Wolfgang K. Seiler

Höhere Mathematik I

Vorlesung an der Universität Mannheim
im Sommersemester 2004

Dieses Skriptum entsteht parallel zur Vorlesung und soll mit möglichst geringer Verzögerung erscheinen. Es ist daher in seiner Qualität auf keinen Fall mit einem Lehrbuch zu vergleichen; insbesondere sind Fehler bei dieser Entstehensweise nicht nur möglich, sondern **sicher**. Dabei handelt es sich wohl leider nicht immer nur um harmlose Tippfehler, sondern auch um Fehler bei den mathematischen Aussagen.

Das Skriptum sollte daher mit Sorgfalt und einem gewissen Mißtrauen gegen seinen Inhalt gelesen werden; falls Sie Fehler finden, teilen Sie mir dies bitte persönlich oder per e-mail (seiler@math.uni-mannheim.de) mit, oder informieren Sie Ihren Übungsgruppenleiter. Auch wenn Sie Teile des Skriptums unverständlich finden, bin ich für entsprechende Hinweise dankbar.

Falls genügend viele Hinweise eingehen, werde ich von Zeit zu Zeit Listen mit Berichtigungen und Verbesserungen zusammenstellen.

Einleitung 1

Literaturhinweise 4

KAPITEL I: VEKTORRÄUME UND LINEARE GLEICHUNGSSYSTEME 9

§1: Zahlen und Körper 10

a) Von den natürlichen zu den komplexen Zahlen 10

b) Der Begriff des Körpers 12

c) Mehr über komplexe Zahlen 16

d) Weitere Körper 18

e) Der Körper mit zwei Elementen 19

§2: Vektoren und Vektorräume 20

a) Vektoren in der Ebene und im Raum 21

b) Definition des Vektorraums 25

c) Erste Beispiele 27

d) Lineare Abbildungen 30

e) Untervektorräume 33

f) Lineare Abhängigkeit 36

g) Die Dimension eines Vektorraums 44

h) Basen 45

i) Dimensionen und lineare Abbildungen 54

§3: Vektorräume und endliche Körper 56

a) Bitfolgen als Vektoren 56

b) Körper von Primzahlordnung 61

c) Der Euklidische Algorithmus 63

d) Das RSA-Verfahren 66

1) Verschlüsselung 73

2) Identitätsnachweis 74

3) Elektronische Unterschriften 74

4) Blinde Unterschriften und elektronisches Bargeld 75

5) Größenordnung der Primzahlen 78

e) Das Verfahren von Diffie und Hellman 80

f) Körper von Zweipotenzordnung 84

g) Der Euklidische Algorithmus für Polynome 89

f) Der Körper mit 256 Elementen und CD-Fehlerkorrektur 94

g) Der Körper mit 256 Elementen in der Kryptographie 97

§4: Matrizen und lineare Gleichungssysteme 98

a) Abbildungsmatrizen 98

b) Rechenregeln für Matrizen 101

c) Matrixdarstellung der komplexen Zahlen 111

d) Das Gaußsche Eliminationsverfahren 114

e) Erste Beispiele 116

f) Die Struktur der Lösungsmenge 130
g) Affine Räume 136
h) Ausblick: Numerische Lösung linearer Gleichungssysteme ... 148
i) Matrixgleichungen und die Berechnung der Inversen 155
j) Spezielle Matrizen 159
1) Diagonalmatrizen 159
2) Dreiecksmatrizen 162
3) Matrizen mit nur einem Eintrag 166
4) Permutationen und Permutationsmatrizen 167
k) Die LR-Zerlegung einer Matrix 172
l) Beispiel eines Basiswechsels 177
m) Basiswechsel im allgemeinen Fall 183
§5: EUKLIDische und HERMITEsche Vektorräume 189
a) Längen und Winkel in $\mathbb{R}^2$ und $\mathbb{R}^3$ 189
b) EUKLIDische Vektorräume 194
c) HERMITEsche Vektorräume 199
d) Die CAUCHY-SCHWARZsche Ungleichung 203
e) Orthonormalbasen 204
f) Die QR-Zerlegung einer Matrix 211
g) Orthogonale und unitäre Matrizen 217
h) Orthogonale Projektionen 221
i) Die Methode der kleinsten Quadrate 226
j) EUKLIDische Vektorräume in der Informationssuche 241
§6: Volumina und Determinanten 245
a) Das Vektorprodukt im $\mathbb{R}^3$ 245
b) Das Spatprodukt 249
c) Forderungen an eine allgemeine Determinante 253
d) Gerade und ungerade Permutationen 256
e) Existenz von Determinanten 261
f) Die Determinante einer Matrix 266
g) Der Entwicklungssatz von LAPLACE 274
h) Die CRAMERsche Regel 283
i) Eigenwerte und Eigenvektoren 284
j) Geschichte und Anwendungen von Determinanten 290
KAPITEL II: MEHRDIMENSIONALE ANALYSIS 292
§1: Funktionen und ihre Eigenschaften 292
a) Darstellungsmöglichkeiten für Funktionen 292
b) Normierte Vektorräume 295
c) Die Ableitung einer Funktion 305
d) TAYLOR-Reihen 323
e) Der Satz über implizite Funktionen 326

§2: Vektorfelder 331

a) Der Begriff des Vektorfelds 334

b) Die JACOBI-Matrix 334

c) Die Divergenz eines Vektorfelds 336

d) Die Rotation eines dreidimensionalen Vektorfelds 338

e) Erste Beispiele 343

1) Das elektrische Feld einer Punktladung 343

2) Das Magnetfeld eines stromdurchflossenen Leiters 345

f) Allgemeine Rechenregeln 349

g) Nichtkartesische Koordinatensysteme 352

1) Polarkoordinaten in $\mathbb{R}^2$ 352

2) Zylinderkoordinaten im $\mathbb{R}^3$ 355

3) Kugelkoordinaten 356

§3: Integralrechnung 358

a) Heuristische Vorüberlegungen 358

1) Integration als Umkehrung der Differentiation 359

2) Integration als Flächenbestimmung 365

3) Integration als Durchschnittsbestimmung 366

b) Integration elementarer Funktionen 368

1) Die Funktion $f(x) = x^2$ 367

2) Die Exponentialfunktion 369

3) Die DIRICHLETsche Sprungfunktion 370

c) Definition des Riemann-Integrals 372

1) Warum lohnt sich ein allgemeinerer Ansatz? 372

2) Wo sollte der bisherige Ansatz modifiziert werden? 373

3) Anwendung des Mittelwertsatzes 375

4) Gleichmäßige Stetigkeit 377

5) Definition einer Approximation für das Integral 379

6) Existenz des RIEMANN-Integrals für stetige Funktionen 383

7) Stückweise stetige Funktionen 388

8) Noch einmal die DIRICHLETsche Sprungfunktion 390

9) Ausblick: Das LEBESGUE-Integral 390

10) Anwendung auf Flächeninhalte 391

d) Erste Integrationsregeln 392

1) Monotonieregel 392

2) Linearität und Zusammensetzung 393

3) Der Mittelwertsatz der Integralrechnung 394

e) Der Hauptsatz der Differential- und Integralrechnung 396

f) Trigonometrische Funktionen, Hyperbelfunktionen und ihre Umkehrfunktionen 399

g) Partielle Integration 411

h) Substitutionsregel 412

1) Der Spezialfall logarithmischer Ableitungen 412

2) Substitutionen mit linearen Funktionen 413

3) Substitutionen mit trigonometrischen und Hyperbelfunktionen 415

4) Integrale der Form $\int h(e^{ax})\,dx$ 417
5) Integrale der Form $\int h(\sin x, \cos x)\,dx$ 419
i) Integration rationaler Funktionen 420
j) Symmetrie 425
k) Einige nicht elementar integrierbare Funktionen 426
1) Der Integralsinus 427
2) Die Fehlerfunktion 427
3) Elliptische Integrale 428
4) Algebraische Integrale 429
l) Uneigentliche Integrale 429
§4: Kurvenintegrale im $\mathbb{R}^n$ 440
a) Kurven und Tangentenvektoren 440
b) Die Bogenlänge einer Kurve 443
c) Integration eines Vektorfelds längs einer Kurve 448
d) Zirkulationsfreie und konservative Vektorfelder 453
§4: Mehrdimensionale Integrationstheorie 459
a) Flächeninhalte und Volumina 459
b) Integration über Normalbereiche 467
c) Die Transformationsformel 473
d) Der Satz von GREEN und der ebene Satz von GAUSS 484
e) Oberflächenintegrale 490
f) Die Sätze von STOKES und GAUSS 501

Einleitung

Die Vorlesung *Höhere Mathematik* soll Ihnen helfen, das für Ihren Studiengang (Technische Informatik *bzw.* Software- und Internettechnologie) notwendige mathematische Rüstzeug zu erwerben.

Im Vordergrund steht somit die Mathematik als Werkzeug.

Den Umgang mit einem Werkzeug lernt man nur durch dessen Gebrauch: Niemand wird zum Schlosser, indem er Bilder von Werkzeugen betrachtet und Abhandlungen über das Drehmoment liest, und niemand, der einfach mathematische Sätze und deren Beweise auswendig lernt, wird damit erfolgreich Probleme der Informatik lösen.

Zum einen beginnen Probleme der Informatik nie mit „$f : \mathbb{R}^3 \to \mathbb{R}$ sei eine mindestens zweimal stetig differenzierbare Funktion“ oder etwas ähnlichem; der Anwender muß sich stets zunächst überlegen, ob sich sein Problem überhaupt mit Mathematik lösen läßt, und wenn ja, mit welcher.

Die Antwort darauf wird selten eindeutig sein: Gelegentlich wird sich dasselbe Problem sowohl mit als auch ohne Mathematik modellieren lassen, und auch wenn man sich für eine mathematische Lösung entscheidet, steht selten eindeutig fest, wie es weiter geht:

Realistische Probleme (im Gegensatz zu Übungs- und Klausuraufgaben) sind fast immer zu komplex, als daß man sie vollständig formal beschreiben könnte. Vor ihrer Übersetzung in Mathematik muß man sie daher vereinfachen und dabei versuchen, die für die jeweilige Problemstellung wesentlichen Eigenschaften trotzdem zu erhalten. Beispielsweise wird man oft Größen, die ihrem Wesen nach ganzzahlig sind, mit reelle Zahlen modellieren, da die Mathematik der reellen Zahlen erheblich

einfacher ist als die der ganzen Zahlen und auch deutlich mehr Methoden zur Verfügung stellt. Bei der Informationsübertragung etwa ist die Anzahl übertragener Bits stets ganzzahlig, aber bei den Größenordnungen, die in einem typischen lokalen Netzwerk (oder auch im Internet) auftreten, macht man keinen großen Fehler, wenn man diese Anzahl als kontinuierliche Größe betrachtet.

Die mathematischen Werkzeuge, die man zur Lösung eines Problems benutzt, hängen wesentlich ab von solchen Modellierungsentscheidungen: Bei Modellierung durch kontinuierliche Größen wird man typischerweise mit analytischen Methoden arbeiten, oft mit Differentialgleichungen; bei Modellierung mit ganzen Zahlen ist man in der diskreten Mathematik, wo eher algebraische und zahlentheoretische Methoden im Vordergrund stehen – von denen zumindest ein Teil übrigens auch analytisch ist.

Selbst wenn die Übersetzung eines Problem in Mathematik (oder besser seine Annäherung an Mathematik) feststeht, gibt es im Allgemeinen verschiedene mathematische Ansätze, die theoretisch allesamt zur korrekten Lösung führen; in der Praxis kann sich aber der Rechenaufwand zweier Verfahren so stark unterscheiden, daß nur eines der beiden wirklich durchführbar ist, oder aber so, daß zwar beide durchführbar sind, das eine aber erheblich mehr kostet als das andere.

In der Praxis eines Informatikers geht es somit nicht in erster Linie darum, ob die Voraussetzungen für einen bestimmten mathematischen Satz erfüllt sind: Die Frage stellt sich erst viel später. Als erstes muß geklärt werden, welche mathematischen Modelle in Frage kommen und welche davon zu Lösungen mit realistischem Aufwand führen.

Wer hier erfolgreich sein will, muß daher vor allem ein Gefühl für die Tragweite und den Aufwand mathematischer Methoden entwickeln; nur so kann er eine geeignete auswählen.

Dieses Gefühl kann man – wie auch die richtige Technik für den Umgang mit einem Hammer oder einer Feile – nur durch praktische Erfahrung erwerben. Ein wesentlicher Bestandteil dieser Vorlesung sind daher die Übungen, die – im Rahmen dessen, was mit Informatikkenntnissen des

zweiten Semesters möglich ist – die notwendige Erfahrung dazu vermitteln sollen. Hauptziel der Vorlesung ist also, daß Sie praktische Fähigkeiten entwickeln, und das ist nur möglich, indem Sie selbst praktische Erfahrung sammeln.

Die Übungen sind somit mindestens genauso wichtig wie die Vorlesung selbst. Es wird jede Woche zwei Übungsblätter für zwei Arten von Übungen geben:

Die wöchentlichen Übungen in kleinen Gruppen sollen Ihnen in erster Linie Gelegenheit geben, Fragen zur Vorlesung und zur konkreten Umsetzung des Vorlesungsstoffs zu stellen. Leider zeigt die Erfahrung, daß sich leider wahrscheinlich nur eine sehr kleine Minderheit von Ihnen getrauen wird, Fragen in der Vorlesung zu stellen oder Themenvorschläge für die Übungen zu machen. Deshalb gibt es jede Woche ein Blatt *Themenvorschläge für die kleinen Übungen*, das Ihnen mögliche Aufgaben vorschlägt. Diese Themenvorschläge sind selbstverständlich nicht verbindlich; in der idealen Übung spielt keiner von ihnen auch nur die geringste Rolle. Wieweit sich eine Übung diesem Idealbild nähert, hängt von Ihnen ab.

In den subidealen realen Übungen werden zunächst Ihre Fragen beantwortet und von Ihnen vorgeschlagene Probleme gelöst; in der (leider meist viel zu langen) Zeit, die noch übrig bleibt, sollten Sie die den Inhalt der Übungen wenigstens dadurch beeinflussen, daß Sie eine Auswahl unter den Themenvorschlägen treffen. Die Themenvorschläge sind im allgemeinen so umfangreich, daß unmöglich alle in einer kleinen Übung behandelt werden können; es liegt an Ihnen, diejenigen auswählen, die Ihnen am ehesten helfen, den Umgang mit den noch nicht so gut verstandenen Konzepten der Vorlesung zu üben.

Als *Hausarbeit* erhalten Sie jede Woche das eigentliche Übungsblatt, das abgegeben und von Ihrem Tutor korrigiert wird. Eine Lösung werde ich jede Woche in den großen Übungen vorrechnen – wie bereits oben ausgeführt, wird dies nur selten die einzig mögliche richtige Lösung sein. Sowohl der Vergleich meiner Lösung mit der Ihrigen als auch vor allem die Rückmeldung durch die Korrektur sollen Ihnen noch vorhandene Schwächen zeigen und zu Fragen in den kleinen Übungen anregen.

Literaturhinweise

Das Angebot von Lehrbüchern zum Thema „Höhere Mathematik“ ist riesig, und fast jedes dieser Bücher kann zumindest für Teile dieser Vorlesung nützlich sein. Ich habe hier vor allem Bücher aufgeführt, die ich selbst schon irgendwann einmal benutzt habe und daher einigermaßen kenne. Die meisten davon sind in der Mathematischen Bibliothek zu finden, teils im allgemeinen Bestand, teils auch in der Lehrbuchsammlung. Die angegebenen Bücher sind in ihrer Art sehr verschieden; bevor Sie sich eines davon kaufen, sollten Sie es sich unbedingt vorher ein Bibliotheksexemplar genau anschauen, ob es Ihnen wirklich zusagt. Wenn Sie bei einen Versender ältere Auflagen zu reduzierten Preisen bekommen können (oft nur die Hälfte), können Sie diese unbesorgt kaufen; die Änderungen zwischen verschiedenen Auflagen sind im allgemeinen so gering, daß sie praktisch kaum ins Gewicht fallen. Die angegebenen Auflagen sind die letzten, die ich gesehen habe; im Buchhandel sind möglicherweise bereits neuere zu finden.

Die beiden erstgenannten Bücher **[MV]** und **[D]** verfolgen wohl am ehesten den gleichen Zweck wie diese Vorlesung:

[MV1] K. MEYBERG, P. VACHENAUER: *Höhere Mathematik I, Differential- und Integralrechnung, Vektor- und Matrizenrechnung,* Springer, ^{6}2001

[MV2] K. MEYBERG, P. VACHENAUER: *Höhere Mathematik II, Differentialgleichungen, Funktionentheorie, Fourier-Analyse, Variationsrechnung,* Springer, ^{4}2001

Dieses zweibändige Werk enthält den gesamten Stoff der *Höheren Mathematik* sowie die dafür relevanten Teile der *Analysis I.* Die Darstellung ist recht kompakt

mit fast vollständigen Beweisen; zu einigen Grundalgorithmen sind Programme angegeben. Für die *Höhere Mathematik I* reicht der erste Band.

[D] H.J. DIRSCHMID: *Mathematische Grundlagen der Elektrotechnik,* Vieweg, 1992

Etwa fünf Pfund Mathematik, die nicht nur diese Vorlesung mehr als abdecken, sondern für viele Studenten für deren gesamtes Berufsleben ausreichen dürften. Der Schwerpunkt liegt eindeutig auf dem Gebiet der Analysis; numerische Mathematik und Statistik sind (wie auch in der Vorlesung) so gut wie nicht vorhanden. Interessant vor allem für Technische Informatiker; für Software- und Internettechnologen, die keine *Höhere Mathematik II* hören, etwas zu viel des Guten. Leider auch sehr teuer, falls überhaupt noch erhältlich.

[P] L. PAPULA: *Mathematik für Ingenieure und Naturwissenschaftler,* Vieweg,
Band 1 [10]2001, Band 2 [10]2001, Band 3 [4]2001, Anwendungsorientierte Übungsaufgaben aus Naturwissenschaft und Technik [4]2000, Mathematische Formelsammlung für Ingenieure und Naturwissenschaftler, [7]2001

Sehr beliebtes und erfolgreiches Lehrbuch mit ausführlicher Darstellung einer beschränkten, aber im wesentlichen ausreichenden Stoffauswahl. Gelegentlich wird mehr Wert auf Anschaulichkeit als auf mathematische Exaktheit gelegt. Der erste Band behandelt vor allem Stoff der Analysis I, lediglich bei der elementaren Vektorrechnung gibt es eine kleine Überschneidung mit dieser Vorlesung. Der größte Teil des HM I-Stoffs ist in Band 2 zu finden, lediglich für die Vektoranalysis braucht man auch Band 3. Auch für die HM II reicht größtenteils Band 2; die in Band 3 sehr ausführlich behandelte Wahrscheinlichkeitstheorie und Statistik wird in der HM II nur am Ende und sehr viel kürzer behandelt. Die „Übungsaufgaben" setzen einiges an Kenntnissen aus Physik und Technik voraus; ein Anhang stellt das notwendige Grundwissen kurz zusammen.

[BHW] BURG/HAF/WILLE: *Höhere Mathematik für Ingenieure,* Teubner 1. Analysis, [6]2003, 2. Lineare Algebra, [4]2002, 3. Gewöhnliche Differentialgleichungen, Distributionen, Integraltransformationen, [4]2002, 4. Vektoranalysis und Funktionentheorie, [2]1994, 5. Funktionalanalysis und Partielle Differentialgleichungen, [2]1993

Enthält deutlich mehr Stoff als **[P]** und geht mathematisch deutlich tiefer; leider ist die Verteilung des Stoffs auf die einzelnen Bände sehr verschieden vom Aufbau dieser Vorlesung: Schon die HM I behandelt Stoff aus den Bänden 1, 2 und 4; in der HM II kommt noch Stoff aus Band 3 dazu.

[BDH] BRAUCH/DREYER/HAACKE: *Mathematik für Ingenieure,* Teubner [9]1995

Dieses Buch richtet sich an Studenten von Fachhochschulen und ist somit nach Ansicht mancher Kollegen nicht für eine Vorlesung an der Universität geeignet. Wer sich allerdings eher für höhere Mathematik als für höheren Dünkel interessiert, findet hier ziemlich kompakt und relativ preisgünstig fast den gesamten Stoff zumindest der HM I; lediglich die Darstellung der Vektoranalysis ist etwas zu knapp.

[W1] T. WESTERMANN: *Mathematik für Ingenieure mit MAPLE I,* Springer [2]2000

[W2] T. WESTERMANN: *Mathematik für Ingenieure mit MAPLE II,* Springer [2]2001

Auch diese beiden Bände wenden sich eher an Studenten von Fachhochschulen; sie passen vom Aufbau her nicht sonderlich gut zur Vorlesung, haben aber den Vorteil, daß sie parallel zum Stoff auch das Computeralgebrasystem MAPLE behandeln und anwenden. Eine CD mit entsprechenden *worksheets* sowie einer eingeschränkten Version von MAPLE V.0 liegt jedem der beiden Bände bei. Interessant vor allem für Studenten, die parallel zur Vorlesung auch den Umgang mit einem Computeralgebrasystem üben wollen und noch keinerlei entsprechende Erfahrung haben. Die *Höhere Mathematik I* behandelt Stoff aus beiden Bänden.

[F] P. FURLAN: *Das gelbe Rechenbuch 1–3,* Verlag Martina Furlan, Dortmund, *o.J.*

Wer vor allem auf Drill und durchgerechnete Beispiele Wert legt, findet hier laut Untertitel „Rechenverfahren der Höheren Mathematik in Einzelschritten erklärt. Mit vielen ausführlich gerechneten Beispielen". Mit den dort vorexerzierten Kochrezepten lassen sich die gängigen Typen von Standardaufgaben lösen; wenn auch nicht immer optimal: Wie immer beim sturen Nachexerzieren von Kochrezepten läuft man Gefahr, sich oftmals zuviel Arbeit zu machen, da sich

konkrete Probleme oft mit etwas Theorie beträchtlich vereinfachen lassen (und, bei realen Problemen, teilweise auch erst dadurch mit vertretbarem Aufwand lösbar werden).

Als Ergänzung zur *Höheren Mathematik I* können, mit diesen Einschränkungen, die ersten beiden Bände nützlich sein – insbesondere auch in der Endphase der Klausurvorbereitung.

Zumindest für *Technische Informatiker,* die im weiteren Verlauf ihres Studiums (und Berufslebens) immer wieder mit mathematischen Problemen konfrontiert werden, empfiehlt sich über kurz oder lang die Anschaffung einer Formelsammlung. Zu einigen der bereits zitierten Werke gibt es einen entsprechenden Band, der in seinen Bezeichnungen und der Stoffauswahl auf das Gesamtwerk abgestimmt ist; ansonsten ist vor allem ein Klassiker zu nennen, der seit Jahrzehnten in immer neuen Auflagen erscheint und mit dem schon Generationen von Naturwissenschaftlern und Ingenieuren gearbeitet haben: „Der BRONSTEIN“, der seit einigen Jahren in zwei konkurrierenden Neubearbeitungen angeboten wird, wobei die erste wahlweise mit oder ohne CD-ROM erhältlich ist.

[BSM] I.N. BRONSTEIN, K.A. SEMENDJAJEW, G. MUSIOL: *Taschenbuch der Mathematik,* Verlag Harri Deutsch, 2000

[BGZ] I.N. BRONSTEIN, G. GROSCHE, E. ZEIDLER: *Teubner-Taschenbuch der Mathematik,* Teubner, 1996

Deutlich weniger ambitioniert und auch billiger ist

[MMWW] G. MERZINGER, G. MÜHLBACH, D. WILLE, T. WIRTH: *Formeln + Hilfen zur Höheren Mathematik,* Binomi ^{4}2001,

eine Formelsammlung die zur HM I und – mit Ausnahme der nicht behandelten FOURIER-Transformation – auch zur HM II ausreicht, vielleicht aber nicht für spätere Anwendungen.

Auch bei den Lehrbüchern seien noch einige „Klassiker“ genannt, die seit Jahrzehnten in immer neuen Auflagen erscheinen und auch heute noch interessant sind. Es handelt sich um Werke aus meist recht vielen Bänden, wobei selbst der Stoff der Vorlesung *Höhere Mathematik I* je

nach Organisation des Gesamtwerks auf bis zu drei Bände verteilt sein kann. Wegen der Vielzahl von Auflagen und Bänden verzichte ich auf die Angabe von Erscheinungsjahren.

[S] W.I. SMIRNOW: *Lehrbuch der Höheren Mathematik,* VEB Deutscher Verlag der Wissenschaften

Ein Klassiker, nach dem Generationen von russischen und (nicht nur ost-)deutschen Naturwissenschaftlern und Ingenieuren ausgebildet wurden; enthält in vier Bänden (von denen die letzten beiden noch in Halbbände unterteilt sind) den gesamten klassischen Stoff der Mathematik für Naturwissenschaftler und Ingenieure (also erheblich mehr, als im zweisemestrigen Kurs *Höhere Mathematik* behandelt werden kann) und ist auch heute noch sehr gut zu lesen. Die HM I behandelt Stoff aus den Bänden I, II und III/1.

[R] R. ROTHE: *Höhere Mathematik,* Teubner

Sieben (dünne) Bände, zu denen allerdings auch Aufgaben- und Formelsammlung gehören, so daß die Darstellung insgesamt eher knapp ist. Für diese Vorlesung relevant sind die Bände II und III.

[Du] A. DUSCHEK: *Höhere Mathematik,* Springer Wien

Die österreichische Variante, vier recht ausführliche Bände, von denen hier vor allem die ersten beiden von Interesse sind.

[A] G. AUMANN: *Höhere Mathematik,* Bibliographisches Institut Mannheim

Drei relativ dünne Taschenbücher, deren erste beide trotzdem fast alles enthalten, was wir in dieser Vorlesung brauchen. Die Darstellung ist natürlich weniger ausführlich als in den dickleibigen Werken, aber das wird nicht jeder Student als Nachteil empfinden.

[C] R. COURANT: *Vorlesungen über Differential- und Integralrechnung I+II,* Springer

Ein Klassiker eines berühmten Mathematikers, auch heute noch sehr lesenswert. Trotz des Titels beschränkt sich das Buch nicht auf Analysis, sondern behandelt auch beispielsweise die Lineare Algebra in einem Umfang, der für diese Vorlesung völlig ausreicht.

Kapitel 1
Vektorräume und lineare Gleichungssysteme

Lineare Strukturen sind sowohl in der Mathematik als auch in ihren Anwendungen allgegenwärtig: Zwar sind die meisten Funktionen nichtlinear, aber fast alles, was man damit anstellt – Differenzieren, Integrieren, FOURIER- oder LAPLACE-transformieren *usw.* – wird sich als lineare Operation herausstellen. Vektorräume bieten einen gemeinsamen Rahmen für alle diese Operationen und sind daher wichtige Hilfsmittel nicht nur innerhalb der Mathematik, wie etwa für Analysis, Geometrie, Differentialgleichungen und Integraltransformationen, sondern auch beispielsweise in der Optimierung, der Signalverarbeitung (z.B. Kodierungstheorie, Kryptographie und Bildverarbeitung), der Optik, Elektrodynamik und Quantenphysik. Die anschaulichsten Vektorräume sind die, deren Elemente anschauliche Vektoren sind; der große Vorteil einer einheitlichen Behandlung dieser Vektoren und zahlreicher anderer Objekte unter dem gemeinsamen Dach des Begriffs „Vektorraum“ besteht darin, daß man viele für Vektoren anschaulich klare Aussagen mit geringem Aufwand so umformulieren kann, daß sie auch in den erheblich schwerer vorstellbaren Vektorräumen gelten, mit denen man es in schwierigeren Anwendungen zu tun hat. Dies ist Teil einer sehr viel allgemeineren Strategie der Mathematik: Abstrakte Mathematik lebt davon, daß anschauliche Phänomene formalisiert werden, um die so entstehende formale Struktur auf andere, weniger anschauliche Phänomene anzuwenden.

Gerade für Anwendungen in der Informationstechnik brauchen wir allerdings nicht nur einen allgemeineren Rahmen für Vektoren:

Physikalische und auch sonstige Größen lassen sich bekanntlich (mit wenigen Ausnahmen) bezüglich ihrer mathematischen Behandlung in

zwei Klassen einteilen: Temperaturen, Stromstärken, Bevölkerungszahlen, Geldmengen *usw.* werden (bezüglich einer festzulegenden Einheit) durch Zahlen beschrieben; Geschwindigkeiten, Kräfte, Bevölkerungswanderungen *usw.* durch Zahlen zusammen mit einer Richtung. Im ersten Fall reden wir von *Skalaren,* im zweiten von *Vektoren.*

Bei den gerade aufgeführten Beispiele handelt es sich bei den Zahlen jeweils um reelle Zahlen oder Teilmengen davon. Wenn wir praktisch rechnen, egal ob mit oder ohne Computer, müssen wir uns allerdings immer auf Teilmengen der reellen Zahlen beschränken, schon weil kein Computer sämtliche Elemente einer überabzählbare Menge darstellen kann.

Es gibt allerdings auch eine Reihe von Beispielen, bei denen wir es mit „Zahlen“ zu tun haben, die nichts mit reellen Zahlen zu tun haben: Für Bits und Bytes lassen sich Addition und Multiplikation so definieren, daß dafür dieselben Rechenregeln gelten wie für Addition und Multiplikation reeller Zahlen, und das nutzt die Informationstechnik aus, um beispielsweise Informationen sicher zu übertragen. Dies betrifft sowohl die Fehlerkorrektur auf einer CD (fehlerkorrigierende Codes) als auch die Sicherung von Information gegen unbefugtes Abhören (Kryptographie). Daher müssen wir nicht nur für Vektoren, sondern auch für Skalare eine gemeinsame Struktur finden, daß möglichst viele Anwendungen unter ihrem Dach vereinigt.

Im Falle der Skalare ist das der aus der Analysis I bekannte Begriff des Körpers; der Begriff des Vektorraums formalisiert das Zusammenspiel zwischen Vektoren und Skalaren. Bevor wir ihn einführen, wollen wir uns zunächst Vektoren und Skalare etwas genauer ansehen.

§1: Zahlen und Körper

a) Von den natürlichen zu den komplexen Zahlen

Im Anfang waren die natürlichen Zahlen $1, 2, 3, \ldots$. Man kann sie addieren und multiplizieren, aber man kann dort weder die Gleichung $5 + x = 2$ noch die Gleichung $5 \cdot x = 2$ lösen. Ersteres Problem führt

auf die Erweiterung der Menge $\mathbb{N}$ der natürlichen Zahlen zur Menge $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ der ganzen Zahlen, in der auch die Subtraktion uneingeschränkt ausführbar ist, nicht aber die Division. Letzteres führte – für positive Zahlen schon rund zwei Tausend Jahre vor der Einführung negativer Zahlen – zur Einführung von Brüchen und letztlich zur Menge $\mathbb{Q}$ der rationalen Zahlen, in dem *alle* Grundrechenarten uneingeschränkt ausführbar sind – mit Ausnahme natürlich der Division durch Null.

Um diese „rationalen Zahlen“ herum bauten PYTHAGORAS (geboren je nach Quelle zwischen *ca.* 600* *v. Chr.* und *ca.* 570* *v. Chr.*, gestorben zwischen *ca.* 509 *v. Chr.* und *ca* 475* *v. Chr.*) und seine Schüler in Süditalien eine ganze Weltanschauung; umso größer war ihr Schock, als sie entdeckten, daß ausgerechnet eine geometrisch so perfekte Größe wie die Diagonale eines Quadrats in keinem rationalen Verhältnis zur Seitenlänge steht, mit anderen Worten, daß die Quadratwurzel aus zwei keine rationale Zahl ist. Der Legende nach entdeckte dies HIPPASOS (*ca.* 520* *v. Chr.* – *a.* 480 *v. Chr.*), der dafür je nach Überlieferung entweder von den anderen Pythagoräern ertränkt wurde oder aber als Strafe der Götter auf einer Schiffsreise ertrank.

Trotzdem ließ sich seine Erkenntnis nicht unbeschränkt Geheimhalten; kombiniert mit dem Interesse an Grenzwerten bei EUDOXOS (408 *v. Chr.* – 355 *v. Chr.*), ARCHIMEDES (287 *v. Chr.* – 212 *v. Chr.*), und vor allem rund 2 000 Jahre später bei NEWTON (1643–1727), LEIBNIZ (1664–1716), EULER (1707–1783) und vielen anderen, führte sie schließlich zu den reellen Zahlen.

Auch hier ist noch nicht alles möglich; beispielsweise hat zwar die Gleichung $x^2 = 3$ eine reelle Lösung (genauer gesagt zwei), nicht aber die Gleichung $x^2 = -3$. Diese könnte gelöst werden, wenn man die Quadratwurzel aus -1 ziehen könnte, denn dann wäre $\sqrt{-3} = \sqrt{-1} \cdot \sqrt{3}$. Genauso überlegt man sich leicht, daß dann jede quadratische Gleichung eine Lösung hätte.

