Höhere Mathematik

Ein Begleiter durch das Studium

Bearbeitet von
Karlheinz Spindler

Nachdruck 2010. Buch. 893 S. Hardcover
ISBN 978 3 8171 1872 4
Format (B x L): 22 x 28,5 cm

Karlheinz Spindler

Höhere Mathematik

Ein Begleiter durch das Studium

Verlag Harri Deutsch

Höhere Mathematik

Verlag Harri Deutsch – Karlheinz Spindler: *Höhere Mathematik – Ein Begleiter durch das Studium* – (978-3-8171-1872-4)

Karlheinz Spindler

Höhere Mathematik

Ein Begleiter durch das Studium

Verlag Harri Deutsch

Verlag Harri Deutsch – Karlheinz Spindler: *Höhere Mathematik – Ein Begleiter durch das Studium* – (978-3-8171-1872-4)

Der Autor

Prof. Dr. Karlheinz Spindler studierte Mathematik, Mechanik und Geschichte an der Technischen Hochschule Darmstadt. Nach Abschluß seines Diploms und des Staatsexamens für das Lehramt an Gymnasien war er als Wissenschaftlicher Mitarbeiter an der TH Darmstadt tätig und wurde dort über ein Thema aus der Strukturtheorie Liescher Algebren promoviert. Anschließend arbeitete er zunächst zwei Jahre lang als Visiting Assistant Professor an der Louisiana State University in Baton Rouge (USA) und dann fünf Jahre lang bei einem Unternehmen der Raumfahrtindustrie am European Space Operations Centre (ESOC) in Darmstadt. Im Jahr 1997 wurde er zum Professor für Mathematik und Datenverarbeitung an die Fachhochschule Wiesbaden (seit dem 1. September 2009 Hochschule RheinMain) berufen. Dort leitet er den Studiengang "Angewandte Mathematik", der im Wintersemester 2010/2011 seinen Betrieb aufnahm und an dessen Konzeption er maßgeblich beteiligt war.

Die Webseite zum Buch

http://www.harri-deutsch.de/1872.html

Der Verlag

Wissenschaftlicher Verlag Harri Deutsch GmbH
Gräfstraße 47
60486 Frankfurt am Main
verlag@harri-deutsch.de
www.harri-deutsch.de

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie.
Detaillierte bibliographische Daten sind im Internet unter http://dnb.d-nb.de abrufbar.

ISBN 978-3-8171-1872-4

Korrigierter Nachdruck der 1. Auflage (2010), 2011

Druck: fgb. freiburger graphische betriebe (www.fgb.de)
Printed in Germany

Meiner Frau und meinen Kindern
in Liebe und Dankbarkeit

Vorwort

Im Englischen ist *Mathematics* (oder, wie Newton noch schrieb, "Mathematicks") ein Pluralwort, und auch im Französischen spricht man von *les mathématiques*. In der Tat hat sich die Mathematik, erwachsen aus den einfachen Anwendungen des Zählens von Gegenständen und des Messens von Größen (und damit den grundlegenden Disziplinen der Arithmetik und der Geometrie) in eine Vielzahl von Einzeldisziplinen aufgefächert, von denen manche auf den ersten Blick nur wenig miteinander verbindet. Diese Tendenz zur Zersplitterung wird verstärkt durch die Vielfalt der Anwendungsgebiete, in denen mathematische Methoden eingesetzt und aus deren Blickwinkel heraus mathematische Begriffe und Verfahren entwickelt werden. Mathematik durchdringt mittlerweile fast alle Lebensbereiche, und mathematische Methoden werden angewandt, um verschiedenste naturwissenschaftliche, technische, wirtschaftliche und gesellschaftliche Prozesse zu beschreiben, zu verstehen und zu optimieren.

Trotz der Vielfalt ihrer Teildisziplinen und ihrer Anwendungsbereiche hat sich die Mathematik eine erstaunliche Einheit bewahrt, und viele der faszinierendsten mathematischen Einsichten und Entdeckungen bestehen gerade darin, Zusammenhänge zwischen Sachverhalten aufzudecken, die auf den ersten Blick nichts miteinander zu tun haben. Diese Einheit kann nur erreicht werden durch einen Prozeß der Abstraktion, der versucht, im Dickicht vieler Einzelfakten nach grundlegenden allgemeinen Strukturen und Prinzipien zu suchen. Erkenntnisse und Einsichten entstehen ja nicht durch das bloße Sammeln einzelner Tatsachen und Beobachtungen, sondern erst durch deren Deutung und Einordnung in einen zugrundeliegenden Sinnzusammenhang. Daß die Suche nach grundlegenden Strukturen und Prinzipien nicht vergeblich ist – daß es diese also überhaupt gibt und daß wir Menschen auch das Potential haben, sie zu finden – liegt zum einen daran, daß die Welt, in der wir leben, kein blindes Chaos ist, sondern ein geschaffener, nach Maß, Zahl und Gewicht geordneter Kosmos, zum andern daran, daß wir selbst Teil dieses geschaffenen Kosmos sind und als vernunftbegabte Wesen die Fähigkeit haben, Gottes Gedanken nach-zudenken und obwaltende Ordnungsprinzipien aufzudecken, wenn auch – aufgrund der Endlichkeit und Beschränktheit unserer Existenz – nur stückweise und in Teilbereichen.

Als Disziplin hat die Mathematik einen eigentümlichen Doppelcharakter; sie ist sowohl Königin der Wissenschaften und Verkörperung reinsten Denkens als auch Dienstmagd vieler Anwendungsdisziplinen und Methodenreservoir zur Lösung verschiedenster Aufgaben der Praxis. Viele Anwendungsdisziplinen (wie etwa Bild- und Signalverarbeitung, Kontrolltheorie, Kontinuumsmechanik und Materialwissenschaften, Codierungstheorie und Kryptographie) haben eine nahezu vollständige Mathematisierung erfahren und erfordern den Einsatz komplexer und tiefgehender mathematischer Methoden. Die Geschichte der Mathematik ist in der Tat gekennzeichnet durch ein Wechselspiel zwischen dem Lösen ganz praktischer Probleme und einem nachfolgenden tieferen Nachdenken über die Natur dieser Probleme, das dann oft eine Eigendynamik entwickelt und aus dem sich "rein theoretische" Fragestellungen ergeben (die dann aber oft in unerwarteter Weise wieder auf die Praxis zurückwirken). Dieses Wechselspiel zeigt sich beispielsweise im Wirken von Carl Friedrich Gauß (1777-1855), der das bemerkenswerte Talent hatte, den mathematischen Kern von Fragestellungen in Anwendungsdisziplinen wie Astronomie, Vermessungswesen und Elektrizitätslehre herauszuschälen und dadurch einerseits mathematische Disziplinen ungemein befruchtete oder gar erst begründete (Ausgleichsrechnung, Wahrscheinlichkeitsrechnung, Differentialgeometrie, Feldtheorie, Topologie), andererseits aber auch seine mathematischen Einsichten nutzte, um außerordentlich effektive Lösungsverfahren für die von ihm untersuchten Anwendungsprobleme zu entwickeln. Tatsächlich wird in vielen Fällen die Lösung eines ganz konkreten Einzelproblems erst durch eine eher abstrakte Herangehensweise ermöglicht, ohne die man vor lauter Bäumen den Wald nicht sieht. Abstraktion in der Mathematik ist also nicht Selbstzweck, sondern Mittel zum begrifflichen Verständnis von Sachverhalten und zum Finden von Lösungen für ganz konkrete Aufgabenstellungen.

Man kann durchaus Mathematik um ihrer selbst willen (sozusagen als "l'art pour l'art") betreiben, aber auch wenn man eher an der Verwendung mathematischer Methoden zur Lösung von Anwendungsproblemen interessiert ist, ist man gut beraten, sich ein klares Verständnis mathematischer Begriffe anzueignen und mathematische Theorien geistig zu durchdringen, statt sie nur rezeptartig anzuwenden. Die mathematisch zunehmend komplexeren Anforderungen in technisch-industriellen Anwendungen erfordern vor allem ein fundamentales Verständnis abstrakter Zusammenhänge. Eine Ausbildung in traditioneller "Ingenieursmathematik", deren Schwerpunkt auf der Vermittlung von Rechenrezepten und deren Umsetzung auf dem Computer liegt, wird diesen Anforderungen immer weniger gerecht. So erfordert etwa die mathematische Modellierung eines physikalischen, chemischen oder technischen Sachverhalts in erster Linie ein großes Repertoire an Mathematik. Ohne dieses Repertoire kann eine Ausbildung in Modellierung *per se* auf nichts zurückgreifen. Insbesondere erfordert der sachgerechte Einsatz mathematischer Methoden zur Lösung von Anwendungsproblemen Verständnis für die Erfassung naturwissenschaftlicher Konzepte durch mathematische Begriffsbildungen: Ableitungen als Änderungsraten, Integration als Aggregation von Einzelgrößen zu einer Gesamtgröße, Differentialformen als Flüsse, Integralsätze als Ausdruck von Bilanzgleichungen, Gruppen zur Beschreibung von Symmetrien, Differentialgleichungen als Entwicklungsgleichungen dynamischer Systeme, und so weiter. Ich habe mich daher beim Schreiben dieses Buches bemüht, die hinter mathematischen Begriffsbildungen steckenden Motivationen deutlich werden zu lassen.

Dies erschien mir um so wichtiger, als zuweilen auch die Anwendung mathematischer Methoden auf neuartige Praxisaufgaben zu einer Neubewertung und Erweiterung längst etablierter mathematischer Begriffe führt. Ein gutes Beispiel hierfür ist die Entwicklung der mathematischen Kontrolltheorie, in der konkrete Anwendungsprobleme zu einer kritischen Befassung und Auseinandersetzung mit grundlegenden Konzepten wie dem Ableitungsbegriff für Funktionen und dem Lösungsbegriff für Differentialgleichungen und damit zur Herausbildung neuer mathematischer Theorien führten (nichtglatte Analysis, schwache Lösungsbegriffe – etwa Viskositätslösungen – für partielle Differentialgleichungen, Behandlung gewöhnlicher Differentialgleichungen mit unstetiger rechter Seite). Das Verhältnis zwischen mathematischer Theoriebildung einerseits, Anwendung mathematischer Methoden auf Praxisprobleme andererseits ist also ein wechselseitiges. Ich überlasse es einem Würdigeren als mir, hierzu noch einige Anmerkungen zu machen, und zitiere (im Kasten rechts) eine am 8. September 1930 in Königsberg gehaltende Rundfunkansprache von David Hilbert (1862-1943), einem der bedeutendsten Mathematiker des frühen 20. Jahrhunderts. (Es handelt sich um einen Auszug aus einer Rede mit dem Titel "Naturerkennen und Logik", die Hilbert beim Kongreß der Vereinigung deutscher Naturwissenschaftler und Ärzte hielt.)

Die Entstehung dieses Buches ist untrennbar verbunden mit den konzeptionellen Vorarbeiten für den Studiengang "Angewandte Mathematik", der im Wintersemester 2010/2011 seinen Betrieb am Studienort Wiesbaden der Hochschule RheinMain aufnahm und innerhalb dessen dieses Buch als Lehrbuch eingesetzt wird. Dennoch handelt es sich nicht um ein Buch über Anwendungen der Mathematik; sein Ziel ist vielmehr die Vermittlung eines soliden und tragfähigen Grundlagenwissens in mathematischen Schlüsseldisziplinen, auf dem eine spätere Einarbeitung in mathematische Spezialdisziplinen oder Anwendungsgebiete problemlos aufbauen kann. Das Buch will nicht nur mathematisches Methodenwissen vermitteln, sondern auch Verständnis für die Herausbildung mathematischer Begriffe und Theorien wecken, Zusammenhänge zwischen verschiedenen mathematischen Disziplinen aufzeigen und den Anwendungsreichtum der Mathematik wenigstens andeuten. Ich hoffe ferner, daß beim Lesen des Buches auch etwas von der Schönheit und Klarheit der Mathematik deutlich wird. Zu sehen, wie sich dicht gewobene Theoriegebäude aus (ganz wenigen und sehr einfachen) geeignet gewählten Grundbegriffen entwickeln lassen und wie sich solche Theoriegebäude zur Beschreibung, zum Verständnis und zur Gestaltung der physischen Welt einsetzen lassen, ist (in den Worten Harro Heusers) "eine geistige Erfahrung höchsten Ranges, um die kein Student betrogen werden darf". Zu dieser geistigen Erfahrung gehört es auch, vertraut zu werden mit der Schärfe mathematischer Begriffsbildungen, der (anfangs oft pedantisch anmutenden) Genauigkeit bei der Formulierung von Definitionen und

Das Instrument, welches die Vermittlung bewirkt zwischen Theorie und Praxis, zwischen Denken und Beobachten, ist die Mathematik; sie baut die verbindende Brücke und gestaltet sie immer tragfähiger. Daher kommt es, daß unsere ganze gegenwärtige Kultur, soweit sie auf der geistigen Durchdringung und Dienstbarmachung der Natur beruht, ihre Grundlagen in der Mathematik findet. Schon Galilei sagt: Die Natur kann nur der verstehen, der ihre Sprache und die Zeichen kennengelernt hat, in der sie zu uns redet; diese Sprache aber ist die Mathematik, und ihre Zeichen sind die mathematischen Figuren. Kant tat den Ausspruch: "Ich behaupte, daß in jeder besonderen Naturwissenschaft nur so viel eigentliche Wissenschaft angetroffen werden kann, als darin Mathematik enthalten ist." In der Tat: Wir beherrschen nicht eher eine naturwissenschaftliche Theorie, als bis wir ihren mathematischen Kern herausgeschält und völlig enthüllt haben. Ohne Mathematik ist die heutige Astronomie und Physik unmöglich; diese Wissenschaften lösen sich in ihren theoretischen Teilen geradezu in Mathematik auf. Diese wie die zahlreichen weiteren Anwendungen sind es, denen die Mathematik ihr Ansehen verdankt, soweit sie solches im weiteren Publikum genießt.

Trotzdem haben es alle Mathematiker abgelehnt, die Anwendungen als Wertmesser für die Mathematik gelten zu lassen. Gauß spricht von dem zauberischen Reiz, den die Zahlentheorie zur Lieblingswissenschaft der ersten Mathematiker gemacht habe, ihres unerschöpflichen Reichtums nicht zu gedenken, woran sie alle anderen Teile der Mathematik so weit übertrifft. Kronecker vergleicht die Zahlentheoretiker mit den Lotophagen, die, wenn sie einmal von dieser Kost etwas zu sich genommen haben, nie mehr davon lassen können. Der große Mathematiker Poincaré wendet sich einmal in auffallender Schärfe gegen Tolstoi, der erklärt hatte, daß die Forderung "die Wissenschaft der Wissenschaft wegen" töricht sei. Die Errungenschaften der Industrie zum Beispiel hätten nie das Licht der Welt erblickt, wenn die Praktiker allein existiert hätten und wenn diese Errungenschaften nicht von uninteressierten Toren gefördert worden wären. Die Ehre des menschlichen Geistes, so sagte der berühmte Königsberger Mathematiker Jacobi, ist der einzige Zweck aller Wissenschaft.

Wir dürfen nicht denen glauben, die heute mit philosophischer Miene und überlegenem Tone den Kulturuntergang prophezeien und sich in dem Ignorabimus gefallen. Für uns gibt es kein Ignorabimus, und meiner Meinung nach auch für die Naturwissenschaft überhaupt nicht. Statt des törichten Ignorabimus heiße im Gegenteil unsere Losung:

Wir müssen wissen. Wir werden wissen.

der Sorgfalt und Strenge bei der Durchführung mathematischer Beweise. Ich habe mich daher beim Schreiben dieses Buches um Lesbarkeit bemüht, aber nicht um den Preis des Verwässerns, des Weglassens von Beweisen oder des Vermeidens "unbequemer" Begriffsbildungen. Ein Mathematikbuch ist anstrengend zu lesen und eignet sich nur in begrenztem Maße als Bettlektüre; dieses Buch ist hier keine Ausnahme. Es liest sich am besten mit Papier und Bleistift in Griffweite, um Rechnungen und Überlegungen aktiv nachzuvollziehen. Die einzelnen Abschnitte können von ihrem logischen Aufbau her in derjenigen Reihenfolge durchgearbeitet werden, in der sie im Buch erscheinen, aber Umstellungen sind auf vielerlei Art möglich; wie in der Praxis verfahren wird, hängt vom jeweiligen Curriculum ab. Was den Aufbau und Inhalt angeht, so will ich kurz einige der Punkte aufführen, die mir beim Schreiben des Buches wichtig waren.

• **Herauspräparieren propädeutischer Kapitel.** Es ist in Mathematikstudiengängen vielfach üblich, einführende Themen (mengentheoretische und aussagenlogische Grundlagen, vollständige Induktion, Zahlbegriff, elementare Kombinatorik usw.) in die Anfängervorlesungen zur Analysis und zur Linearen Algebra zu integrieren, obwohl sie dort thematisch eigentlich gar nicht hingehören. In diesem Buch wurde solches propädeutische Material in separate Kapitel ausgegliedert, die jeweils für sich behandelt werden können. Dies erleichtert auch die Verwendung des Materials in unterschiedlichen Lehrveranstaltungen und Studiengängen.

• **Sorgfältige Grundlegung.** Ich habe viel Wert darauf gelegt, ein durch und durch solides Fundament für spätere mathematische Aktivitäten zu legen. Daher werden auch (vermeintlich) einfache und aus der Schule bekannte Themen wie elementare Zahlentheorie, Bruchrechnung oder Elementargeometrie behandelt. Dabei bietet das Buch weit mehr als nur eine Wiederholung des Schulstoffs: die Darstellung ist mathematisch streng, knüpft Bezüge zu späteren Themen und schält jeweils die zugrundeliegende mathematische Struktur heraus (Ringe und Körper beim Umgang mit Gleichungen, angeordnete Körper beim Umgang mit Ungleichungen, Gruppen beim Umgang mit Symmetrien in kombinatorischen Problemen). Die reellen Zahlen werden in geometrischer Weise eingeführt, wobei der Grenzwertbegriff (der implizit im verwendeten Dedekindschen Schnittaxiom steckt) zunächst vermieden wird; dies erlaubt u.a. eine grenzwertfreie Einführung der Winkelfunktionen. Durch diese sorgfältige Aufbereitung von bereits in der Schule behandelten Themen (wie später dann auch der Differential- und Integralrechnung) ist das Buch auch in Lehramtsstudiengängen einsetzbar.

• **Frühe Einführung abstrakter Begriffe.** Abstrakte Begriffe werden nicht schamhaft vermieden, sondern ganz bewußt und sehr früh explizit gemacht. Ein Beispiel ist etwa der Begriff der Quotientenstruktur, der bereits in der Schule vielfach implizit benutzt wird, ohne klar herausgearbeitet zu werden (Kardinalzahlen als Äquivalenzklassen von Mengen, Brüche als Äquivalenzklassen von Zahlenpaaren, Vektoren als Pfeilklassen). Abstraktion wird in diesem Buch nicht als etwas Unangenehmes und möglichst zu Vermeidendes behandelt, sondern als etwas sehr Wünschenswertes, das begriffliche Klarheit und universelle Einsetzbarkeit mathematischer Methoden überhaupt erst ermöglicht und an das man sich frühzeitig gewöhnen sollte. Insbesondere werden algebraische, ordnungstheoretische und topologische Strukturen früh definiert, und die Existenz solcher Strukturen in verschiedenen Situationen wird systematisch herausgearbeitet.

• **Zahlreiche durchgerechnete Aufgaben und Beispiele.** Das Buch enthält eine Vielzahl komplett durchgerechneter Aufgaben und Beispiele, um die eingeführten Begriffe und Methoden zu verdeutlichen. Ein umfangreicher Aufgabenband zu dem Buch ist in Arbeit.

• **Physikalische Motivation mathematischer Begriffsbildungen.** Viele mathematische Begriffe stammen aus der Physik, und die zugrundeliegende physikalische Intuition ist auch notwendig, um diese Begriffe später zur mathematischen Modellierung realer Systeme heranzuziehen. Dies wird sorgfältig herausgearbeitet, und es wird jeweils klar dargelegt, warum der eingeführte mathematische Begriff tatsächlich das jeweilige physikalische Konzept widerspiegelt und welche physikalische Bedeutung mathematische Sätze haben (Orientierung einer Basis und Dreifingerregel der rechten Hand; materielle, lokale und konvektive Ableitungen als zeitliche Änderungsraten von Feldgrößen; äußere Ableitungen von Differentialformen und deren Zusammenhang mit der Rotation und Divergenz von Vektorfeldern; Stokesscher Integralsatz und Reynoldssches Transporttheorem). Ferner werden zahlreiche Beispiele aus der Mechanik behandelt, wobei wieder viel Wert auf eine sorgfältige Klärung der Begriffe gelegt wird (etwa der Winkelgeschwindigkeit und des Trägheitsmomententensors bei der Bewegung starrer Körper). Das Buch ist daher auch geeignet für die Mathematikausbildung innerhalb eines Studiums der Physik.

• **Weitgehend koordinatenfreies Arbeiten.** Eine Temperaturverteilung ist eine Funktion, die jedem Raumpunkt einen (als reelle Zahl darstellbaren) Temperaturwert zuordnet. Da die Punkte des uns umgebenden Raums nicht von Natur aus (zwecks einfacherer Identifizierbarkeit durch potentielle Beobachter) mit Zahlentripeln versehen sind, ist eine solche Funktion etwas grundsätzlich anderes als eine Funktion $\mathbb{R}^3 \to \mathbb{R}$. Ausgehend von dieser (physikalisch motivierten) Sichtweise werden Begriffe möglichst koordinatenfrei eingeführt; in der Linearen Algebra etwa sind lineare Abbildungen und quadratische Formen fundamentale Begriffe, nicht die Matrizen, durch die diese repräsentiert werden können. Eine koordinatenfreie Einführung mathematischer Begriffe ist nicht nur vorteilhaft für das begriffliche Verständnis, sondern auch bei der Lösung ganz konkreter Aufgaben.

• **Schlüsselrolle der Linearen Algebra.** Der Linearen Algebra kommt eine fundamentale Rolle zu. Zunächst ist bereits die Entwicklung dieser mathematischen Disziplin aus zwei verschiedenen Wurzeln (systematische Untersuchung linearer Gleichungssysteme einerseits, elementargeometrische Vektorrechnung andererseits) Ausdruck einer arithmetisch-geometrischen Synthese, die auch für andere Bereiche grundlegend ist (man denke etwa an kommutative Algebra und algebraische Geometrie); ich habe mich daher bemüht, sowohl arithmetische als auch geometrische Aspekte der Linearen Algebra zu berücksichtigen. Ferner beruht die gesamte Analysis darauf, nichtlineare Begriffe und Methoden mittels Linearisierung auf Lineare Algebra zurückzuführen; dies wird im Rahmen dieses Buches herausgearbeitet (Ableitungen als Linearisierungen von Funktionen an gegebenen Stellen, höhere Ableitungen als multilineare Abbildungen, Mannigfaltigkeiten als deformierte lineare Räume und so weiter). Schließlich lassen sich weite Teile der Funktionalanalysis als Erweiterung der Linearen Algebra auf (geeignet topologisierte) unendlichdimensionale Vektorräume verstehen. Um dies vorzubereiten, werden von vornherein auch unendlichdimensionale Vektorräume betrachtet und frühzeitig Skalarprodukte und Normen auf Vektorräumen behandelt, was die Behandlung einiger Sätze der Funktionalanalysis erlaubt (ohne daß in diesem Buch systematisch Funktionalanalysis betrieben würde).

• **Berücksichtigung numerischer Aspekte.** Bereits bei der Einführung des Grenzwertbegriffs wird dessen genuin numerischer Charakter betont (Approximation einer Größe durch ein Glied einer gegen diese Größe konvergierenden Folge, Wichtigkeit der zugehörigen Fehlerabschätzung, Bedeutung der Konvergenzgeschwindigkeit) und an Beispielen aufgezeigt (Babylonisches Wurzelziehen, Berechnung der Zahlen e und π, numerische Berechnung von Logarithmen). Numerische Aspekte und Methoden sind durchgängig in den Text integriert (Satz von Gerschgorin, Vektor- und Matrixnormen, Fehlerabschätzung bei linearen Gleichungssystemen, Bestapproximation in Skalarprodukträumen, lineare Ausgleichsrechnung, Polynominterpolation, numerische Integration, Newtonverfahren in einer und in mehreren Variablen, Satz von Kantorovitch, Eulerpolygone zur Approximation der Lösung eines Anfangswertproblems und so weiter). Zahlreiche weitere numerische Verfahren werden im Aufgabenband behandelt, insbesondere solche, bei denen auf die Herleitung strenger Konvergenznachweise und Fehlerabschätzungen verzichtet wird.

• **Behandlung der Differentialrechnung vor der Integralrechnung.** Es wird zunächst die Differentialrechnung sowohl in einer als auch in mehreren Variablen (und sogar allgemein in Banachräumen) behandelt, bevor die Integralrechnung entwickelt wird. Dies hat den Vorteil, daß bei der Behandlung der Integrationstheorie bereits Vertrautheit mit Funktionen in mehreren Veränderlichen besteht und eine gewisse mathematische Reife erreicht ist, die die simultane Einführung des Riemannschen und des Lebesgueschen Integralbegriffs erlaubt. Strukturelle Eigenschaften des Integrals können durch diese Vorgehensweise gleich für Funktionen in mehreren Variablen hergeleitet werden. Diese Umstellung wirkt sich kaum auf die übliche Entwicklung der Differentialrechnung aus; lediglich der Beweis der Mittelwertabschätzung ist zu modifizieren (und natürlich muß die Integraldarstellung des Taylor-Restglieds nach hinten gezogen werden). Selbstverständlich kann wahlweise vor dem Studium von Funktionen in mehreren Variablen auch zunächst die Differential- und Integralrechnung in einer Variablen vollständig behandelt werden.

• **Ausführliche Diskussion von Differentialgleichungen.** Trotz seiner Kompaktheit enthält das Kapitel über gewöhnliche Differentialgleichungen mehr Material als in Einführungen üblich: grundlegende Begriffe und elementare Lösungsmethoden, einige motivierende Beispiele, den Existenzsatz von Peano, die Eindeutigkeitssätze von Cauchy und Osgood, Aussagen zum maximalen Lösungsintervall eines Anfangswertproblems, stetige und glatte Abhängigkeit der Lösung von Anfangsbedingungen und Parametern sowie spezielle Lösungsmethoden für lineare Differentialgleichungen (insbesondere mit konstanten oder periodischen Koeffizienten). Anwendungen aus der Physik (Himmelsmechanik, Starrkörperbewegung) sowie ein Kapitel über dynamische Systeme runden die Darstellung ab.

Was den Stil des Buches angeht, so habe ich mich einerseits um Ausführlichkeit bei der Motivation von Begriffsbildungen bemüht (und dabei auch einige außermathematische Abschweifungen nicht gescheut), andererseits aber um einen kompakten Stil bei Beweisen und bei der Entwicklung der behandelten Theorien. Das mag angesichts des Umfangs, den dieses Buch schließlich angenommen hat, nicht sehr glaubwürdig erscheinen, aber ein Buch, das mit der Einführung des Mengenbegriffs beginnt und mit dem Beweis des Riemannschen Abbildungssatzes endet und in dem alle Aussagen ausnahmslos bewiesen werden, erreicht zwangsläufig einen gewissen Umfang – auch ohne Weitschweifigkeit des Autors. Trotz seines Umfangs kann und will dieses Buch nicht für sich reklamieren, einen Gesamtüberblick über die Mathematik zu bieten. Dazu wären Numerische Mathematik und Optimierung systematischer abzuhandeln gewesen, ebenso die Funktionalanalysis und die Theorie dynamischer Systeme (wo etwa Verzweigungsphänomene oder gesteuerte dynamische Systeme gar nicht vorkommen). Einige wichtige Gebiete fehlen völlig, etwa Algebra (Gruppen-, Ring- und Körpertheorie; kommutative Algebra und algebraische Geometrie), Variationsrechnung oder partielle Differentialgleichungen. Zu einer adäquaten Behandlung all dieser Themen hätte ich einen zweiten Band gleichen Umfangs schreiben müssen, und irgendwo war eine Grenze zu ziehen – selbst ein hartgesottener Autor wird nervös,

wenn sein Manuskript (zumal im DIN A 4-Format) einer vierstelligen Seitenzahl zustrebt. Sollten allerdings manche Auslassungen als besonders störend empfunden werden, so bin ich für entsprechende Kommentare jederzeit dankbar.

Rechtschreibung. Ich betrachte die Rechtschreibreform der Jahre 1996 bis 2006 in ihren inhaltlichen Festlegungen und der Art ihrer politischen Durchsetzung als ein Symptom sprachlichen und kulturellen Niedergangs. Das Buch orientiert sich daher an den vor dieser Reform gültigen Regeln; die Lesbarkeit ist dadurch in keiner Weise gefährdet. Dem Verlag danke ich für die Bereitschaft, meinem ausdrücklichen Wunsch nach Verwendung der alten Rechtschreibung nachzukommen.

Typographische Konventionen. Sätze und Definitionen sind *kursiv* gesetzt, um sie vom normalen Text abzuheben. Das Ende eines Beweises ist jeweils mit einem Quadrat ■ markiert, das Ende eines Beispiels oder einer Gruppe von Beispielen mit einer Raute ♦. Bemerkungen werden typischerweise mit einer Raute beendet, aber dann mit einem Quadrat, wenn die Bemerkung den Charakter eines Beweises hat. Die einzelnen Abschnitte sind durchlaufend numeriert, die Bestandteile (Sätze, Definitionen, Beispiele usw.) innerhalb eines Abschnitts ebenfalls; beispielsweise bezeichnet die Nummer (103.7) den Bestandteil 7 des Abschnitts 103. Jeweils vier Abschnitte wurden zu einem Kapitel zusammengefaßt, aber auf eine Numerierung der einzelnen Kapitel wurde verzichtet.

Danksagungen. Dieses Vorwort wäre unvollständig ohne eine Bezeugung tiefen Dankes an diejenigen, die mich beim Schreiben des vorliegenden Buches unterstützten. In erster Linie ist hier Frau Dr. Renate Schappel zu nennen, die sich mit bewundernswerter Energie und Sorgfalt der Herkulesaufgabe annahm, das vollständige Manuskript (teilweise in verschiedenen Versionen) kritisch durchzulesen. Sie deckte eine Unzahl von Fehlern auf, und nichts war vor ihrem kritischen Blick sicher: einfache Tippfehler, Rechenfehler in Beispielen, fehlerhafte oder unvollständige Schlüsse in mathematischen Herleitungen, stilistisch verunglückte Formulierungen, falsche Verweise, selbst Fehler in den Geburts- und Todesjahren von Mathematikern, die im Text genannt werden. Ohne ihre Hilfe wäre dieses Buch schlechterdings nicht denkbar. Mein Dank gilt ferner Frau Prof. Dr. Evgenia Kirillova*, die Teile des Manuskripts las und ebenfalls mit ihren Korrekturen, Kommentaren und Änderungsvorschlägen zur Verbesserung der Darstellung beitrug. Weiterhin danke ich Herrn cand. rer. nat. Claus Meister, der mir immer dann hilfreich zur Seite stand, wenn eher esoterische Befehle des Textverarbeitungssystems benötigt wurden oder wenn es Probleme bei der Erzeugung und Einbindung von Graphiken gab. Bei Herrn Klaus Horn vom Verlag Harri Deutsch bedanke ich mich ganz herzlich für die kompetente verlagsseitige Umsetzung des Werkes und die jederzeit angenehme Zusammenarbeit.

Schließlich gilt mein ganz besonderer Dank meiner Frau und meinen beiden Kindern, die am meisten unter der Entstehung dieses Buches zu leiden hatten – durch einen oft zwar körperlich, aber nicht geistig anwesenden Ehemann und Vater, der zuweilen auch mißmutig wurde, wenn es mit dem Schreiben nicht recht vorwärts ging. Ihnen ist dieses Buch gewidmet.

Schlußbemerkung. Von Richard Feynman (1918-1988), der im Jahr 1965 den Nobelpreis für Physik erhielt, stammt die folgende Bemerkung:

> *There are two kinds of mathematics books; the kind you can't read past the first sentence, and the kind you can't read past the first page.*

Autor und Verlag hegen die Hoffnung, daß diese Bemerkung nicht auf das vorliegende Buch zutrifft (eine nicht ganz unbegründete Hoffnung, wenn jemand bis hierher gelesen haben sollte), und sind für Kommentare, Korrektur-, Verbesserungs- und Ergänzungsvorschläge sowie Wünsche für den in Arbeit befindlichen Aufgabenband jederzeit dankbar.

Wiesbaden, A. D. 2010 Karlheinz Spindler

* Евгения Вадимовна Кириллова; an dieser Stelle erweist sich die Verfügbarkeit kyrillischer Schriftzeichen in meinem Textverarbeitungssystem als unwiderstehliche Versuchung.

INHALT

Verlag Harri Deutsch – Karlheinz Spindler: *Höhere Mathematik – Ein Begleiter durch das Studium* – (978-3-8171-1872-4)

Rechnen mit Grenzwerten

Elementare Funktionen

Metrische Strukturen

Topologische Strukturen

Differentialrechnung in einer Variablen

Differentialrechnung in Banachräumen

Differentialrechnung auf Mannigfaltigkeiten

Inhaltsbestimmung von Mengen

Der Begriff des Integrals

Berechnung von Integralen

Integration auf Mannigfaltigkeiten

Gewöhnliche Differentialgleichungen

Dynamische Systeme

Integraltransformationen

Grundlagen der Stochastik

Anwendung stochastischer Methoden

Funktionentheorie

Die folgende Aufgabe wurde im Jahr 1779 erstmals formuliert und gelöst, und zwar von dem italienischen Mathematiker Gianfrancesco di Fagnano (1715-1797), der sich, ebenso wie sein Vater, Giulio de Toschi di Fagnano (1682-1766), vornehmlich mit Problemen der Geometrie und Analysis beschäftigte.

(38.16) Problem von Fagnano. *Ein Dreieck mit den Ecken A, B und C sei gegeben. Gesucht ist ein möglichst kurzer geschlossener Streckenzug, der die drei Seiten des Dreiecks miteinander verbindet. Wie läßt sich ein solcher Streckenzug finden?*

Lösung. Gesucht sind Punkte A' auf $[B,C]$, B' auf $[C,A]$ und C' auf $[A,B]$ derart, daß

$$(\star) \qquad \overline{A'B'} + \overline{B'C'} + \overline{C'A'}$$

möglichst klein wird. Wir denken uns zunächst den Punkt C' auf $[A,B]$ fest gewählt und überlegen, wie die Punkte A' und B' gewählt werden müssen, damit der Ausdruck $(\star)$ minimal wird. Dazu führen wir zunächst noch die beiden Punkte C_1 und C_2 ein, die aus C' durch Spiegelung an den Geraden $\overline{CA}$ und $\overline{CB}$ entstehen.

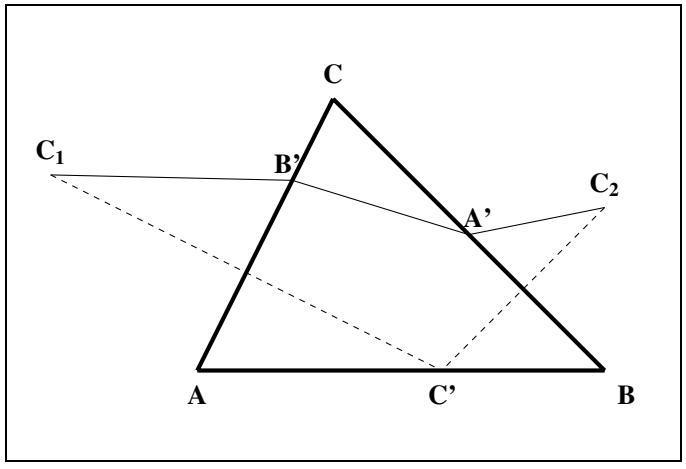

Abb. 38.16: Konstruktion zur Lösung des Problems von Fagnano.

Da ein Liniensegment bei einer Spiegelung seine Länge nicht ändert, läßt sich für beliebige Punkte $A' \in [B,C]$ und $B' \in [A,C]$ der Ausdruck $(\star)$ in der Form

$$\begin{aligned} &\overline{A'B'} + \overline{B'C'} + \overline{C'A'} \\ (\star\star) \qquad &= \overline{A'B'} + \overline{B'C_1} + \overline{C_2A'} \\ &= \overline{C_1B'} + \overline{B'A'} + \overline{A'C_2} \end{aligned}$$

schreiben, stimmt also mit der Länge des Polygonzugs $C_1B'A'C_2$ überein. Dieser hat die minimale Länge $\overline{C_1C_2}$, und diese minimale Länge wird genau dann angenommen, wenn die vier Punkte C_1, B', A' und C_2 auf einer Geraden liegen, wenn also B' und A' die Schnittpunkte der Geraden $\overline{C_1C_2}$ mit den Seiten AC bzw. BC sind, wenn also mit den Bezeichnungen der folgenden Skizze die Beziehungen $B' = B^\star$ und $A' = A^\star$ gelten. Für diese Wahl von A' und B' geht dann $(\star)$ gemäß $(\star\star)$ über in $\overline{C_1B^\star} + \overline{B^\star A^\star} + \overline{A^\star C_2} = \overline{C_1C_2}$.

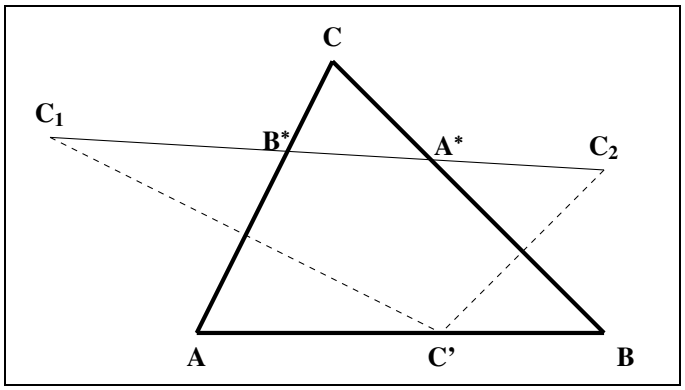

Abb. 38.17: Optimale Wahl von A' und B' bei gegebenem Punkt C'.

Zur endgültigen Lösung der Minimierungsaufgabe stellt sich nun noch die Frage, wie der Punkt C' zu wählen ist, damit die von diesem induzierte Strecke $\overline{C_1C_2}$ minimal wird. Da die Winkel C_1CA und ACC' bzw. $C'CB$ und BCC_2 jeweils durch Spiegelung auseinander hervorgehen und daher gleich sind, gilt $\angle(C_1CC_2) = 2\cdot\angle(ACB) = 2\gamma$; der Winkel C_1CC_2 ist also völlig unabhängig von der Wahl des Punktes C'. Je kürzer nun $\overline{CC'} = \overline{CC_1} = \overline{CC_2}$ ist, desto kürzer ist auch $\overline{C_1C_2}$.

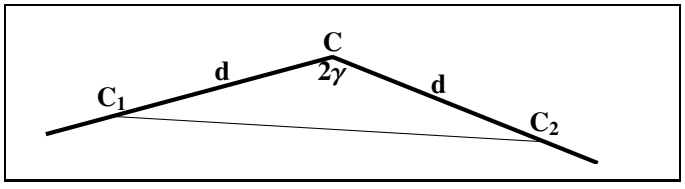

Abb. 38.18: Je kürzer $d := \overline{CC'}$, desto kürzer auch $\overline{C_1C_2}$.

Die kürzestmögliche Strecke $\overline{CC'}$ ergibt sich nun aber, wenn wir für C' den Fußpunkt $C^\star$ des Lotes von C auf die Gerade AB wählen. Damit ist die Lösung des Problems gefunden: Fälle von jeder der drei Ecken A, B, C das Lot auf die jeweils gegenüberliegende Dreiecksseite und bezeichne die entstehenden Lotfußpunkte mit $A^\star$, $B^\star$, $C^\star$; der gesuchte Verbindungsweg kürzester Länge ist dann der Streckenzug $A^\star B^\star C^\star A^\star$. ■

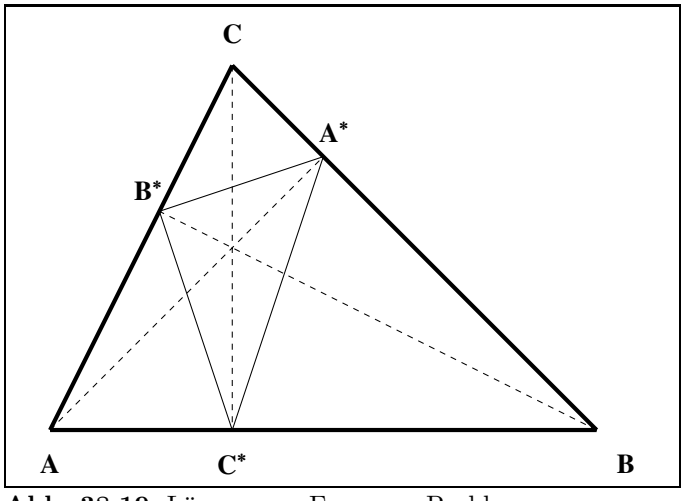

Abb. 38.19: Lösung von Fagnanos Problem.

(38.17) Bemerkung. Wer den Beweis aufmerksam verfolgt hat, wird eine Lücke festgestellt haben; die Argumentation ist nämlich nur für spitzwinklige Dreiecke uneingeschränkt gültig. Hat das Dreieck bei A einen rechten Winkel, so fallen die Punkte $B^\star$ und $C^\star$ mit A zusammen; hat das Dreieck bei A einen stumpfen Winkel, so liegen $B^\star$ und $C^\star$ außerhalb des Dreiecks. In diesen beiden Fällen ist die Lösung des Problems gegeben durch den Streckenzug $AA^\star A$ (Übungsaufgabe!); die Lösung ist also jeweils ein zu einer doppelt durchlaufenen Strecke entartetes Dreieck. ■

Das folgende Problem, das zuerst von den Mathematikern Bonaventura Cavalieri (1598-1647), Pierre de Fermat (1601 oder 1607/1608-1665) und Evangelista Torricelli (1608-1647) studiert wurde, besteht darin, innerhalb eines Dreiecks einen Punkt zu finden, der von den drei Ecken den kürzesten durchschnittlichen Abstand hat. Eine praktische Einkleidung der Aufgabe könnte etwa folgendermaßen lauten: Ein Gelände hat bei A eine Wasserstelle, bei B eine Feuerstelle und bei C einen Vorratsraum. An welcher Stelle P sollte man sein Zelt aufschlagen, wenn man gleich oft zu A, B und C gehen muß und den zurückzulegenden Gesamtweg möglichst kurz halten will? Genauer formuliert lautet die Aufgabe folgendermaßen.

(38.18) Problem von Viviani, Fermat und Torricelli. *Ein Dreieck mit den Ecken A, B und C sei gegeben. Welcher Punkt P des Dreiecks minimiert den Ausdruck $\overline{PA}+\overline{PB}+\overline{PC}$?*

Lösung. Wir greifen uns zunächst einen beliebigen Punkt P in dem Dreieck heraus und drehen dann das Dreieck APC um $\pi/3$ um den Punkt A; das durch die Drehung entstandene Dreieck bezeichnen wir mit AQB'.

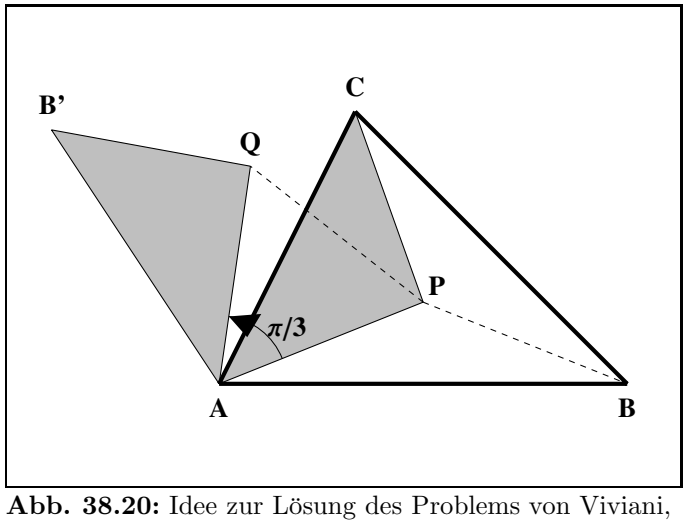

Abb. 38.20: Idee zur Lösung des Problems von Viviani, Fermat und Torricelli.

Wir stellen zunächst fest, daß der Punkt B' vollkommen unabhängig von der Wahl von P ist; das Liniensegment AB' ergibt sich ja durch eine Drehung des Liniensegments AC um $\pi/3$ um den Punkt A. Wegen $\overline{AC}=\overline{AB'}$ ist das Dreieck CAB' gleichschenklig; also stimmen die Basiswinkel $AB'C$ und $B'CA$ überein. Da der Winkel CAB' nach Konstruktion gerade $\pi/3$ ist und die Winkelsumme im Dreieck CAB' gleich π sein muß, ist jeder dieser Basiswinkel ebenfalls gleich $\pi/3$; das Dreieck $AB'C$ ist also sogar gleichseitig. (Der Punkt B' läßt sich folglich konstruieren, indem wir ein gleichseitiges Dreieck über dem Liniensegment AC errichten.)

Nach Konstruktion gilt $AP=AQ$; das Dreieck PAQ ist damit gleichschenklig, hat also bei P und Q gleiche Basiswinkel. Da der Winkel bei A aber nach Konstruktion $\pi/3$ beträgt und die Winkelsumme im Dreieck PAQ gleich π sein muß, ist jeder dieser Basiswinkel ebenfalls gleich $\pi/3$; das Dreieck PAQ ist also sogar gleichseitig, so daß $\overline{PA}=\overline{PQ}$ gilt. Der zu minimierende Ausdruck ist dann

$$\begin{aligned}&\overline{PA}+\overline{PB}+\overline{PC}\\ &=\overline{PQ}+\overline{PB}+\overline{QB'}\\ &=\overline{B'Q}+\overline{QP}+\overline{PB},\end{aligned}$$

stimmt also mit der Länge des Linienzuges $B'QPB$ überein. Die Länge dieses Linienzuges ist aber genau dann minimal, wenn die Punkte B', Q, P und B auf einer Geraden liegen, wenn die Winkel $B'QP$ und QPB also beide gleich π sind; dies ist genau dann der Fall, wenn die Winkel $B'QA$ und APB beide gleich $2\pi/3$ sind.

Ein Punkt P erfüllt also sicher dann die gewünschte Minimalbedingung, wenn P und Q auf der Geraden BB' liegen. Gibt es einen solchen Punkt? Die Antwort ist ja, denn wir können den Winkel $\varphi:=\angle AB'B$ von AC aus abtragen, erhalten einen Schnittpunkt P mit der Geraden BB' und können dann die Strecke CP von B' aus abtragen, um Q zu erhalten. (Es gilt $\varphi=180^0-(\alpha+60^0)-\beta=120^0-\alpha-\beta<\gamma$.)

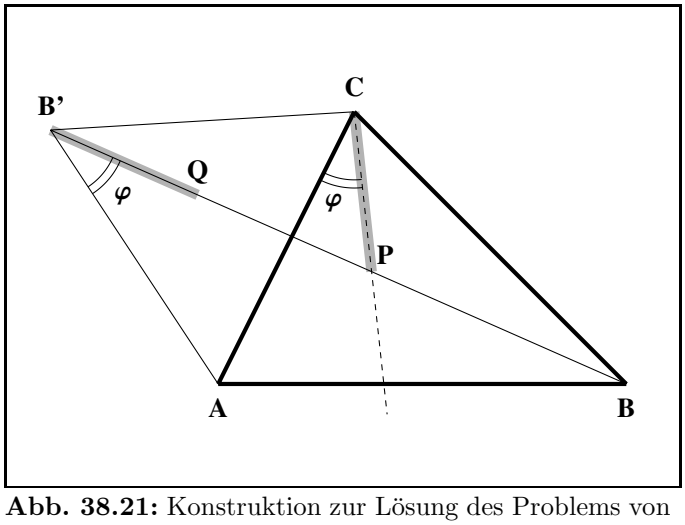

Abb. 38.21: Konstruktion zur Lösung des Problems von Viviani, Fermat und Torricelli.

Daß wir unsere Konstruktion von der Ecke A aus durchführten, war willkürlich; wir hätten genausogut von B oder C aus beginnen können. Dies liefert die folgende Konstruktion des gesuchten Punktes P: Errichte über jeder der

drei Seiten des gegebenen Dreiecks ein gleichseitiges Dreieck und verbinde die dritte Ecke dieses Dreiecks mit der der Seite gegenüberliegenden. Dann schneiden sich die Geraden AA', BB' und CC' in einem Punkt P; dieser Punkt (den man auch den **Torricelli-Punkt** des Dreiecks ABC nennt) ist die eindeutige Lösung der gestellten Optimierungsaufgabe. ■

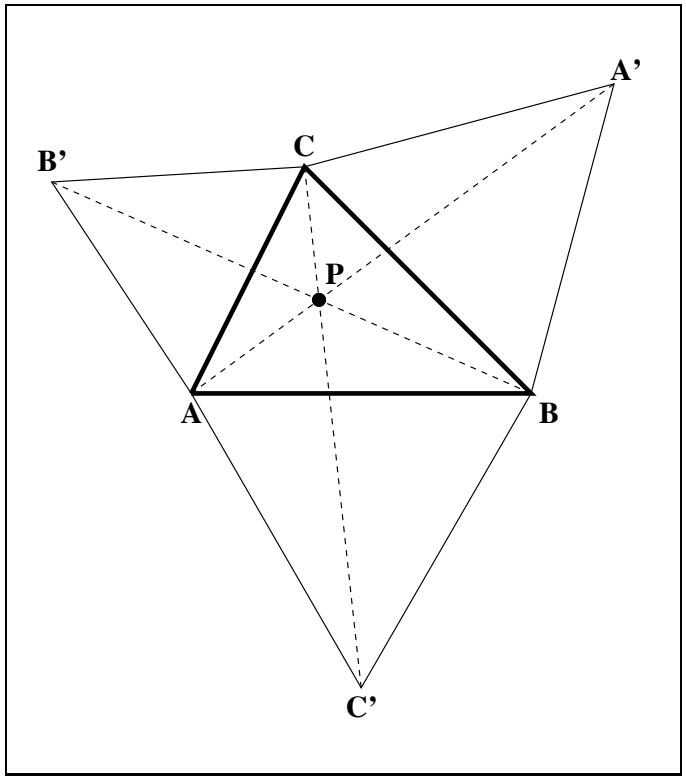

Abb. 38.22: Konstruktion des Torricelli-Punktes eines Dreiecks.

