Würfelexperiments)

Die identisch verteilten Zufallsvariablen X_1 und X_2 mit dem Erwartungswert 3,5 besitzen die Varianz $\text{Var}(X_1) = \text{Var}(X_2) = E([X_1 - E(X_1)]^2) = \frac{1}{6} \sum_{k=1}^6 [k - 3,5]^2 = (17,5)/6 = 2,916\dots$

Wegen der Linearität des Erwartungswerts (Satz 7.9) können wir die Varianz bei Existenz von $E(X^2)$ einfacher darstellen:

$$\begin{aligned} \text{Var}(X) &:= E([X - E(X)]^2) = E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 = E(X^2) - E(X)^2. \end{aligned}$$

Für einen Faktor $\lambda \in \mathbb{R}$ gilt offenbar

$$\text{Var}(\lambda X) = E(\lambda^2 X^2) - E(\lambda X)^2 = \lambda^2(E(X^2) - E(X)^2) = \lambda^2 \text{Var}(X). \quad (7.22)$$

Beispiel 7.51

Wir berechnen die Varianz einer standardnormalverteilten Zufallsvariable X . Wir wissen bereits, dass $E(X) = 0$ ist (siehe Seite 1035). Analog zu (7.21) erhalten wir mittels partieller Integration

$$\begin{aligned} E([X - E(X)]^2) &= E(X^2) - E(X)^2 = E(X^2) = \int_{-\infty}^{\infty} t^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= \frac{1}{\sqrt{2\pi}} \left[\left[-te^{-\frac{t^2}{2}} \right]_{-\infty}^0 + \left[-te^{-\frac{t^2}{2}} \right]_0^{\infty} \right] + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = 0 + P^X([-\infty, \infty]) = 1. \end{aligned}$$

Wir geben die Varianz für die diskreten Verteilungen, mit denen wir uns zuvor beschäftigt haben, an:

- Geometrische Verteilung $G(p)$: $\sigma^2 = \text{Var}(X) = \frac{1-p}{p^2}$.
- Sei $X \sim H(N; M; n)$ hypergeometrisch verteilt. Dann ist

$$\sigma^2 = \text{Var}(X) = n \frac{M}{N} \left(1 - \frac{M}{N} \right) \frac{N-n}{N-1}.$$

- Binomialverteilung $\text{Bi}(n; p)$: $\sigma^2 = \text{Var}(X) = np(1-p)$.
- Poisson-Verteilung $\text{Po}(\lambda)$: $\sigma^2 = \text{Var}(X) = \lambda$.

Die Varianz der hypergeometrischen Verteilung sieht sehr kompliziert aus. Betrachtet man aber das Verhältnis von defekten zu ganzen Teilen $\frac{M}{N}$ als initiale Wahrscheinlichkeit p , auf ein defektes Teil zu stoßen, so wird daraus für $n \geq 2$

$$\sigma^2 = np(1-p) \frac{N-n}{N-1} \leq np(1-p).$$

Rechts steht jetzt die Varianz der Binomialverteilung. Ziehen „ohne Zurücklegen“ hat also eine kleinere Varianz als „Ziehen mit Zurücklegen“. Für $N \rightarrow \infty$ verschwindet dieser Vorteil.

Mittels der Varianz können wir die Wahrscheinlichkeit abschätzen, dass eine Zufallsvariable Werte annimmt, die einen gewissen Abstand zum Erwartungswert haben. Dies ist die wahrscheinlichkeitstheoretische Interpretation der Varianz als Streuung:

Satz 7.10 (Ungleichung von Tschebycheff (1821–1894))

Sei X eine diskrete Zufallsvariable auf (höchstens abzählbarem) Ω , für die nicht nur $E(X)$, sondern auch $E(X^2)$ (und damit $\text{Var}(X)$) existiert. Dann gilt für jedes $\varepsilon > 0$: Die Wahrscheinlichkeit, dass X einen Wert annimmt, der mehr als ε vom Erwartungswert abweicht, ist kleiner oder gleich $\frac{\text{Var}(X)}{\varepsilon^2}$. Je kleiner die Varianz ist, desto kleiner ist auch diese Wahrscheinlichkeit:

$$P(|X - E(X)| > \varepsilon) = P(\{\omega \in \Omega : |X(\omega) - E(X)| > \varepsilon\}) \leq \frac{\text{Var}(X)}{\varepsilon^2}. \quad (7.23)$$

Insbesondere findet man damit Werte, die weiter als eins vom Erwartungswert entfernt sind, höchstens mit der Wahrscheinlichkeit $\min\{1, \text{Var}(X)\}$. Der Beweis ist überraschend einfach, wenn man hier den Satz 7.8 verwendet, mit dem der Erwartungswert als Summe über Elementarereignisse geschrieben werden kann:

Beweis: Sei $A_\varepsilon := \{\omega \in \Omega : |X(\omega) - E(X)| > \varepsilon\}$ das zu bewertende Ereignis, dass X einen Wert liefert, der weiter als ε vom Erwartungswert entfernt ist. Damit ist

$$\begin{aligned} \text{Var}(X) &= E((X - E(X))^2) = \sum_{\omega \in \Omega} (X(\omega) - E(X))^2 P(\{\omega\}) \\ &\geq \sum_{\omega \in A_\varepsilon} (X(\omega) - E(X))^2 P(\{\omega\}) \geq \sum_{\omega \in A_\varepsilon} \varepsilon^2 P(\{\omega\}) = \varepsilon^2 P(A_\varepsilon). \end{aligned}$$

Beim ersten „ $\geq$ “ haben wir wegen $A_\varepsilon \subset \Omega$ nicht-negative Summanden weggelassen, beim zweiten „ $\geq$ “ wird die Definition der Menge A_ε ausgenutzt. Damit ist $P(A_\varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$. $\square$

Wie in der beschreibenden Statistik kann man auch hier den Begriff der Kovarianz definieren, um mittels des Korrelationskoeffizienten im Falle einer stochastischen Abhängigkeit einen (linearen) Zusammenhang zwischen Zufallsvariablen weiter zu klassifizieren.

Definition 7.26 (Korrelation von Zufallsvariablen)

Seien X und Y diskrete Zufallsvariablen, so dass nicht nur die Erwartungswerte, sondern auch $E(X^2)$ und $E(Y^2)$ existieren.

a) Die **Kovarianz** von X und Y ist erklärt durch

$$\text{Cov}(X, Y) := E(XY) - E(X)E(Y).$$

b) Sind $\text{Var}(X) > 0$ und $\text{Var}(Y) > 0$, so heißt

$$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

der **Korrelationskoeffizient** von X und Y .

c) Im Fall $\text{Cov}(X, Y) = 0$ heißen die Zufallsvariablen X und Y **unkorreliert**.

Wir haben in der Definition keine separate Forderung zur Existenz von $E(XY)$ gestellt. In der Tat folgt diese bereits daraus, dass $E(X^2)$ und $E(Y^2)$ existieren: Damit $E(XY)$ existiert und die Kovarianz wohldefiniert ist, müssen wir bei einem höchstens abzählbaren Ω zeigen, dass $\sum_{\omega \in \Omega} |X(\omega)Y(\omega)|P(\{\omega\}) < \infty$ ist. Wegen $(|X(\omega)| - |Y(\omega)|)^2 \geq 0$ ist $|X(\omega)Y(\omega)| \leq \frac{1}{2}(X^2(\omega) + Y^2(\omega))$, so dass

$$\begin{aligned} \sum_{\omega \in \Omega} |X(\omega)Y(\omega)|P(\{\omega\}) &\leq \frac{1}{2} \left(\sum_{\omega \in \Omega} X^2(\omega)P(\{\omega\}) + \sum_{\omega \in \Omega} Y^2(\omega)P(\{\omega\}) \right) \\ &= \frac{1}{2} (E(X^2) + E(Y^2)) < \infty. \end{aligned}$$

Ohne Beweis zitieren wir das folgende Lemma:

Lemma 7.2 (Multiplikationssatz für Erwartungswerte)

Sind die Zufallsvariablen X und Y mit Erwartungswerten $E(X)$ und $E(Y)$ stochastisch unabhängig, so ist

$$E(XY) = E(X)E(Y). \quad (7.24)$$

Damit ist insbesondere $\text{Cov}(X, Y) = 0$, X und Y sind also auch unkorreliert.

Den Bezug zur empirischen Kovarianz (siehe Definition 7.11 auf Seite 985) erhält man über die linke Seite von

$$\begin{aligned} E((X - E(X))(Y - E(Y))) &= E(XY) - 2E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) = \text{Cov}(X, Y). \end{aligned}$$

Wir werden später sehen, dass man über die empirischen Größen die stochastischen schätzen kann. Entsprechend „passt“ der (stochastische) Korrelationskoeffizient zum empirischen. Man kann nachrechnen, dass auch wieder $|\rho(X, Y)| \leq 1$ ist.

Die Bedeutung der Kovarianz erschließt sich aus dem folgenden Satz:

Satz 7.11 (Varianz einer Summe)

Seien X und Y wieder diskrete Zufallsvariablen, für die nicht nur die Erwartungswerte, sondern auch $E(X^2)$ und $E(Y^2)$ existieren. Dann existiert $\text{Var}(X + Y)$ mit

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y).$$

Beweis:

$$\begin{aligned} \text{Var}(X + Y) &= E((X + Y - E(X + Y))^2) = E((X - E(X) + Y - E(Y))^2) \\ &= E((X - E(X))^2) + E((Y - E(Y))^2) + 2E((X - E(X))(Y - E(Y))) \\ &= \text{Var}(X) + \text{Var}(Y) + 2(E(XY) - E(X)E(Y)) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y). \end{aligned}$$

□

Für diskrete Zufallsvariablen $X_1, \dots, X_n$ kann man mittels Vollständiger Induktion den Satz 7.11 erweitern:

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < k \leq n} \text{Cov}(X_i, X_k). \quad (7.25)$$

Mit der Kovarianz können wir also die Streuung einer Summenverteilung ausrechnen.

7.2.7 Gesetz der großen Zahlen

Führt man ein Wahrscheinlichkeitsexperiment sehr oft durch und bildet dann das arithmetische Mittel der beobachteten Werte, so liegt dieses erfahrungsgemäß sehr nah beim Erwartungswert. Das ist das **Gesetz der großen Zahlen**. Würfelt man sehr oft, wird man im Mittel eine Augenzahl von 3,5 erhalten. In diesem Kapitel formulieren wir dieses Gesetz mathematisch, so dass es bewiesen werden kann. Vorweg eine Warnung: Wenn Sie zehnmal die Eins gewürfelt haben, dann ist die Wahrscheinlichkeit, mit dem nächsten Wurf wieder eine 1 zu bekommen, immer noch $\frac{1}{6}$. Der Würfel hat kein Gedächtnis. Das Ereignis, elfmal (unabhängig) eine Eins zu würfeln, ist dagegen nur $\frac{1}{6^{11}}$. Wenn aber bereits zehn Einsen gewürfelt wurden, ist dieses unwahrscheinliche Ereignis, das a priori die Wahrscheinlichkeit $\frac{1}{6^{10}}$ hatte, schon eingetreten. Man kann also beim Roulette nicht reich werden, indem man auf Schwarz setzt, wenn fünfmal Rot gekommen ist. Auch hat man keinen Vorteil, wenn man beim Lotto nur Zahlen ankreuzt, die in der Vergangenheit seltener (oder alternativ: häufiger) gezogen wurden. Wenn die Kugeln nicht manipuliert sind (wovon wir ausgehen können), dann wird jede mit der gleichen Wahrscheinlichkeit unabhängig von der Vergangenheit gezogen.

Wir erwarten, dass sich die axiomatisch eingeführten Wahrscheinlichkeiten wie relative Häufigkeiten verhalten. Für eine sehr große Versuchsanzahl n sollte daher ein Wert x einer Zufallsvariable X ca. $n \cdot P(X = x)$ -mal auftreten.

Bildet man das arithmetische Mittel aller beobachteten Werte von X , so würde man also ungefähr

$$\frac{1}{n} \sum_{x \in X(\Omega)} nP(X = x) \cdot x = E(X)$$

erhalten. Sollte man dies sogar beweisen können, hätte man die Bestätigung, dass die axiomatische Definition von Wahrscheinlichkeiten mit der Vorstellung relativer Häufigkeiten zusammenpasst. Dem ist so, und der entsprechende Satz heißt das Gesetz der großen Zahlen (siehe Satz 7.12).

Um zur mathematischen Formulierung dieses Gesetzes zu gelangen, betrachten wir im Folgenden nicht n Werte einer Zufallsvariable X , sondern stattdessen je einen Wert von n Zufallsvariablen.

Beispiel 7.52 (Fortsetzung des Würfelexperiments)

Wir betrachten die Zufallsvariable $X : \Omega \rightarrow [2, 12]$ definiert über $X := X_1 + X_2$. Damit ist X die Augensumme beider Würfe. Wir berechnen die Wahrscheinlichkeit $P(X = k)$, dass X den Wert k annimmt (vgl. (7.19)):

$$P(X = 2) = P(X^{-1}(\{2\})) = P(\{(1, 1)\}) = \frac{1}{36} = P(X = 12),$$

$$P(X = 3) = P(\{(1, 2), (2, 1)\}) = \frac{2}{36} = P(X = 11),$$

$$P(X = 4) = P(X = 10) = \frac{3}{36}, \quad P(X = 5) = P(X = 9) = \frac{4}{36},$$

$$P(X = 6) = P(X = 8) = \frac{5}{36}, \quad P(X = 7) = \frac{6}{36}, \quad E(X) = 7. \quad \blacksquare$$

Wir können das Beispiel so erweitern, dass wir statt zweimal unendlich oft unabhängig voneinander würfeln (vgl. mit den Überlegungen auf Seite 1026). So entsteht eine Folge von identisch verteilten Zufallsvariablen $(X_n)_{n=1}^\infty$, wobei X_n die Augenzahl des n -ten Wurfs angibt. Diese Zufallsvariablen sind stochastisch unabhängig. Im Sinne von Definition 7.22 auf Seite 1026 meinen wir damit, dass jede endliche Auswahl von Zufallsvariablen der Folge stochastisch unabhängig ist. Im Folgenden reicht es aber schon, wenn je zwei der Zufallsvariablen unabhängig sind.

Betrachtet man nun allgemein n (n sehr groß) (paarweise) stochastisch unabhängige, identisch verteilte Zufallsvariablen $X_1, \dots, X_n$ auf einem Wahrscheinlichkeitsraum und führt ein einziges Experiment mit Ergebnis $\omega \in \Omega$ durch, so werden den Wert $x \in X_1(\Omega)$ ca. $nP(X_1 = x)$ Zufallsvariablen der $X_1, \dots, X_n$ annehmen. Bei der Erläuterung des Erwartungswerts haben wir n Experimente gemacht und jeweils eine Zufallsvariable ausgewertet. Nun machen wir ein Experiment und werten stattdessen n Zufallsvariablen aus. Der Wahrscheinlichkeitsraum ist dabei im Prinzip so aufgebaut,

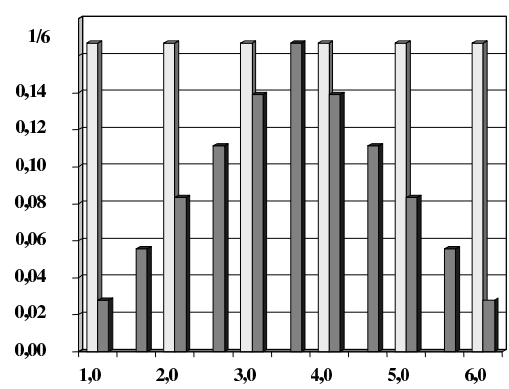


Abb. 7.12: Wahrscheinlichkeiten $P(X_1 = x)$ (hell) und $P(\frac{X_1+X_2}{2} = x)$ (dunkel) der Werte der Zufallsvariablen X_1 und $\frac{X_1+X_2}{2}$ zu unabhängigen Würfelergebnissen X_1 und X_2

dass er die n Experimente als ein Experiment ausdrücken kann (wie das mehrfache Würfeln im obigen Beispiel). Wir erwarten also

$$\frac{1}{n} \sum_{i=1}^n X_i(\omega) \approx \frac{1}{n} \sum_{x \in X_1(\Omega)} n P(X_1 = x) x = E(X_1).$$

Exakt ausgedrückt ist dies der folgende Satz, der in dieser Form (paarweise stochastische Unabhängigkeit der Zufallsvariablen) auf Etemadi (1981) zurückgeht und zuvor schon von Kolmogorov unter stärkeren Voraussetzungen bewiesen wurde. Schaut man auf die Jahreszahl, so erkennt man, dass die Wahrscheinlichkeitsrechnung im Vergleich zur Differenzialrechnung ein relativ junger Bereich der Mathematik ist.

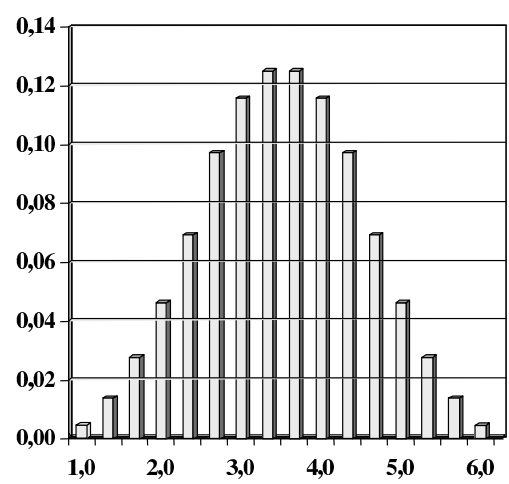


Abb. 7.13: Wahrscheinlichkeiten $P(\frac{X_1+X_2+X_3}{3} = x)$ der Werte der Zufallsvariable $\frac{X_1+X_2+X_3}{3}$, die über das Ergebnis von drei unabhängigen Würfeln mit einem Würfel gebildet ist

Satz 7.12 (Starkes Gesetz der großen Zahlen)

Sei $(X_k)_{k=1}^\infty$ eine Folge identisch verteilter (diskreter) Zufallsvariablen auf Ω , die (nur) paarweise stochastisch unabhängig sind und deren (gemeinsamer) Erwartungswert $E(X_k) = E(X_1)$ definiert ist. Dann genügt $(X_k)_{k=1}^\infty$ dem **starken Gesetz der großen Zahlen**, d. h.

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k(\omega) = E(X_1)$$

für alle $\omega \in \Omega \setminus N$, wobei N ein Ereignis ist, das fast sicher nicht eintritt: $P(N) = 0$.

Der Satz wird verständlicher, wenn man berücksichtigt, dass für identisch verteilte stochastisch unabhängige Zufallsvariablen, die damit insbesondere unkorreliert sind ($\text{Cov}(X_i, X_k) = 0$, $i \neq k$), gilt:

$$\text{Var} \left(\frac{1}{n} \sum_{k=1}^n X_k \right) \stackrel{(7.22)}{=} \frac{1}{n^2} \text{Var} \left(\sum_{k=1}^n X_k \right) \stackrel{(7.25)}{=} \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) = \frac{1}{n} \text{Var}(X_1). \quad (7.26)$$

Die Streuung der Summenverteilung geht also gegen 0, die Werte liegen daher immer näher beim Erwartungswert

$$E \left(\frac{1}{n} \sum_{k=1}^n X_k \right) = \frac{1}{n} \sum_{k=1}^n E(X_k) = \frac{1}{n} n E(X_1) = E(X_1). \quad (7.27)$$

Man beachte, dass die Einschränkung $\omega \in \Omega \setminus N$ notwendig ist. So ergibt sich im Würfelexperiment für $\omega = (1, 1, \dots, 1, \dots)$, also wenn alle unendlich vielen Würfe das Ergebnis 1 haben, $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k(\omega) = \lim_{n \rightarrow \infty} 1 = 1 \neq 3,5 = E(X_1)$. Betrachten wir solche unendlichen Folgen von Würfelausgängen als Elementarereignisse, so hat ein einzelnes Elementarereignis die Wahrscheinlichkeit null. Wir geraten in die Situation eines überabzählbaren Ω , das wir hier nicht weiter behandeln werden.

Der Satz bestätigt das anschauliche Verständnis von Zufall im Sinne der Definition (7.10): Wir führen ein Experiment n -mal durch und betrachten jeweils, ob ein Ereignis $A \subset \mathbb{R}$ eingetreten ist. A trete mit Wahrscheinlichkeit p ein. Es lassen sich n stochastisch unabhängige, identisch verteilte Zufallsvariablen X_k konstruieren, wobei $X_k = 1$ genau das Eintreten des Ereignisses A im k -ten Experiment bezeichnet (sonst ist $X_k = 0$). Offensichtlich ist $E(X_k) = E(X_1) = p$ (außerdem ist $\sum_{k=1}^n X_k \sim \text{Bi}(n; p)$). Dann ist

$$\frac{1}{n} \sum_{k=1}^n X_k = \frac{\text{„Anzahl des Eintretens von } E\text{“}}{n} \quad \text{und} \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k(\omega) = E(X_1) = p$$

für fast alle Versuchsausgänge $\omega \in \Omega$. Die Modellbildung, die auf dem Wahrscheinlichkeitsmaß aus Definition 7.15 basiert, passt also mit einer intuitiven statistischen Abschätzung von Wahrscheinlichkeiten zusammen:

Die relative Häufigkeit des Eintretens eines Ereignisses bei einer sehr großen Anzahl von unabhängigen Experimenten entspricht fast sicher der Wahrscheinlichkeit seines Eintretens.

Beispiel 7.53

Die $\text{Bi}(n; \frac{\pi}{4})$ -verteilte Zufallsvariable X , die im Beispiel 7.42 auf Seite 1028 angibt, wie viele Treffer innerhalb eines Kreises liegen, kann als Summe von unabhängigen Zufallsvariablen X_k geschrieben werden, die wie oben das Treffen des Kreises im k -ten Versuch beschreiben: $X = \sum_{k=1}^n X_k$. Damit strebt für fast alle Versuchsausgänge die Folge $a_n = \frac{1}{n} \sum_{k=1}^n X_k$ gegen den Erwartungswert $E(X_1) = \frac{\pi}{4}$. Das hatten wir schon bei der Einführung von Folgen auf Seite 216 behauptet. ■

Das Gesetz der großen Zahlen eignet sich auch, um in der schließenden Statistik den Erwartungswert von Zufallsvariablen abzuschätzen. Auch die Varianz, die über einen Erwartungswert definiert ist, lässt sich so approximieren (siehe Kapitel 7.3.1).