Es liegt also nahe, auch die reellen Zahlen zu erweitern, indem man ein neues Element hinzufügt, dessen Quadrat -1 ist. Dieses Element wird in der Mathematik und Physik traditionellerweise mit i wie *imaginär*

bezeichnet, da es zur Zeit seiner Einführung als imaginäre, d.h. nur in der Vorstellung vorhandene, Zahl angesehen wurde. In der Elektrotechnik, wo I und i eine feste andere Bedeutung haben, verwendet man stattdessen den Buchstaben j.

Natürlich genügt es nicht, nur die Menge $\mathbb{R} \cup \{i\}$ zu betrachten, denn wir möchten mit den neuen Zahlen auch rechnen und dabei möglichst wenig von den bewährten Rechenregeln für reelle Zahlen aufgeben.

Insbesondere sollten also alle Zahlen der Form $x+iy$ mit $x, y \in \mathbb{R}$ in der neuen Menge liegen, und es sollte möglich sein, mit ihnen wie gewohnt zu rechnen. Für zwei Zahlen $x + iy$ und $u + iv$ müßte also gelten

$$(x + iy) + (u + iv) = (x + u) + (iy + iv) = (x + u) + i(y + v)$$

und

$$(x + iy)(u + iv) = xu + iyu + xiv + (iy)(iv) = xu + iyu + ixv + i^2yv$$
$$= (xu - yv) + (yu + xv)i \,.$$

Damit lassen sich Zahlen dieser Form insbesondere addieren und multiplizieren, ohne daß neue Zahlen entstehen; wir können hoffen, daß sie vielleicht sogar schon für die geplante Erweiterung ausreichen.

Um dies zu überprüfen, nehmen wir die oben heuristisch abgeleiteten Regeln als *Definitionen* von Addition und Multiplikation und untersuchen die entstehende Struktur:

Definition: *a)* Die Menge $\mathbb{C}$ der komplexen Zahlen ist die Menge aller formaler Ausdrücke der Form $x + iy$ mit $x, y \in \mathbb{R}$.
b) Auf $\mathbb{C}$ wird eine Verknüpfung „+“ definiert durch die Vorschrift

$$(x + iy) + (u + iv) = (x + u) + i(y + v) \,.$$

c) Dazu kommt eine Verknüpfung „$\cdot$“, definiert durch

$$(x + iy) \cdot (u + iv) = (xu - yv) + i(xv + yu) \,.$$

d) Für $z = x + iy \in \mathbb{C}$ nennen wir x den *Realteil* und y den *Imaginärteil* von z; in Zeichen

$$x = \mathfrak{Re}\, z \quad \text{und} \quad y = \mathfrak{Im}\, z \,.$$

(Eine komplexe Zahl heißt *komplex,* weil sie aus einem Realteil und einem Imaginärteil *zusammengesetzt* ist.)

b) Der Begriff des Körpers

Wir wollen sehen, daß die komplexen Zahlen mit diesen Verknüpfungen den „üblichen“ Rechenregeln genügen. Das soll heißen, daß wir Addition, Subtraktion, Multiplikation und Division (außer durch Null) uneingeschränkt durchführen können und daß wir auch mit Klammern „wie gewohnt“ umgehen können. Diese vage Beschreibung formalisierte ERNST STEINITZ 1910 durch den Begriff des *Körpers:*

Definition: Ein Körper k ist eine Menge zusammen mit zwei Abbildungen

$$+: k \times k \to k \quad \text{und} \quad \cdot: k \times k \to k\,,$$

genannt *Addition* und *Multiplikation,* für die gilt:

I.1) Das Assoziativgesetz der Addition

$$(a+b)+c = a+(b+c) \qquad \text{für alle } a,b,c \in k$$

I.2) Es gibt ein Element $0 \in k$, so daß gilt

$$a+0 = 0+a = a \qquad \text{für alle } a \in k$$

I.3) Zu jedem Element $a \in k$ gibt es ein Element $a' \in k$, so daß gilt

$$a+a' = a'+a = 0\,.$$

I.4) Das Kommutativgesetz der Addition

$$a+b = b+a \qquad \text{für alle } a,b \in k$$

II.1) Das Assoziativgesetz der Multiplikation

$$(a\cdot b)\cdot c = a\cdot(b\cdot c) \qquad \text{für alle } a,b,c \in k$$

II.2) Es gibt ein von 0 *verschiedenes* Element $1 \in k$, so daß gilt

$$a\cdot 1 = 1\cdot a = a \qquad \text{für alle } a \in k$$

II.3) Zu jedem von 0 verschiedenen Element $a \in k$ gibt es ein Element $a'' \in k$, so daß gilt

$$a\cdot a'' = a''\cdot a = 1\,.$$

II.4) Das Kommutativgesetz der Multiplikation

$$a\cdot b = b\cdot a \qquad \text{für alle } a,b \in k$$

III.) Das Distributivgesetz

$$a\cdot(b+c) = a\cdot b + a\cdot c \qquad \text{für alle } a,b,c \in k$$

Das Element a' aus I.3.) wird üblicherweise als $-a$ bezeichnet und a'' aus II.3) als a^{-1}. Statt $a+(-b)$ schreibt man kurz $a-b$, statt $a\cdot b^{-1}$ entsprechend a/b.

ERNST STEINITZ (1871–1928) wurde in Schlesien geboren und studierte ab 1890 an den Universitäten Breslau und Berlin. 1894 promovierte er in Breslau, ein Jahr später wurde er Privatdozent an der Technischen Hochschule Berlin-Charlottenburg. 1910 wurde er Professor in Breslau, 1920 in der Universität Kiel. In seinem Buch *Algebraische Theorie der Körper* gab er 1910 die erste Definition eines Körpers und bewies viele Sätze, die noch heute zum Standardstoff jeder Algebra-Vorlesung gehören. Auch die Konstruktion der rationalen Zahlen als Äquivalenzklassen von Paaren ganzer Zahlen geht auf ihn zurück.

Demnach bilden also die natürlichen Zahlen keinen Körper, weil dort weder die Addition noch die Multiplikation invertierbar ist, weil also mit anderen Worten weder die Subtraktion noch die Division (durch Zahlen ungleich Null) unbeschränkt möglich ist.

Genauso bilden auch die ganzen Zahlen keinen Körper, denn hier kann man zwar uneingeschränkt subtrahieren, aber außer ± 1 hat keine ganze Zahl ein multiplikatives Inverses.

Die beiden aus der Schule bekannten Standardbeispiele von Körpern sind die rationalen Zahlen $\mathbb{Q}$, d.h. also die Menge aller Brüche mit einem ganzzahligen Zähler und einer natürlichen Zahl als Nenner, und die Menge $\mathbb{R}$ der reellen Zahlen.

Nach dieser Präzisierung ist klar, was wir von den komplexen Zahlen erwarten, und der nächste Satz zeigt, daß unsere Erwartungen auch erfüllt werden:

Satz: Die Menge $\mathbb{C}$ mit den oben definierten Verknüpfungen ist ein Körper.

Beweis: Eigentlich müßten wir alle Körperaxiome einzeln überprüfen; um aber nicht garzu viel Papier zu produzieren, möchte ich mich hier im Sinne des Umweltschutzes auf die interessantesten beschränken.

Völlig uninteressant sind die Axiome, die sich mit der Addition in $\mathbb{C}$ befassen: Da die Addition komponentenweise für Realteil und Imaginärteil definiert ist, folgen alle Axiome sofort aus den entsprechenden Axiomen für $\mathbb{R}$. Das Neutralelement bezüglich der Addition ist natürlich $0 + i0$, und $-(x + iy) = (-x) + i(-y)$.

Die Forderungen an die Multiplikation sind weniger offensichtlich. Unmittelbar einsichtig ist das Kommutativgesetz; das Assoziativgesetz dagegen ist eine eher unangenehme sture Nachrechnerei, die jeder einmal, aber nur einmal in seinem Leben wirklich ausführen sollte. (Ich habe sie glücklicherweise schon hinter mir.) Wir werden im übrigen in Kürze, sobald wir mit Abbildungsmatrizen umgehen können, einen alternativen Beweis finden, der ganz ohne Rechnung auskommt.

Neutralelement bezüglich der Multiplikation kann, wenn alles Sinn haben soll, nur $1 + i0$ sein, und in der Tat sieht man sofort, daß jedes Element $x + iy$ sowohl bei Links- wie auch bei Rechtsmultiplikation hiermit sich selbst liefert.

Bleibt die Existenz eines multiplikativen Inversen, und hier hilft nur ein Trick: Für $x + iy \neq 0 + i0$ ist

$$(x + iy) \cdot (x - iy) = x^2 + y^2 \in \mathbb{R}_{>0}$$

eine positive reelle Zahl; falls also ein Inverses existiert und die üblichen Regeln der Bruchrechnung gelten, ist

$$\frac{1}{x+iy} = \frac{x-iy}{(x+iy)(x-iy)} = \frac{x-iy}{x^2+y^2} = \frac{x}{x^2+y^2} - i \cdot \frac{y}{x^2+y^2} \,.$$

Eine leichte Rechnung zeigt, daß dieses ganz rechts stehende Element von $\mathbb{C}$ in der Tat sowohl bei Links- als auch bei Rechtsmultiplikation mit $x + iy$ das Neutralelement $1 + i0$ liefert.

Die noch verbleibenden Distributivgesetze sind wieder Rechnerei zum Abhaken, die man genau einmal in seinem Leben durchführen sollte,

und auch sie werden zur trivialen Selbstverständlichkeit, wenn wir den komplexen Zahlen später Abbildungsmatrizen zuordnen. ∎

Es ist nun klar, daß die Abbildung

$$\begin{cases} \mathbb{R} & \hookrightarrow \mathbb{C} \\ x & \mapsto x + i0 \end{cases}$$

eine Einbettung des Körpers der reellen in den der komplexen Zahlen definiert; da wir ersteren erweitern wollen, betrachten wir diese Einbettung als Identifikation, d.h. wir identifizieren den „formalen Ausdruck" $x+i0$ mit der reellen Zahl x. Insbesondere sind jetzt also 0 und 1 das additive und das multiplikative Neutralelement. Außerdem schreiben wir kurz i anstelle von $0 + i1$; nach den Rechenregeln, die wir inzwischen kennen, ist der „formale Ausdruck $x + iy$" dann nichts anderes als die mit den Rechenoperationen von $\mathbb{C}$ berechnete komplexe Zahl $x + i \cdot y$.

Damit sind also alle von der reellen Zahlen gewohnte Rechenregeln für die Grundrechenarten auch für komplexe Zahlen gültig. Nicht zu retten sind allerdings die Regeln über die *Ordnungsbeziehung:* Falls es in $\mathbb{C}$ eine mit der algebraischen Struktur kompatible Ordnungsrelation gäbe, müßte $i^2 \geq 0$ sein, was nicht im Sinne des Erfinders ist. Nicht zuletzt aus diesem Grund machen die Körperaxiome keinerlei Aussage über Größer- und Kleinerbeziehungen.

c) Mehr über komplexe Zahlen

Der Erfolg, den wir bei der Herleitung des multiplikativen Inversen durch Erweiterung mit $x - iy$ hatten, verdient genauer untersucht zu werden:

Definition: Für $z = x + iy \in \mathbb{C}$ heißt $\overline{z} = x - iy$ die zu z konjugiert komplexe Zahl.

(Gelegentlich wird $x - iy$ auch als z^* bezeichnet.)

Offensichtlich ist $\overline{z + w} = \overline{z} + \overline{w}$, und auch das entsprechende Resultat für die Multiplikation $\overline{zw} = \overline{z} \cdot \overline{w}$ läßt sich leicht nachrechnen.

Für die Herleitung des Inversen wesentlich war die Tatsache, daß

$$(x+iy)(x-iy) = x^2 + y^2$$

eine reelle Zahl ist, die genau dann verschwindet, wenn sowohl x als auch y und damit auch $x+iy$ verschwinden; wir bezeichnen die Quadratwurzel aus dieser nichtnegativen reellen Zahl als *Betrag* der komplexen Zahl:

Definition: Der Betrag einer komplexen Zahl $z = x + iy$ ist

$$|z| = \sqrt{z \cdot \overline{z}} = \sqrt{x^2 + y^2}\,.$$

Für reelles z stimmt das natürlich genau mit dem gewohnten Betrag einer reellen Zahl überein.

Offensichtlich ist $|z| = 0 \Leftrightarrow z = 0$ und

$$\frac{1}{z} = \frac{\overline{z}}{z \cdot \overline{z}} = \frac{\overline{z}}{|z|^2}\,.$$

Der Betrag hat noch eine weitere nützliche Interpretation: Für zwei komplexe Zahlen $z = x + iy$ und $w = u + iv$ ist

$$|z - w| = \sqrt{(x-u)^2 + (y-v)^2}$$

gerade der EUKLIDische Abstand zwischen den Punkten (x, y) und (u, v) der EUKLIDischen Ebenen $\mathbb{R}^2$. Da die komplexen Zahlen natürlich über die Abbildung

$$\begin{cases} \mathbb{C} & \to \mathbb{R}^2 \\ x+iy & \mapsto (x,y) \end{cases}$$

in Bijektion mit den Punkten der EUKLIDischen Ebenen stehen, können wir die komplexen Zahlen also auch identifizieren mit den Punkten der EUKLIDischen Ebenen, wobei der Betrag der Differenz zwischen zwei Zahlen gerade dem Abstand entspricht. Man bezeichnet den Körper der komplexen Zahlen in diesem Zusammenhang auch als GAUSS*sche Zahlenebene.* Sie war zwar nicht zusammen mit GAUSS auf dem Zehnmarkschein abgebildet, war aber 1977 das Thema der Sondermarke zu seinem zweihundertsten Geburtstag.

CARL FRIEDRICH GAUSS (1777–1855) leistete wesentliche Beiträge zur Zahlentheorie, zur nichteuklidischen Geometrie, zur Differentialgeometrie und Kartographie, zur Fehlerrechnung und Statistik, zur Astronomie und Geophysik *usw.* Als Direktor der Göttinger Sternwarte baute er zusammen mit dem Physiker Weber den ersten Telegraphen. Er leitete die erste Vermessung und Kartierung des Königreichs Hannover und zeitweise auch den Witwenfond der Universität Göttingen; seine hierbei gewonnene Erfahrung benutzte er für erfolgreiche Spekulationen mit Aktien.

Sätze und Verfahren von Gauß werden uns im weiteren Laufe der Vorlesung noch sehr häufig begegnen.

d) Weitere Körper

Wir kennen inzwischen die drei ineinander liegenden Körper

$$\mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}\,;$$

tatsächlich liegen zwischen $\mathbb{Q}$ und $\mathbb{C}$ noch zahlreiche andere Körper, mit denen wir uns hier zwar nicht ausführlich beschäftigen wollen, die aber trotzdem in der Informationsverarbeitung teilweise sehr wichtig sind.

Der Grund dafür liegt darin, daß die reellen Zahlen trotz ihres Namens alles andere als „real“ sind: Sie sind zwar ein sehr erfolgreiches Hilfsmittel zur Behandlung von Problemen aus den Naturwissenschaften, der Technik, den Wirtschaftswissenschaften *usw.*, aber es ist beispielsweise nicht möglich, eine beliebige reelle Zahl mit endlichem Aufwand zu beschreiben – erst recht hat man keine Chance, beliebige reelle Zahlen in einem Computer darzustellen.

Es kommt noch schlimmer: Nach einem 1968 von RICHARDSON bewiesenen Satz läßt sich selbst für zwei endlich beschreibbare reelle Zahlen im allgemeinen nicht entscheiden, ob sie gleich sind oder nicht.

Die Mathematik kennt zwei mehr oder weniger erfolgreiche Auswege aus diesem Dilemma: Standard in den meisten Anwendung ist die Approximation der reellen Zahlen durch sogenannte Gleitkommazahlen, die in einigen Programmiersprachen als *real* bezeichnet werden, in den meisten heute gebräuchlichen aber die korrektere Bezeichnung *float* haben. Mit den Möglichkeiten und Grenzen dieser Strategie beschäftigt

sich die *Numerische Mathematik;* da es darüber eine eigene Vorlesung gibt, werde ich solche Fragen in der *Höheren Mathematik* nur gelegentlich kurz am Rande erwähnen.

Die andere Strategie besteht darin, sich auf einen *Teilkörper* der reellen (oder komplexen) Zahlen zu beschränken, in dem man exakt rechnen kann. Dies ist der (gegenüber der Numerik erheblich aufwendigere) Ansatz der *Computeralgebra,* der beispielsweise bei manchen Problemen der Computergraphik verwendet werden muß, da man hier zum Erhalt der logischen und topologischen Konsistenz der Daten *exakt* wissen muß, ob zwei auf verschiedene Weise berechneten Punkte gleich sind oder nicht. Eine falsche Antwort auf diese Frage führt erstaunlich oft zum Systemabsturz, beispielsweise wegen einer Division durch Null.

Zum Glück gibt es einen Teilkörper von $\mathbb{R}$, den sogenannten Körper der berechenbaren Zahlen, in dem alle Berechnungen exakt und algorithmisch ausgeführt werden können – wenn auch meist sehr teuer. In der Praxis wendet man daher solche Verfahren meist nur dann an, wenn numerische Berechnungen keine hinreichend exakte Antwort garantieren können.

e) Der Körper mit zwei Elementen

Nicht jeder Körper läßt sich in die reellen oder komplexen Zahlen einbetten; das einfachste Gegenbeispiel ist folgendes:

In der digitalen Informationsverarbeitung gibt es fast überall genau zwei Zustände, die – unabhängig von ihrer tatsächlichen technischen Realisierung – üblicherweise mit 0 und 1 bezeichnet werden. Wir wollen aus der Menge $\mathbb{F}_2 = \{0, 1\}$ dieser beiden Zustände einen Körper machen.

Schon bei der Addition gibt es nicht viele Möglichkeiten: Wir müssen eines der beiden Elemente zum Neutralelement machen, wofür wir natürlich sinnvollerweise die Null wählen. Alsdann ist nach Definition der Eigenschaften eines Neutralelements

$$0 + 0 = 0 \quad \text{und} \quad 0 + 1 = 1 + 0 = 1 ;$$

die einzige noch unbekannte Summe ist also $1 + 1$. Wäre $1 + 1 = 1$, müßte nach Subtraktion von 1 auf beiden Seiten, $1 = 0$ sein, was wir nicht wollen, also müssen wir festlegen, daß $1 + 1 = 0$ ist.

Bei der Multiplikation ist alles noch deutlicher festgelegt: In jedem Körper ist für jedes Element x

$$0 \cdot x = (1 - 1) \cdot x = x - x = 0 \quad \text{und} \quad 1 \cdot x = 1 ,$$

also ist

$$0 \cdot 0 = 0, \quad 0 \cdot 1 = 1 \cdot 0 = 0 \quad \text{und} \quad 1 \cdot 1 = 1 .$$

Die Verknüpfungstabellen sehen damit folgendermaßen aus:

+	0	1
0	0	1
1	1	0

·	0	1
0	0	0
1	0	1

Ein Leser, der bereits über Kenntnisse der Logik und/oder der Schaltungstechnik verfügt, wird hier sicherlich bekanntes entdecken:

- Falls man 1 als *wahr* und 0 als *falsch* interpretiert, ist „·“ das logische Und, während „+“ das *exklusive* logische Oder ist. (Für Altphilologen ist dies das lateinische *aut* im Gegensatz zum *vel;* wer sich eher mit Logik oder Schaltalgebra auskennt, sollte zumindest eine der (äquivalenten) Bezeichnungen XOR oder *Antivalenz* schon einmal gehört haben.)
- Falls man ganze Zahlen in Binärdarstellung addieren möchte, ist für jede einzelne Binärstelle $x \cdot y$ der Übertrag, während $x + y$ bis auf den Übertrag der vorherigen Stelle gleich der Binärstelle des Ergebnisses ist. Man bezeichnet daher eine Schaltung, die $x + y$ und $x \cdot y$ berechnet auch als einen *Halbaddierer;* der Volladdierer, der ein Bit plus dem Übertrag des vorherigen Bits verarbeitet, besteht aus zwei Halbaddierern und einem Oder-Gatter.

So seltsam dieser Körper auf den ersten Blick auch aussehen mag, hat er also anscheinend doch das Potential für nützliche Anwendungen; einige davon werden wir schon bald kennenlernen. Wie wir dann sehen werden, gibt es noch eine ganze Reihe weiterer endlicher Körper mit wichtigen Anwendungen in der Kryptographie, der Kodierungstheorie und einer ganzen Reihe weiterer Gebiete der Informationsverarbeitung.

§2: Vektoren und Vektorräume

a) Vektoren in der Ebene und im Raum

Vektoren werden anschaulich dargestellt durch *Pfeile,* d.h. durch gerichtete Verbindungsstrecken zweier Punkte. Sie sind festgelegt durch die Angabe von Anfangs- und Endpunkt, aber auch beispielsweise durch die Angabe von Anfangspunkt, Länge und Richtung, wobei diese Richtung jedoch für Pfeile der Länge Null nicht definiert ist.

Pfeile dieser Art sind nützlich beispielsweise für die Darstellung von elektrischen oder magnetischen Feldern wie etwa den in Abbildung eins dargestellten: dem elektrischen Feld einer abstoßenden Punktladung und dem Magnetfeld eines stromdurchflossenen Leiters.

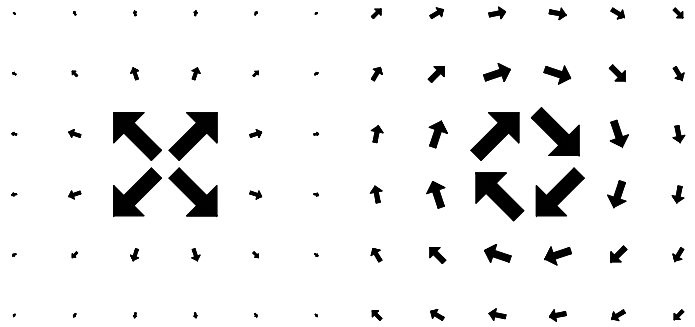


Abb. 1: Zwei elektromagnetische Felder

Zwei Pfeile lassen sich addieren, falls der Endpunkt des ersten gleich dem Anfangspunkt des zweiten ist; die Summe ist dann derjenige Pfeil, der den Anfangspunkt des ersten mit dem Endpunkt des zweiten Pfeils verbindet. Auch läßt sich ein Pfeil mit einer reellen Zahl multiplizieren, wenn wir vereinbaren, daß das Ergebnis jeder Pfeil sein soll, der denselben Anfangspunkt und dieselbe Richtung hat wie der ursprüngliche Pfeil, dessen Länge aber mit der reellen Zahl multipliziert wurde. (Eine

Multiplikation mit einer negativen Zahl soll dabei bedeuten, daß der Pfeil an seinem Anfangspunkt gespiegelt und dann mit dem Betrag der Zahl multipliziert wird.)

Sobald wir uns allerdings dafür interessieren, wie sich ein Teilchen im kombinierten Kraftfeld der Punktladung und des stromdurchflossenen Leiters bewegt, reichen Pfeile nicht mehr aus: Wir haben zwar für jeden Punkt der Ebene (außer dem Nullpunkt) einen Kraftpfeil für jedes Feld, aber natürlich müssen wir in jedem Punkt die beiden *dort* beginnenden Kraftpfeile addieren, was mit Pfeilen nicht geht.

Die Lösung dieses Problems ist wohl bekannt: Die beiden Pfeile werden gemäß dem „Parallelogramm der Kräfte“ kombiniert, d.h. der eine Pfeil wird so verschoben, daß sein Anfangspunkt gleich dem Endpunkt des anderen Pfeils ist. Wie Abbildung zwei zeigt, ist das Ergebnis unabhängig von der Reihenfolge der Summanden, d.h.

$$\vec{v} + \vec{w} = \vec{w} + \vec{v}\,.$$

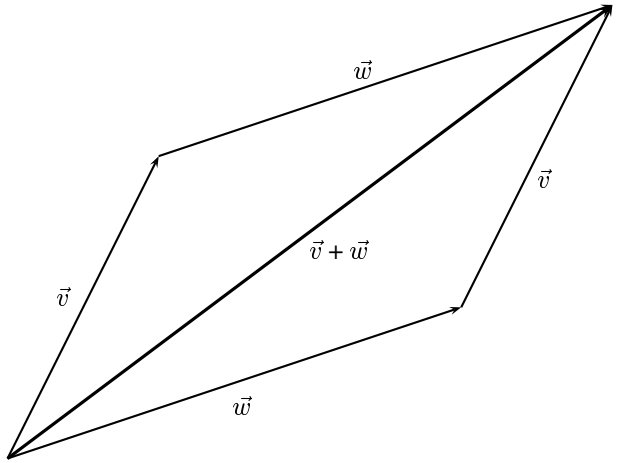


Abb. 2: Das „Parallelogramm der Kräfte“

Wir definieren daher als neuen Begriff einen *Vektor* als etwas, das zwar wie ein Pfeil eine Länge und eine Richtung haben soll, aber keinen

Anfangspunkt. Mathematisch exakt ausgedrückt ist also ein Vektor eine *Äquivalenzklasse* von Pfeilen, wobei zwei Pfeile genau dann äquivalent sind, wenn sie dieselbe Länge und (so die Länge von Null verschieden ist) dieselbe Richtung haben.

Vektoren werden in der Literatur meist durch Fraktur- oder Fettbuchstaben bezeichnet; da sich Fettdruck schlecht an der Tafel realisieren läßt und Frakturbuchstaben meist zu Hörerprotesten führen, verwenden wir hier stattdessen lateinische Buchstaben, die mit einem Pfeil überstrichen sind, also $\vec{u}, \vec{v}, \vec{w}, \ldots$. Die Addition zweier Vektoren wird durch das gewöhnliche Pluszeichen ausgedrückt, wir schreiben also $\vec{v}+\vec{w}$. Aus dem „Parallelogramm der Kräfte“ in Abbildung zwei liest man sofort ab, daß $\vec{v} + \vec{w} = \vec{w} + \vec{v}$ ist. Abbildung drei zeigt die Summe der beiden Kraftfelder aus Abbildung eins.

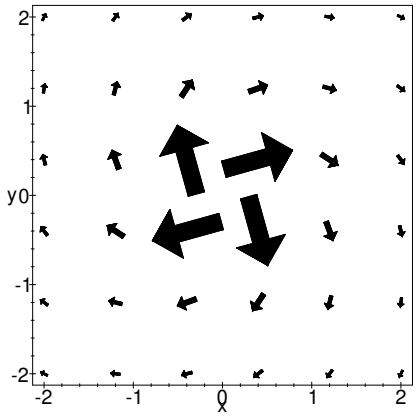


Abb. 3: Die Summe der beiden Felder aus Abbildung eins

Als weitere Eigenschaft der Vektoraddition wollen wir festhalten, daß es zu jedem Vektor $\vec{v}$ einen Vektor $\vec{w}$ gibt, so daß

$$\vec{v} + \vec{w} = \vec{0}$$

der Nullvektor ist; $\vec{w}$ ist einfach der entgegengesetzt orientierte Vektor $\vec{v}$. Wir bezeichnen diesen Vektor kurz als $-\vec{v}$.

Die Addition des Nullvektors ändert natürlich nichts am anderen Summanden, d.h.

$$\vec{v} + \vec{0} = \vec{0} + \vec{v} = \vec{v} \quad \text{für alle Vektoren } \vec{v}\,.$$

Schließlich gilt für die Vektoraddition auch noch das Assoziativgesetz

$$\big(\vec{u} + \vec{v}\big) + \vec{w} = \vec{u} + \big(\vec{v} + \vec{w}\big)\,,$$

wie man sich leicht überzeugt, indem man das Diagramm für die Konstruktion von $\vec{v} + \vec{w}$ an den Endpunkt des Vektors $\vec{u}$ verschiebt; siehe Abbildung vier.

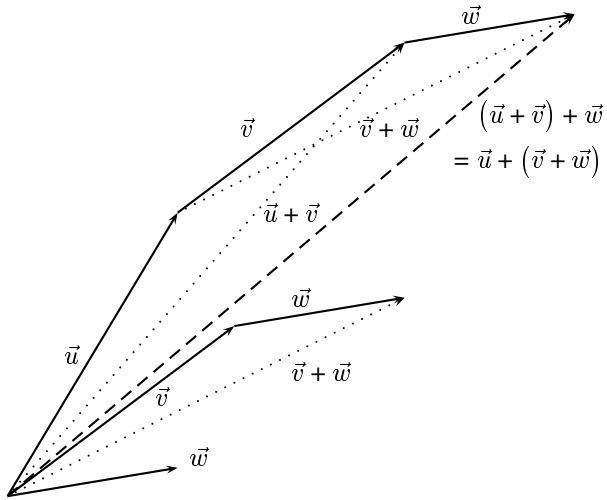


Abb. 4: Das Assoziativgesetz der Vektoraddition

Außer der Addition von Vektoren können wir auch ihre Streckung, d.h. ihre Multiplikation mit einer reellen Zahl, definieren: Ist $\vec{v}$ ein Vektor und $\lambda > 0$ eine positive reelle Zahl, so soll $\lambda\vec{v}$ dieselbe Richtung haben wie $\vec{v}$ und die λ-fache Länge; für $\lambda < 0$ soll $\lambda\vec{v}$ die entgegengesetzte Richtung haben und die $|\lambda|$-fache Länge. Für $\lambda = 0$ schließlich ist $\lambda\vec{v}$ der Nullvektor.

Anwendung des Strahlensatzes auf das Dreieck in Abbildung fünf zeigt, daß für diese Multiplikation das Distributivgesetz

$$\lambda(\vec{v} + \vec{w}) = \lambda\vec{v} + \lambda\vec{w}$$

gilt: Als Strahlen betrachten wir von $\vec{v}$ und $\vec{v}+\vec{w}$ aufgespannten Halbgeraden, und wir schneiden mit den beiden parallelen Geraden durch die eingezeichneten Vektoren $\lambda\vec{w}$ und $\vec{w}$. Dabei sollen die fett eingezeichneten Vektoren die mit λ multiplizierten sein; im Bild ist $\lambda = 0{,}4$.

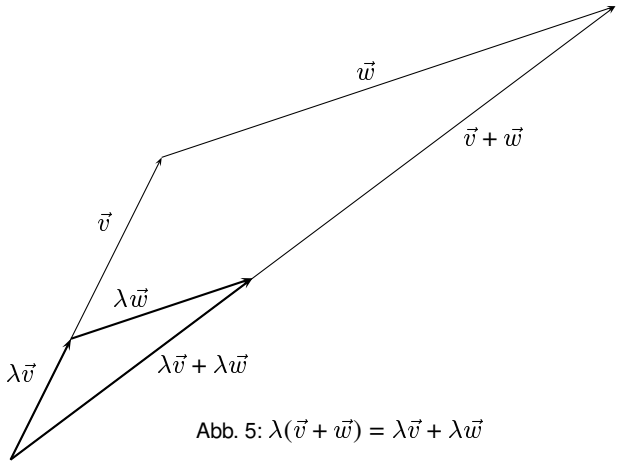


Abb. 5: $\lambda(\vec{v} + \vec{w}) = \lambda\vec{v} + \lambda\vec{w}$

Das andere Distributivgesetz $(\lambda + \mu)\vec{v} = \lambda\vec{v} + \mu\vec{v}$ ist ziemlich trivial: Da sich alles auf der von $\vec{v}$ aufgespannten Geraden abspielt und wir diese mit der reellen Zahlengeraden identifizieren können, läßt sich diese Regel auf das gewöhnliche Distributivgesetz in $\mathbb{R}$ zurückführen, genauso wie sich die Regel $(\lambda\mu)\vec{v} = \lambda(\mu\vec{v})$ auf das gewöhnliche Assoziativgesetz der Multiplikation in $\mathbb{R}$ zurückführen läßt.

b) Definition des Vektorraums

Damit haben wir alle Rechenregeln zusammen, die wir für die Definition eines Vektorraums brauchen. Da wir Vektoren auch mit Zahlen multiplizieren wollen, müssen wir zwei Arten von Objekten betrachten:

Vektoren, die wir weiterhin mit $\vec{v}$, $\vec{w}$ *usw.* bezeichnen, sowie Skalare, für die wir griechische Buchstaben verwenden.

Für die Skalare lassen wir, wie bereits in §1*c)* diskutiert, Elemente eines beliebigen Körpers zu; für den Anfänger ist es wahrscheinlich am einfachsten, sich die Skalare zunächst als reelle Zahlen vorzustellen.