(38.19) Bemerkungen. (a) Die gegebene Lösung zeigt, daß die Winkel $\angle APB$, $\angle BPC$ und $\angle CPA$ jeweils den Wert $2\pi/3 = 120^0$ haben.

(b) Die angegebene Lösung gilt nur, wenn alle Winkel des betrachteten Dreiecks kleiner als 120^0 sind. Hat das Dreieck bei A einen Winkel von 120^0, so fällt der Torricelli-Punkt des Dreiecks mit A zusammen; ist der Winkel bei A sogar größer als 120^0, so liegt der Torricelli-Punkt außerhalb des Dreiecks und liefert nicht die Lösung der gestellten Optimierungsaufgabe. Man kann zeigen (Übungsaufgabe!), daß in diesem Fall die optimale Wahl des Punktes P gegeben ist durch $P := A$.

(c) Das Problem von Viviani, Fermat und Torricelli hat eine interessante mechanische Lösung. Wir deuten A, B und C als Punkte auf einer glatten Tischoberfläche und bohren an diesen Stellen Löcher durch die Fläche. Wir befestigen nun drei gleiche Gewichte an drei (als masselos betrachteten) Schnüren, führen die Schnurenden von unten her durch die gebohrten Löcher, verknoten die Schnüre oberhalb des Tisches und lassen los. Es wird sich ein Gleichgewichtszustand einstellen, bei dem der Schwerpunkt des Systems möglichst tief zu liegen kommt, bei dem also die Gesamtlänge der sich unterhalb des Tischs befindlichen Schnurstücke möglichst groß und damit die Gesamtlänge der sich auf dem Tisch befindlichen Schnurstücke möglichst klein ist. Der Stelle, an der der Knoten sich dann befindet, ist genau der Punkt P, für den die Gesamtlänge $\overline{PA} + \overline{PB} + \overline{PC}$ minimal wird. ■

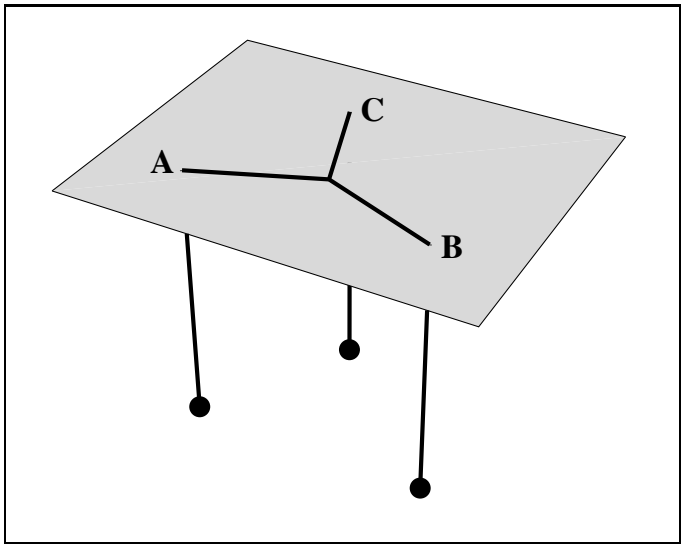

Abb. 38.23: Mechanische Lösung des Problems von Viviani, Fermat und Torricelli.

39. Kreise

Wir definieren Kreise bzw. Sphären als diejenigen geometrischen Örter, die von einem festen Punkt den gleichen Abstand haben.

(39.1) Definition. *Gegeben seien eine Ebene E, ein Punkt M in E und eine Strecke r. Der* **Kreis** *mit Mittelpunkt M und Radius r in E ist die Menge aller Punkte P in E mit der Eigenschaft $\overline{MP} = r$. Die* **abgeschlossene Kreisscheibe** *mit Mittelpunkt M und Radius r in E ist die Menge aller Punkte P in E mit der Eigenschaft $\overline{MP} \leq r$. Die* **offene Kreisscheibe** *mit Mittelpunkt M und Radius r in E ist die Menge aller Punkte P in E mit der Eigenschaft $\overline{MP} < r$.*

(39.2) Definition. *Gegeben seien ein Punkt M im Raum und eine Strecke r. Die* **Sphäre** *mit Mittelpunkt M und Radius r ist die Menge aller Punkte P mit der Eigenschaft, daß die Strecke $\overline{MP}$ mit r übereinstimmt. Die* **abgeschlossene Kugel** *mit Mittelpunkt M und Radius r ist die Menge aller Punkte P mit $\overline{MP} \leq r$. Die* **offene Kugel** *mit Mittelpunkt M und Radius r ist die Menge aller Punkte P mit $\overline{MP} < r$.*

Es ist klar, daß für eine vorgegebene Ebene E und einen Punkt M in E der Kreis bzw. die Kreisscheibe mit Mittelpunkt M und Radius r in E gerade der Durchschnitt von E mit der Sphäre bzw. Kugel mit Mittelpunkt M und Radius r ist. Die Wichtigkeit von Kreisen liegt

und damit $\operatorname{Rad}\beta = V_0 \cap \operatorname{Rad}\beta$, also $\operatorname{Rad}\beta \subseteq V_0$. Insgesamt gilt $V_0 = \operatorname{Rad}\beta$. Weiter sei U ein Unterraum mit $V_+ \subsetneqq U$, auf dem β positiv definit ist. Dann gilt $V = U + V_- + V_0$ und nach der Dimensionsformel für Unterräume daher

$$\begin{aligned}\dim\bigl(U \cap (V_- \oplus V_0)\bigr) &= \dim U + \dim(V_- \oplus V_0) - \dim V \\ &= \dim U - \dim V_+ \;> 0\end{aligned}$$

und damit $U \cap (V_- \oplus V_0) \neq \{0\}$. Das ist aber unmöglich, weil β positiv definit auf U und negativ semidefinit auf $V_- \oplus V_0$ ist. Also ist V_+ ein maximaler Unterraum, auf dem β positiv definit ist. Analog ist V_- ein maximaler Unterraum, auf dem β negativ definit ist.

(b) Es sei V_+ ein maximaler Unterraum, auf dem β positiv definit ist; wir setzen $U := (V_+)^\perp$. Nach (58.10) gilt dann $V = V_+ \oplus U$, und offensichtlich gilt $\operatorname{Rad}\beta \subseteq U$. Ferner ist β negativ semidefinit auf U; gäbe es nämlich ein Element $u \in U$ mit $\beta(u,u) > 0$, so wäre β positiv definit auf $V_+ \oplus \mathbb{R}u$, was der Maximalität von V_+ widerspräche. Anwendung der Cauchy-Schwarzschen Ungleichung auf $-\beta\,|_{U\times U}$ zeigt, daß $\beta(u,v)^2 \leq \beta(u,u)\beta(v,v)$ für alle $u,v \in U$ gilt. Gälte also $\beta(u,u) = 0$ für ein Element $u \in U$, dann auch $\beta(u,v) = 0$ für alle $v \in V$ und damit wegen $V = U^\perp \oplus U$ sogar $\beta(u,v) = 0$ für alle $v \in V$; ist also $\beta(u,u) = 0$ für ein $u \in U$, so gilt $u \in \operatorname{Rad}\beta$. Ist also V_- irgendein Vektorraumkomplement von $\operatorname{Rad}\beta$ in U, so ist β negativ definit auf V_-; damit ist dann $V = V_+ \oplus V_- \oplus \operatorname{Rad}\beta$ eine Standardzerlegung von (V,β). Vollkommen analog zeigt man, daß jeder maximale Unterraum, auf dem β negativ definit ist, als V_--Anteil einer Standardzerlegung von V auftritt.

(c) Da β positiv definit auf V_+ und negativ semidefinit auf $\overline{V}_- \oplus \operatorname{Rad}\beta$ ist, haben wir $V_+ \cap \bigl(\overline{V}_- \oplus \operatorname{Rad}\beta\bigr) = \{0\}$ und damit

$$\begin{aligned}\dim V_+ &= \dim(V_+ + \overline{V}_- + \operatorname{Rad}\beta) - \dim(V_- \oplus \operatorname{Rad}\beta) \\ &\leq \dim V - \dim(\overline{V}_- \oplus \operatorname{Rad}\beta) \;=\; \dim \overline{V}_+ .\end{aligned}$$

Vertauschen wir die Rollen der beiden Standardzerlegungen, so erkennen wir, daß auch $\dim\overline{V}_+ \leq \dim V_+$ gilt, insgesamt also $\dim\overline{V}_+ = \dim V_+$. Vollkommen analog sehen wir die Gleichung $\dim\overline{V}_- = \dim V_-$ ein.

(d) Dies folgt unmittelbar aus (b) und (c). ∎

Wir haben jetzt alles beisammen, um einen vollständigen Klassifikationssatz für reelle symmetrische Bilinearformen aufzustellen.

(58.17) Satz. *Es sei V ein endlichdimensionaler reeller Vektorraum.*

(a) *Jede symmetrische Bilinearform $\beta : V \times V \to \mathbb{R}$ läßt sich durch eine Matrix der Form*

$$\begin{bmatrix} \mathbf{1}_p & 0 & 0 \\ 0 & -\mathbf{1}_q & 0 \\ 0 & 0 & \mathbf{0}_s \end{bmatrix}$$

darstellen. Die Zahlen p,q,s sind dabei eindeutig bestimmt. (Man bezeichnet das Tripel (p,q,s) als den **Index** *und die Differenz $p-q$ als die* **Signatur** *von β.)*

(b) *Zwei symmetrische Bilinearformen auf V sind genau dann äquivalent, wenn sie den gleichen Index besitzen.*

(c) *Zwei symmetrische Bilinearformen auf V sind genau dann äquivalent, wenn sie den gleichen Rang und die gleiche Signatur besitzen.*

Beweis. Die Existenz der Darstellung wurde bereits in (58.12)(c) bewiesen, die Eindeutigkeit folgt aus (58.16). Der Rest der Behauptung ist trivial. ∎

59. Volumenfunktionen

Sind $a,b,c \in \mathfrak{V}$ Vektoren im Vektorraum $\mathfrak{V}$ aller Pfeilklassen, so bezeichnen wir die Punktmenge $\mathfrak{S}(a,b,c) := \{ra + sb + tc \mid 0 \leq r,s,t \leq 1\}$ als den von a, b und c aufgespannten Spat (oder, vornehmer, als das von a, b und c aufgespannte Parallelepiped).

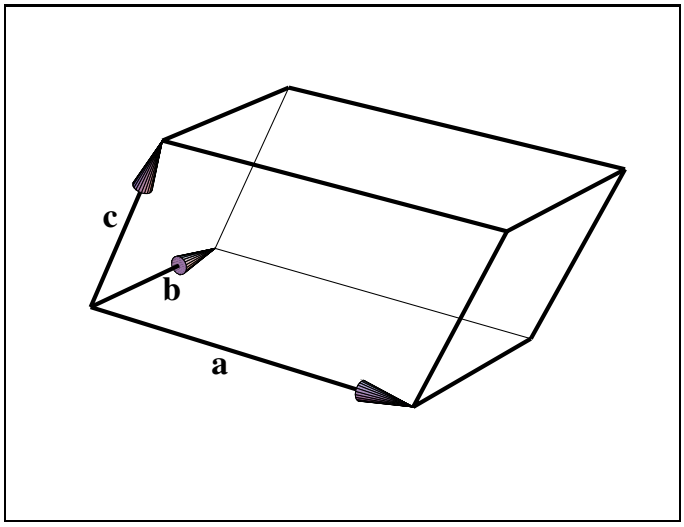

Abb. 59.1: Spat in $\mathfrak{V}$.

Für solche Spate wollen wir nun den Begriff eines orientierten (also vorzeichenbehafteten) Volumens herleiten. Tun wir einmal so, als wüßten wir, was das Volumen $V(a,b,c)$ des Spates $\mathfrak{S}(a,b,c)$ ist.† Sicher ist genau dann $V(a,b,c) = 0$, wenn a, b und c linear abhängig sind, wenn also der Spat $\mathfrak{S}(a,b,c)$ zu einem Parallelogramm, einem Liniensegment oder einem Punkt entartet ist; nehmen wir also an, a, b und c seien linear unabhängig. Wir können dann die Fälle unterscheiden, daß a, b und c (in dieser Reihenfolge!) wie Daumen, Zeigefinger und Mittelfinger

† Etwa aufgrund physikalischer Argumentation: wir legen willkürlich eine Volumeneinheit fest, beispielsweise den Inhalt der Kaffeetasse, die ich beim Schreiben dieser Zeilen vor mir stehen habe, und fragen, wie oft sich diese Volumeneinheit in (eine physikalische Realisierung von) $\mathfrak{S}(a,b,c)$ einfüllen läßt, bevor der Kaffee überläuft.

unserer rechten oder unserer linken Hand zeigen, und nennen das Tripel (a,b,c) dementsprechend ein Rechtssystem oder ein Linkssystem. Als orientiertes Volumen des Spates $\mathfrak{S}(a,b,c)$ bezeichnen wir dann die reelle Zahl $\mathrm{vol}(a,b,c) :=$

$$\begin{cases} V(a,b,c), & \text{falls } (a,b,c) \text{ ein Rechtssystem ist,} \\ 0, & \text{falls } a,\ b \text{ und } c \text{ linear abhängig sind,} \\ -V(a,b,c), & \text{falls } (a,b,c) \text{ ein Linkssystem ist.} \end{cases}$$

Wir beachten, daß $\mathrm{vol}(a,b,c)$ eine Funktion der Vektoren a, b und c ist, nicht nur eine Funktion der Punktmenge $\mathfrak{S}(a,b,c)$. Dieses vorzeichenbehaftete Volumen ist nun algebraisch wesentlich leichter zu handhaben als das gewöhnliche Volumen, denn es hat die charakteristische Eigenschaft, linear von jedem seiner Argumente abzuhängen: die Additivität in jeder Komponente drückt dabei die Volumeninvarianz unter Scherungen aus, die Homogenität den geometrischen Sachverhalt, daß sich das Volumen bei Streckung einer der Seiten um den Faktor λ ver$|\lambda|$facht, wobei für $\lambda < 0$ eine Umkehrung der Orientierung hinzukommt.

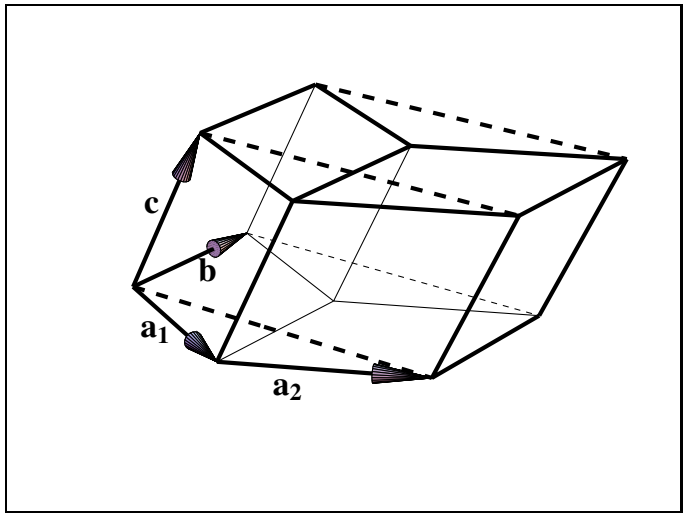

Abb. 59.2: Additivität des orientierten Volumens in jedem Argument: $\mathrm{vol}(a_1{+}a_2,b,c) = \mathrm{vol}(a_1,b,c){+}\mathrm{vol}(a_2,b,c)$.

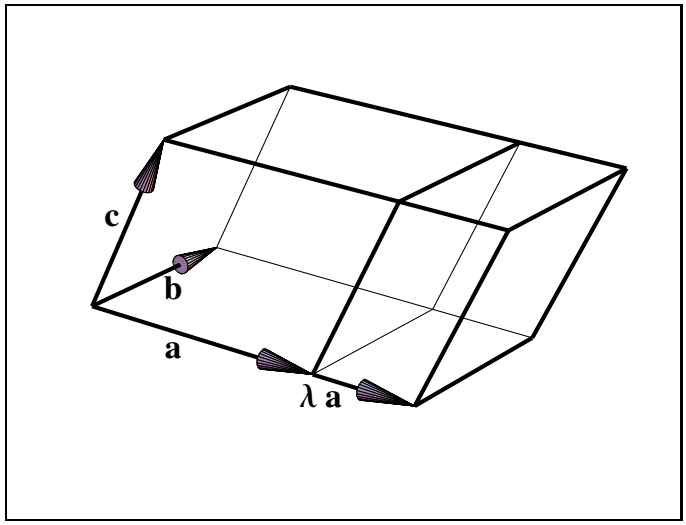

Abb. 59.3: Homogenität des orientierten Volumens in jedem Argument: $\mathrm{vol}(\lambda a,b,c) = \lambda \cdot \mathrm{vol}(a,b,c)$.

Es ist nun eine bemerkenswerte Tatsache, daß eine auf diesen Beobachtungen beruhende Theorie orientierter Volumina rein algebraisch entwickelt werden kann (in beliebigen endlichdimensionalen Vektorräumen und über beliebigen Grundkörpern), ohne daß auf irgendeinen zuvor definierten Volumenbegriff (der oben nur motivationshalber herangezogen wurde) zurückgegriffen werden müßte. Diese Sichtweise, zuerst systematisch vertreten durch Hermann Graßmann (1809-1877) in seiner "Ausdehnungslehre", wird im vorliegenden Abschnitt entwickelt werden; dieser kann als eine geometrische Herleitung des Determinantenbegriffs aufgefaßt werden, die die rein arithmetische Herleitung des Abschnittes 36 ergänzt. Natürlich können die abstrakten Volumenfunktionen, die wir einführen werden, im allgemeinen nicht wirklich als physikalische "Volumina" von Punktmengen interpretiert werden, aber die suggestive geometrische Terminologie wird uns helfen, eine Intuition für abstrakte Volumenfunktionen zu entwickeln.

(59.1) Definition. *Es sei V ein n-dimensionaler Vektorraum über einem Körper K. Eine Abbildung*

$$\mathrm{vol}: \begin{array}{ccc} V \times \cdots \times V & \to & K \\ (v_1,\ldots,v_n) & \mapsto & \mathrm{vol}(v_1,\ldots,v_n) \end{array}$$

heißt **Volumenfunktion** *für V, wenn sie die folgenden Eigenschaften besitzt:*

(1) vol *ist linear in jedem Argument, d.h., für alle Vektoren $v_1,\ldots,v_n,a,b \in V$ und alle Skalare $\lambda,\mu \in K$ gilt die Gleichung* $\mathrm{vol}(v_1,\ldots,\lambda a+\mu b,\ldots,v_n) = \lambda \cdot \mathrm{vol}(v_1,\ldots,a,\ldots,v_n) + \mu \cdot \mathrm{vol}(v_1,\ldots,b,\ldots,v_n)$*;*

(2) *sind $v_1,\ldots,v_n \in V$ linear abhängig, so gilt* $\mathrm{vol}(v_1,\ldots,v_n) = 0$*.*

Eine Volumenfunktion heißt **nichttrivial***, wenn sie nicht identisch Null ist.*

Wir zeigen nun, daß Bedingung (2) durch eine andere Bedingung ersetzt werden kann, die manchmal einfacher zu behandeln ist.

(59.2) Hilfssatz. *Es sei V ein n-dimensionaler Vektorraum über einem Körper K. Es sei* $\mathrm{vol}: V^n \to K$ *eine Funktion, die linear in jedem ihrer n Argumente ist, also die Bedingung* (59.1)(1) *erfüllt. Wir betrachten die folgenden Bedingungen:*

(2) *sind die Vektoren $v_1,\ldots,v_n$ linear abhängig, so ist* $\mathrm{vol}(v_1,\ldots,v_n) = 0$*;*

(2′) *gibt es Indizes $i \neq j$ mit $v_i = v_j$, so ist* $\mathrm{vol}(v_1,\ldots,v_n) = 0$*;*

(2″) *bei Vertauschung zweier Argumente wechselt* vol *das Vorzeichen.*

Dann ist (2) *äquivalent zu* (2′) *und impliziert* (2″)*. Gilt* $\mathrm{char}K \neq 2$*, dann impliziert* (2″) *auch* (2′)*, so daß in diesem Fall alle drei Bedingungen äquivalent sind.*

Beweis. Die Implikation (2) $\Longrightarrow$ (2′) ist trivial. Um die Implikation (2′) $\Longrightarrow$ (2) zu beweisen, nehmen wir

an, die Vektoren $v_1, \dots, v_n \in V$ seien linear abhängig, so daß einer von ihnen als Linearkombination der andern geschrieben werden kann, sagen wir $v_{i_0} = \sum_{i \neq i_0} \lambda_i v_i$. Aufgrund der Bedingungen (1) und (2′) gilt dann

$$\begin{aligned} \operatorname{vol}(v_1, \dots, v_{i_0}, \dots, v_n) &= \operatorname{vol}(v_1, \dots, \sum_{i \neq i_0} \lambda_i v_i, \dots, v_n) \\ &= \sum_{i \neq i_0} \lambda_i \underbrace{\operatorname{vol}(v_1, \dots, v_i, \dots, v_n)}_{=0 \text{ wegen } (2')} = 0 . \end{aligned}$$

Als nächstes beweisen wir die Implikation (2′) $\Longrightarrow$ (2″). Sind $v_1, \dots, v_n \in V$ beliebige Vektoren, so schreiben wir zur Abkürzung $v_{ij} := \operatorname{vol}(v_1, \dots, v_i, \dots, v_j, \dots, v_n)$ und erhalten unter Annahme von (2′) die Gleichung

$$\begin{aligned} 0 &= \operatorname{vol}(v_1, \dots, v_i + v_j, \dots, v_i + v_j, \dots, v_n) \\ &= v_{ii} + v_{ij} + v_{ji} + v_{jj} = 0 + v_{ij} + v_{ji} + 0, \end{aligned}$$

also $v_{ji} = -v_{ij}$. Gilt umgekehrt (2″), so ist stets $v_{ii} = -v_{ii}$ bzw. $2v_{ii} = 0$, was im Falle $\operatorname{char} K \neq 2$ auf $v_{ii} = 0$ führt. ∎

Aus diesem Hilfssatz ergibt sich sofort, wie sich die Werte einer Volumenfunktion ändern, wenn ihre Argumente in irgendeiner Weise permutiert werden.

(59.3) Satz. *Ist* vol *eine Volumenfunktion für einen n-dimensionalen Vektorraum V und ist σ eine beliebige Permutation der Indizes $1, \dots, n$, so gilt für alle Vektoren $v_1, \dots, v_n$ die Beziehung*

$$\operatorname{vol}(v_{\sigma(1)}, \dots, v_{\sigma(n)}) = (\operatorname{sign} \sigma) \operatorname{vol}(v_1, \dots, v_n) .$$

Beweis. Nach (59.2)(2″) ändert vol bei jeder Vertauschung zweier Argumente das Vorzeichen. Läßt sich also σ als Hintereinanderausführung einer geraden Anzahl von Transpositionen darstellen, so gilt $\operatorname{vol}(v_{\sigma(1)}, \dots, v_{\sigma(n)}) = \operatorname{vol}(v_1, \dots, v_n)$; läßt sich dagegen σ als Hintereinanderausführung einer ungeraden Anzahl von Transpositionen darstellen, so gilt $\operatorname{vol}(v_{\sigma(1)}, \dots, v_{\sigma(n)}) = -\operatorname{vol}(v_1, \dots, v_n)$. Das ist aber gerade die Behauptung. ∎

Wir werden nun beweisen, daß jeder endlichdimensionale Vektorraum eine bis auf einen Skalarfaktor eindeutige nichttriviale Volumenfunktion besitzt. Die Idee, eine solche Funktion zu konstruieren, erhalten wir durch die folgende Beobachtung: Identifizieren wir $K^{n \times n}$ mit $K^n \times \cdots \times K^n$, indem wir eine quadratische Matrix $A \in K^{n \times n}$ als das n-Tupel ihrer Spalten auffassen, so ist $\det : K^{n \times n} \to K$ eine nichttriviale Volumenfunktion für K^n, denn die Determinante besitzt die definierende Eigenschaft, in jedem ihrer Argumente linear zu sein. Einen allgemeinen n-dimensionalen K-Vektorraum V werden wir durch Wahl einer Basis mit K^n identifizieren und so auch eine Volumenfunktion für V erhalten.

(59.4) Satz. *Es sei V ein n-dimensionaler Vektorraum über einem Körper K.*

(a) *Es gibt eine nichttriviale Volumenfunktion für V.*

(b) *Ist* vol *eine Volumenfunktion für V und ist $(e_1, \dots, e_n)$ eine fest gewählte Basis von V, so gilt für beliebige Vektoren $v_j = \sum_{i=1}^n v_{ij} e_i$ (mit $1 \leq j \leq n$) die Darstellung*

$$\operatorname{vol}(v_1, \dots, v_n) = \operatorname{vol}(e_1, \dots, e_n) \sum_\sigma (\operatorname{sign} \sigma) v_{\sigma(1)1} v_{\sigma(2)2} \cdots v_{\sigma(n)n},$$

wobei die Summe σ über alle Permutationen der Indizes $1, \dots, n$ läuft.

(c) *Ist* vol *eine nichttriviale Volumenfunktion für V und sind $v_1, \dots, v_n \in V$ linear unabhängig, so gilt* $\operatorname{vol}(v_1, \dots, v_n) \neq 0$. *(Dies ist die Umkehrung von Bedingung* (2) *der Definition einer Volumenfunktion.)*

(d) *Jedes skalare Vielfache einer Volumenfunktion ist wieder eine Volumenfunktion. Sind umgekehrt* vol_1 *und* vol_2 *zwei nichttriviale Volumenfunktionen für V, so gibt es ein Element $\lambda \neq 0$ in K mit* $\operatorname{vol}_2 = \lambda \operatorname{vol}_1$.

Beweis. (a) Wir wählen eine Basis $(e_1, \dots, e_n)$ von V und bezeichnen mit v_{ij} die Koordinaten von v_j bezüglich dieser Basis (so daß $v_j = \sum_{i=1}^n v_{ij} e_i$ für $1 \leq j \leq n$ gilt). Dann definiert

$$\operatorname{vol}(v_1, \dots, v_n) := \det(v_{ij})_{i,j}$$

aufgrund der Eigenschaften der Determinante eine nichttriviale Volumenfunktion für V.

(b) Wir haben zunächst ein kleines Notationsproblem: Da die Summen $v_1 = \sum_{i=1}^n v_{i1} e_i$, …, $v_n = \sum_{i=1}^n v_{in} e_i$ in einem einzigen Term auftreten werden, können wir nicht den gleichen Summationsindex i für alle diese Summen wählen; wir benutzen daher den Summationsindex i_k für die k-te Summe. Unter Ausnutzung der Linearität von vol in jedem Argument sehen wir, daß $\operatorname{vol}(v_1, \dots, v_n)$ gegeben ist durch

$$\begin{aligned} &\operatorname{vol}\Big(\sum_{i_1=1}^n v_{i_1 1} e_{i_1}, \sum_{i_2=1}^n v_{i_2 2} e_{i_2}, \dots, \sum_{i_n=1}^n v_{i_n n} e_{i_n}\Big) \\ &= \sum_{i_1=1}^n \sum_{i_2=1}^n \cdots \sum_{i_n=1}^n v_{i_1 1} v_{i_2 2} \cdots v_{i_n n} \operatorname{vol}(e_{i_1}, \dots, e_{i_n}) \\ &= \sum_\varphi v_{\varphi(1)1} v_{\varphi(2)2} \cdots v_{\varphi(n)n} \operatorname{vol}(e_{\varphi(1)}, \dots, e_{\varphi(n)}), \end{aligned}$$

wobei im letzten Term über alle Abbildungen $\varphi : \{1, \dots, n\} \to \{1, \dots, n\}$ summiert wird. Ist nun eine solche Abbildung φ nicht bijektiv, so gibt es Indizes $i \neq j$ mit $\varphi(i) = \varphi(j)$, was $\operatorname{vol}(e_{\varphi(1)}, \dots, e_{\varphi(n)}) = 0$ zur Folge hat. Also tragen nur bijektive Abbildungen, also die Permutationen der Menge $\{1, \dots, n\}$, etwas zur Summe bei. Folg-

lich gilt

$$\begin{aligned}&\operatorname{vol}(v_1,\dots,v_n)\\ &= \sum_\sigma v_{\sigma(1)1}v_{\sigma(2)2}\cdots v_{\sigma(n)n}\operatorname{vol}(e_{\sigma(1)},\dots,e_{\sigma(n)})\\ &= \sum_\sigma v_{\sigma(1)1}v_{\sigma(2)2}\cdots v_{\sigma(n)n}(\operatorname{sign}\sigma)\operatorname{vol}(e_1,\dots,e_n),\end{aligned}$$

wobei über alle Permutationen σ summiert wird. (In der letzten Gleichung wurde (59.3) benutzt.)

(c) Teil (b) zeigt, daß vol identisch Null ist, wenn vol für irgendeine Basis von V den Wert Null annimmt; dies ist gerade die Behauptung.

(d) Die erste Behauptung folgt sofort aus der Definition einer Volumenfunktion. Sind umgekehrt vol_1 und vol_2 nichttriviale Volumenfunktionen und ist $(e_1,\dots,e_n)$ irgendeine Basis von V, so zeigt die in Teil (b) erhaltene Formel, daß

$$\frac{\operatorname{vol}_1(v_1,\dots,v_n)}{\operatorname{vol}_1(e_1,\dots,e_n)} = \frac{\operatorname{vol}_2(v_1,\dots,v_n)}{\operatorname{vol}_2(e_1,\dots,e_n)}$$

für alle $v_1,\dots,v_n \in V$ gilt. Dann gilt aber $\operatorname{vol}_2 = \lambda\cdot\operatorname{vol}_1$ mit $\lambda := \operatorname{vol}_2(e_1,\dots,e_n)/\operatorname{vol}_1(e_1,\dots,e_n)$. ∎

Satz (59.4) liefert eine explizite Formel für die Determinante einer quadratischen Matrix, die zwar nicht besonders zweckmäßig ist, wenn es um die numerische Berechnung einer Determinante geht, die aber durchaus wichtig für theoretische Zwecke ist.

(59.5) Satz. *Für jede quadratische Matrix $A \in K^{n\times n}$ gilt* $\det(A) = \sum_\sigma(\operatorname{sign}\sigma)a_{\sigma(1)1}a_{\sigma(2)2}\cdots a_{\sigma(n)n}$*, wobei über alle Permutationen der Menge $\{1,\dots,n\}$ summiert wird.*

Beweis. Wir müssen nur Teil (b) von Satz (59.4) mit $V = K^n$, vol = det und der kanonischen Basis $(e_1,\dots,e_n)$ anwenden. ∎

Aus (59.4)(b) ergibt sich, daß das Volumen eines beliebigen Spats schon feststeht, wenn für einen beliebig herausgegriffenen "Einheitsspat" ein "Einheitsvolumen" festgelegt wird. In der Elementargeometrie nimmt man etwa zwei zueinander senkrechte Vektoren e_1 und e_2 der Länge 1 und gibt dem von von diesen aufgespannten "Einheitsquadrat" *per definitionem* den Flächeninhalt 1 (wodurch eine Flächeneinheit festgelegt wird); im räumlichen Fall nimmt man drei paarweise zueinander senkrechte Vektoren e_1, e_2 und e_3 der Länge 1 und gibt dem von diesen aufgespannten "Einheitswürfel" das Volumen 1 (was die Festsetzung einer Volumeneinheit bedeutet). Die Formel in (59.4)(b) zeigt dann, wie sich der Inhalt eines beliebigen Parallelogramms bzw. Spates mit dieser Festsetzung einer Flächen- bzw. Volumeneinheit berechnen läßt.

(59.6) Beispiel. Wir betrachten die Fläche des von zwei Vektoren $a, b \in \mathbb{R}^2$ aufgespannten Parallelogramms.

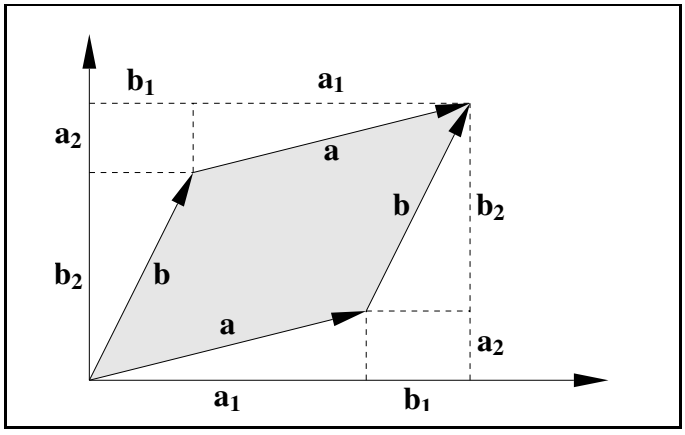

Abb. 59.4: Fläche eines Parallelogramms.

Diese Fläche ergibt sich elementargeometrisch, indem wir von der großen Rechtecksfläche die beiden kleinen Rechtecksflächen sowie die vier Dreiecksflächen subtrahieren; es ergibt sich

$$\begin{aligned}&(a_1+b_1)(a_2+b_2) - b_1a_2 - a_2b_1 - b_1b_2 - a_1a_2\\ &= a_1b_2 - a_2b_1 = \det\begin{bmatrix}a_1 & b_1\\ a_2 & b_2\end{bmatrix}.\end{aligned}$$ ♦

Am Anfang dieses Abschnittes motivierten wir die Einführung von Volumenfunktionen mit dem Begriff des *orientierten Volumens* im Vektorraum $\mathfrak{V}$ aller Pfeilklassen. Die Unterscheidung zwischen verschiedenen Orientierungen in $\mathfrak{V}$ basierte dabei auf der "handgreiflichen" Tatsache, daß drei linear unabhängige Vektoren in $\mathfrak{V}$ wie Daumen, Zeigefinger und Mittelfinger entweder der rechten oder der linken Hand zeigen. Wir werden jetzt den Begriff der Orientierung für einen beliebigen reellen Vektorraum endlicher Dimension einführen. Intuitiv bedeutet dabei die Orientierung eines eindimensionalen Vektorraums einen Richtungssinn (Unterscheidung zwischen links und rechts), die Orientierung eines zweidimensionalen Vektorraums einen Drehsinn (Unterscheidung zwischen Drehungen im und gegen den Uhrzeigersinn) und schließlich die Orientierung eines dreidimensionalen Vektorraums einen Schraubungssinn (Unterscheidung zwischen links- und rechtsläufigen Schraubungen).

(59.7) Definition. *Es sei V ein n-dimensionaler reeller Vektorraum V. Wir definieren eine Äquivalenzrelation auf der Menge aller geordneten Basen von V, indem wir zwei Basen $(v_1,\dots,v_n)$ und $(w_1,\dots,w_n)$* **gleichorientiert** *nennen, falls für eine (und damit jede) nichttriviale Volumenfunktion* vol *für V die reellen Zahlen* $\operatorname{vol}(v_1,\dots,v_n)$ *und* $\operatorname{vol}(w_1,\dots,w_n)$ *das gleiche Vorzeichen haben. Eine* **Orientierung** *von V ist eine Äquivalenzklasse unter dieser Relation.*

Man überzeugt sich schnell davon, daß es genau zwei verschiedene Orientierungen von V gibt. (In der Regel bezeichnet man eine dieser Orientierungen als positiv und die sie repräsentierenden Basen als Rechtssysteme, die andere als negativ und die sie repräsentierenden Basen als Linkssysteme.)

60. Determinante und Spur eines Endomorphismus

Um unsere nächste Definition vorzubereiten, betrachten wir wieder den Vektorraum $\mathfrak{V}$ aller Pfeilklassen. Eine lineare Selbstabbildung $f : \mathfrak{V} \to \mathfrak{V}$ bildet jeden Spat $\mathfrak{S}(a,b,c)$ wieder auf einen Spat ab, nämlich auf $\mathfrak{S}\big(f(a),f(b),f(c)\big)$.

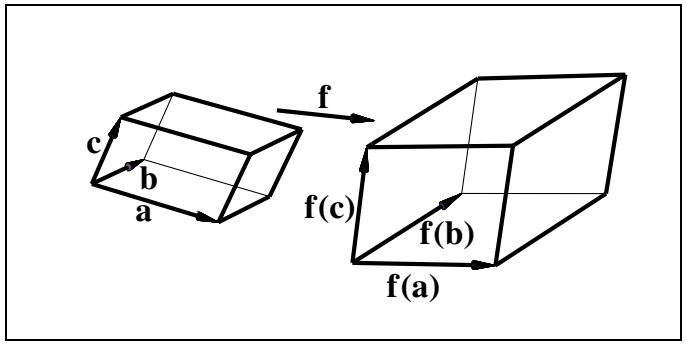

Abb. 60.1: Wirkung eines Endomorphismus auf einen Spat.

Es wird sicher zum geometrischen Verständnis der Abbildung f beitragen, wenn wir etwas über das Verhältnis der Volumina dieser beiden Spate aussagen können. Es wird sich zeigen, daß dieses Verhältnis unabhängig von a, b und c ist und daher eine geometrische Invariante von f darstellt.

(60.1) Hilfssatz. *Ist $f : V \to V$ ein Endomorphismus eines n-dimensionalen Vektorraumes V, so ist der Ausdruck*

$$(\star) \qquad \frac{\mathrm{vol}\big(f(v_1),\dots,f(v_n)\big)}{\mathrm{vol}(v_1,\dots,v_n)}$$

unabhängig von der nichttrivialen Volumenfunktion vol *und der Basis $(v_1,\dots,v_n)$.*

Beweis. Es ist klar, daß $(\star)$ nicht von der Wahl der Volumenfunktion vol abhängt, denn je zwei nichttriviale Volumenfunktionen unterscheiden sich nur um einen von Null verschiedenen Skalarfaktor, der sich in dem Ausdruck $(\star)$ wegkürzt. Ferner können wir zu gegebener Volumenfunktion vol eine Funktion $\overline{\mathrm{vol}}$ durch

$$\overline{\mathrm{vol}}(u_1,\dots,u_n) \;:=\; \mathrm{vol}\big(f(u_1),\dots,f(u_n)\big)$$

definieren. Es ist leicht nachzuprüfen, daß $\overline{\mathrm{vol}}$ wieder eine Volumenfunktion und daher nach (59.4)(d) ein skalares Vielfaches von vol ist, sagen wir $\overline{\mathrm{vol}} = \lambda\,\mathrm{vol}$ mit $\lambda \in K$. Insbesondere nimmt der Ausdruck $(\star)$ den Wert λ für jede Wahl von $v_1,\dots,v_n$ an. ∎

Der gerade bewiesene Hilfssatz erlaubt die Definition der Determinante eines Endomorphismus als eines Verzerrungsfaktors für orientierte Volumina.

(60.2) Definition. *Es sei $f : V \to V$ ein Endomorphismus des n-dimensionalen Vektorraums V. Die* **Determinante** *von f wird dann definiert als*

$$\det(f) \;:=\; \frac{\mathrm{vol}\big(f(v_1),\dots,f(v_n)\big)}{\mathrm{vol}(v_1,\dots,v_n)},$$

wobei vol *irgendeine nichttriviale Volumenfunktion für V und $(v_1,\dots,v_n)$ irgendeine Basis von V ist.*

Die wichtigsten Eigenschaften der Determinante liefert der folgende Satz.

(60.3) Satz. *Es seien $f,g : V \to V$ Endomorphismen eines n-dimensionalen Vektorraumes V; ferner bezeichnen wir mit $\mathbf{1}$ die identische Abbildung von V.*

(a) *Es gilt* $\det(\mathbf{1}) = 1$.

(b) *Für alle $\lambda \in K$ gilt* $\det(\lambda f) = \lambda^n \det(f)$.

(c) *Es gilt* $\det(g \circ f) = \det(g)\det(f)$.

(d) *Genau dann ist f invertierbar, wenn* $\det(f) \neq 0$ *gilt; in diesem Fall ist* $\det(f^{-1}) = \det(f)^{-1}$.

Beweis. Wir wählen eine Basis $(v_1,\dots,v_n)$ des Raumes V und eine nichttriviale Volumenfunktion vol für V. Teil (a) folgt dann sofort wegen $\mathrm{vol}\big(\mathbf{1}(v_1),\dots,\mathbf{1}(v_n)\big) = \mathrm{vol}(v_1,\dots,v_n)$. Zum Beweis von Teil (b) beachten wir, daß wegen der Linearität von vol in jedem Argument die folgende Gleichung gilt:

$$\begin{aligned}&\mathrm{vol}\big((\lambda f)(v_1),\dots,(\lambda f)(v_n)\big) = \mathrm{vol}\big(\lambda\cdot f(v_1),\dots,\lambda\cdot f(v_n)\big)\\ &= \lambda^n\,\mathrm{vol}\big(f(v_1),\dots,f(v_n)\big) = \lambda^n \det(f)\,\mathrm{vol}(v_1,\dots,v_n).\end{aligned}$$

Zum Nachweis von (c) beachten wir

$$\begin{aligned}&\mathrm{vol}\big((g\circ f)(v_1),\dots,(g\circ f)(v_n)\big)\\ &= \mathrm{vol}\,\big(g\big(f(v_1)\big),\dots,g\big(f(v_n)\big)\big)\\ &= \det(g)\,\mathrm{vol}\big(f(v_1),\dots,f(v_n)\big)\\ &= \det(g)\det(f)\,\mathrm{vol}(v_1,\dots,v_n).\end{aligned}$$

Schließlich beweisen wir Teil (d). Ist f invertierbar, so gilt $1 = \det(\mathbf{1}) = \det(f\circ f^{-1}) = \det(f)\det(f^{-1})$ wegen (a) und (c); also gelten die Bedingungen $\det(f) \neq 0$ und $\det(f^{-1}) = \det(f)^{-1}$. Ist f nicht invertierbar, so sind die Vektoren $f(v_1),\dots,f(v_n)$ linear abhängig, was $\mathrm{vol}\big(f(v_1),\dots,f(v_n)\big) = 0$ und damit $\det(f) = 0$ zur Folge hat. ∎

Die Determinante einer quadratischen Matrix $A \in K^{n\times n}$ hat nun zwei Bedeutungen. Fassen wir A als das

n-tupel der Spalten $s_1(A), \dots, s_n(A)$ von A auf, so ist $\det(A) = \mathrm{vol}\big(s_1(A), \dots, s_n(A)\big)$, wobei vol die eindeutige Volumenfunktion auf K^n mit $\mathrm{vol}(e_1, \dots, e_n) = 1$ für die kanonische Basis $(e_1, \dots, e_n)$ gilt; d.h, die Determinante von A ist das orientierte Volumen der Spalten von A. Fassen wir dagegen A als einen Endomorphismus von K^n auf, so gilt $\det(A) = \mathrm{vol}(Ae_1, \dots, Ae_n)/\mathrm{vol}(e_1, \dots, e_n) = \mathrm{vol}(Ae_1, \dots, Ae_n)$. Der Wert von $\det(A)$ ist aber in beiden Interpretationen der gleiche, denn es gilt $Ae_i = s_i(A)$ for $1 \le i \le n$. Etwas allgemeiner gilt das folgende Ergebnis.

(60.4) Satz. *Ist $f : V \to V$ ein Endomorphismus eines n-dimensionalen Vektorraums V und ist A die Matrixdarstellung von f bezüglich irgendeiner Basis $(e_1, \dots, e_n)$ von V, so gilt* $\det(f) = \det(A)$.

Beweis. Für $1 \le i \le n$ sei $v_i := f(e_i)$. Dann ist einerseits $\mathrm{vol}(v_1, \dots, v_n) = \mathrm{vol}\big(f(e_1), \dots, f(e_n)\big) = \det(f)\,\mathrm{vol}(e_1, \dots, e_n)$; andererseits gilt $\mathrm{vol}(v_1, \dots, v_n) = \mathrm{vol}(e_1, \dots, e_n)\det(A)$ nach (59.4)(b). ∎

Das folgende Beispiel illustriert die Bedeutung der Determinante als Verzerrungsfaktor für Volumina (speziell im zweidimensionalen Fall also für Flächeninhalte).

(60.5) Aufgabe. *Ein Sonnenschirm bestehe aus einem fest montierten Ständer, einem mit einem Kugelgelenk am Ständer befestigten Dreharm und der Schirmfläche am Ende des Dreharms, die wir der Einfachheit halber als ebenes Flächenstück (etwa als ein Sechseck) voraussetzen. Wie muß der Dreharm gegenüber der Einfallsrichtung der Sonnenstrahlen ausgerichtet werden, damit die entstehende Schattenfläche am Boden maximal wird?*

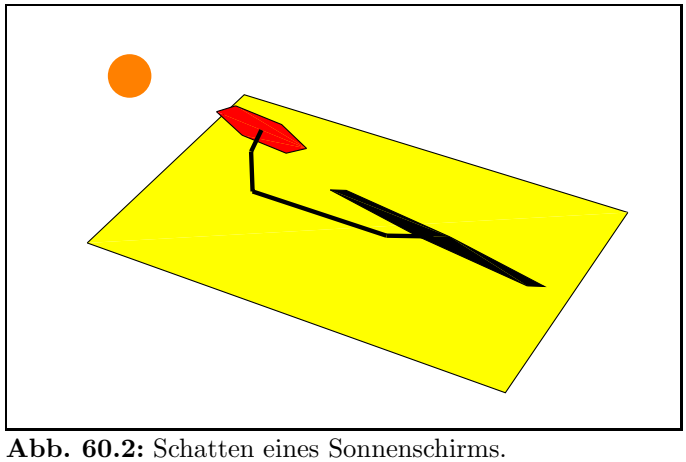

Abb. 60.2: Schatten eines Sonnenschirms.

Lösung. Es sei $\mathfrak{V}$ der Vektorraum aller Pfeilklassen. Wir bezeichnen mit n den himmelwärts zeigenden Einheitsnormalenvektor des Bodens und mit v die Richtung der Sonnenstrahlen (wobei $\|v\| = 1$ sei). Ferner wählen wir eine Orthonormalbasis (b_1, b_2, b_3) so, daß b_1 und b_2 die *Schirmebene* aufspannen (also die Ebene, in der die Schirmfläche liegt), während b_3 in Richtung des Dreharms zeigt (und damit senkrecht auf der Schirmebene steht).

Wir bezeichnen weiterhin mit I die Einbettung der Schirmebene in $\mathfrak{V}$ und mit P die Projektion von $\mathfrak{V}$ in Richtung v auf den Boden; dann ist die Hintereinanderausführung $P \circ I$ eine Abbildung von der Schirmebene in die Bodenebene. Da $|\det(P \circ I)|$ gerade der Faktor ist, um den sich der Flächeninhalt eines beliebigen Parallelogramms (und damit auch der eines allgemeineren Flächenstücks, etwa eines Polygons) unter Anwendung von $P \circ I$ verzerrt, ist die Fläche des Schattens gerade das $|\det(P \circ I)|$-fache der Schirmfläche; zu maximieren ist also $|\det(P \circ I)|$. Eine Basis der Bodenebene ist gegeben durch die beiden Vektoren

$$Pb_i \;=\; b_i - \frac{\langle b_i, n\rangle}{\langle v, n\rangle} v \;=\; b_i + \frac{\langle b_i, n\rangle}{\sin\alpha} v$$

mit $i = 1, 2$; ergänzen wir diese zu der Basis (Pb_1, Pb_2, n), so ist $\det(P \circ I)$ nichts anderes als

$$\begin{aligned}
&\det(Pb_1, Pb_2, n) = \det\left(b_1 + \frac{\langle b_1, n\rangle}{\sin\alpha} v, b_2 + \frac{\langle b_2, n\rangle}{\sin\alpha} v, n\right) \\
&= \det(b_1, b_2, n) + \frac{\langle b_1, n\rangle \det(v, b_2, n) + \langle b_2, n\rangle \det(b_1, v, n)}{\sin\alpha} \\
&= \langle n, b_1 \times b_2\rangle + \frac{-\langle b_1, n\rangle\langle b_2, v \times n\rangle + \langle b_2, n\rangle\langle b_1, v \times n\rangle}{\sin\alpha} \\
&= \langle n, b_1 \times b_2\rangle + \frac{\langle b_1 \times b_2, (v \times n) \times n\rangle}{\sin\alpha} \\
&= \langle n, b_1 \times b_2\rangle + \frac{\langle b_1 \times b_2, -v - (\sin\alpha) n\rangle}{\sin\alpha} \\
&= -\frac{\langle b_1 \times b_2, v\rangle}{\sin\alpha} \;=\; -\frac{\langle b_3, v\rangle}{\sin\alpha}\,.
\end{aligned}$$

Dieser Ausdruck wird dann betragsmäßig maximal, wenn $b_3 = \pm v$ gilt. Die Rechnung liefert also das (intuitiv einleuchtende) Ergebnis, daß die Schattenfläche dann maximal wird, wenn der Dreharm genau in Sonnenrichtung zeigt, die Schirmfläche also senkrecht zur Einfallsrichtung der Sonnenstrahlen liegt. ∎

(60.6) Bemerkung. Ist $f : V \to V$ ein Isomorphismus, so gibt das Vorzeichen von $\det(f)$ an, ob f orientierungserhaltend oder orientierungsumkehrend ist. Sind zwei Basen $(v_1, \dots, v_n)$ und $(w_1, \dots, w_n)$ gegeben, so gibt es genau einen Isomorphismus $f : V \to V$ mit $f(v_i) = w_i$ für $1 \le i \le n$; gilt $\det(f) > 0$, so sind die beiden Basen gleichorientiert, gilt $\det(f) < 0$, so sind sie entgegengesetzt orientiert. ∎

Wir werden neben der Determinante jedem Endomorphismus eines endlichdimensionalen Vektorraums ein zweites Skalar zuordnen, das sich (wenn wir erst den Begriff der Ableitung zu Verfügung haben) gewissermaßen als "infinitesimale Version" der Determinante herausstellen wird.