Hier können wir das starke Gesetz der großen Zahlen nicht beweisen – jedoch eine schwächere Variante. Die in Satz 7.12 beschriebene Konvergenz der Zahlenfolge $(\frac{1}{n} \sum_{k=1}^n X_k(\omega))_{n=1}^\infty$ für jeden festen Punkt aus $\omega \in \Omega \setminus N$ (also fast jedes Elementarereignis ω) wird als starke Konvergenz bezeichnet. Schwächt man diesen Begriff ab, indem man die Konvergenz nicht mehr für feste Punkte $\omega \in \Omega$ verlangt, erhält man das schwache Gesetz der großen Zahlen (Jakob Bernoulli, 1655–1705):

Satz 7.13 (Schwachere Gesetz der großen Zahlen)

Sei $(X_k)_{k=1}^\infty$ eine Folge identisch verteilter diskreter Zufallsvariablen auf Ω , die paarweise unkorreliert sind (was z. B. erfüllt ist, wenn sie paarweise stochastisch unabhängig sind, siehe Seite 1041) und deren Erwartungswerte $E(X_k) = E(X_1)$ sowie $E(X_k^2) = E(X_1^2)$ definiert sind. Dann gilt für jedes $\varepsilon > 0$:

$$\begin{aligned} \lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_{k=1}^n X_k - E(X_1) \right| > \varepsilon \right) \\ = \lim_{n \rightarrow \infty} P \left(\left\{ \omega \in \Omega : \left| \frac{1}{n} \sum_{k=1}^n X_k(\omega) - E(X_1) \right| > \varepsilon \right\} \right) = 0. \end{aligned}$$

Hier hat man keine Aussage über feste Elementarereignisse, sondern erhält für jedes n eine bezüglich P unwahrscheinlicher werdende Menge von Elementarereignissen, die zu Werten fern des Erwartungswerts führen. Für jedes n kann diese Menge aber völlig anders aussehen. Man spricht auch von **stochastischer Konvergenz**. Man kann zeigen, dass aus der starken (punktweisen) Konvergenz die stochastische Konvergenz folgt, das schwache Gesetz der großen Zahlen folgt aus dem starken Gesetz.

Mittels der Kovarianz können wir das schwache Gesetz direkt aus der Tschebycheff-Ungleichung gewinnen:

Beweis: Die Voraussetzungen sind so gewählt, dass (7.26) und (7.27) gelten, die Streuung konvergiert mit wachsendem n gegen null, die Werte der Zufallsvariablen liegen also immer näher beim Erwartungswert. Wir konkretisieren diese Anschauung mit der Ungleichung von Tschebycheff:

$$\begin{aligned} 0 &\leq P\left(\left|\frac{1}{n}\sum_{k=1}^n X_k - E(X_1)\right| > \varepsilon\right) \stackrel{(7.27)}{=} P\left(\left|\frac{1}{n}\sum_{k=1}^n X_k - E\left(\frac{1}{n}\sum_{k=1}^n X_k\right)\right| > \varepsilon\right) \\ &\stackrel{(7.23)}{\leq} \frac{\text{Var}\left(\frac{1}{n}\sum_{k=1}^n X_k\right)}{\varepsilon^2} \stackrel{(7.26)}{=} \frac{\text{Var}(X_1)}{n\varepsilon^2} \rightarrow 0 \quad (n \rightarrow \infty). \quad \square \end{aligned}$$

7.2.8 Zentraler Grenzwertsatz

Das Gesetz der großen Zahlen besagt, dass das arithmetische Mittel von Zufallsvariablen gegen den gemeinsamen Erwartungswert strebt: $\frac{1}{n}\sum_{i=1}^n [X_i - E(X_1)] \rightarrow 0$, $n \rightarrow \infty$. Der Zentrale Grenzwertsatz gibt eine Aussage für $\frac{\sqrt{n}}{n}\sum_{i=1}^n [X_i - E(X_1)]$. Durch den Faktor $\sqrt{n}$ zieht man den Definitionsbereich des Graphen der Verteilungsfunktion bzw. der Dichte auseinander, so dass man genauer sieht, wie die mit dem Gesetz der großen Zahlen beschriebene Konvergenz gegen den Erwartungswert funktioniert. Eine grobe Tendenz kann man bereits in Abbildung 7.13 erkennen.

Durch den Faktor $\sqrt{n}$ sorgt man dafür, dass die Varianz der Summenverteilung identisch verteilter unabhängiger Zufallsvariablen konstant bleibt und nicht gegen 0 geht (vgl. Satz 7.11):

$$\text{Var}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^n X_k\right) = \frac{1}{n}\sum_{k=1}^n \text{Var}(X_k) = \frac{1}{n}n \text{Var}(X_1) = \text{Var}(X_1).$$

Subtrahiert man den gemeinsamen Erwartungswert, so ändert sich daran nichts:

$$\text{Var}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^n [X_k - E(X_1)]\right) = \frac{1}{n}\sum_{k=1}^n \text{Var}(X_k - E(X_1)) = \frac{1}{n}\sum_{k=1}^n \text{Var}(X_k) = \text{Var}(X_1).$$

So kann die Verteilung von $\frac{1}{\sqrt{n}}\sum_{k=1}^n [X_k - E(X_1)]$ gegen eine Grenzverteilung konvergieren. Bevor wir dies (mit dem zusätzlichen Normierungsfaktor $\text{Var}(X_1)^{-\frac{1}{2}}$) präzisieren, vorab ein paar Definitionen:

Die Funktion $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ ist die **Dichte der Standardnormalverteilung** bzw. die Dichte der **Gauß'schen Normalverteilung** (siehe Abbildungen 7.14, 7.15 und Kapitel 7.2.5.3) und ist vom ehemaligen 10-D-Mark-Schein bekannt. Integriert man über die Dichte, so erhält man die Verteilungsfunktion Φ der Standardnormalverteilung (vgl. (7.16) auf Seite 1024)

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

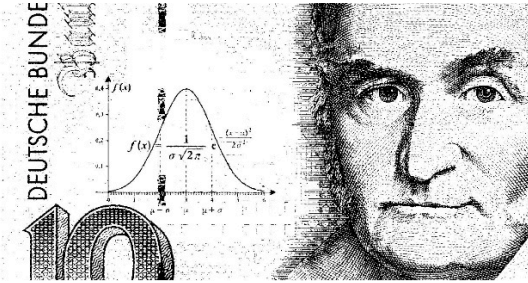


Abb. 7.14: Die Dichtefunktion der Standardnormalverteilung zusammen mit einem Porträt von Gauß auf dem alten 10-D-Mark-Schein (Foto: Deutsche Bundesbank)

Φ ist Verteilungsfunktion einer nicht-diskreten Zufallsvariable, die mehr als nur abzählbar unendlich viele Werte annimmt (vgl. Aufgabe 7.23). Die Wahrscheinlichkeit, dass die Zufallsvariable einen Wert $\leq x$ annimmt, ist also $\Phi(x)$. Einige Werte von Φ sind in Tabelle 7.1 aufgelistet. Die zugehörige **Standardnormalverteilung** (das als Verteilung zugehörige Wahrscheinlichkeitsmaß auf $\mathbb{R}$) mit Erwartungswert 0 (siehe (7.20) auf Seite 1035) und Varianz 1 (siehe (7.23) auf Seite 1039) wird mit $\mathcal{N}(0, 1)$ bezeichnet. Es ist $\lim_{x \rightarrow \infty} \Phi(x) = 1$, und wegen der Symmetrie der Dichte gilt: $\Phi(0) = \frac{1}{2}$ und für $x > 0$:

$$\Phi(x) - \Phi(-x) = 2(\Phi(x) - \Phi(0)) = 2\Phi(x) - 1. \quad (7.28)$$

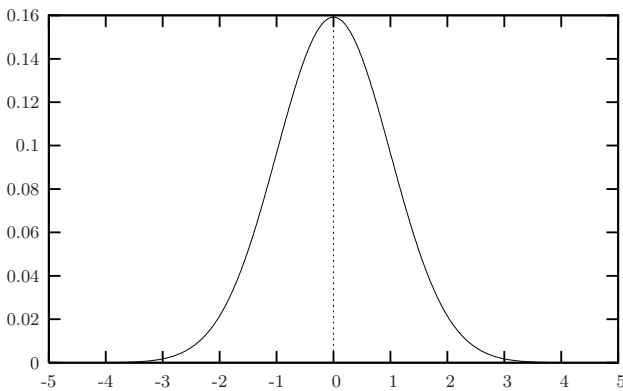


Abb. 7.15: Dichtefunktion $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ der Standardnormalverteilung

In der Natur beobachtet man häufig Verteilungsfunktionen $F^X(x)$, die sehr nah bei $\Phi(x)$ sind, wenn man Φ um den Erwartungswert von X verschiebt und mit der Standardabweichung $\sqrt{\text{Var}(X)}$ streckt. Für die Wahrscheinlichkeit $F^X(x)$, dass man einen Wert $\leq x$ beobachtet, gilt dann

$$\left| F^X(x) - \Phi\left(\frac{x - E(X)}{\sqrt{\text{Var}(X)}}\right) \right| < \varepsilon$$

Tab. 7.1: Näherungswerte der Verteilungsfunktion $\Phi(x)$ der Standardnormalverteilung mit Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1$. Man beachte, dass für eine standardnormalverteilte Zufallsvariable X gilt: $P(X \leq x) = \Phi(x)$ und $P(X \in [-x, x]) = 2 [\Phi(x) - \frac{1}{2}]$

x	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3
$\Phi(x)$	0,5000	0,5398	0,5793	0,6179	0,6554	0,6915	0,7257	0,7580	0,7881	0,8159	0,8413	0,8643	0,8849	0,9032
x	1,4	1,5	1,6	1,7	1,8	1,9	2	2,1	2,2	2,3	2,4	2,5	2,6	2,7
$\Phi(x)$	0,9192	0,9332	0,9452	0,9554	0,9641	0,9713	0,9772	0,9821	0,9861	0,9893	0,9918	0,9938	0,9953	0,9965
x	2,8	2,9	3	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4	4,1
$\Phi(x)$	0,9974	0,9981	0,9987	0,9990	0,9993	0,9995	0,9997	0,9998	0,9998	0,9999	0,9999	1,0000	1,0000	1,0000
$\Phi(x)$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999					
x	1,6449	1,7507	1,8808	2,0537	2,3263	2,5758	3,0902	3,2905	3,7190					

für ein hinreichend kleines ε und alle $x \in \mathbb{R}$. Man sagt in diesem Fall, dass X annähernd $\mathcal{N}(\mathbb{E}(X), \text{Var}(X))$ -verteilt ist, wobei $\mathcal{N}(\mu, \sigma^2)$ die (nicht-diskrete) **Normalverteilung** mit Erwartungswert μ und Varianz σ^2 bezeichnet. Die übliche Darstellung der Verteilungsfunktion erhalten wir mit der Substitution $u = \sigma \cdot t + \mu$, $t = \frac{u-\mu}{\sigma}$, $du = \sigma dt$:

$$\Phi\left(\frac{x-\mu}{\sqrt{\sigma^2}}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-\mu)^2}{2\sigma^2}} du.$$

Die zugehörige Dichtefunktion ist

$$\frac{d}{dx} \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}.$$

Dass die Natur sich häufig so verhält, liegt daran, dass sehr viele Dinge unabhängig (und unkontrollierbar) voneinander ablaufen und man so arithmetische Mittel unabhängiger Experimente beobachtet. In solchen Situationen wird der Zentrale Grenzwertsatz anwendbar sein und das Verhalten der Natur mathematisch begründen.

Ist X annähernd $\mathcal{N}(\mu, \sigma^2)$ -verteilt mit $\mu = \mathbb{E}(X)$ und $\sigma^2 = \text{Var}(X)$, so ist

$$P(X \in [\mu - c, \mu + c]) = P^X([\mu - c, \mu + c]) \approx \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(\frac{-c}{\sigma}\right) = 2\Phi\left(\frac{c}{\sigma}\right) - 1$$

die Wahrscheinlichkeit, dass ein Wert von X nicht weiter als c vom Erwartungswert entfernt liegt. Dies ist umgekehrt gleichzeitig die Wahrscheinlichkeit, dass der Erwartungswert μ im Intervall $[X - c, X + c]$ liegt, also ist

$$\begin{aligned} P(\mu \in [X - c, X + c]) &:= P(\{\omega \in \Omega : \mu \in [X(\omega) - c, X(\omega) + c]\}) \\ &= P^X([\mu - c, \mu + c]) \approx 2\Phi\left(\frac{c}{\sigma}\right) - 1. \end{aligned} \tag{7.29}$$

Auf der Basis eines Experiments (eines Werts $x = X(\omega)$) kann man damit eine Prognose über den Erwartungswert treffen, die mit der angegebenen Wahrscheinlichkeit stimmt. Auf dieser Abschätzung beruhen viele Verfahren der schließenden Statistik, bei denen man weiß oder annimmt, dass man es mit einer Normalverteilung zu tun hat.

Der folgende Satz konkretisiert die eingangs gemachte Überlegung, wie die Konvergenz beim Gesetz der großen Zahlen genauer geschieht, und ist der Grund dafür, dass tatsächlich sehr oft eine Normalverteilung vorliegt:

Satz 7.14 (Zentraler Grenzwertsatz)

Sei $(X_i)_{i=1}^\infty$ eine Folge identisch verteilter (diskreter) Zufallsvariablen auf Ω mit existierenden Erwartungswerten $\mu := E(X_1) = E(X_n)$, wobei die $X_1, \dots, X_n$ für jedes $n \geq 2$ stochastisch unabhängig seien. Weiter sei $0 < \text{Var}(X_1) = \text{Var}(X_n) =: \sigma^2 < \infty$. Wir setzen

$$Y_n := \frac{\sqrt{n}}{\sigma} \left[\frac{1}{n} \sum_{i=1}^n X_i - \mu \right],$$

wobei wir wie oben durch die Subtraktion von μ den Erwartungswert von Y_n auf 0 und durch den Faktor $\frac{\sqrt{n}}{\sigma}$ die Varianz auf 1 normieren.

Dann existiert zu jedem (noch so kleinen) $\varepsilon > 0$ eine Stelle $n_0 \in \mathbb{N}$ unabhängig von $x \in \mathbb{R}$, so dass für alle $n > n_0$ gilt:

$$\left| F^{Y_n}(x) - \Phi(x) \right| < \varepsilon.$$

Für jedes $x \in \mathbb{R}$ gilt also $\lim_{n \rightarrow \infty} F^{Y_n}(x) = \Phi(x)$. Die Verteilungsfunktionen der Zufallsvariablen Y_n streben gegen die Verteilungsfunktion der Standardnormalverteilung. Der Satz sagt darüber hinaus, dass das n_0 sogar unabhängig von x gewählt werden kann und damit diese Konvergenz gleichmäßig ist (vgl. Definition 2.43 auf Seite 404).

Mit dem Zentralen Grenzwertsatz kann man also abschätzen, wie wahrscheinlich es bei großem n ist, dass Y_n einen Wert aus $[a, b]$ annimmt, nämlich ungefähr $\Phi(b) - \Phi(a)$. Die Wahrscheinlichkeit entspricht der Fläche unter der Kurve der Dichte der Standardnormalverteilung im Intervall $[a, b]$, siehe Abbildung 7.15.

Der Beweis würde den Rahmen dieses Kapitels sprengen. Der heute übliche, auf Lyapunov zurückgehende Beweis vom Beginn des 20. Jahrhunderts nutzt die Fourier-Transformation. Dabei erhält man für jedes feste x die Konvergenzaussage. Die Unabhängigkeit der Konvergenz von x wird z. B. in (Bauer, 1968, S. 223) gezeigt.

Beispiel 7.54 (Summenverteilung)

Die Verteilung einer Summe von stochastisch unabhängigen Zufallsvariablen berechnet sich über die Faltung, die aus dem Kapitel über Fourier-Analyse bekannt ist. Wie betrachten wieder einen Wurf mit zwei Würfeln. Dabei möge X_1 das Ergebnis des ersten Würfels und X_2 das (vom ersten stochastisch unabhängige) Ergebnis des zweiten Würfels angeben. Dann ergeben sich die Werte der Summenverteilung wie folgt:

$$P(X_1 + X_2 = 2) = P(X_1 = 1)P(X_2 = 1) = \frac{1}{36},$$

$$P(X_1 + X_2 = 3) = P(X_1 = 2)P(X_2 = 1) + P(X_1 = 1)P(X_2 = 2) = \frac{2}{36},$$

$$\begin{aligned}
 P(X_1 + X_2 = 4) &= P(X_1 = 3)P(X_2 = 1) + P(X_1 = 2)P(X_2 = 2) \\
 &\quad + P(X_1 = 1)P(X_2 = 3) = \frac{3}{36}, \dots
 \end{aligned}$$

Die stochastische Unabhängigkeit haben wir ausgenutzt, da wir Wahrscheinlichkeiten multipliziert haben. Wahrscheinlichkeiten können hier summiert werden, da die zugehörigen Ereignisse disjunkt sind.

Man erkennt, dass dies die Komponenten der Faltung der beiden Vektoren

$$\begin{aligned}
 &(P(X_1 = 1), P(X_1 = 2), P(X_1 = 3), \dots, P(X_1 = 6), 0, 0, \dots, 0) \\
 \text{und } &(P(X_2 = 1), P(X_2 = 2), P(X_2 = 3), \dots, P(X_2 = 6), 0, 0, \dots, 0)
 \end{aligned}$$

sind (siehe Kapitel 6.4.3). Da sich Summenverteilungen über die Faltung berechnen lassen, ist es nicht verwunderlich, dass der übliche Beweis des Zentralen Grenzwertsatzes mit Mitteln der Fourier-Analyse geführt wird. Dabei hat man einen Faltungssatz zur Verfügung und kann statt der Faltung ein einfaches Produkt ausrechnen. ■

Man kennt annähernd die Verteilung von Y_n , obwohl man über die Verteilungen der X_k nichts weiß, außer dass die X_k stochastisch unabhängig sind. Wenn man also die Werte von X_k als Ergebnisse von unabhängigen Experimenten versteht, so kann man bei genügend großer Experimentanzahl die Ergebnisse mit einer Gauß-Verteilung beschreiben!

Betrachtet man zu identisch verteilten, unabhängigen Zufallsvariablen das arithmetische Mittel

$$Z_n := \frac{1}{n} \sum_{k=1}^n X_k,$$

so ist $Y_n = \frac{\sqrt{n}}{\sigma}(Z_n - \mu)$ und damit $Z_n = \frac{\sigma}{\sqrt{n}}Y_n + \mu$. Damit erhalten wir für die Verteilungsfunktion von Z_n :

$$\begin{aligned}
 F^{Z_n}(x) &= P^{Z_n}([-\infty, x]) = P(\{\omega \in \Omega : Z_n(\omega) \leq x\}) \\
 &= P\left(\left\{\omega \in \Omega : \frac{\sigma}{\sqrt{n}}Y_n(\omega) + \mu \leq x\right\}\right) \\
 &= P\left(\left\{\omega \in \Omega : Y_n(\omega) \leq \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}\right\}\right) = F^{Y_n}\left(\frac{\sqrt{n}}{\sigma}(x - \mu)\right).
 \end{aligned}$$

Das in Satz 7.14 zu betrachtende Argument der Verteilungsfunktion hängt jetzt allerdings von n ab. Glücklicherweise ist die Konvergenz aus dem Zentralen Grenzwertsatz gleichmäßig und damit unabhängig vom Argument der Verteilungsfunktion: Zu jedem $\varepsilon > 0$ gibt es ein n_0 , so dass für $n > n_0$ gilt:

$$\left| F^{Z_n}(x) - \Phi\left(\frac{\sqrt{n}}{\sigma}(x - \mu)\right) \right| = \left| F^{Y_n}\left(\frac{\sqrt{n}}{\sigma}(x - \mu)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(x - \mu)\right) \right| < \varepsilon. \quad (7.30)$$

Hier entdecken wir wieder das Gesetz der großen Zahlen: Mit wachsendem n nähert sich die Verteilungsfunktion immer mehr einer Funktion an, die an der Stelle des Erwartungswerts μ von 0 nach 1 springt.

Mit den Gesetz der großen Zahlen haben wir im vorangehenden Unterkapitel gesehen, dass das intuitive Verständnis von Wahrscheinlichkeit als relative Häufigkeit bei unendlicher Wiederholung eines Experiments mit der Theorie übereinstimmt. Dabei haben wir die Wahrscheinlichkeit über den Erwartungswert der Binomialverteilung erhalten. Jetzt können wir diese Konstruktion unter dem Blickwinkel des Zentralen Grenzwertsatzes erneut betrachten.

Satz 7.15 (Grenzwertsatz von de Moivre und Laplace)

Sei die Zufallsvariable X binomialverteilt mit $X \sim \text{Bi}(n; p)$. Dann gilt für jedes $x \in \mathbb{R}$:

$$P(X \leq x) = \sum_{0 \leq k \leq \min\{n, x\}} \binom{n}{k} p^k (1-p)^{n-k} = \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) + \varepsilon_n,$$

wobei ε_n unabhängig von x ist mit $\lim_{n \rightarrow \infty} \varepsilon_n = 0$.

Wählt man also n genügend groß, so kann man statt mit der Binomialverteilung mit der Normalverteilung $\mathcal{N}(np; np(1-p))$ arbeiten:

$$P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$

Beweis: Hier kann man den Grenzwert ohne Wahrscheinlichkeitsrechnung als Fol-gengrenzwert mit den Mitteln der Analysis ausrechnen. Wir wählen aber einen anderen Weg und berechnen ihn wie angekündigt über den Zentralen Grenzwertsatz. Dazu betrachten wir n identisch verteilte und stochastisch unabhängige Zufallsvariablen $X_1, \dots, X_n$ mit $P(X_k = 1) = p$, $P(X_k = 0) = 1 - p$, so dass wir wie im Beispiel zur Berechnung von π auf Seite 1046 die binomialverteilte Zufallsvariable X darstellen als

$$X = \sum_{k=1}^n X_k.$$

Es ist $\mu := E(X_k) = p$ und $\sigma^2 := \text{Var}(X_k) = E(X_k^2) - E(X_k)^2 = p - p^2 = p(1-p)$.