Definition: k sei ein Körper. Eine Menge V heißt *Vektorraum* über k oder k-Vektorraum, wenn es zwei Verknüpfungen

$$+: V \times V \to V \quad \text{und} \quad \cdot: k \times V \to V$$

gibt, so daß gilt:

I.1) Das Assoziativgesetz der Vektoraddition

$$(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w}) \qquad \text{für alle } \vec{u}, \vec{v}, \vec{w} \in V$$

I.2) Es gibt einen Vektor $\vec{0} \in V$, so daß für jeden Vektor $\vec{v} \in V$ gilt

$$\vec{v} + \vec{0} = \vec{0} + \vec{v} = \vec{v}\,.$$

I.3) Zu jedem Vektor $\vec{v} \in V$ gibt es einen Vektor $-\vec{v} \in V$, so daß

$$\vec{v} + (-\vec{v}) = (-\vec{v}) + \vec{v} = \vec{0}\,.$$

I.4) Das Kommutativgesetz der Vektoraddition

$$\vec{u} + \vec{v} = \vec{v} + \vec{u} \qquad \text{für alle } \vec{u}, \vec{v} \in V$$

II.1) Das Distributivgesetz bei der Addition von Skalaren

$$(\lambda + \mu)\vec{v} = \lambda\vec{v} + \mu\vec{v} \quad \text{für alle } \lambda, \mu \in k \text{ und alle } \vec{v} \in V\,.$$

II.2) Das Distributivgesetz bei der Addition von Vektoren

$$\lambda(\vec{v} + \vec{w}) = \lambda\vec{v} + \lambda\vec{w} \quad \text{für alle } \lambda \in k \text{ und alle } \vec{v}, \vec{w} \in V\,.$$

II.3) Kompatibilität von Körper- und Skalarmultiplikation

$$(\lambda\mu)\vec{v} = \lambda(\mu\vec{v}) \quad \text{für alle } \lambda, \mu \in k \text{ und alle } \vec{v} \in V\,.$$

II.4) Multiplikation mit der Eins

$$1\vec{v} = \vec{v} \quad \text{für alle } \vec{v} \in V\,.$$

II.5) Multiplikation mit der Null *bzw.* mit dem Nullvektor

$$0\vec{v} = \vec{0} \quad \text{für alle } \vec{v} \in V \qquad \text{und} \qquad \lambda\vec{0} = \vec{0} \quad \text{für alle } \lambda \in k\,.$$

Bemerkung: Die Forderungen I.1–I.4 in der Definition des Körpers und des Vektorraums sowie die Forderungen II.1–II.4 in der Körperdefinition sind fast identisch, und in der Tat beschreiben sie eine gemeinsame mathematische Struktur, die sogenannte *abelsche Gruppe*. Da wir diese nicht weiter benötigen werden, sei auf Einzelheiten verzichtet.

Vektoren und Vektorräume sind als mathematische Begriffe recht jung: Rechnerische Methoden zur Lösung geometrischer Probleme wurden zwar schon ab etwa 1636 von RENÉ DESCARTES (1596–1650) eingesetzt (kartesische Koordinaten), aber erst gegen Mitte des 19. Jahrhunderts wurden Ansätze entwickelt, um geometrische Objekte *koordinatenfrei* algebraisch zu behandeln. Ein erster Durchbruch war das 1844 erschienene Buch *Die Ausdehnungslehre* von HERMANN GÜNTER GRASSMANN (1809–1877, oberes Bild): Er betrachtete abstrakte Objekte, die unter anderem alle Vektorraumaxiome erfüllten, die darüber hinaus allerdings auch miteinander multipliziert werden konnten, so daß er etwas komplizierteres als einen Vektorraum definiert hatte: eine sogenannte Algebra. Sie spielt noch heute eine große Rolle bei der Charakterisierung der Lage zweier Vektorräume ineinander. 1888 definierte GIUSEPPE PEANO (1858–1932, unteres Bild) in seinem Buch *Calcolo geometrico secondo l'Ausdehnungslehre di H. Grassmann preceduto dalle operazioni della logica deduttiva* Vektorräume (über $\mathbb{R}$) im obigen Sinne; in diesem Buch treten auch erstmalig mengentheoretische Symbole wie $\cap$, $\cap$ und $\in$ auf. Ab etwa 1920 wandte STEFAN BANACH (1892–1945) PEANOs Theorie an auf Funktionenräume und lineare Operatoren.

c) Erste Beispiele

Standardbeispiel sind natürlich die $\mathbb{R}$-Vektorräume $\mathbb{R}^n$. Wir schreiben ihre Elemente als Spaltenvektoren der Form

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}, \quad \vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad \dots$$

und haben die beiden Rechenoperationen

$$\vec{v} + \vec{w} = \begin{pmatrix} v_1 + w_1 \\ \vdots \\ v_n + w_n \end{pmatrix} \quad \text{und} \quad \lambda\vec{v} = \begin{pmatrix} \lambda v_1 \\ \vdots \\ \lambda v_n \end{pmatrix}.$$

Alle Rechenregeln folgen sofort aus den entsprechenden Regeln für reelle Zahlen, und genauso können wir auch für einen beliebigen Körper k die k-Vektorräume k^n definieren.

Auf den ersten Blick seltsam erscheint, daß $\mathbb{R}$ ein $\mathbb{Q}$-Vektorraum ist: Vektoraddition ist die gewöhnliche Addition reeller Zahlen und Multiplikation mit Skalaren die Multiplikation einer reellen Zahl mit einer rationalen. Auch hier folgen alle Vektorraumaxiome sofort aus den üblichen Rechenregeln für reelle Zahlen, für die es natürlich gleichgültig ist, daß hier einige der betrachteten Zahlen sogar rational sind.

Interessanter ist das folgende Beispiel: Für eine natürliche Zahl $n \in \mathbb{N}$ und eine offene Teilmenge U von $\mathbb{R}$, als z.B. ein offenes Intervall $(a,\ b)$ oder $\mathbb{R}$ selbst, definieren wir die Menge

$$\mathcal{C}^n(U,\mathbb{R}) = \left\{ f\colon U \to \mathbb{R} \;\middle|\; f \text{ ist (mindestens) } n\text{-mal stetig differenzierbar} \right\}.$$

Für deren Elemente seien Addition und Skalarmultiplikation punktweise definiert, d.h. für $f, g \in \mathcal{C}^n(U,\mathbb{R})$ und $\lambda \in \mathbb{R}$ setzen wir

$$f+g\colon \begin{cases} U \to \mathbb{R} \\ t \mapsto f(t)+g(t) \end{cases} \quad \text{und} \quad \lambda f\colon \begin{cases} U \to \mathbb{R} \\ t \mapsto \lambda f(t) \end{cases}.$$

Abbildung sechs zeigt für $(a,\ b) = (0,\ 3\pi)$ zu den beiden dünn eingezeichneten Funktionen $f(t) = \sin t$ und $g(t) = \sin 3t$ die dick eingezeichnete Funktion $f+g$, und Abbildung sieben zeigt f zusammen mit der Funktion $2f$.

Damit dies alles wohldefiniert ist, müssen wir uns noch überlegen, daß mit f und g die Funktionen $f+g$ und λf wieder in $\mathcal{C}^n(U,\mathbb{R})$ liegen. Das ist aber klar, denn die Summe zweier stetiger *bzw.* differenzierbarer Funktionen ist wieder stetig *bzw.* differenzierbar, und wegen der Rechenregel $(f+g)' = f' + g'$ gilt dies auch für die höheren Ableitungen. Genauso kann man für λf argumentieren. Da alle Rechenoperationen auf

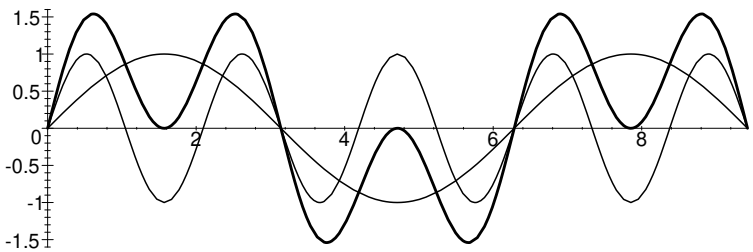


Abb. 6: Die Summe von $f(t) = \sin t$ und $g(t) = \sin 3t$

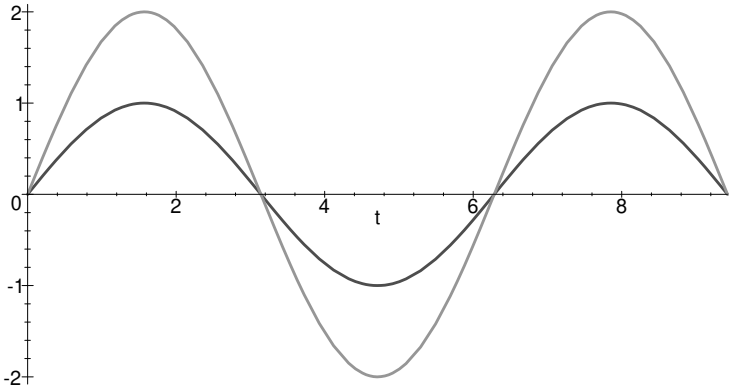


Abb. 7: $f(t) = \sin t$ zusammen mit $2f$

die gewöhnliche reelle Addition und Multiplikation für die Funktionswerte zurückgeführt ist, folgen die Vektorraumaxiome aus den üblichen Rechenregeln für reelle Zahlen: Um etwa das Assoziativgesetz

$$(f + g) + h = f + (g + h)$$

nachzuweisen, müssen wir zeigen, daß für jede reelle Zahl t die Funktionen auf beiden Seiten denselben Wert haben, d.h.

$$\big((f + g) + h\big)(t) = \big(f + (g + h)\big)(t) \quad \text{für alle } t \in \mathbb{R}\,.$$

Dazu rechnen wir beide Seiten aus:

$$\big((f + g) + h\big)(t) = (f + g)(t) + h(t) = \big(f(t) + g(t)\big) + h(t)$$

und

$$\big(f + (g + h)\big)(t) = f(t) + (g + h)(t) = f(t) + \big(g(t) + h(t)\big)\,,$$

und die beiden rechten Seiten stimmen in der Tat überein nach dem Assoziativgesetz für die Addition reeller Zahlen.

Die restlichen Axiome folgen genauso, nur etwas einfacher.

Ganz entsprechend lassen sich auch die Mengen

$$\mathcal{C}^0(U, \mathbb{R}) = \big\{f: U \to \mathbb{R} \;\big|\; f \text{ ist stetig}\big\}$$

und

$$\mathcal{C}^\infty(U, \mathbb{R}) = \big\{f: U \to \mathbb{R} \;\big|\; f \text{ ist beliebig oft differenzierbar}\big\}$$

sowie

$$\mathcal{C}^\omega(U, \mathbb{R}) = \left\{f: U \to \mathbb{R} \;\middle|\; \begin{array}{l} f \text{ ist um jeden Punkt } x \in U \text{ durch} \\ \text{eine TAYLOR-Reihe darstellbar} \end{array}\right\}$$

zu $\mathbb{R}$-Vektorräumen machen. (Wer noch nicht weiß, was eine TAYLOR-Reihe ist, wird es im nächsten Kapitel lernen.)

Als trivialstes Beispiel überhaupt haben wir schließlich noch über jedem Körper k den Nullvektorraum, der nur aus dem Nullvektor $\vec{0}$ besteht.

d) Lineare Abbildungen

Vektorräume werden erst richtig interessant, wenn man mit ihren Elementen etwas mehr tun kann als sie nur zu addieren und mit Skalaren zu multiplizieren. In der Geometrie etwa möchte man Vektoren gelegentlich auch drehen, bei Vektorräumen von differenzierbaren Funktionen möchte man deren Elemente differenzieren und so weiter. Viele derartige Operationen lassen sich unter dem Begriff der linearen Abbildung einordnen:

Definition: *a)* Eine Abbildung $\varphi: V \to W$ heißt *linear,* wenn für alle Vektoren $\vec{u}, \vec{v} \in V$ und alle $\lambda, \mu \in k$ gilt:

$$\varphi(\lambda\vec{u} + \mu\vec{v}) = \lambda\varphi(\vec{u}) + \mu\varphi(\vec{v})\,.$$

b) Unter dem *Kern* von φ verstehen wir die Menge

$$\operatorname{Kern}\varphi \underset{\text{def}}{=} \left\{ \vec{v} \in V \;\middle|\; \varphi(\vec{v}) = \vec{0} \right\} .$$

c) Das *Bild* von φ ist die Menge

$$\operatorname{Bild}\varphi \underset{\text{def}}{=} \left\{ \vec{w} \in W \;\middle|\; \text{Es gibt } \vec{v} \in V \text{ mit } \varphi(\vec{v}) = \vec{w} \right\} .$$

Die beiden allereinfachsten Beispiele für lineare Abbildungen sind für jeden Vektorraum V die identische Abbildung $V \to V$ sowie die Nullabbildung, die jedem Vektor $\vec{v} \in V$ den Nullvektor zuordnet. Letztere kann man wahlweise als Abbildung $V \to V$ oder als Abbildung von V in den Nullvektorraum auffassen.

Ebenfalls völlig trivial ist die Linearität von *Projektionen* wie etwa der Projektion $\mathbb{R}^3 \to \mathbb{R}^2$, die jedem Vektor seine ersten beiden Komponenten zuordnet.

Ein etwas interessanteres Beispiel einer linearen Abbildung ist

$$\varphi\colon \begin{cases} \mathbb{R}^3 & \to \quad \mathbb{R}^2 \\ \begin{pmatrix} x \\ y \\ z \end{pmatrix} & \mapsto \begin{pmatrix} x - y \\ y - z \end{pmatrix} \end{cases} ;$$

sie ist linear, denn

$$\varphi\left(\lambda \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + \mu \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} \right) = \varphi\left(\begin{pmatrix} \lambda x_1 + \mu x_2 \\ \lambda y_1 + \mu y_2 \\ \lambda z_1 + \mu z_2 \end{pmatrix} \right)$$
$$= \begin{pmatrix} \lambda x_1 + \mu x_2 - \lambda y_1 - \mu y_2 \\ \lambda y_1 + \mu y_2 - \lambda z_1 - \mu z_2 \end{pmatrix}$$

und

$$\lambda \cdot \varphi\left(\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \right) + \mu \cdot \varphi\left(\begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} \right) = \lambda \begin{pmatrix} x_1 - y_1 \\ y_1 - z_1 \end{pmatrix} + \mu \begin{pmatrix} x_2 - y_2 \\ y_2 - z_2 \end{pmatrix}$$
$$= \begin{pmatrix} \lambda x_1 - \lambda y_1 + \mu x_2 - \mu y_2 \\ \lambda y_1 - \lambda z_1 + \mu y_2 - \mu z_2 \end{pmatrix} ,$$

was offensichtlich dasselbe ist.

Bei Vektorräumen von Funktionen ist beispielsweise für jeden Punkt $t_0 \in (a,\, b)$ die Auswertungsabbildung

$$\mathcal{C}^n\big((a,\, b), \mathbb{R}\big) \to \mathbb{R}; \quad f \mapsto f(t_0)$$

nach Definition der Vektorraumoperationen von $\mathcal{C}^n\big((a,\, b), \mathbb{R}\big)$ linear, denn $\lambda f + \mu g$ wurde ja gerade so definiert, daß für t_0 wie auch für jeden anderen Punkt aus $(a,\, b)$ gilt

$$\big(\lambda f + \mu g\big)(t_0) = \lambda f(t_0) + \mu g(t_0) .$$

Allgemeiner können wir auch die *Abtastung* einer Funktion betrachten: Für vorgegebene Punkte $t_1, \ldots, t_N \in (a,\, b)$ definieren wir die Abbildung

$$\varphi\colon \begin{cases} \mathcal{C}^n\big((a,\, b), \mathbb{R}\big) & \to \mathbb{R}^N \\ f & \mapsto \begin{pmatrix} f(t_1) \\ \vdots \\ f(t_N) \end{pmatrix} \end{cases} ,$$

die f an N Argumenten auswertet. Anwendung ist beispielsweise die Digitalisierung eines Signals, etwa eines Musikstücks für eine CD. Hier würde die Funktion f die zeitliche Variation des Schalldrucks beschreiben (die man zumindest als stetig annehmen kann, d.h. $n = 0$), und die Punkte t_i wären gleichmäßig über die Länge des Musikstücks verteilt, jeweils 44 100 Stück pro Sekunde.

Diese Abbildung ist linear, weil jede ihrer Komponentenabbildungen $f \mapsto f(t_i)$ linear ist. Wir können die Linearität dieser Digitalisierung eines Signals aber auch inhaltlich interpretieren: Die Eigenschaft

$$\varphi(\lambda f + \mu g) = \lambda\varphi(f) + \mu\varphi(g)$$

bedeutet für positive λ und μ, daß es gleichgültig ist, ob man zwei verschiedene Signale (z.B. Mikrophonkanäle) zunächst in einem analogen Mischpult vereinigt und dann digitalisiert oder zunächst digitalisiert und dann digital mischt. (Dieses setzt natürlich voraus, daß man sowohl analog als auch digital mit perfekter Genauigkeit arbeitet – keine sehr realistische Annahme.)

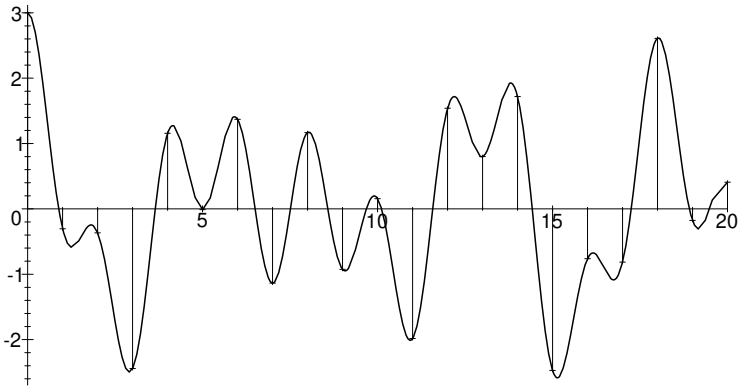


Abb. 8: Abtastung eines Signals

Ein anderes Beispiel einer linearen Abbildung zwischen Vektorräumen von Funktionen ist die Differentiation

$$\mathcal{C}^{n+1}\big((a,b),\mathbb{R}\big) \to \mathcal{C}^{n}\big((a,b),\mathbb{R}\big); \quad f \mapsto \dot{f}\,,$$

denn $(\lambda f + \mu g)' = \lambda f' + \mu g'$. Auch die Abbildung

$$\mathcal{C}^{2}\big((a,b),\mathbb{R}\big) \to \mathcal{C}^{0}\big((a,b),\mathbb{R}\big); \quad f \mapsto f'' + \omega^2 f$$

ist für jedes $\omega \in \mathbb{R}$ linear; ihr Kern besteht genau aus jenen Funktionen $f(t)$, die der Schwingungsdifferentialgleichung

$$f''(t) + \omega^2 f(t) = 0$$

genügen, enthält also beispielsweise die Funktionen $\cos \omega t$ und $\sin \omega t$.

e) Untervektorräume

Kern und Bild einer linearen Abbildung werden im allgemeinen unendliche Mengen sein, so daß selbst bei Vektorräumen wie dem $\mathbb{R}^3$ zunächst nicht ganz klar ist, wie man sie mit endlichem Aufwand beschreiben kann. Bei nichtlinearen Abbildungen kann so etwas in der Tat ein großes Problem sein, aber hier im Linearen reichen unsere vorhandenen Werkzeuge zumindest für Vektorräume wie einen $\mathbb{R}^n$ völlig aus. Zur Klärung der Begriffe beginnen wir mit einer

Definition: Eine Teilmenge $U \subseteq V$ eines k-Vektorraums V heißt *Untervektorraum,* in Zeichen $U \leq V$, wenn U nicht leer ist und mit je zwei Vektoren $\vec{u}, \vec{v} \in U$ und Skalaren $\lambda, \mu \in k$ auch den Vektor $\lambda\vec{u} + \mu\vec{v}$ enthält.

Lemma: $\varphi\colon V \to W$ sei eine lineare Abbildung zwischen zwei k-Vektorräumen. Dann ist Kern φ ein Untervektorraum von V und Bild φ ein Untervektorraum von W.

Beweis: Sind $\vec{u}$ und $\vec{v}$ Elemente des Kerns von $\varphi\colon V \to W$ und $\lambda, \mu \in k$ Skalare, so ist

$$\varphi(\lambda\vec{u} + \mu\vec{v}) = \lambda\varphi(\vec{u}) + \mu\varphi(\vec{u}) = \lambda\vec{0} + \mu\vec{0} = \vec{0}\,,$$

also liegt auch $\lambda\vec{u} + \mu\vec{v}$ im Kern von φ. Außerdem ist dieser nicht leer, denn wegen

$$\varphi(\vec{0}) = \varphi(0 \cdot \vec{0}) = 0 \cdot \varphi(\vec{0}) = \vec{0}$$

liegt der Nullvektor in Kern φ. Also ist der Kern ein Untervektorraum.

Ähnlich ist die Situation für das Bild: Für zwei Vektoren $\vec{v}, \vec{w} \in$ Bild φ gibt es Vektoren $\vec{r}, \vec{s} \in V$, so daß $\varphi(\vec{r}) = \vec{v}$ und $\varphi(\vec{s}) = \vec{w}$ ist. Wegen der Linearität von φ liegt dann für zwei Skalare $\lambda, \mu \in k$ auch

$$\lambda\vec{v} + \mu\vec{w} = \lambda\varphi(\vec{r}) + \mu\varphi(\vec{s}) = \varphi(\lambda\vec{r} + \mu\vec{s})$$

im Bild von φ, das somit ein Untervektorraum von W ist. ∎

Kern und Bild einer linearen Abbildung haben natürlich etwas mit deren Injektivität und Surjektivität zu tun; erinnern wir uns zunächst an die Definition dieser Begriffe:

Definition: *a)* Eine Abbildung $\varphi\colon M \to N$ zwischen zwei Mengen heißt *injektiv,* wenn keine zwei verschiedenen Elemente von M dasselbe Bild haben, d.h. aus der Gleichheit von $\varphi(m_1)$ und $\varphi(m_2)$ folgt für zwei Elemente $m_1, m_2 \in M$, daß $m_1 = m_2$ ist.
b) φ heißt *surjektiv,* wenn es zu jedem $n \in N$ (mindestens) ein $m \in M$ gibt, so daß $\varphi(m) = n$ ist.
c) φ heißt *bijektiv* oder auch „eins-zu-eins (1-1)“, wenn φ injektiv und surjektiv ist.

Lemma: $\varphi: V \to W$ ist genau dann injektiv, wenn Kern φ der Nullvektorraum ist; φ ist genau dann surjektiv, wenn Bild $\varphi = W$ ist.

Beweis: Die zweite Aussage ist zu trivial, als daß man etwas dazu sagen müßte, betrachten wir also die erste. Falls φ injektiv ist, hat insbesondere der Nullvektor nur ein einziges Urbild, d.h. der Kern besteht nur aus dem Nullvektor, der natürlich immer im Kern liegt. Ist umgekehrt Kern φ der Nullraum und haben zwei Vektoren $\vec{u}, \vec{v} \in V$ dasselbe Bild, so ist

$$\varphi(\vec{u} - \vec{v}) = \varphi(\vec{u}) - \varphi(\vec{v}) = \vec{0}\,,$$

d.h. $\vec{u} - \vec{v}$ liegt im Kern und muß daher gleich dem Nullvektor sein, so daß $\vec{u} = \vec{v}$ ist. Dies zeigt die Injektivität von φ. ∎

Als Beispiel einer Anwendung dieses Lemmas betrachten wir noch einmal die Digitalisierung eines Signals: Aufgrund der hoch gelobten CD-Qualität erwarten wir, daß in diesem Fall die Abtastung eine „einigermaßen injektive“ lineare Abbildung ist. Das Wort „einigermaßen injektiv“ ist zwar kein wohldefinierter mathematischer Begriff, aber schon die Tatsache, daß bei einer CD die Abtastwerte nicht als reelle Zahlen gespeichert werden, sondern als 16bit-Zahlen, macht eine „echte“ Injektivität unmöglich. Überlegen wir uns, was sonst noch alles schiefgehen kann.

Nach dem gerade bewiesenen Lemma reicht es, wenn wir den Kern der Abbildung kennen. Dort liegt, bei einer Abtastung mit 44 100 Hz und in Sekunden gemessener Zeit, beispielsweise die Funktion

$$f(t) = \sin\bigl(44\,100\,\pi t\bigr) = \sin\bigl(22\,050 \cdot 2\pi t\bigr)\,;$$

denn für jedes ganzzahlige Vielfache von $1/44\,100$ ist das Argument des Sinus ein ganzzahliges Vielfaches von π, der Sinus also Null.

Die Funktion $f(t)$ entspricht einer reinen Schwingung mit einer Frequenz von 22,05 kHz. Solche Frequenzen sind zwar sehr wichtig für die Navigation von Fledermäusen, sie sind aber unhörbar für Käufer von CDs, so daß uns dieses Element des Kerns nicht weiter stört.

Betrachten wir aber beispielsweise die Funktionen

$$g(t) = \sin\bigl(66\,150\,\pi t\bigr) \quad \text{und} \quad h(t) = \sin\bigl(22\,050\,\pi t\bigr)\,.$$

Für $k \in \mathbb{Z}$ und $t = k/44\,100$ ist

$$g(t) = g\left(\frac{k}{44\,100}\right) = \sin\left(66\,150\pi \cdot \frac{k}{44\,100}\right) = \sin\left(\frac{3k\pi}{2}\right)$$
$$= \begin{cases} 0 & \text{für gerades } k \\ -1 & \text{für } k \equiv 1 \bmod 4 \\ 1 & \text{für } k \equiv 3 \bmod 4 \end{cases}$$

und

$$h(t) = h\left(\frac{k}{44\,100}\right) = \sin\left(22\,050\pi \cdot \frac{k}{44\,100}\right) = \sin\left(\frac{k\pi}{2}\right)$$
$$= \begin{cases} 0 & \text{für gerades } k \\ 1 & \text{für } k \equiv 1 \bmod 4\ , \\ -1 & \text{für } k \equiv 3 \bmod 4 \end{cases}$$

wobei $k \equiv a \bmod b$ bedeuten soll, daß $k - a$ durch b teilbar ist. Damit sind die beiden Funktionen g und h an allen Abtaststellen entgegengesetzt gleich, d.h. die Funktion $g + h$ liegt im Kern von φ.

Dieses Element des Kern stört uns erheblich mehr, denn es hat zur Folge, daß die beiden Funktion $g(t) = \sin\bigl(66\,150\,\pi t\bigr)$ und $-h(t) = -\sin\bigl(22\,050\,\pi t\bigr)$ auf dieselbe Weise digitalisiert werden. g beschreibt aber einen für Menschen unhörbaren Ton mit einer Frequenz von 33,075 kHz, während h mit nur 11,025 kHz durchaus hörbar ist. Abbildung neun zeigt die beiden Schwingungen; die Zeitachse ist der besseren Übersicht wegen in Millisekunden beschriftet, und die Abtastwerte sind durch Quadrate markiert.

Die Digitalisierungsabbildung kann also höchstens dann injektiv sein, wenn wir uns auf Funktionen beschränken, an deren Aufbau keine Schwingungen mit einer Frequenz von 22,05 kHz oder höher beteiligt sind. Was das bedeutet, und ob dann wirklich Injektivität gilt, werden wir in der *Höheren Mathematik II* im Kapitel über harmonische Analyse genauer untersuchen.

f) Lineare Abhängigkeit

Im $\mathbb{R}^3$ definieren zwei Vektoren eine Ebene – es sei denn, sie liegen, wenn man sie am gleichen Anfangspunkt beginnen läßt, auf einer Geraden, d.h einer der beiden Vektoren ist ein Vielfaches des anderen.

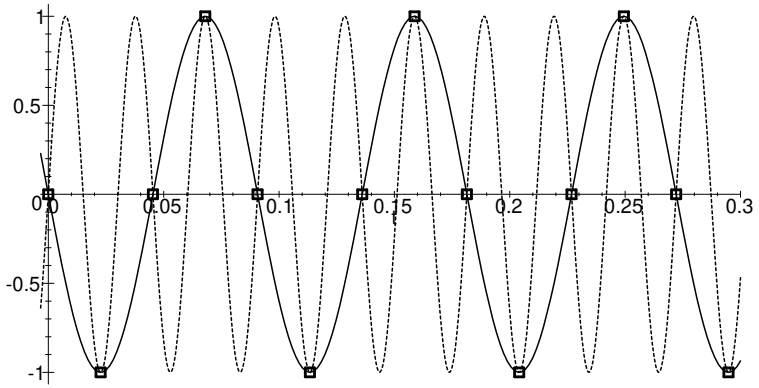


Abb. 9: Zwei verschiedene Signale, die gleich abgetastet werden

Entsprechend spannen drei Vektoren im allgemeinen den gesamten $\mathbb{R}^3$ auf – es sei denn, sie liegen, wenn man sie am gleichen Anfangspunkt beginnen läßt, in einer Ebenen, d.h. einer der drei ist als Summe von Vielfachen der anderen beiden darstellbar.

Der Begriff der *linearen Abhängigkeit* verallgemeinert diese Ausnahmebedingungen so, daß sie auf beliebige Vektorräume angewandt werden können:

Definition: $\vec{v}_1, \ldots, \vec{v}_n$ seien Elemente des k-Vektorraums V.

a) Eine *Linearkombination* von $\vec{v}_1, \ldots, \vec{v}_n$ ist eine Summe der Form

$$\lambda_1 \vec{v}_1 + \cdots + \lambda_n \vec{v}_n$$

mit Skalaren $\lambda_i \in k$; ist diese Summe gleich dem Vektor $\vec{v} \in V$, so sagen wir, $\vec{v}$ sei als Linearkombination von Vektoren aus M darstellbar.

b) Die Menge aller Vektoren, die sich als Linearkombination der Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ darstellen lassen, bezeichnen wir mit $[\vec{v}_1, \ldots, \vec{v}_n]$; wir nennen sie das *Erzeugnis* von $\vec{v}_1, \ldots, \vec{v}_n$.

c) Eine Linearkombination wie in *a)* heißt *nichttrivial,* falls mindestens ein λ_i von Null verschieden ist; ansonsten heißt sie *trivial.*

d) Die Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ heißen *linear unabhängig,* wenn der Nullvektor nicht als nichttriviale Linearkombination von $\vec{v}_1, \ldots, \vec{v}_n$ darstellbar ist, d.h. eine Gleichung der Form

$$\lambda_1 \vec{v}_1 + \cdots + \lambda_n \vec{v}_n = \vec{0}$$

kann nur gelten, wenn alle λ_i verschwinden.

e) Sind $\vec{v}_1, \ldots, \vec{v}_n$ *nicht* linear unabhängig, so bezeichnen wir sie als *linear abhängig.*

f) Eine *Teilmenge* $M \subseteq V$ eines Vektorraums V heißt *linear unabhängig,* wenn jede Auswahl endlich vieler verschiedener Vektoren $\vec{v}_1, \ldots \vec{v}_m$ (für beliebiges $m \in \mathbb{N}$) linear unabhängig ist.

g) Das *Erzeugnis* $[M]$ einer Teilmenge $M \subseteq V$ eines Vektorraums V ist die Menge aller Vektoren aus V, die als Linearkombination aus endlich vielen Vektoren aus V dargestellt werden können.

Beispielsweise sind also die Vektoren

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} \in \mathbb{R}^3$$

linear abhängig, da der zweite das Zweifache des ersten ist, und auch die Vektoren

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \in \mathbb{R}^3$$

sind linear abhängig, denn

$$\lambda \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} + \nu \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda + 2\mu + \nu \\ 3\mu + \nu \\ 0 \end{pmatrix}$$

ist gleich dem Nullvektor wann immer $\nu = -3\mu$ und $\lambda = -2\mu - \nu = \mu$ ist. Eine nichttriviale Darstellung des Nullvektors ist beispielsweise

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} - 3 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Die drei Vektoren

$$\begin{pmatrix}1\\0\\0\end{pmatrix},\quad \begin{pmatrix}0\\1\\0\end{pmatrix},\quad \begin{pmatrix}0\\0\\1\end{pmatrix}\ \in\ \mathbb{R}^3$$

dagegen sind linear unabhängig, denn

$$\lambda_1\begin{pmatrix}1\\0\\0\end{pmatrix}+\lambda_2\begin{pmatrix}0\\1\\0\end{pmatrix}+\lambda_3\begin{pmatrix}0\\0\\1\end{pmatrix}=\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}$$

ist genau dann gleich dem Nullvektor, wenn alle λ_i verschwinden.

Allgemein sind die Vektoren $\vec{v}_1,\dots,\vec{v}_n$ aus einem beliebigen Vektorraum V dann trivialerweise linear abhängig, wenn zwei Vektoren $\vec{v}_i$ und $\vec{v}_j$ (mit $j\neq i$) gleich sind, denn dann ist beispielsweise

$$1\cdot\vec{v}_i+(-1)\cdot\vec{v}_j=\vec{0}$$

eine nichttriviale Darstellung des Nullvektors. Ebenfalls trivial ist die lineare Abhängigkeit, falls einer der Vektoren $\vec{v}_i$ gleich dem Nullvektor ist: Dann ist bereits

$$1\cdot\vec{v}_i=\vec{0}$$

eine solche Darstellung. Eine Menge, die den Nullvektor enthält, ist also stets linear abhängig.

Auch in Vektorräumen von Funktionen können wir leicht Beispiele für lineare Abhängigkeit und Unabhängigkeit finden. In $\mathcal{C}^0(\mathbb{R},\mathbb{R})$ sind etwa Sinus und Kosinus linear unabhängig, denn gäbe es $\lambda_{1/2}\in\mathbb{R}$ mit

$$\lambda_1\sin x+\lambda_2\cos x=0\quad\text{für alle } x\in\mathbb{R}$$

mit $\lambda_1\neq 0$, so wäre

$$\tan x=\frac{\sin x}{\cos x}=-\frac{\lambda_2}{\lambda_1}$$

eine konstante Funktion; wäre $\lambda_2\neq 0$, könnten wir entsprechend auf die Konstanz des Kotangens schließen.