(60.7) Hilfssatz. *Ist $\varphi : V \to V$ ein Endomorphismus eines n-dimensionalen Vektorraumes V, so ist der Ausdruck*

$$(\star) \qquad \frac{\sum_{i=1}^{n} \operatorname{vol}(v_1, \ldots, \varphi v_i, \ldots, v_n)}{\operatorname{vol}(v_1, \ldots, v_n)}$$

unabhängig von der nichttrivialen Volumenfunktion vol *und der Basis $(v_1, \ldots, v_n)$.*

Beweis. Es ist klar, daß $(\star)$ nicht von der Wahl der Volumenfunktion $(\star)$ abhängt, denn je zwei nichttriviale Volumenfunktionen unterscheiden sich nur um einen von Null verschiedenen Skalarfaktor, der sich in dem Ausdruck $(\star)$ wegkürzt. Ferner können wir zu gegebener Volumenfunktion vol eine Funktion $\overline{\text{vol}}$ durch

$$\overline{\text{vol}}(u_1, \ldots, u_n) \;:=\; \sum_{i=1}^{n} \operatorname{vol}(u_1, \ldots, \varphi u_i, \ldots, u_n)$$

definieren. Es ist leicht nachzuprüfen, daß $\overline{\text{vol}}$ wieder eine Volumenfunktion und daher nach (59.4)(d) ein skalares Vielfaches von vol ist, sagen wir $\overline{\text{vol}} = \lambda \operatorname{vol}$ mit $\lambda \in K$. Insbesondere nimmt der Ausdruck $(\star)$ den Wert λ für jede Wahl von $v_1, \ldots, v_n$ an. ∎

(60.8) Definition. *Es sei $f : V \to V$ ein Endomorphismus eines n-dimensionalen Vektorraumes V. Die* **Spur** *(englisch "trace") von f wird dann definiert als*

$$\operatorname{tr}(f) \;:=\; \frac{\sum_{i=1}^{n} \operatorname{vol}\big(v_1, \ldots, f(v_i), \ldots, v_n\big)}{\operatorname{vol}(v_1, \ldots, v_n)},$$

wobei vol *irgendeine nichttriviale Volumenfunktion für V und $(v_1, \ldots, v_n)$ irgendeine Basis von V ist.*

Der folgende Satz zeigt, wie sich die Spur eines Endomorphismus leicht aus einer beliebigen Matrixdarstellung desselben bestimmen läßt.

(60.9) Satz. *Es seien $f : V \to V$ ein Endomorphismus eines endlichdimensionalen Vektorraumes V und A die Matrixdarstellung von f bezüglich irgendeiner Basis $(e_1, \ldots, e_n)$ von V. Dann gilt $\operatorname{tr}(f) = a_{11} + \cdots + a_{nn}$; die Spur von f ist die Summe der Diagonalelemente von A.*

Beweis. Für $1 \le i \le n$ haben wir $f(e_i) = \sum_{j=1}^{n} a_{ji} e_j$; folglich ist $\operatorname{vol}\big(e_1, \ldots, f(e_i), \ldots, e_n\big)$ gleich dem Ausdruck

$$\sum_{j=1}^{n} a_{ji} \underbrace{\operatorname{vol}(e_1, \ldots, e_j, \ldots, e_n)}_{=\,0 \text{ für } j \neq i} \;=\; a_{ii} \operatorname{vol}(e_1, \ldots, e_n),$$

was $\operatorname{vol}\big(e_1, \ldots, f(e_i), \ldots, e_n\big) / \operatorname{vol}(e_1, \ldots, e_n) = a_{ii}$ nach sich zieht. Hieraus folgt sofort die Behauptung. ∎

Die wichtigsten Eigenschaften der Spur liefert der folgende Satz.

(60.10) Satz. *Es seien $f, g : V \to V$ Endomorphismen eines n-dimensionalen Vektorraumes V; ferner bezeichnen wir mit $\mathbf{1}$ die identische Abbildung von V.*

(a) *Es gilt* $\operatorname{tr}(\mathbf{1}) = n \cdot 1$.

(b) *Für $\lambda, \mu \in K$ gilt* $\operatorname{tr}(\lambda f + \mu g) = \lambda \operatorname{tr}(f) + \mu \operatorname{tr}(g)$.

(c) *Es gilt* $\operatorname{tr}(g \circ f) = \operatorname{tr}(f \circ g)$.

Beweis. Wir wählen eine Basis $(v_1, \ldots, v_n)$ von V und eine nichttriviale Volumenfunktion vol für V. Teil (a) folgt dann sofort wegen $\sum_{i=1}^{n} \operatorname{vol}\big(v_1, \ldots, \mathbf{1}(v_i), \ldots, v_n\big) = n \cdot \operatorname{vol}(v_1, \ldots, v_n)$. Um Teil (b) einzusehen, beachten wir, daß wegen der Linearität von vol in jedem Argument die Gleichung

$$\begin{aligned} &\sum_{i=1}^{n} \operatorname{vol}\big(v_1, \ldots, \lambda f(v_i) + \mu g(v_i), \ldots, v_n\big) \\ &= \lambda \sum_{i=1}^{n} \operatorname{vol}\big(v_1, \ldots, f(v_i), \ldots, v_n\big) \\ &+ \mu \sum_{i=1}^{n} \operatorname{vol}\big(v_1, \ldots, g(v_i), \ldots, v_n\big) \end{aligned}$$

gilt. Um schließlich Teil (c) einzusehen, benutzen wir (60.9) und betrachten die Matrixdarstellungen A und B von f und g bezüglich irgendeiner (fest gewählten) Basis von V; dann werden $g \circ f$ und $f \circ g$ durch die Matrizenprodukte BA und AB repräsentiert, und wir erhalten $\operatorname{tr}(g \circ f) = \sum_{i=1}^{n} (BA)_{ii} = \sum_{i=1}^{n} \sum_{k=1}^{n} b_{ik} a_{ki} = \sum_{k=1}^{n} (AB)_{kk} = \operatorname{tr}(f \circ g)$. ∎

82. Stetigkeit

In ingenieurwissenschaftlichen Anwendungen sieht man sich oft vor die Aufgabe gestellt, für eine direkt kontrollierbare Größe x ("Stellgröße") einen Wert x_0 ("Arbeitspunkt") so einzustellen, daß eine von x abhängige Größe $y = f(x)$ ("Zielgröße") einen vorgegebenen Wert y_0 annimmt, daß also $f(x_0) = y_0$ gilt. (Man denke etwa an die Herbeiführung gewünschter Stoffkonzentrationen innerhalb eines chemischen Prozesses durch geeignete Einstellung von Temperaturen und Drücken.) Aufgrund unvermeidlicher Ungenauigkeiten wird der tatsächlich eingestellte Wert $\widehat{x}$ von dem angestrebten Wert x_0 um eine (kleine) Größe $\Delta x = \widehat{x} - x_0$ abweichen, was zur Folge hat, daß auch der Wert der Zielgröße vom gewünschten Zielwert y_0 abweicht, nämlich um $\Delta y = \widehat{y} - y_0 = f(\widehat{x}) - f(x_0)$. Es ist natürlich wünschenswert, daß eine kleine Einstellungsungenauigkeit Δx auch nur eine kleine Zielabweichung Δy nach sich zieht, daß also kleine Fehler bei der Einstellung nicht katastrophale Änderungen der Zielgröße nach sich ziehen. Dies ist eine Eigenschaft der Funktion f, die, grob gesprochen, darin besteht, daß nahe beieinanderliegende Argumente auch auf nahe beieinanderliegende Werte abgebildet werden, daß also geringe Änderungen im Argument nicht zu sprunghaften Änderungen im Wert der Funktion führen können. Dies führt auf die folgende allgemeine Definition.

(82.1) Definition. *Eine Funktion $f : X \to Y$ zwischen metrischen Räumen heißt* **stetig** *in einem Punkt $x_0 \in X$, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt derart, daß aus $d_X(x, x_0) < \delta$ in X schon $d_Y\big(f(x), f(x_0)\big) < \varepsilon$ in Y folgt. Die Funktion $f : X \to Y$ heißt stetig schlechthin, wenn sie in jedem Punkt von X stetig ist.*

In der Folge werden wir oft nicht pedantisch d_X für die Metrik im Urbildbereich und d_Y für die Metrik im Bildbereich schreiben, sondern jeweils einfach d, jedenfalls dann, wenn keine Mißverständnisse zu befürchten sind, welche Metrik gerade gemeint ist. Definition (82.1) definiert Stetigkeit an einer Stelle x_0 als diejenige Eigenschaft einer Funktion f, die besagt, daß $f(x)$ "beliebig dicht" bei $f(x_0)$ liegt, wenn nur x "genügend dicht" ("hinreichend dicht") bei x_0 liegt. In der Definition interpretieren wir ε als die erlaubte "Toleranz" im Zielwert und δ als das zulässige "Spiel" bei der Einstellung des Arbeitspunktes. Die Bedeutung dieser Definition wird auch sofort klar, wenn man sich $y = f(x)$ als physikalische Größe vorstellt, deren Wert man durch Messung der Größe x und anschließendes Einsetzen in die Gleichung $y = f(x)$ ermitteln möchte. Die folgende Frage ist dann naheliegend: Mit welcher Genauigkeit $\delta > 0$ muß ich die Meßgröße x bestimmen, um die Größe y mit einer vorgegebenen Genauigkeit $\varepsilon > 0$ zu ermitteln? (Gibt es zu einem gewissen $\varepsilon > 0$ kein solches $\delta > 0$, so bedeutet dies, daß y nicht beliebig genau durch Messen der Größe x bestimmt werden kann.)

Bevor wir Beispiele für stetige Abbildungen geben, definieren wir noch einen auf Rudolf Otto Sigismund Lipschitz (1832-1903) zurückgehenden verschärften Stetigkeitsbegriff, der meist leicht nachprüfbar ist und sich daher oft gut zum Stetigkeitsnachweis für eine konkret gegebene Funktion eignet.

(82.2) Definition. *Eine Funktion $f : X \to Y$ zwischen metrischen Räumen heißt* **Lipschitz-stetig**, *wenn es eine Konstante L gibt mit $d\big(f(x_1), f(x_2)\big) \le L \cdot d(x_1, x_2)$ für alle $x_1, x_2 \in X$; jede solche Konstante heißt eine* **Lipschitzkonstante** *für f. Wir nennen f eine* **kontrahierende Abbildung** *oder kurz* **Kontraktion**, *wenn es für f eine Lipschitzkonstante $L < 1$ gibt.*

Es ist klar, daß eine auf einer Umgebung eines Punktes $x_0 \in X$ Lipschitz-stetige Funktion in diesem Punkt automatisch stetig ist. Um nämlich bei vorgegebener Toleranz $\varepsilon > 0$ die Bedingung $d\big(f(x), f(x_0)\big) < \varepsilon$ zu erhalten, müssen wir ja nur $\delta := \varepsilon/L$ wählen; aus $d(x, x_0) < \delta$ folgt dann automatisch $d\big(f(x), f(x_0)\big) \le L \cdot \delta = \varepsilon$.

(82.3) Beispiele. (a) Es sei $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$. Jede Funktion $f : \mathbb{K} \to \mathbb{K}$ der Form $f(x) = ax + b$ ist Lipschitz-stetig mit der Lipschitzkonstanten $|a|$, denn es gilt $|f(x_1) - f(x_2)| = |a(x_1 - x_2)| = |a|\,|x_1 - x_2|$ für alle $x_1, x_2 \in \mathbb{K}$.

(b) Sowohl für $f(x) := \sin(x)$ als auch für $f(x) := \cos(x)$ gilt die Abschätzung $|f(b) - f(a)| \le |b - a|$, wie die folgende Abbildung am Einheitskreis unmittelbar zeigt. Sinus und Cosinus sind also als Funktionen von $\mathbb{R}$ nach $\mathbb{R}$ Lipschitz-stetig mit der Lipschitzkonstanten 1.

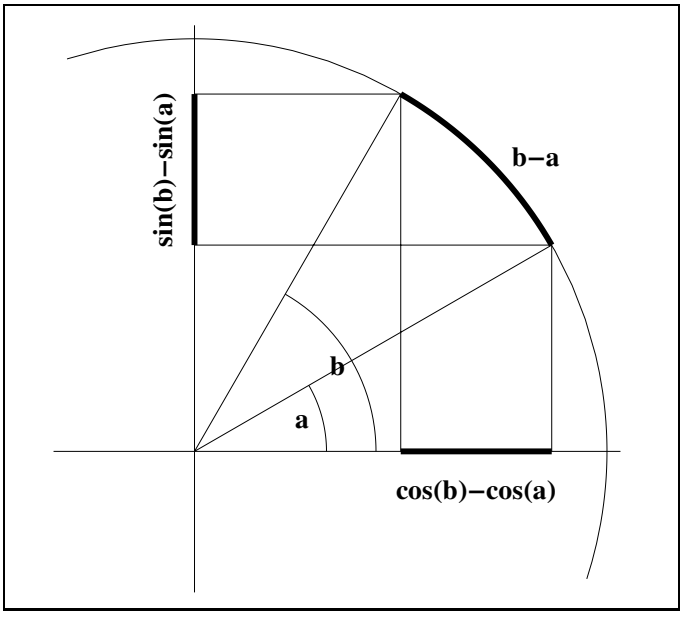

Abb. 82.1: Lipschitz-Stetigkeit von Sinus und Cosinus.

(c) Für jedes $\varepsilon > 0$ ist die durch $f(x) := \sqrt{x}$ definierte Funktion $f : [\varepsilon, \infty) \to \mathbb{R}$ Lipschitz-stetig mit der Lipschitzkonstanten $1/(2\sqrt{\varepsilon})$, denn für alle $x, y \ge \varepsilon$ haben wir $\sqrt{x} + \sqrt{y} \ge \sqrt{\varepsilon} + \sqrt{\varepsilon} = 2\sqrt{\varepsilon}$, für $x \ne y$ folglich

$$\left|\frac{f(x)-f(y)}{x-y}\right| = \left|\frac{\sqrt{x}-\sqrt{y}}{x-y}\right| = \frac{1}{\sqrt{x}+\sqrt{y}} \le \frac{1}{2\sqrt{\varepsilon}} =: L$$

und damit $|f(x)-f(y)| \le L\cdot|x-y|$ für alle $x,y \ge \varepsilon$. (Insbesondere ist damit f stetig in jedem Punkt $x_0 > 0$.) Dagegen ist f als Funktion $f : [0,\infty) \to \mathbb{R}$ *nicht* Lipschitz-stetig, denn der Ausdruck

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

geht für $x \to 0$ und $y \to 0$ offensichtlich gegen Unendlich, bleibt also nicht durch eine Lipschitzkonstante L beschränkt. (Dagegen ist f auch stetig im Punkt $x_0 := 0$, denn zu gegebenem $\varepsilon > 0$ müssen wir ja nur $\delta := \varepsilon^2 > 0$ wählen, um aus $|x-x_0| < \delta$ schon $|f(x)-f(x_0)| < \varepsilon$ folgern zu können.)

(d) Ist V ein beliebiger normierter Raum, so ist die Normabbildung $\|\cdot\| : V \to \mathbb{R}$ Lipschitz-stetig mit der Lipschitzkonstanten 1. Für alle $x,y \in V$ erhalten wir nach der Dreiecksungleichung nämlich einerseits $\|x\| = \|x-y+y\| \le \|x-y\| + \|y\|$, also $\|x\| - \|y\| \le \|x-y\|$, andererseits in völlig analoger Weise auch $\|y\| - \|x\| \le \|y-x\| = \|x-y\|$. Diese beiden Ungleichungen ergeben zusammen die Abschätzung

$$\Big|\,\|x\| - \|y\|\,\Big| \;\le\; \|x-y\|$$

und damit die Behauptung. Insbesondere sind also die Betragsfunktionen auf $\mathbb{R}$ und $\mathbb{C}$ Lipschitz-stetig mit der Lipschitzkonstanten 1.

(e) Wie das in (d) aufgeführte Beispiel der Betragsfunktion $f(x) = |x|$ zeigt, stört ein "Knick" im Graphen einer Funktion nicht die Stetigkeit dieser Funktion. Ein "Sprung" im Funktionsgraphen signalisiert dagegen immer eine Unstetigkeitsstelle; ein solcher Sprung tritt etwa auf, wenn wir das Volumen einer fest gegebenen Wassermenge als Funktion der Temperatur betrachten.

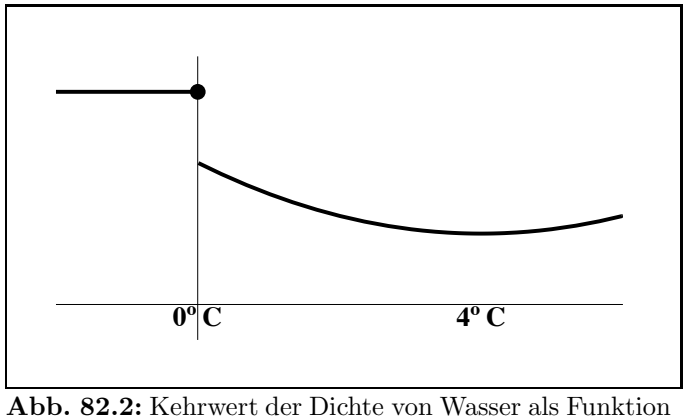

Abb. 82.2: Kehrwert der Dichte von Wasser als Funktion der Temperatur.

Wer die Unstetigkeit dieser Funktion handfest erleben möchte, möge eine randvolle Wasserflasche über Nacht in den Gefrierschrank legen und am nächsten Morgen nachschauen!

(f) Am 1. Januar 2006 war laut Preisverzeichnis der Deutschen Post das Porto y eines Inlandsbriefes (in Cent) als Funktion der Masse x des Briefes (in g) durch die folgende Vorschrift gegeben:

$$f(x) := \begin{cases} 55, & 0 < x \le 20 \text{ ("Standardbrief")},\\ 90, & 20 < x \le 50, \text{ ("Kompaktbrief")},\\ 145, & 50 < x \le 500 \text{ ("Großbrief")},\\ 220, & 500 < x \le 1000 \text{ ("Maxibrief")}. \end{cases}$$

Diese Funktion, aufgefaßt als Abbildung $f : (0,1000] \to \mathbb{R}$, ist unstetig an den Stellen 20, 50 und 500 und stetig an allen anderen Punkten des Definitionsbereichs.

(g) Eine von einer Sprungstelle verschiedene Art der Unstetigkeit offenbart die Funktion $f(x) = \sin(1/x)$, egal, wie wir diese Funktion an der Stelle 0 definieren.

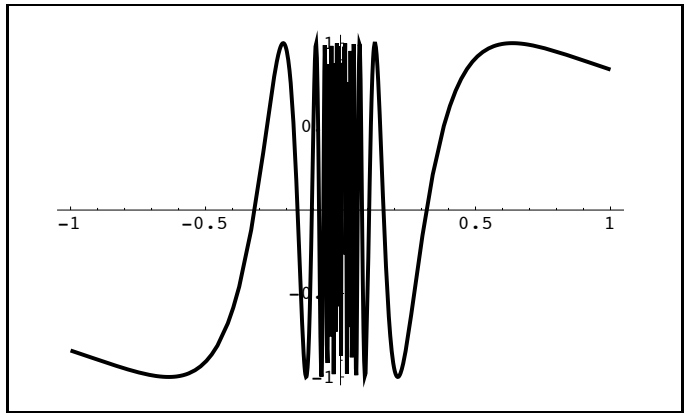

Abb. 82.3: Graph der Funktion $f(x) = \sin(1/x)$.

(h) Eine *lineare* Abbildung $T : V \to W$ zwischen normierten Räumen ist genau dann stetig, wenn sie beschränkt ist, wenn es also eine Konstante C gibt mit $\|Tv\| \le C\|v\|$ für alle $v \in V$. Gibt es eine solche Konstante, so gilt $\|Tv_1 - Tv_2\| = \|T(v_1-v_2)\| \le C\|v_1-v_2\|$, was die Stetigkeit von T zeigt (sogar die Lipschitzstetigkeit mit der Lipschitzkonstanten C). Ist umgekehrt T stetig, so gibt es zu $\varepsilon := 1$ eine Zahl $\delta > 0$ mit $\|Tv\| \le 1$ für alle $v \in V$ mit $\|v\| \le \delta$. Für alle $v \ne 0$ erfüllt dann $\widehat{v} := \delta\cdot v/\|v\|$ die Bedingung $\|\widehat{v}\| = \delta$, folglich $\|T\widehat{v}\| \le 1$; das bedeutet aber $\|Tv\| \le (1/\delta)\|v\|$. Diese letzte Ungleichung gilt natürlich auch für $v = 0$. Mit $C := 1/\delta$ gilt also $\|Tv\| \le C\|v\|$ für alle $v \in V$.

(i) Ist $A \ne \emptyset$ eine Teilmenge eines metrischen Raums (X,d), so ist die Abbildung $X \to \mathbb{R}$ mit $x \mapsto \operatorname{dist}(x,A)$ nach (81.5)(a) Lipschitz-stetig mit der Lipschitzkonstanten $L := 1$. (Dabei sei die Menge $\mathbb{R}$ mit ihrer natürlichen Metrik versehen.) ♦

Wir charakterisieren nun die Stetigkeit einer Abbildung mit Hilfe konvergenter Folgen. Vor dem Durchlesen des formalen Beweises des folgenden Satzes sollte man sich kurz plausibel machen, warum die Aussage dieses Satzes ebenso wie die Definition (82.1) der Stetigkeit die Idee ausdrückt, daß f genau dann stetig an der Stelle x ist, wenn Punkte, die "nahe" bei x liegen, auch Bildwerte haben, die "nahe" bei $f(x)$ liegen.

(82.4) Satz. *Es sei $f : X \to Y$ eine Abbildung zwischen metrischen Räumen. Genau dann ist f stetig an einer Stelle $x \in X$, wenn für jede Folge (x_n) in X mit $x_n \to x$ die Bedingung $f(x_n) \to f(x)$ in Y gilt.*

Beweis. Die Funktion f sei stetig an der Stelle x, und es sei (x_n) eine Folge mit $x_n \to x$; wir müssen zeigen, daß dann $f(x_n) \to f(x)$ gilt. Dazu geben wir uns ein $\varepsilon > 0$ beliebig vor. Es gibt dann ein $\delta > 0$ derart, daß aus $d(\xi, x) < \delta$ in X schon $d\big(f(\xi), f(x)\big) < \varepsilon$ in Y folgt, und wegen $x_n \to x$ gibt es ein $N \in \mathbb{N}$ mit $d(x_n, x) < \delta$ für $n \geq N$. Für alle $n \geq N$ gilt dann $d\big(f(x_n), f(x)\big) < \varepsilon$. Da $\varepsilon > 0$ beliebig war, ist damit $f(x_n) \to f(x)$ gezeigt.

Umgekehrt sei f *nicht* stetig an der Stelle x; wir müssen zeigen, daß es dann eine Folge (x_n) in X gibt mit $x_n \to x$, aber $f(x_n) \not\to f(x)$. Da f nicht stetig an der Stelle x ist, gibt es ein $\varepsilon > 0$, zu dem sich *kein* $\delta > 0$ finden läßt, für das $d(\xi, x) < \delta$ schon $d\big(f(\xi), f(x)\big) < \varepsilon$ impliziert. Für jede Zahl $\delta_n := 1/n$ gibt es also ein Element x_n in X mit $d(x_n, x) < 1/n$, aber $d\big(f(x_n), f(x)\big) \geq \varepsilon$. Die so gefundene Folge (x_n) erfüllt dann $x_n \to x$, aber $f(x_n) \not\to f(x)$. ∎

(82.5) Folgerung. *Alle Potenz-, Wurzel-, Exponential- und Logarithmusfunktionen sind stetig auf ihrem gesamten Definitionsbereich.*

Beweis. Dies folgt mit Hilfe der Folgencharakterisierung (82.4) der Stetigkeit aus (77.14). ∎

Als nächstes zeigen wir, daß jede analytische Funktion auch stetig ist.

(82.6) Satz. *Ist eine Funktion $f : \mathbb{K} \to \mathbb{K}$ in einem Punkt p analytisch, so ist sie in diesem Punkt auch stetig.*

Beweis. Es gelte $f(x) = \sum_{k=0}^{\infty} a_k (x-p)^k$ für $|x-p| < R$. Wählen wir eine Zahl $0 < c < R$, so gilt $\sum_{k=0}^{\infty} |a_k| c^k < \infty$; folglich ist auch $M := \sum_{k=1}^{\infty} |a_k| c^{k-1}$ eine endliche Zahl. Für alle x mit $|x-p| < c$ gilt dann

$$\begin{aligned} |f(x) - f(p)| &= \left| \sum_{k=1}^{\infty} a_k (x-p)^k \right| \\ &= \left| (x-p) \sum_{k=1}^{\infty} a_k (x-p)^{k-1} \right| \\ &\leq |x-p| \sum_{k=1}^{\infty} |a_k| \, |x-p|^{k-1} \\ &\leq |x-p| \sum_{k=1}^{\infty} |a_k| c^{k-1} = M|x-p|. \end{aligned}$$

Diese Abschätzung zeigt sofort, daß f an der Stelle $x = p$ (lokal Lipschitz-stetig und damit) stetig ist. ∎

Als Folgerung ergibt sich ein bemerkenswerter Eindeutigkeitssatz für analytische Funktionen, der besagt, daß eine analytische Funktion schon durch ihre Werte auf einer einzelnen konvergenten Folge festgelegt wird.

(82.7) Eindeutigkeitssatz für analytische Funktionen. *Die Funktionen $f(x) = \sum_{k=0}^{\infty} a_k (x-p)^k$ und $g(x) = \sum_{k=0}^{\infty} b_k (x-p)^k$ seien analytisch im Punkt p, und es sei (x_i) eine gegen p konvergente Folge. Gilt $f(x_i) = g(x_i)$ für alle i, so gilt $a_n = b_n$ für alle $n \in \mathbb{N}_0$.*

Beweis. Wir benutzen Induktion über n. Da f und g nach (82.6) im Punkt p stetig sind, haben wir $a_0 = f(p) = \lim_i f(x_i) = \lim_i g(x_i) = g(p) = b_0$. Ist nun schon gezeigt, daß $a_k = b_k$ für $0 \leq k \leq n-1$ gilt, so stimmen $\sum_{k=n}^{\infty} a_k (x-p)^k$ und $\sum_{k=n}^{\infty} b_k (x-p)^k$ an allen Folgengliedern x_i überein; dies gilt dann (nach Division durch $(x-p)^n$) auch für $F(x) := \sum_{k=n}^{\infty} a_k (x-p)^{k-n}$ und $G(x) := \sum_{k=n}^{\infty} b_k (x-p)^{k-n}$. Anwendung des Stetigkeitssatzes (82.6) auf F und G liefert dann $a_n = F(p) = \lim_i F(x_i) = \lim_i G(x_i) = G(p) = b_n$. ∎

Als nächste Klasse von Funktionen betrachten wir auf einem Intervall $I \subseteq \mathbb{R}$ definierte monotone Funktionen. Solche Funktionen müssen zwar nicht stetig sein, aber man kann für eine monotone Funktion die Menge der möglichen Unstetigkeitsstellen ziemlich genau charakterisieren. Um dies zu tun, geben wir zunächst die folgende Definition.

(82.8) Definition. *Es sei x_0 ein innerer Punkt eines Intervalls $I \subseteq \mathbb{R}$. Eine Funktion $f : I \to V$ hat eine* **Sprungstelle** *im Punkt x_0, wenn die links- und rechtsseitigen Grenzwerte*

$$\lim_{x \to x_0-} f(x) \quad \textit{und} \quad \lim_{x \to x_0+} f(x)$$

zwar beide existieren, aber entweder verschieden sind (in diesem Fall heißt x_0 eine **Unstetigkeitsstelle erster Art***) oder aber gleich sind, aber nicht mit dem Funktionswert $f(x_0)$ übereinstimmen (in diesem Fall sagt man, die Funktion f habe an der Stelle x_0 eine* **hebbare Unstetigkeit***). Existiert mindestens einer der beiden einseitigen Grenzwerte nicht, so nennt man x_0 eine* **Unstetigkeitsstelle zweiter Art**.

Der folgende Satz besagt nun, daß eine monotone Funktion höchstens abzählbar viele Unstetigkeitsstellen hat, und zwar ausschließlich Sprungstellen.

(82.9) Satz. *Eine monotone Funktion $f : I \to \mathbb{R}$ hat keine hebbaren Unstetigkeiten und keine Unstetigkeitsstellen zweiter Art und höchstens abzählbar viele Unstetigkeitsstellen erster Art.*

Beweis. O.B.d.A. sei f monoton wachsend; für monoton fallende Funktionen verläuft der Beweis völlig analog. Es sei $x \in I$ beliebig. Da f monoton wächst, existieren $f_-(x) = \sup_{y<x} f(y)$ und $f_+(x) = \inf_{y>x} f(y)$, und es gilt $f_-(x) \leq f(x) \leq f_+(x)$. Die einzig möglichen Unstetigkeitsstellen von f sind also Sprungstellen. Gilt ferner $x < y$ und ist ξ irgendeine Zahl mit $x < \xi < y$, so gilt

$$(\star) \qquad f_-(x) \leq f_+(x) \leq f(\xi) \leq f_-(y) \leq f_+(y),$$

so daß f_- und f_+ jeweils monoton wachsen. Schließlich sei X die Menge der Unstetigkeitsstellen von f. Dann sind die offenen Intervalle $I_x := \big(f_-(x), f_+(x)\big)$ nach $(\star)$ disjunkt. Da man in jedem dieser Intervalle eine rationale Zahl wählen kann und da es nur abzählbar viele rationale Zahlen gibt, kann es auch nur abzählbar viele solcher Intervalle geben; also ist die Menge X abzählbar. ∎

Wir beweisen nun, daß die Umkehrabbildung einer streng monotonen Funktion wieder stetig ist. Beispielsweise folgt die Stetigkeit der Logarithmus-, Wurzel- und Bogenfunktionen *automatisch* aus der Stetigkeit der Exponential-, Potenz- bzw. Winkelfunktionen.

(82.10) Umkehrsatz für streng monotone Funktionen. *Es seien $I \subseteq \mathbb{R}$ ein Intervall, $f : I \to \mathbb{R}$ eine streng monoton wachsende (fallende) Funktion und $J := f(I) := \{f(x) \mid x \in I\}$ das Bild von f. Dann ist die Umkehrfunktion $f^{-1} : J \to I$ ebenfalls streng monoton wachsend [fallend] und überdies stetig.*

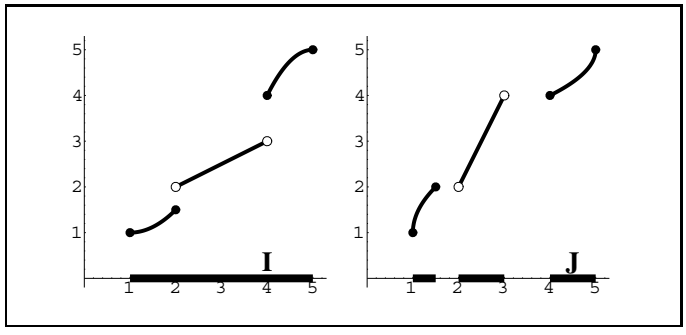

Abb. 82.4: Umkehrfunktion einer streng monotonen Funktion.

Beweis. Wir nehmen an, f sei monoton wachsend; für monoton fallende Funktionen verläuft der Beweis ganz analog. Die Injektivität von f folgt aus der strengen Monotonie; die Surjektivität erzwingen wir dadurch, daß wir f als Funktion $f : I \to J$ auffassen. Damit ist klar, daß die Umkehrabbildung $f^{-1} : J \to I$ existiert; ferner ist f^{-1} wegen der Äquivalenz $x_1 < x_2 \iff f(x_1) < f(x_2)$ selbst wieder streng monoton wachsend. Es bleibt also nachzuweisen, daß f^{-1} stetig ist.

Wir wollen die Stetigkeit von f^{-1} in einem beliebigen Punkt $y_0 \in J$ zeigen. Wir setzen $x_0 := f^{-1}(y_0)$ und geben uns ein beliebiges $\varepsilon > 0$ vor. Wegen der strengen Monotonie von f gilt dann $f(x_0 - \varepsilon) < y_0 < f(x_0 + \varepsilon)$.† Dann können wir ein $\delta > 0$ so wählen, daß auch $f(x_0 - \varepsilon) < y_0 - \delta < y_0 + \delta < f(x_0 + \varepsilon)$ gilt. Für alle $y \in J$ mit $|y - y_0| < \delta$ gilt dann $f(x_0-\varepsilon) < y < f(x_0+\varepsilon)$ und folglich $x_0 - \varepsilon < f^{-1}(y) < x_0 + \varepsilon$ bzw. $-\varepsilon < f^{-1}(y) - f^{-1}(y_0) < \varepsilon$ aufgrund der strengen Monotonie von f^{-1}. Aus $|y - y_0| < \delta$ folgt also $|f^{-1}(y) - f^{-1}(y_0)| < \varepsilon$; dies zeigt die Stetigkeit von f^{-1} im Punkt y_0. ∎

† Ist x_0 linker [rechter] Randpunkt von I, so lassen wir die linke [rechte] Ungleichung einfach weg und führen den folgenden Beweis mit leichten Modifikationen weiter. Wir nehmen hier an, daß x_0 kein Randpunkt von I ist; dann können wir ε immer so klein wählen, daß $x_0 \pm \varepsilon \in I$ gilt.

Wir kehren nun noch einmal zum Beginn dieses Abschnitts zurück, wo wir den Begriff der Stetigkeit einer Funktion f zwischen metrischen Räumen an einer Stelle x_0 dadurch einführten, daß wir zu jeder vorgegebenen Toleranz $\varepsilon > 0$ für den Zielwert $y_0 = f(x_0)$ die Existenz eines zulässigen Spiels $\delta > 0$ derart postulierten, daß eine Abweichung von weniger als δ bei der Einstellung von x_0 zu einer Abweichung von weniger als ε vom Sollzielwert $f(x_0)$ führt. Dabei wird das zulässige Spiel in der Praxis meist nicht nur von der Toleranz ε abhängen (je stringenter die Genauigkeitsanforderung, desto geringer das erlaubte Spiel), sondern auch vom Arbeitspunkt x_0 selbst. (Ein Rennfahrer, der sich einer Kurve mit einer Geschwindigkeit von 300 km/h annähert, kann sich viel geringere Fehler bei der Bewegung des Lenkrades erlauben als der Fahrer eines gewöhnlichen Fahrzeugs im Stadtverkehr.) Das Beispiel der Funktion $f(x) = 1/x$ illustriert diese Abhängigkeit des Spiels vom Arbeitspunkt; als Toleranz wurde $\varepsilon := 0.2$ gewählt, als Arbeitspunkt zunächst $x_0 = 1$, dann $x_0 = 0.5$.

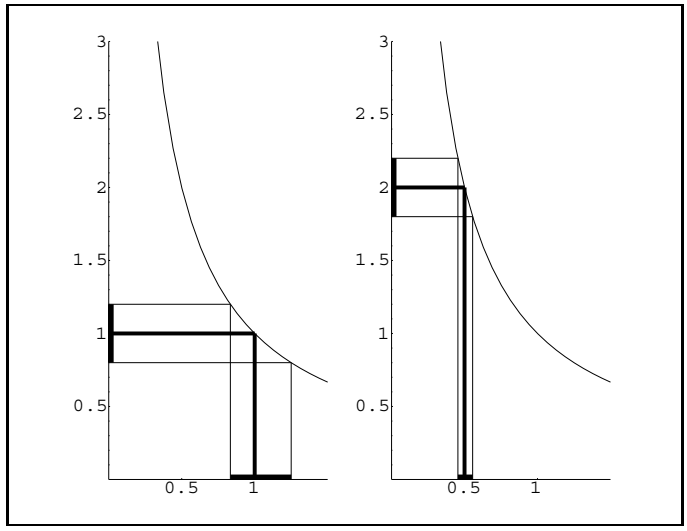

Abb. 82.5: Abhängigkeit des zulässigen Spiels δ vom Arbeitspunkt x_0 bei gleicher Toleranz ε.

Eine Funktion f heißt nun *gleichmäßig stetig* über einen Arbeitsbereich X, wenn zu jeder vorgegebenen erlaubten Toleranz $\varepsilon > 0$ ein (vom speziellen Arbeitspunkt unabhängiges) zulässiges Spiel $\delta > 0$ derart existiert, daß eine Abweichung von einem *beliebigen* Arbeitspunkt $x \in X$ um weniger als δ zu einer Abweichung vom Funktionswert $f(x)$ um weniger als ε führt.

(82.11) Definition. *Eine Abbildung $f : X \to Y$ zwischen metrischen Räumen heißt* **gleichmäßig stetig**, *wenn es zu jeder vorgegebenen Zahl $\varepsilon > 0$ eine Zahl $\delta > 0$ derart gibt, daß aus $d(x_1, x_2) < \delta$ in X stets $d\big(f(x_1), f(x_2)\big) < \varepsilon$ in Y folgt.*

Entscheidend ist, daß das Spiel δ allein in Abhängigkeit von der Toleranz ε (und unabhängig von irgendeinem festen Arbeitspunkt) bestimmt werden kann. Während also Stetigkeit eine lokale Eigenschaft ist (d.h., eine Eigenschaft, die einer Funktion in einem einzelnen Punkt zukommt und nur von den Werten der Funktion in einer beliebig kleinen Umgebung dieses Punktes abhängt), ist gleichmäßige Stetigkeit eine globale Eigenschaft (also eine Eigenschaft, die der Funktion insgesamt – unter Berücksichtigung ihres gesamten Definitionsbereichs – zukommt).

(82.12) Beispiele. (a) Jede Lipschitz-stetige Funktion ist gleichmäßig stetig. Gilt nämlich $d\big(f(x_1),f(x_2)\big) \le L\cdot d(x_1,x_2)$, so kann man zu gegebenem $\varepsilon>0$ stets $\delta := \varepsilon/L$ wählen, um die Implikation $\big(d(x_1,x_2)<\delta \Rightarrow d\big(f(x_1),f(x_2)\big)<\varepsilon\big)$ zu garantieren.

(b) Die durch $f(x)=\sqrt{x}$ definierte Funktion $f:[0,\infty)\to[0,\infty)$ ist gleichmäßig stetig; aus der für alle $x_1,x_2\ge 0$ gültigen Abschätzung

$$|\sqrt{x_1}-\sqrt{x_2}| \;\le\; \sqrt{|x_1-x_2|}$$

folgt nämlich sofort, daß die Wahl $\delta:=\varepsilon^2$ die Gültigkeit der Implikation $\big(|x_1-x_2|<\delta \Rightarrow |f(x_1)-f(x_2)|<\varepsilon\big)$ garantiert.

(c) Die Funktion $f(x)=x^2$ ist gleichmäßig stetig auf jedem Intervall $[-b,b]$ mit festem $b>0$, denn es gilt dann die Lipschitzbedingung

$$\begin{aligned}|f(x_1)-f(x_2)| \;&=\; |x_1^2-x_2^2| \;=\; |x_1+x_2|\,|x_1-x_2|\\ &\le \Big||x_1|+|x_2|\Big|\,|x_1-x_2| \;\le\; 2b|x_1-x_2|\,.\end{aligned}$$

Dagegen ist f nicht gleichmäßig stetig auf ganz $\mathbb{R}$; die Abschätzung

$$\begin{aligned}|f(x_1)-f(x_2)| \;&=\; |x_1+x_2|\,|x_1-x_2|\\ &=\; |2x_1+(x_2-x_1)|\,|x_1-x_2|\\ &\ge\; (2|x_1|-|x_2-x_1|)\,|x_2-x_1|\end{aligned}$$

zeigt nämlich, daß der Abstand $|f(x_1)-f(x_2)|$ beliebig groß werden kann, egal wie dicht x_1 und x_2 beisammen liegen, wenn nur x_1 groß genug gewählt wird.

(d) Die Funktion $f(x)=1/x$ ist gleichmäßig stetig auf jedem Intervall $[a,\infty)$ mit festem $a>0$, denn es gilt dann die Lipschitzbedingung

$$|f(x_1)-f(x_2)| = \left|\frac{1}{x_1}-\frac{1}{x_2}\right| = \frac{|x_1-x_2|}{|x_1|\,|x_2|} \le \frac{1}{a^2}|x_1-x_2|\,.$$

Dagegen ist f auf keinem Intervall der Form $(0,\varepsilon]$ gleichmäßig stetig; die Abschätzung

$$\begin{aligned}|f(x_1)-f(x_2)| \;&=\; \left|\frac{1}{x_1}-\frac{1}{x_2}\right| \;=\; \frac{|x_1-x_2|}{|x_1|\,|x_2|}\\ &\ge \frac{|x_1-x_2|}{|x_1|\,(|x_1-x_2|+|x_1|)}\end{aligned}$$

zeigt nämlich, daß der Abstand $|f(x_1)-f(x_2)|$ beliebig groß werden kann, egal wie dicht x_1 und x_2 beisammen liegen, wenn nur x_1 nahe genug bei 0 liegt.

(e) Sind $f:X\to Y$ und $g:Y\to Z$ zwei gleichmäßig stetige Abbildungen zwischen metrischen Räumen, so ist deren Verkettung $g\circ f: X\to Z$ ebenfalls gleichmäßig stetig. Zu jedem vorgegeben $\varepsilon>0$ gibt es nämlich aufgrund der gleichmäßigen Stetigkeit von g ein $\delta_Y>0$ derart, daß aus $d_Y(y_1,y_2)<\delta_Y$ stets $d_Z\big(g(y_1),g(y_2)\big)<\varepsilon$ folgt; wegen der gleichmäßigen Stetigkeit von f gibt es dann ein $\delta_X>0$ derart, daß aus $d_X(x_1,x_2)<\delta_X$ stets $d_Y\big(f(x_1),f(x_2)\big)<\delta_Y$ folgt. Für alle $x_1,x_2\in X$ mit $d_X(x_1,x_2)<\delta_X$ gilt dann $d_Z\big(g\big(f(x_1)\big),g\big(f(x_2)\big)\big)<\varepsilon$, also $d_Z\big((g\circ f)(x_1),(g\circ f)(x_2)\big)<\varepsilon$. ♦

Wir definieren zum Abschluß dieses Abschnitts noch diejenigen Abbildungen, die metrische Strukturen auf verschiedenen Räumen miteinander identifizieren.

(82.13) Definition. *Eine Abbildung $f:X\to Y$ zwischen metrischen Räumen heißt* **Isometrie** *oder auch* **isometrische Einbettung**, *wenn f abstandserhaltend ist, wenn also $d_Y\big(f(a),f(b)\big)=d_X(a,b)$ für alle $a,b\in X$ gilt. Eine bijektive Isometrie heißt auch* **isometrischer Isomorphismus.**

Eine Isometrie ist stets injektiv, denn aus $f(a)=f(b)$ folgt $0=d_Y\big(f(a),f(b)\big)=d_X(a,b)$ und damit $a=b$. Ist $f:X\to Y$ ein isometrischer Isomorphismus, so ist auch die Umkehrabbildung f^{-1} eine Isometrie. Existiert ein isometrischer Isomorphismus zwischen zwei metrischen Räumen, so sind deren metrische Strukturen ununterscheidbar; jede durch die Metrik ausdrückbare Aussage (etwa hinsichtlich des Abstandes zwischen Punkten oder Teilmengen, der Konvergenz von Folgen oder der Beschränktheit von Mengen) gilt in dem einen Raum genau dann, wenn die entsprechende Aussage in dem anderen Raum gilt. Ein isometrischer Isomorphismus $f:X\to Y$ läßt sich einfach als Umbenennung der Elemente von X auffassen, und (X,d_X) und (Y,d_Y) stellen nur verschiedene Realisierungen des "gleichen" metrischen Raums dar. Das folgende Beispiel zeigt, daß sich jeder metrische Raum isometrisch in einen Funktionenraum einbetten läßt, also isometrisch isomorph zu einem Unterraum eines Funktionenraums ist.

(82.14) Beispiel. Es seien (X,d) ein beliebiger metrischer Raum und $\mathfrak{F}$ die Menge aller Funktionen $f:X\to\mathbb{R}$, die für alle $x,y\in X$ die folgende Bedingung erfüllen:

$$|f(x)-f(y)| \;\le\; d(x,y) \;\le\; f(x)+f(y)\,;$$

dann ist eine Metrik auf $\mathfrak{F}$ definiert durch $D(f,g):=\sup_{x\in X}|f(x)-f(y)|$. Für jedes Element $a\in X$ ist nun eine Funktion $f_a\in\mathfrak{F}$ gegeben durch $f_a(x):=d(x,a)$, und es gilt $D(f_a,f_b)=d(a,b)$ für alle $a,b\in X$. Durch $a\mapsto f_a$ ist also eine Isometrie von X auf einen Unterraum von $\mathfrak{F}$ definiert. (Die Einzelheiten dieses Beispiels sind als Übungsaufgabe zu überprüfen!)

83. Vollständigkeit metrischer Räume

Wir definieren die Vollständigkeit eines metrischen Raums in völliger Analogie zur Definition (73.11) der metrischen Vollständigkeit der Zahlengeraden.

(83.1) Definition. *Ein metrischer Raum heißt* **vollständig***, wenn in ihm jede Cauchyfolge konvergiert.*

Die Bedeutung dieses Begriffes ergibt sich aus dem in vielen praktisch wichtigen Fällen zu beobachtenden Auftreten einer Situation, die etwa folgendermaßen beschrieben werden kann. Man ist daran interessiert, eine gewisse Größe zu ermitteln (Länge einer Kurve, Flächeninhalt einer geometrischen Figur, Nullstelle einer Funktion, Lösung einer irgendwie gearteten Gleichung), kann dies aber nicht direkt tun. Stattdessen ist man in der Lage, Näherungswerte für die gesuchte Größe zu ermitteln und diese sukzessive zu verbessern; man erhält dann eine Folge $(x_1, x_2, x_3, \ldots)$ von Näherungswerten für die eigentlich gesuchte Größe x, und es ist naheliegend zu versuchen, x als Grenzwert der Folge (x_n) zu gewinnen. Wenn nun die Näherungswerte x_n tatsächlich immer bessere Approximationen sind, so werden sie eine Cauchyfolge bilden (was ja gerade heißt, daß für genügend große Indices m und n die Werte x_m und x_n beliebig dicht beieinander liegen), und es ist genau die Eigenschaft der Vollständigkeit, die in dieser Situation garantiert, daß die Folge (x_n) tatsächlich einen Grenzwert besitzt (der dann die Lösung des gesuchten Problems darstellt).

(83.2) Beispiele. (a) Ist $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$, so ist $\mathbb{K}$ mit der natürlichen Metrik $d(x,y) = |x-y|$ vollständig; dagegen ist $\mathbb{Q}$ nicht vollständig.

(b) Der Raum $\mathbb{K}^n$ mit der von der Norm $\|x\| := \sqrt{\sum_{i=1}^n |x_i|^2}$ induzierten Metrik ist vollständig. Ist nämlich $(x^{(k)})_{k=1}^\infty$ eine Cauchyfolge in $\mathbb{K}^n$, so ist für jeden Index $1 \le i \le n$ die Folge $(x_i^{(k)})_{k=1}^\infty$ eine Cauchyfolge in $\mathbb{K}$, folglich wegen der Vollständigkeit von $\mathbb{K}$ konvergent gegen ein Element $x_i \in \mathbb{K}$. Dann konvergiert aber die Folge $(x^{(k)})$ in $\mathbb{K}^n$ gegen das Element $x := (x_1, \ldots, x_n)^T$; vgl. (81.14)(a).

(c) Es seien X eine beliebige Menge, (Y,d) ein vollständiger metrischer Raum und $\mathfrak{B}(X,Y)$ der Raum aller beschränkten Funktionen $f : X \to Y$, versehen mit der Metrik $D(f,g) := \sup_{x\in X} d\big(f(x),g(x)\big)$; wir wollen zeigen, daß $\mathfrak{B}(X,Y)$ vollständig ist, und betrachten eine beliebige Cauchyfolge (f_n) in $\mathfrak{B}(X,Y)$. Dann ist für jedes feste $x \in X$ die Folge $\big(f_n(x)\big)$ eine Cauchyfolge in Y, wegen der Vollständigkeit von Y also konvergent gegen ein Element $f(x) \in Y$. Die Folge (f_n) konvergiert dann punktweise gegen die so definierte Funktion $f : X \to Y$.