Für die Zufallsvariable $\frac{1}{n}X$ gilt nach (7.30) für $t \in \mathbb{R}$:

$$F_n^{\frac{1}{n}X}(t) = P\left(\frac{1}{n}X \leq t\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}(t - \mu)\right) + \varepsilon_n = \Phi\left(\frac{\sqrt{n}}{\sqrt{p(1-p)}}(t - p)\right) + \varepsilon_n,$$

wobei der Fehler ε_n der Approximation unabhängig von t ist mit $\lim_{n \rightarrow \infty} \varepsilon_n = 0$. Setzen wir $x := nt$, dann erhalten wir

$$P(X \leq x) = \Phi\left(\frac{\sqrt{n}}{\sqrt{p(1-p)}}\left(\frac{x}{n} - p\right)\right) + \varepsilon_n = \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) + \varepsilon_n. \quad \square$$

In praktischen Anwendungen rechnet man häufig ab $np(1-p) > 9$ mit der Normal- statt der Binomialverteilung.

Beispiel 7.55

Die Zahl π lässt sich mit einer Monte-Carlo-Simulation ermitteln, siehe Beispiel 7.42 auf Seite 1028. Gibt die Zufallsvariable X_n die Anzahl der Treffer im Kreis bei n Experimenten an, so ist X_n binomialverteilt mit $X_n \sim \text{Bi}(n; \frac{\pi}{4})$ und $E(X_n) = n \cdot \frac{\pi}{4}$. Möchten wir abschätzen, wie viele Versuche n wir benötigen, bis die Wahrscheinlichkeit, dass $\frac{4}{n} \cdot X_n$ weniger als ein vorgegebenes ε von π abweicht, größer als 0,999 ist, so

können wir näherungsweise mit der Verteilungsfunktion $F^{X_n}(x) \approx \Phi\left(\frac{x - n\frac{\pi}{4}}{\sqrt{n\frac{\pi}{4}(1-\frac{\pi}{4})}}\right)$ rechnen.

$$\begin{aligned} P\left(\left|\frac{4}{n} \cdot X_n - \pi\right| < \varepsilon\right) &= P\left(\left|X_n - n\frac{\pi}{4}\right| < \frac{n}{4}\varepsilon\right) = F^{X_n}\left(\frac{n}{4}\varepsilon + n\frac{\pi}{4}\right) - F^{X_n}\left(-\frac{n}{4}\varepsilon + n\frac{\pi}{4}\right) \\ &\approx \Phi\left(\frac{\frac{n}{4}\varepsilon}{\sqrt{n\frac{\pi}{4}(1-\frac{\pi}{4})}}\right) - \Phi\left(-\frac{\frac{n}{4}\varepsilon}{\sqrt{n\frac{\pi}{4}(1-\frac{\pi}{4})}}\right) = 2\Phi\left(\sqrt{n}\frac{\varepsilon}{\sqrt{4\pi - \pi^2}}\right) - 1. \end{aligned}$$

Für $n \rightarrow \infty$ strebt die rechte Seite gegen $2 \cdot 1 - 1 = 1$. Damit kann die Anzahl n nun so mit einem Tabellenwerk für Φ wie Tabelle 7.1 bestimmt werden, dass die Wahrscheinlichkeit größer als 0,999 ist. Dazu wählen wir einen möglichst kleinen Wert x mit $2\Phi(x) - 1 > 0,999 \iff \Phi(x) > 0,9995$, also z. B. $x = 3,4$. Damit gewinnen wir wegen der Monotonie von Φ einen Wert für n aus

$$\sqrt{n}\frac{\varepsilon}{\sqrt{4\pi - \pi^2}} \geq x = 3,4 \implies n \geq (3,4)^2(4\pi - \pi^2)\frac{1}{\varepsilon^2}. \quad \blacksquare$$

Man verwendet Satz 7.15 häufig mit einem zusätzlichen Korrekturterm $\frac{1}{2}$, der die Konvergenz verbessert. Es gilt auch

$$P(X \leq x) = \sum_{0 \leq k \leq x} \binom{n}{k} p^k (1-p)^{n-k} = \Phi\left(\frac{x - np + \frac{1}{2}}{\sqrt{np(1-p)}}\right) + \varepsilon'_n,$$

wobei ε'_n ebenfalls unabhängig von x ist mit $\lim_{n \rightarrow \infty} \varepsilon'_n = 0$. Dies folgt direkt aus dem Satz, denn nach dem Mittelwertsatz der Differenzialrechnung (Satz 2.36) ist

$$\begin{aligned} \left|\Phi\left(\frac{x - np + \frac{1}{2}}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)\right| &= \left|\Phi'(\xi)\frac{1}{2\sqrt{np(1-p)}}\right| \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\xi^2/2) \frac{1}{2\sqrt{np(1-p)}} \leq \frac{1}{\sqrt{2\pi}} \frac{1}{2\sqrt{np(1-p)}} \rightarrow 0, \quad n \rightarrow \infty. \end{aligned}$$

7.2.9 Integrale über Zufallsvariablen*

Eine Zufallsvariable X ist eine Abbildung von der Menge der Elementarereignisse Ω in die reellen Zahlen. Wir wollen hier X über Ω integrieren, um so eine äquivalente Definition des Erwartungswerts zu erhalten. Dabei werden wir außerdem sehen, dass man Zufallsvariablen ohne das Attribut „diskret“ definieren kann, so wie dies in der Literatur auch üblich ist.

Problematisch ist, dass Ω eine Menge von irgend welchen Objekten ist, oft sind die Elementarereignisse keine Zahlen. Denn man verwendet genau deshalb Zufallsvariablen, um mit den Zahlen $X(\omega)$ für $\omega \in \Omega$ zu rechnen. Damit ist aber auch nicht klar, was ein Integral über Ω sein soll. Hier hilft das Lebesgue-Integral aus Kapitel 2.6.8. Beim Lebesgue-Integral wird eine σ -Algebra von Lebesgue-messbaren Mengen definiert. Hier haben wir eine σ -Algebra von Ereignissen (Ereignis-Algebra), die hinsichtlich eines Wahrscheinlichkeitsmaßes P messbar sind. Wir sagen, eine Menge ist P -messbar genau dann, wenn sie Element der Ereignis-Algebra ist.

Wir können nun mit dem Wahrscheinlichkeitsmaß jeder P -messbaren Menge einen Inhalt zwischen 0 und 1 zuordnen. Nun kann man analog zu den Kapiteln 2.6.8.2 und 2.6.8.3 messbare Funktionen und einen Integralbegriff einführen, für den weitgehend die gleichen Eigenschaften wie für das Lebesgue-Integral gelten. Wir verwenden die Notation

$$P\text{-} \int_A X(\omega) d\omega,$$

wobei A ein Ereignis und $X : \Omega \rightarrow \mathbb{R}$ eine P -integrierbare Funktion sei.

Insbesondere gilt für jedes Ereignis A :

$$P(A) = \int_A 1 d\omega.$$

Beispiel 7.56

Wir betrachten ein Würfelexperiment mit $\Omega = \{1, 2, 3, 4, 5, 6\}$. Die Ereignis-Algebra sei die Potenzmenge von Ω , also die Menge aller Teilmengen von Ω . Das Wahrscheinlichkeitsmaß ist hier als Zähldichte über $P(\{\omega\}) = \frac{1}{6}$, $\omega \in \Omega$, definiert. Die Zufallsvariable $X : \Omega \rightarrow \mathbb{R}$ sei 1 bei einem geraden Wurf, also falls das Ereignis $\{2, 4, 6\}$ vorliegt, und X sei 3 bei einem ungeraden Wurf, also wenn das Ereignis $\{1, 3, 5\}$ eintritt. Dann ist

$$\begin{aligned} P\text{-} \int_{\Omega} X(\omega) d\omega &= 1 \cdot P(X^{-1}(\{1\})) + 3 \cdot P(X^{-1}(\{3\})) \\ &= 1 \cdot P(\{2, 4, 6\}) + 3 \cdot P(\{1, 3, 5\}) = \frac{1}{2} + \frac{3}{2} = 2. \end{aligned}$$

■

Wir haben diskrete Zufallsvariablen als Abbildungen $X : \Omega \rightarrow \mathbb{R}$ eingeführt, die nur endlich oder abzählbar unendlich viele Werte annehmen und für deren Werte eine Wahrscheinlichkeit erklärt ist. Die letzte Bedingung entspricht genau der P -Messbarkeit von X . Auf die Bedingung, dass höchstens abzählbar viele Werte angenommen werden, kann verzichtet werden, wenn man das P -Integral verwendet. Ohne diese Bedingung spricht man statt von diskreten Zufallsvariablen nur von **Zufallsvariablen**.

Die Abzählbarkeit ist erforderlich, wenn man beispielsweise den Erwartungswert ohne P -Integral definiert. Für eine diskrete Zufallsvariable ist er definiert als

$$E(X) := \sum_{x \in X(\Omega)} xP(X = x),$$

sofern die Summe absolut konvergiert (siehe 7.24 auf Seite 1033). Ohne die Abzählbarkeit des Wertebereichs $X(\Omega)$ macht die Summe keinen Sinn. Jetzt können wir aber den Erwartungswert mit dem Integral

$$E(X) := P\text{-}\int_{\Omega} X(\omega) d\omega$$

definieren und benötigen dazu lediglich die (weitaus allgemeinere) P -Integrierbarkeit von X . Tatsächlich sind beide Definitionen für einen abzählbaren Wertebereich gleich. So haben wir im obigen Beispiel den Erwartungswert berechnet. Allgemeiner gilt mit (2.66), Seite 380, übertragen auf das P -Integral (beachte, dass die Urbilder der Werte von X disjunkt sind und ihre Vereinigung ganz Ω ergibt):

$$\begin{aligned} P\text{-}\int_{\Omega} X(\omega) d\omega &= \sum_{x \in X(\Omega)} \int_{X^{-1}(\{x\})} X(\omega) d\omega = \sum_{x \in X(\Omega)} \int_{X^{-1}(\{x\})} x d\omega \\ &= \sum_{x \in X(\Omega)} P(X^{-1}(\{x\}))x = \sum_{x \in X(\Omega)} xP(X = x). \end{aligned}$$

Die über den Erwartungswert definierte Varianz $\text{Var}(X) = E((X - E(X))^2)$ ist nun ebenfalls über das P -Integral erklärt:

$$\text{Var}(X) = P\text{-}\int_{\Omega} [X(\omega) - E(X)]^2 d\omega.$$

Beispiel 7.57

Wir betrachten $\Omega = [0, 1]$, und als Ereignis-Algebra wählen wir die σ -Algebra der Lebesgue-messbaren Teilmengen von $[0, 1]$. $P = m$ sei das Lebesgue-Maß m , das wegen $m([0, 1]) = 1$ tatsächlich ein Wahrscheinlichkeitsmaß ist. Weiter sei $X : [0, 1] \rightarrow \mathbb{R}$ eine Lebesgue-integrierbare und damit P -integrierbare Funktion, also eine Zufallsvariable. Es gilt:

$$E(X) = P\text{-}\int_{\Omega} X(\omega) d\omega = L\text{-}\int_0^1 X(t) dt.$$

Wählen wir speziell $X(\omega) := \omega^2$, so nimmt X überabzählbar viele Funktionswerte an, es handelt sich also nicht um eine diskrete Zufallsvariable. Ihr Erwartungswert berechnet sich mit dem klassischen Riemann-Integral:

$$\mathbb{E} \int_0^1 t^2 dt = \int_0^1 t^2 dt = \frac{1}{3}[1^3 - 0^3] = \frac{1}{3}. \quad \blacksquare$$

7.3 Schließende Statistik

Wir setzen in diesem Kapitel die Ergebnisse der Wahrscheinlichkeitsrechnung ein, um auf der Basis unvollständiger Daten Aussagen zu treffen. Dabei können diese Aussagen nur mit einer gewissen Wahrscheinlichkeit gelten. Beispiele sind Wahlprognosen, Qualitätsprüfungen auf der Basis von Stichproben, Wettervorhersagen oder Monte-Carlo-Simulationen analoger Schaltkreise. Die Grundidee dabei ist das Gesetz der großen Zahlen und der Zentrale Grenzwertsatz. Obwohl man die Wahrscheinlichkeiten der Ausgänge eines Experiments nicht kennt, kann man durch hinreichend häufige unabhängige Wiederholung dieses Experiments und Übergang zum arithmetischen Mittel der Ergebnisse etwas über Wahrscheinlichkeiten sagen. Denn das arithmetische Mittel konvergiert gegen den Erwartungswert und ist annähernd normalverteilt – unabhängig davon, welche Verteilung das einzelne Experiment hat.

Das Kapitel fasst kurz einige wichtige Schätzverfahren zusammen. In (Sachs, 2003, Kapitel 4) finden Sie darüber hinaus weitere Beispiele und Verfahren, die in einer ähnlichen Notation dargestellt sind.

In der beschreibenden Statistik haben wir die komplette statistische Masse (Grundgesamtheit) zur Verfügung. Die schließende Statistik dagegen beruht auf den Daten einer Stichprobe, die nur eine (in der Regel kleine) Teilmenge der statistischen Masse ist. Damit die Ergebnisse so aussehen, als hätte man die komplette statistische Masse untersucht, muss die Stichprobe die statistische Masse geeignet repräsentieren. Man spricht von einer **repräsentativen Stichprobe**. So sollten die relativen Häufigkeiten von Merkmalsausprägungen übereinstimmen. Man wählt dazu aber nicht gezielt Merkmalsträger aus der Grundgesamtheit aus, sondern überlässt dies dem Zufall. Dadurch stellen sich bei einer genügend großen Stichprobe automatisch annähernd die Häufigkeiten ein, die in der statistischen Masse vorliegen. Hier muss man aber dem Zufall eine Chance geben und darf die Stichprobe nicht durch eine gezielte Vorauswahl einschränken, vgl. (Quatember, 2008, S. 119).

Wir sind nun nicht an den Merkmalsträgern einer Stichprobe interessiert, sondern an der Ausprägung eines ihrer Merkmale. Daher können wir Zufallsvariablen verwenden, die uns die Merkmalsausprägungen der Stichprobe liefern:

Definition 7.27 (Stichprobe)

Unter einer **Zufallsstichprobe** vom Umfang n versteht man n stochastisch unabhängige und identisch verteilte (diskrete) Zufallsvariablen (**Stichprobenvariablen**) $X_1, \dots, X_n$, wobei X_k die reelle Merkmalsausprägung des k -ten Elements der Stichprobe liefert.

Wir können die Definition so interpretieren: Wir greifen n -mal in die statistische Masse und ziehen jedes mal einen Merkmalsträger. Das Elementarereignis ω möge den Vektor dieser Merkmalsträger beschreiben. Die Stichprobenwerte $X_1(\omega), \dots, X_n(\omega)$ sind dann die Merkmalsausprägungen zu den betrachteten n Merkmalsträgern.

Ist das Ziehen der Merkmalsträger eine Ziehen ohne Zurücklegen, können die Zufallsvariablen bei endlicher Grundgesamtheit nicht stochastisch unabhängig sein (siehe Bemerkung zur hypergeometrischen Verteilung auf Seite 1027). Wir gehen also entweder von einem Ziehen mit Zurücklegen oder einer unendlichen (oder sehr großen) Grundgesamtheit aus. So kann man leichter mit den Zufallsvariablen rechnen.

7.3.1 Punktschätzungen

Auf der Grundlage einer Stichprobe möchte man z. B. angeben, wie viele Wähler sich bei einer bevorstehenden Wahl für eine bestimmte Partei entscheiden. Der gesuchte Prozentsatz entspricht der Wahrscheinlichkeit, dass ein zufälliger Wähler (und damit eine Stichprobenvariable) diese Partei wählt. Ein naheliegender Ansatz ist nun, die relative Häufigkeit aller Wähler der Partei in der Stichprobe als Schätzung dieser Wahrscheinlichkeit zu nehmen. Diese relative Häufigkeit muss nicht mit der (unbekannten) relativen Häufigkeit in der kompletten statistischen Masse übereinstimmen. Die Schätzung ist also mit einem Fehler behaftet, der sicherlich auch vom Umfang der Stichprobe beeinflusst ist. In diesem Abschnitt sehen wir uns zunächst Schätzfunktionen und ihre wichtigen Eigenschaften an. Dabei kümmern wir uns noch nicht um quantitative Abschätzungen des Fehlers. Das holen wir später nach, wenn wir Konfidenzintervalle berechnen (siehe insbesondere Kapitel 7.3.3.2).

Bei Punktschätzungen ist ein fester, aber unbekannter Parameter θ gesucht, der sich über eine Rechenvorschrift aus dem zugrunde liegenden Wahrscheinlichkeitsraum ergibt. θ ist keine Zufallsvariable, sondern eine feste Konstante wie z. B. der Erwartungswert, die Varianz oder eine Wahrscheinlichkeit. Diese Größen sind unabhängig von den Ergebnissen einzelner Experimente und, auch wenn man sie im Allgemeinen nicht kennt, eindeutig festgelegt.

Man möchte nun den gesuchten Parameter aus dem Ergebnis endlich vieler Experimente näherungsweise ermitteln. Die Zufallsvariable X_k möge dazu das Ergebnis des k -ten Versuchs ausdrücken. Damit entsteht eine Stichprobe, die wir mit den Mitteln der beschreibenden Statistik auswerten können. Dabei werden wir sehen, dass wir mit

den Lage- und Streuungsparametern der beschreibenden Statistik die entsprechenden Parameter der Wahrscheinlichkeitsrechnung erhalten.

Definition 7.28 (Schätzer)

Sei $f : \mathbb{R}^n \rightarrow \mathbb{R}$ eine reelle Funktion mit n Variablen. Wir setzen in f die n Zufallsvariablen einer Stichprobe ein. Die so gebildete Funktion (**Stichprobenfunktion**) $\hat{\theta} : \Omega \rightarrow \mathbb{R}$ mit $\hat{\theta} := f(X_1, X_2, \dots, X_n)$ sei selbst bereits wieder eine (diskrete) Zufallsvariable. Wird ihr Wert zur Schätzung eines gesuchten Parameters θ verwendet, so heißt sie **Schätzfunktion** bzw. **Schätzer** für θ .

Man nimmt also Stichprobenwerte $X_1(\omega), \dots, X_n(\omega)$ und setzt diese in eine Rechenvorschrift (eine Funktion) f ein, um den Schätzwert des gesuchten Parameters zu erhalten. Die Stichprobe entsteht zufällig durch das beobachtete Elementarereignis $\omega \in \Omega$.

Beispiel 7.58

Das Stichprobenmittel

$$\hat{\mu} := \frac{1}{n} \sum_{k=1}^n X_k$$

ist eine Schätzfunktion für den (unbekannten, aber festen) Erwartungswert μ der identisch verteilten Zufallsvariablen X_k . Nach dem Gesetz der großen Zahlen konvergieren für $n \rightarrow \infty$ diese arithmetischen Mittel in einem geeigneten Sinne gegen den Erwartungswert. Diese Schätzfunktion hat zwei „schöne“ Eigenschaften:

a) Der Erwartungswert des Schätzers ist gleich dem gesuchten Parameter μ :

$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{k=1}^n X_k\right) = \frac{1}{n} \sum_{k=1}^n E(X_k) = E(X_1) = \mu.$$

Der Schätzer liefert also im Mittel tatsächlich den Wert, den er schätzen soll. Wenn er das nicht täte, dann hätten wir damit einen prinzipiellen Fehler.

b) Die Varianz des Schätzers konvergiert mit wachsendem n gegen null. Da die X_k insbesondere paarweise stochastisch unabhängig und damit unkorreliert sind (siehe Seite 1041), ist nach Satz 7.11

$$\text{Var}\left(\frac{1}{n} \sum_{k=1}^n X_k\right) = \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) = \frac{1}{n} \text{Var}(X_1) \rightarrow 0 \quad (n \rightarrow \infty).$$

Der Schätzer wird damit immer genauer, je größer der Stichprobenumfang wird. Streng genommen handelt es sich für jeden Stichprobenumfang n um eine andere Schätzfunktion, so dass wir eine Folge von Schätzern diskutieren. Den Folgenparameter n schreibt man aber in der Regel nicht an den Schätzer. ■

Definition 7.29 (Eigenschaften von Schätzern)

- a) Eine Schätzfunktion $\hat{\theta}$ heißt **erwartungstreu (unverfälscht, unbiased)** genau dann, wenn ihr Erwartungswert der gesuchte Parameter θ ist, also wenn $E(\hat{\theta}) = \theta$.
- b) Eine erwartungstreu Schätzfunktion $\hat{\theta}$ (genauer: eine Folge $(\hat{\theta}_n)_{n=1}^\infty$ von erwartungstreuen Schätzfunktionen) heißt **konsistent** genau dann, wenn bei wachsendem Stichprobenumfang die Varianz gegen null strebt:

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}) = \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0.$$

Nach der Ungleichung von Tschebycheff (siehe Seite 1040) gilt zum Beispiel für einen erwartungstreuen, konsistenten Schätzer:

$$P(|\hat{\theta} - E(\hat{\theta})| > \varepsilon) = P(|\hat{\theta} - \theta| > \varepsilon) \leq \frac{\text{Var}(\hat{\theta})}{\varepsilon^2},$$

wobei man durch Wahl der Stichprobengröße die rechte Seite beliebig klein gestalten kann.

Beispiel 7.59 (Schätzer für eine Wahrscheinlichkeit)

Wir haben das Kapitel mit der Frage begonnen, wie wir auf der Basis einer Stichprobe die Prozentpunkte einer Partei bei einer Wahl schätzen können. Dazu suchen wir die Wahrscheinlichkeit p , mit der ein beliebiger Wähler die Partei wählt.

Die n unabhängigen, identisch verteilten Stichprobenvariablen $X_1, \dots, X_n$ drücken das Ergebnis einer zufälligen Umfrage aus. Dabei nimmt X_k den Wert 1 an, wenn der k -te befragte Wähler die Partei wählen möchte. Sonst ist sie 0. Die Wahrscheinlichkeit, dass 1 angenommen wird, ist das gesuchte, aber unbekannte p , also $P(X_k = 1) = p$. Für den Erwartungswert gilt: $E(X_k) = 1 \cdot p + 0 \cdot (1 - p) = p$.