Genauso sieht man, daß die Funktionen $\sin^2 x$ und $\cos^2 x$ linear unabhängig sind, denn auch die Quadrate von Tangens und Kotangens

sind nicht konstant. Dagegen sind die drei Funktionen $\sin^2 x$, $\cos^2 x$ und 1 (konstante Funktion) linear abhängig, denn

$$\sin^2 x+\cos^2 x-1=0\quad\text{für alle } x\in\mathbb{R}\,.$$

Elementare Beispiele von Linearkombinationen sind etwa die Zerlegung eines Vektors in seine Komponenten entlang der Achsen eines gegebenen Koordinatensystems, also etwa

$$\lambda\begin{pmatrix}1\\0\\0\end{pmatrix}+\mu\begin{pmatrix}0\\1\\0\end{pmatrix}+\nu\begin{pmatrix}0\\0\\1\end{pmatrix}=\begin{pmatrix}\lambda\\\mu\\\nu\end{pmatrix},$$

oder die „übliche“ Darstellung eines Polynoms durch Potenzen der Variable:

$$f(x)=\lambda_0+\lambda_1 x+\lambda_2 x^2+\lambda_3 x^3\,.$$

Im Vektorraum $V=\mathbb{R}[x]$ aller reeller Polynome in x ist demgemäß das Erzeugnis $[1,x,x^2,x^3]$ der Untervektorraum aller Polynome vom Grad höchstens drei.

Auch für Erzeugnisse unendlicher Mengen gibt es einfache Beispiele in $\mathbb{R}[x]$; beispielsweise ist das Erzeugnis

$$[1,x^2,x^4,x^6,x^8,\dots]$$

der Menge aller gerader Potenzen gleich die Menge aller Polynome, in denen nur gerade x-Potenzen vorkommen, also (wie man sich leicht überlegt) gleich der Menge aller gerader Polynome, d.h. der Polynome $f\in\mathbb{R}[x]$ mit $f(-x)=f(x)$ für alle $x\in\mathbb{R}$.

Da Konstanten und x-Potenzen stetige Funktionen sind, können wir auch im Vektorraum $\mathcal{C}^0(\mathbb{R},\mathbb{R})$ aller stetiger Funktionen das Erzeugnis derselben Menge betrachten, und wieder erhalten wir die Menge aller gerader Polynome. Das mag auf den ersten Blick verwundern, da einige vielleicht erwartet hätten, daß auch die Funktion

$$\cos x=1-\frac{x^2}{2}+\frac{x^4}{24}-\frac{x^6}{720}+-\cdots=\sum_{i=0}^{\infty}(-1)^i\frac{x^{2i}}{(2i)!}$$

in $[1, x^2, x^4, x^6, x^8, \dots]$ liegt, aber dies ist eine *unendliche* Summe, und $[M]$ war ausdrücklich definiert als die Menge aller Linearkombinationen, in denen jeweils nur *endlich* viele Elemente aus M auftreten.

In der Musik (und in der Signalverarbeitung) spielen Linearkombinationen von Sinus- und Kosinusschwingungen eine große Rolle: Der Aufbau eines Tons aus Grundschwingung und Oberschwingungen ist mathematisch betrachtet einfach eine Linearkombination

$$f(t) = \sum_{i=1}^{n} \sin 2\pi i\nu t\,,$$

wobei ν die (Grund-)Frequenz des Tons ist. Bei einem Orchester, das den Kammerton a auf 440 Hz festlegt, sind also alle möglichen Klänge, die dieser Ton auf den verschiedenen Instrumenten annehmen kann, Funktionen aus dem Erzeugnis

$$[\sin 440 \cdot 2\pi t,\ \sin 880 \cdot 2\pi t,\ \sin 1320 \cdot 2\pi t,\ \dots] \leq \mathcal{C}^0(\mathbb{R},\, \mathbb{R})\,.$$

Abbildung zehn zeigt den Ton, den die g-Saite einer Geige produziert zusammen mit der (kaum sichtbaren) Grundschwingung von 196 Hz sowie den ersten acht Oberschwingungen; außerdem ist zum Vergleich gestrichelt eine reine Schwingung der Frequenz 196 Hz eingezeichnet. Wie man sieht, spielen in diesem Beispiel die Oberschwingungen mit der doppelten und der dreifachen Grundfrequenz die größte Rolle.

(Wer selbst Töne aus Grund- und Oberschwingungen synthetisieren möchte, findet unter `http://www.gac.edu/~huber/fourier/` ein Java-Applet, das die entsprechenden Summenkurven zeichnen und die dazugehörigen Töne hörbar machen kann.)

Linearkombinationen sind somit ein einfaches Mittel, um aus relativ wenigen einfachen Funktionen oder Vektoren kompliziertere aufzubauen. Insbesondere bieten sie auch die Möglichkeit, Untervektorräume mit endlichem Aufwand zu beschreiben: Im $\mathbb{R}^n$ etwa ist jeder Untervektorraum mit Ausnahme des Nullraums $\{\vec{0}\}$ eine unendliche Menge, aber wie wir bald sehen werden, läßt sich jeder dieser Untervektorräume als Erzeugnis von endlich vielen Vektoren darstellen.

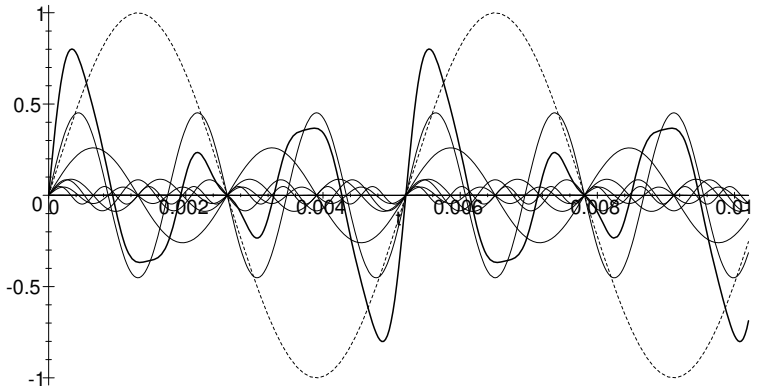


Abb. 10: Ton der g-Saite einer Geige und seine Komponenten

Als ersten Schritt dazu wollen wir uns überlegen, daß die Teilmenge $[M]$ stets ein Untervektorraum ist:

Lemma: Für jede Teilmenge M eines k-Vektorraums V ist $[M] \leq V$ ein Untervektorraum von V; es ist der kleinste Untervektorraum von V, der M enthält.

Beweis: Nach dem Untervektorraumkriterium müssen wir zeigen, daß $[M]$ nicht leer ist und mit je zwei Vektoren $\vec{u}, \vec{v} \in [M]$ und zwei Skalaren $\lambda, \mu \in k$ auch den Vektor $\lambda\vec{u} + \mu\vec{v}$ enthält.

Die erste Eigenschaft ist (fast) trivial: Ist $\vec{v}$ irgendein Vektor aus M, so ist $1\vec{v} = \vec{v}$ eine Linearkombination von $\vec{v}$, liegt also in $[M]$, und insbesondere ist damit $M \subseteq [M]$. Die einzige kleine Schwierigkeit ergibt sich, wenn $M = \emptyset$ die leere Menge ist. Hier müssen wir uns auf die übliche Konvention berufen, daß leere Summen gleich Null sein sollen, eine „Linearkombination“ aus null Vektoren als entsprechend gleich dem Nullvektor, der somit auch im Falle $M = \emptyset$ in $[M]$ liegt.

Nun seien

$$\vec{u} = \lambda_1\vec{u}_1 + \dots + \lambda_n\vec{u}_n \quad \text{und} \quad \vec{v} = \lambda_1\vec{v}_1 + \dots + \lambda_m\vec{v}_m$$

zwei Linearkombinationen von Vektoren aus M. Da wir zu jeder Linearkombination Summanden der Form $0\vec{w}$ hinzufügen können, ohne etwas

an der Summe zu ändern, können wir die beiden Linearkombinationen auch in der Form

$$\vec{u} = \alpha_1\vec{w}_1 + \cdots + \alpha_\ell\vec{w}_\ell \quad \text{und} \quad \vec{v} = \beta_1\vec{w}_1 + \cdots + \beta_\ell\vec{w}_\ell$$

schreiben, wobei

$$\{\vec{w}_1, \ldots, \vec{w}_\ell\} = \{\vec{u}_1, \ldots, \vec{u}_n\} \cup \{\vec{v}_1, \ldots, \vec{v}_m\}$$

ist mit irgendeiner beliebigen Nummerierung der Elemente. Dann ist aber klar, daß auch

$$\lambda\vec{u} + \mu\vec{w} = (\lambda\alpha_1 + \mu\beta_1)\vec{w}_1 + \cdots + (\lambda\alpha_\ell + \mu\beta_\ell)\vec{w}_\ell$$

eine Linearkombination von Vektoren aus M ist und somit in $[M]$ liegt.

Schließlich müssen wir noch zeigen, daß $[M]$ der *kleinste* Untervektorraum von V ist, der M enthält. Wir wissen bereits, daß $[M]$ ein Untervektorraum von V ist, der M enthält; um zu sehen, daß es der kleinste ist, betrachten wir irgendeinen Untervektorraum U von V ist, der M enthält. Dann ist U insbesondere ein Vektorraum, enthält also mit je zwei Vektoren auch deren sämtliche Linearkombinationen. Induktiv folgt, daß er mit jeder endlichen Anzahl von Vektoren auch deren sämtliche Linearkombinationen enthält, also enthält er mit M auch alle Vektoren aus $[M]$. Damit ist $[M] \subseteq U$ für jeden Untervektorraum U, der M enthält, und $[M]$ ist somit in der Tat der kleinste solche Untervektorraum von V. ∎

Vektorräume wurden früher und werden auch gelegentlich noch heute als *lineare Räume* bezeichnet; da $[M]$ somit der kleinste lineare Raum ist, der M enthält, nennt man $[M]$ auch die *lineare Hülle* von M.

Am ökonomischsten ist die Darstellung eines Untervektorraums $U \leq V$ in der Form $U = [M]$ dann, wenn M möglichst wenig Elemente enthält. Wir wollen uns als nächstes überlegen, daß dies höchstens dann der Fall sein kann, wenn M linear unabhängig ist:

Lemma: Falls die Menge $M \subseteq V$ linear abhängig ist, gibt es ein Element $\vec{v} \in M$, das sich als Linearkombination der übrigen, d.h. von Vektoren aus $M \setminus \{\vec{v}\}$, schreiben läßt. Insbesondere ist dann auch

$$[M] = \left[M \setminus \{\vec{v}\}\right] .$$

Beweis: Wenn M linear abhängig ist, gibt es eine nichttriviale Linearkombination von Vektoren $\vec{v}_i \in M$, so daß

$$\lambda_1\vec{v}_1 + \cdots + \lambda_n\vec{v}_n = \vec{0}$$

ist mit Körperelementen λ_i, die nicht alle gleich Null sind. Sei zum Beispiel $\vec{\lambda}_j \neq 0$. Dann kann obige Gleichung nach $\vec{v}_j$ aufgelöst werden; für $1 < j < n$ etwa ist

$$\vec{v}_j = -\frac{\lambda_1}{\lambda_j}\vec{v}_1 - \cdots - \frac{\lambda_{j-1}}{\lambda_j}\vec{v}_{j-1} - \frac{\lambda_{j+1}}{\lambda_j}\vec{v}_{j+1} - \cdots - \frac{\lambda_n}{\lambda_j}\vec{v}_n ,$$

und entsprechend läßt sich $\vec{v}_j$ auch im Falle $j = 1$ oder $j = n$ als Linearkombination der übrigen $\vec{v}_i$ schreiben. ∎

g) Die Dimension eines Vektorraums

Die Dimension eines Vektorraums soll natürlich so definiert werden, daß $\mathbb{R}^n$ die Dimension n hat; wir müssen die Zahl n also irgendwie als Eigenschaft von (Mengen von) Vektoren aus $\mathbb{R}^n$ rekonstruieren.

Offensichtlich kann jeder Vektor aus $\mathbb{R}^n$ als Linearkombination der n Einheitsvektoren geschrieben werden:

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = a_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + a_n \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} .$$

Es fällt aber schwer sich eine Menge aus weniger als n Vektoren vorzustellen, aus der sich ebenfalls *jeder* Vektor aus $\mathbb{R}^n$ als Linearkombination darstellen läßt.

Diese Eigenschaft machen wir uns zunutze, um allgemein Dimensionen zu definieren:

Definition: *a)* Eine Teilmenge $M \subseteq V$ eines k-Vektorraums V heißt *Erzeugendensystem,* wenn $[M] = V$ ist.
b) Wir sagen, der k-Vektorraum V sei *endlichdimensional,* wenn er ein endliches Erzeugendensystem hat; ansonsten bezeichnen wir V als *unendlichdimensional.*

c) Wir sagen, der endlichdimensionale k-Vektorraum V habe die Dimension n, in Zeichen $n = \dim_k V$ oder kurz $n = \dim V$, wenn er ein n-elementiges Erzeugendensystem enthält, aber kein Erzeugendensystem mit weniger als n Elementen.
d) Dem Nullvektorraum $\{\vec{0}\}$ ordnen wir (formal) die Dimension Null zu.

Als Beispiel eines unendlichdimensionalen Vektorraums haben wir den Vektorraum aller reeller Polynome. Hätte dieser nämlich ein endliches Erzeugendensystem, bestehend etwa aus den Polynomen f_1 bis f_n, so ließe sich sich jedes Polynom als Linearkombination

$$f = \lambda_1 f_1 + \cdots + \lambda_n f_n$$

schreiben. Auf diese Weise aber erhält man nur Polynome, deren Grad nicht größer ist als der größte Grad eines f_i. Damit sind auch alle Vektorräume $\mathcal{C}^k\big((a, b), \mathbb{R}\big)$ unendlichdimensional, denn sie enthalten insbesondere alle Polynome.

Endlichdimensional sind natürlich die reellen Vektorräume $\mathbb{R}^n$, denn $\mathbb{R}^n$ wird von seinen n Einheitsvektoren erzeugt. Da wir aber noch nicht sicher wissen, daß es kein Erzeugendensystem mit *weniger* als n Vektoren gibt, können wir im Augenblick nur sagen, daß die Dimension von $\mathbb{R}^n$ *höchstens* n ist.

Für $\mathbb{R}^2$ sieht man leicht, daß sie genau zwei ist: Ansonsten gäbe es nämlich ein Erzeugendensystem aus nur einem Vektor, d.h. alle Vektoren aus $\mathbb{R}^2$ wären proportional zueinander, was natürlich nicht der Fall ist. Für beliebiges n müssen wir ähnlich argumentieren mit linearer Abhängigkeit anstelle von Proportionalität; die Methoden dazu entwickelt der nächste Abschnitt.

h) Basen

Im $\mathbb{R}^3$ läßt sich jeder Vektor

$$\vec{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = a \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + b \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + c \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

auf genau eine Weise als Linearkombination der drei Einheitsvektoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

schreiben; dies entspricht der Tatsache, daß wir im $\mathbb{R}^3$ drei Koordinaten haben.

Mit dem Begriff der *Basis* soll dieser Sachverhalt (soweit möglich) auf beliebige Vektorräume verallgemeinert werden. Da bei der Beschreibung eines Punkts durch seine Koordinaten deren Reihenfolge wesentlich ist, werden wir Basen meist nicht einfach als Mengen auffassen, sondern als geordnete Systeme, im endlichen Fall also als Tupel:

Definition: Ein System $\mathcal{B}$ von Vektoren $\vec{b}_1, \vec{b}_2, \ldots$ eines k-Vektorraums V heißt *Basis* von V, wenn gilt:
1.) Die Menge der $\vec{b}_i$ erzeugt den Vektorraum V, *und*
2.) $\mathcal{B}$ ist linear unabhängig.

Wenn es nicht auf die Reihenfolge ankommt, bezeichnen wir gelegentlich auch die Menge der Basisvektoren als Basis; in diesem Sinne ist also eine Basis einfach ein linear unabhängiges Erzeugendensystem.

Es ist klar, daß die Einheitsvektoren $\vec{e}_1, \vec{e}_2, \vec{e}_3$ eine Basis des $\mathbb{R}^3$ bilden. Ihre wesentliche Eigenschaft der eindeutigen Darstellbarkeit eines jeden Vektors als Linearkombination teilt sie mit jeder anderen Basis:

Lemma: Ist $\mathcal{B}$ eine Basis eines Vektorraums V, so läßt sich jeder Vektor $\vec{v} \in V$ auf genau eine Weise als Linearkombination

$$\vec{v} = \lambda_1 \vec{b}_1 + \cdots + \lambda_r \vec{b}_r$$

von Basisvektoren $\vec{b}_i$ aus $\mathcal{B}$ darstellen.

Beweis: Nach der ersten Eigenschaft aus der Definition einer Basis müssen die Basisvektoren V erzeugen, also läßt sich jeder Vektor $\vec{v} \in V$ als Linearkombination von endlich vielen Elementen aus $\mathcal{B}$ darstellen. Auch wenn wir von zwei solchen Darstellungen ausgehen, ist die Menge

der daran beteiligten Vektoren aus $\mathcal{B}$ noch endlich; wir können also annehmen, daß es r Vektoren $\vec{b}_1, \ldots, \vec{b}_r$ gibt, so daß

$$\begin{aligned}\vec{v} &= \lambda_1\vec{b}_1 + \cdots + \lambda_r\vec{b}_r \\ &= \mu_1\vec{b}_1 + \cdots + \mu_r\vec{b}_r\end{aligned}$$

ist, wobei wir einfach λ_i oder μ_i gleich Null setzen, wenn $\vec{b}_i$ in der entsprechenden Darstellung nicht vorkommt. Subtrahieren wir die beiden Darstellungen voneinander, erhalten wir eine Darstellung des Nullvektors als Linearkombination

$$\vec{0} = (\lambda_1 - \mu_1)\vec{b}_1 + \cdots + (\lambda_r - \mu_r)\vec{b}_r$$

von Basisvektoren. Da diese nach der zweiten definierenden Eigenschaft einer Basis linear unabhängig sind, müssen alle Koeffizienten $\lambda_i - \mu_i$ verschwinden. Damit sind die beiden betrachteten Darstellungen von $\vec{v}$ als Linearkombination der Vektoren aus $\mathcal{B}$ gleich, mit anderen Worten: Es gibt genau eine solche Darstellung. ∎

Lemma: Ein Erzeugendensystem eines k-Vektorraum V ist genau dann eine Basis, wenn es minimal ist.

Beweis: Das Erzeugendensystem $\mathcal{B}$ sei eine Basis. Um zu zeigen, daß es minimal ist, müssen wir uns überlegen, daß jeder Basisvektor $\vec{v}$ aus $\mathcal{B}$ wirklich notwendig ist, daß also $\mathcal{B}$ ohne diesen Vektor $\vec{v}$ kein Erzeugendensystem mehr ist.

Falls es eines wäre, könnte insbesondere der Vektor $\vec{v}$ als Linearkombination der restlichen Vektoren aus $\mathcal{B}$ geschrieben werden. Gleichzeitig hat er aber die Darstellung $\vec{v} = \vec{v}$, deren rechte Seite man auch als Linearkombination von Elementen aus $\mathcal{B}$ auffassen kann. Somit ist seine Basisdarstellung nicht eindeutig, im Widerspruch zum gerade bewiesenen Lemma. Daher muß $\mathcal{B}$ minimal sein.

Umgekehrt sei $\mathcal{B}$ ein minimales Erzeugendensystem. Um zu zeigen, daß es eine Basis ist, reicht der Nachweis der linearen Unabhängigkeit von $\mathcal{B}$.

Sei also $\lambda_1\vec{b}_1 + \cdots + \lambda_n\vec{b}_n = \vec{0}$ eine Darstellung des Nullvektors als Linearkombination von Elementen aus $\mathcal{B}$. Falls darin einer der Koeffizienten λ_i nicht verschwindet, läßt sich der zugehörige Vektor $\vec{v}_i$ als Linearkombination der restlichen $\vec{v}_j$ schreiben. Dann reicht aber bereits $\mathcal{B}$ ohne $\vec{v}_i$ zur Erzeugung aus, $\mathcal{B}$ ist also nicht minimal. Somit müssen alle λ_i verschwinden, $\mathcal{B}$ ist also linear unabhängig und damit eine Basis. ∎

Lemma: Eine System von linear unabhängigen Elementen eines Vektorraums ist genau dann eine Basis, wenn es maximal ist.

Beweis: $\mathcal{B}$ sei eine Basis. Dann läßt sich jeder Vektor $\vec{v} \in V$ als Linearkombination der Elemente von $\mathcal{B}$ schreiben, nimmt man also $\vec{v}$ zu $\mathcal{B}$ hinzu, ist das System nicht mehr linear unabhängig.

Umgekehrt sei $\mathcal{B}$ maximal linear unabhängig, und $\vec{v}$ sei ein beliebiger Vektor; wir müssen zeigen, daß er als Linearkombination der Vektoren aus $\mathcal{B}$ darstellbar ist. Das ist trivial, falls $\vec{v}$ bereits zu $\mathcal{B}$ gehört.

Andernfalls ist $\mathcal{B}$ zusammen mit $\vec{v}$ linear abhängig, da $\mathcal{B}$ ja als *maximal* linear unabhängig vorausgesetzt war. Somit gibt es ein nichttriviale Linearkombination

$$\lambda\vec{v} + \lambda_1\vec{b}_1 + \cdot + \lambda_n\vec{b}_n = \vec{0}$$

mit Vektoren $\vec{b}_i \in \mathcal{B}$. Darin muß $\lambda \neq 0$ sein, denn sonst wäre $\mathcal{B}$ linear abhängig. Also liegt

$$\vec{v} = -\frac{\lambda_1}{\lambda}\vec{b}_1 + \cdots + -\frac{\lambda_n}{\lambda}\vec{b}_n$$

in $[\mathcal{B}]$, und $\mathcal{B}$ ist ein Erzeugendensystem. ∎

Basen lassen sich somit auch charakterisieren als minimale Erzeugendensysteme oder als maximale Systeme linear unabhängiger Vektoren.

Als nächstes stellt sich die Frage, wann es Basen gibt. Glücklicherweise hat *jeder* Vektorraum eine Basis; der Beweis ist allerdings für unendlichdimensionale Vektorräume logisch nicht ganz einfach. Für diese Vorlesung wollen wir uns daher mit einem Beweis für endlichdimensionale

Vektorräume begnügen. Wir beweisen dazu den etwas allgemeineren, tatsächlich ebenfalls für beliebige Vektorräume gültigen

Basisergänzungssatz: $M \subset V$ sei eine linear unabhängige Teilmenge des endlichdimensionalen Vektorraums V. Dann gibt es eine Basis $\mathcal{B}$ von V, die M enthält.

Beweis: Da V nach Voraussetzung endlichdimensional ist, gibt es zunächst einmal überhaupt eine endliche Menge $E \subset V$, die V erzeugt. Falls E einen oder mehrere der Vektoren aus M enthält, entfernen wir diese; was übrigbleibt, sei die Menge N, d.h. $N = E \setminus M$.

Damit sind M und N zwei disjunkte Teilmengen von V, deren Vereinigung die Menge E enthält und somit insbesondere ein Erzeugendensystem von V ist.

Konkret sei $M = \{\vec{b}_1, \ldots, \vec{b}_r\}$ und $N = \{\vec{v}_1, \ldots, \vec{v}_s\}$; dann wird V also erzeugt von

$$M \cup N = \{\vec{b}_1, \ldots, \vec{b}_r, \vec{v}_1, \ldots, \vec{v}_s\} .$$

Wir beweisen die Behauptung durch Induktion nach der Elementanzahl s von N.

Für $s = 0$ bilden die Elemente von M, in irgendeiner Weise angeordnet, bereits eine Basis, und wir sind fertig.

Für $s > 0$ sind wir fertig, falls $M \cup N$ linear unabhängig ist; denn dann ist $M \cup N$ eine Basis von V, die M enthält.

Andernfalls gibt es Elemente $\lambda_1, \ldots, \lambda_r$ und $\mu_1, \ldots, \mu_s$, die nicht alle gleichzeitig verschwinden, so daß

$$\lambda_1\vec{b}_1 + \cdots + \lambda_r\vec{b}_r + \mu_1\vec{v}_1 + \cdots + \mu_s\vec{v}_s = \vec{0}$$

ist. In dieser Gleichung können nicht alle μ_i verschwinden, denn sonst wären die $\vec{b}_i$ linear abhängig, im Widerspruch zur Voraussetzung. Also gibt es (mindestens) ein $\mu_i \neq 0$, und der zugehörige Vektor $\vec{v}_i$ läßt sich als Linearkombination der restlichen $\vec{v}_j$ und der $\vec{b}_\ell$ ausdrücken.

Damit wird $V = [M \cup (N \setminus \{\vec{v}_i\})]$ auch von der um $\vec{v}_i$ verminderten Menge erzeugt, und wir haben nur noch $s - 1$ Vektoren $\vec{v}_j$. Daher gibt es nach Induktionsannahme eine Basis von V, die M enthält. ∎

Korollar: Jeder endlichdimensionale Vektorraum $V \neq \{\vec{0}\}$ hat eine Basis.

Beweis: Man wende den obigen Satz an auf eine Menge M, die aus einem einzigen Vektor $\vec{v} \neq \vec{0}$ besteht. ∎

Um die Sonderrolle des Nullvektorraums zu eliminieren, vereinbaren wir, daß er die leere Menge als Basis haben soll; dies ist kompatibel mit der üblichen Interpretation von leeren Summen und leeren Aussagen.

Wie bereits erwähnt, gelten sowohl der Basisergänzungssatz als auch das obige Korollar für beliebige Vektorräume, d.h. also auch im Falle unendlicher Dimension. Für interessierte Leser sei kurz erwähnt, wie man hier vorgeht. Das wesentliche neue Hilfsmittel ist das ZORNsche Lemma, benannt nach dem deutschen Mathematiker MAX ZORN (1906–1993), der es, nachdem er Deutschland wegen der nationalsozialistischen Politik verlassen mußte, um 1935 an der amerikanischen Yale Universität bewies. Es besagt folgendes:

Gegeben sei eine nichtleere partiell geordnete Menge $\mathcal{M}$, d.h. für manche Paare von Elementen $A, B \in \mathcal{M}$ ist eine Relation $A < B$ erklärt mit der Eigenschaft, daß mit $A < B$ und $B < C$ auch $A < C$ gilt, wohingegen nie $A < A$ ist. Diese partiell geordnete Menge habe die zusätzliche Eigenschaft, daß es zu jeder Kette

$$A_1 < A_2 < A_3 < \cdots$$

von Elementen aus $\mathcal{M}$ ein Element A_∞ gebe mit der Eigenschaft, daß $A_i < A_\infty$ für alle i. Dann gibt es in $\mathcal{M}$ ein *maximales* Element, d.h. ein Element B, zu dem es kein $C \in \mathcal{M}$ gibt mit $B < C$.

Dieses Lemma kann nicht aus den üblichen Axiomen der Mengenlehre hergeleitet werden, sondern ist äquivalent zum sogenannten *Auswahlaxiom.* Für dieses bewies um 1940 der österreichische Mathematiker KURT GÖDEL (1906–1978), seit 1940 im amerikanischen Exil in Princeton, daß sowohl dieses Axiom als auch seine Negation mit den restlichen Axiomen der Mengenlehre kompatibel ist; das gleiche gilt demnach auch für das ZORNsche Lemma. Man kann daher wählen, ob man eine Mathematik mit oder ohne ZORNsches Lemma bevorzugt. Die meisten Mathematiker haben sich für „mit“ entschieden, es gibt aber auch welche, die das ZORNsche Lemma ablehnen.

Aus dem ZORNschen Lemma folgt der Basisergänzungssatz recht einfach: Als Menge $\mathcal{M}$ nehmen wir die Menge aller linear unabhängiger Teilmengen $A \subset V$, die M enthalten; die partielle Ordnungsrelation sei die gewöhnliche (echte) Teilmengenbeziehung. Die Kettenbedingung des ZORNschen Lemmas ist offensichtlich erfüllt, denn für eine Kette

$$M \subset A_1 \subset A_2 \subset A_3 \subset \cdots$$

aus linear unabhängigen Mengen A_i, die M enthalten, ist auch

$$A_\infty = \bigcup_{i \geq 1} A_i$$

eine linear unabhängige Teilmenge von V, die M enthält, da jede endliche Menge von Vektoren aus A_∞ bereits in einer der Mengen A_m liegt. Also gibt es nach dem ZORNschen Lemma ein maximales Element $\mathcal{B} \in \mathcal{M}$. Diese Menge $\mathcal{B}$ ist linear unabhängig, da sie in $\mathcal{M}$ liegt, und sie ist eine Basis, denn gäbe es einen Vektor $\vec{v} \notin [\mathcal{B}]$, so wäre auch die Menge $\mathcal{C} = \mathcal{B} \cup \{\vec{v}\}$ linear unabhängig, im Widerspruch zur Maximalität von $\mathcal{B}$. Damit ist der Basisergänzungssatz bewiesen, und das Korollar folgt wie oben.

Um wenigstens anhand eines Beispiels zu sehen, daß auch unendlichdimensionale Vektorräume Basen haben, betrachten wir den Vektorraum V aller Polynome mit reellen Koeffizienten. Da sich ein Polynom P vom Grad d als

$$P = a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d$$

schreiben läßt, erzeugt das System $\mathcal{B}$ der x-Potenzen $1, x, x^2, x^3, \ldots$ diesen Vektorraum. Jede Linearkombination des Nullvektors, des Polynoms $P \equiv 0$ also, aus Elementen von $\mathcal{B}$ wäre ein Polynom

$$\lambda_0 + \lambda_1 x + \cdots + \lambda_n x^n \,,$$

dessen Koeffizienten zumindest teilweise von Null verschieden sind, während es selbst identisch Null wäre. Da es kein solches Polynom gibt, ist $\mathcal{B}$ linear unabhängig, also eine Basis von V.

Die Schwierigkeiten, die bei unendlichdimensionalen Vektorräumen auftreten können, sieht man, wenn man in diesem Beispiel die Polynome durch Potenzreihen (egal ob formal oder konvergent) ersetzt: Da Potenzreihen *unendliche* Summen sind, während bei Linearkombinationen nur *endliche* Summen erlaubt sind, bilden nun die x-Potenzen kein Erzeugendensystem mehr. Nach dem Basisergänzungssatz, der, auch wenn wir das nicht bewiesen haben, auch für unendlichdimensionale Vektorräume gilt, gibt es eine Menge von Potenzreihen, die zusammen mit der obigen Menge $\mathcal{B}$ eine Basis bilden; explizit angeben konnte diese Menge aber noch niemand, genauso wenig wie eine explizite Basis für einen der Räume $\mathcal{C}^n(\mathbb{R}, \mathbb{R})$.

Kehren wir also zurück zum überschaubareren endlichdimensionalen Fall, und beweisen wir dort zunächst die anschaulich fast selbstverständliche Aussage, daß jede Basis eines n-dimensionalen Vektorraums aus n Vektoren besteht. Dazu benötigen wir eine leichte Verschärfung

des Basisergänzungssatzes; er geht zurück auf ERNST STEINITZ, den wir bereits von der Körperdefinition aus §1*b)* kennen:

Austauschsatz von STEINITZ: M sei eine endliche linear unabhängige Teilmenge des endlichdimensionalen Vektorraums V, und $\mathcal{B}$ sei eine Basis von V. Dann gibt es eine Teilmenge $\mathcal{B}'$ von $\mathcal{B}$, so daß $M \cup \mathcal{B}'$ eine Basis von V ist. Diese hat genauso viele Elemente wie $\mathcal{B}$.

Mit anderen Worten: Man kann Vektoren aus $\mathcal{B}$ finden, die sich Stück für Stück gegen die Vektoren aus M *austauschen* lassen.

Der *Beweis* ist dem des Basisergänzungssatzes sehr ähnlich; mit Rücksicht auf die Anzahlaussage führen wir ihn aber durch Induktion nach der Elementanzahl m von M.

Für $m = 0$ ist $M = \emptyset$ und wir setzen einfach $\mathcal{B}' = \mathcal{B}$.

Für $m \geq 1$ entfernen wir einen Vektor $\vec{v}$ aus M und wenden den Satz auf die Menge $M' = M \setminus \{\vec{v}\}$ an. Für diese gilt er nach Induktionsannahme, es gibt also eine Teilmenge $\mathcal{C}'$ von $\mathcal{B}$, so daß $\mathcal{C} = M' \cup \mathcal{C}'$ eine Basis von V ist mit gleicher Elementanzahl wie $\mathcal{B}$. Bezüglich dieser Basis habe $\vec{v}$ die Darstellung

$$\vec{v} = \lambda_1 \vec{v}_1 + \cdots + \lambda_{m-1} \vec{v}_{m-1} + \mu_1 \vec{c}_1 + \cdots + \mu_r \vec{c}_r \,,$$

wobei $M' = \{\vec{v}_1, \ldots, \vec{v}_{m-1}\}$ und $\mathcal{C}' = \{\vec{c}_1, \ldots, \vec{c}_r\}$ sein soll.