Wir zeigen zunächst, daß die Abbildung f beschränkt und damit ein Element von $\mathfrak{B}(X,Y)$ ist. Es seien $a, b \in X$ beliebige Elemente von X. Wegen $f_n(a) \to f(a)$ und $f_n(b) \to f(b)$ für $n \to \infty$ gibt es zu $\varepsilon := 1$ ein $m \in \mathbb{N}$ mit $d\big(f(a), f_n(a)\big) \le 1$ und $d\big(f(b), f_n(b)\big) \le 1$ für alle $n \ge m$. Da f_m eine beschränkte Funktion ist, erhalten wir dann die Abschätzung

$$\begin{aligned} &d\big(f(a), f(b)\big) \\ &\le d\big(f(a), f_m(a)\big) + d\big(f_m(a), f_m(b)\big) + d\big(f_m(b), f(b)\big) \\ &\le 1 + \operatorname{diam}\big(\operatorname{Bild}(f_m)\big) + 1 \;=\; 2 + \operatorname{diam}\big(\operatorname{Bild}(f_m)\big); \end{aligned}$$

da $a, b \in X$ beliebig gewählt waren, gilt also die Abschätzung $\operatorname{diam}\big(\operatorname{Bild}(f)\big) \le 2 + \operatorname{diam}\big(\operatorname{Bild}(f_m)\big) < \infty$, die zeigt, daß die Funktion f beschränkt ist.

Wir wollen weiter zeigen, daß für $n \to \infty$ nicht nur $f_n \to f$ punktweise gilt, sondern sogar $D(f_n, f) \to 0$; ist dies geschafft, so ist $\mathfrak{B}(X,Y)$ als vollständig nachgewiesen. Es sei $\varepsilon > 0$ beliebig vorgegeben. Da (f_n) eine Cauchyfolge ist, gibt es einen Index $N \in \mathbb{N}$ mit $D(f_m, f_n) \le \varepsilon$ für alle $m, n \ge N$ und damit $d\big(f_m(x), f_n(x)\big) \le \varepsilon$ für alle $m, n \ge N$ und jedes fest gewählte Element $x \in X$. Für $m \to \infty$ folgt dann $d\big(f(x), f_n(x)\big) \le \varepsilon$ für alle $n \ge N$ und jedes fest gewählte $x \in X$, folglich $D(f, f_n) \le \varepsilon$ für alle $n \ge N$. Da $\varepsilon > 0$ beliebig war, ist damit die Konvergenzaussage $D(f_n, f) \to 0$ gezeigt.

(d) Ein metrischer Raum, dessen Metrik nur ganzzahlige Werte annehmen kann, ist automatisch vollständig. Eine Folge (x_n) in einem solchen Raum ist nämlich genau dann eine Cauchyfolge, wenn sie "schließlich konstant" ist, wenn es also einen Index $n_0 \in \mathbb{N}$ gibt mit $x_n = x_{n_0}$ für alle $n \ge n_0$, und eine solche Folge ist natürlich konvergent. ♦

Wir wollen nun herausfinden, wann ein Unterraum eines vollständigen metrischen Raums selbst wieder vollständig ist (was uns eine ganze Reihe weiterer Beispiele für vollständige metrische Räume liefern wird). Dazu führen wir zunächst den Begriff der *Abgeschlossenheit* einer Teilmenge eines metrischen Raums ein.

(83.3) Definition. *Es seien (X,d) ein metrischer Raum und A eine Teilmenge von X. Der* **Abschluß** *$\overline{A}$ von A ist die Menge aller derjenigen Elemente von X, die sich als Grenzwert einer Folge von Elementen in A darstellen lassen. Die Elemente von $\overline{A}$ heißen* **Berührpunkte** *von A. Die Menge A heißt* **abgeschlossen** *in X, wenn $\overline{A} = A$ gilt. Die Menge A heißt* **dicht** *in X, wenn $\overline{A} = X$ gilt.*

Wir beobachten, daß stets $A \subseteq \overline{A}$ gilt, denn jedes Element $a \in A$ läßt sich ja als Grenzwert der konstanten Folge $(a, a, a, \ldots)$ in A auffassen. Der folgende Satz gibt nun die gewünschte Charakterisierung der vollständigen Unterräume eines vollständigen metrischen Raums.

(83.4) Satz. *Es sei (X,d) ein vollständiger metrischer Raum. Ein Teilraum $A \subseteq X$ von X (mit der von d induzierten Metrik) ist genau dann vollständig, wenn er abgeschlossen ist.*

Beweis. Wir nehmen an, A sei abgeschlossen. Ist (a_n) eine Cauchyfolge in A, dann auch eine Cauchyfolge in X, wegen der vorausgesetzten Vollständigkeit von X also konvergent in X, sagen wir $a_n \to x$ mit $x \in X$. Da A abgeschlossen ist, muß der Grenzwert in A liegen; sagen wir $x = a \in A$. Dann gilt aber $a_n \to a$ in A. Damit ist gezeigt, daß jede Cauchyfolge in A konvergiert; folglich ist A vollständig.

Differentialrechnung in einer Variablen

89. Ableitungsbegriff und Ableitungsregeln

Ausgangspunkt der Differentialrechnung ist die Frage nach der lokalen Änderungsrate einer Funktion an einer gegebenen Stelle ihres Definitionsbereichs. Bevor wir diese Frage präzisieren, betrachten wir ein einfaches Beispiel.

(89.1) Beispiel. Die Körpertemperatur eines Patienten A betrage 41.6^0C, die eines zweiten Patienten B dagegen 41.4^0C. Kennt man nur diese Zahlen, so wird man vermuten, daß A in kritischerem Zustand sei als B. Als zusätzliche Information seien nun aber die Fieberkurven beider Patienten während der vorausgegangenen drei Tage (dargestellt in dem folgenden Diagramm) bekannt. Jetzt wird deutlich, daß A zwar (noch) eine höhere Temperatur hat als B, daß diese Temperatur aber abnehmende Tendenz hat, während diejenige von B stark ansteigt; der Zustand von B ist also zum gegenwärtigen Zeitpunkt weit besorgniserregender als der von A.

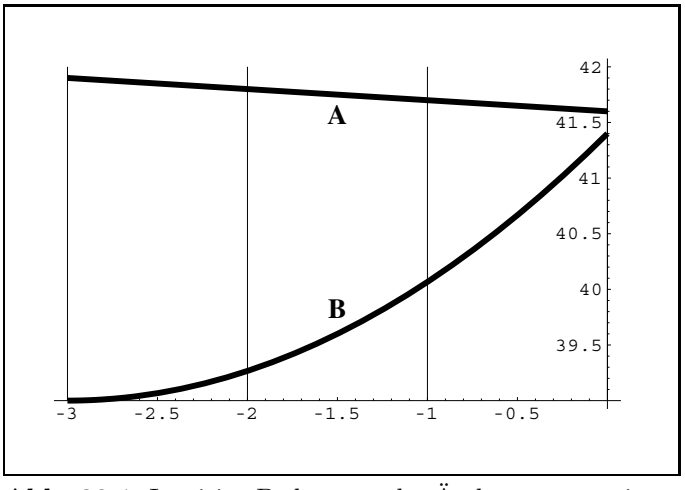

Abb. 89.1: Intuitive Bedeutung der Änderungsrate einer Funktion.

Dieses Beispiel zeigt, daß neben dem Wert $f(x_0)$ einer Funktion an einer bestimmten Stelle x_0 ihres Definitionsbereichs auch die "Änderungstendenz" oder "Änderungsrate" von f an dieser Stelle von Interesse ist. (Wir haben hier zunächst reellwertige Funktionen im Auge, werden aber sehen, daß sich unsere Überlegungen problemlos auf vektorwertige Funktionen verallgemeinern lassen.) Um diese "Änderungsrate" quantitativ zu erfassen, fragen wir, wie stark sich der Funktionswert $f(x_0)$ ändert, wenn das Argument x_0 ein klein wenig verändert wird. Eine solche Änderung Δx im Argument führt zu einer Änderung $\Delta f = f(x_0+\Delta x)-f(x_0)$ im Funktionswert, und der Quotient

$$\frac{\Delta f}{\Delta x} \;=\; \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$$

drückt die Änderung im Funktionswert relativ zur Änderung im Argument der Funktion aus. Konvergiert dieser Quotient für $\Delta x \to 0$ gegen einen Grenzwert, so können wir diesen als lokale Änderungsrate der Funktion f an der Stelle x_0 interpretieren.

(89.2) Definition. *Es seien $I \subseteq \mathbb{R}$ ein offenes Intervall und $f : I \to \mathbb{R}$ eine Funktion. Die Funktion f heißt* **differentiierbar**† *an der Stelle $x_0 \in I$, falls der Grenzwert*

$$f'(x_0) \;:=\; \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0} \;=\; \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h}$$

existiert; dieser heißt dann die **Ableitung** *von f an der Stelle x_0. Existiert $f'(x)$ an jedem Punkt $x \in I$, so heißt die Funktion $f' : I \to \mathbb{R}$ die* **Ableitungsfunktion** *von f.*

Zuweilen ist es nützlich, bei der Bestimmung der Änderungsrate einer Funktion sich nur von einer Seite her an die interessierende Stelle x_0 anzunähern; dies führt auf den Begriff einseitiger Ableitungen, die folgendermaßen definiert werden.

(89.3) Definition. *Es sei $x_0 \in \mathbb{R}$ eine reelle Zahl.*

(a) *Es sei $f : [x_0, b) \to \mathbb{R}$ eine Funktion. Existiert der folgende Grenzwert, so wird dieser als* **rechtsseitige Ableitung** *von f an der Stelle x_0 bezeichnet:*

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

(b) *Es sei $f : (a, x_0] \to \mathbb{R}$ eine Funktion. Existiert der folgende Grenzwert, so wird dieser als* **linksseitige Ableitung** *von f an der Stelle x_0 bezeichnet:*

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

Offenbar ist eine Funktion f genau dann an einem inneren Punkt x_0 ihres Definitionsbereichs differentiierbar, wenn sowohl $f'_+(x_0)$ als auch $f'_-(x_0)$ existieren und übereinstimmen; es ist dann $f'(x_0) = f'_+(x_0) = f'_-(x_0)$. Es folgt eine geometrische Deutung des Ableitungsbegriffs.

(89.4) Deutung. Für eine Funktion $f : \mathbb{R} \to \mathbb{R}$ können wir die Ableitung $f'(x_0)$ folgendermaßen als Tangentensteigung deuten. Die *Tangente* an die Kurve $y = f(x)$ im Punkt $\big(x_0, f(x_0)\big)$ ist die Grenzlage von *Sekanten* durch den Punkt $\big(x_0, f(x_0)\big)$ und benachbarte Punkte $\big(x, f(x)\big)$, wobei x beliebig nahe an x_0 heranrückt. Da die Steigung der Sekante durch die Punkte $\big(x_0, f(x_0)\big)$ und $\big(x, f(x)\big)$ den Wert $\big(f(x)-f(x_0)\big)/(x-x_0)$ hat, ist die Steigung der gesuchten Tangente gegeben durch den Grenzwert $\lim_{x\to x_0}\big(f(x)-f(x_0)\big)/(x-x_0) = f'(x_0)$; vorausgesetzt, dieser Grenzwert existiert überhaupt. Die gesuchte

† Wir benutzen statt der üblichen Schreibweise "differenzierbar" die altmodische, aber sprachlich korrektere Schreibweise "differentiierbar".

Tangente hat dann die Steigung $f'(x_0)$ und muß durch $\big(x_0, f(x_0)\big)$ gehen, ist also gegeben durch die Gleichung $y = f'(x_0)(x - x_0) + y_0$.

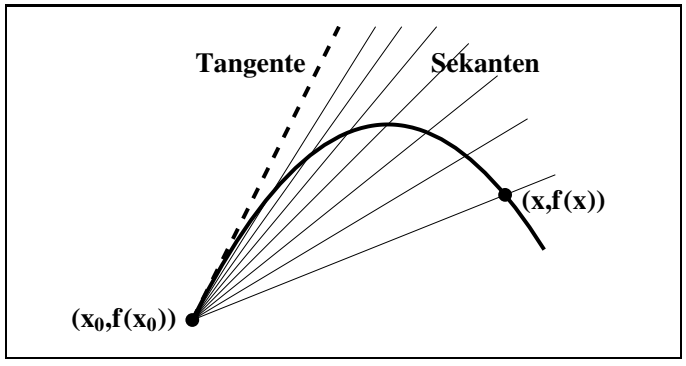

Abb. 89.2: Tangente als Grenzlage von Sekanten.

Um mit Ableitungen sinnvoll operieren zu können, müssen wir sie berechnen können; wir zeigen daher zunächst, wie sich für einige wichtige Funktionen die Ableitungen "per Hand" berechnen lassen.

(89.5) Beispiel: Potenzfunktionen. Die Funktion $f(x) = x^n$ ist auf ganz $\mathbb{R}$ differentiierbar, und es gilt $f'(x) = nx^{n-1}$. Nach der binomischen Formel ist nämlich

$$\begin{aligned}\frac{f(x+h)-f(x)}{h} &= \frac{(x+h)^n - x^n}{h}\\ = \binom{n}{1}x^{n-1} + \binom{n}{2}x^{n-2}h + \cdots + \binom{n}{n}h^{n-1},\end{aligned}$$

und dieser Ausdruck strebt für $h \to 0$ gegen nx^{n-1}. ⧫

(89.6) Beispiel: Kehrwertabbildung. Die Funktion $f(x) = 1/x$ ist an jeder Stelle $x \neq 0$ ihres Definitionsbereichs differentiierbar, und es gilt $f'(x) = -1/x^2$. Dies folgt aus den Umformungen

$$\begin{aligned}\frac{f(x+h)-f(x)}{h} &= \frac{1}{h}\left(\frac{1}{x+h} - \frac{1}{x}\right)\\ = \frac{x-(x+h)}{h(x+h)x} &= \frac{-h}{h(x+h)x} = \frac{-1}{x(x+h)},\end{aligned}$$

denn die rechte Seite geht für $h \to 0$ offensichtlich gegen $-1/x^2$. ⧫

(89.7) Beispiel: Wurzelfunktion. Die Funktion $f(x) := \sqrt{x}$ ist an jeder Stelle $x > 0$ differentiierbar mit $f'(x) = 1/(2\sqrt{x})$, denn durch Erweitern mit $\sqrt{x+h} + \sqrt{x}$ erhalten wir

$$\frac{\sqrt{x+h}-\sqrt{x}}{h} = \frac{(x+h)-x}{h\cdot\big(\sqrt{x+h}+\sqrt{x}\big)} = \frac{1}{\sqrt{x+h}+\sqrt{x}},$$

und dieser Ausdruck strebt für $h \to 0$ gegen $1/(2\sqrt{x})$. An der Stelle $x = 0$ existiert dagegen der (rechtsseitige) Grenzwert für $h \to 0$ nicht; es gilt $\big(f(h) - f(0)\big)/h = 1/\sqrt{h} \to \infty$ für $h \to 0+$ (was man so deuten kann, daß die Wurzelfunktion an der Stelle $x = 0$ eine "unendlich große Steigung" hat, ihr Graph also eine vertikale Tangente). ⧫

(89.8) Beispiel: Logarithmusfunktionen. Die Funktion $f(x) = \log_b x$ ist auf ganz $(0, \infty)$ differentiierbar, und es gilt $f'(x) = (\log_b e)/x$. Es gilt nämlich

$$\begin{aligned}\frac{\log_b(x+h)-\log_b(x)}{h} &= \frac{\log_b(1+h/x)}{h}\\ = \frac{1}{x}\cdot\frac{x}{h}\log_b(1+\frac{h}{x}) &= \frac{1}{x}\cdot\log_b(1+\frac{h}{x})^{x/h},\end{aligned}$$

und dieser Ausdruck strebt für $h \to 0$ wegen (78.11) und (77.14) gegen $(1/x)\cdot\log_b e$. ⧫

(89.9) Beispiel: Sinus- und Kosinusfunktion. Wir beweisen zunächst die Grenzwertbeziehungen

$$\lim_{h\to 0}\frac{\sin h}{h} = 1 \quad\text{und}\quad \lim_{h\to 0}\frac{\cos h - 1}{h} = 0,$$

welche besagen, daß die Sinus- und die Kosinusfunktion an der Stelle 0 differentiierbar sind mit $\sin'(0) = 1$ und $\cos'(0) = 0$. Zum Beweis der ersten Aussage beachten wir, daß für $0 < h < \pi/4$ aufgrund der Abschätzungen in (79.2) die Beziehung $\sin(h) < h < \tan(h)$ gilt, nach Division durch $\sin(h)$ also $1 < h/\sin(h) < 1/\cos(h)$, nach Kehrwertbildung daher $1 > \sin(h)/h > \cos(h)$.

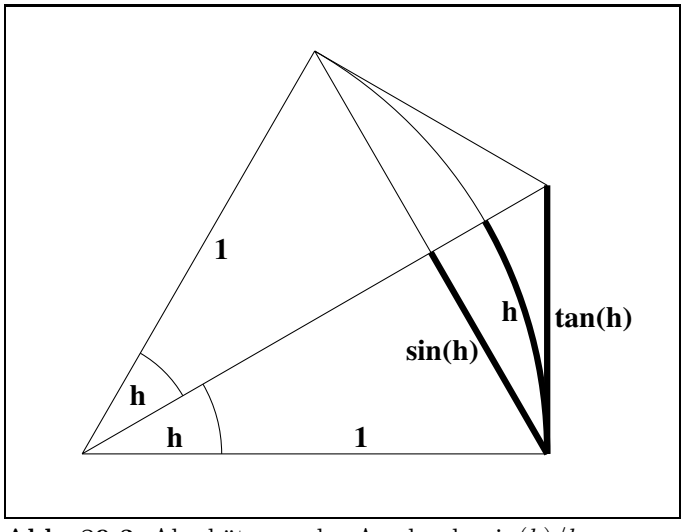

Abb. 89.3: Abschätzung des Ausdrucks $\sin(h)/h$.

Da sich diese Ungleichung nicht ändert, wenn h durch $-h$ ersetzt wird, folgt hieraus, daß $\cos(h) < \sin(h)/h < 1$ für $0 < |h| < \pi/4$ gilt; für $h \to 0$ folgt hieraus die erste Behauptung nach dem Einschnürungskriterium. Die zweite Aussage wird nun mit Hilfe der Additionstheoreme auf die erste zurückgeführt; es gilt nämlich

$$\begin{aligned}\frac{\cos h - 1}{h} &= \frac{\cos h - \cos 0}{h} = \frac{-2\sin^2(h/2)}{h}\\ &= -\frac{\sin^2(h/2)}{h/2} = -\big(\frac{\sin(h/2)}{h/2}\big)^2\cdot\frac{h}{2},\end{aligned}$$

und dieser Ausdruck geht für $h \to 0$ gegen Null. An einer beliebigen Stelle $x \in \mathbb{R}$ erhalten wir aufgrund des Additionstheorems für die Sinusfunktion

$$\frac{\sin(x+h)-\sin(x)}{h} = \frac{\sin x\cos h+\cos x\sin h-\sin x}{h}$$
$$= \sin x\cdot\frac{\cos h-1}{h}+\cos x\cdot\frac{\sin h}{h}\to\sin x\cdot 0+\cos x\cdot 1$$

für $h \to 0$, für die Kosinusfunktion dagegen

$$\frac{\cos(x+h)-\cos(x)}{h} = \frac{\cos x\cos h-\sin x\sin h-\cos x}{h}$$
$$= \cos x\cdot\frac{\cos h-1}{h}-\sin x\cdot\frac{\sin h}{h}\to\cos x\cdot 0-\sin x\cdot 1$$

für $h \to 0$. Die Aussagen $\sin'(x) = \cos(x)$ und $\cos'(x) = -\sin(x)$ an einer beliebigen Stelle x lassen sich also mit Hilfe der Additionstheoreme auf die entsprechenden Aussagen an der Stelle $x = 0$ zurückführen. ♦

(89.10) Beispiel: Exponentialfunktion. Wir behaupten, daß die Exponentialfunktion ihre eigene Ableitung ist, daß also $\exp' = \exp$ gilt. Zum Nachweis beweisen wir zunächst, daß die Ableitung $\exp'(0)$ an der Stelle 0 existiert und den Wert 1 hat. Für alle $h \in \mathbb{R}$ und alle $n \in \mathbb{N}$ gilt

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

Wegen

$$\binom{n}{k}\cdot\frac{1}{n^k} = \frac{1}{k!}\cdot\underbrace{\frac{n-k+1}{n}}_{\le 1}\cdot\underbrace{\frac{n-k+2}{n}}_{\le 1}\cdots\underbrace{\frac{n-k+k}{n}}_{\le 1}\le\frac{1}{k!}$$

erhalten wir für $|h| \le 1$ also

$$\left|\frac{(1+h/n)^n-1}{h}-1\right| \le \sum_{k=2}^{n}\binom{n}{k}\frac{|h|^{k-1}}{n^k}$$
$$\le \sum_{k=2}^{n}\frac{1}{k!}|h|^{k-1} \le (\sum_{k=2}^{\infty}\frac{1}{k!})\cdot|h|.$$

(Die Bedingung $|h| \le 1$ wurde nur in der letzten Ungleichung benutzt.) Mit $n \to \infty$ folgt hieraus zunächst

$$\left|\frac{\exp(h)-1}{h}-1\right| \le (\sum_{k=2}^{\infty}\frac{1}{k!})\cdot|h|,$$

und mit $h \to 0$ folgt dann $\lim_{h\to 0}\big(\exp(h)-1\big)/h = 1$. An einer beliebigen Stelle $x \in \mathbb{R}$ gilt aufgrund der Funktionalgleichung der Exponentialfunktion dann

$$\exp'(x) = \lim_{h\to 0}\frac{\exp(x+h)-\exp(x)}{h}$$
$$= \exp(x)\cdot\lim_{h\to 0}\frac{\exp(h)-1}{h};$$

die Aussage $\exp'(x) = \exp(x)$ an einer beliebigen Stelle x läßt sich also mit Hilfe der Funktionalgleichung der Exponentialfunktion auf die entsprechende Aussage an der Stelle $x = 0$ zurückführen. ♦

(89.11) Gegenbeispiel: Betragsfunktion. Die Betragsfunktion $f(x) := |x|$ ist nicht differentiierbar an der Stelle $x_0 := 0$, denn es gilt

$$\frac{f(x)-f(x_0)}{x-x_0} = \frac{|x|-|0|}{x-0} = \frac{|x|}{x} = \begin{cases} 1, & \text{falls } x>0,\\ -1, & \text{falls } x<0.\end{cases}$$

Der Grenzwert dieses Ausdrucks für $x \to 0$ existiert offensichtlich nicht. ♦

Auch wenn eine Funktion f an einer Stelle x_0 eine Sprungstelle hat, existiert dort die Ableitung nicht; dies folgt unmittelbar aus dem folgenden Ergebnis.

(89.12) Notiz. *Ist f differentiierbar an der Stelle x_0, dann auch stetig.*

Beweis. Aus $x \to x_0$ folgt $f(x) \to f(x_0)$, denn

$$f(x)-f(x_0) = \frac{f(x)-f(x_0)}{x-x_0}\cdot(x-x_0) \to f'(x_0)\cdot 0 = 0.$$

■

Für einige wenige Funktionen haben wir oben die Ableitung direkt nach der Definition berechnet. Wir werden nun Regeln herleiten, mit deren Hilfe wir die Ableitungen kompliziert aufgebauter Funktionen aus den Ableitungen ihrer Einzelbestandteile berechnen können. Dies wird uns in die Lage versetzen, die Ableitung jeder beliebigen aus elementaren Funktionen zusammengesetzten Funktion zu bestimmen.

(89.13) Ableitungsregeln. *Es seien $f, g : \mathbb{R} \to \mathbb{R}$ reelle Funktionen und $c \in \mathbb{R}$ eine Konstante.*

(a) **Faktorregel:** *Ist f differentiierbar in x_0, dann auch cf, und es gilt*

$$(cf)'(x_0) = c\cdot f'(x_0).$$

(b) **Summenregel:** *Sind f und g differentiierbar in x_0, dann auch $f+g$, und es gilt*

$$(f+g)'(x_0) = f'(x_0)+g'(x_0).$$

(c) **Produktregel:** *Sind f und g differentiierbar in x_0, dann auch fg, und es gilt*

$$(fg)'(x_0) = f'(x_0)g(x_0)+f(x_0)g'(x_0).$$

(d) **Quotientenregel:** *Sind f und g differentiierbar in x_0 und gilt $g(x_0) \neq 0$, so ist auch f/g differentiierbar in x_0, und es gilt*

$$(\frac{f}{g})'(x_0) = \frac{f'(x_0)g(x_0)-f(x_0)g'(x_0)}{g(x_0)^2}.$$

(e) **Kettenregel:** *Ist f differentiierbar in x_0 und g differentiierbar in $f(x_0)$, so ist die Verkettung $g \circ f$ differentiierbar in x_0 mit*

$$(g \circ f)'(x_0) \;=\; g'\big(f(x_0)\big) f'(x_0)\,.$$

(f) **Umkehrregel:** *Ist f differentiierbar in x_0 mit $f'(x_0) \neq 0$ und besitzt f eine Umkehrfunktion f^{-1} in einer Umgebung von $y_0 := f(x_0)$, so ist f^{-1} differentiierbar an der Stelle y_0, und es gilt*

$$(f^{-1})'(y_0) \;=\; \frac{1}{f'(x_0)} \;=\; \frac{1}{f'\big(f^{-1}(y_0)\big)}\,.$$

Faktorregel: $(cf)' = c \cdot f'$
Summenregel: $(f+g)' = f' + g'$
Produktregel: $(fg)' = f'g + fg'$
Quotientenregel: $(f/g)' = (f'g - fg')/g^2$
Kettenregel: $(g \circ f)' = (g' \circ f) \cdot f'$
Umkehrregel: $(f^{-1})' = 1/(f' \circ f^{-1})$

Beweis. Zum Nachweis von (a) schreiben wir

$$\frac{(cf)(x_0+h) - (cf)(x_0)}{h} \;=\; c \cdot \frac{f(x_0+h) - f(x_0)}{h}$$

und erkennen, daß dieser Ausdruck für $h \to 0$ gegen $c \cdot f'(x_0)$ strebt. Zum Nachweis von (b) schreiben wir

$$\frac{(f+g)(x_0+h) - (f+g)(x_0)}{h}$$
$$= \frac{f(x_0+h) - f(x_0)}{h} + \frac{g(x_0+h) - g(x_0)}{h}$$

und erkennen, daß dieser Ausdruck für $h \to 0$ gegen $f'(x_0)+g'(x_0)$ strebt. Zum Nachweis von (c) schreiben wir den Differenzenquotienten $\big((fg)(x_0+h) - (fg)(x_0)\big)/h$ in der Form

$$\frac{f(x_0+h) - f(x_0)}{h} \cdot g(x_0+h) + f(x_0) \cdot \frac{g(x_0+h) - g(x_0)}{h}$$

und erkennen, daß dieser Ausdruck für $h \to 0$ gegen $f'(x_0)g(x_0)+f(x_0)g'(x_0)$ strebt. (Es gilt $g(x_0+h) \to g(x_0)$ aufgrund der in (89.12) bewiesenen Stetigkeit von g im Punkt x_0.) Zum Nachweis von (d) schreiben wir den Differenzenquotienten $\big((f/g)(x_0+h) - (f/g)(x_0)\big)/h$ in der Form

$$\frac{\dfrac{f(x_0+h) - f(x_0)}{h} \cdot g(x_0) - f(x_0) \cdot \dfrac{g(x_0+h) - g(x_0)}{h}}{g(x_0+h)g(x_0)}$$

und erkennen, daß dieser Ausdruck für $h \to 0$ gegen $\big(f'(x_0)g(x_0) - f(x_0)g'(x_0)\big)/g(x_0)^2$ strebt. (Es gilt $g(x_0+h) \to g(x_0)$ aufgrund der in (89.12) bewiesenen Stetigkeit von g im Punkt x_0.) Beim Nachweis von (e) würden wir gern einfach

$$\frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = \frac{g\big(f(x)\big) - g\big(f(x_0)\big)}{f(x) - f(x_0)} \cdot \frac{f(x) - f(x_0)}{x - x_0}$$

schreiben und dann argumentieren, auf der rechten Seite konvergiere der rechte Faktor für $x \to x_0$ gegen $f'(x_0)$, während der linke Faktor wegen $f(x) \to f(x_0)$ gegen $g'\big(f(x_0)\big)$ konvergiere. Wir müssen aber ein wenig aufpassen, denn es könnte beliebig nahe bei x_0 Punkte x mit $f(x) = f(x_0)$ geben, und dann wäre der linke Faktor gar nicht definiert. Wir betrachten daher eine beliebige Folge (x_n) mit $x_n \to x_0$ und $x_n \neq x_0$ für alle $n \in \mathbb{N}$. Falls nötig, zerlegen wir die Folge in zwei Teilfolgen (x'_n) und (x''_n) derart, daß $f(x'_n) \neq f(x_0)$ und $f(x''_n) = f(x_0)$ für alle $n \in \mathbb{N}$ gilt. Für die Elemente der ersten Folge gilt dann

$$\frac{(g \circ f)(x'_n) - (g \circ f)(x_0)}{x'_n - x_0} \;=$$
$$\frac{g\big(f(x'_n)\big) - g\big(f(x_0)\big)}{f(x'_n) - f(x_0)} \cdot \frac{f(x'_n) - f(x_0)}{x'_n - x_0};$$

für $n \to \infty$ geht dieser Ausdruck gegen $g'\big(f(x_0)\big)f'(x_0)$. Existiert eine unendliche Teilfolge (x''_n) wie angegeben, so folgt einerseits $f'(x_0) = \lim_n \big(f(x''_n) - f(x_0)\big)/(x''_n - x_0) = \lim_{n\to\infty} 0 = 0$, andererseits auch $g\big(f(x''_n)\big) = g\big(f(x_0)\big)$ für alle n, so daß auch $\big((g \circ f)(x''_n) - (g \circ f)(x_0)\big)/(x''_n - x_0) \to 0 = g'\big(f(x_0)\big) \cdot f'(x_0)$ für diese zweite Teilfolge gilt.

(f) Es sei (y_n) eine Folge im Definitionsbereich von f^{-1} mit $y_n \to y_0$. Wir setzen $x_n := f^{-1}(y_n)$; da f^{-1} nach (82.10) automatisch stetig ist, folgt hieraus $x_n \to x_0$. Es ergibt sich

$$\frac{f^{-1}(y_n) - f^{-1}(y_0)}{y_n - y_0} = \frac{x_n - x_0}{f(x_n) - f(x_0)} = \left[\frac{f(x_n) - f(x_0)}{x_n - x_0}\right]^{-1},$$

und dieser Ausdruck geht für $n \to \infty$ gegen $1/f'(x_0)$. ∎

(89.14) Beispiele. (a) Die Ableitung eines Polynoms $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ ist aufgrund der Faktor- und der Summenregel gegeben durch $f'(x) = a_1 + 2a_2 x + \cdots + n a_n x^{n-1}$.

(b) Die Funktion $f(x) = \sin x \cos x$ hat nach der Produktregel die Ableitung $f'(x) = \cos^2 x - \sin^2 x$. Das gleiche Ergebnis erhält man auch, wenn man $f(x) = \frac{1}{2}\sin(2x)$ schreibt und dann die Ableitung nach der Kettenregel bildet: $f'(x) = \frac{1}{2}\cos(2x) \cdot 2 = \cos(2x) = \cos^2 x - \sin^2 x$.

(c) Die Ableitung von $f(x) = \tan x = \sin x / \cos x$ ist nach der Quotientenregel gegeben durch $f'(x) = (\cos^2 x + \sin^2 x)/\cos^2 x = 1/\cos^2 x$ bzw. $f'(x) = 1 + \tan^2 x$.

(d) Die Ableitung der Funktion $f(x) = (x^2+1)^2$ erhält man nach der Kettenregel als $f'(x) = 2(x^2+1)\cdot 2x = 4x(x^2+1) = 4x^3 + 4x$. Das gleiche Ergebnis erhält man auch, indem man zuerst ausmultipliziert und dann direkt ableitet; aus $f(x) = x^4 + 2x^2 + 1$ folgt nämlich $f'(x) = 4x^3 + 4x$.

(e) Die Ableitung von $f(x) = e^x$ erhält man, indem man f als Umkehrfunktion von $g(x) := \ln x$ auffaßt und dann die Umkehrregel anwendet; es ergibt sich $f'(x) = (g^{-1})'(x) = 1/g'\big(f(x)\big) = 1/\big(1/f(x)\big) = f(x) = e^x$. Die Funktion f ist also – wie bereits in (89.10) durch direkte Rechnung hergeleitet – ihre eigene Ableitung.

(f) Die Ableitung der für $x > 0$ definierten Funktion $f(x) = x^x = e^{x \ln x}$ ist nach Teil (e) und der Ketten- und der Produktregel gegeben durch $f'(x) = e^{x \ln x} \cdot (\ln x + x/x) = x^x(1 + \ln x)$. ♦

Es folgt ein Beispiel dafür, daß die Ableitungsfunktion einer differentiierbaren Funktion nicht zwangsläufig stetig sein muß.

(89.15) Beispiel. Wir betrachten die Funktion

$$f(x) \;:=\; \begin{cases} x^2 \sin(1/x), & \text{falls } x \neq 0; \\ 0, & \text{falls } x = 0. \end{cases}$$

Man überzeugt sich leicht davon, daß f auf ganz $\mathbb{R}$ differentiierbar ist mit

$$f'(x) \;=\; \begin{cases} 2x \sin(1/x) - \cos(1/x), & \text{falls } x \neq 0; \\ 0, & \text{falls } x = 0. \end{cases}$$

Wir sehen, daß f' an der Stelle 0 zwar definiert, aber nicht stetig ist. (Die folgenden Abbildungen zeigen die Graphen von f und f' mit einer Skalierung um den Faktor 10.)

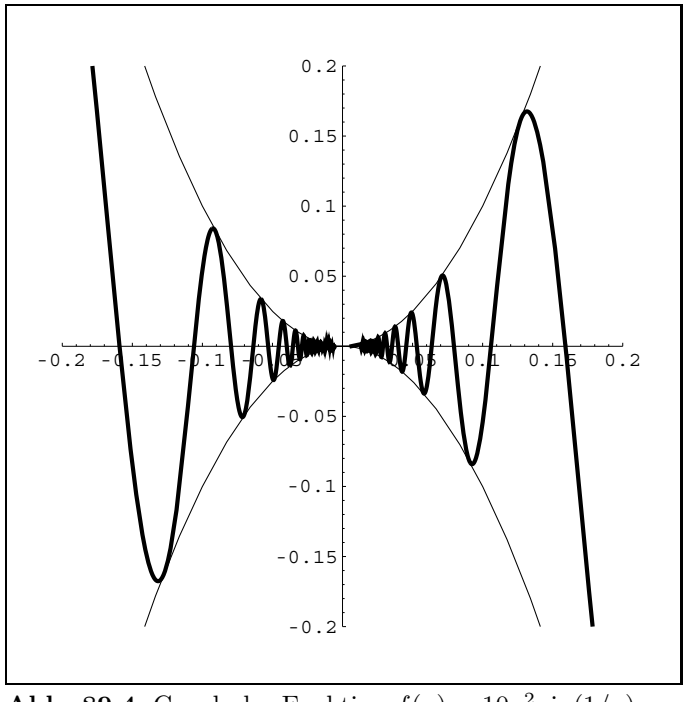

Abb. 89.4: Graph der Funktion $f(x) = 10x^2 \sin(1/x)$.

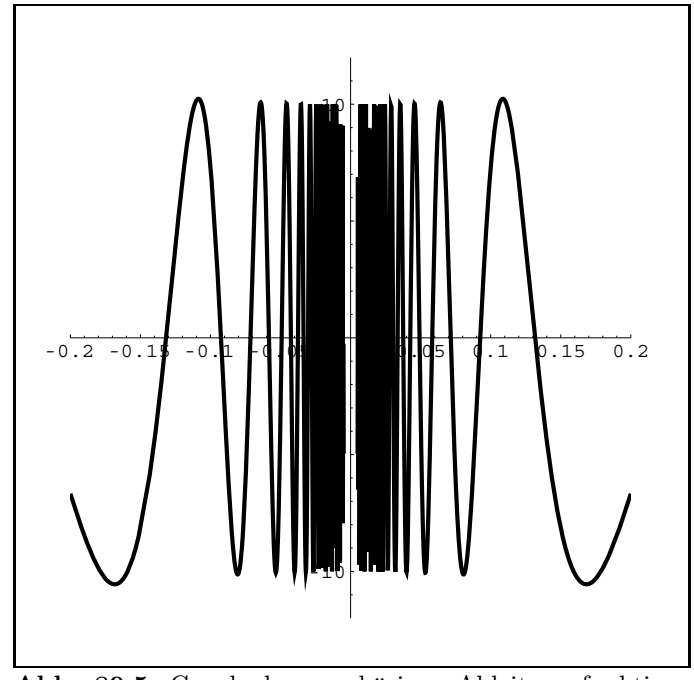

Abb. 89.5: Graph der zugehörigen Ableitungsfunktion $f'(x) = 20x \sin(1/x) - 10 \cos(1/x)$.

Aus der Funktion dieses Beispiels läßt sich leicht eine Funktion konstruieren, die an einer Stelle x_0 zwar eine positive Ableitung besitzt (deren Graph an dieser Stelle also positive Steigung hat), die aber auf keiner Umgebung von x_0 monoton wächst, nämlich $f(x) := x + 10x^2 \sin(1/x)$ für $x \neq 0$ und $f(0) := 1$. Anhand dieser Funktion sieht man auch, daß in der Umkehrregel (89.13) die Existenz der Umkehrfunktion explizit vorausgesetzt werden muß und nicht schon aus der Bedingung $f'(x_0) \neq 0$ folgt. ♦

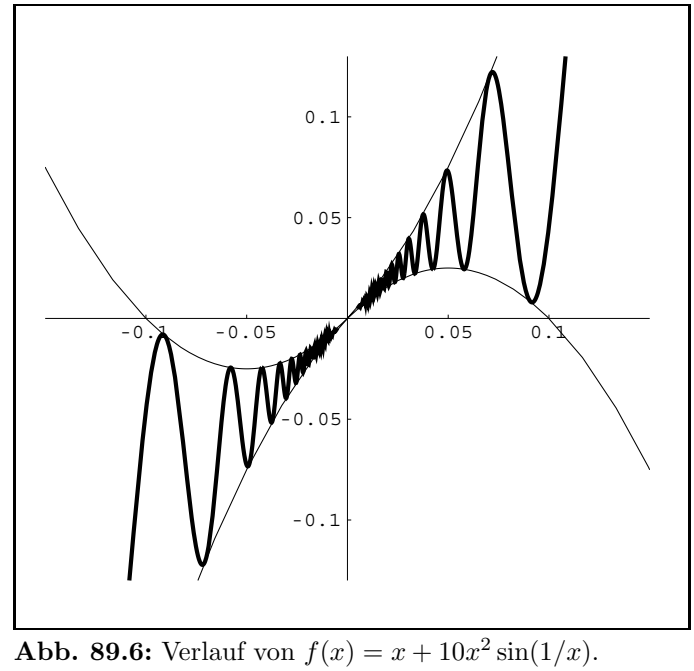

Abb. 89.6: Verlauf von $f(x) = x + 10x^2 \sin(1/x)$.

(89.16) Beispiel: Allgemeine Exponentialfunktion. Es sei $a > 0$ beliebig. Die durch $f(x) = a^x = e^{x \ln a}$ definierte Funktion $f : \mathbb{R} \to \mathbb{R}$ ist nach (89.10) und der Kettenregel auf ganz $\mathbb{R}$ differentiierbar mit $f'(x) = e^{x \ln a} \cdot \ln a = a^x \ln a$. ♦

(89.17) Beispiel: Allgemeine Potenzfunktion. Es sei $a \in \mathbb{R}$ beliebig. Die durch $f(x) = x^a = e^{a \ln(x)}$ definierte Funktion $f : (0, \infty) \to \mathbb{R}$ ist nach (89.10) und der Kettenregel gegeben durch $f'(x) = e^{a \ln(x)} \cdot (a/x) = x^a \cdot a \cdot x^{-1} = ax^{a-1}$. ♦

(89.18) Beispiele: Hyperbel- und Areafunktionen. Mit Hilfe der Exponentialfunktion definierten wir in (80.1) die Hyperbelfunktionen

$$\begin{aligned} \sinh(x) &:= (e^x - e^{-x})/2, \\ \cosh(x) &:= (e^x + e^{-x})/2, \\ \tanh(x) &:= \sinh(x)/\cosh(x). \end{aligned}$$

Wegen $\exp' = \exp$ nach (89.10) erhalten wir unter Benutzung der Kettenregel zunächst

$$\begin{aligned} \sinh'(x) &= (e^x + e^{-x})/2 = \cosh(x) \quad \text{und} \\ \cosh'(x) &= (e^x - e^{-x})/2 = \sinh(x) \end{aligned}$$

und unter Benutzung der Quotientenregel und der Identität $\cosh^2 - \sinh^2 = 1$ dann

$$\tanh'(x) = \frac{\cosh(x)^2 - \sinh(x)^2}{\cosh(x)^2} = \frac{1}{\cosh(x)^2}$$

bzw. $\tanh'(x) = 1 - \tanh(x)^2$. Da die Hyperbelfunktionen mit Hilfe der Exponentialfunktion definiert wurden, konnten wir in (80.7) ihre Umkehrfunktionen durch den natürlichen Logarithmus ausdrücken und erhielten

$$\begin{aligned} \operatorname{arsinh}(x) &= \ln(x + \sqrt{x^2 + 1}), \\ \operatorname{arcosh}(x) &= \ln(x + \sqrt{x^2 - 1}), \\ \operatorname{artanh}(x) &= \frac{1}{2} \ln \frac{1 + x}{1 - x}. \end{aligned}$$

Wenden wir auf diese Gleichungen die Kettenregel und die Quotientenregel an, so erhalten wir

$$\begin{aligned} \operatorname{arsinh}'(x) &= 1/\sqrt{x^2 + 1}, \\ \operatorname{arcosh}'(x) &= 1/\sqrt{x^2 - 1}, \\ \operatorname{artanh}'(x) &= 1/(1 - x^2). \end{aligned}$$

Das gleiche Ergebnis läßt sich auch durch Anwendung der Umkehrregel aus den Ableitungen der Hyperbelfunktionen herleiten (Übungsaufgabe!). ♦

(89.19) Beispiel: Bogenfunktionen. Die Umkehrfunktionen der trigonometrischen Funktionen haben die folgenden Ableitungen:

$$\begin{aligned} \arcsin'(x) &= 1/\sqrt{1 - x^2}, \\ \arccos'(x) &= -1/\sqrt{1 - x^2}, \\ \arctan'(x) &= 1/(1 + x^2). \end{aligned}$$

Die Differentiierbarkeit aller drei Funktionen ist zunächst klar nach (89.13)(f). Um die Ableitungen zu berechnen, betrachten wir zunächst die Funktion $y = \arcsin(x)$. Dann ist $x = \sin(y)$, nach der Kettenregel also $1 = \cos(y) \cdot y' = \sqrt{1 - \sin^2 y} \cdot y' = \sqrt{1 - x^2} \cdot y'$ und damit $y' = 1/\sqrt{1 - x^2}$. (Wegen $y \in [-\pi/2, \pi/2]$ ist $\cos(y) \geq 0$ und damit $\cos y = +\sqrt{1 - \sin^2 y}$.) Analog erhalten wir für $y = \arccos(x)$ die Beziehung $x = \cos y$, nach Ableiten also $1 = -\sin(y) \cdot y' = -\sqrt{1 - \cos^2 y} \cdot y' = -\sqrt{1 - x^2} \cdot y'$ und damit $y' = -1/\sqrt{1 - x^2}$. (Wegen $y \in [0, \pi]$ ist $\sin(y) \geq 0$ und damit $\sin y = +\sqrt{1 - \cos^2 y}$.) Für $y = \arctan(x)$ gilt schließlich $x = \tan(y)$, nach Ableiten folglich $1 = \tan'(y) \cdot y' = (1 + \tan^2 y) \cdot y' = (1 + x^2) \cdot y'$ und damit $y' = 1/(1 + x^2)$. ♦

Der nächste Satz behandelt nun statt einer einzelnen Funktion eine ganze Klasse von Funktionen; er zeigt, daß jede analytische Funktion differentiierbar ist und daß die Ableitung durch gliedweises Ableiten einer Potenzreihendarstellung gebildet werden darf.

(89.20) Satz. *Es gelte $f(x) = \sum_{k=0}^{\infty} a_k (x - p)^k$ für $|x - p| < R$. Dann ist f an jeder Stelle x mit $|x - p| < R$ differentiierbar, und es gilt*

$$f'(x) = \sum_{k=1}^{\infty} k a_k (x - p)^{k-1}.$$

Beweis. Wir wählen einen beliebigen Punkt q mit $|q - p| < R$. Für alle x mit $|x - q| < r := R - |q - p|$ gilt nach (76.7) dann eine Potenzreihendarstellung $f(x) = f(q) + \sum_{k=1}^{\infty} b_k (x - q)^k$; für jedes solche x ist dann

$$\frac{f(x) - f(q)}{x - q} = \sum_{k=1}^{\infty} b_k (x - q)^{k-1}.$$

Für $x \to q$ konvergiert die rechte Seite nach dem Stetigkeitssatz (82.6) gegen $b_1 = \sum_{n=1}^{\infty} n a_n (q - p)^{n-1}$; also existiert

$$f'(q) = \lim_{x \to q} \frac{f(x) - f(q)}{x - q} = \sum_{k=1}^{\infty} k a_k (q - p)^{k-1}.$$

Da q beliebig in $B_R(p)$ gewählt war, ist dies schon die Behauptung.

Alternativer Beweis. Wir geben noch einen Beweis, der weder den Transformationssatz noch den Stetigkeitssatz benutzt, sondern das gewünschte Ergebnis durch eine direkte Rechnung liefert. O.B.d.A. dürfen wir $p = 0$ annehmen, also $f(x) = \sum_{k=0}^{\infty} a_k x^k$ ansetzen. Wegen $\sqrt[k]{k+1} \to 1$ für $k \to \infty$ gilt $\limsup \sqrt[k]{(k+1)|a_{k+1}|} = \limsup \sqrt[k]{|a_{k+1}|} = \limsup \sqrt[k]{|a_k|}$; die gliedweise abgeleitete Reihe $g(x) := \sum_{k=1}^{\infty} k a_k x^{k-1}$ hat also den gleichen Konvergenzradius wie die ursprüngliche Reihe. Für $n \in \mathbb{N}$ sei $f_n(x) := \sum_{k=0}^{n} a_k x^k$ die n-te Partialsumme von f; dann ist $f_n'(x) = \sum_{k=0}^{n} k a_k x^{k-1}$.

Wir halten nun einen Punkt x_0 mit $|x_0| < \rho$ fest und geben uns ein $\varepsilon > 0$ vor; wir wollen zeigen, daß dann

$$|f(x) - f(x_0) - g(x_0)(x - x_0)| \;<\; |x - x_0| \cdot \varepsilon$$

bzw.

$$\left| \frac{f(x) - f(x_0)}{x - x_0} - g(x_0) \right| \;<\; \varepsilon$$

gilt, wenn nur x genügend nahe bei x_0 liegt. (Gelingt dieser Nachweis für beliebiges $\varepsilon > 0$, so ist gezeigt, daß f an der Stelle x_0 differentiierbar ist mit $f'(x_0) = g(x_0)$.) Zunächst wählen wir eine Zahl r mit $|x_0| < r < \rho$ und dann ein $n \in \mathbb{N}$ mit $\sum_{k=n+1}^{\infty} k|a_k|r^{k-1} < \varepsilon/4$, was wegen der absoluten Konvergenz der Potenzreihe g möglich ist. Ferner wählen wir ein $\delta > 0$ derart, daß

$$|f_n(x) - f_n(x_0) - f_n'(x_0)(x - x_0)| \;<\; |x - x_0| \cdot \frac{\varepsilon}{2}$$

für $|x - x_0| < \delta$ gilt, was wegen der Differentiierbarkeit des Polynoms f_n an der Stelle x_0 möglich ist. Der Ausdruck $f(x) - f(x_0) - g(x_0)(x - x_0)$ ist nun die Summe der beiden Terme

$$T_1 \;=\; f_n(x) - f_n(x_0) - f_n'(x_0)(x - x_0)$$

und

$$\begin{aligned} T_2 \;&=\; \sum_{k=n+1}^{\infty} a_k x^k - \sum_{k=n+1}^{\infty} a_k x_0^k - \sum_{k=n+1}^{\infty} k a_k x_0^{k-1}(x - x_0) \\ &=\; \sum_{k=n+1}^{\infty} \big(a_k(x^k - x_0^k) - k a_k x_0^{k-1}(x - x_0) \big) \\ &=\; (x - x_0) \sum_{k=n+1}^{\infty} a_k \left(\sum_{i+j=k-1} x^i x_0^j - k x_0^{k-1} \right). \end{aligned}$$

Für $|x - x_0| < \delta$ ist $|T_1| < |x - x_0| \cdot \varepsilon/2$, und für $|x| \le r$ ist $\left| \left(\sum_{i+j=k-1} x^i x_0^j\right) - k x_0^{k-1} \right| \le \left(\sum_{i+j=k-1} r^i r^j\right) + k r^{k-1} = 2k r^{k-1}$ und damit

$$|T_2| \;\le\; |x - x_0| \sum_{k=n+1}^{\infty} 2k|a_k| r^{k-1} \;<\; |x - x_0| \cdot \frac{\varepsilon}{2}\,.$$

Für $|x - x_0| < \min\{\delta, r - |x_0|\} =: \delta^\star$ ist dann

$$|f(x) - f(x_0) - g(x_0)(x - x_0)| \le |T_1| + |T_2| < |x - x_0| \cdot \varepsilon,$$

was zu zeigen war. ∎

(89.21) Beispiele. (a) Wie wir in (89.10) schon auf anderem Wege ermittelten, besitzt die Exponentialfunktion $\exp(x) = \sum_{k=0}^{\infty} x^k/k!$ die Ableitung

$$\exp'(x) = \sum_{k=1}^{\infty} \frac{k x^{k-1}}{k!} = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = \sum_{k=0}^{\infty} \frac{x^k}{k!} = \exp(x)\,.$$

(b) Die für $|x| < 1$ definierte Funktion $B_\alpha(x) = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$ hat die Ableitung

$$\begin{aligned} B_\alpha'(x) \;&=\; \sum_{k=1}^{\infty} k \binom{\alpha}{k} x^{k-1} \;=\; \sum_{k=0}^{\infty} (k+1) \binom{\alpha}{k+1} x^k \\ &=\; \alpha \sum_{k=0}^{\infty} \binom{\alpha - 1}{k} x^k \;=\; \alpha \cdot B_{\alpha-1}(x); \end{aligned}$$

für $\mathbb{K} = \mathbb{R}$ ist dies nichts anderes als die Gleichung $(\mathrm{d}/\mathrm{d}x)(1 + x)^\alpha = \alpha(1 + x)^{\alpha-1}$. ♦

Der folgende Satz liefert eine allgemeine Aussage über die Differentiierbarkeit konvexer Funktionen.