Wir haben bereits das Stichprobenmittel $\hat{\mu} := \frac{1}{n} \sum_{k=1}^n X_k$ als erwartungstreuen und konsistenten Schätzer für den Erwartungswert und damit für die gesuchte Wahrscheinlichkeit kennengelernt. Über die Güte dieser Schätzung machen wir uns in Abschnitt 7.3.3.2 Gedanken, siehe insbesondere Beispiel 7.63 auf Seite 1067. ■

Neben der Schätzfunktion für den Erwartungswert ist eine Schätzfunktion für die Varianz wichtig. Dazu nehmen wir die Formel für die empirische Varianz. Hier wird gleichzeitig der Erwartungswert μ der unabhängigen, identisch verteilten $X_1, \dots, X_n$ mitgeschätzt:

$$\hat{\sigma}^2 := \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{k=1}^n X_k \right)^2. \quad (7.31)$$

Auch dieser Schätzer ist erwartungstreu und konsistent. Wir rechnen die Erwartungstreue nach und benutzen dabei, dass für unabhängige Variablen X_k und X_l (für die die verwendeten Erwartungswerte existieren mögen) die Regel (7.24), also

$E(X_k X_l) = E(X_k) E(X_l)$, gilt. Da die Zufallsvariablen außerdem identisch verteilt sind, können wir $E(X_k) = E(X_1)$ und $E(X_k^2) = E(X_1^2)$ nutzen:

$$\begin{aligned}
 & E\left(\frac{1}{n-1} \sum_{l=1}^n \left(X_l - \frac{1}{n} \sum_{k=1}^n X_k\right)^2\right) \\
 &= \frac{1}{n-1} \sum_{l=1}^n \left[E(X_l^2) - \frac{2}{n} E\left(\sum_{k=1}^n X_l X_k\right) + \frac{1}{n^2} E\left(\sum_{k=1}^n \sum_{i=1}^n X_k X_i\right) \right] \\
 &= \frac{1}{n-1} \sum_{l=1}^n \left[E(X_l^2) - \frac{2}{n} E(X_l^2) - \frac{2}{n} \sum_{k \neq l} E(X_l) E(X_k) + \frac{1}{n^2} \sum_{k=1}^n E(X_k^2) \right. \\
 &\quad \left. + \frac{1}{n^2} \sum_{k=1}^n \sum_{i \neq k} E(X_k) E(X_i) \right] \\
 &= \frac{1}{n-1} \sum_{l=1}^n \left[\left[1 - \frac{2}{n} + \frac{n}{n^2}\right] E(X_1^2) - \left[\frac{2(n-1)}{n} - \frac{n(n-1)}{n^2}\right] E(X_1)^2 \right] \\
 &= \frac{n}{n-1} \left[\frac{n^2 - 2n + n}{n^2} E(X_1^2) - \frac{2n^2 - 2n - n^2 + n}{n^2} E(X_1)^2 \right] \\
 &= E(X_1^2) - E(X_1)^2 = \text{Var}(X_1).
 \end{aligned}$$

Dies ist der Grund, warum wir bei der empirischen Varianz durch $n-1$ und nicht durch n geteilt haben! Setzt man statt des Schätzers für den Erwartungswert einen bekannten Erwartungswert μ ein, muss man durch n teilen.

Werden wir also n unabhängige Zufallsexperimente mittels der beschreibenden Statistik aus, so ist die dabei berechnete empirische Varianz eine gute Schätzung für die stochastische Varianz.

Wenn wir später die Güte von Schätzungen untersuchen, benötigen wir dazu die Varianz der Stichprobenvariablen. Da diese in der Regel nicht bekannt ist, greifen wir auf die empirische Varianz zurück (siehe Kapitel 7.3.3.3).

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 notwendigerweise wie in der Definition 7.28 ein Parameter des Wahrscheinlichkeitsraums. Vielmehr wird dies nur angenommen. Die Annahme einer Wahrscheinlichkeitsverteilung lässt sich allerdings mit statistischen Mitteln überprüfen (siehe Kapitel 7.3.4.3).

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.

Beispiel 7.60

Nimmt man 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 $f''(\frac{1}{3}) = -6 < 0$ wird hier ein lokales Maximum angenommen. An den Intervallrändern $p = 1$ und $p = 0$ ist $f(p) = 0$. Da die stetige Funktion f auf $[0, 1]$ ein globales Maximum annimmt, liegt dieses bei $p = \frac{1}{3}$ mit Wert $\frac{4}{9}$. Damit ist $\frac{1}{3}$ die Maximum-Likelihood-Schätzung für p . ■

7.3.2 Begriffe der Fehlerrechnung*

Man misst eine physikalische Größe häufig n -mal, um sie durch Bildung des arithmetischen Mittels genauer angeben zu können. Jeder Messwert x_k ist das Ergebnis einer Zufallsvariable X_k , $1 \leq k \leq n$, wobei die X_k identisch verteilt und stochastisch unabhängig sind. Dabei schätzt man den (gemeinsamen) Erwartungswert $E(X_1)$. Macht man keinen systematischen Fehler (z. B. dejustiertes Messgerät), so sollte der Erwartungswert mit der gesuchten Größe übereinstimmen.

Die (gemeinsame) Varianz der Zufallsvariablen X_k lässt sich annähernd als s^2 mit dem Schätzer (7.31) bestimmen, der die empirische Varianz der gegebenen n Messwerte berechnet. Die empirische Standardabweichung $s = \sqrt{s^2}$ ist eine Schätzung der Standardabweichung der Zufallsvariablen und heißt in der Fehlerrechnung **mittlerer Fehler der Einzelmessungen**. Dadurch, dass man zum arithmetischen Mittel übergeht, reduziert man diesen Fehler. Die Zufallsvariable $\bar{X} := \frac{1}{n} \sum_{k=1}^n X_k$ hat die Varianz $\frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) = \frac{1}{n} \text{Var}(X_1)$ und die Standardabweichung $\sigma = \sqrt{\text{Var}(X_1)/n}$, die durch $\frac{s}{\sqrt{n}}$ geschätzt wird. Wie beim Gesetz der großen Zahlen und beim Zentralen Grenzwertsatz diskutiert, konvergiert die Standardabweichung gegen null für $n \rightarrow \infty$.

Definition 7.30 (Unsicherheit)

Man bezeichnet

$$u := \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{k=1}^n x_k\right)^2}}{\sqrt{n}} \quad (7.32)$$

als **mittleren Fehler des Mittelwerts** oder **Unsicherheit des Mittelwerts**.

Die Unsicherheit ist eine Schätzung der Standardabweichung des arithmetischen Mittels $\bar{X}$. Diese Zufallsvariable $\bar{X}$ ist als Konsequenz des Zentralen Grenzwertsatzes für große n annähernd normalverteilt, so dass man mit u die Standardabweichung und mit u^2 die Varianz dieser Normalverteilung annähernd kennt. Mit der Unsicherheit weiß man also sehr viel über das Verhalten der Mittelwerte.

Vielfach kann eine Größe nicht direkt gemessen werden, sondern sie berechnet sich über eine reellwertige Funktion $f(y_1, y_2, \dots, y_m)$ aus m anderen Messgrößen. Nun möchte man auf Basis der Daten der einzelnen Messgrößen auch Schätzungen für die zusammengesetzte Größe angeben. Dies haben wir bereits im Kontext der mehrdimensionalen Differenzialrechnung in Kapitel 4.2.4 diskutiert.

Seien $x_{1,1}, x_{1,2}, \dots, x_{1,n_1}$ die Messwerte der ersten Messgröße (die in y_1 eingesetzt werden) mit zugehörigem arithmetischem Mittel $\bar{x}_1$. Entsprechend werden die Messwerte der weiteren Größen bezeichnet. Die Werte für die m -te Messgröße seien $x_{m,1}, x_{m,2}, \dots, x_{m,n_m}$ mit Mittelwert $\bar{x}_m$. Die m Messgrößen mögen die Unsicherheiten $u_1, \dots, u_m$ haben.

Als arithmetisches Mittel der zusammengesetzten Größe verwenden wir

$$\frac{1}{n_1 n_2 n_3 \cdots n_m} \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \cdots \sum_{k_m=1}^{n_m} f(x_{1,k_1}, x_{2,k_2}, \dots, x_{m,k_m}),$$

wobei wir die Werte aller Einzelmessungen miteinander kombiniert haben.

Nun wollen wir die Unsicherheit u dieses Mittelwerts angeben. Das geht, wenn man die Funktion f in der Nähe des Vektors $(\bar{x}_1, \dots, \bar{x}_m)$ näherungsweise durch eine abgebrochene Taylor-Entwicklung erster Ordnung schreiben kann (siehe Satz 4.6 auf Seite 613, vgl. Kapitel 4.2.4)

$$f(y_1, y_2, \dots, y_m) \approx f(\bar{x}_1, \dots, \bar{x}_m) + \sum_{i=1}^m \frac{\partial f}{\partial y_i}(\bar{x}_1, \dots, \bar{x}_m)(\bar{x}_i - y_i)$$

und wenn die einzelnen Messgrößen unabhängig voneinander und damit unkorreliert sind. In diesem Fall gilt das Gauß'sche **Fehlerfortpflanzungsgesetz**

$$u \approx \sqrt{\sum_{k=1}^m \left(\frac{\partial f}{\partial y_k}(\bar{x}_1, \dots, \bar{x}_m) \cdot u_k \right)^2}. \quad (7.33)$$

Man kann dieses Gesetz zeigen, indem man mittels der abgebrochenen Taylor-Entwicklung das arithmetische Mittel näherungsweise darstellt als $f(\bar{x}_1, \dots, \bar{x}_m)$. Dann setzt man das Ergebnis für das arithmetische Mittel in (7.32) ein. Außerdem teilt man

dabei nicht durch $n - 1$, sondern durch n (was für große Werte kaum einen Unterschied macht), wobei wir für n nun $n_1 \cdot n_2 \cdots n_m$ setzen müssen:

$$u^2 \approx \frac{1}{(n_1 n_2 n_3 \cdots n_m)^2} \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \cdots \sum_{k_m=1}^{n_m} [f(x_{1,k_1}, x_{2,k_2}, \dots, x_{m,k_m}) - f(\bar{x}_1, \dots, \bar{x}_m)]^2.$$

Daraus ergibt sich durch erneute Taylor-Entwicklung von f und unter Ausnutzung der Unabhängigkeit der einzelnen Messgrößen das Gesetz (7.33).

7.3.3 Intervallschätzungen

Wir haben bislang nicht den möglichen Fehler betrachtet, den wir mit einer Punktschätzung machen. Hier helfen Intervallschätzungen, bei denen man keinen einzelnen Wert, sondern ein Intervall bestimmt, in dem der gesuchte Parameter mit angegebener hoher Wahrscheinlichkeit $1 - \alpha$ liegt. Beispielsweise ist die Prognose „Mit einer Wahrscheinlichkeit von 99 % gewinnt eine bestimmte Partei 29 % bis 31 % der Stimmen“ viel aussagekräftiger als „Voraussichtlich gewinnt die Partei 30 % der Stimmen“.

Definition 7.31 (Konfidenzintervall)

Sei $0 \leq \alpha \leq 1$ eine kleine Wahrscheinlichkeit. Ein Intervall, welches mit der großen Wahrscheinlichkeit $1 - \alpha$ den gesuchten Parameter θ beinhaltet, heißt **Konfidenzintervall (Vertrauensintervall) zum Niveau $1 - \alpha$ für θ** .

Man vertraut also diesem Intervall und kann das Ausmaß des Vertrauens (Konfident = Vertrauter) mit der Wahrscheinlichkeit $1 - \alpha$ angeben. Es ist üblich (wenn auch vielleicht etwas kompliziert aussehend), die Wahrscheinlichkeit hier mit $1 - \alpha$ zu bezeichnen, so dass α die (kleine) Fehlerwahrscheinlichkeit ist, dass der Parameter nicht im Konfidenzintervall liegt. Je größer $1 - \alpha$ gewählt wird, desto größer ist auch das Intervall. Für die Anwendungen reicht es in der Regel aus, wenn man ein möglichst kleines Intervall, das den Parameter mit **mindestens** der Wahrscheinlichkeit $1 - \alpha$ überdeckt, findet. Die Überdeckungswahrscheinlichkeit muss in der Regel nicht genau gleich $1 - \alpha$ sein.

7.3.3.1 Konfidenzintervall für den Erwartungswert bei großem Stichprobenumfang und bekannter Varianz

In diesem Abschnitt erweitern wir die Punktschätzung für den Erwartungswert zu einem Konfidenzintervall und geben damit die Genauigkeit der Schätzung an.

Als Stichprobenvariablen sind $X_1, \dots, X_n$ stochastisch unabhängig und identisch verteilt. Wir erstellen eine Intervallschätzung für den (gemeinsamen) Erwartungswert μ . Hierzu verwenden wir die bekannte Punktschätzung mit der Schätzfunktion

$$\hat{\mu} := Z_n := \frac{1}{n} \sum_{k=1}^n X_k$$

auf der Zufallsstichprobe $X_1, \dots, X_n$, wobei wir die (gemeinsame) Varianz σ^2 der Zufallsvariablen kennen. Es ist

$$\text{Var}(Z_n) = \frac{1}{n} \text{Var}(X_1) = \frac{\sigma^2}{n}.$$

Der Schätzer liefert uns eine Punktschätzung, aus der wir eine Intervallschätzung machen, indem wir das Intervall $[Z_n - c, Z_n + c]$ für ein $c > 0$ benutzen. Man beachte, dass das Intervall über die Stichprobe gebildet wird und damit das Ergebnis eines Zufallsexperiments ist. Wir konstruieren hier ein **zweiseitiges Konfidenzintervall**, da wir mit $Z_n - c$ sowohl eine untere als auch mit $Z_n + c$ eine obere Intervallgrenze ermitteln. Um dieses Konfidenzintervall zum Niveau $1 - \alpha$ zu erhalten, ist der Wert c nun so zu wählen, dass die Wahrscheinlichkeit, dass $\mu \in [Z_n - c, Z_n + c]$ ist, gleich $1 - \alpha$ ist. Diese Wahrscheinlichkeit ist gleich der Wahrscheinlichkeit von $|Z_n - \mu| \leq c$ bzw. von $Z_n \in [\mu - c, \mu + c]$.

Ist der Stichprobenumfang genügend groß ($n \geq 30$), können wir wegen des Zentralen Grenzwertsatzes für $Z_n = \frac{1}{n} \sum_{k=1}^n X_k$ die Approximation (7.30) benutzen:

$$P^{Z_n}([a, b]) = F^{Z_n}(b) - F^{Z_n}(a) \approx \Phi\left(\frac{\sqrt{n}}{\sigma}(b - \mu)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(a - \mu)\right).$$

Damit ist also (mit (7.28))

$$P^{Z_n}([\mu - c, \mu + c]) = \Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(-c)\right) = 2\Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - 1.$$

Demnach ist c so zu wählen, dass

$$2\Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - 1 = 1 - \alpha \iff \Phi\left(\frac{\sqrt{n}}{\sigma}c\right) = 1 - \frac{\alpha}{2}$$

ist. Dann ist die Wahrscheinlichkeit von $\mu \in [Z_n - c, Z_n + c]$ genau $1 - \alpha$.

Man kann in Abhängigkeit von $1 - \frac{\alpha}{2}$ folgende Urbilder x für Φ bestimmen, d. h. $\Phi(x) = 1 - \frac{\alpha}{2} = [1 - \alpha + 1]/2$ ($1 - \frac{\alpha}{2}$ -Quantil der Standardnormalverteilung, vgl. Tabelle 7.1 auf Seite 1049):

$$1 - \alpha = 0,9 \implies x \approx 1,645, \quad 1 - \alpha = 0,95 \implies x \approx 1,960, \quad 1 - \alpha = 0,99 \implies x \approx 2,576.$$

Damit erhält man

$$c = x \cdot \frac{\sigma}{\sqrt{n}} = x \cdot \sqrt{\text{Var}(Z_n)}.$$

Algorithmus zur Berechnung eines $1 - \alpha$ Konfidenzintervalls für den Erwartungswert bei bekannter Varianz σ^2 aus n Stichprobenwerten:

- a) Berechne einen geschätzten Erwartungswert als Wert von Z_n , d. h. als arithmetisches Mittel der n Stichprobenwerte.
- b) Zu α bestimme x mit $\Phi(x) = 1 - \frac{\alpha}{2}$ über eine Tabelle wie Tabelle 7.1 oder ein Computer-Algebra-Programm, z. B. $x \approx 1,645$ für $1 - \alpha = 0,9$.
- c) Berechne c als $c = \frac{\sigma}{\sqrt{n}}x$, also z. B. für $1 - \alpha = 0,9$: $c \approx \frac{\sigma}{\sqrt{n}}1,645$.

Damit liegt der unbekannte Erwartungswert μ mit der Wahrscheinlichkeit $1 - \alpha$ im Intervall $[Z_n - c, Z_n + c]$.

Beispiel 7.61

Mit einem Kerbschlagbiegeversuch wird in der Werkstoffprüfung das Zähigkeitsverhalten von Werkstoffen bestimmt. Dazu wird eine Probe eingekerbt. Ein Pendelhammer zerschlägt die Probe, indem er mit einer bestimmten kinetischen Energie auf die ungekerbte Rückseite der Probe aufschlägt. Durch die Verformung der Probe wird kinetische Energie absorbiert, und der Pendelhammer schwingt auf der anderen Seite weniger hoch.

Für Blöcke aus Stahl werden bei zehn Kerbschlagversuchen die folgenden Werte für die Kerbschlagarbeit W in Joule gemessen:

64,3; 64,6; 64,5; 64,7; 64,5; 64,1; 64,6; 64,8; 64,3; 64,2.

Die Kerbschlagarbeit W sei normalverteilt mit $\sigma = 1$. Wir bestimmen das Konfidenzintervall zum Niveau 0,95 für die mittlere Kerbschlagarbeit μ . Die benötigten Daten sind $n = 10$, $\alpha = 0,05$ sowie x mit $\Phi(x) = 0,975$, also $x \approx 1,96$ und $\bar{W} = 64,46$. Mit $c = x\sigma/\sqrt{n} = \frac{1,96 \cdot 1}{\sqrt{10}} = 0,62$ erhalten wir das Konfidenzintervall zum Niveau 0,95 aus der Ungleichung

$$\bar{W} - c \leq \mu \leq \bar{W} + c \iff 64,46 - 0,62 \leq \mu \leq 64,46 + 0,62 \iff \mu \in [63,84, 65,08].$$

Auf Basis der Messungen haben wir einen (hochplausiblen) Bereich für die Kerbschlagarbeit des verwendeten Stahls gefunden. ■

Die Definition des Konfidenzintervalls lässt auch einseitig unbegrenzte Konfidenzintervalle zu. Ersetzen wir in $[Z_n - c, Z_n + c]$ einen der beiden Randpunkte durch ∞ , so erhalten wir ein **einseitiges Konfidenzintervall** $]-\infty, Z_n + c]$ oder $[Z_n - c, \infty[$ zum vergrößerten Niveau $1 - \alpha + \frac{\alpha}{2} = 1 - \frac{\alpha}{2}$. Möchte man auch hier das Niveau $1 - \alpha$ haben, so kann man c wie oben berechnen und muss nur statt α den Wert 2α verwenden.

Insbesondere sehen wir, dass es viele Konfidenzintervalle zu einem vorgegebenen Niveau gibt. Eine sinnvolle Wahl hängt von der Anwendung des Intervalls ab.

7.3.3.2 Konfidenzintervall für eine Wahrscheinlichkeit

Das Vorgehen im vorangehenden Absatz ist unabhängig vom konkreten Schätzer für den Erwartungswert, solange dieser erwartungstreu und (annähernd) normalverteilt

ist. Betrachten wir statt der zuvor benutzten Z_n einen beliebigen Schätzer $\hat{\theta}$ für den Erwartungswert mit Varianz σ^2 , der diese Voraussetzungen erfüllt, so ist nach (7.29)

$$P(\mathbf{E}(\hat{\theta}) \in [\hat{\theta} - c, \hat{\theta} + c]) \approx 2\Phi\left(\frac{c}{\sigma}\right) - 1.$$

Mit einem $x > 0$, für das $\Phi(x) = 1 - \frac{\alpha}{2}$ gilt, und mit $c = x \cdot \sigma = x \cdot \sqrt{\text{Var}(\hat{\theta})}$ ist analog zum vorangehenden Abschnitt wieder

$$P\left(\mathbf{E}(\hat{\theta}) \in \left[\hat{\theta} - x \cdot \sqrt{\text{Var}(\hat{\theta})}, \hat{\theta} + x \cdot \sqrt{\text{Var}(\hat{\theta})}\right]\right) \approx 1 - \alpha,$$

d. h. $\left[\hat{\theta} - x \cdot \sqrt{\text{Var}(\hat{\theta})}, \hat{\theta} + x \cdot \sqrt{\text{Var}(\hat{\theta})}\right]$ ist ein Konfidenzintervall zum Niveau $1 - \alpha$ für den zu schätzenden Parameter θ_0 , der wegen der Erwartungstreue von $\hat{\theta}$ mit $\mathbf{E}(\hat{\theta})$ übereinstimmt.

Dies können wir anwenden, um ein Konfidenzintervall zum Niveau $1 - \alpha$ für eine unbekannte Wahrscheinlichkeit p für ein Ereignis zu berechnen. Wie in Beispiel 7.59 machen wir dazu n unabhängige Experimente und zählen, wie oft das Ereignis eingetreten ist. Dies ist der Wert einer $\text{Bi}(n; p)$ -verteilten Zufallsvariable X_n (Trefferzahl beim n -maligen Ziehen mit Zurücklegen) mit $\mathbf{E}(X_n) = np$ und bekannter $\text{Var}(X_n) = np(1-p)$.