Da $M = M' \cup \{\vec{v}\}$ linear unabhängig ist, muß in dieser Darstellung mindestens ein $\vec{\mu}_i$ von Null verschieden sein. Daher läßt sich der zugehörige Vektor $\vec{c}_i$ als Linearkombination aus den restlichen $\vec{c}_j$, den $\vec{v}_\ell$ und dem Vektor $\vec{v}$ schreiben, d.h. auch die durch den *Austausch* von $\vec{c}_i$ durch $\vec{v}$ entstehende Menge

$$M' \cup \left(\mathcal{C}' \setminus \{\vec{c}_i\}\right) \cup \{\vec{v}\} = M \cup \left(\mathcal{C}' \setminus \{\vec{c}_i\}\right)$$

erzeugt ganz V. Diese Menge ist auch linear unabhängig und somit eine Basis, denn ist

$$\alpha \vec{v} + \sum_{\ell=1}^{m-1} \alpha_\ell \vec{v}_\ell + \sum_{\substack{j=1 \\ j \neq i}}^{n} \beta_j \vec{c}_j = \vec{0} \,,$$

so muß zunächst α verschwinden, da $\vec{v}$ sonst als Linearkombination der $\vec{v} \in M'$ und der $\vec{c}_j$ mit $j \neq i$ dargestellt werden könnte, was wir oben durch die Wahl eines i mit $\mu_i \neq 0$ ausgeschlossen haben. Also steht hier nur eine Linearkombination von Elementen einer Basis, so daß alle α_ℓ und β_j verschwinden müssen. Mit

$$\mathcal{B}' = \left(\mathcal{C}' \setminus \{\vec{c}_i\}\right)$$

ist somit die Behauptung des Satzes erfüllt. ∎

Aus dem STEINITZschen Austauschsatz folgt

Satz: *a)* Jede Basis $\mathcal{B}$ eines n-dimensionalen Vektorraums V besteht aus n Vektoren.
b) Jede Teilmenge von V mit mehr als n Elementen ist linear abhängig.
c) Keine Teilmenge von V mit weniger als n Elementen ist ein Erzeugendensystem.

Beweis: a) Da V die Dimension n hat, gibt es ein Erzeugendensystem $M = \{\vec{v}_1, \ldots, \vec{v}_n\}$ mit n-Elementen, aber keines mit weniger Elementen. Also ist M ein minimales Erzeugendensystem und somit eine Basis.

Nun sei $\mathcal{B} = \{\vec{b}_1, \ldots, \vec{b}_m\}$ irgendeine andere Basis von V. Nach dem Austauschsatz läßt sich M zu einer Basis von V ergänzen, die genauso viele Elemente hat wie $\mathcal{B}$. Da es keine Basis geben kann, die M echt enthält, muß M genauso viele Elemente enthalten wie $\mathcal{B}$, also n.

b) Jede linear unabhängige Teilmenge läßt sich zu einer Basis ergänzen, und jede Basis besteht aus n Vektoren. Also kann eine linear unabhängige Menge höchstens n Vektoren enthalten.

c) Das ist die Definition der Dimension. ∎

Nach diesem Satz läßt sich die Dimension eines Vektorraums einfach dadurch bestimmen, daß man eine Basis findet und deren Elemente zählt. Insbesondere hat $\mathbb{R}^n$ als $\mathbb{R}$-Vektorraum die Dimension n, da die n Einheitsvektoren eine Basis bilden.

Weniger offensichtlich ist, daß $\mathbb{R}$ als $\mathbb{Q}$-Vektorraum unendlichdimensional ist: Dazu betrachten wir die unendliche Menge M aller Logarithmen

$\ln p$ der Primzahlen. Wäre diese Menge linear abhängig, gäbe es eine nichttriviale Linearkombination

$$\lambda_1 \ln p_1 + \cdots + \lambda_r \ln p_r = 0$$

mit $\lambda_i \in \mathbb{Q}$. Multipliziert man diese Gleichung mit dem Hauptnenner der λ_i, so erhält man eine entsprechende Gleichung mit Koeffizienten $\mu_i \in \mathbb{Z}$. Dann ist

$$\mu_1 \ln p_1 + \cdots + \mu_r \ln p_r = \ln(p_1^{\mu_1} \cdot \cdots \cdot p_r^{\mu_r}) = 0$$

gleichbedeutend mit

$$p_1^{\mu_1} \cdot \cdots \cdot p_r^{\mu_r} = 1\,,$$

was wegen der Eindeutigkeit der Primzerlegung in $\mathbb{Z}$ nur gelten kann, wenn alle μ_i und damit auch alle λ_i verschwinden.

Also ist $\mathbb{R}$ als $\mathbb{Q}$-Vektorraum unendlichdimensional, und dies erklärt, warum Computer so große Schwierigkeiten mit reellen Zahlen haben: Exakt rechnen kann ein Computer nur in Teilmengen von $\mathbb{R}$, die endlichdimensionale $\mathbb{Q}$-Vektorräume sind – und selbst da gibt es zumindest theoretisch noch das Problem der potentiell beliebig großen Zähler und Nenner.

i) Dimensionen und lineare Abbildungen

Als nächstes wollen wir uns mit Dimensionen von Untervektorräumen, insbesondere auch Kernen und Bildern beschäftigen. Anschaulich klar und auch recht einfach zu beweisen ist der folgende

Satz: Für einen echten Untervektorraum $U < V$ eines endlichdimensionalen Vektorraums V ist $\dim U < \dim V$.

Beweis: Eine Basis von U ist auch in V linear unabhängig, läßt sich also ergänzen zu einer Basis von V. Da die Dimension eines Vektorraums gleich der Elementanzahl einer beliebigen Basis ist, folgt sofort, daß $\dim U \leq \dim V$ sein muß, und wenn beide gleich sind, ist $U = V$. ∎

Hier haben wir ganz wesentlich benutzt, daß V endlichdimensional ist; in einen unendlichdimensionalen Vektorraum gibt es stets Untervektorräume, die ebenfalls unendlichdimensinal sind, im Vektorraum V

aller reeller Polynome in x beispielsweise den Untervektorraum aller Polynome in x^2.

Satz: Für endlichdimensionale Vektorräume V, W und eine lineare Abbildung $\varphi\colon V \to W$ ist $\dim \text{Bild}\,\varphi = \dim V - \dim \text{Kern}\,\varphi$.

Beweis: $\vec{b}_1, \ldots, \vec{b}_r$ sei eine Basis von Kern φ; falls φ injektiv ist, setzen wir $r = 0$. Nach dem Basisergänzungssatz oder (falls $r = 0$) wegen der Existenz von Basen lassen sich dann $n - r$ Vektoren $\vec{b}_{r+1}, \ldots, \vec{b}_n$ finden mit $n = \dim V$, so daß $\vec{b}_1, \ldots, \vec{b}_n$ eine Basis von V ist.

Das Bild eines beliebigen Vektors $\vec{v} = \lambda_1\vec{b}_1 + \cdots + \lambda_n\vec{b}_n$ ist dann

$$\varphi(\vec{v}) = \lambda_1\varphi(\vec{b}_1) + \cdots + \lambda_n\varphi(\vec{b}_n) = \lambda_{r+1}\varphi(\vec{b}_{r+1}) + \cdots + \lambda_n\varphi(\vec{b}_n)\,,$$

da $\vec{b}_1 \ldots, \vec{b}_r$ ja auf den Nullvektor abgebildet werden. Also wird Bild φ von den Vektoren $\varphi(\vec{b}_{r+1}), \ldots, \varphi(\vec{b}_n)$ erzeugt.

Diese Vektoren sind auch linear unabhängig in W, denn ist

$$\lambda_{r+1}\varphi(\vec{b}_{r+1}) + \cdots + \lambda_n\varphi(\vec{b}_n) = \varphi(\lambda_{r+1}\vec{b}_{r+1} + \cdots + \lambda_n\vec{b}_n) = \vec{0}\,,$$

so liegt $\lambda_{r+1}\vec{b}_{r+1} + \cdots + \lambda_n\vec{b}_n$ im Kern von φ.

Im Fall einer injektiven Abbildung ist $\lambda_{r+1}\vec{b}_{r+1} + \cdots + \lambda_n\vec{b}_n$ daher gleich dem Nullvektor, und damit müssen alle $\lambda_i = 0$ sein, denn die $\vec{b}_i$ sind als Basisvektoren insbesondere linear unabhängig.

Falls φ nicht injektiv ist, können wir nur sagen, daß der Vektor $\lambda_{r+1}\vec{b}_{r+1} + \cdots + \lambda_n\vec{b}_n$ im Kern von φ liegt; er ist also darstellbar als Linearkombination der Basisvektoren $\vec{b}_1, \ldots, \vec{b}_r$ des Kerns:

$$\lambda_{r+1}\vec{b}_{r+1} + \cdots + \lambda_n\vec{b}_n = \lambda_1\vec{b}_1 + \cdots + \lambda_r\vec{b}_r\,.$$

Auch daraus folgt wegen der linearen Unabhängigkeit der $\vec{b}_i$, daß alle λ_i Null sein müssen.

Damit ist $\{\varphi(\vec{b}_{r+1}), \ldots, \varphi(\vec{b}_n)\}$ eine Basis von Bild φ, d.h.

$$\dim \text{Bild}\,\varphi = n - r = \dim V - \dim \text{Kern}\,\varphi\,,$$

wie behauptet. ∎

Wir werden diese Aussage im folgenden als *Dimensionsformel* bezeichnen. Da wir hier mit Dimensionen rechnen, ist klar, daß sie nicht auf unendlichdimensionale Vektorräume verallgemeinert werden kann: Sind etwa sowohl V als auch Kern φ unendlichdimensional, kann Bild φ jede beliebige Dimension haben – einschließlich null und unendlich. Der sogenannte *Homomorphiesatz* macht eine genauere Aussage über Bild φ, die auch für unendlichdimensionale Vektorräume gilt. Da wir die zu seiner Formulierung benötigten Begriffe nur teilweise kennen und für diese Vorlesung auch nicht brauchen, sei auf Einzelheiten verzichtet.

Korollar: Eine lineare Selbstabbildung $\varphi\colon V \to V$ eines endlichdimensionalen Vektorraums V ist genau dann injektiv, wenn sie surjektiv ist.

Beweis: φ ist genau dann injektiv, wenn $\dim \text{Kern}\,\varphi = 0$ ist und genau dann surjektiv, wenn $\dim \text{Bild}\,\varphi = \dim V$ ist. Diese beiden Dimensionsaussagen sind nach dem gerade bewiesenen Satz äquivalent. ∎

Man beachte, daß es in diesem Korollar sehr wesentlich ist, daß wir von einem *endlichdimensionalen* Vektorraum ausgehen: Für den Vektorraum V aller reeller Polynome ist die Abbildung

$$\varphi\colon V \to V; \quad \sum_{i=0}^{d} a_i x^i \mapsto \sum_{i=0}^{d} a_i x^{2i}$$

linear *(warum?)* und injektiv, aber nicht surjektiv. Umgekehrt ist die Ableitung

$$\psi\colon V \to V; \quad f \mapsto f'$$

linear und surjektiv, aber nicht injektiv.

§3: Vektorräume und endliche Körper

Bislang hatten wir in fast allen Beispielen nur Vektorräume über dem Körper der reellen Zahlen betrachtet; in der Informationsverarbeitung treten aber oftmals auch Probleme auf, für die Vektorräume über endlichen Körpern nützlich sind. Als einfachstes Beispiel kennen wir bereits aus §1*e)* den Körper $\mathbb{F}_2 = \{0, 1\}$ der Bits; erstes Thema dieses Paragraphen sind Vektorräume über $\mathbb{F}_2$.

a) Bitfolgen als Vektoren

Mit einem einzigen Bit läßt sich nicht viel Information darstellen und verarbeiten; interessant wird es erst mit Bitfolgen. Natürlich können wir Folgen von N Bits als Elemente des Vektorraums $\mathbb{F}_2^N$ betrachten. Da im Körper $\mathbb{F}_2$ die Summen $0+0$ und $1+1$ beide gleich 0 sind, hat dieser Vektorraum die Eigenschaft

$$\vec{v} + \vec{v} = \vec{0} \quad \text{für alle } \vec{v} \in \mathbb{F}_2^N \,,$$

jeder Vektor ist also zu sich selbst invers, und genau wie auch in $\mathbb{F}_2$ gibt es keinen Unterschied zwischen plus und minus.

Der Vektorraum $\mathbb{F}_2^N$ hat eine sehr einfache Struktur: Die Vektoraddition ist in jeder Komponente einfach die logische Antivalenz, und bitweise logische Antivalenz für ganze Wörter gehört zu den Grundbefehlen der meisten Prozessoren und auch Programmiersprachen. Bei einer Maschine mit 32 Bit-Prozessor läßt sich also eine Vektoraddition in $\mathbb{F}_2^{32}$ mit einem einzigen Befehl ausführen; in C oder C++ wäre der entsprechende Ausdruck gleich `a^b`.

Noch einfacher ist die Multiplikation mit einem Skalar, denn es gibt nur zwei Skalare: Multiplikation mit Eins ändert nichts, Multiplikation mit Null hat immer die Bitfolge aus lauter Nullen als Ergebnis.

Das Rechnen in $\mathbb{F}_2^N$ ist also sehr einfach und effizient, und es kann schon in dieser ganz trivialen Form auch nützlich sein:

Eine Anwendung ist etwa die Fehlererkennung in der Informationsübertragung: Dazu werden Daten beispielsweise oft zusammen mit einem „Paritätsbit“ übertragen, d.h. jede Folge von sieben Bits wird um ein achtes „Prüfbit“ erweitert, so daß im entstehenden Byte immer eine gerade Anzahl von Einsen vorkommt; es hat also gerade Parität. Vor der Übertragung wird also auf jede Folge von sieben Bit die lineare Abbildung

$$\varphi\colon \begin{cases} \mathbb{F}_2^7 \to \mathbb{F}_2^8 \\ (x_1, \ldots, x_7) \mapsto (x_1, \ldots, x_7,\ x_1 + \cdots + x_7) \end{cases}$$

angewendet. Auch die Überprüfung, ob ein gegebenes Byte tatsächlich gerade Parität hat, läßt sich mit einer linearen Abbildung realisieren:

Die Bytes mit gerader Parität sind offenbar gerade die aus dem Kern der linearen Abbildung

$$\psi\colon \begin{cases} \mathbb{F}_2^8 \to \mathbb{F}_2 \\ (x_1, \ldots, x_8) \mapsto (x_1 + \cdots + x_8) \end{cases}$$

Mit etwas mehr Aufwand kann man Fehler nicht nur erkennen, sondern auch korrigieren: Als Beispiel dafür konstruieren wir eine Abbildung

$$\varphi\colon \mathbb{F}_2^{nm} \to \mathbb{F}_2^{(n+1)(m+1)}$$

wie folgt: Wir schreiben die Elemente von $\mathbb{F}_2^{nm}$ in der Form

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{pmatrix}$$

und bilden ein solches Element ab auf

$$\varphi(X) = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} & x_{1,n+1} \\ x_{21} & x_{22} & \ldots & x_{2n} & x_{2,n+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} & x_{m,n+1} \\ x_{m+1,1} & x_{m+1,2} & \ldots & x_{m+1,n} & x_{m+1,n+1} \end{pmatrix},$$

wobei

$$x_{i,n+1} = \sum_{j=1}^{n} x_{ij} \quad \text{und} \quad x_{m+1,j} = \sum_{i=1}^{m} x_{ij}$$

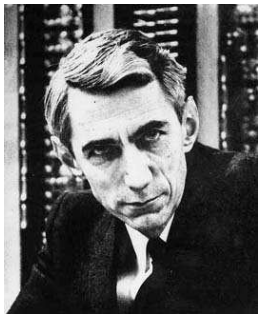
sein soll. Es braucht uns dabei nicht stören, daß $x_{n+1,m+1}$ hier auf zwei verschiedene Weisen definiert ist: Wie man sich leicht überlegt, führen beide Definitionen ausgeschrieben zu

$$x_{n+1,m+1} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \,.$$

Hier gibt es also $n+m+1$ Prüfbits; in $\varphi(X)$ sind alle Zeilensummen und alle Spaltensummen Null. Falls nun durch einen Übertragungsfehler das Bit x_{ij} (und sonst keines) verfälscht wurde, ist genau in der i-ten Zeile

und der j-ten Spalte die entsprechende Summe gleich eins, es ist also klar, daß x_{ij} korrigiert werden muß.

Mit entsprechend größerem Aufwand lassen sich auch mehr Fehler korrigieren; tatsächlich können nach zwei Sätzen von SHANNON, wenn man nur genügend lange Codewörter zuläßt, mit beliebig geringem (relativem) Aufwand beliebig hohe (vorgegebene) Fehlerraten korrigiert werden – vorausgesetzt natürlich, diese Raten sind echt kleiner als $1/2$. Bei einer Fehlerrate von $1/2$ kommen nur Zufallsbits ohne jeglichen Informationsgehalt an.



CLAUDE ELWOOD SHANNON (1916–2001) wurde in Petoskey im US-Bundesstaat Michigan geboren; 1936 verließ er die University of Michigan mit sowohl einem Bachelor der Mathematik als auch einem Bachelor der Elektrotechnik, um am M.I.T. weiterzustudieren. Seine 1938 geschriebene Diplomarbeit *A symbolic analysis of relay and switching circuits* bildet die Grundlage der digitalen Informationsverarbeitung auf der Grundlage der hier entwickelten Schaltlogik; seine Dissertation 1940 befaßte sich mit Anwendungen der Algebra auf die MENDELSchen Gesetze. Danach arbeitete er bis 1956 bei den Bell Labs, wo er während des zweiten Weltkriegs insbesondere über die Sicherheit kryptographischer Systeme forschte. Seine *Mathematical theory of cryptography* wurde aus Geheimhaltungsgründen erst 1949 zur Veröffentlichung freigegeben. Seine wohl bekannteste Arbeit ist die 1948 erschienene *Mathematical theory of communication,* in der er die fehlerfreie Übertragung von Nachrichten über einen gestörten Kanal untersuchte. Von 1956 bis zu seiner Emeritierung 1978 lehrte er am M.I.T., das er dadurch zur führenden Universität auf dem Gebiet der Informationstheorie und Kommunikationstechnik machte. Zu seinen zahlreichen Arbeiten zählt auch eine über die mathematische Theorie der Jongliermuster, anhand derer Jongleure eine Reihe neuer Muster gefunden haben; auch konstruierte er mehrere Jonglierroboter.

Beim nächsten Beispiel geht es um die Sicherung von Information gegen *absichtliche* Manipulation und unberechtigtes Mithören:

Während des kalten Kriegs hielten viele (wohl zu Recht) die Gefahr eines Atomkriegs aus Versehen für erheblich größer als die eines absichtlichen Atomkriegs. Um ersteren weniger wahrscheinlich zu machen, einigten sich die beiden Großmächte im Juni 1963 in Genf darauf, das sogenannte *Rote Telephon* einzurichten; es funktioniert seit dem 30. August 1963.

Natürlich handelt es sich dabei nicht wirklich um ein Telephon, denn zu keinem Zeitpunkt des kalten Krieges reichten die Sprachkenntnisse eines amerikanischen Präsidenten oder eines Generalsekretärs der KPdSU auch nur für ein direktes Gespräch über das Wetter.

Tatsächlich war das *Rote Telephon* eine Fernschreibverbindung mit je vier Fernschreibern (geliefert von Siemens Mannheim) an beiden Enden: jeweils zwei mit lateinischem und zwei mit kyrillischem Alphabet. Bislang verbrachten sie ihre meiste Zeit damit, stündliche Testnachrichten zu drucken wie amerikanische Baseball-Ergebnisse oder TURGENJEWs *Aufzeichnungen einer Jägers.*

Aus Sicherheitsgründen wurden zwei Leitungen eingerichtet, eine entlang der Route Washington-London-Kopenhagen-Stockholm-Helsinki-Moskau, die andere via Tanger. Natürlich war es unmöglich, diese Leitungen auf ihrer ganzen Länge zu überwachen, so daß niemand ausschließen konnte, daß irgendwo zwischen Moskau und Washington eine vertrauliche Kommunikation abgehört oder – schlimmer noch – eine gefälschte Nachricht eingespielt wurde.

Zum Schutz davor wurde die gesamte Kommunikation verschlüsselt. Wegen der hohen Sicherheitsanforderungen konnte dazu allerdings keines der üblicherweise in heutiger Office-Software eingebauten Verfahren verwendet werden: Wer noch irgendwelche Illusionen über die Sicherheit gängiger kommerzieller Programme hat, sollte unter

`http://pwcrack.com`

nachlesen, für welche vergleichsweise bescheidenen Beträge spezialisierte Unternehmen dazu bereit sind, „vergessene“ Paßwörter zu rekonstruieren.

Das *Rote Telephon* benutzte stattdessen eine Variante eines alten, absolut sicheren, Verschlüsselungsverfahrens, des sogenannten *one time pads:* Von Zeit zu Zeit tauschten die beiden Seiten per Kurier Magnetbänder mit zufallserzeugten Bitfolgen aus. Jedesmal, wenn eine Nachricht übermittelt werden sollte, übersetzte der Fernschreiber diese in eine Bitfolge, d.h. in einen Vektor $\vec{v}$ aus einem Vektorraum $\mathbb{F}_2^N$. Aus den ersten N bislang noch nicht benutzten Bits auf dem Magnetband wurde dazu ein

weiterer Vektor $\vec{w} \in \mathbb{F}_2^N$ gebildet, und tatsächlich übertragen wurde die Summe $\vec{s} = \vec{v} + \vec{w}$.

Am anderen Ende der Leitung, wo eine Kopie des Magnetbands vorlag, war $\vec{w}$ bekannt, so daß die Nachricht

$$\vec{v} = \vec{v} + \vec{0} = \vec{v} + (\vec{w} + \vec{w}) = (\vec{v} + \vec{w}) + \vec{w} = \vec{s} + \vec{w}$$

rekonstruiert werden konnte.

Ein Lauscher ohne Magnetband konnte nur die Länge N der Nachricht ermitteln, was bei den seitenlangen in Diplomatensprache formulierten Texten, die über diese Leitung liefen, so gut wie keine konkrete Information lieferte. In der Tat können auch schon sehr kurze Nachrichten gleicher Länge völlig verschiedenen Inhalt haben: Im Deutschen etwa besteht der Satz „Herzlichen Glückwunsch zu Ihrem sehr guten Klausurergebnis!“ aus genauso vielen Zeichen wie „Mit 3 von 2000 Punkten haben Sie das schlechteste Ergebnis.“ Entsprechend hat auch jemand, der irgendeinen Vektor $\vec{s}$ in die Leitung einspielt, so gut wie keine Chance, daß nach Addition von $\vec{w}$ daraus verständlicher Text wird, so daß die Manipulation mit an Sicherheit grenzender Wahrscheinlichkeit entdeckt wird.

Diese Art der Kommunikation ist also sehr sicher, aber leider auch sehr aufwendig: Wer einfach ein Buch im Internet bestellen will, hat üblicherweise keine Möglichkeit, vorher über Kurier ein Magnetband oder eine CD-ROM mit dem Versandhaus auszutauschen, bevor er seine Kontendaten dorthin schickt. Für Alltagsanwendungen braucht man daher Verfahren, die einfacher anwendbar sind. Leider sind die wirklich guten darunter mathematisch deutlich anspruchsvoller als der *one time pad;* zwei davon werden wir im Laufe dieses Paragraphen noch kennenlernen.

b) Körper von Primzahlordnung

Der Körper mit zwei Elementen ist nur einen von vielen endlichen Körpern; beispielsweise gibt es zu jeder Primzahl p einen solchen Körper; wir bezeichnen ihn in Analogie zu $\mathbb{F}_2$ mit $\mathbb{F}_p$.

Als Menge ist $\mathbb{F}_p \underset{\text{def}}{=} \{0, 1, \ldots, p-1\}$; Addition und Multiplikation werden definiert durch die Vorschriften

$$a \oplus b \underset{\text{def}}{=} (a + b) \bmod p \quad \text{und} \quad a \odot b \underset{\text{def}}{=} ab \bmod p\,,$$

d.h. für führen zunächst die entsprechenden Operationen für ganze Zahlen aus und betrachten dann den Divisionsrest des Ergebnisses bei Division durch die Primzahl p. Für $p = 2$ führt das auf die inzwischen wohlbekannten Rechenoperationen von $\mathbb{F}_2$.

Da das Kommutativ- und das Assoziativgesetz für Addition und Multiplikation ganzer Zahlen gelten und zwei gleiche Zahlen insbesondere den gleichen Divisionsrest modulo p haben, gelten diese Gesetze auch für $\oplus$ und $\odot$; aus demselben Grund gilt auch das Distributivgesetz, und 0 und 1 sind Neutralelemente bezüglich $\oplus$ und $\odot$.

Das zu a inverse Element bezüglich der Addition ist für $a = 0$ natürlich a selbst, ansonsten $p - a$, denn

$$a \oplus (p - a) = a + (p - a) \bmod p = p \bmod p = 0\,.$$

Multiplikative Inverse lassen sich nicht so leicht finden, aber immerhin läßt sich relativ einfach sehen, daß die existieren: Für $a \in \mathbb{F}_p \setminus \{0\}$ betrachten wir die sämtlichen Elemente $a \cdot x$ mit $x \in \mathbb{F}_p$. Ist $a \cdot x = a \cdot y$, so ist $a \cdot (x - y) = 0$, d.h. $a(x - y)$ ist durch p teilbar. Da p eine Primzahl ist (das verwenden wir hier zum ersten Mal!), muß dann auch mindestens einer der beiden Faktoren durch p teilbar sein. a aber ist eine Zahl zwischen 1 und $p - 1$, also sicherlich nicht durch p teilbar. Somit teilt p die Differenz $x - y$. Da x, y als Elemente von $\mathbb{F}_p$ höchstens gleich $p - 1$ sind, hat ihre Differenz höchstens Betrag $p - 1$, also kann sie nur dann durch p teilbar sein, wenn sie verschwindet. Somit gilt: Für $x, y \in \mathbb{F}_p$ ist $a \cdot x = a \cdot y$ genau dann, wenn $x = y$. Die p Elemente $a \cdot x$ sind somit allesamt verschieden; da es in $\mathbb{F}_p$ nur p Elemente gibt, läßt sich also jedes von diesen in der Form $a \cdot x$ mit einem geeigneten $x \in \mathbb{F}_p$ schreiben. Dies gilt insbesondere für die Eins, und damit ist auch die Existenz multiplikativer Inverser als letztes des Körpereaxiome nachgewiesen.

$\mathbb{F}_p$ ist also in der Tat ein Körper.

Der Aufwand für das Rechnen in diesem Körper ist am geringsten für die Addition: Da $a+b$ für $a, b \in \mathbb{F}_p$ zwischen Null und $2p-2$ liegt, ist

$$a \oplus b = \begin{cases} a+b & \text{falls } a+b < p \\ a+b-p & \text{sonst} \end{cases},$$

hier braucht man also keine Division, und dasselbe gilt natürlich auch für die Subtraktion:

$$a \ominus b = \begin{cases} a-b & \text{falls } a-b \geq 0 \\ a-b+p & \text{sonst} \end{cases}.$$

Zur Berechnung des Produkts zweier Elemente von $\mathbb{F}_p$ brauchen wir eine Multiplikation ganzer Zahlen und eine Division mit Rest; hier ist also der Aufwand deutlich höher.

Am aufwendigsten ist die Division in $\mathbb{F}_p$: Bislang können wir nur durch a dividieren, indem wir alle Produkte von a mit Elementen aus $\mathbb{F}_p$ systematisch durchprobieren. Da p bei einigen kryptographischen Anwendungen durchaus mehr als sechshundert Dezimalstellen haben kann, brauchen wir dringend eine effizientere Alternative; diese wird uns der EUKLIDische Algorithmus liefern.

Im folgenden werde ich, wenn keine Verwechslungsgefahr mit ganzen Zahlen besteht, die Rechenoperationen in $\mathbb{F}_p$ meist einfach mit + und $\cdot$ anstelle von $\oplus$ und $\odot$ bezeichnen.

c) Der Euklidische Algorithmus

Beginnen wir mit dem einfachsten Fall, für den der Algorithmus schon im zehnten Buch von EUKLIDs Elementen zu finden ist: Wir suchen den größten gemeinsamen Teiler zweier nichtnegativer ganzer Zahlen a und b, d.h. die größte ganze Zahl d, die sowohl a als auch b teilt. Für $a = b = 0$ gibt es kein *größtes* solches d; hier setzen wir $d = 0$. Wir schreiben kurz

$$d = \mathrm{ggT}(a, b)\,.$$

Grundidee des EUKLIDischen Algorithmus ist die Anwendung der Division mit Rest: Für je zwei natürliche Zahlen x und y gibt es nichtnegative ganze Zahlen q und r, so daß

$$x = qy + r \quad \text{und} \quad 0 \leq r < y$$

ist. Alsdann ist

$$\mathrm{ggT}(x, y) = \mathrm{ggT}(y, r)\,,$$

denn wegen der beiden Gleichungen

$$x = qy + r \quad \text{und} \quad r = x - qy$$

teilt jeder gemeinsame Teiler von x und y auch r, und jeder gemeinsame Teiler von y und r teilt auch x. Da außerdem offensichtlich

$$\mathrm{ggT}(x, 0) = x \quad \text{für alle } x \in \mathbb{N}_0$$

ist, können wir den ggT leicht rekursiv berechnens, indem wir die Regel $\mathrm{ggT}(x, y) = \mathrm{ggT}(y, r)$ so lange anwenden, bis $r = 0$ und damit der ggT gleich y ist. In Scheme ist der EUKLIDische Algorithmus somit einfach die folgende Vorschrift:

```
(define (ggT x y) (if (= y 0) x (ggT y (modulo x y))))
```

Der Algorithmus bricht ab, da das zweite Argument y in jedem Rekursionsschritt kleiner wird, aber stets eine nichtnegative ganze Zahl ist; nach endlich vielen Schritten muß also $y = 0$ sein und der ggT wird als $\mathrm{ggT}(x, 0) = x$ berechnet. Die Korrektheit des Ergebnisses ist auch klar, denn die Gleichung

$$\mathrm{ggT}(x, y) = \mathrm{ggT}(y, x \bmod y)$$

besagt ja gerade, daß sich am ggT nichts ändert, wenn wir (x, y) durch (y, r) ersetzen, wobei r den Rest bei der Division von x durch y bezeichnet.

Formuliert man den Algorithmus nicht in Scheme, sondern in natürlicher Sprache, läßt er sich nicht mehr so kompakt darstellen; dann sieht der EUKLIDische Algorithmus zur Berechnung des ggT zweier nichtnegativer ganzer Zahlen x und y zum Beispiel wie folgt aus:

Schritt 0: Setze $r_0 = x$ und $r_1 = y$

Schritt i, $i \geq 1$**:** Falls $r_i = 0$ ist, endet der Algorithmus mit dem Ergebnis

$$\mathrm{ggT}(x, y) = r_{i-1}\,;$$

andernfalls dividiere man r_{i-1} mit Rest durch r_i und bezeichne den Divisionsrest mit r_{i+1}.

(Für eine Implementierung in egal welcher Programmiersprache bieten sich natürlich einige offensichtliche Optimierungen an.)

Bislang ist noch nicht zu sehen, wie uns der EUKLIDische Algorithmus bei der Division in einem Körper $\mathbb{F}_p$ helfen kann; dies leistet erst seine (auch schon bei EUKLID zu findende) *Erweiterung:*

Die Gleichung $u = qv + r$ zur Division mit Rest läßt sich auch umschreiben als $r = u - qv$; der Divisionsrest ist also eine ganzzahlige Linearkombination des Dividenden u und des Divisors v. Falls sich diese wiederum als Linearkombination der beiden Ausgangszahlen x und y darstellen lassen, erhalten wir eine entsprechende Darstellung für r:

$$u = ax + by \quad \text{und} \quad v = cx + dy \quad \Longrightarrow r = (a - qc)x + (b - qd)z\,.$$

Wir können also ausgehen von den Darstellungen

$$x = 1 \cdot x + 0 \cdot y \quad \text{und} \quad y = 0 \cdot x + 1 \cdot y\,,$$

bei jeder Division im EUKLIDischen Algorithmus den Divisionsrest als ganzzahlige Linearkombination von x und y darstellen und damit auch den ggT als den letzten nichtverschwindenen solchen Rest.