(89.22) Satz. *Es seien $I \subseteq \mathbb{R}$ ein offenes Intervall und $f : I \to \mathbb{R}$ eine konvexe Funktion. Dann existieren an jeder Stelle $x \in I$ sowohl die links- als auch die rechtsseitige Ableitung; diese sind an höchstens abzählbar vielen Stellen voneinander verschieden.*

Beweis. Es sei $x \in I$ fest gewählt. Da nach (71.47) die auf $I \setminus \{x\}$ definierte Funktion

$$y \mapsto \frac{f(y) - f(x)}{y - x}$$

monoton wächst, existieren

$$f_-'(x) = \sup_{y<x} \frac{f(y) - f(x)}{y - x} \text{ und } f_+'(x) = \inf_{y>x} \frac{f(y) - f(x)}{y - x},$$

und es gilt $f_-'(x) \le f_+'(x)$. Gilt ferner $x < y$, so ist

$$(\star)\;\; f_-'(x) \;\le\; f_+'(x) \;\le\; \frac{f(y) - f(x)}{y - x} \;\le\; f_-'(y) \;\le\; f_+'(y),$$

so daß f_-' und f_+' auf ihrem Definitionsbereich jeweils monoton wachsen. Schließlich sei X die Menge derjenigen Punkte $x \in I$, für die $f'(x)$ nicht existiert, für die also $f_-'(x) < f_+'(x)$ gilt. Dann sind die Intervalle $I_x := \big(f_-'(x), f_+'(x)\big)$ mit $x \in X$ nach $(\star)$ disjunkt. Da man in jedem dieser Intervalle eine rationale Zahl wählen kann und da es nur abzählbar viele rationale Zahlen gibt, kann es auch nur abzählbar viele solcher Intervalle geben; also ist die Menge X abzählbar. ∎

Wir gehen noch einmal zurück zur Definition des Ableitungsbegriffs. Es sei $I \subseteq \mathbb{R}$ ein offenes Intervall, und es sei $f : I \to \mathbb{R}$ eine Funktion, deren Ableitung $f'(x_0)$ an jedem Punkt $x_0 \in I$ existiere. Dann können wir die Ableitungsfunktion $f' : I \to \mathbb{R}$ betrachten und natürlich auch nach deren Ableitung fragen; existiert diese, so nennen wir sie die *zweite Ableitung* von f und bezeichnen sie mit dem Symbol $f'' = (f')'$. Die höheren Ableitungen werden dann induktiv definiert durch $f''' := (f'')'$, $f^{(4)} := f'''' := (f''')'$, $f^{(5)} := (f^{(4)})'$, und so weiter; allgemein ist also $f^{(n+1)} := (f^{(n)})'$. (Die Existenz von

$f^{(n+1)}(x_0)$ an einer festen Stelle x_0 setzt dabei die Existenz von $f^{(n)}$ auf einer ganzen Umgebung von x_0 voraus.) Nach (89.12) garantiert die Existenz von $f^{(n)}$ auf einem Intervall I bereits die Stetigkeit von $f, f', \dots, f^{(n-1)}$ auf diesem Intervall; die n-te Ableitung $f^{(n)}$ selbst ist aber nicht zwangsläufig stetig.

(89.23) Definition höherer Ableitungen. *Es seien $I \subseteq \mathbb{R}$ ein offenes Intervall und $f : I \to \mathbb{R}$ eine Funktion. Als "nullte Ableitung" $f^{(0)}$ von f bezeichnen wir die Funktion f selbst. Ist die n-te Ableitung $f^{(n)}$ von f bereits definiert, so definieren wir die $(n+1)$-te Ableitung von f an einer Stelle x_0 durch $f^{(n+1)}(x_0) := (f^{(n)})'(x_0)$, falls $f^{(n)}$ auf einer Umgebung von x_0 definiert und an der Stelle x_0 differentiierbar ist. Wir sagen, f sei n-mal differentiierbar auf I, wenn die Ableitungen $f', \dots, f^{(n)}$ auf ganz I definiert sind. Ist $f^{(n)}$ sogar noch stetig, so nennen wir f von der Klasse C^n und schreiben $f \in C^n(I)$.*

(89.24) Beispiel. *Es gelte $f(x) = \sum_{k=0}^{\infty} a_k (x-p)^k$ mit dem Konvergenzradius R. Dann ist f an jeder Stelle x mit $|x-p| < R$ unendlich oft differentiierbar; für alle $n \in \mathbb{N}$ gilt*

$$\begin{aligned} f^{(n)}(x) &= \sum_{k=n}^{\infty} a_k \cdot k(k-1)\cdots\big(k-(n-1)\big)(x-p)^{k-n} \\ &= \sum_{k=0}^{\infty} a_{n+k}(k+n)(k+n-1)\cdots(k+1)(x-p)^k\,. \end{aligned}$$

Insbesondere gilt $f^{(n)}(p) = a_n \cdot n!$, also $a_n = \dfrac{f^{(n)}(p)}{n!}$.

Beweis. Dies folgt mit vollständiger Induktion über n sofort aus (89.20). ∎

Wir fragen nun, ob sich die in (89.13) hergeleiteten Ableitungsregeln auf höhere Ableitungen verallgemeinern lassen. Als triviale Anwendung der Summen- und der Faktorregel ergibt sich die folgende Aussage: Sind $f, g : I \to \mathbb{R}$ an einer Stelle x_0 jeweils n-mal differentiierbar, dann auch alle Linearkombinationen $\alpha f + \beta g$ mit $\alpha, \beta \in \mathbb{R}$, und es gilt $(\alpha f + \beta g)^{(n)}(x_0) = \alpha \cdot f^{(n)}(x_0) + \beta \cdot g^{(n)}(x_0)$. Wiederholte Anwendung der Produktregel führt auf die folgende **Leibnizregel**, die eine Formel für die höheren Ableitungen des Produkts zweier Funktionen liefert.

(89.25) Formel von Leibniz. *Sind die Funktionen $f, g : I \to \mathbb{R}$ an der Stelle x_0 bis zur n-ten Ordnung differentiierbar, so ist auch das Produkt $f \cdot g$ an der Stelle x_0 bis zur n-ten Ordnung differentiierbar, und es gilt*

$$(f \cdot g)^{(n)}(x_0) = \sum_{k=0}^{n} \binom{n}{k} f^{(k)}(x_0) g^{(n-k)}(x_0)\,.$$

Beweis. Dies folgt unter Anwendung der Produktregel mit vollständiger Induktion über n (Übungsaufgabe!). ∎

Ähnlich wie für die Produktregel gibt es auch für die Kettenregel eine Verallgemeinerung auf höhere Ableitungen. Ausgehend von der Kettenregel $(g \circ f)' = (g' \circ f) \cdot f'$ erhalten wir zunächst $(g \circ f)'' = (g'' \circ f) \cdot (f')^2 + (g' \circ f) \cdot f''$, anschließend

$$(g \circ f)''' = (g''' \circ f) \cdot (f')^3 + (g'' \circ f) \cdot 3f'f'' + (g' \circ f) \cdot f'''\,,$$

dann

$$\begin{aligned} (g \circ f)'''' &= (g'''' \circ f) \cdot (f')^4 + (g''' \circ f) \cdot 6(f')^2 f'' \\ &+ (g'' \circ f)\big(4f'f''' + 3(f'')^2\big) + (g' \circ f) f''''\,, \end{aligned}$$

und so weiter. Die allgemeine Formel, benannt nach dem italienischen Mathematiker Francesco Faà di Bruno (1825-1888), ist im folgenden Satz angegeben.

(89.26) Formel von Faà di Bruno. *Gegeben seien Funktionen $f : I \to J$ und $g : J \to \mathbb{R}$. Existieren $f^{(n)}(x_0)$ an einer Stelle $x_0 \in I$ und $g^{(n)}\big(f(x_0)\big)$ an der Stelle $f(x_0)$, so existiert auch $(g \circ f)^{(n)}(x_0)$, und es gilt*

$$\begin{aligned} (g \circ f)^{(n)}(x_0) &= \sum_{k=0}^{n} \sum_{\substack{k_1+k_2+\cdots+k_n=k,\\ k_1+2k_2+\cdots+nk_n=n}} \frac{n!}{k_1!k_2!\cdots k_n!} \times \cdots \\ &\cdots \times g^{(k)}\big(f(x_0)\big)\Big(\frac{f'(x_0)}{1!}\Big)^{k_1}\Big(\frac{f''(x_0)}{2!}\Big)^{k_2}\cdots\Big(\frac{f^{(n)}(x_0)}{n!}\Big)^{k_n} \end{aligned}$$

bzw.

$$(g \circ f)^{(n)} = \sum \left(\frac{n!}{k_1! \cdots k_n!} \big(g^{(k_1+\cdots+k_n)} \circ f\big) \cdot \prod_{i=1}^{n} \Big(\frac{f^{(i)}}{i!}\Big)^{k_i} \right),$$

wobei die Summe über alle n-Tupel $(k_1, \dots, k_n) \in \mathbb{N}_0^n$ läuft, die die Bedingung $k_1 + 2k_2 + \cdots + nk_n = n$ erfüllen.

Beweis.† Eine triviale Induktion zeigt, daß es Polynome $P_{n,k}$ in n Variablen derart gibt, daß

$$(\star) \qquad (g \circ f)^{(n)} = \sum_{k=0}^{n} (g^{(k)} \circ f) \cdot P_{n,k}(f', f'', \dots, f^{(n)})$$

für alle Funktionen f und g gilt, die die genannten Voraussetzungen erfüllen. (Die Induktion zeigt, daß diese Polynome rekursiv gegeben sind durch $P_{0,0}(x) = 1$ und $P_{n+1,k}(x_1, \dots, x_n, x_{n+1}) = x_1 \cdot P_{n,k-1}(x_1, \dots, x_n) + \sum_{i=1}^{n} x_{i+1} \cdot (\partial_i P_{n,k})(x_1, \dots, x_n)$, wenn wir mit $P_{n,0}$ und $P_{n,n+1}$ jeweils das Nullpolynom und mit ∂_i die Ableitung nach dem i-ten Argument bezeichnen, aber das ist unwesentlich für die Zwecke des Beweises.) Aus $(\star)$ ergibt sich, daß $(g \circ f)^{(n)}(x_0)$ nur von den Werten $f^{(k)}(x_0)$ und $g^{(k)}\big(f(x_0)\big)$ mit $0 \le k \le n$ abhängt. Um die Gültigkeit der

† Karlheinz Spindler: *A short proof of the formula of Faà di Bruno*; Elemente der Mathematik **60** (2005), S. 33-35. (Gegenüber dem in diesem Artikel angegebenen Beweis wurden die Rollen von f und g vertauscht.)

angegebenen Formel an der festen Stelle x_0 zu beweisen, dürfen wir daher f und g durch beliebige Funktionen F und G ersetzen, die die gleichen Ableitungen bis zur Ordnung n wie f und g an den Stellen x_0 bzw. $f(x_0)$ haben. Es genügt also, die Formel von Faà di Bruno für Polynome zu beweisen! Dabei dürfen wir o.B.d.A. $x_0 = 0$ und $f(x_0) = 0$ annehmen, haben also $g(x) = b_0 + b_1x + \cdots + b_nx^n$ and $f(x) = a_1x + a_2x^2 + \cdots + a_nx^n$ mit $b_k = g^{(k)}(0)/k!$ und $a_k = f^{(k)}(0)/k!$ für alle k. Die zu beweisende Formel reduziert sich dann auf die Aussage, daß der Koeffizient von x^n in der Entwicklung von $g\big(f(x)\big)$ nach Potenzen von x gegeben ist durch

$$\sum_{k=0}^{n} \sum_{\substack{k_1+k_2+\cdots+k_n=k,\\ k_1+2k_2+\cdots+nk_n=n}} \frac{k!}{k_1!k_2!\cdots k_n!} b_k\, a_1^{k_1} a_2^{k_2} \cdots a_n^{k_n}\,.$$

Das ist aber trivial! Wenden wir nämlich die multinomische Formel

$$(X_1+\cdots+X_n)^k = \sum_{k_1+\cdots+k_n=k} \frac{k!}{k_1!k_2!\cdots k_n!} X_1^{k_1} X_2^{k_2} \cdots X_n^{k_n}$$

mit $X_k := a_kx^k$ an, so erhalten wir

$$g\big(f(x)\big) \;=\; \sum_{k=0}^{n} b_k(a_1x + a_2x^2 + \cdots + a_nx^n)^k \;=$$
$$\sum_{k=0}^{n} b_k \sum_{k_1+\cdots+k_n=k} \frac{k!}{k_1!k_2!\cdots k_n!} a_1^{k_1} a_2^{k_2} \cdots a_n^{k_n} x^{k_1+2k_2+\cdots+nk_n}\,.$$

∎

In der bisherigen Diskussion haben wir immer nur Funktionen $f : I \to \mathbb{R}$ betrachtet, die auf einem Intervall I definiert sind und reelle Werte annehmen. In vielen Fällen von praktischem Interesse will man aber nicht skalarwertige Funktionen (etwa Temperaturen) untersuchen, sondern vektorwertige Funktionen (etwa Positionen oder Geschwindigkeiten), und natürlich ist auch für solche Funktionen die Frage nach der Änderungsrate von Interesse. Im nächsten Abschnitt beschäftigen wir uns daher mit dem Ableitungsbegriff für vektorwertige Funktionen.

90. Differentiation vektorwertiger Funktionen

Der im vorigen Abschnitt definierte Begriff der Differentiierbarkeit reellwertiger Funktionen $f : I \to \mathbb{R}$ (wobei $I \subseteq \mathbb{R}$ ein offenes Intervall sei) läßt sich vollkommen problemlos auf vektorwertige Funktionen $f : \mathbb{R} \to W$ übertragen, vorausgesetzt, der betrachtete Vektorraum W trägt eine topologische Struktur, so daß Grenzwertbetrachtungen möglich sind. Zumeist wird dabei W ein endlichdimensionaler reeller Vektorraum sein (in physikalischen Anwendungen etwa der dreidimensionale Vektorraum $\mathfrak{V}$ aller Pfeilklassen), aber wir geben die Definition gleich für beliebige topologische Vektorräume.

(90.1) Definition. *Es seien $I \subseteq \mathbb{R}$ ein offenes Intervall, W ein reeller topologischer Vektorraum und $f : I \to W$ eine Funktion von I in W.*

(a) *Existiert an einer Stelle $x_0 \in I$ der folgende Grenzwert, so wird dieser als* **rechtsseitige Ableitung** *von f an der Stelle x_0 bezeichnet:*

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

(b) *Existiert an einer Stelle $x_0 \in I$ der folgende Grenzwert, so wird dieser als* **linksseitige Ableitung** *von f an der Stelle x_0 bezeichnet:*

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

(c) *Die Funktion f heißt* **differentiierbar** *an der Stelle $x_0 \in I$, falls der Grenzwert*

$$f'(x_0) \;:=\; \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0} \;=\; \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h}$$

existiert, falls also $f'_+(x_0)$ und $f'_-(x_0)$ existieren und übereinstimmen; es ist dann $f'(x_0) = f'_+(x_0) = f'_-(x_0)$. Der Vektor $f'(x_0) \in W$ heißt in diesem Fall die **Ableitung** *von f an der Stelle x_0.*

(d) *Existiert $f'(x)$ an jedem Punkt $x \in I$, so heißt die Funktion $f' : I \to W$ die* **Ableitungsfunktion** *von f.*

(e) *Wir bezeichnen $f^{(0)} := f$ als "nullte Ableitung" von f. Ist für ein $n \in \mathbb{N}$ die Funktion $f^{(n)} : I \to W$ definiert und existiert $(f^{(n)})'(x_0)$, so schreiben wir $f^{(n+1)}(x_0) := (f^{(n)})'(x_0)$ und nennen diesen Vektor die $(n+1)$-te Ableitung von f an der Stelle x_0.*

Genau wie in (89.12) ergibt sich, daß Differentiierbarkeit automatisch Stetigkeit zur Folge hat. Bevor wir einige Interpretationen des Ableitungsbegriffs geben, wollen wir noch auf eine auf Newton zurückgehende Bezeichnungskonvention hinweisen, die in der Physik üblich ist und die wir auch im Rahmen dieses Skriptums verwenden wollen.

(90.2) Bemerkung. Hat die unabhängige Variable die Bedeutung einer Zeitvariablen, so bezeichnet man die Ableitung einer Funktion oft mit einem Punkt statt mit einem Strich; die unabhängige Variable selbst wird dann mit dem Buchstaben t bezeichnet (was an lateinisch "tempus" für "Zeit" erinnern soll). Ist also $t \mapsto x(t)$ eine vektorwertige Funktion, so schreiben wir $\dot{x}(t_0)$ statt $x'(t_0)$ und $\ddot{x}(t_0)$ statt $x''(t_0)$. ♦

(90.3) Beispiel. Wir identifizieren den uns umgebenden affinen Raum mit dem Vektorraum $\mathfrak{V}$ aller Pfeilklassen. Für eine Funktion $x : \mathbb{R} \to \mathfrak{V}$ können wir die Ableitung $\dot{x}(t_0)$ folgendermaßen als Geschwindigkeitsvektor deuten. Wir fassen $x(t)$ als Position eines sich im Raum bewegenden Teilchens zur Zeit t auf, so daß $t \mapsto x(t)$ die von dem Teilchen durchlaufene Kurve beschreibt. Dann

∞, so daß Satz (95.10) die Existenz eines globalen Minimums garantiert. Die partiellen Ableitungen von F sind gegeben durch

$$\frac{\partial F}{\partial a}(a,b) \;=\; 2(a-b)+2\big((a-6)^2+b^2\big)\cdot 2(a-6) \text{ und}$$
$$\frac{\partial F}{\partial b}(a,b) \;=\; -2(a-b)+2\big((a-6)^2+b^2\big)\cdot 2b;$$

das globale Minimum von f kann nur an einer gemeinsamen Nullstelle (a,b) dieser partiellen Ableitungen angenommen werden. Dies führt auf

$$a-b \;=\; -2(a-6)\big((a-6)^2+b^2\big) \;=\; 2b\big((a-6)^2+b^2\big),$$

also auf die Gleichungen $a=6-b$ und $6-2b=4b^3$ bzw. $2b^3+b-3=0$. Die letzte Gleichung hat eine einzige reelle Nullstelle, nämlich $b_0=1$; dann ist $a_0=6-b_0=5$. Der Abstand wird also in den Punkten $(5,1)$ auf dem ersten und $(1,-1)$ auf dem zweiten Parabelbogen angenommen; er beträgt $\sqrt{F(5,1)}=\sqrt{20}=2\sqrt{5}\approx 4.472$ Längeneinheiten. ♦

(95.16) Aufgabe. *Aus zwei rechteckigen und zwei dreieckigen Brettern soll ein Futtertrog mit vorgegebenem Fassungsvermögen $V = 500\ell$ so gebildet werden, daß möglichst wenig Material verbraucht wird. Wie sind die Abmessungen des Trogs zu wählen?*

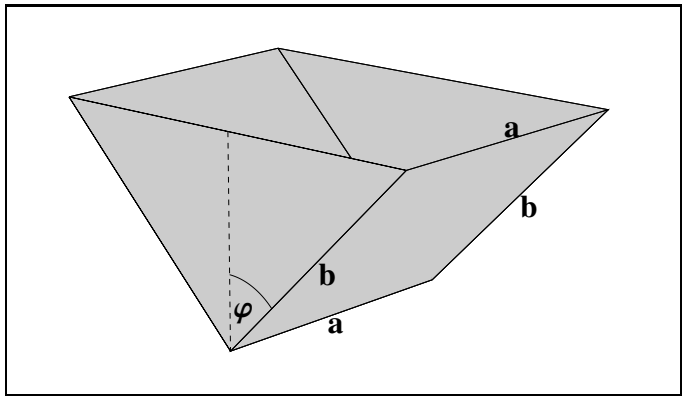

Abb. 95.9: Trog mit zu minimierendem Materialverbrauch.

Erste Lösung. Wir bezeichnen mit a und b die Länge bzw. Breite der beiden rechteckigen Bretter und mit 2φ den Winkel, unter dem diese beiden Bretter zusammenstoßen. Zu minimieren ist dann die Summe der beiden Rechtecksflächen und der beiden Dreiecksflächen, also der Ausdruck

$$F \;=\; 2ab+2b^2\sin(\varphi)\cos(\varphi) \;=\; 2ab+b^2\sin(2\varphi)$$

unter der Nebenbedingung $V = ab^2\sin(\varphi)\cos(\varphi)$ bzw. $\sin(2\varphi) = 2V/(ab^2)$; diese zieht die Ungleichung $0 < 2V/ab^2 \le 1$ nach sich. Zu minimieren ist also die Funktion

$$F(a,b) \;:=\; 2ab+b^2\cdot\frac{2V}{ab^2} \;=\; 2ab+\frac{2V}{a}$$

auf der Menge $D := \{(a,b)\in\mathbb{R}^2 \mid a>0, b>0, 2V\le ab^2\}$. Wegen $(\partial F/\partial b)(a,b) = 2a > 0$ für alle $(a,b)\in D$ besitzt F im Innern von D keine lokalen Extrema; es bleibt also nur die Möglichkeit eines Randminimums. Auf dem Rand von D ist $2V = ab^2$, also $a = 2V/b^2$ und damit $F(a,b) = F(2V/b^2,b) = b^2+4V/b =: f(b)$. Wegen $f(b)\to\infty$ für $b\to 0+$ und für $b\to\infty$ und $f(b)>0$ für alle $b>0$ besitzt f ein Minimum in $(0,\infty)$, welches das gesuchte Minimum von F auf D ist. Die Gleichung $f'(b) = 2b-4V/b^2$ führt auf $b^3 = 2V$ und damit auf die eindeutige Lösung $b = \sqrt[3]{2V}$. Einsetzen in die Gleichung $a = 2V/b^2$ und dann in $\sin(2\varphi) = 2V/(ab^2)$ liefert

$$a \;=\; b \;=\; \sqrt[3]{2V} \quad\text{und}\quad \varphi = \pi/4$$

als optimale Lösung. Der ideale Trog hat also quadratische Seitenbretter, die rechtwinklig aufeinanderstoßen.

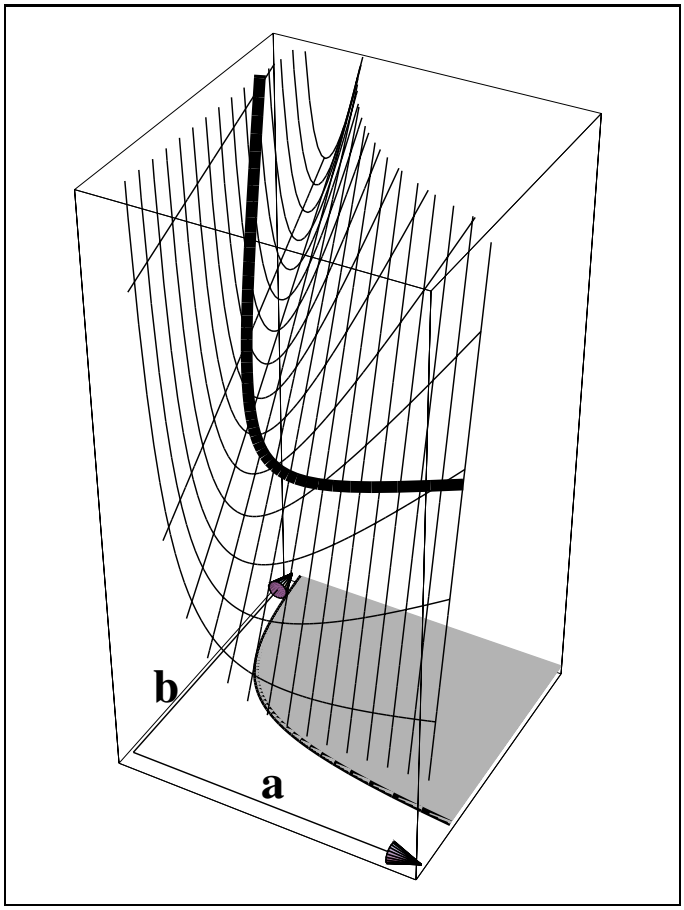

Abb. 95.10: Graph der Funktion F; der Definitionsbereich D ist grau schraffiert.

Zweite Lösung. Mit $a = V/(b^2\sin\varphi\cos\varphi)$ erhalten wir F als Funktion von b und φ, nämlich

$$\begin{aligned} F(b,\varphi) \;&=\; 2b^2\sin\varphi\cos\varphi+\frac{2V}{b\sin\varphi\cos\varphi} \\ &=\; b^2\sin(2\varphi)+\frac{4V}{b\sin(2\varphi)}\,. \end{aligned}$$

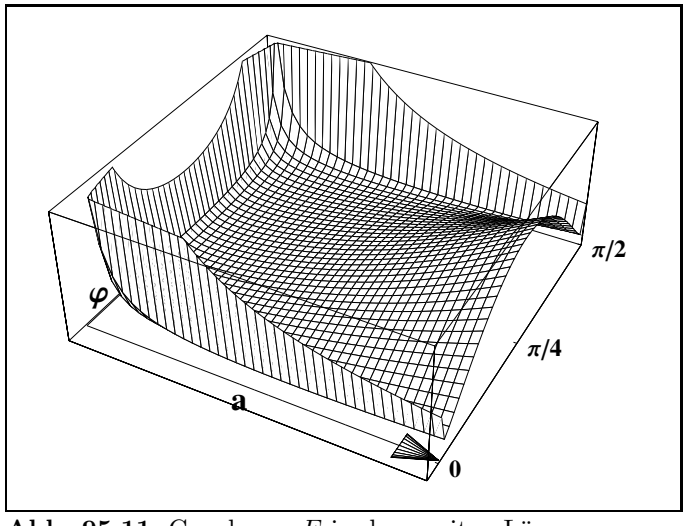

Abb. 95.11: Graph von F in der zweiten Lösung.

Die partiellen Ableitungen von F ergeben sich zu

$$\frac{\partial F}{\partial b}(b,\varphi) \;=\; 2\cdot\frac{b^3\sin^2(2\varphi)-2V}{b^2\sin(2\varphi)},$$
$$\frac{\partial F}{\partial \varphi}(b,\varphi) \;=\; 2\cos(2\varphi)\cdot\frac{b^3\sin^2(2\varphi)-4V}{b\sin^2(2\varphi)}.$$

Nullsetzen liefert $b^3\sin^2(2\varphi)=2V$ und dann $\cos(2\varphi)=0$. Dies impliziert zunächst $\varphi=\pi/4$, dann $b^3=2V$ (also $b=\sqrt[3]{2V}$ und schließlich $a=V/(b^2\sin\varphi\cos\varphi)=b$. ♦

(95.17) Aufgabe. *Gegeben seien Datenpunkte (x_i,y_i) mit $1\le i\le n$ mit paarweise verschiedenen x-Werten x_i. Die* **Ausgleichsgerade** *ist diejenige Gerade der Form $y=ax+b$, für die der Ausdruck*

$$F(a,b) \;:=\; \sum_{i=1}^{n}(ax_i+b-y_i)^2$$

minimal wird. Wie läßt sich diese Gerade aus den gegebenen Datenpunkten bestimmen?

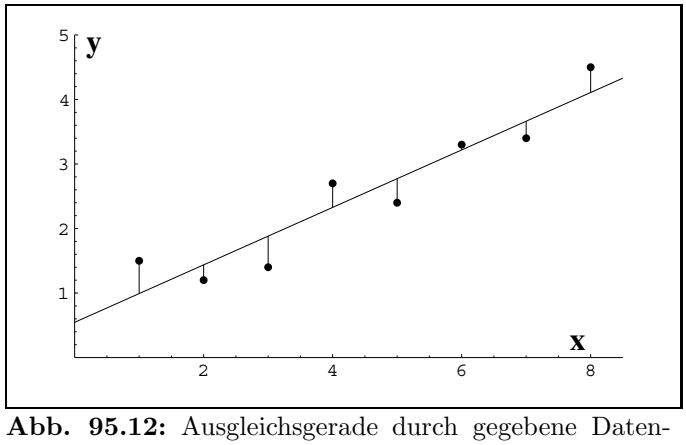

Abb. 95.12: Ausgleichsgerade durch gegebene Datenpunkte.

Lösung. Wegen $F(a,b)\to\infty$ für $\|(a,b)\|\to\infty$ ist nach Satz (95.10) zunächst einmal klar, daß es tatsächlich reelle Zahlen a und b gibt, die den Ausdruck $F(a,b)$ minimieren; an der Stelle $(a,b)\in\mathbb{R}^2$ müssen dann die partiellen Ableitungen $\partial F/\partial a$ und $\partial F/\partial b$ beide verschwinden; d.h., die folgenden Gleichungen müssen gelten:

$$0 \;=\; (\partial F/\partial a)(a,b) \;=\; 2\sum_{i=1}^{n}(ax_i+b-y_i)x_i,$$
$$0 \;=\; (\partial F/\partial b)(a,b) \;=\; 2\sum_{i=1}^{n}(ax_i+b-y_i).$$

Dies führt auf das lineare Gleichungssystem

$$(\sum_i x_i^2)a+(\sum_i x_i)b=\sum_i x_iy_i,\quad (\sum_i x_i)a+nb=\sum_i y_i,$$

dessen Koeffizientendeterminante gegeben ist durch

$$n\sum_i x_i^2-(\sum_i x_i)^2=\frac{1}{2}\sum_{i,j}(x_i-x_j)^2$$

und das daher eine eindeutige Lösung besitzt, wenn wenigstens zwei der auftretenden x-Werte voneinander verschieden sind (wenn also nicht alle Datenpunkte auf einer senkrechten Geraden liegen). Als Beispiel betrachten wir die folgende Meßreihe.

x	1	2	3	4	5	6	7	8
y	1.5	1.2	1.4	2.7	2.4	3.3	3.4	4.5

Hier ergeben sich die beiden Gleichungen $204a+36b=110.5$ und $36a+8b=20.4$, deren eindeutige gemeinsame Lösung gegeben ist durch $a=0.445238$ und $b=0.546429$. Die Ausgleichsgerade durch die betrachteten Datenpunkte hat also die Gleichung $y=0.445238\cdot x+0.546429$. ♦

Bei der Bestimmung der Ausgleichsgeraden wird von jedem der Datenpunkte aus die Entfernung zu demjenigen Punkt der Geraden bestimmt, der in y-Richtung von dem betrachteten Punkt aus liegt. Je nach Problemstellung kann es sinnvoller sein, den Abstand jedes einzelnen Punktes zum jeweils zugehörigen Lotfußpunkt zu ermitteln und dann die Quadratsumme dieser Abstände zu minimieren. Diese Aufgabe behandeln wir gleich für beliebige Dimensionen.

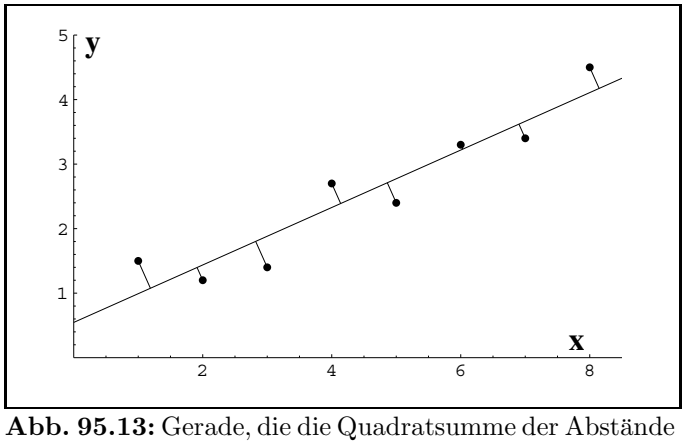

Abb. 95.13: Gerade, die die Quadratsumme der Abstände zu gegebenen Datenpunkten minimiert.

(95.18) Aufgabe. *Gegeben seien N Punkte $x_1, \dots, x_n$ in einem endlichdimensionalen Euklidischen Raum V. Bestimme diejenige Hyperebene H in V, für die der Ausdruck $\sum_{i=1}^N \operatorname{dist}(x_i, H)$ minimiert wird, wobei $\operatorname{dist}(x, H)$ den Abstand eines Punktes $x \in V$ zu H bezeichnet.*

Lösung. Jede Hyperebene H wird durch eine Gleichung der Form $\langle x - a, n\rangle = 0$ beschreiben, wobei a den Ortsvektor eines Aufpunktes und n einen Normalenvektor von H bezeichnet; da der Abstand eines beliebigen Punktes x dann durch $|\langle x-a, n\rangle|/\|n\|$ gegeben ist, besteht die Aufgabe darin, Vektoren $a \in V$ und $n \in V \setminus \{0\}$ so zu finden, daß

$$f(a,n) \;:=\; \sum_{i=1}^N \frac{\langle x_i - a, n\rangle^2}{\|n\|^2}$$

minimiert wird. Setzen wir $\widehat{n} := n/\|n\|$ und bezeichnen wir mit $\lambda\widehat{n}$ den Anteil von a in Richtung von n gemäß der Zerlegung $V = \mathbb{R}n \oplus n^\perp$, so gilt offensichtlich $f(a,n) = f(\lambda\widehat{n}, \widehat{n})$; wegen der Kompaktheit von $\{u \in V \mid \|u\| = 1\}$ und wegen $|f(\lambda\widehat{n},\widehat{n})| \to \infty$ für $|\lambda| \to \infty$ garantiert Satz (95.10) die Existenz eines globalen Minimums von f. Wird dieses Minimum an einer Stelle (a,n) angenommen, so gilt $0 = (\nabla_a f)(a,n) = \sum_{i=1}^N 2\langle x_i - a, \widehat{n}\rangle(-\widehat{n}) = -2\langle \sum_{i=1}^N (x_i - a), \widehat{n}\rangle\widehat{n}$ und damit $\langle \overline{x} - a, \widehat{n}\rangle = 0$, wenn $\overline{x} := (\sum_{i=1}^N x_i)/N$ den Schwerpunkt der "Punktwolke" $\{x_1, \dots, x_N\}$ bezeichnet. Wir erhalten also das (intuitiv plausible) Ergebnis, daß die gesuchte Hyperebene durch den Schwerpunkt der Punkte x_i verlaufen muß. Wir können also $a := \overline{x}$ wählen; schreiben wir $\xi_i := x_i - \overline{x}$, so erhalten wir

$$\begin{aligned} f(a,n) \;&=\; f(\overline{x},\widehat{n}) \;=\; \sum_{i=1}^N \langle \xi_i, \widehat{n}\rangle^2 \\ &= \sum_{i=1}^N \left\langle \langle \xi_i, \widehat{n}\rangle \xi_i, \widehat{n} \right\rangle \;=\; \sum_{i=1}^N \left\langle (\xi_i \otimes \xi_i)(\widehat{n}), \widehat{n} \right\rangle; \end{aligned}$$

der Einheitsnormalenvektor $\widehat{n}$ ist also so zu wählen, daß mit $T := \sum_{i=1}^N \xi_i \otimes \xi_i$ der Ausdruck $\langle T\widehat{n}, \widehat{n}\rangle$ minimal wird. Gemäß (67.6) besitzt nun V eine Orthonormalbasis aus Eigenvektoren von T; stellen wir $\widehat{n}$ in Koordinaten bezüglich einer solchen Basis dar, so erkennen wir sofort, daß $\langle T\widehat{n}, \widehat{n}\rangle$ genau dann minimal wird, wenn $\widehat{n}$ ein Eigenvektor von T zum kleinsten Eigenwert ist. Eine Hyperebene löst also genau dann die gestellte Aufgabe, wenn sie durch den Schwerpunkt $\overline{x}$ der gegebenen Datenpunkte x_i verläuft und wenn ihr Normalenvektor ein Eigenvektor von $\sum_i (x_i - \overline{x}) \otimes (x_i - \overline{x})$ zum kleinsten Eigenwert ist. ♦

Die letzte Beispielaufgabe, die wir in diesem Abschnitt behandeln wollen, hat ihren Ursprung in der medizinischen Bildverarbeitung. In manchen medizinischen Anwendungen ist es wichtig, markante Punkte im Körper eines individuellen Patienten mit solchen Punkten in einem standardisierten Modell ("Durchschnittspatient", Gehirnatlas) abzugleichen, und zwar so, daß Kollinearität und Proportionen erhalten bleiben. Sind $P_1, \dots, P_N$ die gegebenen Punkte des individuellen Patienten und $Q_1, \dots, Q_N$ die entsprechenden Punkte im Standardmodell, so läuft dies darauf hinaus, eine affine Transformation T so zu finden, daß die Punkte $T(P_i)$ so nahe wie möglich bei den Punkten Q_i zu liegen kommen, daß also T die Punktfolge $(P_1, \dots, P_N)$ so gut wie möglich auf die Punktfolge $(Q_1, \dots, Q_N)$ abbildet.

(95.20) Aufgabe. *Es seien V ein endlichdimensionaler Skalarproduktraum und $(P_1, \dots, P_N)$ und $(Q_1, \dots, Q_N)$ geordnete Mengen von Punkten in V. Gesucht ist diejenige affine Transformation $T : V \to V$, die den folgenden Ausdruck minimiert*

$$\Phi(T) \;:=\; \sum_{i=1}^n \|T(P_i) - Q_i\|^2 .$$

Lösung. Wir schreiben $Tv = Av + b$ mit einer linearen Abbildung $A : V \to V$ und einem Translationsvektor b; ferner identifizieren wir die Punkte P_i und Q_i mit ihren Ortsvektoren $p_i = \overrightarrow{OP_i}$ und $q_i = \overrightarrow{OQ_i}$. Weil dann

$$\begin{aligned} f(A,b) \;&:=\; \sum_{i=1}^N \|Ap_i + b - q_i\|^2 \\ &=\; \sum_{i=1}^N \big(\|Ap_i - q_i\|^2 + 2\langle Ap_i - q_i, b\rangle + \|b\|^2\big) \end{aligned}$$

minimal wird, müssen die partiellen Ableitungen nach A und b verschwinden. Zunächst gilt $0 = (\nabla_b f)(A,b) = 2\sum_{i=1}^N (Ap_i - q_i) + 2Nb$ und damit $b = -\sum_{i=1}^N (Ap_i - q_i)/N$. Folglich ist T gegeben durch $Tv = Av - (1/N) \cdot \sum_{i=1}^N (Ap_i - q_i)$, also

$$Tv \;=\; A\left(v - \frac{1}{N}\sum_{i=1}^N p_i\right) + \frac{1}{N}\sum_{i=1}^N q_i \;=\; A(v - \widehat{p}) + \widehat{q},$$

wenn wir mit $\widehat{P}$ und $\widehat{Q}$ die Schwerpunkte der Punktmengen $\{P_1, \dots, P_N\}$ und $\{Q_1, \dots, Q_N\}$ bezeichnen. Dies liefert die (anschaulich plausible) Aussage, daß T den Schwerpunkt $\widehat{P}$ der ersten Punktmenge auf den Schwerpunkt $\widehat{Q}$ der zweiten Punktmenge abbildet. Nach einer Translation im Bild- und im Urbildraum dürfen wir annehmen, daß beide Punktmengen den Schwerpunkt im Nullpunkt haben; wir dürfen also $\widehat{p} = \widehat{q} = 0$ annehmen und müssen daher nur noch die Abbildung

$$f(A) \;:=\; \frac{1}{2}\sum_{i=1}^N \|Ap_i - q_i\|^2$$

minimieren. Gemäß (94.15)(c) ist der Gradient von f gegeben durch $(\nabla f)(A) = \sum_{i=1}^N (Ap_i - q_i) \otimes p_i$. Da A optimal ist, gilt dann $\mathbf{0} = (\nabla f)(A) = \sum_{i=1}^N (Ap_i - q_i) \otimes p_i = \sum_{i=1}^N (Ap_i) \otimes p_i - \sum_{i=1}^N q_i \otimes p_i = A(\sum_{i=1}^N p_i \otimes p_i) - \sum_{i=1}^N q_i \otimes$

p_i; es gilt also $Tv = A(v-\widehat{p})+\widehat{q}$, wobei die lineare Abbildung $A : V \to V$ die Gleichung

$$A\big(\sum_{i=1}^{N}(p_i-\widehat{p})\otimes(p_i-\widehat{p})\big) \;=\; \sum_{i=1}^{N}(q_i-\widehat{q})\otimes(p_i-\widehat{p})$$

erfüllt. Ist $d := \dim(V)$ und nehmen wir an, daß sich unter den N Punkten P_i eine Zahl von $d+1$ Punkten in allgemeiner Lage befindet (so daß $\{P_1,\dots,P_N\}$ nicht in einem niedrigerdimensionalen affinen Unterraum von V enthalten ist), so sind d der Vektoren $p_i-\widehat{p}$ linear unabhängig, woraus folgt, daß $\sum_{i=1}^{N}(p_i-\widehat{p})\otimes(p_i-\widehat{p})$ invertierbar ist; in diesem Fall hat dann die Aufgabe eine eindeutige Lösung, nämlich

$$A = \left(\sum_{i=1}^{N}(q_i-\widehat{q})\otimes(p_i-\widehat{p})\right)\left(\sum_{i=1}^{N}(p_i-\widehat{p})\otimes(p_i-\widehat{p})\right)^{-1}.$$

♦

(95.21) Beispiel. *Welche affine Abbildung $T : \mathbb{R}^2 \to \mathbb{R}^2$ bildet die Punkte $P_1 = (5,0)$, $P_2 = (7,0)$, $P_3 = (8,2)$, $P_4 = (7,3)$, $P_5 = (6,1)$ in dem Sinne bestmöglich auf die Punkte $Q_1 = (0,1)$, $Q_2 = (2,3)$, $Q_3 = (1,4)$, $Q_4 = (-1,5)$, $Q_5 = (-2,3)$ ab, daß $\sum_{i=1}^{5}\big(T(P_i)-Q_i\big)^2$ minimal wird?*

Lösung. Die Schwerpunkte der beiden Punktwolken sind $\widehat{P} = (6.6, 1.2)$ und $\widehat{Q} = (0, 3.2)$. Wir berechnen $A = \big(\sum_{i=1}^{N}(q_i-\widehat{q})\otimes(p_i-\widehat{p})\big)\big(\sum_{i=1}^{N}(p_i-\widehat{p})\otimes(p_i-\widehat{p})\big)^{-1}$ gemäß (95.20) und erhalten

$$A \;=\; \frac{1}{119}\begin{bmatrix}153 & -129\\ 68 & 85\end{bmatrix};$$

die (eindeutig bestimmte) affine Transformation mit der angegebenen Optimalitätseigenschaft ist daher gegeben durch $Tv = A(v-\widehat{p})+\widehat{q}$, also

$$T(x,y) \;=\; \frac{1}{119}\begin{bmatrix}153 & -129\\ 68 & 85\end{bmatrix}\begin{bmatrix}x-6.6\\ y-1.2\end{bmatrix} + \begin{bmatrix}0\\ 3.2\end{bmatrix}. \quad ♦$$

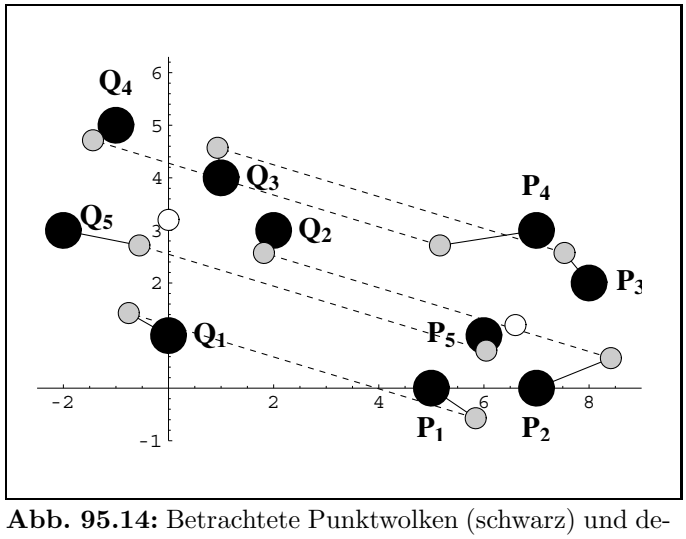

Abb. 95.14: Betrachtete Punktwolken (schwarz) und deren optimaler Abgleich (grau). Die Kreise markieren die Schwerpunkte der beiden Punktwolken.

96. Auflösen von Gleichungen

In diesem Abschnitt wollen wir Mittel der Differentialrechnung benutzen, um Gleichungen zu lösen bzw. Aussagen über die Lösbarkeit gegebener Gleichungen zu gewinnen. Wir beginnen mit Gleichungen der Form $f(x) = 0$, wobei jetzt $f : V \hookrightarrow V$ eine (partielle) Selbstabbildung eines beliebigen Banachraums sein kann. Wie im eindimensionalen Fall kann man durch lokale Linearisierung versuchen, eine solche (nichtlineare) Gleichung iterativ zu lösen. Hat man einen Näherungswert $x^{(0)}$ für die gesuchte Lösung, so ersetzt man f durch die affine Approximation $\widehat{f}(x) := f(x^{(0)}) + f'(x^{(0)})(x-x^{(0)})$ und sucht statt einer Nullstelle von f eine Nullstelle von $\widehat{f}$; ist die lineare Abbildung $f'(x^{(0)})$ invertierbar, so gibt es genau eine solche, nämlich

$$x^{(1)} \;:=\; x^{(0)} - f'(x^{(0)})^{-1}f(x^{(0)}).$$

Ist $x^{(0)}$ "genügend nahe" bei der gesuchten Nullstelle, so darf man hoffen, daß $x^{(1)}$ eine bessere Annäherung an die gesuchte Nullstelle ist als $x^{(0)}$ und daß man bei Iteration des Verfahrens eine Folge besser und besser werdender Näherungswerte erhält.

(96.1) Definition. *Es seien V ein reeller Banachraum und $f : V \to V$ eine differentiierbare Abbildung. Das* **Newtonverfahren** *zur Bestimmung einer Lösung der Gleichung $f(x) = 0$ ist definiert durch Vorgabe eines Startwertes $x^{(0)} \in V$ und die Rekursionsvorschrift*

$$\boxed{x^{(n+1)} \;:=\; x^{(n)} - f'(x^{(n)})^{-1}f(x^{(n)}).}$$

Als **vereinfachtes Newtonverfahren** *zu dem Startwert $x^{(0)} \in V$ bezeichnet man die Iterationsvorschrift*

$$\boxed{x^{(n+1)} \;:=\; x^{(n)} - f'(x^{(0)})^{-1}f(x^{(n)}).}$$

Das vereinfachte Newtonverfahren wird manchmal benutzt, um den Rechenaufwand zur Bestimmung von $f'(x^{(n)})^{-1}$ zu vermeiden (die etwa im Fall $V = \mathbb{R}^n$ die Auswertung von n^2 partiellen Ableitungen und die Inversion einer $(n\times n)$-Matrix erfordert).

(96.2) Beispiel. Wir suchen Lösungen des Systems der beiden Gleichungen $4x^2+9y^2 = 36$ und $13x^2+10xy+13y^2 = 72$, also Nullstellen der Funktion $f : \mathbb{R}^2 \to \mathbb{R}^2$, die gegeben ist durch

$$f(x,y) \;:=\; \begin{bmatrix}4x^2+9y^2-36\\ 13x^2+10xy+13y^2-72\end{bmatrix}.$$

Die Newtoniteration mit dem Startwert $(x_0,y_0) := (3,-1)$ liefert die Ergebnisse in der folgenden Tabelle.

98. Optimierung auf Mannigfaltigkeiten

In diesem Kapitel wollen wir Methoden finden, um Minima und Maxima einer auf einer Mannigfaltigkeit M definierten reellwertigen Funktion $f : M \to \mathbb{R}$ zu bestimmen. Solche Optimierungsaufgaben stellen sich praktisch meist so, daß wir die Minima und Maxima einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$ unter gewissen Nebenbedingungen finden wollen; lassen sich diese Nebenbedingungen als ein reguläres Gleichungssystem $g_1(x) = \cdots = g_m(x) = 0$ schreiben, so sind wir genau in der angegebenen Situation. Ist dabei die betrachtete Mannigfaltigkeit M durch eine Parametrisierung $\varphi : U \to M$ gegeben, so läuft die gestellte Aufgabe einfach darauf hinaus, die Funktion $f \circ \varphi : U \to \mathbb{R}$ zu untersuchen; die Aufgabe reduziert sich dann auf ein Optimierungsproblem ohne Nebenbedingungen, das wir mit den bereits entwickelten Methoden der Differentialrechnung angehen können.