Daraus wird mit $\hat{p} := \frac{1}{n}X_n$ ein erwartungstreuer Schätzer für die unbekannte Wahrscheinlichkeit p : Es ist $\mathbf{E}(\hat{p}) = \frac{1}{n}np = p$ und $\text{Var}(\hat{p}) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$. Um nun zum Konfidenzintervall zu kommen, nutzen wir aus, dass X_n $\text{Bi}(n; p)$ -verteilt ist. Mit dem Satz von de Moivre (Satz 7.15) lässt sich ihre Verteilung durch die einer Normalverteilung annähern. Genauer folgt, dass es eine Nullfolge $(\varepsilon_n)_{n=1}^\infty$ gibt, so dass für alle $t \in \mathbb{R}$:

$$P(\hat{p} \leq t)$$

$$= P\left(\frac{1}{n}X_n \leq t\right) = P(X_n \leq nt) = \Phi\left(\frac{nt - np}{\sqrt{np(1-p)}}\right) + \varepsilon_n = \Phi\left(\frac{t - p}{\sqrt{\frac{p(1-p)}{n}}}\right) + \varepsilon_n.$$

Damit ist der Schätzer $\hat{p}$ annähernd $\mathcal{N}(p; p(1-p)/n)$ -verteilt (bis auf einen kleinen Fehler ε_n), und wir können das eingangs angegebene Konfidenzintervall für den Erwartungswert einer normalverteilten Zufallsvariable anwenden. Für p erhalten wir also mit der aus Tabelle 7.1 zu bestimmenden Zahl x , die $\Phi(x) = 1 - \frac{\alpha}{2}$ erfüllt, das Konfidenzintervall

$$\left[\hat{p} - x \cdot \sqrt{\frac{p(1-p)}{n}}, \hat{p} + x \cdot \sqrt{\frac{p(1-p)}{n}}\right].$$

Das Problem dabei ist, dass hier noch p vorkommt. Man darf allerdings p durch dessen Punktschätzung $\hat{p}$ ersetzen ohne einen zu großen Fehler zu machen. Damit erhält man das gängige Wald-Intervall

$$\left[\hat{p} - x \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + x \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Allerdings ist die Wahrscheinlichkeit, dass kleine oder große gesuchte Wahrscheinlichkeiten p in diesem Intervall liegen, deutlich kleiner als $1 - p$. Eine Berücksichtigung der für p eingesetzten Punktschätzung führt zum genaueren Wilson-Intervall, siehe z. B. Dalitz (2017):

$$\left[\frac{\hat{p} + \frac{x^2}{2n} - x \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{x^2}{4n^2}}}{1 + \frac{x^2}{n}}, \frac{\hat{p} + \frac{x^2}{2n} + x \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{x^2}{4n^2}}}{1 + \frac{x^2}{n}} \right].$$

Beispiel 7.62

Um die Wahrscheinlichkeit zu schätzen, mit der ein Zug Verspätung hat, werden innerhalb einer Stunde alle verspäteten (100) und alle pünktlichen Züge (1 900) gezählt. Damit ist die relative Häufigkeit $\hat{p} = \frac{100}{2\,000} = 0,05$ eine Schätzung der Wahrscheinlichkeit. Wir suchen ein Intervall, das mit 99 %-iger Sicherheit ($\alpha = 0,01$) die tatsächliche Wahrscheinlichkeit p überdeckt. Aus $\Phi(x) = 1 - 0,005 = 0,995$ erhalten wir $x = 2,5758$ und das Wald-Konfidenzintervall

$$\left[0,05 - 2,5758 \cdot \sqrt{\frac{0,05 \cdot 0,95}{2\,000}}, 0,05 + 2,5758 \cdot \sqrt{\frac{0,05 \cdot 0,95}{2\,000}} \right] \approx [0,037; 0,062].$$

Völlig analog erhält man ein Konfidenzintervall für das Abschneiden einer Partei bei einer Wahl (siehe Beispiel 7.59). ■

Beispiel 7.63 (Bestimmung der Stichprobengröße)

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. ■

7.3.3.3 Konfidenzintervall für den Erwartungswert bei unbekannter Varianz

Bei der Ermittlung des Konfidenzintervalls haben wir in Abschnitt 7.3.3.1 die bekannte Varianz σ^2 der Zufallsvariable verwendet. Realistischer ist jedoch, dass weder Erwartungswert noch Varianz bekannt sind und man die Varianz mit dem bereits bekannten Schätzer

$$\hat{\sigma}^2 := \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{k=1}^n X_k \right)^2$$

ermitteln muss. Sind die identisch verteilten, unabhängigen Zufallsvariablen X_i schon normalverteilt, so kann man analog zur Situation eines bekannten σ^2 vorgehen, muss aber statt der Normalverteilung die **Student'sche t -Verteilung** nutzen:

Zu $1 - \alpha$ bestimme einen Wert x mit $F_{n-1}(x) = 1 - \frac{\alpha}{2}$, wobei F_{n-1} die Verteilungsfunktion der t -Verteilung mit $n - 1$ Freiheitsgraden ist. Diese ähnelt der Standardnormalverteilung, die zugehörige Dichte ist insbesondere eine gerade Funktion, also achsensymmetrisch mit $F_n(0) = \frac{1}{2}$. Der Wert x kann auch ohne Kenntnisse dieser Funktion nachgeschlagen oder mit einem Mathematikprogramm berechnet werden, siehe Tabelle 7.2. Mit der Zahl x ist das gesuchte Konfidenzintervall

$$\left[Z_n - x \cdot \frac{\hat{\sigma}}{\sqrt{n}}, Z_n + x \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right].$$

Beispiel 7.64

Gegeben seien $n = 5$ Messdaten für den maximalen Durchmesser d von Kartoffeln in Zentimeter (als Werte von 5 unabhängigen, identisch normalverteilten Zufallsvariablen):

$$10; \quad 10,12; \quad 8,7; \quad 9,9; \quad 9,5.$$

Tatsächlich ist Gauß durch einen Größenvergleich von Kartoffeln auf die Standardnormalverteilung gestoßen. Wir berechnen das Konfidenzintervall für den Erwartungswert des Durchmessers zum Niveau $1 - \alpha = 0,99$. Es ist $Z_5 = \bar{d} = 9,644$ und

$$\hat{\sigma} = s = \sqrt{\frac{1}{4}((10 - 9,644)^2 + \dots + (9,5 - 9,644)^2)} \approx 0,57678.$$

Aus $F_4(x) = 1 - \frac{\alpha}{2} = 0,995$ kann man mit Tabelle 7.2 den Wert $x \approx 4,604$ bestimmen und erhält weiter $x \cdot \frac{\hat{\sigma}}{\sqrt{n}} \approx 4,604 \cdot \frac{0,57678}{\sqrt{5}} \approx 1,188$. Das Konfidenzintervall lautet

$$[9,644 - 1,188, 9,644 + 1,188] \approx [8,456, 10,832]. \quad \blacksquare$$

Für große Werte von n kann die t -Verteilung F_n näherungsweise durch die handlichere Standardnormalverteilung Φ ersetzt werden. Groß bedeutet in der Praxis etwa $n \geq 30$. Dann sind auch die arithmetischen Mittel Z_n selbst in der Situation annähernd normalverteilt, in der die Verteilung der X_i nicht bekannt oder keine Normalverteilung ist. Bei genügend großem Stichprobenumfang kann man wie in Kapitel 7.3.3.1 rechnen, wenn man das dort bekannte σ^2 durch $\hat{\sigma}^2$ ersetzt.

Tab. 7.2: Quantile der Verteilungsfunktionen $F_n(x)$ der t -Verteilung zu $n = 4, 10, 20$ und 30 Freiheitsgraden

$F_4(x)$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	2,1318	2,3329	2,6008	2,9985	3,7469	4,6041	7,1732	8,6103	13,0337
$F_{10}(x)$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	1,8125	1,9481	2,1202	2,3593	2,7638	3,1693	4,1437	4,5869	5,6938
$F_{20}(x)$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	1,7247	1,8443	1,9937	2,1967	2,5280	2,8453	3,5518	3,8495	4,5385
$F_{30}(x)$	0,95	0,96	0,97	0,98	0,99	0,995	0,999	0,9995	0,9999
x	1,6973	1,8120	1,9546	2,1470	2,4573	2,7500	3,3852	3,6460	4,2340

7.3.3.4 Statistische Prozesslenkung *

Die Statistische Prozesslenkung (Statistical Process Control, SPC) wurde bereits 1931 von Walter Shewhard konzipiert und hat sich zu einem der wichtigsten Werkzeuge der Qualitätssicherung entwickelt. SPC ermöglicht Aussagen über den Zustand eines Fertigungsprozesses auf Basis von Stichproben. Die Ergebnisse der Stichproben werden mit Regelkarten erfasst und bewertet, siehe z. B. Logothetis (1992), Theden und Colson (2002). Die Mathematik steckt im Aufbau der Regelkarte und muss von einem Arbeiter, der Kontrollen durchführt, nicht verstanden werden. Damit geben wir uns hier aber nicht zufrieden.

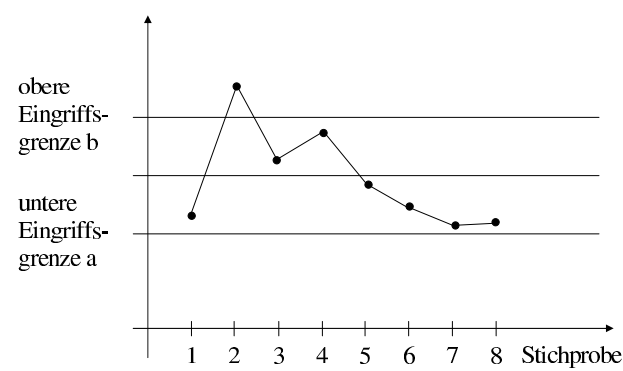


Abb. 7.16: Eine $\bar{x}$ -Regelkarte

Bei der Durchführung der SPC werden in regelmäßigen Abständen (z. B. stündlich) Zufallsstichproben aus dem Prozess gezogen. Oft betrachtet man $k = 20$ Zufallsstichproben. Jede Stichprobe i hat einen möglichst großen Umfang n , ausgedrückt durch die Zufallsvariablen $X_{i,1}, \dots, X_{i,n}$. Zu jeder Stichprobe i wird als Schätzer des Erwartungswerts

$$\bar{X}_i := \frac{1}{n} \sum_{l=1}^n X_{i,l}$$

verwendet. Die damit gewonnenen k Punktschätzungen werden dann in eine sogenannte $\bar{x}$ -Regelkarte eingetragen, siehe Abbildung 7.16 und für weitere Typen von Regelkarten z. B. Theden und Colsman (2002).

Auf der Basis der Werte aller Zufallsstichproben $X_{1,1}, \dots, X_{k,n}$ kann man zudem eine Intervallschätzung für den Erwartungswert (bei unbekannter Varianz) vornehmen, wobei ein Konfidenzintervall $[a, b]$ zum (hohen) Niveau $1 - \alpha$ bestimmt wird. Liegen nun einzelne Punktschätzungen des Erwartungswerts außerhalb dieses Konfidenzintervalls, dann ist das ein Anzeichen dafür, dass evtl. mit dem Prozess etwas nicht stimmt (z. B. durch Verstellung von Maschinen liegt keine identische Verteilung vor).

Wir sehen uns das Verfahren etwas genauer an: Für jede Stichprobe i sind die $X_{i,l}$, $1 \leq l \leq n$, stochastisch unabhängig und identisch verteilt. Bei genügend großem n sind damit die $\bar{X}_i$ nach dem Zentralen Grenzwertsatz annähernd $\mathcal{N}(\mu, \sigma^2/n)$ -verteilt. Ist der Prozess in Ordnung, so geht man davon aus, dass nicht nur die Zufallsvariablen einer Stichprobe, sondern die aller Stichproben stochastisch unabhängig und identisch verteilt sind. Als Punktschätzung für den Erwartungswert μ verwenden wir daher:

$$\hat{\mu} := \overline{\bar{X}} := \frac{1}{k} \sum_{i=1}^k \bar{X}_i = \frac{1}{kn} \sum_{i=1}^k \sum_{l=1}^n X_{i,l}.$$

Um einen Anhaltspunkt für die Varianz zu bekommen, wird die Spannweite R_i der i -ten Stichprobe berechnet:

$$R_i := \max \{X_{i,l} - X_{i,k} : l, k \in \{1, 2, \dots, n\}\}.$$

Damit approximieren wir die Varianz σ^2 bzw. die Standardabweichung σ durch

$$\hat{\sigma} := \frac{\bar{R}}{d_2(n)} \text{ mit } \bar{R} := \frac{1}{k} \sum_{i=1}^k R_i,$$

wobei man den Wert für d_2 der Tabelle 7.3 entnehmen kann. Die Zufallsvariablen $\bar{X}_i$

Tab. 7.3: Werte für die Statistische Prozesslenkung, siehe z. B. (Logothetis, 1992, S. 237)

n	2	3	4	5	6	7	8	9	10
$d_2(n)$	1,128	1,693	2,059	2,326	2,534	2,704	2,847	2,970	3,078

haben damit eine Verteilungsfunktion, die durch $\Phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}/\sqrt{n}}\right)$ approximiert wird.

Die Ergebnisse der Untersuchung werden in Regelkarten eingetragen. Bei der $\bar{X}$ -Karte werden die Schranken (Quantile) $\hat{\mu} - 3\frac{\hat{\sigma}}{\sqrt{n}}$ und $\hat{\mu} + 3\frac{\hat{\sigma}}{\sqrt{n}}$ sowie die einzelnen $\bar{X}$ -Werte eingezeichnet. Verfeinerungen des Ansatzes benutzen z. B. weitere Schranken bei $\mu - 2\frac{\hat{\sigma}}{\sqrt{n}}$ und $\mu + 2\frac{\hat{\sigma}}{\sqrt{n}}$ und teilen die Skala damit in grüne, gelbe und rote Zonen ein.

Die Wahrscheinlichkeit α , dass bei dieser Normalverteilung ein Wert $\bar{X}_i$ außerhalb des Intervalls $[\hat{\mu} - 3\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + 3\frac{\hat{\sigma}}{\sqrt{n}}]$ und damit in der roten Zone liegt, ist ungefähr

$$\begin{aligned}\alpha_1 &= 1 - \left[\Phi\left(\frac{\hat{\mu} + 3\frac{\hat{\sigma}}{\sqrt{n}} - \hat{\mu}}{\hat{\sigma}/\sqrt{n}}\right) - \Phi\left(\frac{\hat{\mu} - 3\frac{\hat{\sigma}}{\sqrt{n}} - \hat{\mu}}{\hat{\sigma}/\sqrt{n}}\right) \right] \\ &= 1 - [\Phi(3) - \Phi(-3)] = 1 - 2\left[\Phi(3) - \frac{1}{2}\right] = 2 - 2\Phi(3) \approx 2 - 2 \cdot 0,9986 = 0,0028,\end{aligned}$$

also gleich 0,28 %. Das ist so gering und damit unwahrscheinlich, dass die Ursache dieses dennoch eingetretenen $\bar{X}$ untersucht werden sollte. Hier kann sich etwas verstellt haben, so dass die Zufallsvariablen unterschiedlicher Stichproben doch nicht identisch verteilt sind. In einer Regelkarte wird der Bereich außerhalb dieses Intervalls rot markiert.

Liegt dagegen ein Wert zwar noch innerhalb dieses Intervalls aber in Randnähe, so kann eine Warnung sinnvoll sein. Die Wahrscheinlichkeit für einen Wert außerhalb von $[\hat{\mu} - 2\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + 2\frac{\hat{\sigma}}{\sqrt{n}}]$ ist ungefähr $\alpha_2 = 2 - 2\Phi(2) \approx 2 - 2 \cdot 0,9772 = 0,0456$. Die beiden Streifen $[\hat{\mu} - 3\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} - 2\frac{\hat{\sigma}}{\sqrt{n}}[$ und $] \hat{\mu} + 2\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + 3\frac{\hat{\sigma}}{\sqrt{n}}]$ sind der gelbe Bereich der Regelkarte.

Beispiel 7.65

Wir wenden die Statistische Prozesslenkung im Form einer $\bar{X}$ -Regelkarte für die sieben Stichproben aus Tabelle 7.4 an und erhalten den geschätzten Erwartungswert $\hat{\mu} =$

Tab. 7.4: Stichprobendaten für eine Regelkarte

Stichprobe	Mittelwert	Spannweite	Stichprobe	Mittelwert	Spannweite
(10, 9, 8, 11, 12)	10	4	(10, 10, 10, 10, 10)	10	0
(1, 17, 10, 11, 11)	10	16	(9, 9, 9, 9, 9)	9	0
(1, 2, 3, 4, 5)	3	4	(11, 11, 11, 11, 11)	11	0
(15, 16, 17, 18, 19)	17	4			

$\frac{1}{7}(10+10+3+17+10+9+11) = 10$ und mit $\bar{R} = 4$ die geschätzte Standardabweichung $\hat{\sigma}/\sqrt{n} = 4/(\sqrt{5} \cdot 2,326) = 0,77$. Damit sind alle Mittelwerte außerhalb von $[10 - 2 \cdot 0,77, 10 + 2 \cdot 0,77] = [8,46, 11,54]$ kritisch (im gelben oder roten Bereich). Dies sind 3 und 17. ■

7.3.4 Hypothesentests

Häufig möchte man nicht nur Parameter schätzen, sondern möchte mittels eines Tests auf Basis einer Stichprobe wissen, ob man eine Aussage (**Nullhypothese**) H_0 akzeptieren oder zu Gunsten einer Alternativhypothese H_1 „ H_0 gilt nicht“ ablehnen

sollte. Dazu sind die Hypothese sowie die Testvorschrift vor Durchführung der Stichprobe festzulegen (damit man sie nicht nachträglich der Beobachtung anpassen kann). Diese Regel wird häufig verletzt, um z. B. bei einer Studie irgendein Ergebnis in die Stichprobe hereinzuinterpretieren.

7.3.4.1 Ablehnung einer Hypothese mittels Konfidenzintervall

Als Beispiel betrachten wir die Nullhypothese, dass ein Parameter θ , der sich über eine Rechenvorschrift aus dem zugrunde liegenden Wahrscheinlichkeitsraum ergibt, gleich einem Wert θ_0 sei. Hier ist θ wie bei der Betrachtung von Schätzern eine Kennzahl des Wahrscheinlichkeitsraums, die unabhängig von den Ausgängen einzelner Experimente ist und die man in der Regel nicht kennt. Man kann also nur vermuten, dass dieser Parameter den Wert θ_0 hat. Diese Nullhypothese lautet:

$$H_0 : \quad \theta = \theta_0.$$

Wir kennen zwar θ nicht, können aber die Kennzahl wie zuvor über eine Stichprobe schätzen. Dazu berechnen wir über einen Schätzer $\hat{\theta}$ ein Konfidenzintervall zum (hohen) Niveau $1 - \alpha$ (häufig: $\alpha = 0,05$). Der wahre Parameter θ liegt mit der Wahrscheinlichkeit $1 - \alpha$ in diesem statistisch ermittelten Intervall. Die Wahrscheinlichkeit, dass θ_0 nicht im Konfidenzintervall liegt, aber die Nullhypothese $\theta = \theta_0$ gilt, ist damit nicht mehr als α (z. B. 0,05). Damit ist die Nullhypothese abzulehnen, wenn sich θ_0 nicht im Konfidenzintervall befindet. Man sagt, die Abweichung von der Nullhypothese ist **signifikant**, und spricht auch von einem **Signifikanztest**. Falls aber θ_0 im Konfidenzintervall liegt, hat man nicht die Hypothese verifiziert – sie kann immer noch mit großer Wahrscheinlichkeit falsch sein. Wir können sie lediglich auf Basis der Stichprobe nicht ablehnen.

Man hat es hier mit zwei möglichen Fehlertypen zu tun:

- Eine korrekte Nullhypothese abzulehnen, nennt man **Fehler erster Art**. Diesen Fehler macht man, wenn trotz $\theta = \theta_0$ der Wert θ_0 außerhalb des berechneten Konfidenzintervalls liegt. Die Wahrscheinlichkeit, dass das passiert, ist α . Im Beispiel 7.32 auf Seite 1014 liefert ein Test zur Hypothese „Der Mensch ist gesund.“ einen großen Fehler erster Art von $\alpha = 0,2$. α heißt das **Signifikanzniveau** des Tests.
- Eine falsche Nullhypothese nicht abzulehnen, nennt man **Fehler zweiter Art**. Im Beispiel 7.32 ist dieser Fehler 0,01.

Häufig besteht bei Tests ein Trade-Off zwischen diesen beiden Fehlern, d. h., dass man nicht beide Fehler gleichzeitig klein halten kann.

7.3.4.2 Vergleich zweier geschätzter Wahrscheinlichkeiten

In Abschnitt 7.3.3.2 haben wir das Konfidenzintervall zum Niveau $1 - \alpha$ für eine Wahrscheinlichkeit p_1 eines Ereignisses E_1 kennengelernt, die mit einem Schätzer $\hat{p}_1$ ermittelt wurde. Dabei zählt $\hat{p}_1$, wie oft das Ereignis E_1 beim n_1 -maligen Ziehen mit Zurücklegen eintritt und dividiert diese Summe durch n_1 . Die Verteilung von $\hat{p}_1$ lässt sich bei genügend großem n_1 durch $\mathcal{N}(p_1; p_1(1 - p_1)/n_1)$ approximieren. Ist $\hat{p}_2$ eine zu $\hat{p}_1$ stochastisch unabhängige Zufallsvariable, die analog die Wahrscheinlichkeit eines Ereignisses E_2 auf Basis von n_2 Werten schätzt, so interessiert uns, ob wir die Nullhypothese $N_0: p_1 = p_2$ anhand der Ergebnisse der Schätzer ablehnen müssen. Wir fragen uns also, ob der durch den Wahrscheinlichkeitsraum vorgegebene Parameter $p_1 - p_2$ mit dem Wert 0 übereinstimmt: $N_0: p_1 - p_2 = 0$. Dazu ermitteln wir zu $\hat{p}_1 - \hat{p}_2$ ein Konfidenzintervall zum hohen Niveau $1 - \alpha$ für $p_1 - p_2$. Liegt der Wert 0 nicht im Intervall, dann ist die Wahrscheinlichkeit von $p_1 - p_2 = 0$, also $p_1 = p_2$, höchstens α . Bei kleinem α ist damit die Nullhypothese abzulehnen.