In Scheme läßt sich das beispielsweise folgendermaßen formulieren:

```
(define (erwEuklid x y)
   (define (iteration u v a b c d)
      (if (= v 0) (list u a b)
       (let ((q (quotient u v)))
       (iteration v (remainder u v) c d
                  (- a (* q c)) (- b (* q d))))))
   (iteration x y 1 0 0 1))
```

In natürlicher Sprache kann das etwa so ausgedrückt werden:

Schritt 0: Setze $r_0 = a$, $r_1 = b$, $\alpha_0 = \beta_1 = 1$ und $\alpha_1 = \beta_0 = 0$. Mit $i = 1$ ist dann

$$r_{i-1} = \alpha_{i-1}a + \beta_{i-1}b \quad \text{und} \quad r_i = \alpha_i a + \beta_i b\,.$$

Diese Relationen werden in jedem der folgenden Schritte erhalten:

Schritt i, $i \geq 1$**:** Falls $r_i = 0$ ist, endet der Algorithmus mit

$$\mathrm{ggT}(a, b) = r_{i-1} = \alpha_{i-1}a + \beta_{i-1}b\,.$$

Andernfalls dividiere man r_{i-1} mit Rest durch r_i mit dem Ergebnis

$$r_{i-1} = q_i r_i + r_{i+1}\,.$$

Dann ist

$$\begin{aligned} r_{i+1} = -q_i r_i + r_{i-1} &= -q_i(\alpha_i a + \beta_i b) + (\alpha_{i-1}a + \beta_{i-1}b) \\ &= (\alpha_{i-1} - q_i\alpha_i)a + (\beta_{i-1} - q_i\beta_i)b\,; \end{aligned}$$

man setze also

$$\alpha_{i+1} = \alpha_{i-1} - q_i\alpha_i \quad \text{und} \quad \beta_{i+1} = \beta_{i-1} - q_i\beta_i\,.$$

Genau wie oben folgt, daß der Algorithmus für alle natürlichen Zahlen a und b endet und daß am Ende der richtige ggT berechnet wird; außerdem sind die α_i und β_i so definiert, daß in jedem Schritt $r_i = \alpha_i a + \beta_i b$ ist, insbesondere ist also im letzten Schritt der ggT als Linearkombination der Ausgangszahlen dargestellt.

Als Beispiel wollen wir den ggT von 200 und 148 als Linearkombination darstellen. Im nullten Schritt haben wir 200 und 148 als die trivialen Linearkombinationen

$$200 = 1 \cdot 200 + 0 \cdot 148 \quad \text{und} \quad 148 = 0 \cdot 200 + 1 \cdot 148\,.$$

Im ersten Schritt dividieren wir, da 148 nicht verschwindet, 200 mit Rest durch 148:

$$200 = 1 \cdot 148 + 52 \Longrightarrow 52 = 1 \cdot 200 - 1 \cdot 148$$

Da auch $52 \neq 0$, dividieren wir im zweiten Schritt 148 durch 52 mit Ergebnis $148 = 2 \cdot 52 + 44$, d.h.

$$44 = 148 - 2 \cdot (1 \cdot 200 - 1 \cdot 148) = 3 \cdot 148 - 2 \cdot 200$$

Auch $44 \neq 0$, wir dividieren also weiter: $52 = 1 \cdot 44 + 8$ und

$$\begin{aligned} 8 = 52 - 44 &= (1 \cdot 200 - 1 \cdot 148) - (3 \cdot 148 - 2 \cdot 200) \\ &= 3 \cdot 200 - 4 \cdot 148\,. \end{aligned}$$

Im nächsten Schritt erhalten wir $44 = 5 \cdot 8 + 4$ und

$$\begin{aligned} 4 = 44 - 5 \cdot 8 &= (3 \cdot 148 - 2 \cdot 200) - 5 \cdot (3 \cdot 200 - 4 \cdot 148) \\ &= 23 \cdot 148 - 17 \cdot 200\,. \end{aligned}$$

Bei der Division von acht durch vier schließlich erhalten wir Divisionsrest Null; damit ist vier der ggT von 148 und 200 und kann in der angegebenen Weise linear kombiniert werden.

d) Das RSA-Verfahren

Als praktische Anwendungen der Körper $\mathbb{F}_p$ und des erweiterten EUKLIDischen Algorithmus möchte ich zwei Beispiele aus der Kryptographie betrachten.

Wir kennen bereits den absolut sicheren *one time pad;* leider können wir ihn aber nur anwenden, wenn vorher riesige Mengen an Schlüsselbits ausgetauscht wurden. Dies ist kein Problem für die diplomatischen Dienste großer Staaten, ist aber völlig unrealistisch für die meisten Fälle von Kommunikation zwischen Privatleuten. Hier ist oft nicht einmal der sichere Austausch eines kurzen, für längere Zeit gültigen Schlüssels wirklich praktikabel.

Vor dem Hintergrund dieses Problems veröffentlichten 1976 MARTIN HELLMAN, damals Assistenzprofessor in Stanford, und sein Forschungsassistent WHITFIELD DIFFIE eine Arbeit mit dem Titel *New directions in cryptography* (IEEE Trans. Inform. Theory **22**, 644–654), in der sie vorschlugen, den Vorgang der *Ver*schlüsselung und den der *Ent*schlüsselung völlig voneinander zu trennen: Es sei schließlich nicht notwendig, daß der Sender einer verschlüsselten Nachricht auch in der Lage sei, diese zu *ent*schlüsseln. Der Vorteil eines solchen Verfahrens wäre, daß jeder potentielle Empfänger nur einen einzigen Schlüssel bräuchte und dennoch sicher sein könnte, daß nur er selbst seine Post entschlüsseln kann. Der Schlüssel müßte nicht einmal geheimgehalten werden, da es ja nicht schadet, wenn jedermann Nachrichten *ver*schlüsseln kann. In einem Netzwerk mit n Teilnehmern bräuchte man also nur n Schlüssel, um es jedem Teilnehmer zu erlauben, mit jeden anderen so zu kommunizieren, und diese Schlüssel könnten sogar in einem öffentlichen Verzeichnis stehen. Bei einem symmetrischen Kryptosystem wäre der gleiche Zweck nur erreichbar mit $\frac{1}{2}n(n-1)$ Schlüsseln, die zudem noch durch ein sicheres Verfahren wie etwa ein persönliches Treffen oder durch vertrauenswürdige Boten ausgetauscht werden müßten.

DIFFIE und HELLMAN machten nur sehr vage Andeutungen, wie so ein System mit öffentlichen Schritten aussehen könnte; klar war nur, daß es mit einer sogenannten *Einwegfunktion* arbeiten mußte, d.h. mit einer Funktion, die jedermann leicht berechnen kann, deren Umkehrfunktion

aber nicht berechenbar ist.

Mathematisch betrachtet kann es eine solche Funktion nicht geben: Da Verschlüsselungsfunktionen immer Funktionen zwischen endlichen Mengen sind und aus offensichtlichen Gründen injektiv sein müssen, kann *im Prinzip* jeder durch Probieren das Umkehrproblem lösen. Wenn man aber wüßte, daß jeder Versuch, die inverse Funktion zur Verschlüsselung zu finden, weit jenseits der Leistungsgrenze heutiger Computer liegt, wäre ein solches Verfahren trotzdem *praktisch* sicher.

Tatsächlich wäre es sogar so sicher, daß nicht einmal der legitime Empfänger seine Post lesen könnte; anwendbar wird es erst, wenn die Entschlüsselung *für genau eine Person* möglich ist auf Grund geheimer Information, über die sonst niemand verfügt. Solche Einwegfunktionen bezeichnet man als *Einwegfunktionen mit Falltür.* DIFFIE und HELLMAN kannten kein Beispiel einer solche Funktion, und es gab unter vielen Experten große Skepsis bezüglich der Möglichkeit, je eine zu finden.

Wie inzwischen bekannt ist, gab es damals bereits Systeme, die auf solchen Funktionen beruhten; sie waren allerdings nicht in der offenen Literatur dokumentiert: Die britische *Communications-Electronics Security Group* (CESG) hatte sich bereits Ende der sechziger Jahre mit diesem Problem befaßt, um die Probleme des Militärs mit dem Schlüsselmanagement zu lösen, aufbauend auf (impraktikablen) Ansätzen von AT&T zur Sprachverschlüsselung, die während des zweiten Weltkriegs untersucht wurden. Die Briten sprachen nicht von Kryptographie mit öffentlichen Schlüsseln, sondern von *nichtgeheimer Verschlüsselung,* aber das Prinzip war das gleiche.

Erste Ideen dazu sind in einer auf Januar 1970 datierten Arbeit von JAMES H. ELLIS zu finden, ein praktikables System in einer auf den 20. November 1973 datierten Arbeit von CLIFF C. COCKS. Wie im Milieu üblich, gelangte nichts über diese Arbeiten an die Öffentlichkeit; erst 1997 veröffentlichten die *Goverment Communications Headquarters* (GCHQ), zu denen CESG gehört, einige Arbeiten aus der damaligen Zeit auf ihrer Website. Die Links sind inzwischen verschwunden, man findet einige der Arbeiten aber immer noch, wenn man auf

http://www.cesg.gov.uk/ unter *CESG Publications* direkt nach ELLIS oder COCKS sucht.

Im akademischen Bereich gab es ein Jahr nach Erscheinen der Arbeit von DIFFIE und HELLMAN das erste Kryptosystem mit öffentlichen Schlüsseln: Drei Professoren am Massachussetts Institute of Technology fanden nach rund vierzig erfolglosen Ansätzen 1977 schließlich jenes System, das heute nach ihren Anfangsbuchstaben mit RSA bezeichnet wird: RON RIVEST, ADI SHAMIR und LEN ADLEMAN.

Das System wurde 1983 von der eigens dafür gegründeten Firma RSA Computer Security Inc. patentiert und mit großem kommerziellem Erfolg vermarktet. Das Patent lief zwar im September 2000 aus, die Firma ist aber weiterhin erfolgreich im Kryptobereich tätig.

RSA ist übrigens identisch mit dem von COCKS vorgeschlagenen System, so daß einige Historiker auch Zweifel an den Behauptungen der GCHQ haben. Die Beschreibung durch RIVEST, SHAMIR und ADLEMAN erschien 1978 unter dem Titel *A method for obtaining digital signatures and public-key cryptosystems* in Comm. ACM **21**, 120–126.

Grundidee ist die Verwendung einer Abbildung der Form

$$\{0,1,\dots,N-1\} \to \{0,1,\dots,N-1\}; \quad x \mapsto x^e \bmod N\,.$$

Sie ist einfach auszurechnen, für geeignetes e auch injektiv, und das Problem, ihre Umkehrabbildung effizient zu berechnen ist zumindest für hinreichend allgemeine große Werte von N ungelöst. Wie Abbildung 11 zeigt, sieht eine Potenzfunktion modulo N bei weitem nicht so harmlos aus wie eine Potenzfunktion in $\mathbb{R}$ – es ist zunächst nicht einmal wirklich klar, wann sie injektiv ist.

Für den Fall, daß $N = p$ eine Primzahl ist und wir somit in einem der endlichen Körper $\mathbb{F}_p$ arbeiten, lassen sich die injektiven Potenzfunktionen charakterisieren und auch umkehren. Ausgangspunkt dazu ist der folgende Satz:

Kleiner Satz von Fermat: Für jedes $a \in \mathbb{Z}$ ist

$$a^p \equiv a \bmod p\,;$$

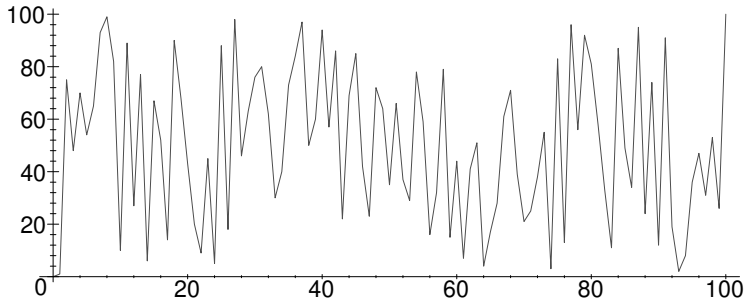


Abb. 11: Die Funktion $x \mapsto x^{17}$ in $\mathbb{F}_{101}$

ist a nicht durch p teilbar, ist auch

$$a^{p-1} \equiv 1 \bmod p\,.$$

Beweis: Für zwei Zahlen $x, y \in \mathbb{Z}$ ist

$$(x+y)^p = \binom{p}{0}x^p + \binom{p}{1}x^{p-1}y + \cdots + \binom{p}{p-1}xy^{p-1} + \binom{p}{p}y^p$$

mit

$$\binom{p}{j} = \frac{p!}{j!(p-j)!}\,.$$

Für $j = 0$ und $j = p$ ist $\binom{p}{0} = \binom{p}{p} = 1$, ansonsten sind j und $p-j$ beide kleiner als p. Im Nenner steht daher dann kein Faktor p, der das p aus dem Zähler wegkürzen könnte, so daß alle $\binom{p}{j}$ mit $1 \le j \le p-1$ durch p teilbar sind, und

$$(x+y)^p \equiv x^p + y^p \bmod p\,.$$

Induktiv folgt, daß eine entsprechende Gleichung auch für Summen mit mehr als zwei Summanden gilt, insbesondere gilt also für beliebig viele Summanden eins, daß

$$(1+\cdots+1)^p \equiv 1^p + \cdots + 1^p = 1 + \cdots + 1 \bmod p\,.$$

Damit ist $a^p \equiv a \bmod p$ für jede natürliches Zahl a. Für $a = 0$ gilt die Beziehung auch, und da für jedes positive a

$$0 = \big(a + (-a)\big)^p = a^p + (-a)^p = a + (-a)^p \bmod p$$

ist, gilt sie auch für negative a.

Falls a nicht durch p teilbar ist, gibt es, da $\mathbb{F}_p$ ein Körper ist, ein b mit $ba \equiv 1 \bmod p$; mit diesem b ist

$$a^{p-1} \equiv (ba)a^{p-1} = ba^p \equiv ba \equiv 1 \bmod p\,.$$

■

Der französische Mathematiker PIERRE DE FERMAT (1601–1665) wurde in Beaumont-de-Lomagne im Departement Tarn et Garonne geboren. Bekannt ist er heutzutage vor allem für seine 1994 von ANDREW WILES bewiesene Vermutung, wonach die Gleichung $x^n + y^n = z^n$ für $n \geq 3$ keine ganzzahlige Lösung mit $xyz \neq 0$ hat. Dieser „große" Satzes von FERMATs, von dem FERMAT lediglich in einer Randnotitz behauptete, daß er ihn beweisen könne, erklärt den Namen der obigen Aussage. Obwohl er sich sein Leben lang sehr mit Mathematik beschäftigte und wesentliche Beiträge zur Zahlentheorie, Wahrscheinlichkeitstheorie und Analysis lieferte, war er hauptberuflich Jurist.

Korollar: Falls die natürliche Zahl e keinen Teiler mit $p-1$ gemeinsam hat, gibt es eine natürliche Zahl d, so daß

$$\left(a^e\right)^d \equiv a \bmod p$$

für alle $a \in \mathbb{Z}$. Insbesondere ist die Abbildung $x \mapsto x^e$ von $\mathbb{F}_p$ nach $\mathbb{F}_p$ dann injektiv.

Beweis: Nach dem erweiterten EUKLIDischen Algorithmus läßt sich der größte gemeinsame Teiler eins von e und $p-1$ als ganzzahlige Linearkombination dieser beiden Zahlen schreiben, es gibt also ganze Zahlen d und r, so daß

$$d \cdot e + r \cdot (p-1) = 1$$

ist. Hier könnte d eventuell noch negativ sein, aber indem wir nötigenfalls noch ein Vielfaches der Gleichung

$$(p-1) \cdot e - e \cdot (p-1) = 0$$

dazuaddieren, können wir annehmen, daß d positiv ist (und r dann natürlich negativ). Für nicht durch p teilbares a ist dann

$$\left(a^e\right)^d = a^{de} = a^{1-r(p-1)} = a \cdot \left(a^{p-1}\right)^{-r} \equiv a \cdot 1^{-r} = a \bmod p\,,$$

wie behauptet.

Für durch p teilbares a sind beide Seiten der zu beweisenden Kongruenz durch p teilbar, so daß sie auch in diesem Fall richtig ist.

■

In diesem Fall gibt uns also der erweiterte EUKLIDische Algorithmus eine natürliche Zahl d, so daß die Umkehrfunktion zum Potenzieren mit e einfach das Potenzieren mit d ist. Den Exponenten d kann jeder berechnen, der p und e (und den erweiterten EUKLIDischen Algorithmus) kennt; da man p und e zum Verschlüsseln braucht, liefert das also kein Kryptosystem mit öffentlichen Schlüsseln, sondern höchstens bei ein klassisches System. (Unter dem Namen POHLIG/HELLMAN-Verfahren wird es für einige wenige Spezialfälle wie das Pokerspiel via Internet auch tatsächlich verwendet.)

Das RSA-Verfahren arbeitet nicht modulo einer Primzahl p, sondern modulo dem Produkt $N = pq$ zweier Primzahlen p und q. Hier führt der kleine Satz von FERMAT zu folgenden Aussagen:

Satz: Ist $N = pq$ Produkt zweier verschiedener Primzahlen p und q, so gilt
a) Für jede ganze Zahl a ist $a^{1+(p-1)(q-1)} \equiv a \bmod N$.
b) Falls die natürliche Zahl e keinen Teiler mit $(p-1)(q-1)$ gemeinsam hat, gibt es eine natürliche Zahl d, so daß

$$\left(a^e\right)^d \equiv a \bmod p$$

für alle $a \in \mathbb{Z}$.
c) Die Berechnung von $(p-1)(q-1)$ aus N ist äquivalent zur Faktorisierung von N.

Beweis: a) Nach dem kleinen Satz von FERMAT ist für jedes nicht durch p teilbare a und jedes r

$$a^{1+r(p-1)} \equiv a \bmod p\,.$$

Für Vielfache von p sind beide Seiten durch p teilbar, so daß die Gleichung tatsächlich für alle ganzen Zahlen a gilt.

Entsprechendes gilt natürlich auch für die Primzahl q, und damit ist für jede ganze Zahl s

$$a^{1+s(p-1)(q-1)} \equiv a \bmod p \quad \text{und} \quad a^{1+(p-1)(q-1)} \equiv a \bmod q\,.$$

Diese beiden Kongruenzen besagen, daß die Differenz zwischen linker und rechter Seite sowohl durch p als auch durch q teilbar ist, also auch durch deren Produkt N, d.h.

$$a^{1+s(p-1)(q-1)} \equiv a \bmod N\,.$$

b) Geht wegen *a)* genau wie beim Primzahlfall im obigen Korollar.

c) Falls die Faktorisierung $N = pq$ bekannt ist, läßt sich natürlich leicht $(p-1)(q-1)$ berechnen. Ist umgekehrt

$$(p-1)(q-1) = pq - p - q + 1 = N + 1 - (p+q)$$

bekannt, so kennt man außer dem Produkt N auch die Summe M der beiden Primfaktoren, und diese sind die leicht berechenbaren Lösungen der quadratischen Gleichung $x^2 - Mx + N = 0$. ∎

Zur praktischen Durchführung des RSA-Verfahrens wählt man sich zwei verschiedene Primzahlen p, q, die unbedingt geheim gehalten werden müssen, und eine natürliche Zahl e, die keinen gemeinsamen Teiler mit $(p-1)(q-1)$ hat. Dann veröffentlicht man die beiden Zahlen $N = pq$ und e als öffentlichen Schlüssel.

Sodann berechnet man zu $(p-1)(q-1)$ gemäß obigem Satz nach dem EUKLIDischen Algorithmus eine Zahl d, so daß

$$\left(a^e\right)^d \equiv a \bmod N$$

für alle a; diese Zahl ist der geheime Schlüssel.

1) Verschlüsselung: Jeder, der den öffentlichen Schlüssel (N, e) kennt, kann Nachrichten verschlüsseln: Er bricht die Nachricht auf in Blöcke, die durch ganze Zahlen zwischen null und $N-1$ dargestellt werden können, berechnet für jeden so dargestellten Block den Chiffretext $b \equiv a^e \bmod N$, der als Zahl zwischen null und $N-1$ interpretiert und an den Inhaber des geheimen Schlüssels geschickt wird. (Wir schreiben

in solchen Fällen in Zukunft kurz $b = a^e \bmod N$, wobei „mod“ in diesem Zusammenhang als die Berechnung des Divisionsrests bei Division durch N zu interpretieren ist.)

Der Empfänger berechnet $b^d \bmod N$; da

$$b^d \equiv a^{ed} \equiv a \bmod N$$

ist, entschlüsselt dies die Nachricht.

2) Identitätsnachweis: Im Gegensatz zu symmetrischen Kryptoverfahren endet die Nützlichkeit des RSA-Verfahrens nicht mit der bloßen Möglichkeit einer Verschlüsselung; das Verfahren kann beispielsweise auch benutzt werden, um in Zugangskontrollsystemen, vor Geldautomaten oder bei einer Bestellung im Internet die Identität einer Person Zu beweisen: Nur der Inhaber des geheimen Schlüssels d kann zu einem gegebenen a eine Zahl b berechnen, für die $b^e \equiv a \bmod N$ ist.

Falls also der jeweilige Gegenüber eine Zufallszahl a erzeugt und als Antwort das zugehörige b verlangt, kann er anhand eines öffentlichen Schlüsselverzeichnisses die Richtigkeit von b überprüfen und sich so von der Identität seines Partners überzeugen. Im Gegensatz zu Kreditkarteninformation oder Paßwörtern ist dieses Verfahren auch immun gegen Abhören: Falls jedesmal ein neues zufälliges a erzeugt wird, nützt ein einmal abgehörtes b nichts.

Trotzdem ist das Verfahren in dieser Form nicht als Ersatz zur Übertragung von Kreditkarteninformation oder ähnlichem geeignet, da der Gegenüber anhand des öffentlichen Schlüssels jederzeit zu einer willkürlich gewählten Zahl b die Zahl $a = b^e \bmod N$ erzeugen kann um dann zu behaupten, er habe b als Antwort darauf empfangen. Man müßte also beispielsweise noch zusätzlich verlangen, daß die Zahl a eine spezielle Form hat, etwa daß die vordere Hälte der Ziffernfolge identisch mit der hinteren Hälfte ist.

3) Elektronische Unterschriften: Praktische Bedeutung hat vor allem eine weitere Variante: die elektronische Unterschrift. Her geht es darum, daß der Empfänger erstens davon überzeugt wird, daß eine Nachricht

tatsächlich vom behaupteten Absender stammt, und daß er dies zweitens auch einem Dritten gegenüber beweisen kann. (In Deutschland sind solche elektronischen Unterschriften, sofern gewisse formale Voraussetzungen erfüllt sind, rechtsverbindlich.)

Um einen Nachrichtenblock a mit $0 \leq a < N$ zu unterschreiben, berechnet der Inhaber des öffentlichen Schlüsseln (N, e) mit seinem geheimen Schlüssel d die Zahl

$$b = a^d \bmod N$$

und sendet das Paar (a, b) an den Empfänger. Dieser überprüft, ob

$$b^e \equiv a \bmod N\,;$$

falls ja, akzeptiert er dies als unterschriebene Nachricht a. Da er ohne Kenntnis des geheimen Schlüssels d nicht in der Lage ist, den Block (a, b) zu erzeugen, kann er auch gegenüber einem Dritten beweisen, daß der Absender die Nachricht a unterschrieben hat.

Für kurze Nachrichten ist dieses Verfahren in der vorgestellten Form praktikabel; in vielen Fällen kann man sogar auf die Übermittlung von a verzichten, da $b^e \bmod N$ für ein falsch berechnetes b mit an Sicherheit grenzender Wahrscheinlichkeit keine sinnvolle Nachricht ergibt.

Falls die übermittelte Nachricht geheimgehalten werden soll, müssen a und b natürlich noch vor der Übertragung mit dem öffentlichen Schlüssel des Empfängers oder nach irgendeinem anderen Kryptoverfahren verschlüsselt werden.

Bei langen Nachrichten ist die Verdoppelung der Nachrichtenlänge nicht mehr akzeptabel, und selbst, wenn man auf die Übertragung von a verzichten kann, ist das Unterschreiben jedes einzelnen Blocks sehr aufwendig. Deshalb unterschreibt man meist nicht die Nachricht selbst, sondern einen daraus extrahierten Hashwert. Dieser Wert muß natürlich erstens von der gesamten Nachricht abhängen, und zweitens muß es für den Empfänger (praktisch) unmöglich sein, zwei Nachrichten zu erzeugen, die zum gleichen Hashwert führen. Mit der Theorie (und Praxis) dieser sogenannten kollisionsfreien Hashfunktionen werden wir uns später beschäftigen.

4) Blinde Unterschriften und elektronisches Bargeld: Einer der erfolgversprechendsten Ansätze zum Aushebeln eines Kryptosystems besteht darin, sich auf die Dummheit seiner Mitmenschen zu verlassen.

So sollte es durch gutes Zureden nicht schwer sein, jemanden zu Demonstrationszwecken zum Unterschreiben einer sinnlosen Nachricht zu bewegen: Eine Folge von Nullen und Einsen ohne sinnvolle Interpretation hat schließlich keine rechtliche Wirkung.

Nun muß eine sinnlose Nachricht aber nicht unbedingt eine Zufallszahl sein: Sie kann sorgfältig präpariert sein. Sei dazu etwa m eine Nachricht, die ein Zahlungsversprechen enthält, (N, e) der öffentliche Schlüssel des Opfers und r eine Zufallszahl zwischen 2 und $N - 2$. Dann wird

$$x = m \cdot r^e \bmod N$$

wie eine Zufallsfolge aussehen, für die man eine Unterschrift

$$u = x^d \bmod N = (mr^e)^d \bmod N = m^d r \bmod N$$

bekommt. Multiplikation mit r^{-1} macht daraus eine Unterschrift unter die Zahlungsverpflichtung m.

Das angegebene Verfahren kann nicht nur von Trickbetrügern benutzt werden; blinde Unterschriften sind auch die Grundlage von *digitalem Bargeld.*

Zahlungen im Internet erfolgen meist über Kreditkarten; die Kreditkartengesellschaften haben also einen recht guten Überblick über die Ausgaben ihrer Kunden und machen teilweise auch recht gute Geschäfte mit Kundenprofilen.

Digitales Bargeld will die Anonymität von Geldscheinen mit elektronischer Übertragbarkeit kombinieren und so ein anonymes Zahlungssystem z.B. für das Internet bieten.

Es wir ausgegeben von einer Bank, die für jede angebotene Stückelung einen öffentlichen Schlüssel (N, e) bekanntgibt. Eine Banknote ist eine mit dem zugehörigen geheimen Schlüssel unterschriebene Seriennummer.

Die Seriennummer kann natürlich nicht einfach *jede* Zahl sein; sonst wäre jede Zahl kleiner N eine Banknote. Andererseit dürfen die Seriennummern aber auch nicht von der Bank vergeben werden, denn sonst wüßte diese, welcher Kunde Scheine mit welchem Seriennummern hat. Als Ausweg wählt man Seriennummern einer sehr speziellen Form: Ist $N > 10^{150}$, kann man etwa als Seriennummer eine 150-stellige Zahl wählen, deren Ziffern spiegelsymmetrisch zur Mitte sind, d.h. ab der 76. Ziffer werden die vorherigen Ziffern rückwärts wiederholt. Die Wahrscheinlichkeit, daß eine zufällige Zahl x nach Anwendung des öffentlichen Exponenten auf so eine Zahl führt, ist 10^{-75} und damit vernachlässigbar.

Seriennummern werden von den Kunden zufällig erzeugt. Für jede solche Seriennummer m erzeugt der Kunde eine Zufallszahl r, schickt $mr^e \bmod N$ an die Bank und erhält (nach Belastung seines Kontos) eine Unterschrift u für diese Nachricht zurück. Wie oben berechnet er daraus durch Multiplikation mit r^{-1} die Unterschrift $v = m^d \bmod N$ für die Seriennummer N, und mit diesem Block kann er bezahlen.

Der Zahlungsempfänger berechnet $v^e \bmod N$; falls dies die Form einer gültigen Seriennummer hat, kann er sicher sein, einen von der Bank unterschriebenen Geldschein vor sich zu haben. Er kann allerdings noch nicht sicher sein, daß dieser Geldschein nicht schon einmal ausgegeben wurde.

Deshalb muß er die Seriennummer an die Bank melden, die mit ihrer Datenbank bereits ausbezahlter Seriennummern vergleicht. Falls sie darin noch nicht vorkommt, wird sie eingetragen und der Händler bekommt sein Geld; andernfalls verweigert sie die Zahlung.

Schon bei nur 10^{75} möglichen Nummern liegt die Wahrscheinlichkeit dafür, daß zwei Kunden, die eine (wirklich) zufällige Zahl wählen, dieselbe Nummer erzeugen, bei etwa $10^{-37{,}5}$. Die Wahrscheinlichkeit, mit jeweils einem Spielschein fünf Wochen lang hintereinander sechs Richtige im Lotto zu haben, liegt dagegen bei $\binom{49}{6}^{-5} \approx 5 \cdot 10^{-35}$, also etwa um den Faktor sechzig höher. Zwei gleiche Seriennummern sind also praktisch auszuschließen, wenn auch theoretisch möglich.

Das System kann nur funktionieren, wenn in diesem Fall der zweite Geldschein mit derselben Seriennummer nicht anerkannt wird, so daß der zweite Kunde sein Geld verliert. Dies muß als eine zusätzliche Gebühr gesehen werden, die mit an Sicherheit grenzender Wahrscheinlichkeit nie fällig wird.

Da digitales Bargeld überdies nur in kleinen Stückelungen sinnvoll ist (Geldscheinen im Millionenwert wären auf Grund ihrer Seltenheit nicht wirklich anonym und würden, wegen der damit verbundenen Möglichkeiten zur Geldwäsche auch in keinem seriösen Wirtschaftssystem angeboten), wäre der theoretisch mögliche Verlust auch nicht sehr groß.

5) Größenordnung der Primzahlen: Das RSA-Verfahren ist nur sicher, wenn die Primzahlen p und q hinreichend groß gewählt werden. Als erstes müssen wir uns also die Frage stellen, was „hinreichend groß“ bedeuten soll.

Ein treu sorgender Staat bleibt seinen Bürgern auf eine so wichtige Frage natürlich keine Antwort schuldig: Zwar gibt es noch keine oberste Bundesbehörde für Primzahlen, aber das Bundesamt für Sicherheit in der Informationstechnik (BSI) und die Regulierungsbehörde für Telekommunikation und Post erarbeiten jedes Jahr ein gemeinsames Dokument mit dem Titel *Geeignete Kryptoalgorithmen gemäß §17 (2) SigV;* es ist zu finden unter `www.regtp.de` über die Linkkette *Elektronische Signatur → Veröffentlichungen → Geeignete Algorithmen.*

SigV steht für die aufgrund des Signaturgesetztes SigG erlassene Signaturverordnung; beide gemeinsam legen fest, daß elektronische Unterschriften in Deutschland grundsätzlich zulässig und rechtsgültig sind, sofern gewisse Bedingungen erfüllt sind. Zu diesen Bedingungen gehört unter anderem, daß das Verfahren und die Schlüssellänge gemeinsam einen „geeigneten Kryptoalgorihmus“ im Sinne der jeweils gültigen Veröffentlichung der Regulierungsbehörde ist.

Da Rechner immer schneller und leistungsfähiger werden und auch auf der theoretisch-algorithmischen Seite fast jedes Jahr kleinere oder größere Fortschritte zu verzeichnen sind, gelten die jeweiligen Empfehlungen nur für etwa sechs Jahre. Offiziell geht es dabei jeweils nur um die

Empfehlung geeigneter Algorithmen für elektronische Unterschriften und deren Schlüssellängen, aber wie die Entwicklung der letzten Jahre zeigte, drehen sich die Diskussionen, die zu den jeweiligen Empfehlungen führen, tatsächlich fast ausschließlich um die jeweils notwendige Schlüssellänge für RSA.

Natürlich hat in einer Demokratie bei so einer wichtigen Frage auch die Bevölkerung ein Mitspracherecht; deshalb beginnt das BSI jeweils zunächst einen Entwurf, zu dem es um Kommentare bittet; erst einige Monate später wird die endgültige Empfehlung verkündet und im Bundesanzeiger veröffentlicht.

Die interessierte Öffentlichkeit, von der die Kommentare zu den Entwürfen kommen, besteht naturgemäß in erster Linie aus Anbietern von Kryptographie-Software, und als erfahrene Experten für Datensicherheit wissen diese, daß ein Verfahren nur dann wirklich geeignet sein kann, wenn es die eigene Firma im Angebot hat. (Am geeignetsten sind natürlich die Verfahren, die keines der Konkurrenzunternehmen anbietet.)

Derzeit erhältliche Hardware-Implementierungen von RSA unterstützen typischerweise Schlüssellängen von bis zu 1024 Bit; größere Schlüssel sind vor allem in *public domain* Software wie PGP zu finden. Dies erklärt, warum es in den letzten Jahren recht lebhafte Diskussionen gab:

Bis Ende 2000 galten 768 Bit als ausreichende Größe für das Produkt N der beiden Primzahlen, aber die Richtlinien für 2000 legten fest, daß danach bis Mitte 2005 eine Mindestgröße von 1024 Bit erforderlich sei, danach bis Ende 2005 sogar 2048 Bit.

Anbieterproteste führten dazu, daß nach den Richtlinien von 2001 eine Schlüssellänge von 1024 dann doch noch bis Ende 2006 sicher war; die Schlüssellänge 2048 war nur noch „empfohlen“, also nicht mehr verbindlich.