(98.1) Aufgabe. *In dem Gebirge* $z = f(x, y)$ *mit*

$$f(x, y) := (x^2 - y^2)\exp\bigl(-(x^2 + y^2)\bigr)$$

soll ein Rundwanderweg angelegt werden, dessen senkrechte Projektion auf die xy*-Ebene ein Kreis mit fest gegebenem Radius* r *und Mittelpunkt* $(0, 0)$ *sei. Wo liegen die höchsten und tiefsten Punkte auf diesem Rundweg?*

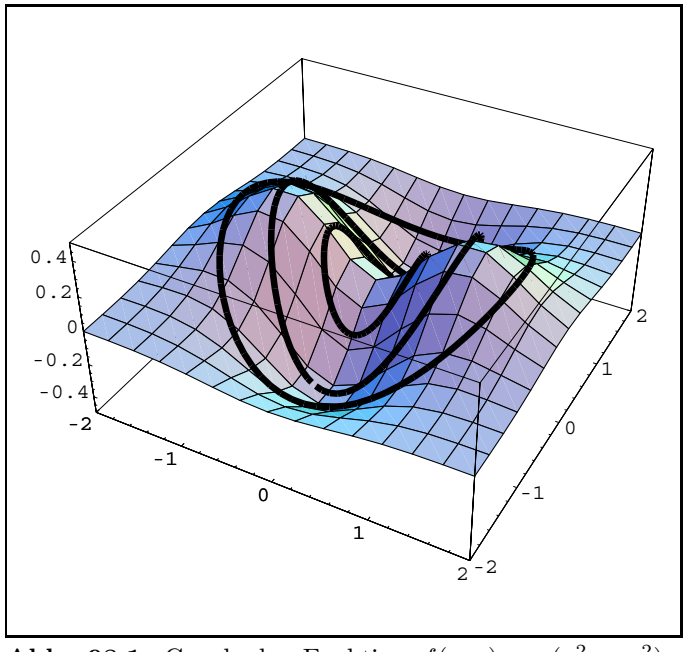

Abb. 98.1: Graph der Funktion $f(x, y) = (x^2 - y^2) \cdot \exp\bigl(-(x^2 + y^2)\bigr)$ mit einigen Rundwanderwegen.

Lösung. Gesucht sind hier die Minima und Maxima der Funktion $f(x, y) = (x^2-y^2)\exp\bigl(-(x^2+y^2)\bigr)$ unter der Nebenbedingung $x^2 + y^2 = r^2$. Die Menge $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\}$ (also die xy-Projektion des Wanderwegs) besitzt die globale Parametrisierung

$$\alpha(t) \;=\; \bigl(x(t), y(t)\bigr) \;=\; \bigl(r\cos(t), r\sin(t)\bigr)$$

mit $0 \le t < 2\pi$; gesucht sind daher die Minima und Maxima der Funktion $\widehat{f}(t) := f\bigl(\alpha(t)\bigr) = r^2(\cos^2 t - \sin^2 t)e^{-r^2}$, also

$$\widehat{f}(t) \;=\; r^2 e^{-r^2}\cos(2t).$$

Wegen $-1 \le \cos(2t) \le 1$ hat $\widehat{f}$ den Minimalwert $-r^2e^{-r^2}$ (dieser wird angenommen für $t = \pi/2$ und $t = 3\pi/2$, also an den Stellen $\alpha(\pi/2) = (0, r)$ und $\alpha(3\pi/2) = (0, -r)$) und den Maximalwert $r^2e^{-r^2}$ (dieser wird angenommen für $t = 0$ und $t = \pi$, also an den Stellen $\alpha(0) = (r, 0)$ und $\alpha(\pi) = (-r, 0)$). ∎

(98.2) Aufgabe. *An welchen Punkten* (x, y, z) *der Einheitssphäre* $x^2 + y^2 + z^2 = 1$ *wird der Ausdruck* $\exp\bigl(2(x + y)z\bigr)$ *minimal bzw. maximal?*

Lösung. Zu optimieren ist die Funktion $f(x, y, z) = \exp\bigl(2(x + y)z\bigr)$ unter der Nebenbedingung $g(x, y, z) = 0$ mit $g(x, y, z) := x^2+y^2+z^2-1$. (Wegen der Stetigkeit von f und der Kompaktheit der Einheitssphäre steht dabei die Existenz eines Maximums bzw. Minimums von vornherein fest.) Wir parametrisieren die Sphäre durch Kugelkoordinaten (Länge u, Breite v) vermöge

$$\varphi(u, v) \;=\; \begin{bmatrix} x(u, v) \\ y(u, v) \\ z(u, v) \end{bmatrix} \;=\; \begin{bmatrix} \cos u\cos v \\ \sin u\cos v \\ \sin v \end{bmatrix}$$

und erhalten dann die in lokalen Koordinaten dargestellte Funktion

$$\begin{aligned} \widehat{f}(u, v) \;&:=\; f\bigl(\varphi(u, v)\bigr) \;=\; \exp\bigl(2\sin v\cos v(\cos u + \sin u)\bigr) \\ &=\; \widehat{f}(u, v) \;=\; \exp\bigl(\sin(2v)\cdot(\cos u + \sin u)\bigr)\,. \end{aligned}$$

Als partielle Ableitungen erhalten wir

$$\begin{aligned} \frac{\partial\widehat{f}}{\partial u} \;&=\; \widehat{f}(u, v)\cdot\sin(2v)\cdot(\cos u - \sin u), \\ \frac{\partial\widehat{f}}{\partial v} \;&=\; \widehat{f}(u, v)\cdot 2\cos(2v)\cdot(\cos u + \sin u). \end{aligned}$$

Beide partiellen Ableitungen sind genau dann gleichzeitig Null, wenn entweder die Bedingungen $\sin(2v) = 0$ und $\cos u = -\sin u$ oder aber die Bedingungen $\cos(2v) = 0$ und $\cos u = \sin u$ erfüllt sind; dies führt auf die acht kritischen Punkte

$$\begin{aligned} p_{1,2} \;&=\; \pm(0,\,0,\,1), \quad p_{3,4} \;=\; \pm\left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0\right), \\ p_{5,6} \;&=\; \pm\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}}\right), \quad p_{7,8} \;=\; \pm\left(\frac{1}{2}, \frac{1}{2}, \frac{-1}{\sqrt{2}}\right), \end{aligned}$$

unter denen die beiden Stellen, an denen die Parametrisierung singulär wird (nämlich Nord- und Südpol), bereits enthalten sind. Wegen $f(p_i) = 1$ für $1 \le i \le 4$ sowie $f(p_5) = f(p_6) = \exp(\sqrt{2}) \approx 4.113$ und $f(p_7) = f(p_8) = \exp(-\sqrt{2}) \approx 0.243$ nimmt die Funktion ihr globales Maximum also in den Punkten $p_{5,6}$ und ihr globales Minimum in den Punkten $p_{7,8}$ an. ∎

Da sich jede Mannigfaltigkeit lokal parametrisieren läßt, können wir zumindest im Prinzip jede Optimierungsaufgabe auf einer Mannigfaltigkeit M so lösen wie die Aufgaben in den beiden vorangegangenen Beispielen. (Existiert keine globale Parametrisierung von M, so müssen wir mehrere verschiedene lokale Parametrisierungen betrachten, und zwar so, daß die zugehörigen Kartenbereiche ganz M überdecken; das ist aber kein prinzipielles Problem.) Ist allerdings eine Mannigfaltigkeit M durch ein reguläres Gleichungssystem $g_1(x) = \cdots = g_m(x) = 0$ gegeben, so ist es oft mühsam oder aufwendig, eine Parametrisierung von M zu finden und dann mit dieser zu arbeiten. Es stellt sich daher die Frage, ob man ein Optimierungsproblem auf M nicht direkt unter Benutzung der Funktionen g_i lösen kann. Der folgende Satz zeigt, daß dies tatsächlich möglich ist.

(98.3) Satz von Lagrange. *Es seien $M \subseteq \mathbb{R}^n$ eine Mannigfaltigkeit und $f : M \to \mathbb{R}$ eine C^1-Funktion. Nimmt f an der Stelle $p \in M$ ein lokales Minimum oder Maximum an und hat M in einer Umgebung von p eine reguläre Darstellung $g_1(x) = \cdots = g_m(x) = 0$, so gibt es Zahlen $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ (sogenannte* **Lagrange-Multiplikatoren***) mit*

$$(\star) \quad (\nabla f)(p) \;=\; \lambda_1 \cdot (\nabla g_1)(p) + \cdots + \lambda_m \cdot (\nabla g_m)(p)\,.$$

Beweis. Für jede in M verlaufende Kurve $t \mapsto \alpha(t)$ mit $\alpha(0) = p$ hat die Funktion $t \mapsto f\big(\alpha(t)\big)$ ein Minimum bzw. Maximum an der Stelle $t = 0$; also gilt

$$0 \;=\; \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} f\big(\alpha(t)\big) \;=\; \langle (\nabla f)(p), \dot\alpha(0)\rangle\,.$$

Da α beliebig war, liegt also $(\nabla f)(p)$ in $(T_pM)^\perp$. Nach (97.17) wird aber $(T_pM)^\perp$ aufgespannt von den Vektoren $(\nabla g_i)(p)$ mit $1 \le i \le m$. ∎

(98.4) Bemerkung. Wir wollen für den Spezialfall einer Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ und einer einzelnen Nebenbedingung $g(x, y) = 0$ eine geometrische Deutung der Lagrangeschen Methode geben. Da aufgrund der vorausgesetzten Regularität $\nabla g \neq 0$ gilt, beschreibt die Gleichung $g(x, y) = 0$ eine Kurve in $\mathbb{R}^2$. Wenn diese in irgendeinem Punkt p transversal zu der Höhenlinie von f durch p verläuft (diese also kreuzt), kann f entlang dieser Kurve kein Minimum oder Maximum annehmen; wir müssen ja von p aus nur ein kleines Stückchen in der einen oder andern Richtung der Kurve laufen, um Höhenlinien von f zu kleineren bzw. größeren Werten zu erreichen. Nimmt also umgekehrt f in p ein Maximum oder Minimum unter der Nebenbedingung $g = 0$ an, so ist dies nur möglich, wenn die Höhenlinien von f and g an der Stelle p tangential zueinander verlaufen. Das bedeutet aber, daß die Gradienten von f und g an dieser Stelle linear abhängig sind (also in die gleiche oder in die exakt gegengesetzte Richtung zeigen).

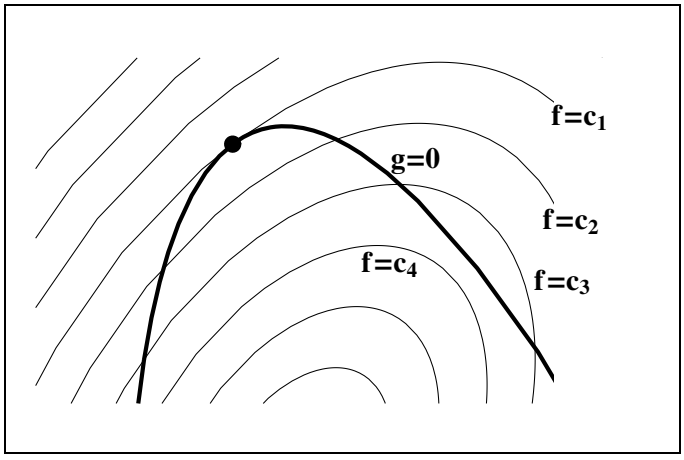

Abb. 98.2: Geometrische Deutung der Methode.

Wir sehen auch, daß aus der Existenz einer Zahl λ mit $(\nabla f)(p) = \lambda \cdot \nabla g(p)$ noch nicht folgt, daß f auf der Menge $g = 0$ tatsächlich ein Extremum annimmt; die Existenz Lagrangescher Multiplikatoren ist also notwendig, aber nicht hinreichend für das Vorliegen eines Extremums.

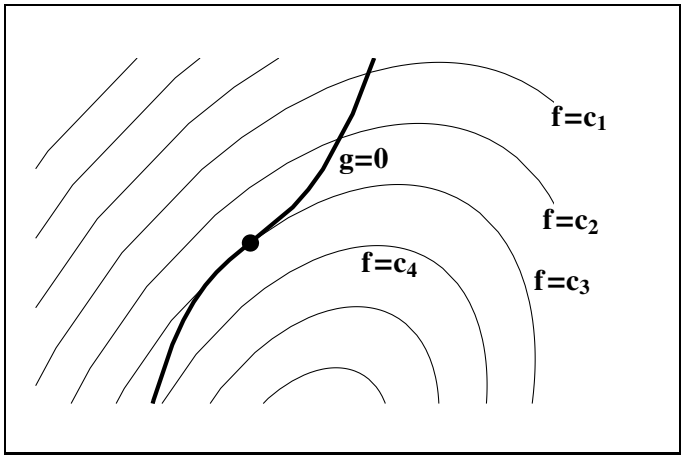

Abb. 98.3: Existenz Lagrangescher Multiplikatoren nicht hinreichend für das Vorliegen eines Extremums.

Auf die Voraussetzung der Regularität der Darstellung $g_1(x) = \cdots = g_m(x) = 0$ kann nicht verzichtet werden.

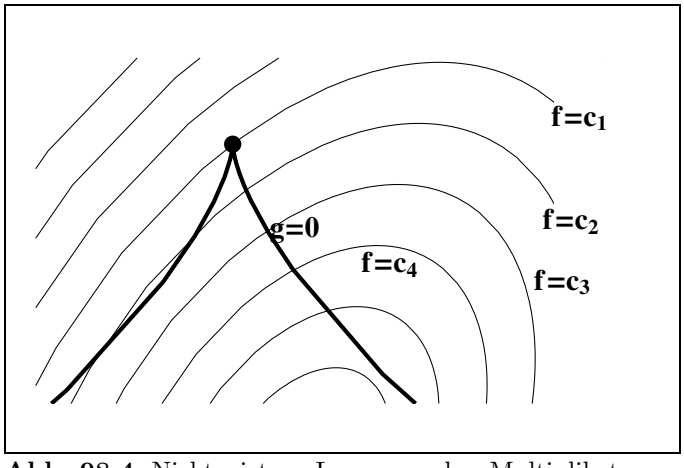

Abb. 98.4: Nichtexistenz Lagrangescher Multiplikatoren trotz vorliegenden Extremums bei fehlender Regularität (im Extrempunkt p gilt $(\nabla g)(p) = 0$).

Wir wollen nun sehen, wie wir Satz (98.3) in Optimierungsproblemen praktisch anwenden können. Die Vektorgleichung $(\star)$ in (98.3) besteht aus n skalaren Gleichungen; zusammen mit den m Nebenbedingungen $g_1(x) = \cdots = g_m(x) = 0$ haben wir dann also $n+m$ Gleichungen für die $n+m$ Unbekannten $x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m$. Die Lagrange-Multiplikatoren λ_i sind dabei nur Mittel zum Zweck; sie zu berechnen ist nur insoweit nötig, als sie zur Ermittlung der Werte $x_1, \ldots, x_n$ führen, und man wird in der Regel versuchen, sie möglichst frühzeitig aus den Gleichungen zu eliminieren.

(98.5) Beispiel. Wir rechnen noch einmal die bereits in (98.2) behandelte Aufgabe, diesmal mit Hilfe der Lagrangeschen Methode. Die Gleichung $\nabla f = \lambda \cdot \nabla g$ nimmt hier die Form

$$2\exp\bigl(2(x+y)z\bigr)\begin{bmatrix} z \\ z \\ x+y \end{bmatrix} \;=\; 2\lambda \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

an. Ist $\lambda = 0$, so erhalten wir $z = 0$ und $x + y = 0$; dies führt auf die Lösungen $p_{3,4}$ in (98.2). Im Fall $\lambda \neq 0$ schließen wir zunächst auf $z \neq 0$ (sonst wäre $x = y = z = 0$ entgegen der Bedingung $x^2 + y^2 + z^2 = 1$) und dann auf $x + y \neq 0$; es ergeben sich dann die Gleichungen

$$\frac{x}{z} \;=\; \frac{y}{z} \;=\; \frac{z}{x+y},$$

die auf $y = x$ und $x/z = z/(2x)$, also $2x^2 = z^2$, führen. Wegen $1 = x^2 + y^2 + z^2 = 4x^2$ ist $x = \pm 1/2$; wir erhalten also $x = y = \pm 1/2$ und $z = \pm 1/\sqrt{2}$, damit also die Punkte $p_5, \ldots, p_8$ aus (98.2). (Wir beachten, daß die Punkte $p_{1,2}$ hier nicht vorkommen; sie sind keine kritischen Punkte von f, sondern traten in (98.2) nur deswegen als kritische Punkte von $\widehat{f}$ auf, weil sie singuläre Stellen der gewählten Parametrisierung durch Kugelkoordinaten sind.) ⧫

(98.6) Aufgabe. *Wie muß man zwei Vektoren $u, v \in \mathbb{R}^2$ der Länge 1 wählen, damit die Determinante der (2×2)-Matrix mit den Spalten u und v möglichst groß bzw. möglichst klein wird?*

Lösung. Schreiben wir $u = (a, b)^T$ und $v = (c, d)^T$, so ist die Funktion $f(a, b, c, d) := ad - bc$ unter den Nebenbedingungen $g_1(a, b, c, d) := a^2 + b^2 - 1 = 0$ und $g_2(a, b, c, d) := c^2 + d^2 - 1 = 0$ zu optimieren. Nach Lagrange machen wir den Ansatz $\nabla f = \lambda_1 \nabla g_1 + \lambda_2 \nabla g_2$, also

$$\begin{bmatrix} d \\ -c \\ -b \\ a \end{bmatrix} \;=\; \lambda_1 \begin{bmatrix} 2a \\ 2b \\ 0 \\ 0 \end{bmatrix} \;+\; \lambda_2 \begin{bmatrix} 0 \\ 0 \\ 2c \\ 2d \end{bmatrix};$$

zusammen mit den Nebenbedingungen $a^2 + b^2 = 1$ und $c^2 + d^2 = 1$ sind dies sechs skalare Gleichungen für die sechs Unbekannten $a, b, c, d, \lambda_1, \lambda_2$. Aus diesen Gleichungen ergeben sich zunächst die Bedingungen $1 = a^2 + b^2 = 4\lambda_2^2(c^2 + d^2) = 4\lambda_2^2$ und $1 = c^2 + d^2 = 4\lambda_1^2(a^2 + b^2) = 4\lambda_1^2$, also $\lambda_1 = \pm 1/2$ und $\lambda_2 = \pm 1/2$. Je nach Vorzeichen ergibt sich für die Matrix $A = (u \mid v)$ dann

$$A \;=\; \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \quad \text{oder} \quad A \;=\; \begin{bmatrix} a & b \\ b & -a \end{bmatrix};$$

im ersten Fall ist A eine Drehung mit der (maximal möglichen) Determinante 1, im zweiten Fall ist A eine Spiegelung mit der (minimal möglichen) Determinante -1.

Das gleiche Ergebnis hätten wir auch ohne die Benutzung der Lagrangeschen Methode erhalten können: wegen $a^2+b^2 = c^2+d^2 = 1$ gibt es Winkel α und β mit $a = \cos\alpha$, $b = \sin\alpha$, $c = \cos\beta$ und $d = \sin\beta$; wegen

$$\det\begin{bmatrix} \cos\alpha & \cos\beta \\ \sin\alpha & \sin\beta \end{bmatrix} = \cos\alpha\sin\beta - \sin\alpha\cos\beta = \sin(\beta - \alpha)$$

nimmt die Determinante ihren maximalen Betrag 1 an, wenn $\beta - \alpha = (\pi/2) + 2k\pi$ gilt, ihren minimalen Betrag -1 für $\beta - \alpha = -(\pi/2) + 2k\pi$. Dies führt auf die Lösungen

$$A \;=\; \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \quad \text{bzw.} \quad A \;=\; \begin{bmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{bmatrix},$$

was mit dem oben erhaltenen Ergebnis übereinstimmt. ⧫

(98.7) Aufgabe. *Wie muß man vier Stäbe mit vorgegebenen Längen a, b, c und d zu einem Viereck legen, damit dieses einen möglichst großen Flächeninhalt bekommt?*

Lösung. Wir legen die vier Stäbe irgendwie zu einem Viereck und bezeichnen mit α, β, γ und δ die entstehenden Winkel (die natürlich alle zwischen 0 und π liegen und ferner die Bedingung $\alpha + \beta + \gamma + \delta = 2\pi$ erfüllen müssen.)

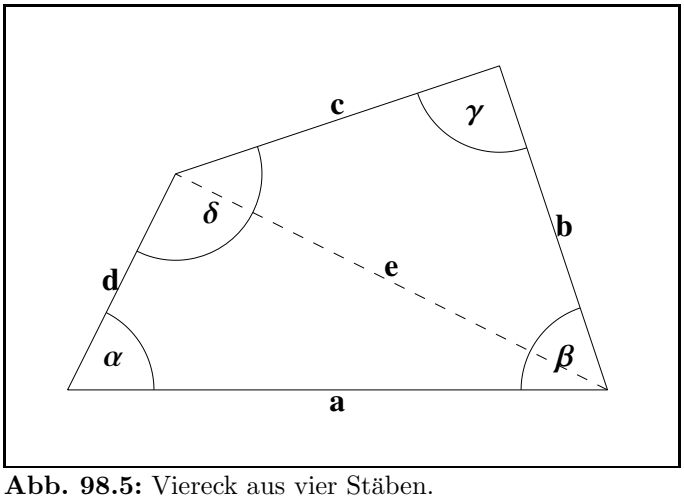

Abb. 98.5: Viereck aus vier Stäben.

Der Flächeninhalt des Vierecks ist dann

$$f(\alpha, \gamma) \;:=\; \frac{1}{2}ad\sin\alpha + \frac{1}{2}bc\sin\gamma.$$

Für die entstehende Diagonale e haben wir nach dem Kosinussatz dann einerseits $e^2 = a^2+d^2-2ad\cos\alpha$, andererseits $e^2 = b^2+c^2-2bc\cos\gamma$; also gilt die Nebenbedingung $g(\alpha,\beta) = 0$ mit

$$g(\alpha,\beta) \;:=\; 2bc\cos\gamma - 2ad\cos\alpha + a^2 + d^2 - b^2 - c^2.$$

Um f unter dieser Nebenbedingung zu maximieren, machen wir nach Lagrange den Ansatz $\nabla f = \lambda\cdot\nabla g$, also

$$\begin{bmatrix} (ad/2)\cdot\cos\alpha \\ (bc/2)\cdot\cos\gamma \end{bmatrix} \;=\; \lambda\cdot\begin{bmatrix} 2ad\sin\alpha \\ -2bc\sin\gamma \end{bmatrix}$$

und damit $\cos\alpha = 4\lambda\sin\alpha$ und $\cos\gamma = -4\lambda\sin\gamma$. Ist $\lambda = 0$, so folgt hieraus $\cos\alpha = \cos\gamma = 0$ und damit $\alpha = \gamma = \pi/2$; ist $\lambda \neq 0$, so ergibt sich $\tan\gamma = -\tan\alpha$ und wegen $0 < \alpha,\gamma < \pi$ damit $\alpha+\gamma = \pi$. In jedem Fall gilt $\gamma = \pi-\alpha$, damit $\sin\gamma = \sin\alpha$ und $\cos\gamma = -\cos\alpha$. Die Bedingung $g(\alpha,\beta) = 0$ führt dann auf $\cos\alpha = (a^2+d^2-b^2-c^2)/(2ad+2bc)$ und $\cos\gamma = -\cos\alpha = (b^2+c^2-a^2-d^2)/(2bc+2ad)$. Wir erhalten also

$$\alpha \;=\; \arccos\left(\frac{a^2+d^2-b^2-c^2}{2(ad+bc)}\right),$$
$$\gamma \;=\; \arccos\left(\frac{b^2+c^2-a^2-d^2}{2(bc+ad)}\right).$$

In völlig analoger Weise ergeben sich die Beziehungen

$$\beta \;=\; \arccos\left(\frac{a^2+b^2-c^2-d^2}{2(ab+cd)}\right),$$
$$\delta \;=\; \arccos\left(\frac{c^2+d^2-a^2-b^2}{2(cd+ab)}\right).$$

Die Bedingungen $\alpha+\gamma = \pi$ und $\beta+\delta = \pi$ besagen, daß die vier Ecken des gesuchten Vierecks auf einem Kreis liegen müssen, wenn der Flächeninhalt maximal werden soll. ♦

(98.8) Aufgabe. *Gegeben sei eine Matrix $A \in \mathbb{R}^{n\times n}$. Für welchen Vektor x der Länge 1 wird der Ausdruck $\langle x, Ax\rangle$ maximal bzw. minimal?*

Lösung. Gesucht sind Minimum und Maximum der quadratischen Form $f(x) := \langle x, Ax\rangle$ auf der Menge $S = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$, also unter der Nebenbedingung $g(x) := \|x\|^2 - 1 = 0$. (Deren Existenz ist wegen der Stetigkeit von f und der Kompaktheit von S von vornherein gesichert.) Wegen $f(x+th) = \langle x, Ax\rangle + t\langle h, Ax\rangle + t\langle x, Ah\rangle + t^2\langle h, Ah\rangle$ erhalten wir $(\mathrm{d}/\mathrm{d}t)|_{t=0} f(x+th) = \langle h, Ax\rangle + \langle x, Ah\rangle = \langle h, Ax\rangle + \langle A^Tx, h\rangle = \langle (A+A^T)x, h\rangle$ und damit $(\nabla f)(x) = (A+A^T)x$. Wird also das Maximum bzw. Minimum von f an einer Stelle $x_0 \in S$ angenommen, so gibt es nach dem Satz von Lagrange eine Zahl $\lambda \in \mathbb{R}$ mit $(\nabla f)(x_0) = \lambda(\nabla g)(x_0)$, also $(A+A^T)x_0 = 2\lambda x_0$ und damit $\frac{1}{2}(A+A^T)x_0 = \lambda x_0$. Das bedeutet aber, daß x_0 ein Eigenvektor von $\frac{1}{2}(A+A^T)$ zum Eigenwert λ ist. (Wegen $\|x_0\| = 1$ ist x_0 nicht der Nullvektor.) Der minimale bzw. maximale Wert von f wird also angenommen, wenn wir als Argument einen Eigenvektor zum minimalen bzw. maximalen Eigenwert von $\frac{1}{2}(A+A^T)$ einsetzen. (Da von vornherein die Existenz von Maximum und Minimum feststeht, liefert die Lösung dieser Aufgabe einen alternativen Beweis der Existenz reeller Eigenwerte symmetrischer Matrizen.) ♦

(98.9) Problem des Apollonius. *Gegeben seien eine Ellipse und ein Punkt P, der nicht auf der Ellipse liegt. Wie viele Möglichkeiten gibt es, ein Lot von P aus auf die Ellipse zu fällen?*

Lösung. Wir können annehmen, daß die Ellipse in Normalenform $(x/a)^2+(y/b)^2 = 1$ gegeben ist; den Punkt P bezeichnen wir mit $P = (\xi,\eta)$. Gesucht sind dann die kritischen Punkte der Funktion $f(x,y) := (x-\xi)^2+(y-\eta)^2$ unter der Nebenbedingung $g(x,y) = 0$ mit $g(x,y) := (x/a)^2 + (y/b)^2 - 1$. Nach Lagrange gibt es an jedem kritischen Punkt eine Zahl λ mit $\nabla f = \lambda\nabla g$, also

$$\begin{bmatrix} 2(x-\xi) \\ 2(y-\eta) \end{bmatrix} \;=\; \lambda\begin{bmatrix} 2x/a^2 \\ 2y/b^2 \end{bmatrix}$$

bwz. $x-\xi = \lambda x/a^2$ und $y-\eta = \lambda y/b^2$. Auflösen dieser beiden Gleichungen nach x und y liefert

$$(1) \qquad x \;=\; \frac{a^2\xi}{a^2-\lambda} \quad\text{und}\quad y \;=\; \frac{b^2\eta}{b^2-\lambda}.$$

Setzen wir dies in die Nebenbedingung $(x/a)^2+(y/b)^2 = 1$ ein, so ergibt sich die Bedingung

$$\varphi(\lambda) = 1 \quad\text{mit}\quad \varphi(\lambda) := \left(\frac{a\xi}{a^2-\lambda}\right)^2 + \left(\frac{b\eta}{b^2-\lambda}\right)^2;$$

sobald λ aus dieser Bedingung ermittelt wurde, finden wir den zugehörigen Lotfußpunkt (x,y) aus den Gleichungen (1). Die Funktion φ hat Polstellen bei $\lambda = a^2$ und $\lambda = b^2$ und erfüllt die Bedingungen $\varphi > 0$ sowie $\varphi(\lambda) \to 0$ für $\lambda \to \pm\infty$. Die Gleichung $\varphi(\lambda) = 1$ hat daher entweder zwei, drei oder vier Lösungen, je nachdem, ob an der einzigen kritischen Stelle $\lambda_\star$ von φ die Bedingung $\varphi(\lambda_\star) > 1$, $\varphi(\lambda_\star) = 1$ oder $\varphi(\lambda_\star) < 1$ gilt.

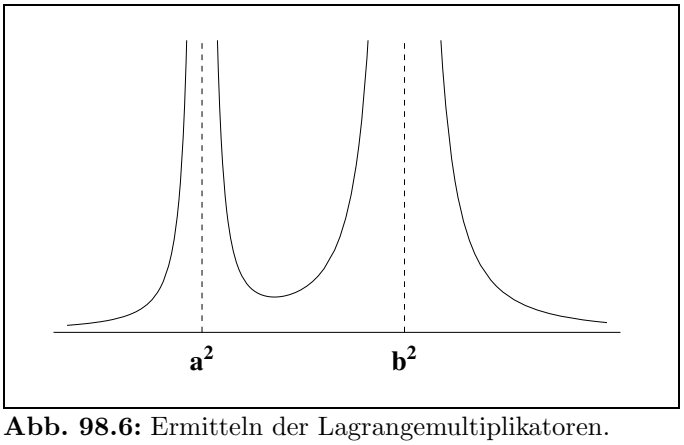

Abb. 98.6: Ermitteln der Lagrangemultiplikatoren.

Die kritische Stelle $\lambda_\star$ erhalten wir durch Nullsetzen der Ableitung

$$\varphi'(\lambda) \;=\; \frac{2a^2\xi^2}{(a^2-\lambda)^3} + \frac{2b^2\eta^2}{(b^2-\lambda)^3};$$

dabei ergibt sich (wenn wir die Sonderfälle $\xi = 0$ und $\eta = 0$ ausnehmen) die Gleichung

$$\frac{a^2-\lambda_\star}{a^{2/3}\xi^{2/3}} \;=\; -\frac{b^2-\lambda_\star}{b^{2/3}\eta^{2/3}} \;=:\; C, \tag{2}$$

wobei $a^2-b^2 = (a^2-\lambda_\star)-(b^2-\lambda_\star) = C\big((a\xi)^{2/3}+(b\eta)^{2/3}\big)$ und damit

$$C \;=\; \frac{a^2-b^2}{(a\xi)^{2/3}+(b\eta)^{2/3}} \tag{3}$$

gilt. Aus (2) folgt $a^2\xi^2(\lambda_\star-b^2)^3 = b^2\eta^2(a^2-\lambda_\star)^3$ bzw. $a^{2/3}\xi^{2/3}(\lambda_\star-b^2) = b^{2/3}\eta^{2/3}(a^2-\lambda_\star)$ mit der eindeutigen Lösung

$$\lambda_\star \;=\; \frac{a^2b^{2/3}\eta^{2/3}+b^2a^{2/3}\xi^{2/3}}{a^{2/3}\xi^{2/3}+b^{2/3}\eta^{2/3}}\,. \tag{4}$$

Es gibt nun genau dann drei mögliche Lotfußpunkte, wenn die Gleichung $\varphi(\lambda_\star) = 1$ erfüllt ist; diese Gleichung läßt sich, wenn wir die aufgrund von (2) gültigen Beziehungen $a^2-\lambda_\star = Ca^{2/3}\xi^{2/3}$ und $b^2-\lambda_\star = -Cb^{2/3}\eta^{2/3}$ benutzen, in der Form $1 = (a\xi)^{2/3}/C^2 + (b\eta)^{2/3}/C^2$ schreiben. Umformen liefert $(a\xi)^{2/3}+(b\eta)^{2/3} = C^2 = (a^2-b^2)^2/\big((a\xi)^{2/3}+(b\eta)^{2/3}\big)^2$, folglich $\big((a\xi)^{2/3}+(b\xi)^{2/3}\big)^3 = (a^2-b^2)^2$ und damit

$$\left(\frac{a\xi}{a^2-b^2}\right)^{2/3} + \left(\frac{b\eta}{a^2-b^2}\right)^{2/3} \;=\; 1. \tag{5}$$

Die Menge aller Punkte (ξ,η), die diese Gleichung erfüllen, nennt man die der Ellipse zugeordnete *Astroide*.

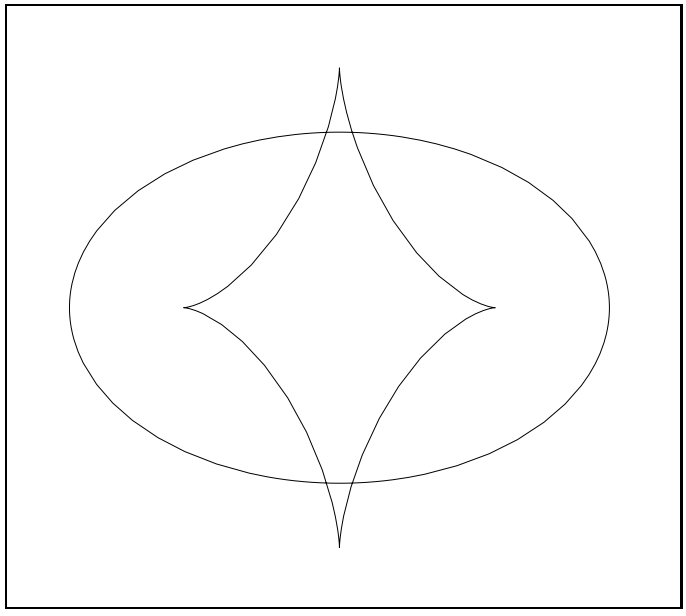

Abb. 98.7: Ellipse und zugeordnete Astroide.

Wir erhalten also das folgende Ergebnis: Liegt der betrachtete Punkt $P = (\xi,\eta)$ auf der Astroide mit der Gleichung (5) (mit Ausnahme der vier Spitzen, die zu den zuvor ausgeschlossenen Sonderfällen $\xi = 0$ bzw. $\eta = 0$ gehören), so lassen sich von P aus genau drei Lote auf die betrachtete Ellipse fällen. Analog zeigt man, daß für einen Punkt außerhalb der Astroide (bzw. in einer deren Spitzen) die Bedingung $\varphi(\lambda_\star) > 1$ gilt (so daß es genau zwei mögliche Lote auf die Ellipse gibt), während für einen Punkt innerhalb der Astroide die Bedingung $\varphi(\lambda_\star) < 1$ gilt (so daß es genau vier mögliche Lote auf die Ellipse gibt). Dieses Ergebnis war bereits Apollonius von Perge (um 262-190 v. Chr.) bekannt – was beinahe unglaublich ist, wenn man sich vergegenwärtigt, daß ihm die von uns benutzten mathematischen Techniken nicht zur Verfügung standen! ♦

Wir wollen nun noch ein Ergebnis herleiten, das in verschiedenen Anwendungsbereichen (etwa in der Bildverarbeitung) eine Rolle spielt, nämlich die bestmögliche Approximation eines Vektorraumendomorphismus durch einen orthogonalen (bzw. im komplexen Fall unitären) Operator. Um dieses Ergebnis herzuleiten, benötigen wir zunächst den folgenden Hilfssatz.

(98.10) Hilfssatz. *Es sei V ein Euklidischer Raum. Bezeichnen wir mit $\mathfrak{A}$ die Menge aller schiefadjungierten und mit $\mathfrak{B}$ die Menge aller selbstadjungierten Endomorphismen von V, so gilt bezüglich des Frobeniusproduktes die direkte Zerlegung $\operatorname{Hom}(V,V) = \mathfrak{A}\oplus\mathfrak{B}$ als orthogonale direkte Summe reeller Vektorräume.*

Beweis. Es ist leicht nachzurechnen, daß $\mathfrak{A}$ und $\mathfrak{B}$ *reelle* Untervektorräume von $\operatorname{Hom}(V,V)$ sind. Da wir jeden Endomorphismus $T : V \to V$ in der Form

$$T \;=\; \underbrace{(1/2)(T+T^\star)}_{\in\,\mathfrak{A}} + \underbrace{(1/2)(T-T^\star)}_{\in\,\mathfrak{B}}$$

schreiben können, gilt $\operatorname{Hom}(V,V) = \mathfrak{A}+\mathfrak{B}$; ferner ist $\mathfrak{A}\cap\mathfrak{B} = \{0\}$, denn für $T \in \mathfrak{A}\cap\mathfrak{B}$ gilt sowohl $T^\star = T$ als auch $T^\star = -T$ und damit $T = \mathbf{0}$. Damit gilt $\operatorname{Hom}(V,V) = \mathfrak{A}\oplus\mathfrak{B}$. Es bleibt zu zeigen, daß die Summe orthogonal ist. Für $A \in \mathfrak{A}$ und $B \in \mathfrak{B}$ gilt nun $\langle\!\langle A,B\rangle\!\rangle = \operatorname{tr}(A^\star B) = \operatorname{tr}\big((-A)B^\star\big) = -\operatorname{tr}(AB^\star) = -\langle\!\langle A,B\rangle\!\rangle$ und damit $\langle\!\langle A,B\rangle\!\rangle = 0$. ■

Wir sind nun in der Lage, die folgende Aufgabe zu lösen.

(98.11) Aufgabe. *Es seien V ein Euklidischer Raum und $T : V \to V$ ein Endomorphismus von V. Finde denjenigen unitären Operator $U : V \to V$, der T bezüglich der Frobeniusnorm auf $\operatorname{Hom}(V,V)$ bestmöglich approximiert!*

Lösung. Es sei $\mathbb{U}$ die Menge der unitären Endomorphismen von V; zu minimieren ist dann die Funktion $f : \mathbb{U} \to \mathbb{R}$, die definiert ist durch $f(U) := \|T-U\|^2 =$

$\|T\|^2 - 2\langle\langle T,U\rangle\rangle + \|U\|^2 = \|T\|^2 - 2\langle\langle T,U\rangle\rangle + \dim(V)$ (wobei wir die Beziehung $\|U\|^2 = \operatorname{tr}(U^\star U) = \operatorname{tr}(\mathbf{1}) = \dim(V)$ ausnutzten). Offensichtlich ist die Minimierung von f äquivalent zur Maximierung von

$$\varphi(U) \;:=\; \langle\langle T,U\rangle\rangle\;.$$

Wegen der Kompaktheit von $\mathbb{U}$ ist dabei die Existenz eines globalen Maximums (sagen wir an einer Stelle U_0) von vornherein gesichert. Für jede zweimal differentiierbare Kurve $t \mapsto U(t)$ in $\mathbb{U}$ mit $U(0) = U_0$ gelten dann die Bedingungen

$$(1)\qquad 0 \;=\; \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \varphi\big(U(t)\big) \;=\; \langle\langle T, \dot U(0)\rangle\rangle,$$

$$(2)\qquad 0 \;\geq\; \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \varphi\big(U(t)\big) \;=\; \langle\langle T, \ddot U(0)\rangle\rangle.$$

Wir wählen speziell $U(t) := U_0 \exp(tX)$ mit einem beliebigen schiefadjungierten Operator X und erhalten $\dot U(t) = U_0 \exp(tX)X$ und $\ddot U(t) = U_0 \exp(tX)X^2$, folglich $\dot U(0) = U_0 X$ und $\ddot U(0) = U_0 X^2$. Zunächst gilt nach (1) die Gleichung $0 = \langle\langle U_0 X, T\rangle\rangle = \operatorname{tr}(X^\star U_0^\star T) = -\langle\langle X, U_0^\star T\rangle\rangle$. Da X beliebig war, bedeutet dies, daß $U_0^\star T$ senkrecht auf jedem schiefadjungierten Operator steht, nach dem Hilfssatz also selbstadjungiert ist. Also gilt $U_0^\star T = S$ bzw. $T = U_0 S$ mit $S^\star = S$; es ist dann $S^2 = S^\star S = T^\star U_0 U_0^\star T = T^\star T$, also S eine Quadratwurzel aus $T^\star T$. Sind also $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ die Eigenwerte von $T^\star T$, so sind die (betragsmäßig angeordneten) Eigenwerte von S gerade die Zahlen $\lambda_i = \pm\sqrt{\sigma_i}$; da der zu maximierende Ausdruck gerade

$$\varphi(U_0) \;=\; \langle\langle T, U_0\rangle\rangle \;=\; \operatorname{tr}(U_0^\star T) \;=\; \operatorname{tr}(S) \;=\; \sum_{i=1}^n \lambda_i$$

ist, wird das Maximum angenommen, wenn wir jeweils $\lambda_i = \sqrt{\sigma_i}$ wählen, also S als positiv semidefinite Quadratwurzel aus $T^\star T$ wählen. Ist T und damit S invertierbar, so erhalten wir die eindeutige Lösung

$$U_0 \;=\; T(\sqrt{T^\star T})^{-1}\,.$$

Wegen (2) gilt ferner für jedes schiefadjungierte X die Bedingung

$$(\star)\quad 0 \;\geq\; \langle\langle T, U_0 X^2\rangle\rangle \;=\; \langle\langle U_0 S, U_0 X^2\rangle\rangle \;=\; \langle\langle S, X^2\rangle\rangle.$$

Wählen wir speziell $X := u \otimes v - v \otimes u$, wobei $u, v \in V$ irgendwelche aufeinander senkrecht stehenden Vektoren sind, so erhalten wir

$$\begin{aligned} X^2 &= (u \otimes v - v \otimes u) \circ (u \otimes v - v \otimes u) \\ &= \langle u, v\rangle u \otimes v + \langle v, u\rangle v \otimes u - \|u\|^2 v \otimes v - \|v\|^2 u \otimes u \\ &= -\|u\|^2 v \otimes v - \|v\|^2 u \otimes u; \end{aligned}$$

Einsetzen in $(\star)$ zeigt dann, daß

$$\begin{aligned} 0 \;&\leq\; \|u\|^2 \langle\langle S, v \otimes v\rangle\rangle + \|v\|^2 \langle\langle S, u \otimes u\rangle\rangle \\ &=\; \|u\|^2 \langle Sv, v\rangle + \|v\|^2 \langle Su, u\rangle \end{aligned}$$

gilt, wann immer u und v aufeinander senkrecht stehen. Ist nun $(e_1, \ldots, e_n)$ eine Orthonormalbasis von V aus Eigenwerten von S zu den Eigenvektoren $\lambda_1, \ldots, \lambda_n$ und wählen wir speziell $v = e_k$ und $u = e_\ell$, so ergibt sich $0 \leq \lambda_k + \lambda_\ell$ für alle $k \neq \ell$, was nur möglich ist, wenn entweder alle $\lambda_i \geq 0$ sind (was der oben gefundenen Lösung entspricht) oder aber wenn $\lambda_i > 0$ für $1 \leq i \leq n-1$ und $\lambda_n < 0$ gilt, wenn also der betragsmäßig kleinste Eigenwert negativ ist (was aber nicht den maximal möglichen Wert von $\varphi(U)$ für $U \in \mathbb{U}$ liefert). ⧫

Wir haben bei der bisherigen Diskussion von Optimierungsaufgaben auf Mannigfaltigkeiten nur notwendige, nicht aber hinreichende Bedingungen für das Vorliegen eines Minimums oder Maximums formuliert; in den einzelnen Beispielaufgaben waren immer *ad hoc*-Überlegungen nötig, um tatsächlich einzusehen, ob ein Minimum oder Maximum vorlag. Wir wollen nun (im Fall $r = 2$) die in (95.4) formulierten hinreichenden Kriterien, bei denen die Eigenschaften der Hesse-Matrix der zu optimierenden Funktion f eine entscheidende Rolle spielen, auf die jetzt behandelte Situation übertragen. Dazu müssen wir uns zunächst überlegen, wie sich die Hessesche Matrix einer Funktion in n Variablen unter einem Koordinatenwechsel verhält.

(98.12) Bemerkung. Wir betrachten eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und führen eine Koordinatentransformation $x_i = x_i(y_1, \ldots, y_n)$ durch. Anwendung der Kettenregel zeigt dann die Gültigkeit der Gleichung

$$\frac{\partial^2 f}{\partial y_i \partial y_j} \;=\; \sum_{k,\ell=1}^n \frac{\partial x_k}{\partial y_i}\frac{\partial x_\ell}{\partial y_j}\frac{\partial^2 f}{\partial x_k \partial x_\ell} + \sum_{k=1}^n \frac{\partial f}{\partial x_k}\frac{\partial^2 x_k}{\partial y_i \partial y_j}\,.$$

An einem kritischen Punkt von f gilt nun $\nabla f = 0$, so daß die zweite Summe in diesem Ausdruck verschwindet; die Hessesche Matrix B von f in den neuen Koordinaten hängt also mit der Hesseschen Matrix A in den alten Koordinaten über die Gleichung $B = J^T A J$ zusammen, wobei $J = (\partial x_i/\partial y_j)_{i,j}$ die Jacobi-Matrix der betrachteten Transformation ist. ⧫

(98.13) Satz. *Es sei $g_1(x) = \cdots = g_m(x) = 0$ eine um den Punkt $p \in \mathbb{R}^n$ reguläre Darstellung einer Mannigfaltigkeit M, und es sei $f : \mathbb{R}^n \to \mathbb{R}$ eine in einer Umgebung von p differentiierbare Funktion. Wir definieren $g : \mathbb{R}^n \to \mathbb{R}^m$ durch $g(x) := \big(g_1(x), \ldots, g_m(x)\big)^T$.*

(a) *Genau dann ist p ein kritischer Punkt der Funktion $f\,|_M$, wenn es ein Element $\lambda_0 \in \mathbb{R}^m$ derart gibt, daß (p, λ_0) ein kritischer Punkt der* **Lagrange-Funktion** *$L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ ist, die definiert wird durch*

$$L(x,\lambda) \;:=\; f(x) + \langle \lambda, g(x)\rangle \;=\; f(x) + \sum\nolimits_{i=1}^m \lambda_i g_i(x)\,.$$

(b) *Es seien A die Hessesche Matrix von $f\,|_M$ im Punkt p (ausgedrückt in irgendwelchen lokalen Koordinaten um p), und es sei H die Hessesche Matrix von L*

an der Stelle (p,λ_0). Ist (p,q,s) der Index von A, so ist $(p+m,q+m,s)$ der Index von H. (*Siehe* (58.17) *zum Begriff des Index.*)

(c) *Hat H genau $d+m$ positive Eigenwerte, so ist p ein lokales Maximum von $f\,|_M$. Hat H genau $d+m$ negative Eigenwerte, so ist p ein lokales Minimum von $f\,|_M$. Hat H mehr als m positive und mehr als m negative Eigenwerte, so ist p ein Sattelpunkt von $f\,|_M$.*

Beweis. (a) Genau dann ist (p,λ_0) ein kritischer Punkt von L, wenn sowohl $(\nabla_\lambda L)(x,\lambda)=g(x)$ als auch $(\nabla_x L)(x,\lambda) = (\nabla f)(x) + \langle\lambda,(\nabla g)(x)\rangle = (\nabla f)(x) + \sum_{i=1}^m \lambda_i(\nabla g_i)(x)$ an dieser Stelle verschwindet, was nach dem Satz von Lagrange genau dann der Fall ist, wenn p ein kritischer Punkt von $f\,|_M$ ist.

(b) Nach (97.18) können wir in einer Umgebung von p lokale Koordinaten $(y_1,\dots,y_n)$ so wählen, daß $y_1,\dots,y_d$ lokale Koordinaten auf M und $y_{d+1},\dots,y_{d+m}=y_n$ gerade die Werte der Funktionen g_i sind; in diesen lokalen Koordinaten lautet die Nebenbedingung dann einfach $y_{d+1}=\dots=y_n=0$, und die Lagrangefunktion nimmt die Form $L(y,\lambda)=f(y)+\lambda_1 y_{d+1}+\dots+\lambda_m y_{d+m}$ an. In den neuen Koordinaten nimmt die Hessesche Matrix von L (die nach (98.12) kongruent zur Hessematrix in den ursprünglichen Koordinaten ist) dann die Form

$$H \;=\; \begin{bmatrix} A & B & \mathbf{0} \\ B^T & C & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \end{bmatrix}$$

an, wobei $A=(\partial^2 f/\partial y_i\partial y_j)_{i,j=1}^d$ die Hessesche Matrix von $f\,|_M$ ist, wobei B die Matrix der partiellen Ableitungen $\partial^2 f/\partial y_i\partial y_k$ mit $1\le i\le d$ und $d+1\le k\le n$ bezeichnet und wobei $C=(\partial^2 f/\partial y_i\partial y_j)_{i,j=d+1}^n$ ist. Mit

$$T := \begin{bmatrix} \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ -B^T & -\frac{1}{2}C & \mathbf{1} \end{bmatrix} \text{ und } S := \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2}\cdot\mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & \mathbf{1} \\ 0 & -\mathbf{1} & \mathbf{1} \end{bmatrix}$$

sowie $P:=TS$ gilt dann

$$P^T \begin{bmatrix} A & B & 0 \\ B^T & C & \mathbf{1} \\ 0 & \mathbf{1} & 0 \end{bmatrix} P = S^T \begin{bmatrix} A & 0 & 0 \\ 0 & 0 & \mathbf{1} \\ 0 & \mathbf{1} & 0 \end{bmatrix} S = \begin{bmatrix} A & 0 & 0 \\ 0 & -\mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} \end{bmatrix};$$

also hat H genau m mehr positive bzw. negative Eigenwerte als A, während der Eigenwert 0 in A und in H mit gleicher Vielfachheit auftritt.