Wir benötigen zur Berechnung des Konfidenzintervalls die Verteilung von $\hat{p}_1 - \hat{p}_2$. Dabei hilft der Additionssatz der Normalverteilung (Sachs, 2003, S. 126), auf den wir hier nicht näher eingehen können, da wir nur diskrete Zufallsvariablen betrachten. Für unsere Zwecke reicht: Sind X_1 bzw. X_2 stochastisch unabhängige Zufallsvariablen, deren Verteilung durch $\mathcal{N}(\mu_1, \sigma_1^2)$ bzw. $\mathcal{N}(\mu_2, \sigma_2^2)$ approximiert werden kann, dann kann die Verteilung von $X_1 + X_2$ durch $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ approximiert werden.

Damit ist $\hat{p}_1 - \hat{p}_2$ annähernd $\mathcal{N}(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})$ verteilt. (Beachte: $-\hat{p}_2$ ist annähernd $\mathcal{N}(-p_2, \frac{p_2(1-p_2)}{n_2})$ -verteilt.)

Ein $1 - \alpha$ -Konfidenzintervall ergibt sich für die Zahl $x > 0$ mit $\Phi(x) = 1 - \frac{\alpha}{2}$ zu

$$\left[\hat{p}_1 - \hat{p}_2 - x \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + x \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right].$$

Hier kann man wieder p_1 durch $\hat{p}_1$ und p_2 durch $\hat{p}_2$ ersetzen ohne einen zu großen Fehler zu machen.

Jetzt können wir am Ende des Kapitels die Fragestellung der Einleitung beantworten (siehe Seite 970):

Beispiel 7.66

In einem Softwareprojekt wird die Anzahl der nach einer gewissen Zeit gefundenen Fehler pro Klasse als Qualitätsmaß eingesetzt. Zur Klasse A mit 1 000 Programmzeilen wurden 10 Fehler gefunden, zur Klasse B mit 2 000 Programmzeilen wurden 18 Fehler gefunden. Damit haben wir Schätzwerte für die Fehlerwahrscheinlichkeiten (Wahrscheinlichkeit eines Fehlers pro Programmzeile) p_1 und p_2 : $\hat{p}_1(\omega) = \frac{10}{1\,000} = 0,01$

und $\hat{p}_2(\omega) = \frac{18}{2\,000} = 0,009$. Zu $\alpha = 0,05$ ist $x = 1,96$, und damit erhalten wir das (gerundete) Konfidenzintervall

$$\left[0,01 - 0,009 - 1,96 \sqrt{\frac{0,01(1-0,01)}{1\,000} + \frac{0,009(1-0,009)}{2\,000}}, 0,001 + 1,96 \cdot 0,0038 \right],$$

also $[-0,006448, 0,008448]$. Da 0 in diesem Intervall enthalten ist, kann die Nullhypothese einer gleichen Wahrscheinlichkeit zum Signifikanzniveau $\alpha = 0,05$ nicht abgelehnt werden. Wir können nicht schließen, dass das Team, das Klasse A erstellt hat, schlechter arbeitet als das Team von Klasse B. ■

7.3.4.3 Test auf eine Wahrscheinlichkeitsverteilung

Neben dem Vergleich zweier Wahrscheinlichkeiten kann man auch eine komplette vorliegende (aber unbekannte) Wahrscheinlichkeitsverteilung gegen eine angenommene testen. Die Nullhypothese beim sogenannten **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 7.17).

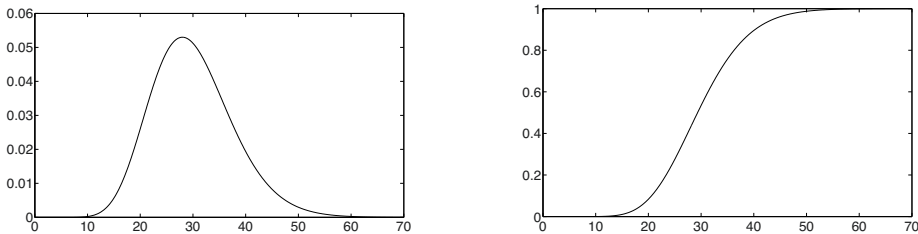


Abb. 7.17: Dichtefunktion und Verteilungsfunktion der χ^2 -Verteilung zum Freiheitsgrad 30

Sei $x_{1-\alpha}$ die Stelle, an der die χ^2 -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. Genauso 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 $X_1, \dots, X_n$ definiert ist. Für

jeden Wert $n \in \mathbb{N}$ erhält man eine andere Verteilung. Die Werte der Verteilungen findet man in Tabellenwerken (siehe Tabelle 7.5).

Tab. 7.5: 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 $X_1(\omega), \dots, X_n(\omega)$ beobachtet, und es wird gezählt, wie oft jeder der m Zufallswerte angenommen wird. Diese Anzahlen seien $n_1, \dots, n_m$. 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 zusammen. 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 7.67

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

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

Dieser Wert ist (erheblich) größer als der Wert 11,0705 des 0,95-Quantils (siehe Tabelle 7.5 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. ■

7.4 Aufgaben

Aufgabe 7.1 (Paradoxon von Simpson)

An einer Hochschule bewirbt sich für die Fächer Soziologie und Elektrotechnik die folgende Anzahl von Männern und Frauen, vgl. Krämer (1992):

	Soziologie	Elektrotechnik	Σ
Männer	320	180	500
Frauen	480	20	500
	800	200	1 000

Die Annahmequoten seien wie folgt:

	Soziologie	Elektrotechnik	
Männer	10 %	40 %	$(0,1 \cdot 320 + 0,4 \cdot 180)/500 = 104/500 > 0,2$
Frauen	12,5 %	50 %	$(0,125 \cdot 480 + 0,5 \cdot 20)/500 = 0,14$

Es erhalten also nur 14 % der Bewerberinnen im Gegensatz zu mehr als 20 % bei den männlichen Bewerbern einen Studienplatz. Kann man daraus schließen, dass Bewerberinnen diskriminiert werden?

Aufgabe 7.2

Für ein Merkmal X werden bei einer Umfrage die Ausprägungen $x_1 = 1$, $x_2 = 2$, $x_3 = 2$, $x_4 = 5$ und $x_5 = 10$ erfasst. Geben Sie die

- a)** Verteilungsfunktion an, und berechnen Sie **b)** das arithmetische Mittel,
c) den Median, **d)** das 0,25- und das 0,75-Quantil und **e)** die Varianz.
f) Zeichnen Sie einen Box-Plot.

Aufgabe 7.3

In einem Prozess werden 2 500 Aktionen durchgeführt. Hierbei verlaufen 500 Aktionen fehlerhaft. Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Aktion zu einem Fehler führt, wenn die Fehlerwahrscheinlichkeit für jede Aktion gleich groß ist? Wie groß ist die Wahrscheinlichkeit, dass eine Aktion fehlerfrei durchgeführt wird?

Aufgabe 7.4

Wie groß ist die Wahrscheinlichkeit, mit drei Würfeln mit einem Wurf dreimal die Augenzahl 6 zu würfeln?

Aufgabe 7.5 (Das klassische Kartenspiel-Modell)

*Ein Skatblatt besteht aus 32 Karten in den vier Farben (absteigende Wertigkeit) Kreuz, Pik, Herz und Karo. Zu jeder Farbe gibt es die Karten (absteigende Reihenfolge) Ass, König, Dame, Bube, 10, 9, 8, 7. Zusammen ergibt sich damit die Kartenreihenfolge Kreuz-Ass, Pik-Ass, Herz-Ass, Karo-Ass, Kreuz-König,... Aus allen Karten wird zufällig eine Karte gezogen, **zurückgelegt** und dann eine weitere gezogen. Jedes mögliche Kartenpaar wird mit gleicher Wahrscheinlichkeit gezogen. Geben Sie einen geeigneten Wahrscheinlichkeitsraum (Ω, P) zur Beschreibung des Experiments an, und beschreiben Sie die drei folgenden Ereignisse:*

- a)** Die erste gezogene Karte ist Kreuz-Ass und die zweite ist Karo-7.
b) Die erste Karte hat eine kleinere Wertigkeit als die zweite.
c) Beide Karten haben die Farbe Herz.

Bestimmen Sie die Wahrscheinlichkeit dieser Ereignisse.

Aufgabe 7.6

Zehn Personen einer Reisegruppe geben in einem Hotel je einen Rucksack und einen Koffer am Empfang ab, da die Zimmer noch nicht fertig sind. Später erhalten sie zufällig je einen Rucksack und zufällig einen Koffer zurück. Wie groß ist die Wahrscheinlichkeit, dass alle Reisenden sowohl den richtigen Rucksack als auch den richtigen Koffer erhalten?

Aufgabe 7.7

- a) Wie viele Wörter kann man aus den Buchstaben des Worts „MATHEMATIK“ unter Verwendung aller Buchstaben bilden?
- b) Aus einer Gruppe von 30 Personen soll ein siebenköpfiger Ausschuss gebildet werden. Wie viele Möglichkeiten gibt es? (Die Reihenfolge der Personen im Ausschuss spielt keine Rolle.)

Aufgabe 7.8

Wie viele unterschiedliche Farbmuster können entstehen, wenn man 4 grüne, 3 blaue und zwei rote Steine nebeneinander legt?

Aufgabe 7.9

Es werden n gleiche Kugeln auf k Urnen verteilt, so dass keine Urne leer bleibt. Wir notieren die Anzahl der Kugeln pro Urne. Wie viele verschiedene Möglichkeiten gibt es hier?

Hinweis: Wir legen die Kugeln nebeneinander und betrachten die Zwischenräume. Wie viele Möglichkeiten gibt es, $k - 1$ Trenner in verschiedenen Zwischenräumen zu positionieren?

Aufgabe 7.10

Zur Markierung von Werkstücken mit Farbstrichen stehen n Farben zur Verfügung. Bei einer Markierung wird die Reihenfolge der Striche nicht ausgewertet (z. B. weil das Werkstück gedreht sein könnte). Wie viele Markierungen sind möglich, wenn ein Werkstück

- a) zwei verschiedenfarbige Striche und
- b) drei Striche, die untereinander auch gleichfarbig sein können,

erhält? Wie viele Farben braucht man im Falle von 20 Werkstücken bei a) und b) mindestens?

Aufgabe 7.11

Zum Bau einer Mauer stehen 5 grüne, 3 blaue und 2 rote Steine einer vollen Breite zur Verfügung. Außerdem gibt es noch 4 Steine der halben Breite, davon sind 2 blau und je einer rot und einer grün.

Die Mauer soll vier große Steine breit und drei Steine hoch sein. Dabei sollen sich vertikale Fugen nicht über zwei Reihen erstrecken. Es ergibt sich somit ein Muster wie in der Abbildung 7.18.

- a) Wie viele verschiedene Muster können gemauert werden? Eine Vereinfachung des Werts bis zu einer Dezimalzahl ist nicht erforderlich.
- b) Wie wahrscheinlich ist es, dass der erste Stein unten links blau ist?

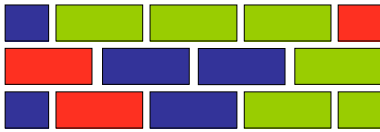


Abb. 7.18: Eine Mauer mit drei Reihen und einer Breite von vier großen Steinen

Aufgabe 7.12

In einer Fabrik sollen n Aufträge produziert werden. Jeder Auftrag kann auf einer von m Produktionsstraßen vollständig erstellt werden (Einstufenfertigung). Gleichzeitig kann auf einer Produktionsstraße immer nur ein Auftrag bearbeitet werden, die Produktion läuft also pro Straße sequentiell. Unterschiedliche Straßen produzieren parallel. Begründen Sie, dass es

$$n! \cdot \binom{n+m-1}{m-1}$$

verschiedene Zuordnungen der Aufträge zu Produktionsstraßen gibt.

Aufgabe 7.13

In einer Urne befinden sich 50 Kugeln. Davon sind 20 grün, 25 blau und 5 rot. Es werden drei Kugeln der Reihe nach mit Zurücklegen gezogen.

- Wie groß ist die Wahrscheinlichkeit, dass die drei Kugeln in der Reihenfolge grün, blau, rot gezogen wurden?
- Wie groß ist die Wahrscheinlichkeit, dass die drei Kugeln eine unterschiedliche Farbe haben?

Aufgabe 7.14

Ein viermotoriges Flugzeug stürzt ab, wenn mehr als zwei Triebwerke ausfallen. Wir nehmen an, dass der Ausfall von Triebwerken stochastisch unabhängig ist. Die Wahrscheinlichkeit, dass ein Triebwerk bei einem Flug ausfällt, sei 10 % ($= 0,1$). Wie groß ist dann die Wahrscheinlichkeit, dass das Flugzeug abstürzt?

Aufgabe 7.15

Ein Auto wird als Modell „Harlekin“ mit Karosserieteilen in unterschiedlichen Farben angeboten. Es stehen die Farben Blau, Grün, Gelb und Rot zur Verfügung. Die Türen (vier Stück), das Heck, die Motorhaube und die beiden vorderen Kotflügel erhalten je eine andere Farbe als die benachbarten Teile. Außerdem unterscheidet sich die Farbe des Dachs von den Türen. Benachbart sind Motorhaube und Kotflügel, Kotflügel und vordere Türen, vordere Türen und hintere Türen, hintere Türen und Heck. Die Farben auf der rechten und linken Wagenseite sind gleich.

- Wie viele verschiedene Fahrzeuglackierungen gibt es?
- Alle Lackierungen verkaufen sich gleich gut. Wie wahrscheinlich ist es, dass ein zufällig gesichtetes Fahrzeug ein blaues Dach und ein rotes Heck hat?

- c) Sind die Ereignisse „Auto hat blaues Dach“ und „Auto hat rotes Heck“ stochastisch unabhängig?

Aufgabe 7.16

Die Gesamtproduktion eines Werkes verteilt sich wie folgt auf drei Maschinen:

Maschine A: 25 %, Maschine B: 35 %, Maschine C: 40 %. Jede Maschine produziert einen gewissen Ausschuss. Der Anteil fehlerhafter Teile beträgt bei Maschine A: 5 %, Maschine B: 4 % und Maschine C: 2 %. Ein zufällig ausgewähltes Teil ist fehlerhaft. Mit welcher Wahrscheinlichkeit wurde es auf Maschine A produziert?

Aufgabe 7.17

Eine Kreisringscheibe wird in sechs gleich große Stücke unterteilt. Die Vorderseiten werden eingefärbt. Drei Stücke werden rot, zwei blau und eins grün. Die Stücke werden mit der farbigen Seite nach oben auf einen Tisch gelegt.

- a) Auf wie viele verschiedene Arten kann der Ring durch Verschieben der Teile zusammengesetzt werden? Zwei Ringe heißen verschieden, wenn ihre Farbmuster durch Drehen nicht in Einklang gebracht werden können.
- b) Wie wahrscheinlich ist es, dass bei einer rein zufälligen Anordnung der Farben zwei rote Stücke gegenüberliegen, wie wahrscheinlich ist es, dass zwei blaue Stücke gegenüberliegen, und wie wahrscheinlich ist es, dass beide Fälle gemeinsam eintreten?

Aufgabe 7.18

Ein Zug kommt nachmittags mit der Wahrscheinlichkeit 0,5 verspätet an, und sein Zugführer trinkt mit Wahrscheinlichkeit 0,7 mittags ein Bier. Wenn er ein Bier getrunken hat, hat der Zug nachmittags sogar mit Wahrscheinlichkeit 0,8 Verspätung.

- a) Mit welcher Wahrscheinlichkeit ist der Zug nachmittags verspätet, wenn der Zugführer kein Bier getrunken hat?
- b) Warum ist das Ergebnis aus Teil a) sinnlos?

Aufgabe 7.19

Gegeben sind zwei Urnen. Die erste enthalte zwei schwarze und drei weiße (ansonsten gleiche) Kugeln, die zweite eine schwarze und eine weiße. Zunächst werden zufällig zwei Kugeln aus der ersten Urne gezogen und in die zweite gelegt. Es handelt sich also um ein Ziehen ohne Zurücklegen. Dann werden aus der zweiten Urne zwei Kugeln zufällig ebenfalls ohne Zurücklegen gezogen.

- a) Wie wahrscheinlich ist es, dass aus der ersten Urne eine schwarze und eine weiße Kugel gezogen wurde?
- b) Wie wahrscheinlich ist es, dass auch aus der zweiten Urne eine schwarze und eine weiße Kugel gezogen wurde?

Aufgabe 7.20

Sei X eine diskrete Zufallsvariable, die die Werte 1, 2, 3, 11 mit der Wahrscheinlichkeit $P(X = 1) = \frac{1}{4}$, $P(X = 2) = \frac{1}{2}$, $P(X = 3) = \frac{1}{8}$ und $P(X = 11) = \frac{1}{8}$ annimmt. Berechnen Sie den Erwartungswert und die Varianz von X .

Aufgabe 7.21

Eine Urne enthalte n von 1 bis n (n gerade) durchnummerierte Kugeln. Es werden k Kugeln mit Zurücklegen sukzessive und unabhängig voneinander gezogen.

- a) Wie groß ist die Wahrscheinlichkeit, dass alle k gezogenen Kugeln unterschiedliche Nummern tragen?
- b) Wie groß ist die Wahrscheinlichkeit, dass von den k Kugeln genau r gerade sind?

Aufgabe 7.22

In einem Reservat befindet sich eine unbekannte Anzahl N Tiere. Um eine Schätzung für die Anzahl zu gelangen, geht man so vor: Es werden M Tiere gefangen und markiert. Diese markierten („defekten“) Tiere werden wieder ausgesetzt. Dann wird eine Zeit gewartet, so dass sich die Tiere gut durchmischen. Danach werden n Tiere gefangen. Darunter befinden sich m markierte. Ein vernünftiger Schätzwert ist dann die Zahl $N = N_0$, für die die Wahrscheinlichkeit, genau m „defekte“ Tiere beim Ziehen von n Tieren ohne Zurücklegen zu erhalten, maximal ist. Bestimmen Sie den Schätzwert. Beachten Sie, dass N eine natürliche Zahl ist.

Aufgabe 7.23

Auf einem geeigneten Wahrscheinlichkeitsraum sei X eine Zufallsvariable mit $X(\Omega) = [0, 1]$ und $P^X([0, x]) = x$ für $0 \leq x \leq 1$.

- a) Berechnen Sie $P^X(\{x\})$ für $x \in [0, 1]$.
- b) Versuchen Sie, den Erwartungswert von X zu berechnen. Warum klappt das nicht?

Aufgabe 7.24 (Nadelexperiment von Buffon)

Der französische Adelige Georges-Louis Leclerc de Buffon hat 1727 ein einfaches Experiment zur Bestimmung der Zahl π erfunden: Er hat eine Nadel der Länge l auf ein liniertes Blatt fallen lassen, wobei der Abstand zweier Linien $a > l$ war. Bestimmen Sie die Wahrscheinlichkeit, dass die Nadel eine Linie trifft. Wie kann man damit π bestimmen? Als Modell für dieses Experiment eignet sich die Menge

$$\Omega = \left\{ (x, \varphi) : 0 \leq x < a, -\frac{\pi}{2} \leq \varphi < \frac{\pi}{2} \right\},$$

wobei ein Elementarereignis (x, φ) bedeutet, dass ein linker Endpunkt der Nadel die Entfernung x von einer unteren Linie hat und dass die Nadel bezogen auf die Linien um φ gedreht ist.

Aufgabe 7.25

Mit einer Maschine werden Schrauben hergestellt. Dabei gibt es gewisse Fertigungstoleranzen. Die Zufallsvariable X misst die Länge der Schrauben und ist annähernd $\mathcal{N}(\mu, \sigma^2)$ -verteilt. Dabei ist der Erwartungswert μ die angestrebte Länge der Schrauben. Die Varianz σ^2 kann durch Justierung beeinflusst werden. Als fehlerhaft gelten alle Schrauben mit Länge $l < \mu - \tau$ oder $l > \mu + \tau$. Dabei ist $\tau > 0$ die maximale Toleranz. Wie muss σ^2 durch Justierung der Maschine eingestellt werden, damit 99,9999998 % aller Schrauben fehlerfrei sind?

Hinweis: Aufgrund der Normalverteilung ist $P^X(\mu - 6\sigma, \mu + 6\sigma) \approx 0,999999998$.

Aufgabe 7.26

Ein Zufallsgenerator produziert eine Folge von Zufallswerten, die einer (gemeinsamen) Verteilung mit Erwartungswert 0 und Varianz 1 entsprechen (Rauschsignal). Summiert man n so generierte Folgenglieder auf, beobachtet man, dass bei größeren Werten von n die Summe mit über 99 % Wahrscheinlichkeit zwischen $-3\sqrt{n}$ und $3\sqrt{n}$ liegt. Begründen Sie diese Beobachtung. **Hinweis:** $\Phi(3) \approx 0,9987$ (vgl. Tabelle 7.1).

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)
- de Jong T. (2012) Analysis. Pearson, München. (Kompakte Darstellung mit übersichtlichen Beweisen, Umfang ähnlich dem Kapitel 2)
- 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)
- Dürschnabel 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)
- Goebbels St. und Rethmann J. (2014) Mathematik für Informatiker. Springer Vieweg, Heidelberg.
- Göllmann L., Hübl R., Pulham S., Ritter S., Schon H., Schüffler K., Voß U. und Vossen G. (2017) Mathematik für Ingenieure: Verstehen, Rechnen, Anwenden. Band 1 und 2. Springer Vieweg, Heidelberg.
- 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 Spektrum-Verlag, Mannheim.
- Meyberg K. und Vachenaue 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)
- Ritter S. und Voß U. (2015) Erfolgreich Starten ins Ingenieurstudium. Springer Vieweg, Heidelberg.