Im April 2002 erschien der erste Entwurf für die 2002-Richtlinien; danach war für 2006 und 2007 nur eine Mindestlänge von 2048 Bit wirklich sicher. Einsprüche führten im September 2002 zu einem revidierten Entwurf, wonach 2006 doch noch 1024 Bit reichen, 2007 aber mindestens

1536 notwendig werden. Die Mindestlänge von 2048 Bit wurde wieder zur „Empfehlung“ zurückgestuft.

Am 2. Januar 2003 erschienen endlich die offiziellen Richtlinien des Jahres 2002; veröffentlich wurden sie am 11. März 2003 im Bundesanzeiger Nr. 48, S. 4202–4203. Danach reichen 1024 Bit auch noch bis Ende 2007, erst 2008 werden 1280 Bit erforderlich. 2048 Bit bleiben dringend empfohlen.

Die neuesten Richtlinien stammen vom 2. Januar 2004 (Bundesanzeiger Nr. 30 vom 13. Februar 2004, S. 2537–2538). Sie verlangen ab 2009 mindestens 1536 Bit und empfehlen weiterhin 2048.

Die beiden Primfaktoren p, q sollen zufällig und unabhängig voneinander erzeugt werden und aus einem Bereich stammen, in dem

$$\varepsilon_1 < |\log_2 p - \log_2 q| < \varepsilon_2$$

gilt. Als *Anhaltspunkte* werden dabei die Werte

$$\varepsilon_1 = 0{,}5 \quad \text{und} \quad \varepsilon_2 = 30$$

vorgeschlagen; ist p die kleinere der beiden Primzahlen, soll also

$$\sqrt{2}p < q < 2^{30}p \approx 10^9 p$$

sein, d.h. die beiden Primzahlen sollten zwar ungefähr dieselbe Größenordnung haben, aber nicht zu nahe beieinander liegen. Der Grund dafür ist ein von FERMAT entdecktes Faktorisierungsverfahren auf Grundlage der dritten binomischen Formel: Falls für eine Zahl N und eine natürliche Zahl y die Zahl $N + y^2$ eine Quadratzahl x^2 ist, ist $N = x^2 - y^2 = (x + y)(x - y)$, womit zwei Faktoren gefunden sind. Probiert man alle kleinen natürlichen Zahlen y systematisch durch, führt dieses Verfahren offensichtlich umos schneller zum Erfolg, je näher die beiden Faktoren von N beieinander liegen.

e) Das Verfahren von Diffie und Hellman

Kurz nach der Veröffentlichung des RSA-Algorithmus fanden auch DIFFIE und HELLMAN ein Verfahren, das im Gegensatz zu RSA sogar ganz ohne vorvereinbarte Schlüssel auskommt: Zwei Personen vereinbaren

über eine unsichere Leitung einen Schlüssel, den anschließend nur sie kennen.

Ausgangspunkt ist wieder das Potenzieren im Körper $\mathbb{F}_p$; hier betrachten wir aber die Exponentialfunktion $x \mapsto a^x$ zu einer geeigneten Basis a. Ihre Umkehrfunktion bezeichnet man als *Index* oder *diskreten Logarithmus* zur Basis a:

In $\mathbb{R}$ ist der Logarithmus zur Basis a die Umkehrfunktion der Funktion $x \mapsto a^x$, und genauso definieren wir ihn auch für endliche Körper:

$$y = a^x \Longrightarrow x = \log_a y\,.$$

Trotz dieser formalen Übereinstimmung gibt es es allerdings große Unterschiede zwischen reellen Logarithmen und ihren Analoga in endlichen Körpern: Während reelle Logarithmen sanft ansteigende stetige Funktionen sind, die man leicht mit beliebig guter Genauigkeit annähern kann, sieht der diskrete Logarithmus typischerweise so aus, wie es in der Abbildung zu sehen ist. Auch ist im Reellen der Logarithmus zur Basis $a > 1$ für jede positive Zahl definiert; in endlichen Körpern ist es viel schwerer zu entscheiden, ob ein bestimmter Logarithmus existiert: Modulo sieben etwa sind 2, 4 und 1 die einzigen Zweierpotenzen, so daß 3, 5 und 6 keine Zweierlogarithmen haben. Ein Satz aus der Algebra besagt allerdings, daß es stets Elemente a gibt, für die a^x jeden Wert außer der Null annimmt, die sogenannten primitiven Wurzeln. In $\mathbb{F}_7$ wären dies etwa drei und fünf.

Die Berechnung der Potenzfunktion durch sukzessives Quadrieren und Multiplizieren ist auch in endlichen Körpern einfach, für ihre Umkehrfunktion, den diskreten Logarithmus gibt es aber derzeit nur deutlich schlechtere Verfahren. Die derzeit besten Verfahren zur Berechnung von diskreten Logarithmen in Körpern mit N Elementen erfordern etwa denselben Aufwand wie die Faktorisierung eines RSA-Moduls der Größenordnung N. Diese Diskrepanz zwischen Potenzfunktion und Logarithmen kann kryptologisch ausgenutzt werden.

Als Körper verwendet man entweder Körper von Zweipotenzordnung, die wir weiter unten betrachten werden, oder Körper von Primzahlordnung. Da es für viele interessante Körper von Zweipotenzordnung

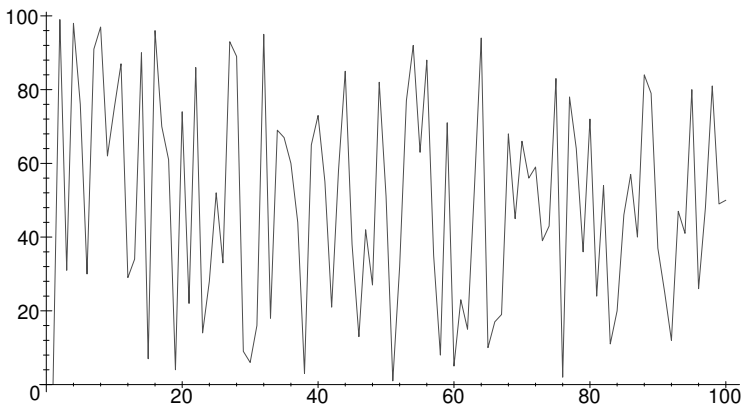


Abb. 12: Die Funktion $\log_{51} x$ in $\mathbb{F}_{101}$

bereits Chips gibt, die dort diskrete Logarithmen berechnen, dürften Körper von Primzahlordnung bei ungefähr gleicher Elementanzahl wohl etwas sicherer sein: Es gibt einfach viel mehr Primzahlen als Zweierpotenzen, und jeder Fall erfordert einen neue Hardwareentwurf. Falls man die Primzahlen hinreichend häufig wechselt, dürfte sich dieser Aufwand für kaum einen Gegner lohnen.

Da Körper von Primzahlordnung auch einfacher sind als solche von Primzahlpotenzordnung, wollen wir uns zunächst auf diese beschränken; die spätere Übertragung des Algorithmus auf Körper von Zweipotenzordnung sollte dem Leser keine Schwierigkeiten machen.

Beim DIFFIE-HELLMAN-Verfahren, dem ältesten auf der Grundlage diskreter Logarithmen, geht es wie gesagt darum, daß zwei Teilnehmer, die weder über gemeinsame Schlüsselinformation noch über eine sichere Leitung verfügen, einen Schlüssel vereinbaren wollen.

Dazu einigen sie sich zunächst (über die unsichere Leitung) auf eine Primzahl p und eine natürliche Zahl a derart, daß die Potenzfunktion $x \mapsto a^x$ möglichst viele Werte annimmt.

Als nächstes wählt Teilnehmer A eine Zufallszahl $x < p$ und B entsprechend $y < p$; A schickt $u = a^x \bmod p$ an B und erhält dafür

$v = a^y \bmod p$.

Sodann berechnet A die Zahl

$$v^x \bmod p = \left(a^y\right)^x \bmod p = a^{xy} \bmod p$$

und B entsprechend

$$u^y \bmod p = \left(a^x\right)^y \bmod p = a^{xy} \bmod p\,;$$

beide haben also auf verschiedene Weise dieselbe Zahl berechnet, die sie nun als Schlüssel in einem klassischen Kryptosystem verwenden können: Beispielsweise könnten die letzten 128, 196 oder 256 Bit der Zahl als AES-Schlüssel dienen. (Den *Advanced Encryption Standard* werden wir weiter unten kennenlernen.)

Ein Gegner, der den Datenaustausch abgehört hat, kennt die Zahlen p, a, u und v; um $a^{xy} \bmod p$ zu finden, muß er den diskreten Logarithmus von u oder v berechnen.

Mit den besten heute bekannten Algorithmen ist die möglich, wenn p eine Primzahl von bis zu etwa 512 Bit ist, also ungefähr 155 Dezimalstellen hat; auch in diesem Fall dauert die Berechnung allerdings selbst bei massiver Parallelisierung über das Internet mehrere Monate, gefolgt von einer Schlußrechnung auf einem Supercomputer.

Da diskrete Logarithmen auch für die in Deutschland rechtlich bindenden digitalen Unterschriften verwendet wird, befaßt sich die Regulierungsbehörde für Telekommunikation und Post in ihrem Bericht über sichere Kryptoverfahren auch damit; bislang hat sie für die Länge der Primzahl noch immer dieselbe Mindestlänge vorgeschrieben wie für einen RSA-Modul.

Natürlich gibt es keine Garantie, daß kein Gegner mit einem besseren als den bislang bekannten Verfahren diskrete Logarithmen oder Faktorisierungen auch in weitaus größeren Körpern berechnen kann. Dazu bräuchte er allerdings einen Durchbruch entweder auf der mathematischen oder auf der technischen Seite, für den weit und breit keine Grundlage zu sehen ist.

Falls sich allerdings die sogenannten *Quantencomputer* realisieren lassen, werden alle heute bekannten Verfahren der Kryptographie mit

öffentlichen Schlüsseln, egal ob mit diskreten Logarithmen, RSA oder elliptischen Kurven, unsicher sein. Bislang können Quantencomputer kaum mit drei Bit rechnen, und nicht alle Experten sind davon überzeugt, daß es je welche geben wird, die mit mehreren Tausend Bit rechnen können.

f) Körper von Zweipotenzordnung

Heutige Computer sind nicht für das Rechnen mit sechshundertstelligen Zahlen optimiert, sondern für den Umgang mit Bits und Bytes. Es liegt daher nahe, auch nach Körpern zu suchen, die dies ausnutzen können. Wie wir bald sehen werden, lassen sich in der Tat alle Vektorräume $\mathbb{F}_2^n$ zu Körpern machen können; für $n = 8$ spielt das beispielsweise eine große Rolle für die Fehlerkorrektur von CDs sowie für den neuen Kryptographiestandard AES.

Beginnen wir mit dem einfachsten Fall $\mathbb{F}_2^2$! Wir wissen schon, wie $\mathbb{R}^2$ zum Körper gemacht werden kann: Wir wählen eine Basis $\{1, i\}$ und müssen dann nur noch festlegen, was i^2 sein soll.

Entsprechend können wir auch für $\mathbb{F}_2^2$ eine Basis $\{1, \alpha\}$ wählen; dann läßt sich jedes Element von $\mathbb{F}_2^2$ schreiben als $a + b\alpha$. Da es nur viel Elemente gibt, können wir diese leicht explizit angeben: Es sind

$$0, \quad 1, \quad \alpha \quad \text{und} \quad 1 + \alpha\,.$$

Die Addition dieser Elemente ist klar: Schließlich haben wir bereits einen Vektorraum.

Zur Definition der Multiplikation hatten wir bei der Konstruktion von $\mathbb{C}$ festgelegt, daß $i^2 = -1$ sein sollte, d.h. also gleich einem Element, das in $\mathbb{R}$ kein Quadrat ist. Ein solches Element gibt es in $\mathbb{F}_2$ nicht: Jedes Element ist sein eigenes Quadrat. Daher muß entweder $\alpha^2 = \alpha$ oder $\alpha^2 = 1 + \alpha$ sein.

Wäre $\alpha^2 = \alpha$, so wäre $\alpha(\alpha - 1) = 0$, d.h. $\alpha = 0$ oder $\alpha = 1$, was wir natürlich nicht wollen. Also müssen wir

$$\alpha^2 = \alpha + 1$$

setzen. Damit ist dann alles klar, und wir erhalten die folgende Additions- und Multiplikationstafel, die uns insbesondere auch zeigen, daß wir tatsächlich einen Körper konstruiert haben:

+	0	1	α	$1+\alpha$
0	0	1	α	$1+\alpha$
1	1	0	$1+\alpha$	α
α	α	$1+\alpha$	0	1
$1+\alpha$	$1+\alpha$	α	1	0

und

·	0	1	α	$1+\alpha$
0	0	0	0	0
1	0	1	α	$1+\alpha$
α	0	α	$1+\alpha$	1
$1+\alpha$	0	$1+\alpha$	1	α

Dieser Körper wird üblicherweise mit $\mathbb{F}_4$ bezeichnet: Das $\mathbb{F}$ steht für *finite,* und vier ist ist die Anzahl der Elemente.

Allgemein bezeichnet man einen endlichen Körper mit q Elementen, so es einen gibt, als $\mathbb{F}_q$; in einigen Büchern auch als GF(q), wobei GF für GALOIS *field* steht nach dem französischen Mathematiker EVARISTE GALOIS (1811–1832) und dem englischen Wort *field* für *Körper.*

Man kann zeigen, daß es genau dann einen solchen Körper gibt, wenn q eine Primzahlpotenz ist, und daß dieser Körper dann bis auf Isomorphie eindeutig bestimmt ist.

Uns interessiert vor allem der Fall, daß $q = 2^n$ eine Zweierpotenz ist. Die Addition von $\mathbb{F}_q$ ist dann die Vektoraddition in $\mathbb{F}_2^n$, und genau wie oben geht es darum, eine Multiplikation zu definieren.

Der einfachste Weg dorthin führt über Polynome: Wir identifizieren den ersten Vektor der Standardbasis mit der Eins von $\mathbb{F}_{2^n} = \mathbb{F}_2^n$, bezeichnen

den zweiten als α und definieren die α-Potenzen bis zur $(n-1)$-ten als die weiteren Basisvektoren:

$$1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \alpha = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \alpha^2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \alpha^{n-1} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Damit läßt sich jedes Element von $\mathbb{F}_{2^n}$ als Polynom

$$c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_{n-1}\alpha^{n-1}$$

schreiben mit $c_i \in \mathbb{F}_2$, und wir können Produkte via Poly-nommultiplikation definieren, sobald wir wissen, was die höheren Potenzen von α sind.

Tatsächlich reicht es bereits, wenn wir nur die Potenz α^n kennen: Diese muß in der Form

$$\alpha^n = p_0 + p_1\alpha + p_2\alpha^2 + \cdots + p_{n-1}\alpha^{n-1}$$

mit $p_i \in \mathbb{F}_2$ darstellbar sein, und sobald wir die Koeffizienten p_i kennen, können wir rekursiv auch alle weiteren α-Potenzen ausrechnen: Beispielsweise ist

$$\begin{aligned} \alpha^{n+1} = \alpha \cdot \alpha^n &= p_0\alpha + p_1\alpha^2 + p_2\alpha^3 + \cdots + p_{n-2}\alpha^{n-2} + p_{n-1}\alpha^n \\ &= p_0\alpha + p_1\alpha^2 + p_2\alpha^3 + \cdots + p_{n-2}\alpha^{n-1} \\ &\quad + p_{n-1}\left(p_0 + p_1\alpha + p_2\alpha^2 + \cdots + p_{n-1}\alpha^{n-1}\right) \\ &= p_{n-1}p_0 + (p_{n-1}p_1 + p_2)\alpha + (p_{n-1}p_2 + p_3)\alpha^2 + \cdots \\ &\quad + (p_{n-1}p_{n-2} + p_{n-1})\alpha^{n-2} + (p_{n-1}^2 + p_n)\alpha^{n-1}, \end{aligned}$$

und entsprechend geht es weiter für die höheren Potenzen.

Wie wir schon beim Körper mit vier Elementen gesehen haben, können wir die Koeffizienten p_i nicht beliebig aus $\mathbb{F}_2$ wählen; nur in einem Fall ergab sich dort wirklich ein Körper.

Um zu sehen, welche Bedingungen wir an die p_i stellen müssen, nehmen wir an, wir hätten bereits Koeffizienten gefunden, für die sich ein

Körper $\mathbb{F}_{2^n}$ ergibt, und untersuchen, was wir dann über die p_i aussagen können.

Wir können die Gleichung

$$\alpha^n = p_0 + p_1\alpha + p_2\alpha^2 + \cdots + p_{n-1}\alpha^{n-1}$$

auch so auffassen, daß α eine Nullstelle des Polynoms

$$\begin{aligned} P(x) &= x^n - p_0 - p_1 x - p_2 x^2 - \cdots - p_{n-1}x^{n-1} \\ &= x^n + p_0 + p_1 x + p_2 x^2 + \cdots + p_{n-1}x^{n-1} \end{aligned}$$

im Körper $\mathbb{F}_{2^n}$ sein soll. (Das zweite Gleichheitszeichen kommt daher, daß es beim Rechnen im Körper $\mathbb{F}_2$ und in den Vektorräumen $\mathbb{F}_2^n$ keinen Unterschied gibt zwischen *plus* und *minus:* Für jeden Vektor $\vec{v} \in \mathbb{F}_2^n$ ist $\vec{v} + \vec{v} = \vec{0}$.)

Wenn wir nun ein Element von $\mathbb{F}_{2^n}$ als Polynom in α schreiben, ist diese Darstellung offensichtlich nicht eindeutig, denn beispielsweise ist

$$f(\alpha) = f(\alpha) + P(\alpha) = (f + P)(\alpha)$$

und allgemeiner gilt sogar für *jedes* Polynom g mit Koeffizienten in $\mathbb{F}_2$, daß

$$(f + g \cdot P)(\alpha) = f(\alpha) + g(\alpha) \cdot P(\alpha) = 0$$

ist. Offensichtlich ist $f(\alpha) = h(\alpha)$, wann immer das Polynom $f - h$ durch P teilbar ist.

Dies liefert einen neuen und schnelleren Zugang zur Multiplikation in $\mathbb{F}_{2^n}$: Um das Produkt zweier Elemente $f(\alpha)$ und $g(\alpha)$ auszurechnen, berechnen wir das Produktpolynom $f \cdot g$ und dividieren es mit Rest durch P, d.h.

$$(f \cdot g) : P = q \quad \text{Rest } h \qquad \text{oder} \qquad f \cdot g = q \cdot P + r\,.$$

Dann ist

$$(f \cdot g)(\alpha) = r(\alpha)\,,$$

und da r ein Polynom vom Grad höchstens $n - 1$ ist, kann $r(\alpha)$ direkt mit einem Vektor aus $\mathbb{F}_2^n$ identifiziert werden.

Rechnen wir etwa im Fall $n = 3$ mit dem Polynom

$$P = x^3 + x + 1\,,$$

so wird das Produkt der Vektoren

$$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

folgendermaßen bestimmt: Die beiden Vektoren lassen sich als Linearkombination der Potenzen von α schreiben als

$$1 + 0 \cdot \alpha + 1 \cdot \alpha^2 = 1 + \alpha^2 \quad \text{und} \quad 0 + 1 \cdot \alpha + 1 \cdot \alpha^2 = \alpha + \alpha^2\,;$$

das Produkt der beiden zugehörigen Polynome

$$f = 1 + x^2 \quad \text{und} \quad g = x + x^2 \quad \text{ist} \quad x + x^2 + x^3 + x^4\,.$$

Division durch P ergibt

$$(x^4 + x^3 + x^2 + x) : (x^3 + x + 1) = x + 1 \quad \text{Rest } x + 1\,,$$

d.h.

$$(1 + \alpha^2)(\alpha + \alpha^2) = 1 + \alpha$$

oder

$$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}\,.$$

Falls man das Polynom P als Produkt zweier Polynome f und g schreiben kann, die beide positiven Grad haben, haben beide insbesondere auch höchstens Grad $n - 1$, definieren also nichtverschwindende Elemente $f(\alpha)$ und $g(\alpha)$ aus $\mathbb{F}_{2^n}$. Deren Produkt ist aber $P(\alpha) = 0$, was in einem Körper natürlich nicht vorkommen darf. Damit haben wir eine erste Bedingung an P gefunden: P muß *irreduzibel* sein im Sinne der folgenden Definition:

Definition: Ein nichtkonstantes Polynom $P \in k[x]$ mit Koeffizienten aus einem Körper k heißt *reduzibel,* wenn es zwei nichtkonstante Polynome $f, g \in k[x]$ gibt, so daß $P = f \cdot g$ ist. Andernfalls heißt P *irreduzibel* (über k).

(Der Zusatz *über k* ist notwendig: Beispielsweise ist x^2+1 irreduzibel über $\mathbb{R}$, aber reduzibel über $\mathbb{C}$, denn dort ist $(x^2+1)=(x+i)(x-i)$. Da meist klar ist, über welchem Körper man arbeitet, wird der Zusatz aber oft weggelassen: Bei uns etwa geht es im Augenblick ausschließlich um Polynome über $\mathbb{F}_2$, so daß dieser Körper nicht ständig erwähnt werden muß.)

Wenn wir uns noch einmal die Konstruktion des Körpers mit vier Elementen anschauen, sehen wir, daß Irreduzibilität zumindest dort auch reicht: Von den vier Polynomen zweiten Grades über $\mathbb{F}_2$ ist genau eines irreduzibel, nämlich das, mit dem wir den Körper $\mathbb{F}_4$ definiert haben:

Ansatz für α^2	Polynom	Problem
$\alpha^2=0$	$f=x^2=x\cdot x$	$\alpha\cdot\alpha=0$
$\alpha^2=1$	$f=x^2+1=(x+1)\cdot(x+1)$	$(\alpha+1)(\alpha+1)=0$
$\alpha^2=\alpha$	$f=x^2+x=x(x+1)$	$\alpha\cdot(\alpha+1)=0$
$\alpha^2=\alpha+1$	$f=x^2+x+1$	*keine Probleme*

Tatsächlich reicht die Irreduzibilität von P *immer* zur Definition eines Körpers; damit wir das zeigen können, müssen wir aber zunächst den EUKLIDischen Algorithmus auf Polynome verallgemeinern.

g) Der Euklidische Algorithmus für Polynome

Dazu sei k ein Körper, z.B. der Körper $\mathbb{F}_2$ mit zwei Elementen; außerdem seien

$$A=x^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$$

und

$$B=x^m+b_{m-1}x^{m-1}+\cdots+b_1x+b_0$$

Polynome mit Koeffizienten a_i, b_i aus k; wir bezeichnen

$$n=\deg A \quad \text{und} \quad m=\deg B$$

als die Grade von A und B.

Dann läßt sich das Polynom A mit Rest durch B dividieren, d.h. man kann Polynome Q, R bestimmen, für die

$$A=QB+R \quad \text{ist mit} \quad \deg R<\deg B\,.$$

Mit dieser Division lassen sich sowohl der gewöhnliche als auch der erweiterte EUKLIDische Algorithmus sofort verallgemeinern auf Polynome; da der Grad von R kleiner ist als der von B und Grade als nichtnegative ganze Zahlen nicht unbegrenzt kleiner werden können, folgt daß der Algorithmus auch für Polynome stets nach endlich vielen Schritten endet.

Das Ergebnis kann allerdings in manchen Fällen unerwartet ausfallen: Betrachten wir etwa über dem Körper $\mathbb{Q}$ der rationalen Zahlen die beiden Polynome

$$P=X^8+X^6-3X^4-3X^3+8X^2+2X-5$$

und

$$Q=3X^6+5X^4-4X^2-9X+21\,.$$

Division von P durch Q führt auf den Quotienten $X^2/3-2/9$ und Divisionsrest

$$R_2=-\frac{5}{9}X^4+\frac{1}{9}X^2-\frac{1}{3}\,.$$

Division von Q durch R_2 ergibt

$$R_3=-\frac{117}{25}X^2-9X+\frac{441}{25}\,,$$

bei der Division von R_2 durch R_3 bleibt Rest

$$R_4=\frac{233150}{6591}X-\frac{102500}{2197}\,,$$

und bei der letzten Division verbleibt als Rest der ggT

$$R_5=\frac{1288744821}{543589225}\,.$$

Da beide Ausgangspolynome ganzzahlige Koeffizienten haben, erscheint ein ggT mit einem so großen Nenner seltsam. In der Tat ist jedes Polynom durch jede von Null verschiedene Konstante teilbar; ist also ein Polynom P Teiler eines Polynoms Q, so ist auch jedes von Null verschiedene skalare Vielfache von P Teiler von Q. Somit können wir hier nicht sinnvoll von *dem* größten gemeinsamen Teiler zweier Polynome reden: Jede von Null verschiedene Konstante ist ein ggT. Meist sagt man in einem solchem Fall, „der" ggT sei Eins. Die Computeralgebra kennt Modifikationen des EUKLIDischen Algorithmus, mit denen sich ohne den Umweg über Brüche stets ein möglichst einfacher ggT berechnen läßt.

Wir haben bislang noch nicht definiert, wann ein Polynom *größer* sein soll als ein anderes: Bei zwei natürlichen Zahlen ist klar, welche größer ist, aber schon bei reellen Polynomen ist alles andere als klar, ob etwa $x+2$ größer sein soll als $2x+1$ oder umgekehrt. Wir werden dieses Problem ignorieren und einfach sagen, P sei *ein* größter gemeinsamer

Teiler von A und B, wenn P ein gemeinsamer Teiler ist und jeder andere gemeinsamen Teiler ein Teiler von P ist.

Der größte gemeinsame Teiler, den uns der EUKLIDISCHE Algorithmus für Polynome liefert, hat diese Eigenschaft, denn da dieser ggT als Linearkombination von A und B geschrieben werden kann, muß jedes Polynom, das sowohl A als auch B teilt, auch den ggT teilen.

Problematischer ist, daß es viele solche größten gemeinsamen Teiler geben kann: Zumindest jedes von Null verschiedene skalare Vielfache eines ggT ist selbst einer. Zum Glück ist das aber auch schon alles, was passieren kann: Sind nämlich P und Q zwei größte gemeinsame Teiler von A und B, so muß nach Definition P ein Teiler von Q sein und umgekehrt. Da der Grad eines Teilers stets kleiner oder gleich dem des Polynoms ist, haben die beiden also insbesondere denselben Grad, und ihr Quotient, egal in welcher Reihenfolge, hat Grad Null und ist somit eine Konstante.

Der größte gemeinsame Teiler zweier Polynome über einem Körper ist also eindeutig bis auf Multiplikation mit einer nichtverschwindenden Konstanten; diese Konstante kann nach Belieben gewählt werden und wird meist so gewählt, daß das Ergebnis in irgendeinem Sinne einfach wird.

Auf das obige Beispiel angewendet heißt das, daß mit

$$R_5 = \frac{1288744821}{543589225}$$

auch eins ein ggT von A und B ist und man daher im allgemeinen sagen würde, „der“ ggT von A und B sei eins. Es ist ein wohlbekanntes (und umgehbares) Problem der Computeralgebra, daß der EUKLIDische Algorithmus diese einfache Lösung in einer so komplizierten Form liefert; da uns vor allem Polynome über endlichen Körpern interessieren, braucht uns das nicht weiter zu kümmern.

Kehren wir zurück zum Ausgangsproblem: Wir wollen den Vektorraum $\mathbb{F}_2^n$ zu einem Körper machen. Da es in $\mathbb{F}_2$ genau ein von Null verschiedenes Element gibt, spielt die obige Diskussion hier keine Rolle: Für Polynome über $\mathbb{F}_2$ existiert *der* ggT. Trotzdem war diese Diskussion nicht umsonst, denn erstens werden wir im nächsten Kapitel im Zusammenhang mit der Integration rationaler Funktionen den EUKLIDischen

Algorithmus auch auf reelle Polynome anwenden, und zweitens sei zumindest kurz erwähnt, daß die folgende Konstruktion auch für eine beliebige Primzahl p Körper mit p^n Elementen liefert. Sie werden allerdings in der Informationstechnik nur selten benutzt: Dort interessieren praktisch nur die Körper $\mathbb{F}_{2^n}$ und die Körper $\mathbb{F}_p$, denn das Rechnen in $\mathbb{F}_{p^n}$ ist umständlicher als das Rechnen in einem Körper $\mathbb{F}_q$ mit einer Primzahl q der Größenordnung p^n und bietet für $p \neq 2$ keinerlei Vorteile. Lediglich für $p = 2$, wo die Vektorraumstruktur von $\mathbb{F}_2^N$ so gut an die heutige Computer-Hardware angepaßt wird, bieten Körper von Zweierpotenzordnung oft (wenn auch keinesfalls immer!) Vorteile über Körper von Primzahlordnung.

In Abschnitt *e)* hatten wir die ganzen Zahlen modulo p zu einem Körper gemacht; der einzige nichttriviale Schritt dabei war die Existenz des multiplikativen Inversen, die wir aus der linearen Kombinierbarkeit des ggT folgerten und daraus, daß der ggT einer Zahl mit einer Primzahl gleich eins ist, falls die Zahl kein Vielfaches der Primzahl ist.

Genauso wollen wir jetzt Körper definieren, indem wir Polynome über einem festen Körper k modulo einem vorgegebenen Polynom P betrachten: Für ein beliebiges Polynom A über k ist A mod P gleich dem Rest bei der Division von A durch P.

Falls A kleineren Grad als P hat, ist natürlich einfach A mod $P = A$; zum konkreten Rechnen können wir daher ausgehen vom Vektorraum V aller Polynome vom Grad höchstens d, wobei $d + 1$ der Grad von P ist. Die Addition ist die gewöhnliche Addition von Polynomen, das Nullpolynom ist Neutralelement, und $-A$ ist invers zu A.

Das Produkt AB zweier Polynome $A, B \in V$ kann größeren Grad als d haben; wir setzen daher

$$A \odot B = AB \bmod P\,;$$

dies ist ein Polynom vom Grad höchstens d, und es ist klar, daß die so definierte Multiplikation kommutativ und assoziativ ist und das Distributivgesetz erfüllt. Das konstante Polynom 1 ist Neutralelement auch bezüglich dieser Multiplikation.

Ein inverses Polynom zu A ist ein Polynom B, für das $A \odot B = 1$ ist, d.h.

$$AB = 1 + CP \quad \text{oder} \quad AB + CP = 1$$

für ein geeignetes Polynom C. Zu vorgegebenen Polynomen A und P gibt es solche Polynome B und C genau dann, wenn der ggT von A und P gleich eins ist; alsdann lassen sich B und C nach dem EUKLIDischen Algorithmus berechnen.

Wenn wir möchten, daß jedes Polynom A, dessen Grad kleiner als deg P ist, ein Inverses hat, müssen wir sicherstellen, daß A und P immer teilerfremd sind; dies ist offensichtlich genau dann der Fall, wenn P keinen nichttrivialen Teiler hat, also irreduzibel ist.

Falls es ein irreduzibles Polynom P vom Grad n mit Koeffizienten aus k gibt, läßt sich der Vektorraum k^n also zu einem Körper machen, indem wir ein n-tupel $(a_0, \ldots, a_{n-1})$ mit dem Polynom

$$a_{n-1}X^{n-1} + a_{n-2}X^{n-2} + \cdots + a_1X + a_0$$

identifizieren und die Multiplikation als Multiplikation von Polynomen modulo P erklären.

Betrachten wir noch einmal das altbekannte Beispiel der komplexen Zahlen: Für $n = 2$ gibt es irreduzible Polynome vom Grad n über $\mathbb{R}$, beispielsweise das Polynom $P = X^2 + 1$. Da

$$\begin{aligned}(a_1X + a_0)(b_1X + b_0) &= a_1b_1X^2 + (a_0b_1 + a_1b_0)X + a_0b_0 \\ &\equiv (a_0b_1 + a_1b_0)X + (a_0b_0 - a_1b_1) \bmod X^2 + 1\end{aligned}$$

ist, folgt $(a_0, a_1) \odot (b_0, b_1) = (a_0b_0 - a_1b_1, a_0b_1 + a_1b_0)$, wir erhalten also den Körper der komplexen Zahlen. Weitere Beispiele über $\mathbb{R}$ gibt es nicht, denn ein irreduzibles reelles Polynom muß entweder Grad eins oder Grad zwei haben, und da jedes irreduzible quadratische Polynom zwei konjugiert komplexe Nullstellen hat, entstehen dabei immer die komplexen Zahlen – lediglich die Basis über $\mathbb{R}$ ändert sich.

Über endlichen Körpern ist die Situation etwas komplizierter: Hier wissen wir nicht einmal, für welche n es überhaupt ein irreduzibles Polynom

N	Polynome vom Grad N	*davon irreduzibel*	*in Prozent*
2	4	1	25.0 %
3	8	2	25.0 %
4	16	3	18.8 %
5	32	6	18.8 %
6	64	9	14.1 %
7	128	18	14.1 %
8	256	30	11.7 %
9	512	56	10.9 %
10	1024	99	9.7 %
11	2048	186	9.1 %
12	4096	335	8.2 %
13	8192	630	7.7 %
14	16384	1161	7.1 %
15	32768	2182	6.7 %
16	65536	4080	6.2 %

vom Grad n gibt. Tatsächlich gibt es sogar ziemlich viele solche Polynome; die Tabelle zeigt deren Anzahl über $\mathbb{F}_2$ für $n \leq 16$.