(c) Ist $d+m$ die Anzahl der positiven (negativen) Eigenwerte von H, so ist nach (b) gerade d die Anzahl der positiven (negativen) Eigenwerte von A; d.h., A ist positiv (negativ) definit. Hat H mehr als m positive und mehr als m negative Eigenwerte, so hat A nach (b) mindestens einen positiven und mindestens einen negativen Eigenwert und ist damit indefinit. Die Behauptungen folgen dann sofort aus (95.4). ∎

(98.14) Beispiel. *Welches sind die lokalen Maxima und Minima der Funktion $f(x,y,z):=x^3+y^3+z^3$ unter der Nebenbedingung $(1/x)+(1/y)+(1/z)=1$?*

Lösung. Wir setzen $g(x,y,z) := (1/x)+(1/y)+(1/z)-1$ und bezeichnen mit M die Nullstellenmenge von g; die Lagrangefunktion ist dann gegeben durch

$$L(x,y,z,\lambda) \;=\; x^3+y^3+z^3+\lambda\left(\frac{1}{x}+\frac{1}{y}+\frac{1}{z}-1\right).$$

Nullsetzen des Gradienten

$$(\nabla L)(x,y,z,\lambda) \;=\; \begin{bmatrix} 3x^2-\lambda/x^2 \\ 3y^2-\lambda/y^2 \\ 3z^2-\lambda/z^2 \\ (1/x)+(1/y)+(1/z)-1 \end{bmatrix}$$

führt auf die Gleichungen $3x^4=3y^4=3z^4=\lambda$ und $(1/x)+(1/y)+(1/z)=1$ mit den Lösungen $x=y=z=3$, $\lambda=243$ und $x=y=-z=1$, $\lambda=3$ sowie den beiden Lösungen, die aus der letzteren durch Vertauschung der Rollen von x und z bzw. y und z hervorgehen. Die Funktion f hat also auf der Nullstellenmenge von g die vier kritischen Punkte $(3,3,3)$, $(1,1,-1)$, $(1,-1,1)$ und $(-1,1,1)$. Um den Charakter dieser kritischen Punkte festzustellen, betrachten wir die Hessesche Matrix $(\operatorname{Hess} L)(x,y,z,\lambda)$, die gegeben ist durch

$$\begin{bmatrix} 6x+2\lambda/x^3 & 0 & 0 & -1/x^2 \\ 0 & 6y+2\lambda/y^3 & 0 & -1/y^2 \\ 0 & 0 & 6z+2\lambda/z^3 & -1/z^2 \\ -1/x^2 & -1/y^2 & -1/z^2 & 0 \end{bmatrix}.$$

Nun hat

$$(\operatorname{Hess} L)(3,3,3,243) \;=\; \begin{bmatrix} 36 & 0 & 0 & -1/9 \\ 0 & 36 & 0 & -1/9 \\ 0 & 0 & 36 & -1/9 \\ -1/9 & -1/9 & -1/9 & 0 \end{bmatrix}$$

das charakteristische Polynom $p(\lambda)=(\lambda-36)^2(\lambda^2-36\lambda-1/27)$ und damit die Eigenwerte $\lambda_{1,2}=36$ und $\lambda_{3,4}=18\pm\sqrt{18^2+1/27}$, also drei positive Eigenwerte und einen negativen Eigenwert und damit die Signatur 2. Nach Satz (98.13) hat also auch die Hesseform von $f\,|_M$ die Signatur 2, ist also positiv definit; an der Stelle $(3,3,3)$ liegt daher ein lokales Minimum von $f\,|_M$ vor. Dagegen hat

$$(\operatorname{Hess} L)(1,1,-1,3) \;=\; \begin{bmatrix} 12 & 0 & 0 & -1 \\ 0 & 12 & 0 & -1 \\ 0 & 0 & -12 & -1 \\ -1 & -1 & -1 & 0 \end{bmatrix}$$

das charakteristische Polynom $p(\lambda)=(\lambda-12)\,q(\lambda)$ mit $q(\lambda):=\lambda^3-147\lambda-12$ und damit den Eigenwert 12 sowie jeweils einen Eigenwert in jedem der Intervalle $(-\infty,-1)$, $(-1,0)$ und $(0,\infty)$ (weil dort q jeweils einen Vorzeichenwechsel hat). Also hat $(\operatorname{Hess} L)(1,1,-1,3)$ zwei positive und zwei negative Eigenwerte; nach (98.13) hat dann die Hesseform von $f\,|_M$ einen positiven und einen negativen Eigenwert. An der Stelle $(1,1,-1)$ liegt also ein Sattelpunkt von $f\,|_M$ vor. Gleiches gilt für $(1,-1,1)$ und $(-1,1,1)$. ♦

Das Maximum von $|f'''|$ wird nun am Rand des Integrationsintervalls (also bei $x = 0$ oder $x = 1$) oder an den Nullstellen von f'''' (also bei $x = \pm\sqrt{3\pm\sqrt{6}}$) angenommen. Die einzige Nullstelle von f'''' innerhalb des Integrationsintervalls ist $\xi = \sqrt{3-\sqrt{6}} \approx 0.74196$; wegen $f'''(0) = 0$, $f'''(1) \approx 1.21306$ und $f'''(\xi) \approx 1.38012$ gilt also $\max_{0\le x\le 1} |f'''(x)| = f'''(\xi) \approx 1.38012$. Die erste Fehlerabschätzung für die Simpsonregel lautet also

$$|\text{Fehler}| \;\le\; \frac{1^4}{192\,m^3}\cdot 1.38012 \;=\; \frac{1.38012}{192\,m^3}\,.$$

Da wir $m = 2$ gewählt hatten, ist der Fehler also garantiert kleiner als $1.38012/(192\cdot 8) \approx 0.00089852$; wir haben also $\int_0^1 e^{-x^2/2}\mathrm{d}x = 0.855651 \pm 0.00089852$.

Das Maximum von $|f''''|$ wird entweder an den Randpunkten des Integrationsintervalls (also bei $x = 0$ oder $x = 1$) oder an den Nullstellen von $f^{(5)}$ (also bei $x = 0$ oder $x = \pm\sqrt{5\pm\sqrt{10}}$) angenommen. Von den Nullstellen liegt nur 0 im Integrationsintervall; wegen $f''''(0) = 3$ und $f''''(1) = -2/\sqrt{e}$ gilt dann $\max_{0\le x\le 1} |f''''(x)| = 3$. Die zweite Fehlerabschätzung für die Simpsonregel lautet also

$$|\text{Fehler}| \;\le\; \frac{1^5}{2880\,m^4}\cdot 3 \;=\; \frac{1}{960\,m^4}\,.$$

Da wir $m = 2$ gewählt hatten, ist der Fehler also garantiert kleiner als $1/(960\cdot 16) \approx 0.0000651042$; wir haben also sogar $\int_0^1 e^{-x^2/2}\mathrm{d}x = 0.855651 \pm 0.0000651042$. In diesem Fall (wie überhaupt in den meisten Fällen, wenn hohe Genauigkeit gefordert wird) ist die zweite Fehlerabschätzung also besser als die erste. ♦

Auch bei der praktischen Anwendung der Simpsonregel muß man sich zunächst überlegen, wie fein die Partition des Integrationsintervalls gewählt werden muß, um eine geforderte Genauigkeit garantieren zu können.

(110.15) Beispiel. Wir wollen mit Hilfe der Simpsonregel das Integral $\int_0^1 e^{-x^2/2}\mathrm{d}x$ mit einer Genauigkeit von $\varepsilon = 10^{-6}$ berechnen. In (110.14) haben wir bereits hergeleitet, daß bei einer Zerlegung in m gleich große Doppelstreifen der Fehler kleiner als $1.38012/(192m^3)$ und auch kleiner als $1/(960m^4)$ sein muß. Ein Fehler kleiner als ε kann daher garantiert werden, wenn wir m so groß wählen, daß entweder

$$m \;>\; \sqrt[3]{\frac{1.38012}{192\,\varepsilon}} \;=\; \sqrt[3]{\frac{1\,380\,120}{192}} \;\approx\; 19.3$$

oder

$$m \;>\; \sqrt[4]{\frac{1}{960\,\varepsilon}} \;=\; \sqrt[4]{\frac{1\,000\,000}{960}} \;\approx\; 5.7$$

gilt. Es genügen also $m = 6$ Doppelstreifen bzw. $n = 12$ Einzelstreifen. (Man vergleiche dies mit dem in (110.10) erhaltenen Ergebnis für die Trapezregel!) ♦

111. Berechnung von Mehrfachintegralen

In den vorhergehenden Abschnitten entwickelten wir schlagkräftige Methoden, um Integrale mit einem eindimensionalen Integrationsbereich zu berechnen. Wir wollen nun einen (in voller Allgemeinheit erstmals von dem italienischen Mathematiker Guido Fubini (1879-1943) formulierten) Satz beweisen, der es erlaubt, ein (n_1+n_2)-dimensionales Integral als Iteration eines n_1-dimensionalen und eines n_2-dimensionalen Integrals darzustellen. Durch $(n-1)$-fache Anwendung dieses Satzes läßt sich dann ein n-dimensionales Integral durch iterierte Berechnung n eindimensionaler Integrale effektiv ermitteln. Bevor wir verschiedene Versionen dieses Satzes formulieren und beweisen, wollen wir kurz die wesentliche Idee verdeutlichen und betrachten dazu zwei Intervalle $I, J \subseteq \mathbb{R}$ und eine Funktion $f : I\times J \to [0,\infty)$; das Integral $\iint_{I\times J} f(x,y)\,\mathrm{d}(x,y)$ ist dann das Volumen zwischen der xy-Ebene und dem Graphen von f. Für jedes feste $x \in I$ ist $Q(x) := \int_J f(x,y)\mathrm{d}y$ die Fläche des Querschnitts dieses Volumens mit der Ebene $\{x\}\times\mathbb{R}\times\mathbb{R}$. Ersetzen wir diesen (zweidimensionalen) Querschnitt gedanklich durch eine "unendlich dünne" Scheibe mit der Fläche $Q(x)$ und einer "unendlich kleinen" Dicke $\mathrm{d}x$ und damit dem "unendlich kleinen" Volumen $Q(x)\,\mathrm{d}x$, so ergibt sich das Gesamtvolumen als die Summe dieser unendlich vielen unendlich kleinen Scheibenvolumina: $V = \int_I Q(x)\,dx$, also

$$\iint\limits_{I\times J} f(x,y)\,\mathrm{d}(x,y) \;=\; \int_I \left(\int_J f(x,y)\,\mathrm{d}y\right)\,\mathrm{d}x\,.$$

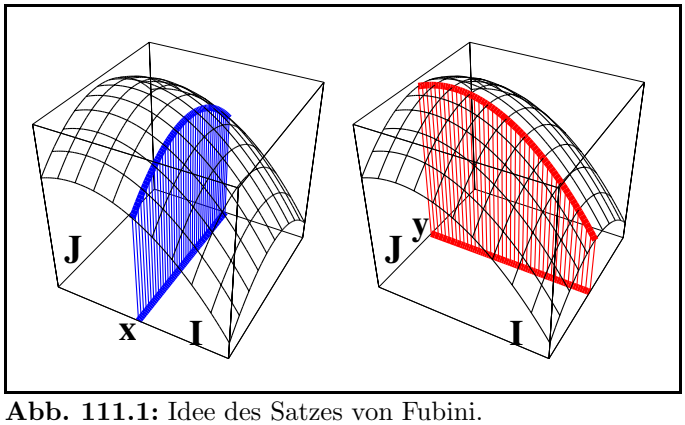

Abb. 111.1: Idee des Satzes von Fubini.

Analog können wir für jedes feste $y \in J$ den Querschnitt des betrachteten Volumens mit der Ebene $\mathbb{R}\times\{y\}\times\mathbb{R}$ ins Auge fassen. Dieser hat den Flächeninhalt $\widehat{Q}(y) := \int_I f(x,y)\,\mathrm{d}x$; ersetzen wir ihn durch eine unendlich dünne Scheibe mit "infinitesimaler Dicke" $\mathrm{d}y$, so ergibt sich $V = \int_J \widehat{Q}(y)\,\mathrm{d}y$, also

$$\iint\limits_{I\times J} f(x,y)\,\mathrm{d}(x,y) \;=\; \int_J \left(\int_I f(x,y)\,\mathrm{d}x\right)\,\mathrm{d}y\,.$$

(Es kommt also nicht darauf, ob wir das Volumen durch Scheiben in Längs- oder in Querrichtung zerlegen, wenn

wir das Gesamtvolumen als Summe von Scheibenvolumina berechnen wollen.) Die Umsetzung dieser anschaulich einleuchtenden Idee in mathematisch hieb- und stichfeste Sachverhalte erfordert einige technische Voraussetzungen. Wir beweisen zunächst eine (schwache) Version des Satzes von Fubini für Riemannsche Integrale.

(111.1) Satz von Fubini (Version für Riemannsche Integrale). *Es seien U, V, W endlichdimensionale reelle Vektorräume, $X \subseteq U$ und $Y \subseteq V$ Jordanmeßbare Teilmengen sowie $f : X \times Y \to W$ eine Riemannintegrierbare Funktion.*

(a) *Existiert für jedes $x \in X$ das Riemannsche Integral $F(x) := \int_Y f(x,y)\,\mathrm{d}y$, so ist $F : X \to W$ Riemannintegrierbar mit $\iint_{X\times Y} f = \int_X F$, also*

$$\iint\limits_{X\times Y} f(x,y)\,\mathrm{d}(x,y) \;=\; \int_X \left(\int_Y f(x,y)\,\mathrm{d}y\right)\,\mathrm{d}x\,.$$

(b) *Existiert für jedes $y \in Y$ das Riemannsche Integral $G(y) := \int_X f(x,y)\,\mathrm{d}x$, so ist $G : Y \to W$ Riemannintegrierbar mit $\iint_{X\times Y} f = \int_Y G$, also*

$$\iint\limits_{X\times Y} f(x,y)\,\mathrm{d}(x,y) \;=\; \int_Y \left(\int_X f(x,y)\,\mathrm{d}x\right)\,\mathrm{d}y\,.$$

Beweis. Es genügt, Teil (a) zu beweisen, denn (b) folgt sofort aus (a), indem wir die Rollen von X und Y vertauschen. Wir bezeichnen mit μ und ν die Jordaninhalte auf U bzw. V. Das Integral $\iint_{X\times Y} f$ ist Grenzwert Riemannscher Summen $\sum_{i,j} f(\xi_i,\eta_j)\mu(X_i)\nu(Y_j) = \sum_i \Big(\sum_j f(\xi_i,\eta_j)\nu(Y_j)\Big)\,\mu(X_i)$. Wegen der vorausgesetzten Integrierbarkeit von F ist die innere Summe hierbei eine Riemannsche Summe für $F(\xi_i)$; also ist $\iint_{X\times Y} f$ Grenzwert der Riemannschen Summen $\sum_i F(\xi_i)\mu(X_i)$.

Etwas genauer geben wir uns $\varepsilon > 0$ beliebig vor. Da f Riemann-integrierbar ist, gibt es ein $\delta > 0$ mit $\|\iint_{X\times Y} f - \Sigma\| < \varepsilon/2$ für jede Riemannsche Summe Σ von f zu einer Zerlegung von $X \times Y$ mit Feinheit kleiner als δ. Wir behaupten, daß dann $\|\iint_{X\times Y} f - \Sigma^\star\| < \varepsilon$ für jede Riemannsche Summe $\Sigma^\star$ von F gilt, die zu einer Zerlegung von X mit Feinheit kleiner als $\delta/2$ gehört. Ist dies gezeigt, so folgt, daß $F : X \to W$ Riemann-integrierbar ist mit $\int_X F = \iint_{X\times Y} f$, und wir sind fertig.

Es sei $(X_i)_{i=1}^m$ eine Zerlegung von X mit Feinheit kleiner als $\delta/2$. Wir wählen beliebige Elemente $\xi_i \in X_i$. Da nach Voraussetzung jedes der Integrale $F(\xi_i) = \int_Y f(\xi_i,y)\,\mathrm{d}y$ existiert, gibt es eine Zerlegung $(Y_j)_{j=1}^n$ von Y, die so fein ist, daß $\|F(\xi_i) - \sum_j f(\xi_i,\eta_j)\nu(Y_j)\| < \varepsilon/\big(2\mu(X)\big)$ für alle $\eta_j \in Y_j$ gilt (was simultan für die endlich vielen Indices $1 \le i \le m$ geht) und daß außerdem der Durchmesser jeder der Mengen $X_i \times Y_j$ kleiner als δ ist. Dann bilden einerseits die Mengen $X_i \times Y_j$ eine Zerlegung von $X \times Y$ mit Feinheit kleiner als δ, was

$$(1)\quad \Big\|\iint\limits_{X\times Y} f(x,y)\,\mathrm{d}(x,y) - \sum_{i,j} f(\xi_i,\eta_j)\,\mu(X_i)\nu(Y_j)\Big\| \;<\; \frac{\varepsilon}{2}$$

nach sich zieht. Andererseits gilt

$$\begin{aligned}
&\Big\|\sum_{i,j} f(\xi_i,\eta_j)\mu(X_i)\nu(Y_j) - \sum_i F(\xi_i)\mu(X_i)\Big\| \\
&= \Big\|\sum_i \Big(\sum_j f(\xi_i,\eta_j)\nu(Y_j) - \int_Y f(\xi_i,y)\,\mathrm{d}y\Big)\mu(X_i)\Big\| \\
&\le \sum_i \Big\|\sum_j f(\xi_i,\eta_j)\nu(Y_j) - \int_Y f(\xi_i,y)\,\mathrm{d}y\,\Big\|\,\mu(X_i) \\
&< \frac{\varepsilon}{2\mu(X)} \cdot \sum_i \mu(X_i) \;=\; \frac{\varepsilon}{2}\,.
\end{aligned} \tag{2}$$

Aus (1) und (2) folgt mit Hilfe der Dreiecksungleichung die Abschätzung $\|\iint_{X\times Y} f - \sum_i F(\xi_i)\mu(X_i)\| < \varepsilon$ und damit die Behauptung. ∎

(111.2) Bemerkung. Die Voraussetzung der Riemann-Integrierbarkeit von $f(x,\cdot)$ für festes x bzw. von $f(\cdot,y)$ für festes y ist ziemlich stark und beileibe nicht immer erfüllt. (Im allgemeinen garantiert weder die Existenz des Integrals $\iint_{X\times Y} f(x,y)\,\mathrm{d}(x,y)$ die Existenz der iterierten Integrale $\int_X(\int_Y f(x,y)\,\mathrm{d}y)\,\mathrm{d}x$ und $\int_Y(\int_X f(x,y)\,\mathrm{d}x)\,\mathrm{d}y$ noch umgekehrt.) Eine Voraussetzung, die aber in vielen praktisch wichtigen Fällen die Anwendung von Satz (111.1) ermöglicht, ist die Stetigkeit von f, denn eine stetige Funktion f erfüllt die Voraussetzungen von Teil (a) und Teil (b) dieses Satzes. ♦

(111.3) Beispiele. (a) Wir wollen das Volumen des räumlichen Bereichs berechnen, der oberhalb des Rechtecks $[0,1] \times [2,4]$ und unterhalb der Fläche $z = x^2 + y$ liegt, also das Integral

$$I \;:=\; \iint\limits_{[0,1]\times[2,4]} (x^2+y)\,\mathrm{d}(x,y)\,.$$

Anwendung des Satzes von Fubini liefert einerseits

$$\begin{aligned}
I \;&=\; \int_0^1 \left(\int_2^4 (x^2+y)\,\mathrm{d}y\right)\,\mathrm{d}x \;=\; \int_0^1 \left[x^2y + \frac{y^2}{2}\right]_{y=2}^4 \mathrm{d}x \\
&=\; \int_0^1 (4x^2 + 8 - 2x^2 - 2)\,\mathrm{d}x \;=\; \int_0^1 (2x^2+6)\,\mathrm{d}x \\
&=\; \left[\frac{2x^3}{3} + 6x\right]_{x=0}^1 \;=\; \frac{2}{3} + 6 - 0 \;=\; 6\frac{2}{3},
\end{aligned}$$

andererseits

$$\begin{aligned}
I \;&=\; \int_2^4 \left(\int_0^1 (x^2+y)\,\mathrm{d}x\right)\,\mathrm{d}y \;=\; \int_2^4 \left[\frac{x^3}{3} + yx\right]_{x=0}^1 \mathrm{d}y \\
&=\; \int_2^4 (\tfrac{1}{3} + y - 0)\,\mathrm{d}y \;=\; \int_2^4 (y + \tfrac{1}{3})\,\mathrm{d}y \\
&=\; \left[\frac{y^2}{2} + \frac{y}{3}\right]_{y=2}^4 \;=\; 8 + \frac{4}{3} - 2 - \frac{2}{3} \;=\; 6\frac{2}{3}\,.
\end{aligned}$$

(b) Sind $X \subseteq U$ und $Y \subseteq V$ Jordan-meßbare Mengen und ist $f : X \times Y \to \mathbb{R}$ eine stetige Funktion der Form $f(x,y) = f_1(x)f_2(y)$, so gilt

$$\iint\limits_{X\times Y} f(x,y)\,\mathrm{d}(x,y) \;=\; \left(\int_X f_1(x)\,\mathrm{d}x\right)\left(\int_Y f_2(y)\,\mathrm{d}y\right).$$

♦

In vielen Fällen ist der Integrationsbereich nicht ein Quader $X \times Y$, sondern irgendeine Jordan-meßbare Teilmenge $\Omega \subseteq U \times V$. Auch in diesem Fall ist der Satz von Fubini anwendbar: wir können nämlich Jordan-meßbare Mengen $X \subseteq U$ und $Y \subseteq V$ mit $\Omega \subseteq X \times Y$ wählen und dann das Integral $\iint_\Omega f$ in der Form $\iint_{X\times Y} f\chi_\Omega$ schreiben, wobei χ_Ω die charakteristische Funktion der Menge Ω bezeichnet. Die folgenden Beispiele zeigen, wie diese Vorgehensweise praktisch angewandt wird.

(111.4) Beispiele. (a) Wir wollen das Integral $I := \iint_D xy^2\,\mathrm{d}(x,y)$ berechnen, wobei D das Dreieck mit den Ecken $(0,0)$, $(1,0)$ und $(1,1)$ sei; es gilt

$$\begin{aligned} D \;&=\; \{(x,y) \in \mathbb{R}^2 \mid 0 \le x \le 1,\ 0 \le y \le x\} \\ &=\; \{(x,y) \in \mathbb{R}^2 \mid 0 \le y \le 1,\ y \le x \le 1\}. \end{aligned}$$

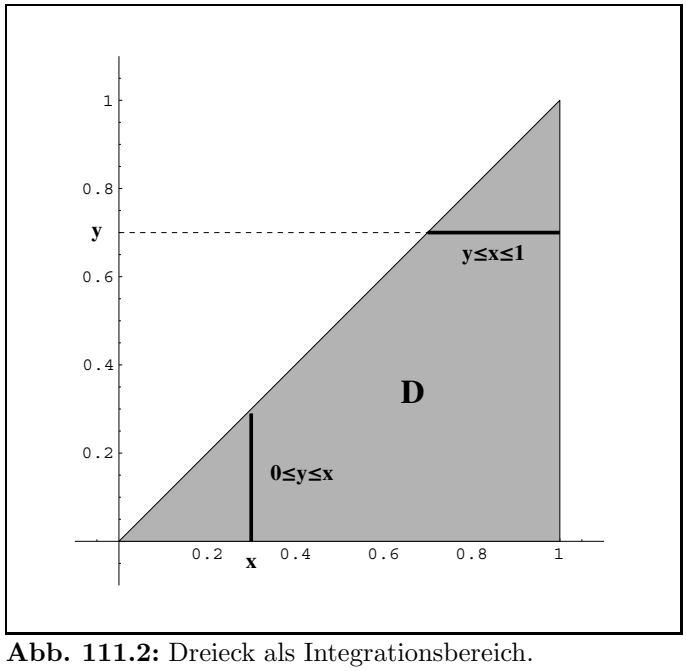

Abb. 111.2: Dreieck als Integrationsbereich.

Wir erhalten bei Verwendung von x als äußerer und y als innerer Integrationsvariablen einerseits

$$\begin{aligned} I \;&=\; \int_0^1 \left(\int_0^x xy^2\,\mathrm{d}y\right)\mathrm{d}x \;=\; \int_0^1 \left[\frac{xy^3}{3}\right]_{y=0}^{x}\mathrm{d}x \\ &=\; \int_0^1 \frac{x^4}{3}\,\mathrm{d}x \;=\; \left[\frac{x^5}{15}\right]_{x=0}^{1} \;=\; \frac{1}{15}, \end{aligned}$$

bei Verwendung von y als äußerer und x als innerer Integrationsvariablen andererseits

$$\begin{aligned} I \;&=\; \int_0^1 \left(\int_y^1 xy^2\,\mathrm{d}x\right)\mathrm{d}y \;=\; \int_0^1 \left[\frac{x^2y^2}{2}\right]_{x=y}^{1}\mathrm{d}y \\ &=\; \int_0^1 \frac{y^2-y^4}{2}\,\mathrm{d}y \;=\; \left[\frac{y^3}{6}-\frac{y^5}{10}\right]_{y=0}^{1} \;=\; \frac{1}{6}-\frac{1}{10} \;=\; \frac{1}{15}. \end{aligned}$$

(b) Wir wollen das Integral $I := \iint_D f(x,y)\mathrm{d}(x,y)$ für den Integranden $f(x,y) := (x^2+y^2)/6$ und den Integrationsbereich

$$\begin{aligned} D \;&:=\; \{(x,y) \in \mathbb{R}^2 \mid 0 \le x \le 2,\ 0 \le y \le x^2\} \\ &=\; \{(x,y) \in \mathbb{R}^2 \mid 0 \le y \le 4,\ \sqrt{y} \le x \le 2\} \end{aligned}$$

berechnen. Lassen wir zunächst x als äußere Variable und dann y als innere Variable laufen, so erhalten wir

$$\begin{aligned} I \;&=\; \int_0^2 \int_0^{x^2} \frac{x^2+y^2}{6}\mathrm{d}y\,\mathrm{d}x \;=\; \int_0^2 \left[\frac{x^2y}{6}+\frac{y^3}{18}\right]_{y=0}^{x^2}\mathrm{d}x \\ &=\; \int_0^2 \left(\frac{x^4}{6}+\frac{x^6}{18}\right)\mathrm{d}x \;=\; \left[\frac{x^5}{30}+\frac{x^7}{126}\right]_{x=0}^{2} \\ &=\; \frac{2^5}{30}+\frac{2^7}{126} \;=\; \frac{656}{315} \;\approx\; 2.08254. \end{aligned}$$

Ändern wir die Integrationsreihenfolge – lassen wir also zunächst y als äußere und dann x als innere Variable laufen –, so erhalten wir

$$\begin{aligned} I \;&=\; \int_0^4 \int_{\sqrt{y}}^2 \frac{x^2+y^2}{6}\mathrm{d}x\,\mathrm{d}y \;=\; \int_0^4 \left[\frac{x^3}{18}+\frac{y^2x}{6}\right]_{x=\sqrt{y}}^{2}\mathrm{d}y \\ &=\; \int_0^4 \left(\frac{8}{18}+\frac{2y^2}{6}-\frac{y^{3/2}}{18}-\frac{y^{5/2}}{6}\right)\mathrm{d}y \\ &=\; \left[\frac{4y}{9}+\frac{y^3}{9}-\frac{y^{5/2}}{45}-\frac{y^{7/2}}{21}\right]_{y=0}^{4} \;=\; \frac{656}{315}. \end{aligned}$$

(c) Es sei $I := \iint_D \sqrt{x}\,y\,\mathrm{d}(x,y)$, wobei

$$\begin{aligned} D \;&=\; \{(x,y) \in \mathbb{R}^2 \mid 0 \le x \le 1,\ x^2 \le y \le \sqrt{x}\} \\ &=\; \{(x,y) \in \mathbb{R}^2 \mid 0 \le y \le 1,\ y^2 \le x \le \sqrt{y}\} \end{aligned}$$

den Bereich zwischen den beiden Parabelbögen $y = x^2$ und $y = \sqrt{x}$ bezeichne. Lassen wir zuerst x laufen, so ergibt sich

$$\begin{aligned} I \;&=\; \int_0^1 \int_{x^2}^{\sqrt{x}} \sqrt{x}\,y\,\mathrm{d}y\,\mathrm{d}x \;=\; \int_0^1 \left[\sqrt{x}\cdot\frac{y^2}{2}\right]_{y=x^2}^{\sqrt{x}}\mathrm{d}x \\ &=\; \int_0^1 \left[\frac{x^{3/2}}{2}-\frac{x^{9/2}}{2}\right]\mathrm{d}x \;=\; \left[\frac{x^{5/2}}{5}-\frac{x^{11/2}}{11}\right]_{x=0}^{1} \\ &=\; \frac{1}{5}-\frac{1}{11} \;=\; \frac{6}{55} \;\approx\; 0.109091. \end{aligned}$$

Lassen wir dagegen zuerst y laufen, so ergibt sich

$$\begin{aligned} I &= \int_0^1 \int_{y^2}^{\sqrt{y}} \sqrt{x}y \,\mathrm{d}x\,\mathrm{d}y = \int_0^1 \left[\frac{2x^{3/2}y}{3}\right]_{x=y^2}^{\sqrt{y}} \mathrm{d}y \\ &= \frac{2}{3}\int_0^1 (y^{7/4} - y^4)\,\mathrm{d}y = \frac{2}{3}\left[\frac{4y^{11/4}}{11} - \frac{y^5}{5}\right]_{y=0}^1 \\ &= \frac{2}{3}\left[\frac{4}{11} - \frac{1}{5}\right] = \frac{2}{3}\cdot\frac{9}{55} = \frac{6}{55}. \end{aligned}$$

♦

Wählen wir im Satz von Fubini als Integranden f die charakteristische Funktion χ_Ω einer Jordan-meßbaren Menge $\Omega \subseteq X \times Y$, so ergibt sich der folgende Satz, der auf den italienischen Mathematiker Bonaventura Cavalieri (1598-1647) zurückgeht, der im übrigen Prior eines Klosters des Ordens der Jesuaten (nicht Jesuiten!) war.

(111.5) Satz von Cavalieri (Version für Riemannsche Integrale) *Es seien U und V endlichdimensionale reelle Vektorräume und $\Omega \subseteq U \times V$ eine Jordan-meßbare Menge. Gibt es eine Jordan-meßbare Menge $A \subseteq U$ und für jedes $a \in A$ eine Jordan-meßbare Menge $Q_a \subseteq V$ mit $\Omega = \bigcup_{a\in A}\{a\} \times Q_a$, so gilt (wenn wir mit μ, ν und $\mu \times \nu$ die Jordaninhalte auf U, V und $U \times V$ bezeichnen) die Gleichung*

$$(\mu \times \nu)(\Omega) = \int_A \nu(Q_a)\,\mathrm{d}a\,.$$

Beweis. Wähle Jordan-meßbare Mengen $X \subseteq U$ und $Y \subseteq V$ mit $\Omega \subseteq X \times Y$ und betrachte die Funktion $f = \chi_\Omega : X \times Y \to \mathbb{R}$. Für jedes $x \in X$ existiert dann $F(x) := \int_Y f(x,y)\,\mathrm{d}y$; es gilt nämlich $F(x) = \nu(Q_a)$ für $x = a \in A$ und $F(x) = 0$ für $x \notin A$. Nach dem Satz von Fubini ist daher

$$\begin{aligned} (\mu \times \nu)(\Omega) &= \iint_{X\times Y} \chi_\Omega(x,y)\,\mathrm{d}(x,y) \\ &= \int_X \left(\int_Y \chi_\Omega(x,y)\,\mathrm{d}y\right)\mathrm{d}x \\ &= \int_X \nu(Q_x)\chi_A(x)\,\mathrm{d}x \\ &= \int_A \nu(Q_a)\,\mathrm{d}a\,. \end{aligned}$$

■

Der Satz von Cavalieri besagt also, daß sich der Inhalt einer Menge in einem höherdimensionalen Raum durch Aufintegrieren der Inhalte niedrigdimensionaler Schnitte dieser Menge ermitteln läßt, was in den beiden folgenden Abbildungen illustriert wird. (Dabei haben wir $\dim(U) = 1$ und $\dim(V) = 2$ in Abbildung 111.3, dagegen $\dim(U) = 2$ und $\dim(V) = 1$ in Abbildung 111.4.)

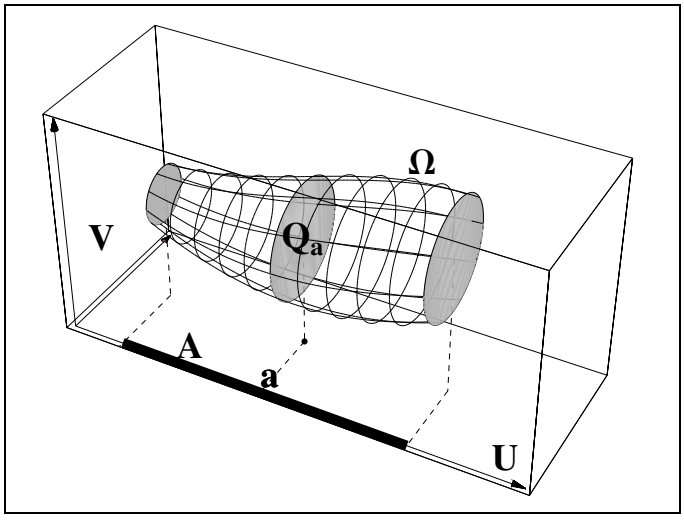

Abb. 111.3: Darstellung eines dreidimensionalen Körpers als Vereinigung zweidimensionaler Schnitte.

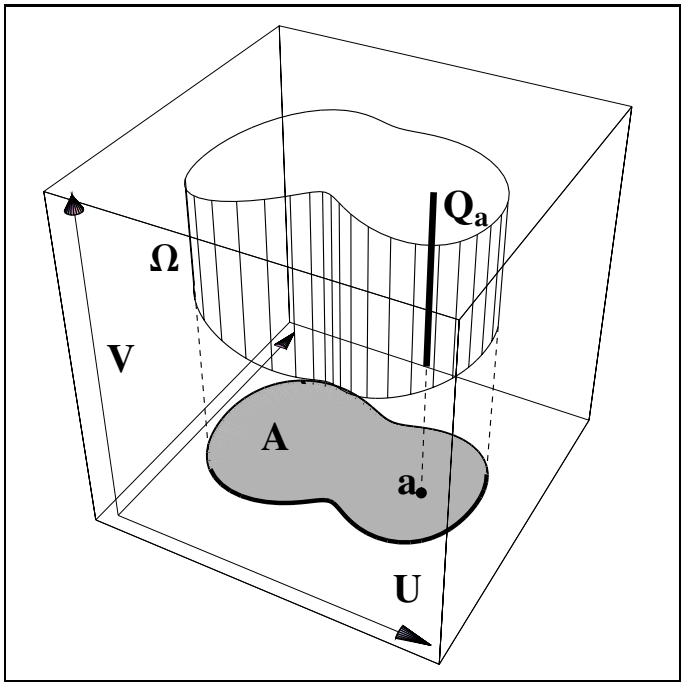

Abb. 111.4: Darstellung eines dreidimensionalen Körpers als Vereinigung eindimensionaler Schnitte.

Das Volumen des Körpers in Abbildung 111.3 ergibt sich demnach, indem man die Flächeninhalte der Mengen Q_a über alle $a \in A$ aufintegriert, was plausibel ist, wenn man sich den Körper aus lauter dünnen Scheiben mit "unendlich kleiner Dicke" – man denke an idealisierte Bierdeckel – zusammengesetzt denkt, deren "infinitesimale Volumina" aufzusummieren sind. Analog ergibt sich das Volumen des Körpers in Abbildung 111.4, indem man die Längen der Mengen Q_a über alle $a \in A$ aufintegriert, was plausibel ist, wenn man sich den Körper aus lauter dünnen Stäben mit "unendlich kleiner Grundfläche" – man denke an idealisierte Zahnstocher – zusammengesetzt denkt, deren "infinitesimale Volumina" dann aufzusummieren sind. Nach dieser heuristischen Erklärung der Aussage des Satzes folgen einige Beispiele.

(111.6) Beispiel. Die in (102.4) hergeleiteten Formeln für die Volumina von Zylindern und Kegeln ergeben sich sofort durch Anwendung des Satzes von Cavalieri. (Man überzeugt sich leicht davon, daß der in (102.4) gegebene Beweis ein Spezialfall des Beweises für den Satz von Cavalieri ist.) Ist beispielsweise ein Kreiskegel mit dem Öffnungswinkel $\alpha \in (0, \pi/2)$ und der Höhe h gegeben und zählen wir eine Variable x von der Kegelspitze aus in Richtung der Symmetrieachse, so ist an jeder Stelle x der Querschnitt senkrecht zur Symmetrieachse ein Kreis mit dem Radius $x \sin\alpha$, also dem Flächeninhalt $Q(x) = \pi x^2 \sin^2\alpha$. Nach dem Satz von Cavalieri ist das Volumen des Kegels dann gegeben durch

$$V = \int_0^h \pi x^2 \sin^2\alpha \,\mathrm{d}x = \pi \sin^2\alpha \int_0^h x^2 \,\mathrm{d}x = \frac{\pi \sin^2\alpha}{3} h^3. \quad \blacklozenge$$

(111.7) Beispiel. Wir betrachten eine Kugel vom Radius r und eine Achse durch den Mittelpunkt dieser Kugel. Zählen wir vom Mittelpunkt aus eine Koordinate x entlang dieser Achse, so ist der Querschnitt der Kugel an der Stelle x ein Kreis mit dem Radius $\sqrt{r^2 - x^2}$ und damit dem Flächeninhalt $Q(x) = \pi(r^2 - x^2)$.

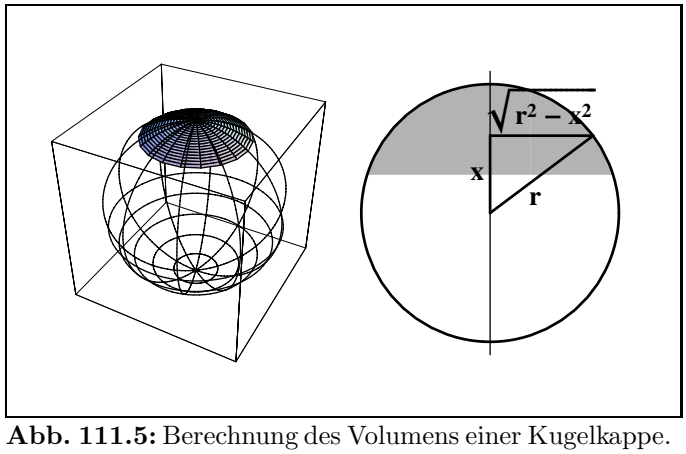

Abb. 111.5: Berechnung des Volumens einer Kugelkappe.

Das Volumen einer Kugelkappe der Höhe h ist nach dem Satz von Cavalieri daher

$$\int_{r-h}^{r} Q(x)\,\mathrm{d}x = \int_{r-h}^{r} \pi(r^2 - x^2)\,\mathrm{d}x = \left[\pi(r^2 x - \frac{x^3}{3})\right]_{x=r-h}^{r}$$
$$= \pi\left(r^3 - \frac{r^3}{3} - r^2(r-h) + \frac{(r-h)^3}{3}\right) = \frac{\pi}{3} h^2 (3r - h).$$

Speziell für $h = 2r$ ergibt sich damit $(4\pi/3) \cdot r^3$ als Volumen einer Kugel vom Radius r. $\blacklozenge$

(111.8) Beispiel. Eine (als Riemann-integrierbar vorausgesetzte) Funktion $f : [a, b] \to [0, \infty)$ sei gegeben; wir betrachten den Rotationskörper, der durch Drehung der Kurve $y = f(x)$ um die x-Achse entsteht. Der Querschnitt dieses Körpers senkrecht zur Drehachse an einer beliebigen Stelle $a \le x \le b$ ist dann ein Kreis mit dem Radius $f(x)$ und damit dem Flächeninhalt $Q(x) = \pi f(x)^2$; nach dem Satz von Cavalieri ist das Volumen des Rotationskörpers dann gegeben durch

$$V = \int_a^b Q(x)\,\mathrm{d}x = \int_a^b \pi f(x)^2\,\mathrm{d}x = \pi \cdot \int_a^b f(x)^2\,\mathrm{d}x.$$
$\blacklozenge$

(111.9) Bemerkung. Die Bezeichnung "Satz von Cavalieri" für den in (111.5) formulierten Sachverhalt ist etwas anachronistisch, denn der Integralbegriff stand Cavalieri gar nicht zur Verfügung. Was dieser explizit formulierte, war das folgende **Prinzip von Cavalieri:** *Wenn zwei Körper in jeder Höhe die gleiche Querschnittsfläche haben, so besitzen sie das gleiche Volumen.* Dieses Prinzip wurde implizit bereits von Archimedes (287-212 v. Chr.) zur Bestimmung des Kugelvolumens benutzt; wir wollen kurz den Beweisgang des Archimedes nachvollziehen. †

Dazu betrachten wir zwei Körper: als ersten Körper eine Halbkugel vom Radius R, als zweiten Körper den Restkörper, der übrigbleibt, wenn wir aus einem Zylinder mit Radius R und Höhe R den Kegel entfernen, dessen Grundseite mit dem Zylinderdeckel und dessen Spitze mit dem Mittelpunkt des Zylinderbodens zusammenfällt.

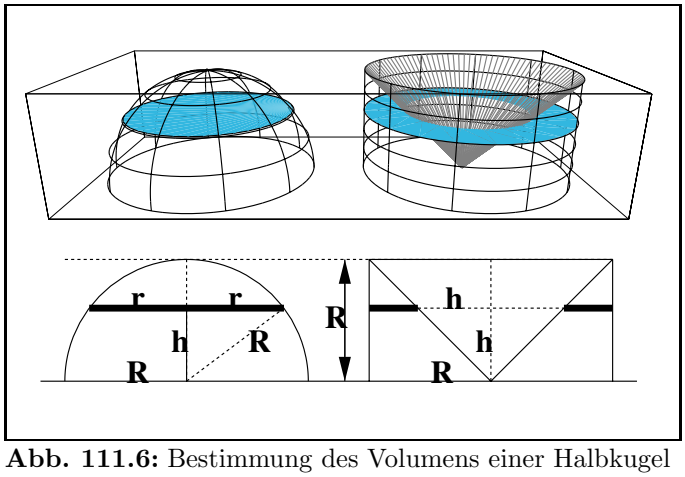

Abb. 111.6: Bestimmung des Volumens einer Halbkugel nach Archimedes.

Diese beiden Körper haben nun in jeder Höhe h mit $0 \le h \le R$ gleiche Querschnittsflächen; der erste Körper hat nämlich als Querschnitt einen Kreis mit dem Radius $r = \sqrt{R^2 - h^2}$ und damit die Querschnittsfläche $\pi r^2 = \pi(R^2 - h^2)$; der zweite Körper hat als Querschnitt einen Kreisring mit dem Außenradius R und dem Innenradius h und damit die Querschnittsfläche $\pi R^2 - \pi h^2$. Nach dem Prinzip von Cavalieri haben daher beide Körper das gleiche Volumen. Das Volumen des zweiten Körpers ergibt sich, indem wir von dem Zylindervolumen $\pi R^2 \cdot R = \pi R^3$ das Kegelvolumens $(1/3) \cdot \pi R^2 \cdot R = \pi R^3/3$ subtrahieren; es ergibt sich $(2/3)\pi R^3$. Das Volumen einer Halbkugel vom Radius R ist daher $(2/3)\pi R^3$ (so daß sich die Volumina

† Die Schrift, in der Archimedes seine Methode beschreibt, wurde erst im Jahr 1907 (wieder)entdeckt. Siehe J. L. Heiberg, H. G. Zeuthen: *Eine neue Schrift des Archimedes*; Bibl. Math., 3. Folge, Band 7, S. 321-363 (1907).

von Kegel, Halbkugel und Zylinder wie 1 : 2 : 3 verhalten); das Volumen einer Vollkugel vom Radius R ist folglich $(4\pi/3)\cdot R^3$.

Archimedes war so stolz auf seine Entdeckung, daß er verfügte, einen Zylinder mit einbeschriebener Kugel und einer entsprechenden Inschrift auf seinem Grab anzubringen (Plutarch, Leben des Marcellus **17**). Cicero beschreibt (Tusculanae disputationes **5**, 64b-66a), wie er im Jahr 75 v. Chr. während seiner Zeit als Quästor in Sizilien das Grab des Archimedes entdeckte und freilegen ließ; vielleicht auch deswegen, weil er den Niedergang der Mathematik bei den Römern schmerzlich empfand. In loc. cit. **1**, 5 schreibt er: "Bei ihnen [den Griechen] wurde die Geometrie in höchsten Ehren gehalten, nichts war glorreicher als die Mathematik; wir aber [die Römer] haben die Nützlichkeit dieser Kunst auf das Ausmessen und Berechnen beschränkt."† ♦

Wir wollen nun eine maßtheoretische Version der Sätze von Fubini und Cavalieri angeben und beweisen. Dabei wird eine Schwäche der zuvor hergeleiteten Versionen für Riemannsche Integrale behoben: die Bedingung, daß für jedes feste $x \in X$ das Integral $\int_Y f(x,y)\,\mathrm{d}y$ existiert, wird nun nicht mehr als *Voraussetzung* gefordert, sondern als Teil der *Behauptung* nachgewiesen. Die Reihenfolge der Herleitung ist anders als zuvor: wir beweisen zunächst den Satz von Cavalieri und leiten aus diesem den Satz von Fubini her (für den er einen Spezialfall darstellt). Ziel ist es also, das Maß einer Menge $Q \subseteq X\times Y$ auszudrücken durch die Maße der Schnitte $Q_x = \{y \in Y \mid (x,y) \in Q\}$ bzw. $Q^y := \{x \in X \mid (x,y) \in Q\}$.

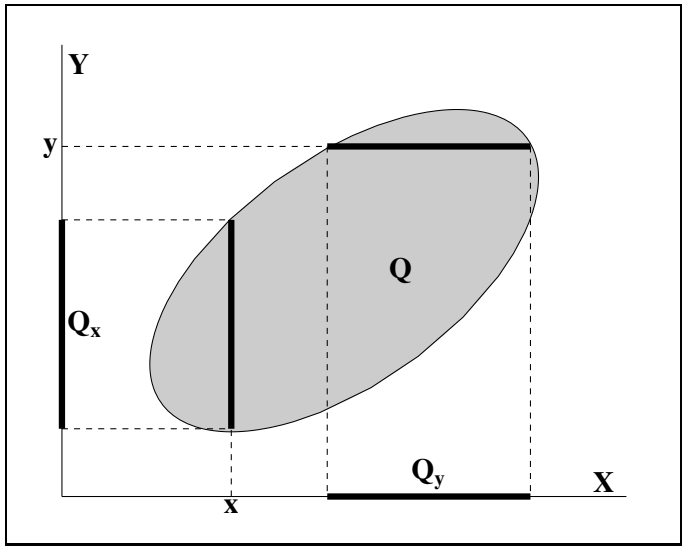

Abb. 111.7: Schnitte in einem Produktraum.

† "The Romans were so uninterested in mathematics that Cicero's act of respect in cleaning up Archimedes' grave was perhaps the most memorable contribution of any Roman to the history of mathematics." (So etwas sarkastisch George F. Simmons, *Calculus Gems*, McGraw-Hill, New York 1992, S. 38.)

Die betrachtete Situation ist die folgende: wir haben Maße μ und ν auf Mengen X bzw. Y und betrachten das Produktmaß $\mu\times\nu$ auf $X\times Y$ (das eindeutig bestimmt ist, wenn X und Y σ-endlich sind, was wir voraussetzen werden). Ist dann $Q \subseteq X\times Y$ eine $(\mu\times\nu)$-meßbare Teilmenge, so sind gemäß (104.24) für alle $x \in X$ und alle $y \in Y$ die Schnitte

$$Q_x := \{y \in Y \mid (x,y) \in Q\} \quad \text{und}$$
$$Q^y := \{x \in X \mid (x,y) \in Q\}$$

meßbar bezüglich ν bzw. μ, so daß es sinnvoll ist, die Funktionen $\varphi : X \to [0,\infty]$ und $\psi : Y \to [0,\infty]$ mit $\varphi(x) := \nu(Q_x)$ und $\psi(y) := \mu(Q^y)$ zu betrachten. Die maßtheoretische Variante des Satzes von Cavalieri lautet dann folgendermaßen.

(111.10) Satz von Cavalieri. *Es seien $(X,\mathfrak{A},\mu)$ und $(Y,\mathfrak{B},\nu)$ zwei σ-endliche Maßräume, und es sei $\mu\times\nu$ das Produktmaß auf $X\times Y$. Für jede $(\mu\times\nu)$-meßbare Teilmenge $Q \subseteq X\times Y$ sind dann die Funktionen $\varphi_Q(x) := \nu(Q_x)$ und $\psi_Q(y) := \mu(Q^y)$ meßbar, und es gilt*

$$(\mu\times\nu)(Q) \;=\; \int_X \varphi_Q\,\mathrm{d}\mu \;=\; \int_Y \psi_Q\,\mathrm{d}\nu\,.$$

Bemerkung. Das Integral $\int_X \varphi_Q\,\mathrm{d}\mu$ nimmt ausgeschrieben die Form

$$\int_X \nu(Q_x)\,\mathrm{d}\mu(x) \;=\; \int_X \left[\int_Y \chi_Q(x,y)\,\mathrm{d}\nu(y)\right]\mathrm{d}\mu(x)$$

an, das Integral $\int_Y \psi_Q\,\mathrm{d}\nu$ dagegen die Form

$$\int_Y \mu(Q^y)\,\mathrm{d}\nu(y) \;=\; \int_Y \left[\int_X \chi_Q(x,y)\,\mathrm{d}\mu(x)\right]\mathrm{d}\nu(y)\,.$$

Beweis. Es sei $\mathfrak{F}$ die Familie aller $(\mu\times\nu)$-meßbaren Teilmengen von $X\times Y$, für die φ_Q und ψ_Q meßbar sind mit $\int_X \varphi_Q\,\mathrm{d}\mu = \int_Y \psi_Q\,\mathrm{d}\nu$; wir müssen zeigen, daß *jede* $(\mu\times\nu)$-meßbare Teilmenge von $X\times Y$ in $\mathfrak{F}$ liegt.