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

Weitere Referenzen

- Adams R. A. (1975) Sobolev Spaces. Academic Press, New York.
- Alt H. W. (1992) Lineare Funktionalanalysis. Springer, Berlin.
- Arens T. et al. (2012) Mathematik. Spektrum, Heidelberg. (Umfassende, mathematisch genaue Darstellung der Mathematik, schöne Illustrationen)
- Bauer, H. (1990) Maß- und Integrationstheorie. W. de Gruyter, Berlin. (Standardwerk zur Integration mit Maßen wie das Lebesgue-Maß)
- 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. und Nessel R. J. (1971) Fourier Analysis and Approximation. Birkhäuser, Basel. (Klassische Speziallektüre zur Approximationstheorie und Fourieranalysis)
- Ciarlet P. G. (1990) Basic error estimates for elliptic problems. In: Ciarlet P. G., Lions, J.-L. (Hrsg.) Handbook of Numerical Analysis, II, North-Holland, Amsterdam, S. 5–196.
- Dalitz Ch. (2017) Konstruktionsmethoden für Konfidenzintervalle. Technischer Bericht 1-2017 des Fachbereichs Elektrotechnik und Informatik der Hochschule Niederrhein, Krefeld, <http://www.hs-niederrhein.de/fb03/>
- 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)
- Goebbels St., Pohle-Fröhlich R. und Rethmann J. (2016) Planarization of CityGML models using a linear program. In: Proceedings Operations Research (OR 2016 Hamburg), Springer, Berlin, S. 591–597. (Beispiel zur linearen Optimierung)
- Goebbels St. und Pohle-Fröhlich R. (2015) Context-Sensitive Filtering of Terrain Data based on Multi Scale Analysis. In: Proceedings International Conference on Computer Graphics Theory and Applications (GRAPP 2015), Berlin, S. 106–113.
- Goebbels, St. (2014) Mathematik der Z-Transformation. Technischer Bericht 2-2014 des Fachbereichs Elektrotechnik und Informatik der Hochschule Niederrhein, Krefeld, <http://www.hs-niederrhein.de/fb03/>
- 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. (2006) Funktionalanalysis. Teubner, Wiesbaden.
- Heuser H. (2004) Lehrbuch der Analysis Teil 2. Teubner, Wiesbaden. (Differenzial- und Integralrechnung mit mehreren Variablen, Fourier-Analyse, Geschichte der Analysis)
- Hochstättler W. (2010) Algorithmische Mathematik, Springer, Berlin. (Schöne Einführung in nicht-lineare und lineare Optimierung)

- Hohloch E. et al. (2009) Brücken zur Mathematik Band 1–7. Cornelsen, Berlin.
- Jung M. und Langer U. (2001) Methode der finiten Elemente für Ingenieure. Teubner, Stuttgart.
- 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)
- Meister A. (2011) Numerik linearer Gleichungssysteme. Vieweg+Teubner, Wiesbaden.
- Natanson I. P. (1955) Konstruktive Funktionentheorie. Akademie-Verlag, Berlin. (Klassische Speziallektüre zur Approximationstheorie und Fourieranalysis)
- Persson P.-O. und Strang G. (2004) A simple mesh generator in MATLAB. SIAM Review 46, S. 329–345.
- Schatz A. H. (1980) A weak discrete maximum principle and stability of finite element method in L_∞ on plane polygonal domains. Math. Comp. 34, S. 77–91.
- Schenk J. und Rigoll G. (2010) Mensch-Maschine-Kommunikation. Springer, Berlin.
- Schüffler K. (1991) Mathematik in der Wirtschaftswissenschaft. Hanser, München.
- Sokal, A. D. (2011) A really simple elementary proof of the uniform boundedness theorem. Am. Math. Mon. 118 (5), S. 450–452.
- 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)
- von Hippel, P. T. (2005) Mean, Median, and Skew: Correcting a Textbook Rule. Journal of Statistics Education 13, S. 965–971
- Zygmund A. (2002) Trigonometric Series. Cambridge University Press, Cambridge. (Der Klassiker zu Fourier-Reihen von 1935 in der dritten Auflage)

Index

A

A-posteriori-Abschätzung, 320, 567
A-priori-Abschätzung, 320, 567
Abbildung, 8
Abel'sche Gruppe, 46
abgeleitete Funktion, 294
abgeschlossen, 256, 670
abgeschlossene Menge, 277
abgeschlossenes Intervall, 51
Ableitung, 289, 294
absolut konvergent, 243
absolute Homogenität, 558
absoluter Fehler, 309
Abstand, 486
Abszisse, 77
Abtastfrequenz, 937, 942
Abtastsatz, 941
abzählbar, 39
Additionstheoreme, 119
Additivität, 483, 500
Additivität in beiden Argumenten, 485
Adjungierte, 173
adjungierte Matrix, 173
ähnlich, 520
affin-lineare Funktion, 78
Aktivierungsfunktion, 604
algebraische Gleichungen, 72
algebraischen Funktionen, 101
Aliasing, 934
Allmenge, 5
Allquantor, 21
alternierende harmonische Reihe, 240
alternierenden Folge, 213
Amplitude, 122
Amplitudengang, 896
Amplitudenspektrum, 822, 863
analytisch, 419
Analytische Geometrie, 453
Anfangsbedingung, 713, 751
Anfangspunkt, 674
Anfangswertproblem, 707, 713, 751
Anfangswertsatz, 892
Ansatz vom Typ der rechten Seite, 730
Ansatzfunktionen, 793
Anti-Kommutativgesetz, 444
Antisymmetrie, 27
Apfelmännchen, 215
Approximation, 52
Approximationsprozess, 582
Approximationstheorie, 952
äquivalente Normen, 559
Äquivalenz, 20
Arbeit, 435
Arbeitspunkt, 308

Area-Kosinushyperbolikus, 131
Area-Kotangenshyperbolikus, 131
Area-Sinushyperbolikus, 130
Area-Tangenshyperbolikus, 131
Argument, 8
arithmetische Folge, 213
arithmetisches Mittel, 69, 977
Arkus-Funktionen, 125
Arkuskosinus, 126
Arkussinus, 125
Arkustangens, 126
assoziativ, 429, 467, 468
Assoziativgesetze, 6, 14, 45, 429
Asymptote, 393
auf (Abbildung), 10
Aufpunkt, 460
aufspannen, 471
Aussage, 12
Aussageform, 18
aussagenlogische Formeln, 13
äußeres Lebesgue-Maß, 377
äußeres Produkt, 444
Austauschsatz, 479
Axiome, 23

B

Balkendiagramm, 973
Banach'scher Fixpunktsatz, 318, 567
Banach-Raum, 560
bandbegrenzt, 941
Bandmatrizen, 176
Bar Chart, 973
Barometrische Höhenformel, 107
Bartlett-Fenster, 952
Basis, 36, 478
Basislösung, 641
Basismatrix, 518
Basistransformation, 518
Basisvariablen, 641
Baum, 530
bedingte Wahrscheinlichkeit, 1009
Begleitmatrix, 772
benachbart, 530
Bernoulli'sche Ungleichung, 68
Bernoulli-Differenzialgleichung, 741
beschränkt, 796
beschränkt, 82, 214, 567
beschränkt nach oben, 47, 82, 213
beschränkt nach unten, 48, 82, 213
beschränkte Variation, 677
beschreibende Statistik, 969
Bessel'sche Ungleichung, 830
bestimmt divergent, 229, 239, 259, 263

Betrag, 65, 135, 427, 433, 486
 Betragsfunktion, 89
 Betragsgleichungen, 73
 Betragsungleichung, 75
 Beweis, 23
 Beweis durch Widerspruch, 23
 BIBO-stabil, 580
 bijektiv, 10
 Bild, 8, 508
 Bilinearform, 796
 bimodale Verteilung, 981
 Binet-Formel, 551
 Binomialkoeffizient, 33
 binomialverteilt, 1027
 Binomialverteilung, 1027
 Binomische Formel, 64
 Binomischer Satz, 64
 biquadratische Gleichung, 72
 Bisektionsverfahren, 279
 Blätter, 531
 Blindwiderstand, 146
 Bogenlänge, 676
 Box-Plot, 983
 Branch-and-Bound, 648
 Brücke, 35

C

Caesar-Code, 11
 Cantor-Diagonalisierung, 40
 Cardan'sche Formeln, 150
 Cauchy-Bedingung, 241
 Cauchy-Folge, 235, 560
 Cauchy-Hadamard-Formel, 413
 Cauchy-Integralformel, 687
 Cauchy-Integralsatz, 687
 Cauchy-Kriterium, 235, 241
 Cauchy-Produkt, 244
 Cauchy-Riemann-Differenzialgl., 609
 Cauchy-Schwarz'sche Ungleichung, 437, 487
 CG-Verfahren, 624
 Charakteristiken, 602
 charakteristische Funktion, 379
 charakteristische Gleichung, 540
 charakteristisches Polynom, 539
 Chi-Quadrat-Test, 1074
 Chi-Quadrat-Verteilung, 1074
 Cholesky-Zerlegung, 628
 Cramer'sche Regel, 197
 Currying, 591
 Cuts, 648

D

Dampfdruck, 110
 De Morgan'sche Regeln, 6, 15
 Dedekind'sche Schnitte, 49

Defekt, 510
 Definitionsbereich, 8
 Dekadischer Logarithmus, 105
 Delta-Distribution, 896
 Delta-Umgebung, 256, 593
 Determinante, 182
 diagonalisierbar, 548
 Diagonalmatrix, 155
 dicht, 57, 582
 Dichte, 1024
 Dichte der Standardnormalverteilung, 1025, 1047
 Differenz, 4, 429
 Differenzenquotient, 289
 Differenzenverfahren, 793
 Differenzial, 308
 Differenzialgleichung, 711
 Differenzialgleichungssystem, 749
 Differenzialquotient, 289
 differenzierbar, 289, 599
 Dimension, 481
 Dimensionssatz, 510, 511, 527
 direkter Beweis, 23
 Dirichlet-Funktion, 263
 Dirichlet-Kern, 838
 Dirichlet-Randbedingung, 697
 disjunkt, 2
 Disjunktion, 13
 disjunktive Normalform, 17
 diskrete Fourier-Transformation, 906, 911, 918
 diskrete Merkmale, 971
 diskretes Spektrum, 862
 Diskretisierung, 905
 Diskriminante, 55
 Distributionen, 899
 Distributivgesetz, 45, 444
 Distributivgesetze, 6, 14, 429, 437, 468
 divergent, 217, 239
 Divergenz, 678
 doppelt-logarithmischen Darstellung, 110
 Drehmoment, 443
 Drehmomentvektor, 443
 Dreiecksmatrix, 176
 Dreiecksungleichung, 68, 429, 487, 558
 Dreiecksungleichung nach unten, 68
 Dualer Logarithmus, 105
 dualer Simplex-Algorithmus, 640
 Dualraum, 570
 Dualsystem, 15
 Durchschnitt, 4

E

e-Funktion, 103
 Eigenfrequenz, 783
 Eigenraum, 540

Eigenvektor, 538
 Eigenwert, 538
 eineindeutig, 9
 Einerkomplement, 28
 einfach, 674
 einfach-logarithmische Darstellungen, 109
 einfache Funktion, 379
 Einheitsmatrix, 155
 Einheitsvektor, 428, 486
 Einheitswurzeln, 909
 einseitig differenzierbar, 293
 einseitiges Konfidenzintervall, 1065
 Einzelschritt-Verfahren, 180
 Element, 154
 elementare Funktion, 343
 elementarer Inhalt, 650
 Elementarereignis, 998
 Elementarereignisraum, 998
 Elemente, 2
 elementfremd, 2
 elliptisch, 791, 796
 elliptische Projektion, 797
 empirische Korrelationskoeffizient, 987
 empirische Kovarianz, 985
 empirische Regressionsgerade, 992
 empirische Standardabweichung, 984
 empirische Varianz, 983
 empirische Verteilungsfunktion, 976
 Endknoten, 530
 Endpunkt, 674
 Endwertsatz, 892
 Entwicklungsmittelpunkt, 401
 Ereignis, 998
 Ereignis-Algebra, 998, 999
 erfüllbar, 14
 erwartungstreu, 1059
 Erwartungswert, 1033
 Erweitern, 35
 erweiterte Matrix, 163
 erzeugen, 471
 Erzeugendensystem, 472
 Euklid'sche Norm, 486
 Euklid'scher Algorithmus, 31
 Euklid'scher Raum, 483
 Euler'sche Gleichung, 139
 Euler'scher Multiplikator, 746
 Euler-Cauchy-Polygonzugmethode, 715
 exakt, 744
 Existenzquantor, 22
 Exponent, 36
 Exponentialform, 139
 Exponentialfunktion, 103
 Exponentialfunktion zur Basis a , 104
 Exponentialreihe, 400
 Extremum, 280

F

Fakultät, 32
 Falk-Schema, 159
 Faltung, 834, 885, 919
 Faltungssatz, 837, 922
 Fast Fourier Transform, 923
 Fehler der besten Approximation, 798
 Fehler erster Art, 1015, 1072
 Fehler zweiter Art, 1015, 1072
 Fehlerfortpflanzungsgesetz, 1062
 Fehlerrechnung, 309, 614
 Fejér-Kern, 846
 Fensterfunktion, 938, 950
 FFT, 923
 Fibonacci-Zahlen, 231
 Finite-Elemente-Methode, 793
 Fixpunkt, 318, 567
 Fläche, 691
 Flussintegral, 693
 Flussmenge, 693
 Folge, 212
 Folgenglieder, 212
 Folgenhäufungspunkt, 232
 Folgenkompaktheit, 234
 Folgerung, 19
 Format, 154
 Formel von Cauchy-Hadamard, 413
 Fourier-Analyse, 817
 Fourier-Analyse, 817
 Fourier-Koeffizienten, 817, 877
 Fourier-Lagrange-Koeffizienten, 906
 Fourier-Matrix, 911
 Fourier-Reihe, 817, 877
 Fourier-Synthese, 817
 Fourier-Transformation, 862, 878
 Fourier-transformierbar, 862
 Fourier-Transformierte, 862
 Fourier-Umkehrtransformation, 865
 freier Vektor, 433
 Frequenz, 122
 Frequenzbereich, 838, 862
 Frequenzgang, 896
 Fundamentalsatz, 340
 Fundamentalsatz der Algebra, 147
 Fundamentalsystem, 754
 Funktion, 8
 Funktional, 899
 Funktionalanalysis, 557
 Funktionaldeterminante, 606
 Funktionenfolge, 402
 Funktionenreihe, 405
 Funktionsgraph, 7, 77
 Funktionswert, 8

G

Gammafunktion, 364

ganze Zahlen, 26
 ganzrationale Funktionen, 90
 ganzrationale Gleichungen, 71
 ganzrationale Gleichungen höherer Ordnung, 71
 Gauß'sche Normalgleichungen, 997
 Gauß'sche Normalverteilung, 1047
 Gauß-Seidel-Verfahren, 180
 Gauß-Algorithmus, 162
 Gauß-Jordan-Verfahren, 163
 Gebiet, 682
 gebrochen-rationale Funktion, 100
 gebrochen-rationale Gleichungen, 72
 gebrochen-rationale Ungleichung, 75
 geometrische Folge, 213
 geometrische Reihe, 237, 239
 geometrische Summe, 60
 geometrische Verteilung, 1022
 geometrisches Mittel, 69, 982
 gerade, 81
 Gerade, 91
 Gerüst, 532
 Gesamtschritt-Verfahren, 179
 geschlossen, 530
 geschlossene Kurve, 674
 Gesetz der großen Zahlen, 1042
 gestörte Fourier-Koeffizienten, 906
 Gewichtsfunktion, 896
 Gibbs-Effekt, 856
 Gittergenerierung, 808
 Glättungskerne, 836
 gleich, 133, 156
 gleichmäßig konvergent, 404
 gleichmäßig stetig, 284
 gleichverteilte Zufallsvariable, 1025
 globales Maximum, 280, 618
 globales Minimum, 280, 618
 goldener Schnitt, 56, 231, 551
 Gomory-Cuts, 648
 goniometrische Form, 138
 größter gemeinsamer Teiler, 31
 Grad, 90
 Gradient, 597
 Gradientenverfahren, 603
 Gram-Schmidt'sche
 Orthonormierungsverfahren, 493
 Graph, 529
 Grauwert, 977
 Grauwerte, 981
 Green'sche Formel, 696
 Green'sche Funktion, 697
 Grenzfunktion, 402
 Grenzwert, 217, 239, 258, 259
 größte untere Schranke, 48
 Grundfrequenz, 860
 Grundgesamtheit, 970, 998
 Gruppe, 46

H

Haar-Wavelet, 957
 Halbnormen, 671
 halboffenes Intervall, 51
 Halteproblem, 24
 Hamming-Fenster, 953, 968
 Hann-Fenster, 953, 968
 Harmonische, 821
 harmonische Funktion, 696
 harmonische Reihe, 240
 harmonische Schwingung, 122, 124
 häufigster Wert, 980
 Hauptachsentransformation, 991
 Hauptdiagonale, 155
 Hauptraum, 553
 Hauptsatz der Differenzial- und
 Integralrechnung, 340
 Hauptvektor, 553
 Hauptwert, 125, 144
 Heaviside-Funktion, 271
 hebbare Unstetigkeitsstelle, 285
 hermitesch, 174, 485
 Hesse'sche Normalform, 463
 Hesse-Matrix, 612
 Hilbert'sches Hotel, 41
 Hilbert-Raum, 560
 Hilfssatz, 23
 hinreichende Bedingung, 19
 Hochpunkt, 280
 Hölder-Ungleichung, 564
 holomorph, 419
 Homöomorphismus, 277
 homogen, 153
 homogene Differenzialgleichung, 741
 homogene Koordinaten, 504
 homogene Lösung, 751
 homogenes Gleichungssystem, 523
 Homogenität, 437, 483, 500
 Horn-Schunck-Gleichung, 673
 Horner-Schema, 28, 97
 Hyperbelfunktionen, 128, 129
 Hyperbelkosinus, 129
 Hyperbelkotangens, 129
 Hyperbelsinus, 129
 Hyperbeltangens, 129
 hyperbolisch, 791
 hypergeometrisch verteilt, 1023
 hypergeometrische Verteilung, 1023

I

identisch verteilt, 1023
 identische Abbildung, 501
 IDFT, 912
 imaginäre Einheit, 132
 Imaginärteil, 133
 Impedanz, 146

Implikation, 19
Impulsfunktion, 898
Index, 55
indirekter Beweis, 23
Infimum, 48
Inhalt, 377
inhomogen, 153
inhomogenes Gleichungssystem, 523
injektiv, 9
innere Knoten, 801
Innere-Punkte-Verfahren, 646
innerer Punkt, 256, 593
inneres Produkt, 435
instabil, 891
Integral, 331, 651
Integralfunktion, 339
Integralgleichung, 718
Integraltransformation, 862, 880
integrierbar, 331, 651
integrierender Faktor, 746
Interpolationspolynom, 92, 360
Intervalle, 51, 650
Intervallschachtelung, 51
Invariante, 31
inverse diskrete Fourier-Transformation, 912
inverse Matrix, 169
inverses Element, 429, 467
Inversionsmethode, 1030
invertierbar, 169
irrational, 50, 52

J

Jacobi-Matrix, 606
Jacobi-Verfahren, 179
Jordan-Kurve, 674
Jordan-Normalform, 555

K

Kästchensatz, 190
Kanten, 529
Karnaugh-Veitch-Diagramm, 17
Kehrwertfunktion, 100
Kern, 508
Kette von k Hauptvektoren, 553
Kettenregel, 297, 607
kleinste obere Schranke, 47
kleinstes gemeinsames Vielfaches, 32
Knoten, 529, 801
Knotenfunktion, 802
Koeffizienten, 90, 157
Koeffizientenmatrix, 162
Koeffizientenvergleich, 94
koerziv, 796
kollinear, 428
Kombinationen, 33, 1004

Kombinationen mit Wiederholung, 1005
Kombinationen ohne Wiederholung, 1004
kommutativ, 467
kommutative Gruppe, 46
Kommutativgesetze, 6, 14, 45, 429, 437
kompakt, 670
komplanar, 475
Komplement, 2
komplexe n -te Wurzeln, 143
komplexe Zahlen, 131
komplexer Widerstand, 146
Komponente, 434
Komponenten, 155, 491
Kondition, 573, 574
Konfidenzintervall, 1063
konjugiert komplexe Zahl, 134
konjugierte Gradienten, 624
Konjunktion, 13
konjunktive Normalform, 17
konkav, 388
konsistent, 1059
Konsistenz, 573
konstante Folgen, 213
Kontingenztafel, 985
kontinuierliches Spektrum, 862
Kontraktion, 318
konvergent, 217, 239, 258, 259, 402, 560, 592, 594
Konvergenzradius, 412
konvex, 388, 613
Koordinaten, 432, 491
Koordinatenform, 462
Koordinatengleichung, 453, 461
Koordinatentransformation, 518
Körper, 45
korrekt gestelltes Problem, 714
Korrelationskoeffizient, 1041
Kosinus, 113
Kosinus-Form, 822
Kosinus-Reihe, 819
Kosinus-Satz, 118
Kosinushyperbolicus, 129
Kotangens, 113
Kotangenshyperbolicus, 129
Kovarianz, 1041
Kovarianzmatrix, 988
Kreis, 530
Kreisfrequenz, 122
Kreissatz von Gerschgorin, 547
Kreuzprodukt, 4
kritische Stellen, 100, 287
Kronecker-Delta, 490
kubische Splinefunktion, 312
kubisches Polynom, 96
Kurve, 674
Kurvendiskussion, 388
Kurvenintegral, 681

Kürzen, 35

L

Länge, 433
 Lagrange'sches Interpolationspolynom, 93
 Lagrange'sches Knotenpolynom, 93
 Lagrange-Darstellung, 384
 Lagrange-Multiplikator, 633
 Länge, 530
 Länge der Kurve, 677
 Laplace-Differenzialgleichung, 685
 Laplace-Experiment, 1001
 Laplace-Operator, 680
 Laplace-Transformierte, 879
 Laplace-transformierbar, 879
 Lastvektor, 801
 Leakage, 938
 Lebesgue-Integral, 374
 Lebesgue-integrierbar, 380
 Lebesgue-Maß, 375, 377
 Lebesgue-messbar, 375, 377, 378, 669
 Lebesgue- σ -Algebra, 375
 Leck-Effekt, 938, 946
 leere Menge, 2
 Lehr'sches Dämpfungsmaß, 783
 Leibniz'sche Regel für Parameterintegrale, 656
 Leibniz'sche Zinseszinsformel, 61
 Leibniz-Formel für Determinanten, 193
 Leibniz-Kriterium, 254
 Leibniz-Regel, 310
 Leitkoeffizient, 90
 Lemma, 23
 Lemma von Céa, 798
 Lemma von Lax und Milgram, 796
 Lenz'sche Regel, 117
 Levenberg-Marquardt-Verfahren, 638
 lexikographisch, 27
 Lexikographische Ordnung, 27
 Likelihood-Funktion, 1060
 Limes, 217
 linear, 500
 linear abhängig, 474
 linear normierter Raum, 558
 linear unabhängig, 474, 533
 lineare Abbildung, 500
 lineare Funktion, 500
 lineare Gleichungen, 71
 lineare Hülle, 471
 lineare Optimierung, 638
 lineare Programmierung, 638
 lineare Ungleichung, 74
 lineares Funktional, 502, 570
 lineares Gleichungssystem, 153
 Linearfaktoren, 147
 linearisierte Fehlerschätzung, 615