(1) Jedes Rechteck $Q = A\times B$ mit $A \in \mathfrak{A}$ und $B \in \mathfrak{B}$ liegt in $\mathfrak{F}$. Für $Q = A\times B$ haben wir nämlich

$$Q_x = \begin{cases} B, & \text{falls } x \in A \\ \emptyset, & \text{falls } x \notin A \end{cases} \quad \text{und} \quad Q^y = \begin{cases} A, & \text{falls } y \in B \\ \emptyset, & \text{falls } y \notin B \end{cases}$$

und daher $\varphi_Q = \nu(B)\chi_A$ und $\psi_Q = \mu(A)\chi_B$, folglich

$$(\star)\qquad \begin{aligned} \int_X \varphi_Q(x)\,\mathrm{d}\mu(x) \;&=\; \int_A \nu(B)\,\mathrm{d}\mu(x) \;=\; \mu(A)\nu(B) \\ &=\; \int_B \mu(A)\,\mathrm{d}\nu(y) \;=\; \int_Y \psi_Q(y)\,\mathrm{d}\nu(y). \end{aligned}$$

(2) Ist $Q_1 \subseteq Q_2 \subseteq Q_3 \subseteq \cdots$ eine aufsteigende Folge von Mengen Q_i in $\mathfrak{F}$, so liegt auch die Vereinigung $Q := \bigcup_{i=1}^\infty Q_i$ in $\mathfrak{F}$. Schreiben wir nämlich kurz $\varphi_i = \varphi_{Q_i}$, $\varphi = \varphi_Q$, $\psi_i = \psi_{Q_i}$ und $\psi = \psi_Q$, so gelten

$$\left|\frac{f(x)-f(y)}{x-y}\right| = \left|\frac{\sqrt{x}-\sqrt{y}}{x-y}\right| = \frac{1}{\sqrt{x}+\sqrt{y}} \le \frac{1}{2\sqrt{\varepsilon}} =: L$$

und damit $|f(x)-f(y)| \le L\cdot|x-y|$ für alle $x,y \ge \varepsilon$. (Insbesondere ist damit f stetig in jedem Punkt $x_0 > 0$.) Dagegen ist f als Funktion $f : [0,\infty) \to \mathbb{R}$ *nicht* Lipschitz-stetig, denn der Ausdruck

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

geht für $x \to 0$ und $y \to 0$ offensichtlich gegen Unendlich, bleibt also nicht durch eine Lipschitzkonstante L beschränkt. (Dagegen ist f auch stetig im Punkt $x_0 := 0$, denn zu gegebenem $\varepsilon > 0$ müssen wir ja nur $\delta := \varepsilon^2 > 0$ wählen, um aus $|x-x_0| < \delta$ schon $|f(x)-f(x_0)| < \varepsilon$ folgern zu können.)

(d) Ist V ein beliebiger normierter Raum, so ist die Normabbildung $\|\cdot\| : V \to \mathbb{R}$ Lipschitz-stetig mit der Lipschitzkonstanten 1. Für alle $x,y \in V$ erhalten wir nach der Dreiecksungleichung nämlich einerseits $\|x\| = \|x-y+y\| \le \|x-y\| + \|y\|$, also $\|x\| - \|y\| \le \|x-y\|$, andererseits in völlig analoger Weise auch $\|y\| - \|x\| \le \|y-x\| = \|x-y\|$. Diese beiden Ungleichungen ergeben zusammen die Abschätzung

$$\Big|\,\|x\| - \|y\|\,\Big| \;\le\; \|x-y\|$$

und damit die Behauptung. Insbesondere sind also die Betragsfunktionen auf $\mathbb{R}$ und $\mathbb{C}$ Lipschitz-stetig mit der Lipschitzkonstanten 1.

(e) Wie das in (d) aufgeführte Beispiel der Betragsfunktion $f(x) = |x|$ zeigt, stört ein "Knick" im Graphen einer Funktion nicht die Stetigkeit dieser Funktion. Ein "Sprung" im Funktionsgraphen signalisiert dagegen immer eine Unstetigkeitsstelle; ein solcher Sprung tritt etwa auf, wenn wir das Volumen einer fest gegebenen Wassermenge als Funktion der Temperatur betrachten.

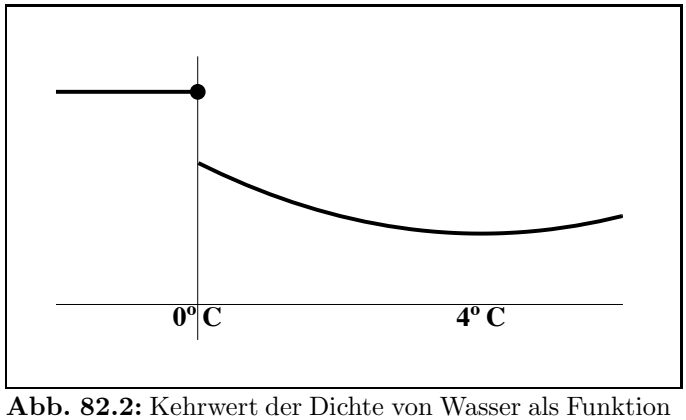

Abb. 82.2: Kehrwert der Dichte von Wasser als Funktion der Temperatur.

Wer die Unstetigkeit dieser Funktion handfest erleben möchte, möge eine randvolle Wasserflasche über Nacht in den Gefrierschrank legen und am nächsten Morgen nachschauen!

(f) Am 1. Januar 2006 war laut Preisverzeichnis der Deutschen Post das Porto y eines Inlandsbriefes (in Cent) als Funktion der Masse x des Briefes (in g) durch die folgende Vorschrift gegeben:

$$f(x) := \begin{cases} 55, & 0 < x \le 20 \text{ ("Standardbrief")},\\ 90, & 20 < x \le 50, \text{ ("Kompaktbrief")},\\ 145, & 50 < x \le 500 \text{ ("Großbrief")},\\ 220, & 500 < x \le 1000 \text{ ("Maxibrief")}. \end{cases}$$

Diese Funktion, aufgefaßt als Abbildung $f : (0,1000] \to \mathbb{R}$, ist unstetig an den Stellen 20, 50 und 500 und stetig an allen anderen Punkten des Definitionsbereichs.

(g) Eine von einer Sprungstelle verschiedene Art der Unstetigkeit offenbart die Funktion $f(x) = \sin(1/x)$, egal, wie wir diese Funktion an der Stelle 0 definieren.

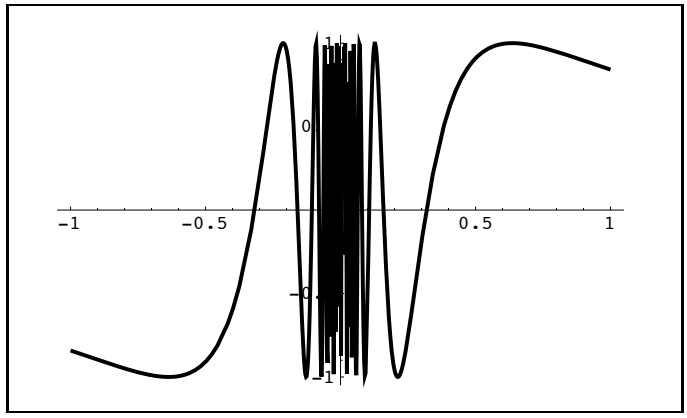

Abb. 82.3: Graph der Funktion $f(x) = \sin(1/x)$.

(h) Eine *lineare* Abbildung $T : V \to W$ zwischen normierten Räumen ist genau dann stetig, wenn sie beschränkt ist, wenn es also eine Konstante C gibt mit $\|Tv\| \le C\|v\|$ für alle $v \in V$. Gibt es eine solche Konstante, so gilt $\|Tv_1 - Tv_2\| = \|T(v_1-v_2)\| \le C\|v_1-v_2\|$, was die Stetigkeit von T zeigt (sogar die Lipschitzstetigkeit mit der Lipschitzkonstanten C). Ist umgekehrt T stetig, so gibt es zu $\varepsilon := 1$ eine Zahl $\delta > 0$ mit $\|Tv\| \le 1$ für alle $v \in V$ mit $\|v\| \le \delta$. Für alle $v \ne 0$ erfüllt dann $\widehat{v} := \delta\cdot v/\|v\|$ die Bedingung $\|\widehat{v}\| = \delta$, folglich $\|T\widehat{v}\| \le 1$; das bedeutet aber $\|Tv\| \le (1/\delta)\|v\|$. Diese letzte Ungleichung gilt natürlich auch für $v = 0$. Mit $C := 1/\delta$ gilt also $\|Tv\| \le C\|v\|$ für alle $v \in V$.

(i) Ist $A \ne \emptyset$ eine Teilmenge eines metrischen Raums (X,d), so ist die Abbildung $X \to \mathbb{R}$ mit $x \mapsto \operatorname{dist}(x,A)$ nach (81.5)(a) Lipschitz-stetig mit der Lipschitzkonstanten $L := 1$. (Dabei sei die Menge $\mathbb{R}$ mit ihrer natürlichen Metrik versehen.) ♦

Wir charakterisieren nun die Stetigkeit einer Abbildung mit Hilfe konvergenter Folgen. Vor dem Durchlesen des formalen Beweises des folgenden Satzes sollte man sich kurz plausibel machen, warum die Aussage dieses Satzes ebenso wie die Definition (82.1) der Stetigkeit die Idee ausdrückt, daß f genau dann stetig an der Stelle x ist, wenn Punkte, die "nahe" bei x liegen, auch Bildwerte haben, die "nahe" bei $f(x)$ liegen.

(82.4) Satz. *Es sei $f : X \to Y$ eine Abbildung zwischen metrischen Räumen. Genau dann ist f stetig an einer Stelle $x \in X$, wenn für jede Folge (x_n) in X mit $x_n \to x$ die Bedingung $f(x_n) \to f(x)$ in Y gilt.*

Beweis. Die Funktion f sei stetig an der Stelle x, und es sei (x_n) eine Folge mit $x_n \to x$; wir müssen zeigen, daß dann $f(x_n) \to f(x)$ gilt. Dazu geben wir uns ein $\varepsilon > 0$ beliebig vor. Es gibt dann ein $\delta > 0$ derart, daß aus $d(\xi, x) < \delta$ in X schon $d\big(f(\xi), f(x)\big) < \varepsilon$ in Y folgt, und wegen $x_n \to x$ gibt es ein $N \in \mathbb{N}$ mit $d(x_n, x) < \delta$ für $n \geq N$. Für alle $n \geq N$ gilt dann $d\big(f(x_n), f(x)\big) < \varepsilon$. Da $\varepsilon > 0$ beliebig war, ist damit $f(x_n) \to f(x)$ gezeigt.

Umgekehrt sei f *nicht* stetig an der Stelle x; wir müssen zeigen, daß es dann eine Folge (x_n) in X gibt mit $x_n \to x$, aber $f(x_n) \not\to f(x)$. Da f nicht stetig an der Stelle x ist, gibt es ein $\varepsilon > 0$, zu dem sich *kein* $\delta > 0$ finden läßt, für das $d(\xi, x) < \delta$ schon $d\big(f(\xi), f(x)\big) < \varepsilon$ impliziert. Für jede Zahl $\delta_n := 1/n$ gibt es also ein Element x_n in X mit $d(x_n, x) < 1/n$, aber $d\big(f(x_n), f(x)\big) \geq \varepsilon$. Die so gefundene Folge (x_n) erfüllt dann $x_n \to x$, aber $f(x_n) \not\to f(x)$. ∎

(82.5) Folgerung. *Alle Potenz-, Wurzel-, Exponential- und Logarithmusfunktionen sind stetig auf ihrem gesamten Definitionsbereich.*

Beweis. Dies folgt mit Hilfe der Folgencharakterisierung (82.4) der Stetigkeit aus (77.14). ∎

Als nächstes zeigen wir, daß jede analytische Funktion auch stetig ist.

(82.6) Satz. *Ist eine Funktion $f : \mathbb{K} \to \mathbb{K}$ in einem Punkt p analytisch, so ist sie in diesem Punkt auch stetig.*

Beweis. Es gelte $f(x) = \sum_{k=0}^{\infty} a_k (x-p)^k$ für $|x-p| < R$. Wählen wir eine Zahl $0 < c < R$, so gilt $\sum_{k=0}^{\infty} |a_k| c^k < \infty$; folglich ist auch $M := \sum_{k=1}^{\infty} |a_k| c^{k-1}$ eine endliche Zahl. Für alle x mit $|x-p| < c$ gilt dann

$$\begin{aligned} |f(x) - f(p)| &= \left| \sum_{k=1}^{\infty} a_k (x-p)^k \right| \\ &= \left| (x-p) \sum_{k=1}^{\infty} a_k (x-p)^{k-1} \right| \\ &\leq |x-p| \sum_{k=1}^{\infty} |a_k| \, |x-p|^{k-1} \\ &\leq |x-p| \sum_{k=1}^{\infty} |a_k| c^{k-1} = M|x-p|. \end{aligned}$$

Diese Abschätzung zeigt sofort, daß f an der Stelle $x = p$ (lokal Lipschitz-stetig und damit) stetig ist. ∎

Als Folgerung ergibt sich ein bemerkenswerter Eindeutigkeitssatz für analytische Funktionen, der besagt, daß eine analytische Funktion schon durch ihre Werte auf einer einzelnen konvergenten Folge festgelegt wird.

(82.7) Eindeutigkeitssatz für analytische Funktionen. *Die Funktionen $f(x) = \sum_{k=0}^{\infty} a_k (x-p)^k$ und $g(x) = \sum_{k=0}^{\infty} b_k (x-p)^k$ seien analytisch im Punkt p, und es sei (x_i) eine gegen p konvergente Folge. Gilt $f(x_i) = g(x_i)$ für alle i, so gilt $a_n = b_n$ für alle $n \in \mathbb{N}_0$.*

Beweis. Wir benutzen Induktion über n. Da f und g nach (82.6) im Punkt p stetig sind, haben wir $a_0 = f(p) = \lim_i f(x_i) = \lim_i g(x_i) = g(p) = b_0$. Ist nun schon gezeigt, daß $a_k = b_k$ für $0 \leq k \leq n-1$ gilt, so stimmen $\sum_{k=n}^{\infty} a_k (x-p)^k$ und $\sum_{k=n}^{\infty} b_k (x-p)^k$ an allen Folgengliedern x_i überein; dies gilt dann (nach Division durch $(x-p)^n$) auch für $F(x) := \sum_{k=n}^{\infty} a_k (x-p)^{k-n}$ und $G(x) := \sum_{k=n}^{\infty} b_k (x-p)^{k-n}$. Anwendung des Stetigkeitssatzes (82.6) auf F und G liefert dann $a_n = F(p) = \lim_i F(x_i) = \lim_i G(x_i) = G(p) = b_n$. ∎

Als nächste Klasse von Funktionen betrachten wir auf einem Intervall $I \subseteq \mathbb{R}$ definierte monotone Funktionen. Solche Funktionen müssen zwar nicht stetig sein, aber man kann für eine monotone Funktion die Menge der möglichen Unstetigkeitsstellen ziemlich genau charakterisieren. Um dies zu tun, geben wir zunächst die folgende Definition.

(82.8) Definition. *Es sei x_0 ein innerer Punkt eines Intervalls $I \subseteq \mathbb{R}$. Eine Funktion $f : I \to V$ hat eine* **Sprungstelle** *im Punkt x_0, wenn die links- und rechtsseitigen Grenzwerte*

$$\lim_{x \to x_0-} f(x) \quad \textit{und} \quad \lim_{x \to x_0+} f(x)$$

zwar beide existieren, aber entweder verschieden sind (in diesem Fall heißt x_0 eine **Unstetigkeitsstelle erster Art***) oder aber gleich sind, aber nicht mit dem Funktionswert $f(x_0)$ übereinstimmen (in diesem Fall sagt man, die Funktion f habe an der Stelle x_0 eine* **hebbare Unstetigkeit***). Existiert mindestens einer der beiden einseitigen Grenzwerte nicht, so nennt man x_0 eine* **Unstetigkeitsstelle zweiter Art**.

Der folgende Satz besagt nun, daß eine monotone Funktion höchstens abzählbar viele Unstetigkeitsstellen hat, und zwar ausschließlich Sprungstellen.

(82.9) Satz. *Eine monotone Funktion $f : I \to \mathbb{R}$ hat keine hebbaren Unstetigkeiten und keine Unstetigkeitsstellen zweiter Art und höchstens abzählbar viele Unstetigkeitsstellen erster Art.*

Beweis. O.B.d.A. sei f monoton wachsend; für monoton fallende Funktionen verläuft der Beweis völlig analog. Es sei $x \in I$ beliebig. Da f monoton wächst, existieren $f_-(x) = \sup_{y<x} f(y)$ und $f_+(x) = \inf_{y>x} f(y)$, und es gilt $f_-(x) \leq f(x) \leq f_+(x)$. Die einzig möglichen Unstetigkeitsstellen von f sind also Sprungstellen. Gilt ferner $x < y$ und ist ξ irgendeine Zahl mit $x < \xi < y$, so gilt

$$(\star) \qquad f_-(x) \leq f_+(x) \leq f(\xi) \leq f_-(y) \leq f_+(y),$$

so daß f_- und f_+ jeweils monoton wachsen. Schließlich sei X die Menge der Unstetigkeitsstellen von f. Dann sind die offenen Intervalle $I_x := \big(f_-(x), f_+(x)\big)$ nach $(\star)$ disjunkt. Da man in jedem dieser Intervalle eine rationale Zahl wählen kann und da es nur abzählbar viele rationale Zahlen gibt, kann es auch nur abzählbar viele solcher Intervalle geben; also ist die Menge X abzählbar. ∎

Wir beweisen nun, daß die Umkehrabbildung einer streng monotonen Funktion wieder stetig ist. Beispielsweise folgt die Stetigkeit der Logarithmus-, Wurzel- und Bogenfunktionen *automatisch* aus der Stetigkeit der Exponential-, Potenz- bzw. Winkelfunktionen.

(82.10) Umkehrsatz für streng monotone Funktionen. *Es seien $I \subseteq \mathbb{R}$ ein Intervall, $f : I \to \mathbb{R}$ eine streng monoton wachsende (fallende) Funktion und $J := f(I) := \{f(x) \mid x \in I\}$ das Bild von f. Dann ist die Umkehrfunktion $f^{-1} : J \to I$ ebenfalls streng monoton wachsend [fallend] und überdies stetig.*

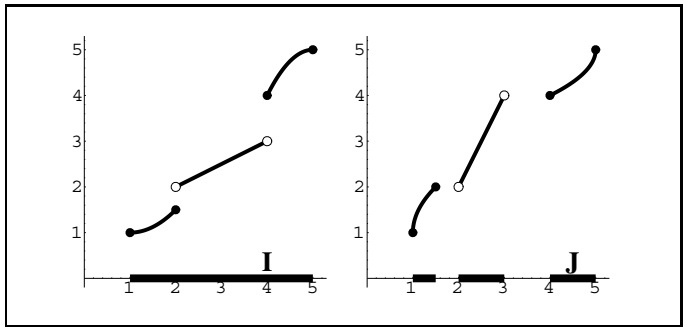

Abb. 82.4: Umkehrfunktion einer streng monotonen Funktion.

Beweis. Wir nehmen an, f sei monoton wachsend; für monoton fallende Funktionen verläuft der Beweis ganz analog. Die Injektivität von f folgt aus der strengen Monotonie; die Surjektivität erzwingen wir dadurch, daß wir f als Funktion $f : I \to J$ auffassen. Damit ist klar, daß die Umkehrabbildung $f^{-1} : J \to I$ existiert; ferner ist f^{-1} wegen der Äquivalenz $x_1 < x_2 \iff f(x_1) < f(x_2)$ selbst wieder streng monoton wachsend. Es bleibt also nachzuweisen, daß f^{-1} stetig ist.

Wir wollen die Stetigkeit von f^{-1} in einem beliebigen Punkt $y_0 \in J$ zeigen. Wir setzen $x_0 := f^{-1}(y_0)$ und geben uns ein beliebiges $\varepsilon > 0$ vor. Wegen der strengen Monotonie von f gilt dann $f(x_0 - \varepsilon) < y_0 < f(x_0 + \varepsilon)$.† Dann können wir ein $\delta > 0$ so wählen, daß auch $f(x_0 - \varepsilon) < y_0 - \delta < y_0 + \delta < f(x_0 + \varepsilon)$ gilt. Für alle $y \in J$ mit $|y - y_0| < \delta$ gilt dann $f(x_0-\varepsilon) < y < f(x_0+\varepsilon)$ und folglich $x_0 - \varepsilon < f^{-1}(y) < x_0 + \varepsilon$ bzw. $-\varepsilon < f^{-1}(y) - f^{-1}(y_0) < \varepsilon$ aufgrund der strengen Monotonie von f^{-1}. Aus $|y - y_0| < \delta$ folgt also $|f^{-1}(y) - f^{-1}(y_0)| < \varepsilon$; dies zeigt die Stetigkeit von f^{-1} im Punkt y_0. ∎

† Ist x_0 linker [rechter] Randpunkt von I, so lassen wir die linke [rechte] Ungleichung einfach weg und führen den folgenden Beweis mit leichten Modifikationen weiter. Wir nehmen hier an, daß x_0 kein Randpunkt von I ist; dann können wir ε immer so klein wählen, daß $x_0 \pm \varepsilon \in I$ gilt.

Wir kehren nun noch einmal zum Beginn dieses Abschnitts zurück, wo wir den Begriff der Stetigkeit einer Funktion f zwischen metrischen Räumen an einer Stelle x_0 dadurch einführten, daß wir zu jeder vorgegebenen Toleranz $\varepsilon > 0$ für den Zielwert $y_0 = f(x_0)$ die Existenz eines zulässigen Spiels $\delta > 0$ derart postulierten, daß eine Abweichung von weniger als δ bei der Einstellung von x_0 zu einer Abweichung von weniger als ε vom Sollzielwert $f(x_0)$ führt. Dabei wird das zulässige Spiel in der Praxis meist nicht nur von der Toleranz ε abhängen (je stringenter die Genauigkeitsanforderung, desto geringer das erlaubte Spiel), sondern auch vom Arbeitspunkt x_0 selbst. (Ein Rennfahrer, der sich einer Kurve mit einer Geschwindigkeit von 300 km/h annähert, kann sich viel geringere Fehler bei der Bewegung des Lenkrades erlauben als der Fahrer eines gewöhnlichen Fahrzeugs im Stadtverkehr.) Das Beispiel der Funktion $f(x) = 1/x$ illustriert diese Abhängigkeit des Spiels vom Arbeitspunkt; als Toleranz wurde $\varepsilon := 0.2$ gewählt, als Arbeitspunkt zunächst $x_0 = 1$, dann $x_0 = 0.5$.

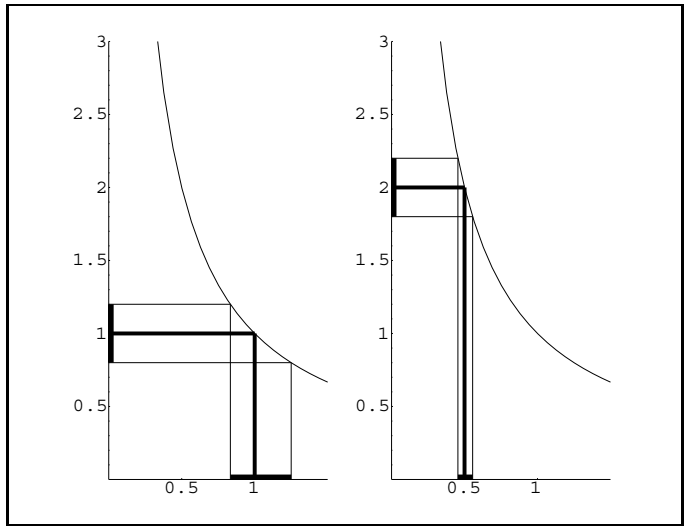

Abb. 82.5: Abhängigkeit des zulässigen Spiels δ vom Arbeitspunkt x_0 bei gleicher Toleranz ε.

Eine Funktion f heißt nun *gleichmäßig stetig* über einen Arbeitsbereich X, wenn zu jeder vorgegebenen erlaubten Toleranz $\varepsilon > 0$ ein (vom speziellen Arbeitspunkt unabhängiges) zulässiges Spiel $\delta > 0$ derart existiert, daß eine Abweichung von einem *beliebigen* Arbeitspunkt $x \in X$ um weniger als δ zu einer Abweichung vom Funktionswert $f(x)$ um weniger als ε führt.

(82.11) Definition. *Eine Abbildung $f : X \to Y$ zwischen metrischen Räumen heißt* **gleichmäßig stetig**, *wenn es zu jeder vorgegebenen Zahl $\varepsilon > 0$ eine Zahl $\delta > 0$ derart gibt, daß aus $d(x_1, x_2) < \delta$ in X stets $d\big(f(x_1), f(x_2)\big) < \varepsilon$ in Y folgt.*

Entscheidend ist, daß das Spiel δ allein in Abhängigkeit von der Toleranz ε (und unabhängig von irgendeinem festen Arbeitspunkt) bestimmt werden kann. Während also Stetigkeit eine lokale Eigenschaft ist (d.h., eine Eigenschaft, die einer Funktion in einem einzelnen Punkt zukommt und nur von den Werten der Funktion in einer beliebig kleinen Umgebung dieses Punktes abhängt), ist gleichmäßige Stetigkeit eine globale Eigenschaft (also eine Eigenschaft, die der Funktion insgesamt – unter Berücksichtigung ihres gesamten Definitionsbereichs – zukommt).

(82.12) Beispiele. (a) Jede Lipschitz-stetige Funktion ist gleichmäßig stetig. Gilt nämlich $d\big(f(x_1),f(x_2)\big) \le L\cdot d(x_1,x_2)$, so kann man zu gegebenem $\varepsilon>0$ stets $\delta := \varepsilon/L$ wählen, um die Implikation $\big(d(x_1,x_2)<\delta \Rightarrow d\big(f(x_1),f(x_2)\big)<\varepsilon\big)$ zu garantieren.

(b) Die durch $f(x)=\sqrt{x}$ definierte Funktion $f:[0,\infty)\to[0,\infty)$ ist gleichmäßig stetig; aus der für alle $x_1,x_2\ge 0$ gültigen Abschätzung

$$|\sqrt{x_1}-\sqrt{x_2}| \;\le\; \sqrt{|x_1-x_2|}$$

folgt nämlich sofort, daß die Wahl $\delta:=\varepsilon^2$ die Gültigkeit der Implikation $\big(|x_1-x_2|<\delta \Rightarrow |f(x_1)-f(x_2)|<\varepsilon\big)$ garantiert.

(c) Die Funktion $f(x)=x^2$ ist gleichmäßig stetig auf jedem Intervall $[-b,b]$ mit festem $b>0$, denn es gilt dann die Lipschitzbedingung

$$\begin{aligned}|f(x_1)-f(x_2)| \;&=\; |x_1^2-x_2^2| \;=\; |x_1+x_2|\,|x_1-x_2|\\ &\le \Big||x_1|+|x_2|\Big|\,|x_1-x_2| \;\le\; 2b|x_1-x_2|\,.\end{aligned}$$

Dagegen ist f nicht gleichmäßig stetig auf ganz $\mathbb{R}$; die Abschätzung

$$\begin{aligned}|f(x_1)-f(x_2)| \;&=\; |x_1+x_2|\,|x_1-x_2|\\ &=\; |2x_1+(x_2-x_1)|\,|x_1-x_2|\\ &\ge\; (2|x_1|-|x_2-x_1|)\,|x_2-x_1|\end{aligned}$$

zeigt nämlich, daß der Abstand $|f(x_1)-f(x_2)|$ beliebig groß werden kann, egal wie dicht x_1 und x_2 beisammen liegen, wenn nur x_1 groß genug gewählt wird.

(d) Die Funktion $f(x)=1/x$ ist gleichmäßig stetig auf jedem Intervall $[a,\infty)$ mit festem $a>0$, denn es gilt dann die Lipschitzbedingung

$$|f(x_1)-f(x_2)| = \left|\frac{1}{x_1}-\frac{1}{x_2}\right| = \frac{|x_1-x_2|}{|x_1|\,|x_2|} \le \frac{1}{a^2}|x_1-x_2|\,.$$

Dagegen ist f auf keinem Intervall der Form $(0,\varepsilon]$ gleichmäßig stetig; die Abschätzung

$$\begin{aligned}|f(x_1)-f(x_2)| \;&=\; \left|\frac{1}{x_1}-\frac{1}{x_2}\right| \;=\; \frac{|x_1-x_2|}{|x_1|\,|x_2|}\\ &\ge \frac{|x_1-x_2|}{|x_1|\,(|x_1-x_2|+|x_1|)}\end{aligned}$$

zeigt nämlich, daß der Abstand $|f(x_1)-f(x_2)|$ beliebig groß werden kann, egal wie dicht x_1 und x_2 beisammen liegen, wenn nur x_1 nahe genug bei 0 liegt.

(e) Sind $f:X\to Y$ und $g:Y\to Z$ zwei gleichmäßig stetige Abbildungen zwischen metrischen Räumen, so ist deren Verkettung $g\circ f: X\to Z$ ebenfalls gleichmäßig stetig. Zu jedem vorgegeben $\varepsilon>0$ gibt es nämlich aufgrund der gleichmäßigen Stetigkeit von g ein $\delta_Y>0$ derart, daß aus $d_Y(y_1,y_2)<\delta_Y$ stets $d_Z\big(g(y_1),g(y_2)\big)<\varepsilon$ folgt; wegen der gleichmäßigen Stetigkeit von f gibt es dann ein $\delta_X>0$ derart, daß aus $d_X(x_1,x_2)<\delta_X$ stets $d_Y\big(f(x_1),f(x_2)\big)<\delta_Y$ folgt. Für alle $x_1,x_2\in X$ mit $d_X(x_1,x_2)<\delta_X$ gilt dann $d_Z\big(g\big(f(x_1)\big),g\big(f(x_2)\big)\big)<\varepsilon$, also $d_Z\big((g\circ f)(x_1),(g\circ f)(x_2)\big)<\varepsilon$. ♦

Wir definieren zum Abschluß dieses Abschnitts noch diejenigen Abbildungen, die metrische Strukturen auf verschiedenen Räumen miteinander identifizieren.

(82.13) Definition. *Eine Abbildung $f:X\to Y$ zwischen metrischen Räumen heißt* **Isometrie** *oder auch* **isometrische Einbettung**, *wenn f abstandserhaltend ist, wenn also $d_Y\big(f(a),f(b)\big)=d_X(a,b)$ für alle $a,b\in X$ gilt. Eine bijektive Isometrie heißt auch* **isometrischer Isomorphismus.**

Eine Isometrie ist stets injektiv, denn aus $f(a)=f(b)$ folgt $0=d_Y\big(f(a),f(b)\big)=d_X(a,b)$ und damit $a=b$. Ist $f:X\to Y$ ein isometrischer Isomorphismus, so ist auch die Umkehrabbildung f^{-1} eine Isometrie. Existiert ein isometrischer Isomorphismus zwischen zwei metrischen Räumen, so sind deren metrische Strukturen ununterscheidbar; jede durch die Metrik ausdrückbare Aussage (etwa hinsichtlich des Abstandes zwischen Punkten oder Teilmengen, der Konvergenz von Folgen oder der Beschränktheit von Mengen) gilt in dem einen Raum genau dann, wenn die entsprechende Aussage in dem anderen Raum gilt. Ein isometrischer Isomorphismus $f:X\to Y$ läßt sich einfach als Umbenennung der Elemente von X auffassen, und (X,d_X) und (Y,d_Y) stellen nur verschiedene Realisierungen des "gleichen" metrischen Raums dar. Das folgende Beispiel zeigt, daß sich jeder metrische Raum isometrisch in einen Funktionenraum einbetten läßt, also isometrisch isomorph zu einem Unterraum eines Funktionenraums ist.

(82.14) Beispiel. Es seien (X,d) ein beliebiger metrischer Raum und $\mathfrak{F}$ die Menge aller Funktionen $f:X\to\mathbb{R}$, die für alle $x,y\in X$ die folgende Bedingung erfüllen:

$$|f(x)-f(y)| \;\le\; d(x,y) \;\le\; f(x)+f(y)\,;$$

dann ist eine Metrik auf $\mathfrak{F}$ definiert durch $D(f,g):=\sup_{x\in X}|f(x)-f(y)|$. Für jedes Element $a\in X$ ist nun eine Funktion $f_a\in\mathfrak{F}$ gegeben durch $f_a(x):=d(x,a)$, und es gilt $D(f_a,f_b)=d(a,b)$ für alle $a,b\in X$. Durch $a\mapsto f_a$ ist also eine Isometrie von X auf einen Unterraum von $\mathfrak{F}$ definiert. (Die Einzelheiten dieses Beispiels sind als Übungsaufgabe zu überprüfen!)

83. Vollständigkeit metrischer Räume

Wir definieren die Vollständigkeit eines metrischen Raums in völliger Analogie zur Definition (73.11) der metrischen Vollständigkeit der Zahlengeraden.

(83.1) Definition. *Ein metrischer Raum heißt* **vollständig***, wenn in ihm jede Cauchyfolge konvergiert.*

Die Bedeutung dieses Begriffes ergibt sich aus dem in vielen praktisch wichtigen Fällen zu beobachtenden Auftreten einer Situation, die etwa folgendermaßen beschrieben werden kann. Man ist daran interessiert, eine gewisse Größe zu ermitteln (Länge einer Kurve, Flächeninhalt einer geometrischen Figur, Nullstelle einer Funktion, Lösung einer irgendwie gearteten Gleichung), kann dies aber nicht direkt tun. Stattdessen ist man in der Lage, Näherungswerte für die gesuchte Größe zu ermitteln und diese sukzessive zu verbessern; man erhält dann eine Folge $(x_1, x_2, x_3, \ldots)$ von Näherungswerten für die eigentlich gesuchte Größe x, und es ist naheliegend zu versuchen, x als Grenzwert der Folge (x_n) zu gewinnen. Wenn nun die Näherungswerte x_n tatsächlich immer bessere Approximationen sind, so werden sie eine Cauchyfolge bilden (was ja gerade heißt, daß für genügend große Indices m und n die Werte x_m und x_n beliebig dicht beieinander liegen), und es ist genau die Eigenschaft der Vollständigkeit, die in dieser Situation garantiert, daß die Folge (x_n) tatsächlich einen Grenzwert besitzt (der dann die Lösung des gesuchten Problems darstellt).

(83.2) Beispiele. (a) Ist $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$, so ist $\mathbb{K}$ mit der natürlichen Metrik $d(x,y) = |x-y|$ vollständig; dagegen ist $\mathbb{Q}$ nicht vollständig.

(b) Der Raum $\mathbb{K}^n$ mit der von der Norm $\|x\| := \sqrt{\sum_{i=1}^n |x_i|^2}$ induzierten Metrik ist vollständig. Ist nämlich $(x^{(k)})_{k=1}^\infty$ eine Cauchyfolge in $\mathbb{K}^n$, so ist für jeden Index $1 \le i \le n$ die Folge $(x_i^{(k)})_{k=1}^\infty$ eine Cauchyfolge in $\mathbb{K}$, folglich wegen der Vollständigkeit von $\mathbb{K}$ konvergent gegen ein Element $x_i \in \mathbb{K}$. Dann konvergiert aber die Folge $(x^{(k)})$ in $\mathbb{K}^n$ gegen das Element $x := (x_1, \ldots, x_n)^T$; vgl. (81.14)(a).

(c) Es seien X eine beliebige Menge, (Y,d) ein vollständiger metrischer Raum und $\mathfrak{B}(X,Y)$ der Raum aller beschränkten Funktionen $f : X \to Y$, versehen mit der Metrik $D(f,g) := \sup_{x\in X} d\big(f(x),g(x)\big)$; wir wollen zeigen, daß $\mathfrak{B}(X,Y)$ vollständig ist, und betrachten eine beliebige Cauchyfolge (f_n) in $\mathfrak{B}(X,Y)$. Dann ist für jedes feste $x \in X$ die Folge $\big(f_n(x)\big)$ eine Cauchyfolge in Y, wegen der Vollständigkeit von Y also konvergent gegen ein Element $f(x) \in Y$. Die Folge (f_n) konvergiert dann punktweise gegen die so definierte Funktion $f : X \to Y$.

Wir zeigen zunächst, daß die Abbildung f beschränkt und damit ein Element von $\mathfrak{B}(X,Y)$ ist. Es seien $a, b \in X$ beliebige Elemente von X. Wegen $f_n(a) \to f(a)$ und $f_n(b) \to f(b)$ für $n \to \infty$ gibt es zu $\varepsilon := 1$ ein $m \in \mathbb{N}$ mit $d\big(f(a), f_n(a)\big) \le 1$ und $d\big(f(b), f_n(b)\big) \le 1$ für alle $n \ge m$. Da f_m eine beschränkte Funktion ist, erhalten wir dann die Abschätzung

$$\begin{aligned} &d\big(f(a), f(b)\big) \\ &\le d\big(f(a), f_m(a)\big) + d\big(f_m(a), f_m(b)\big) + d\big(f_m(b), f(b)\big) \\ &\le 1 + \operatorname{diam}\big(\operatorname{Bild}(f_m)\big) + 1 \;=\; 2 + \operatorname{diam}\big(\operatorname{Bild}(f_m)\big); \end{aligned}$$

da $a, b \in X$ beliebig gewählt waren, gilt also die Abschätzung $\operatorname{diam}\big(\operatorname{Bild}(f)\big) \le 2 + \operatorname{diam}\big(\operatorname{Bild}(f_m)\big) < \infty$, die zeigt, daß die Funktion f beschränkt ist.

Wir wollen weiter zeigen, daß für $n \to \infty$ nicht nur $f_n \to f$ punktweise gilt, sondern sogar $D(f_n, f) \to 0$; ist dies geschafft, so ist $\mathfrak{B}(X,Y)$ als vollständig nachgewiesen. Es sei $\varepsilon > 0$ beliebig vorgegeben. Da (f_n) eine Cauchyfolge ist, gibt es einen Index $N \in \mathbb{N}$ mit $D(f_m, f_n) \le \varepsilon$ für alle $m, n \ge N$ und damit $d\big(f_m(x), f_n(x)\big) \le \varepsilon$ für alle $m, n \ge N$ und jedes fest gewählte Element $x \in X$. Für $m \to \infty$ folgt dann $d\big(f(x), f_n(x)\big) \le \varepsilon$ für alle $n \ge N$ und jedes fest gewählte $x \in X$, folglich $D(f, f_n) \le \varepsilon$ für alle $n \ge N$. Da $\varepsilon > 0$ beliebig war, ist damit die Konvergenzaussage $D(f_n, f) \to 0$ gezeigt.

(d) Ein metrischer Raum, dessen Metrik nur ganzzahlige Werte annehmen kann, ist automatisch vollständig. Eine Folge (x_n) in einem solchen Raum ist nämlich genau dann eine Cauchyfolge, wenn sie "schließlich konstant" ist, wenn es also einen Index $n_0 \in \mathbb{N}$ gibt mit $x_n = x_{n_0}$ für alle $n \ge n_0$, und eine solche Folge ist natürlich konvergent. ♦

Wir wollen nun herausfinden, wann ein Unterraum eines vollständigen metrischen Raums selbst wieder vollständig ist (was uns eine ganze Reihe weiterer Beispiele für vollständige metrische Räume liefern wird). Dazu führen wir zunächst den Begriff der *Abgeschlossenheit* einer Teilmenge eines metrischen Raums ein.

(83.3) Definition. *Es seien (X,d) ein metrischer Raum und A eine Teilmenge von X. Der* **Abschluß** *$\overline{A}$ von A ist die Menge aller derjenigen Elemente von X, die sich als Grenzwert einer Folge von Elementen in A darstellen lassen. Die Elemente von $\overline{A}$ heißen* **Berührpunkte** *von A. Die Menge A heißt* **abgeschlossen** *in X, wenn $\overline{A} = A$ gilt. Die Menge A heißt* **dicht** *in X, wenn $\overline{A} = X$ gilt.*

Wir beobachten, daß stets $A \subseteq \overline{A}$ gilt, denn jedes Element $a \in A$ läßt sich ja als Grenzwert der konstanten Folge $(a, a, a, \ldots)$ in A auffassen. Der folgende Satz gibt nun die gewünschte Charakterisierung der vollständigen Unterräume eines vollständigen metrischen Raums.

(83.4) Satz. *Es sei (X,d) ein vollständiger metrischer Raum. Ein Teilraum $A \subseteq X$ von X (mit der von d induzierten Metrik) ist genau dann vollständig, wenn er abgeschlossen ist.*

Beweis. Wir nehmen an, A sei abgeschlossen. Ist (a_n) eine Cauchyfolge in A, dann auch eine Cauchyfolge in X, wegen der vorausgesetzten Vollständigkeit von X also konvergent in X, sagen wir $a_n \to x$ mit $x \in X$. Da A abgeschlossen ist, muß der Grenzwert in A liegen; sagen wir $x = a \in A$. Dann gilt aber $a_n \to a$ in A. Damit ist gezeigt, daß jede Cauchyfolge in A konvergiert; folglich ist A vollständig.

Wir nehmen umgekehrt an, A sei nicht abgeschlossen. Dann gibt es eine Folge (a_n), die in X gegen einen Grenzwert $x \in X \setminus A$ konvergiert. Dann ist aber (a_n) eine Cauchyfolge in A, die nicht in A konvergiert; die Existenz einer solchen Folge zeigt, daß A nicht vollständig ist. ∎

Um Satz (83.4) anwenden zu können, müssen wir in der Lage sein festzustellen, wann ein Punkt im Abschluß einer Menge A liegt. Der folgende Satz gibt einige Charakterisierungen, die allesamt (wenn auch auf verschiedene Arten) ausdrücken, daß genau dann $x \in \overline{A}$ gilt, wenn sich x beliebig genau durch Elemente in A approximieren läßt.

(83.5) Satz. *Es sei $A \neq \emptyset$ eine Teilmenge eines metrischen Raums (X,d) und $x \in X$ ein Element in x. Dann sind die folgenden Aussagen äquivalent:*

(1) $x \in \overline{A}$*;*

(2) *für jedes $\varepsilon > 0$ gibt es ein $a \in A$ mit $d(x,a) < \varepsilon$;*

(3) *für jede Umgebung U von x gilt $U \cap A \neq \emptyset$;*

(4) $\operatorname{dist}(x,A) = 0$.

Beweis. (1)⇒(2): Es gibt eine Folge (a_n) in A mit $a_n \to x$. Ist nun $\varepsilon > 0$ beliebig vorgegeben, so gibt es ein $N \in \mathbb{N}$ mit $d(a_n,x) < \varepsilon$ für alle $n \geq N$.

(2)⇒(1): Zu $\varepsilon_n := 1/n$ gibt es ein Element $a_n \in A$ mit $d(x,a_n) < \varepsilon_n = 1/n$; dann ist (a_n) eine Folge in A, die gegen x konvergiert.

Die Äquivalenz (2) ⇔ (3) folgt unmittelbar aus der Definition des Umgebungsbegriffs; die Äquivalenz (2) ⇔ (4) folgt sofort daraus, wie der Begriff des Abstands zweier Mengen definiert wurde. ∎

(83.6) Beispiele. (a) Der Abschluß eines offenen Intervalls (a,b) in $\mathbb{R}$ ist das abgeschlossene Intervall $[a,b]$.

(b) Es gilt $\overline{\mathbb{Q}} = \overline{\mathbb{R} \setminus \mathbb{Q}} = \mathbb{R}$; d. h., sowohl die rationalen als auch die irrationalen Zahlen liegen dicht in $\mathbb{R}$. (Jede reelle Zahl läßt sich sowohl durch rationale als auch durch irrationale Zahlen beliebig gut approximieren.)

(c) In einem normierten Raum X gilt $\overline{B_r(p)} = K_r(p)$, wenn wir mit $B_r(p)$ und $K_r(p)$ die offene bzw. abgeschlossene Kugel mit Radius $r > 0$ um einen Punkt $p \in X$ bezeichnen. Ist (x_n) eine Folge in $B_r(p)$ mit $x_n \to x$, so gilt einerseits $d(x_n,p) < r$ für alle $n \in \mathbb{N}$, andererseits $d(x,p) = \lim_{n\to\infty} d(x_n,p)$ und damit $d(x,p) \leq r$, also $x \in K_r(p)$; damit ist die Inklusion $\overline{B_r(p)} \subseteq K_r(p)$ gezeigt. (Der Beweis zeigt, daß diese Inklusion in einem beliebigen metrischen Raum gilt, nicht nur in einem normierten Raum.) Umgekehrt sei $x \in K_r(p)$, also $d(x,p) \leq r$. Definieren wir

$$x_n \;:=\; p + \frac{n-1}{n}(x-p) \;=\; x + \frac{1}{n}\,(p-x),$$

so gilt einerseits $d(x_n,p) = \|x_n - p\| < \|x-p\| \leq r$ und damit $x_n \in B_r(p)$, andererseits $d(x_n,x) = \|x_n - x\| = (1/n)\|x-p\| \to 0$ und damit $x_n \to x$. Damit ist auch die umgekehrte Inklusion $K_r(p) \subseteq \overline{B_r(p)}$ gezeigt.

(d) Es seien X und Y metrische Räume, $\mathfrak{B}(X,Y)$ der Raum aller beschränkten Funktionen $f : X \to Y$ (vgl. (81.2)(b)) sowie $C(X,Y)$ der Unterraum aller *stetigen* beschränkten Funktionen $f : X \to Y$. (Der Buchstabe C soll an "continuity"="Stetigkeit" erinnern.) Wir behaupten, daß $C(X,Y)$ abgeschlossen in $\mathfrak{B}(X,Y)$ ist. Mit anderen Worten: gilt $f_n \to f$ in $\mathfrak{B}(X,Y)$ und sind alle Funktionen f_n stetig, so ist auch die Grenzfunktion f stetig.

Wir geben uns ein beliebiges Element $x_0 \in X$ und eine Zahl $\varepsilon > 0$ vor. Wegen $D(f_n,f) \to 0$ gibt es einen Index m mit $D(f_n,f) < \varepsilon/3$ für alle $n \geq m$. Da f_m stetig an der Stelle x_0 ist, gibt es eine Zahl $\delta > 0$ mit $d_Y\big(f_m(x), f_m(x_0)\big) < \varepsilon/3$ für alle $x \in X$ mit $d_X(x,x_0) < \delta$. Aus $d_X(x,x_0) < \delta$ folgt dann $d_Y\big(f(x),f(x_0)\big) \leq d_Y\big(f(x),f_m(x)\big) + d_Y\big(f_m(x),f_m(x_0)\big) + d_Y\big(f_m(x_0),f(x_0)\big)$ und damit

$$\begin{aligned} &d_Y\big(f(x),f(x_0)\big) \\ &\leq D(f,f_m) + d_Y\big(f_m(x),f_m(x_0)\big) + D(f_m,f) \\ &\leq (\varepsilon/3) + (\varepsilon/3) + (\varepsilon/3) \;=\; \varepsilon. \end{aligned}$$

Da $\varepsilon > 0$ beliebig vorgegeben war, ist damit die Stetigkeit von f an der Stelle x_0 gezeigt.

(e) Es sei $C[0,1]$ der Vektorraum aller stetigen Funktionen $f : [0,1] \to \mathbb{R}$, versehen mit der Supremumsnorm. Bezeichnen wir mit A die Menge aller Polynomfunktionen $p : [0,1] \to \mathbb{R}$ mit $p(0) = 0$ und mit f die Wurzelfunktion $f(x) = \sqrt{x}$, so gilt $f \in \overline{A}$; etwas expliziter formuliert: es gibt eine Folge von Polynomfunktionen p_n, die gleichmäßig auf $[0,1]$ gegen f konvergiert. (Wie der Beweis zeigt, können diese Polynomfunktionen p_n so gewählt werden, daß $p_n(0) = 0$ für alle n gilt.)

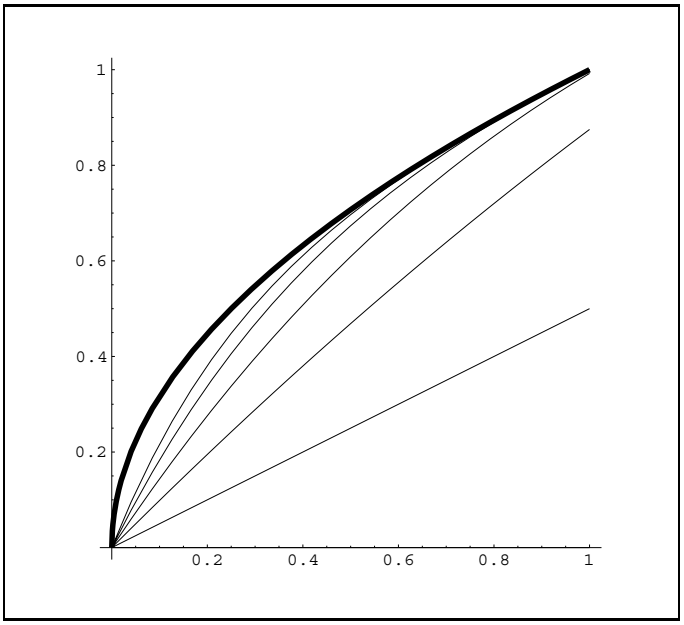

Abb. 83.1: Approximation der Wurzelfunktion durch Polynome.

Der Nachweis wird geführt, indem wir die fraglichen Polynome p_n explizit konstruieren, und zwar durch $p_0(x) := 0$ und die Rekursionsformel

$$p_{n+1}(x) \;:=\; p_n(x) + \frac{x - p_n(x)^2}{2}\,.$$