Linearität, 334, 500
 Linearkombination, 157, 470, 478
 Linienelement, 715
 linksgekrümmt, 388
 linksseitig bestimmt divergent, 264
 linksseitig differenzierbar, 293
 linksseitig konvergent, 264
 linksseitig stetig, 270
 linksseitige punktierte Umgebung, 256
 linksseitiger Grenzwert, 264
 Lipschitz-Bedingung, 717
 Logarithmische Ableitung, 304
 logarithmisches Integrieren, 350
 Logarithmus zur Basis a , 104
 lokales Extremum, 388
 lokales Maximum, 280, 618
 lokales Minimum, 280, 618
 Lot, 441
 Lotvektor, 496
 L_p -Normen, 562
 LR-Zerlegung, 177
 LU-Zerlegung, 177

M

m -fache Nullstelle, 94
 m -Tupel, 156
 Mächtigkeit, 127
 MacLaurin-Entwicklung, 383
 MacLaurin-Reihe, 399
 Majorante, 246
 Mandelbrot-Menge, 214
 Masche, 530
 Masse, 532
 mathematisch positiver Sinn, 114
 Mathematische Modellierung, 705
 mathematisches Pendel, 709
 Matrix, 154
 Matrix-Norm, 571
 maximaler Baum, 532
 Maximalfehlerabschätzung, 615
 Maximum, 47
 Maximum-Likelihood-Schätzer, 1060
 Maximumprinzip, 697
 Maxwell'sche Gleichungen, 699
 Median, 979
 Menge, 2
 Merkmale, 971
 Merkmalsausprägungen, 971
 Merkmalsträger, 970
 Merkmalswerte, 971
 Messbarkeit, 1019
 Methode der kleinsten Quadrate, 992
 Minimum, 48
 Minkowski-Ungleichung, 564
 Minorante, 246
 Mittelpunktregel, 360

Mittelwert, 1033
 Mittelwertsatz der Differenzialrechnung, 315
 Mittelwertsatz der Integralrechnung, 337
 mittlere Geschwindigkeit, 290
 mittlerer Fehler der Einzelmessungen, 1061
 mittlerer Fehler des Mittelwerts, 1061
 Modalwert, 980
 modulo, 46
 Modus Ponens, 23
 Momentangeschwindigkeit, 290
 Monome, 90
 monoton, 81, 215
 monoton fallend, 81, 215
 monoton wachsend, 81, 215
 Monte-Carlo-Methode, 216
 Multilinearform, 191
 Multiplikation mit einem Skalar, 485
 Multiskalenanalyse, 956

N

n -mal stetig differenzierbar, 310
 n -te Einheitswurzel, 906
 n -te Potenz, 36
 n -te Wurzel, 37, 56
 Nabla-Operator, 597
 Nadelexperiment von Buffon, 1081
 natürliche Zahlen, 25
 natürlicher Logarithmus, 103
 Negation, 13
 negativ definit, 622
 negative Korrelation, 986
 Neumann-Randbedingung, 696
 neutrales Element, 429, 467
 neutrales Element der Addition, 134
 neutrales Element der Multiplikation, 134
 Newton-Cotes-Formeln, 361
 Newton-Verfahren, 306
 Nitsche-Trick, 807
 nodale Basis, 802
 Norm, 486, 557
 normal, 552
 Normalbereich, 658
 Normaleneinheitsvektor, 463
 Normalparabel, 78
 Normalverteilung, 1049
 normierte Zeilenstufenform, 513
 normierter Vektorraum, 558
 normiertes Polynom, 90
 notwendige Bedingung, 19
 Nullabbildung, 502
 Nullfolge, 218
 Nullhypothese, 1071
 Nullmatrix, 155
 Nullphasenwinkel, 122
 Nullpunkt, 432
 Nullraum, 510

Nullstelle mit Vielfachheit, 147
 Nullstellen, 80
 Nullstellensatz von Bolzano, 279
 Nullvektor, 428, 468
 Nullvektorraum, 468

O

obere Schranke, 47, 213
 Oberflächenintegral, 691
 Obersumme, 329, 650
 offen, 256, 530, 593
 offene Mengen, 277
 offenes Intervall, 51
 Operatornorm, 567
 Ordinate, 77
 Ordnung, 749
 Ordnung des Pols, 288
 Ordnungsrelation, 27
 ordnungsvollständiger geordneter Körper, 48
 orthogonal, 175, 436, 488
 orthogonale Projektion, 441, 495
 Orthogonalität, 826
 Orthogonalsystem, 490
 Orthonormalbasis, 490
 Orthonormalsystem, 490, 492
 Ortskurve, 896
 Ortsvektor, 432
 Oversampling, 933

P

p-q-Formel, 55
 p-Quantil, 980
 Paare, 4
 parabolisch, 791
 Paradoxon von Simpson, 1076
 Parallelepipet, 448
 Parameterdarstellung, 674, 691
 Parameterform, 454, 460
 Parseval'sche Gleichung, 830, 865
 Partialbruchzerlegung, 352
 partielle Ableitung, 597
 partielle Differenzialgleichung, 602, 685, 788
 partielle Integration, 345
 partikuläre Lösung, 726, 751
 Pascal'sches Dreieck, 34
 Pearson'scher Korrelationskoeffizient, 987
 Periode, 38, 83
 periodische Funktion, 83
 permanente Schwingung, 786
 Permutation, 32
 Phase, 122
 Phasengang, 896
 Phasenspektrum, 822
 Phasenwinkel, 122
 Pivot-Suche, 166

Poincaré-Ungleichung, 672
 Poisson-Gleichung, 685, 799
 Poisson-Summutationsformel, 867
 Poisson-verteilt, 1030
 Pol, 285
 Polarform, 138
 Polarkoordinaten, 138
 Polstelle, 285
 Polynomdivision, 95
 Polynome, 90
 polynomiale Regression, 996
 Population, 970
 positiv definit, 622
 positive Definitheit, 483, 485
 positive Korrelation, 986
 Potenzen, 36
 Potenzfunktion, 101
 Potenzial, 683, 744
 Potenzialgleichung, 685
 Potenzmenge, 2
 Potenzregeln, 37
 Potenzreihe, 401
 Prädikat, 18
 Prädikatenlogik, 18
 Primfaktorzerlegung, 30
 primitiv, 909
 primitive Periode, 83
 Primzahl, 30
 Principal Component Analysis, 991
 Prinzip gleichgradiger Beschränktheit, 577
 Produktregel, 295
 Produktzeichen, 60
 Projektion, 384, 441, 818
 Prozentrechnung, 37
 Punkt-Richtungsform, 454, 460
 Punkt-Steigungsform, 91
 punktierte δ -Umgebung, 256
 punktsymmetrisch zum Ursprung, 81
 punktweise konvergent, 402
 Punktwolke, 985

Q

Quadranten, 77
 quadratische Gleichung, 71
 quadratische Matrix, 155
 quadratische Ungleichung, 74
 Quadraturformel, 359
 qualitative Merkmale, 971
 quantitative Merkmale, 971
 Quantoren, 21
 Quartile, 980
 Quartilsabstand, 983
 quellenfreies Feld, 679
 Quotientenkriterium, 250
 Quotientenregel, 295

R

Radiocarbonmethode, 725
 Rand, 593
 Randbedingung, 713, 789
 Randhäufigkeiten, 985
 Randknoten, 801
 Randpunkt, 593
 Rang, 510, 514
 rationale Funktionen, 90
 rationale Zahlen, 35
 Raum, 466
 Realteil, 133
 Rechteckimpuls, 863
 Rechteckregel, 360
 rechtsgekrümmt, 388
 rechtsseitig bestimmt divergent, 264
 rechtsseitig differenzierbar, 293
 rechtsseitig konvergent, 264
 rechtsseitig stetig, 270
 rechtsseitige punktierte Umgebung, 256
 rechtsseitige Umgebung, 256
 rechtsseitiger Grenzwert, 264
 Rechtssystem, 431
 reelle Funktionen, 77
 reelle Zahlen, 49
 Reflexivität, 27
 Regressionsgerade, 992
 regulär, 169, 193, 516
 reguläre Distribution, 899
 Reihe, 227, 236, 238
 rektifizierbar, 677
 rekursiv, 32
 relativer Fehler, 309, 615
 relativer linearisierter Fehler, 615
 relatives Maximum, 280, 618
 relatives Minimum, 280, 618
 repräsentative Stichprobe, 971, 1056
 Residuensatz, 687
 Residuum, 624, 995
 Resonanz, 779
 Restklasse, 46
 Restklassenring, 46
 Ricatti-Differenzialgleichung, 742
 Richardson-Verfahren, 576
 Richtung, 427
 Richtungsableitung, 601
 Richtungsfeld, 715
 Richtungsvektor, 453
 Richtungsvektoren, 460
 Riemann-Integral, 331, 651
 Riemann-integrierbar, 331, 651
 Riemann'sches Lokalisationsprinzip, 841
 Riemann-Lebesgue-Lemma, 848, 864
 Riemann-Oberintegral, 331, 651
 Riemann-Unterintegral, 331, 651
 Riemann-Zwischensumme, 332
 Riesz'scher Darstellungssatz, 584

Ring, 46
Ritz-Galerkin-Verfahren, 795
Ritz-Projektion, 797
RSA, 46
Rückweisungsmethode, 1031
Runge-Kutta-Verfahren, 721

S

Sägezahnfunktion, 84
Sampling, 905
Sarrus'sche Regel, 184
Sattelpunkt, 314, 389
Satz, 23
Satz über die implizite Funktion, 610
Satz von Arzela, Osgood und Lebesgue, 411
Satz von Banach-Schauder, 583
Satz von Bayes, 1016
Satz von Bolzano-Weierstraß, 234
Satz von Carleson, 845
Satz von Cayley-Hamilton, 544
Satz von de Moivre und Laplace, 1052
Satz von Fermat, 314, 618
Satz von Fubini, 653
Satz von Fubini-Tonelli, 654
Satz von Gauß, 694
Satz von Green, 688
Satz von Hahn-Banach, 583
Satz von Heine-Borel, 256
Satz von L'Hospital, 323, 325
Satz von Laplace, 1001
Satz von Meyers und Serrin, 672
Satz von Moivre, 141
Satz von Picard und Lindelöf, 716
Satz von Plancherel, 865
Satz von Pythagoras, 118, 433, 486, 490
Satz von Rolle, 315
Satz von Schwarz, 611
Satz von Stokes, 698
Satz von Taylor, 384, 613
Satz von Thales, 439
Satz von Vieta, 97, 148
Säulendiagramm, 973
Schätzer, 1058
Schätzfunktion, 1058
schiefsymmetrisch, 174
Schlupfvariablen, 641
schnelle Faltung, 928
schnelle Wavelet-Transformation, 961
Schönhage-Strassen-Algorithmus, 928
schwach differenzierbar, 584
schwache Ableitung, 584, 670
schwache Form, 800
schwaches Problem, 794
Schwartz-Raum, 865
Schwebung, 124
Schwingungsdauer, 122

Sekantenverfahren, 307
selbstadjungiert, 174
Sensitivität, 1014
Shannon Sampling Theorem, 941
Shannon-Nyquist-Bedingung, 942
sicheres Ereignis, 998
Siebformel von Sylvester und Poincaré, 1001
Sigma-Algebra, 999
sigmoide Funktion, 604
signifikant, 1072
Signifikanzniveau, 1072
Signifikanztest, 1072
Signumfunktion, 89
Simplex-Algorithmus, 642
Simpson-Regel, 361
singulär, 169, 193
Sinus, 113
Sinus Cardinalis, 271
Sinus-Form, 822
Sinus-Reihe, 819
Sinus-Satz, 119
Sinushyperbolikus, 129
Skalar, 156, 427
skalare Funktionen, 673
skalare Multiplikation, 468
skalares Vielfaches, 156
Skalarfeld, 672
Skalarmultiplikation, 468
Skalarprodukt, 435, 468, 483, 485
Skalierungsfunktion, 956
Skalierungsgleichung, 960
Sobolev-Raum, 570, 585
SOR-Verfahren, 181
Spalten, 154
Spaltenindex, 154
Spaltenmatrix, 155
Spaltenraum, 510
Spaltensummennorm, 571
Spaltenvektor, 155
Spannweite, 983
Spat, 448
Spatprodukt, 448
Spektralnrm, 571
Spektralradius, 571
Spektrum, 825
spezielle Lösung, 726
Spezifität, 1014
Spline, 93
Splines, 312
Sprungantwort, 896
Sprungstelle, 285
Stabdiagramm, 973
stabil, 274, 891
Stabilität, 573, 714
Stammfunktion, 339
Standard-Einheitsvektoren, 434
Standardabweichung, 1038

Standardbasis, 479
 Standardform, 639
 Standardnormalverteilung, 1048
 Standardskalarprodukt, 484
 starke Korrelation, 988
 Starkes Gesetz der großen Zahlen, 1045
 stationäre Schwingung, 786
 stationärer Zustand, 537
 statistische Einheiten, 970
 statistische Elemente, 970
 statistische Masse, 970
 Steifigkeitsmatrix, 794, 801
 stetig, 270, 277, 596
 stetig differenzierbar, 294, 612, 674
 stetig ergänzbar, 285
 stetig fortsetzbar, 285
 stetige Merkmale, 971
 Stetigkeit, 566, 596
 Stetigkeit im Mittel, 848
 Stichprobe, 971
 Stichprobenfunktion, 1058
 Stichprobenvariable, 1057
 Stirling'sche Formel, 368
 stochastisch unabhängig, 1009, 1026
 stochastische Konvergenz, 1046
 Strahlensatz, 429
 streng diagonaldominant, 576
 streng monoton fallend, 81, 215
 streng monoton wachsend, 81, 215
 Streudiagramm, 985
 striktes globales Maximum, 280, 618
 striktes globales Minimum, 280, 618
 striktes lokales Maximum, 280, 618
 striktes lokales Minimum, 280, 618
 Strukturvariablen, 641
 stückweise lineare Funktion, 78
 Student'sche t -Verteilung, 1068
 Stützstellen, 359
 Substitutionsregel, 347, 661
 Summe, 429
 Summenzeichen, 58
 Supremum, 47
 surjektiv, 10
 Symmetrie, 483
 symmetrisch, 174, 796
 symmetrisch zur y -Achse, 81

T

Tangens, 113
 Tangenshyperbolicus, 129
 Tangente, 676
 Tangenteneinheitsvektor, 676
 Tangentialebene, 601
 Tautologie, 14
 Taylor-Entwicklung, 383
 Taylor-Polynom, 383

Taylor-Reihe, 399
 Taylor-Summe, 383
 Teile und herrsche, 924
 Teilerhebung, 971
 Teilgraph, 529
 Teilmenge, 2
 Teilüberdeckung, 257
 Teleskopprodukt, 60
 Teleskopsumme, 59
 Thaleskreis, 439
 Tiefpass, 734
 Tiefpassfilter, 840, 875
 Tiefpunkt, 280
 Topologie, 256, 277
 Topologischer Raum, 277
 Tortendiagramm, 973
 total differenzierbar, 599, 606
 total geordnet, 27
 totale Wahrscheinlichkeit, 1013
 Totalerhebung, 971
 Totales Differenzial, 600
 Träger, 802
 Transformationsmatrix, 518
 transiente Schwingung, 786
 transitiv, 20
 Transitivität, 27
 transzendente Funktionen, 102
 Trapezregel, 360
 Trennung der Variablen, 723, 735
 Triangulierung, 808
 Tridiagonalmatrix, 177
 trigonometrische Interpolation, 930
 trigonometrisches Polynom, 818
 triviale Linearkombination, 474
 Tupel, 4

U

überabzählbar, 54
 überbestimmtes Gleichungssystem, 168
 Übergangsfunktion, 896
 Übertragungsfunktion, 895
 Übertragungsprinzip, 266, 594
 Umkehrabbildung, 10
 Umkehrfunktion, 10, 77
 unbestimmt divergent, 239
 unbestimmtes Integral, 339
 unbiased, 1059
 Undersampling, 934
 uneigentlich integrierbar, 362, 363
 uneigentliches Integral, 362
 unerfüllbar, 14
 ungerade, 81
 Ungleichung von
 Poincaré, 672
 Ungleichung von Tschebycheff, 1040
 Ungleichungen, 65, 66

unimodale Verteilung, 981
 unitär, 176
 unkorreliert, 988, 1041
 unmögliches Ereignis, 998
 Unsicherheit des Mittelwerts, 1061
 Unsicherheitsintervall, 615
 Unstetigkeitsstelle erster Art, 285
 Unstetigkeitsstelle zweiter Art, 285
 unterbestimmtes Gleichungssystem, 167
 untere Schranke, 48
 Unterraum, 472
 Untersumme, 329, 650
 Untervektorraum, 472
 unverfälscht, 1059
 Urbild, 8
 Ursprung, 432

V

Vandermonde-Matrix, 195
 Varianz, 1038
 Variation der Konstanten, 726
 Variationen, 1003
 Variationen mit Wiederholung, 1004
 Variationen ohne Wiederholung, 1003
 Vektor, 155, 427, 444
 Vektoren, 468
 Vektorfeld, 673
 Vektorprodukt, 444
 Vektorraum, 466
 Venn-Diagramme, 4
 verallgemeinerter Mittelwertsatz der
 Differenzialrechnung, 317
 Verbindungszweige, 532
 Vereinigung, 4
 Verfahren des steilsten Abstiegs, 603
 Verfahren von Heun, 720
 Verfeinerungsgleichung, 960
 Vergleichskriterium, 246, 366, 405
 Verkettung, 87
 verschiebeinvariant, 921
 Versuch, 998
 Verteilung, 1021
 Verteilungsfunktion, 1023
 Vertrauensintervall, 1063
 Vollständigkeitsaxiom, 49
 Volladdierer, 16
 Vollerhebung, 971
 vollständig, 44, 560
 Vollständige Induktion, 41
 von höchstens exponentiellem Wachstum,
 881
 Vorkonditionierung, 574
 Vorzeichenfunktion, 89

W

Wahrscheinlichkeitsraum, 1000
 Wahrheitswerte, 12
 Wahrscheinlichkeitsmaß, 1000
 Wavelet-Transformation, 955
 Weg, 530
 Wegunabhängigkeit, 682
 Weierstraß-Funktion, 406
 Wellengleichung, 789
 Wendepunkt, 388
 Wertebereich, 8
 Wertedifferenz, 308
 Wertemenge, 8, 213
 wesentliches Supremum, 563
 Whittaker'sche Rekonstruktion, 942
 Winkel, 438, 488
 Winkelgeschwindigkeit, 122
 Wirbelfluss, 698
 wirbelfreies Feld, 680
 Wirkwiderstand, 146
 Wronski-Matrix, 754
 Wurzel, 531
 Wurzelfunktion, 101
 Wurzelgleichungen, 72
 Wurzelkriterium, 253
 Wurzeln, 36

Z

Z-Transformation, 249
 Z-Transformierte, 249
 Zähldichte, 1001
 Zahlengerade, 44
 Zeigerdiagramm, 123
 Zeilen, 154
 Zeilenindex, 154
 Zeilenmatrix, 155
 Zeilenraum, 511
 Zeilensummennorm, 571
 Zeilenvektor, 155
 Zeitbereich, 862
 zeitinvariant, 921
 zeitinvariantes Übertragungssystem, 894
 Zeitreihendarstellung, 974
 Zentraler Grenzwertsatz, 1050
 Zentralwert, 979
 Zerlegung, 329, 650
 Ziehen mit Zurücklegen, 1004
 Ziehen ohne Zurücklegen, 1003
 Zielfunktion, 638
 Zufall, 997
 Zufallsexperiment, 997
 Zufallsstichprobe, 1057
 Zufallsvariable, 1019
 Zufallsvariablen, 1055
 zulässige Lösung, 641
 zusammenhängend, 530, 682

Zwangsbedingungen, 631
Zwei-Punkte-Form, 91
zweidimensionale Häufigkeitstabelle, 985
Zweierkomplement, 29
Zweige, 529
Zweigefolge, 530
zweiseitiges Konfidenzintervall, 1064
Zwischenpunktwahl, 332
Zwischenwertsatz, 278
Zylinderkoordinaten, 665



Willkommen zu den Springer Alerts

Jetzt
anmelden!

- Unser Neuerscheinungs-Service für Sie:
aktuell *** kostenlos *** passgenau *** flexibel

Springer veröffentlicht mehr als 5.500 wissenschaftliche Bücher jährlich in gedruckter Form. Mehr als 2.200 englischsprachige Zeitschriften und mehr als 120.000 eBooks und Referenzwerke sind auf unserer Online Plattform SpringerLink verfügbar. Seit seiner Gründung 1842 arbeitet Springer weltweit mit den hervorragendsten und anerkanntesten Wissenschaftlern zusammen, eine Partnerschaft, die auf Offenheit und gegenseitigem Vertrauen beruht.

Die SpringerAlerts sind der beste Weg, um über Neuentwicklungen im eigenen Fachgebiet auf dem Laufenden zu sein. Sie sind der/die Erste, der/die über neu erschienene Bücher informiert ist oder das Inhaltsverzeichnis des neuesten Zeitschriftenheftes erhält. Unser Service ist kostenlos, schnell und vor allem flexibel. Passen Sie die SpringerAlerts genau an Ihre Interessen und Ihren Bedarf an, um nur diejenigen Information zu erhalten, die Sie wirklich benötigen.

Mehr Infos unter: springer.com/alert