Mit etwas mehr Algebra zeigt man leicht, daß es über jedem endlichen Körper irreduzible Polynome jedes beliebigen (positiven) Grads gibt und daß zwei solche Polynome *im wesentlichen* (d.h. bis auf Isomorphie) zum selben Körper führen. Wir wollen uns darauf beschränken, für den uns hauptsächlich interessierenden Fall des Körpers $\mathbb{F}_{256}$ konkrete Polynome zu betrachten: Konkretes Rechnen mit Bitfolgen setzt schließlich ohnehin immer ein konkretes Polynom voraus.

f) Der Körper mit 256 Elementen und CD-Fehlerkorrektur

Zunächst müssen wir diesen Körper definieren, d.h. ein irreduzibles Polynom vom Grad acht über $\mathbb{F}_2$ finden. Wie die obige Tabelle zeigt, haben wir dazu dreißig Möglichkeiten. Diese führen zwar, abstrakt betrachtet, alle auf denselben Körper, aber das praktische Rechnen in diesem Körper hängt natürlich stark von der Wahl des Polynoms ab. Insbesondere wird

die Geschwindigkeit umso höher, je weniger Terme das Polynom hat.

Dreizehn der dreißig Polynome bestehen aus sieben nichtverschwindenden Termen, die restlichen siebzehn aus fünf; Wir wählen natürlich eines der letzteren. Alle diese Polynome haben, wie jedes Polynom vom Grad acht über $\mathbb{F}_2$, den führenden Term X^8; danach folgen vier weitere Terme. Bei der Reduktion modulo einem solchen Polynom $P = X^8 + \mathit{Rest}$ benutzt man, daß dann

$$X^8 \equiv \mathit{Rest}, \qquad X^9 \equiv X \cdot \mathit{Rest}, \quad \ldots$$

ist; dies wird umso häufiger mehrfach angewandt werden müssen, je höheren Grad die Terme in *Rest* haben. Am effizientesten kann man also rechnen, wenn das Polynom *Rest* den kleinstmöglichen Grad hat, und wenn zudem auch noch die hinteren Terme von *Rest* möglichst kleinen Grad haben. Inspektion der siebzehn Polynome mit fünf Termen zeigt, daß das Polynom

$$x^8 + x^4 + x^3 + x + 1$$

in dieser Hinsicht optimal ist; es wird beim *Advanced Encryption Standard* AES verwendet, den wir im nächsten Abschnitt kurz betrachten werden. Für die Fehlerkorrektur auf CDs verwendet man das leicht verschiedene Polynom $x^8 + x^4 + x^3 + x^2 + 1$.

Überlegen wir uns zunächst, worum es bei dieser Fehlerkorrektur geht:

Bei der Fertigung einer CD läßt sich die Fehlerwahrscheinlichkeit pro Bit auf etwa eins zu einer Million herunterdrücken; da aber bei einer Audio-CD rund vier Millionen Bit pro Sekunde verarbeitet werden, treten trotzdem jede Menge Fehler auf, die teilweise verheerende Folgen haben können: Falls beispielsweise das Wort 0000 0000 0001 0101 versehentlich als 1000 0000 0100 0101 interpretiert wird, wird aus einem Pianissimo ein ohrenbetäubender Knacklaut von ca. 90 dB.

Nun sollten allerdings nicht nur Fertigungsfehler korrigiert werden, sondern zumindest in gewissem Umfang auch Fehler durch Fingerabdrücke, Staubkörner *usw.*

Da die Spur bei einer CD etwa einen halben Mikrometer breit ist und die einzelnen Pits zwischen $0{,}833\mu$ und $3{,}56\mu$ lang sind, wohingegen

ein Fingerabdruck bereits Linien mit einer Breite von 15μ erzeugt und eine beim Staubwischen übriggebliebene Baumwollfaser gar eine Breite von 150μ hat, ist klar, daß sich solche Fehler nicht im Bitbereich bewegen.

Dies wird auf einer CD zum einen dadurch berücksichtigt, daß man die Information nicht linear anordnet (die geraden Bytes werden gegenüber den ungeraden verzögert), zum anderen dadurch, daß man anstelle des Bits das Byte als grundlegende Einheit betrachtet, d.h. man arbeitet mit dem Körper $\mathbb{F}_{256}$.

Die Fehlerkorrektur arbeitet mit Prüfbytes, die (wie Paritätsbits) durch lineare Abbildungen definiert sind. Zu einem Vektor aus 24 Bytes werden in zwei Stufen insgesamt acht Prüfbytes berechnet, und zwar werden zunächst vier Prüfbytes angehängt derart, daß der entstehende Vektor aus $\mathbb{F}_{256}^{28}$ im Kern der linearen Abbildung

$$\varphi\colon \mathbb{F}_{256}^{28} \to \mathbb{F}_{256}^{4}; \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{27} \\ x_{28} \end{pmatrix} \mapsto \begin{pmatrix} \sum_{i=1}^{28} x_i \\ \sum_{i=1}^{28} \alpha^{28-i} x_i \\ \sum_{i=1}^{28} \alpha^{2(28-i)} x_i \\ \sum_{i=1}^{28} \alpha^{3(28-i)} x_i \end{pmatrix}$$

liegt, danach werden vier weitere Bytes angehängt derart, daß der entstehende Vektor aus $\mathbb{F}_{256}^{32}$ im Kern der linearen Abbildung

$$\psi\colon \mathbb{F}_{256}^{32} \to \mathbb{F}_{256}^{4}; \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{31} \\ x_{32} \end{pmatrix} \mapsto \begin{pmatrix} \sum_{i=1}^{32} x_i \\ \sum_{i=1}^{32} \alpha^{32-i} x_i \\ \sum_{i=1}^{32} \alpha^{2(32-i)} x_i \\ \sum_{i=1}^{32} \alpha^{3(32-i)} x_i \end{pmatrix}$$

liegt. $\alpha \in \mathbb{F}_{256}$ bezeichnet dabei wie üblich jenes Element, für das die Eins zusammen mit α bis α^7 eine $\mathbb{F}_2$-Basis von $\mathbb{F}_{256}$ ist, und das Nullstelle des definierenden Polynoms ist.

Durch Kombination dieser Prüfbytes mit einer geschickten (nichtlinearen) Anordnung der Bytes auf der Spirale lassen sich selbst Fehler einer Länge von etwa 4 000 Bit beheben – teils durch echte Korrektur,

teils durch bloße Fehlererkennung und Interpolation aus unverfälschten Daten. Versuche von Physikern der University of Maryland haben ergeben, daß eine CD eingebohrte Löcher mit einem Durchmesser von 0,8 mm problemlos verkraftet, und selbst ein Lochdurchmesser von 1,5 mm führt kaum zu Knackgeräuschen. Einzelheiten findet man unter `www.physics.umd.edu/deptinfo/facilities/lecdem/h4-67.htm` .

g) Der Körper mit 256 Elementen in der Kryptographie

Zwar lehnt es die Internationale Standardisierungsorganisation ISO ab, ein Kryptoverfahren zu standardisieren (Ein Grund dafür ist die dann befürchtete Bündelung krimineller Energie auf dieses Verfahren), aber das amerikanische Handelsministerium hat am 2. Januar 1997 die Suche nach einem Nachfolgealgorithmus für den nach heutigen Standards nicht mehr sicheren DES *(Data Encryption Standard)* international ausgeschrieben. Federführend für die Auswahl war das *National Institute of Standards and Technology* (NIST) in Gaithersburg, Maryland, das am 2. Oktober 2000 den Algorithmus Rijndael der beiden flämischen Kryptologen JOAN DAEMEN und VINCENT RIJMEN auswählte. (Als Aussprachehilfe für Personen, die kein Niederländisch, Flämisch, Surinamer oder Afrikaans sprechen, geben diese folgende englische Approximationen des Wortes „Rijndael“: „Reign Dahl“, „Rain Doll“ und „Rhine Dahl“.) Es steht zu erwarten, daß Rijndael mittelfristig auch außerhalb der USA zu *dem* Standardverfahren in der Kryptographie wird.

Grundidee sind, wie bei allen Kryptoverfahren, die beiden SHANNONschen Forderungen der *Diffusion* und *Konfusion:* Ersteres bedeutet, daß sich schon die Änderung eines einzigen Klartextbits an vielen, möglichst weit entfernten Stellen bemerkbar machen muß, das zweite bedeutet in erster Linie eine hohe Nichtlinearität der Verschlüsselungsabbildung, so daß diese ohne Kenntnis des Schlüssels nicht invertiert werden kann.

Nichtlinearität erreicht Rijndael durch die Abbildung $\mathbb{F}_{256} \to \mathbb{F}_{256}$, die die Null auf sich selbst abbildet und jedes andere Element auf sein multiplikatives Inverses. Über mehrere Runden hinweg wird diese Abbildung auf Byte-Ebene immer wieder mit linearen Abbildungen und Vektoradditionen auf $\mathbb{F}_{256}^4$ und Shift-Operationen auf $\mathbb{F}_{256}^{16}$ oder noch größeren

Vektorräumen kombiniert. Alle linearen Abbildungen sind $\mathbb{F}_{256}$-linear, was auf Bitebene noch einmal eines Konfusionseffekt hat. Die einzelnen Operationen hängen ab von einem Schlüsselvektor, der Element von $\mathbb{F}_{256}^{16}$, $\mathbb{F}_{256}^{24}$ oder $\mathbb{F}_{256}^{32}$ sein kann und somit 128, 192 oder 256 Bit lang ist. Einzelheiten findet man unter `http://www.rijndael.com` .

§3: Matrizen und lineare Gleichungssysteme

Die abstrakte Art und Weise, wie wir Vektoren und lineare Abbildungen bisher betrachtet haben, hat zwar den Vorteil, daß wir damit viele Probleme behandeln können, die nichts mit den gewohnten anschaulichen Vektoren zu tun haben; sie hat aber auch den Nachteil, daß wir bislang noch sehr wenige nichttriviale Beispiele konkret durchrechnen können. Dieser Paragraph soll die wichtigsten Hilfsmittel zum Rechnen in endlichdimensionalen Vektorräumen bereitstellen.

a) Abbildungsmatrizen

Basen sind nicht nur nützlich, um Vektoren darzustellen, sie können auch den Umgang mit linearen Abbildungen vereinfachen. Der Grund liegt im folgenden Lemma:

Lemma: V und W seien k-Vektorräume, und $\mathcal{B}$ sei eine Basis von V. Dann ist jede lineare Abbildung $\varphi: V \to W$ eindeutig bestimmt durch die Bilder $\varphi(\vec{b})$ der Basisvektoren $\vec{b} \in \mathcal{B}$. Umgekehrt läßt sich jede Abbildung $\varphi: \mathcal{B} \to W$ eindeutig fortsetzen zu einer linearen Abbildung $\varphi: V \to W$

Beweis: Jeder Vektor $\vec{v} \in V$ läßt sich in eindeutiger Weise als Linearkombination

$$\vec{v} = \lambda_1 \vec{b}_1 + \cdots + \lambda_n \vec{b}_n \quad \text{mit} \quad \vec{b}_1, \ldots, \vec{b}_n \in \mathcal{B}$$

darstellen, und für eine lineare Abbildung φ muß dann

$$\varphi(\vec{v}) = \lambda_1 \varphi(\vec{b}_1) + \cdots + \lambda_n \varphi(\vec{b}_n)$$

sein. ∎

Besonders nützlich ist dies im Fall endlichdimensionaler Vektorräume.

Sei etwa V ein m-dimensionaler k-Vektorraum und W ein n-dimensionaler; wir wählen Basen

$$\mathcal{B} = (\vec{b}_1, \ldots, \vec{b}_m) \quad \text{und} \quad \mathcal{C} = (\vec{c}_1, \ldots, \vec{c}_n)\,.$$

Eine lineare Abbildung $\varphi\colon V \to W$ ist, wie wir gerade gesehen haben, eindeutig bestimmt durch die Bilder der Basisvektoren $\vec{b}_j$; diese wiederum lassen sich als Linearkombinationen der Basisvektoren $\vec{c}_i$ schreiben:

$$\varphi(\vec{b}_j) = a_{1j}\vec{c}_1 + a_{2j}\vec{c}_2 + \cdots + a_{nj}\vec{c}_n \quad \text{mit} \quad a_{ij} \in k\,.$$

Somit ist φ bei gegebenen Basen eindeutig bestimmt durch die $n \cdot m$ Skalare a_{ij}. Wir fassen diese zusammen zu einer *Matrix:*

Definition: *a)* Eine $n \times m$-Matrix A über dem Körper k ist eine zweidimensionale Anordnung von Körperelementen $a_{ij} \in k$ in n Zeilen und m Spalten, d.h.

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix}\,.$$

b) Die Menge aller $n \times m$-Matrizen A über k bezeichnen wir mit $k^{n\times m}$.

Die Matrix A zur linearen Abbildung φ bezeichen wir als *Abbildungsmatrix* von φ bezüglich der (geordneten) Basen $\mathcal{B}$ und $\mathcal{C}$; für eine lineare Abbildung $\varphi\colon V \to V$, deren Zielraum gleich derUrbildmenge ist, wählen wir im allgemeinen nur eine Basis $\mathcal{B}$ von V, setzen also $\mathcal{C} = \mathcal{B}$, und reden dann von der Abbildungsmatrix bezüglich der Basis $\mathcal{B}$.

Wenn $\mathcal{B}$ und $\mathcal{C}$ vorgegeben sind, gibt es offensichtlich für jede Matrix $A \in k^{n\times m}$ eine lineare Abbildung $\varphi\colon V \to W$ mit A als Abbildungsmatrix, nämlich diejenige lineare Abbildung, für die

$$\varphi(\vec{b}_j) = a_{1j}\vec{c}_1 + a_{2j}\vec{c}_2 + \cdots + a_{nj}\vec{c}_n$$

ist. Bei gegebenen Basen entsprechen sich lineare Abbildungen und Matrizen also eineindeutig: Zu jeder linearen Abbildung gibt es *genau* eine Matrix und zu jeder Matrix *genau* eine lineare Abbildung.

Ein wesentlicher Punkt ist hier, daß es sich bei einer linearen Abbildung $\varphi\colon V \to W$ im allgemeinen um eine Abbildung zwischen *unendlichen* Mengen handelt und solche Abbildungen nur selten mit endlichem Aufwand beschrieben werden können. (Wie sieht etwa eine „allgemeine“ Abbildung $\varphi\colon \mathbb{R} \to \mathbb{R}$ aus?) Eine lineare Abbildung zwischen endlichdimensionalen Vektorräumen ist, wie wir gerade gesehen haben, nach Wahl von Basen durch die endlich vielen Einträge der Abbildungsmatrix eindeutig bestimmt und damit einer algorithmischen Behandlung zugänglich.

Matrizen als zweidimensionale Zahlenschemata sind natürlich erheblich älter als Vektorräume und lineare Abbildungen; erste Spuren aus dem zweiten vorchristlichen Jahrhundert finden sich bereits in den *Neun Büchern der Rechenkunst* 九章算術 aus der chinesischen Han-Dynastie. Rechenregeln für den Umgang mit Matrizen tauchen ab dem 16. Jahrhundert bei den verschiedensten Autoren auf.

Als Beispiel betrachten wir den Vektorraum V aller reeller Polynome vom Grad höchstens vier und den Vektorraum W aller reeller Polynome vom Grad höchstens drei zusammen mit der linearen Abbildung

$$\varphi\colon V \to W; \quad f \mapsto f'\,.$$

Bevor wir eine Abbildungsmatrix berechnen können, brauchen wir zunächst Basen der beiden Vektorräume, zum Beispiel die „üblichen“ Basen

$$\mathcal{B} = (1, X, X^2, X^3, X^4) \quad \text{und} \quad \mathcal{C} = (1, X, X^2, X^3)\,.$$

Dann ist

$$\begin{aligned} \varphi(1) &= 0 &&= 0\cdot 1 + 0\cdot X + 0\cdot X^2 + 0\cdot X^3 \\ \varphi(X) &= 1 &&= 1\cdot 1 + 0\cdot X + 0\cdot X^2 + 0\cdot X^3 \\ \varphi(X^2) &= 2X &&= 0\cdot 1 + 2\cdot X + 0\cdot X^2 + 0\cdot X^3 \\ \varphi(X^3) &= 3X^2 &&= 0\cdot 1 + 0\cdot X + 3\cdot X^2 + 0\cdot X^3 \quad \text{und} \\ \varphi(X^4) &= 4X^3 &&= 0\cdot 1 + 0\cdot X + 0\cdot X^2 + 4\cdot X^3\,, \end{aligned}$$

die Abbildungsmatrix ist also

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \in \mathbb{R}^{4\times 5}\,.$$

Hier wie auch im allgemeinen Fall stehen in den *Spalten* der Abbildungsmatrix die Koeffizienten der Bilder der Basisvektoren von V, ausgedrückt bezüglich der Basis von W.

b) Rechenregeln für Matrizen

Wir haben Matrizen eingeführt, um mit linearen Abbildungen konkret rechnen zu können; als erstes sollten wir uns dazu überlegen, wie man mit *Matrizen* rechnen kann.

Zu zwei $n \times m$-Matrizen

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \text{ und } B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix}$$

können wir deren Summe

$$A + B \underset{\text{def}}{=} \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1m}+b_{1m} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2m}+b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \dots & a_{nm}+b_{nm} \end{pmatrix}$$

definieren, und für einen Skalar $\lambda \in k$ auch das Produkt

$$\lambda A \underset{\text{def}}{=} \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1m} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{n1} & \lambda a_{n2} & \dots & \lambda a_{nm} \end{pmatrix} .$$

Das legt die Vermutung nahe, daß $k^{n\times m}$ mit diesen beiden Verknüpfungen ein Vektorraum sein könnte, und in der Tat gilt:

Lemma: $k^{n\times m}$ ist ein k-Vektorraum der Dimension nm.

Beweis: Eigentlich gibt es nichts zu beweisen, denn wir haben einfach die Elemente aus dem Vektorraum $k^{(nm)}$ anders hingeschrieben, ohne daß dabei an den Rechenoperationen etwas zu ändern. Für diejenigen, die den Umgang mit Vektorraumaxiomen und das Rechnen mit Matrizen

noch etwas üben wollen, sei aber trotzdem noch einmal ein ausführlicher Beweis gegeben:

Beide Rechenoperationen sind so definiert, daß, wenn wir den ij-Eintrag für sich alleine betrachten, dort die entsprechende Rechenoperation im Grundkörper k ausgeführt wird. Da für die Rechenoperationen im Grundkörper alle bei der Definition eines Vektorraums geforderten Rechenregeln gelten, gelten sie auch in $k^{n\times m}$.

Beispielsweise ist also

$$\begin{aligned} A + B &= \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1m}+b_{1m} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2m}+b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \dots & a_{nm}+b_{nm} \end{pmatrix} \\ &= \begin{pmatrix} b_{11}+a_{11} & b_{12}+a_{12} & \dots & b_{1m}+a_{1m} \\ b_{21}+a_{21} & b_{22}+a_{22} & \dots & b_{2m}+a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1}+a_{n1} & b_{n2}+a_{n2} & \dots & b_{nm}+a_{nm} \end{pmatrix} = B + A \end{aligned}$$

und

$$\begin{aligned} \lambda(\mu A) &= \lambda \begin{pmatrix} \mu a_{11} & \mu a_{12} & \dots & \mu a_{1m} \\ \mu a_{21} & \mu a_{22} & \dots & \mu a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mu a_{n1} & \mu a_{n2} & \dots & \mu a_{nm} \end{pmatrix} \\ &= \begin{pmatrix} \lambda(\mu a_{11}) & \lambda(\mu a_{12}) & \dots & \lambda(\mu a_{1m}) \\ \lambda(\mu a_{21}) & \lambda(\mu a_{22}) & \dots & \lambda(\mu a_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(\mu a_{n1}) & \lambda(\mu a_{n2}) & \dots & \lambda(\mu a_{nm}) \end{pmatrix} \\ &= \begin{pmatrix} (\lambda\mu) a_{11} & (\lambda\mu) a_{12} & \dots & (\lambda\mu) a_{1m} \\ (\lambda\mu) a_{21} & (\lambda\mu) a_{22} & \dots & (\lambda\mu) a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ (\lambda\mu) a_{n1} & (\lambda\mu) a_{n2} & \dots & (\lambda\mu) a_{nm} \end{pmatrix} = (\lambda\mu) A . \end{aligned}$$

Nullelement der Addition ist natürlich die Nullmatrix

$$\begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix},$$

deren sämtliche Einträge Null sind, und

$$-\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} = \begin{pmatrix} -a_{11} & -a_{12} & \dots & -a_{1m} \\ -a_{21} & -a_{22} & \dots & -a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \dots & -a_{nm} \end{pmatrix}.$$

Schließlich müssen wir uns noch überlegen, daß $k^{n\times m}$ die Dimension nm hat; wir wir am Ende von §1*h)* gesehen haben, ist das gleichbedeutend damit, daß es eine Basis aus nm Matrizen gibt.

Als eine solche Basis wählen wir die Menge aller Matrizen E_{ij}, die so definiert sind, daß E_{ij} an der Stelle ij eine Eins stehen hat und sonst lauter Nullen.

In $k^{4\times 5}$ wäre also etwa $E_{23} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$.

In einem beliebigen $k^{n\times m}$ läßt sich jede Matrix eindeutig als Linearkombination der E_{ij} schreiben, denn

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}E_{ij},$$

und ist

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{ij}E_{ij} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{nm} \end{pmatrix}$$

die Nullmatrix, so müssen offensichtlich alle λ_{ij} verschwinden, die E_{ij} sind also auch linear unabhängig. Damit bilden sie eine Basis von $k^{n\times m}$, und $\dim_k k^{n\times m} = nm$.

∎

Die Basis mit den Matrizen E_{ij} ist zwar sicherlich die einfachste Basis für den Vektorraum aller $n \times m$-Matrizen, aber nicht immer die beste: Matrizen bieten sich beispielsweise auch an, um digitalisierte Bilder darzustellen, und zumindest in digitalen Kameras oder Scannern entsteht das Bild wirklich als eine Matrix von Grauwerten *bzw.* als drei Matrizen von Farb- oder sonstigen Werten, dargestellt in der Basis aus den E_{ij}. Für die Übertragung oder Speicherung ist das aber selten optimal, da hier für jede Komponente der Basisdarstellung dieselbe Genauigkeit erforderlich ist. Daher werden die Bilder etwa für die Speicherung im JPEG-Format zunächst in eine andere Basis umgerechnet, bezüglich derer viele Koeffizienten nahe bei Null liegen. Bei der Diskretisierung und Quantisierung enstehen dann viele Koeffizienten, für die nur wenige oder gar keine Bit benötigt werden, was im Zusammenspiel mit anderen Verfahren wie *run length encoding* und HUFFMAN-Codierung zu Komprimierungsfaktoren um die zwanzig oder dreißig ohne nennenswerten Qualitätsverlust führt.

Auch für zwei lineare Abbildungen $\varphi, \psi: V \to W$ können wir eine Summe definieren, und für $\lambda \in k$ auch ein Produkt $\lambda\varphi$ durch

$$\varphi + \psi: \begin{cases} V \to W \\ \vec{v} \mapsto \varphi(\vec{v}) + \psi(\vec{v}) \end{cases} \quad \text{und} \quad \lambda\varphi: \begin{cases} V \to W \\ \vec{v} \mapsto \lambda\varphi(\vec{v}) \end{cases};$$

sind V und W endlichdimensional und sind A, B die Abbildungsmatrizen von φ, ψ, so hat $\varphi + \psi$ offenbar die Abbildungsmatrix $A + B$ und λA ist die von $\lambda\varphi$.

Auch hier ist klar, daß die sämtlichen linearen Abbildungen $V \to W$ einen Vektorraum bilden, da einfach für jeden Vektor $\vec{v} \in V$ die Vektorraumoperationen von W auf die Bildvektoren angewendet werden; dieser Vektorraum wir mit $\mathrm{Hom}_k(V, W)$ bezeichnet nach dem Wort *Homomorphismus,* das man gelegentlich anstelle von *lineare Abbildung* gebraucht.

Im endlichdimensionalen Fall hat $\mathrm{Hom}_k(V,W)$ wegen der eineindeutigen Entsprechung von linearen Abbildungen und Matrizen als Dimension das Produkt der Dimensionen von V und von W. Die Basismatrix $E_{ij} \in k^{n\times m}$ entspricht dabei bezüglich der Basen $\mathcal{B}$ von V und $\mathcal{C}$ von W jener linearen Abbildung, die alle Vektoren aus $\mathcal{B}$ mit Ausnahme des j-ten auf den Nullvektor abbildet; der j-te Basisvektor geht auf den i-ten Basisvektor aus $\mathcal{C}$. Bei den linearen Abbildungen $k^5 \to k^4$ etwa entspräche obige Matrix E_{23} der linearen Abbildung

$$k^5 \to k^4; \quad \begin{pmatrix} u \\ v \\ w \\ x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 0 \\ w \\ 0 \\ 0 \end{pmatrix} .$$

Lineare Abbildungen lassen sich nicht nur addieren und mit Skalaren multiplizieren; sie lassen sich auch, wie alle Abbildungen, hintereinanderausführen: Sind $\psi\colon U \to V$ und $\varphi\colon V \to W$ lineare Abbildungen, so ist auch

$$\varphi\circ\psi\colon U \to W; \quad \vec{v} \mapsto \varphi\big(\psi(\vec{v})\big)$$

eine lineare Abbildung.

Falls alle beteiligten Vektorräume endlichdimensional sind, können wir endliche Basen wählen; das seien etwa

$$\mathcal{B} = \big\{\vec{b}_1,\dots,\vec{b}_m\big\} \subset U, \quad \mathcal{C} = \big\{\vec{c}_1,\dots,\vec{c}_n\big\} \subset V$$

und

$$\mathcal{D} = \big\{\vec{d}_1,\dots,\vec{d}_p\big\} \subset W ,$$

d.h.

$$\dim_k U = m, \quad \dim_k V = n \quad \text{und} \quad \dim_k W = p .$$

Dann haben wir Abbildungsmatrizen $A \in k^{p\times n}$ von φ und $B \in k^{n\times m}$ von ψ; wir wollen die Abbildungsmatrix $C \in k^{p\times m}$ von $\varphi\circ\psi\colon U \to W$ berechnen.

Nach Definition der Abbildungsmatrizen $A = (a_{ij})$ von φ und $B = (b_{j\ell})$ von ψ ist

$$\varphi\circ\psi(\vec{b}_i) = \varphi\big(\psi(\vec{b}_i)\big) = \varphi\left(\sum_{j=1}^{n} b_{ji}\vec{c}_j\right) = \sum_{j=1}^{n} b_{ji}\varphi(\vec{c}_j)$$

$$= \sum_{j=1}^{n} b_{ji}\sum_{\ell=1}^{p} a_{\ell j}\vec{d}_\ell = \sum_{\ell=1}^{p}\left(\sum_{j=1}^{n} a_{\ell j}b_{ji}\right)\vec{d}_\ell .$$

Für die Abbildungsmatrix $C = (c_{i\ell})$ von $\varphi\circ\psi$ ist nach Definition

$$\varphi\circ\psi(\vec{b}_i) = \sum_{\ell=1}^{p} c_{\ell i}\vec{d}_\ell ,$$

also ist

$$c_{\ell i} = \sum_{j=1}^{n} a_{\ell j}b_{ji} .$$

Definition: Für zwei Matrizen $A = (a_{i\ell}) \in k^{p\times n}$ und $B = (b_{\ell j}) \in k^{n\times m}$ bezeichnen wir die Matrix $C = (c_{ij}) \in k^{p\times m}$ mit

$$c_{ij} = \sum_{\ell=1}^{n} a_{i\ell}b_{\ell j}$$

als *Produktmatrix* $C = AB$ von A und B.

Für praktische Rechnungen empfiehlt es sich als Eselsbrücke, den zweiten Faktor des Produkts höher zu stellen nach dem Schema

$$\begin{array}{c} \\ \\ \end{array}\begin{pmatrix} \cdots & b_{1j} & \cdots \\ & \vdots & \\ \cdots & b_{mj} & \cdots \end{pmatrix}$$

$$\begin{pmatrix} \vdots & & \vdots \\ a_{i1} & \cdots & a_{im} \\ \vdots & & \vdots \end{pmatrix} \quad \begin{pmatrix} & \vdots & \\ \cdots & \sum_{\ell=1}^{m} a_{i\ell}b_{\ell j} & \cdots \\ & \vdots & \end{pmatrix} \quad ;$$

dadurch behält man den Überblick, welcher Rechenschritt jeweils als nächster auszuführen ist.

Im Gegensatz zu den meisten bislang aufgetretenen Produkten ist dieses Matrixprodukt im allgemeinen *nicht* kommutativ: Falls nicht zufälligerweise $n = p$ sein sollte, ist das Matrixprodukt BA nicht einmal definiert, geschweige denn gleich AB. Allgemein ist Kommutativität bei der Hintereinanderausführung von Abbildungen eine sehr seltene Ausnahmeerscheinung; schließlich ist auch $\sin(x^2) \neq \sin^2 x$ für fast jedes x, und so ist auch bei Matrizen, selbst wenn beide Produkte definiert sind, im allgemeinen $AB \neq BA$. Beispielsweise ist

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} ,$$

aber

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} .$$

Ansonsten gelten aber doch die meisten bekannten Rechenregeln, beispielsweise das *Assoziativitätsgesetz*

$$A(BC) = (AB)C \quad \text{für alle } A \in k^{n\times m},\ B \in k^{m\times p},\ C \in k^{p\times q} .$$

Es ist durchaus möglich (und verglichen mit manch anderen Dingen sogar nicht einmal so extrem aufwendig), dieses Gesetz nach obiger Formel explizit nachzurechnen. Bevor wir uns das antun, sollten wir uns aber daran erinnern, wo das Matrixprodukt eigentlich herkommt: Matrizen entsprechen umkehrbar eindeutig linearen Abbildungen, und das Matrixprodukt entspricht deren Hintereinanderausführung. Für die Hintereinanderausführung von Abbildungen (egal ob linear oder nicht) ist das Assoziativgesetz aber trivial: Sind etwa

$$\varphi\colon k^m \to k^n, \quad \psi\colon k^n \to k^p \quad \text{und} \quad \omega\colon k^q \to k^p$$

drei lineare Abbildungen mit Abbildungsmatrizen A, B, C, so ist für jeden Vektor $\vec{v} \in k^q$ sowohl

$$\big(\varphi \circ (\psi \circ \omega)\big)(\vec{v}) = \varphi\big(\psi \circ \omega)(\vec{v})\big) = \varphi\Big(\psi\big(\omega(\vec{v})\big)\Big)$$

als auch

$$\big((\varphi \circ \psi) \circ \omega\big)\big(\vec{v}\big) = (\varphi \circ \psi)\big(\omega(\vec{v})\big) = \varphi\Big(\psi\big(\omega(\vec{v})\big)\Big) ,$$

d.h. für die Hintereinanderausführung von Abbildungen (egal ob linear oder nicht) ist das Assoziativgesetz

$$\varphi \circ (\psi \circ \omega) = (\varphi \circ \psi) \circ \omega$$

automatisch erfüllt.

Da nun $A(BC)$ die Abbildungsmatrix von $\varphi \circ (\psi \circ \omega)$ ist und $(AB)C$ die von $(\varphi \circ \psi) \circ \omega$, und da diese beiden Abbildungen übereinstimmen, müssen auch die Abbildungsmatrizen gleich sein, wir haben also gezeigt, daß

$$A(BC) = (AB)C \quad \text{für alle } A \in k^{n\times m},\ B \in k^{m\times p},\ C \in k^{p\times q} ,$$

ohne daß wir ein einziges Matrixprodukt explizit ausrechnen mußten.

Genauso folgen auch die Rechenregeln

$$A(B_1 + B_2) = AB_1 + AB_2 \quad \text{und} \quad (A_1 + A_2)B = A_1B + A_2B$$

aus den entsprechenden Rechenregeln

$$\varphi \circ (\psi_1 + \psi_2) = \varphi \circ \psi_1 + \varphi \circ \psi_2 \quad \text{und} \quad (\varphi_1 + \varphi_2) \circ \psi = \varphi_1 \circ \psi + \varphi_2 \circ \psi ,$$

die sich wiederum leicht und ohne Rechnung überprüfen lassen:

$$\begin{aligned} \big(\varphi \circ (\psi_1 + \psi_2)\big)(\vec{v}) &= \varphi\big(\psi_1(\vec{v}) + \psi_2(\vec{v})\big) = \varphi\big(\psi_1(\vec{v})\big) + \varphi\big(\psi_2(\vec{v})\big) \\ &= (\varphi \circ \psi_1)(\vec{v}) + (\varphi \circ \psi_2)(\vec{v}) = (\varphi \circ \psi_1 + \varphi \circ \psi_2)(\vec{v}) \end{aligned}$$

und

$$\begin{aligned} \big((\varphi_1 + \varphi_2) \circ \psi\big)(\vec{v}) &= (\varphi_1 + \varphi_2)\big(\psi(\vec{v})\big) = \varphi_1\big(\psi(\vec{v})\big) + \varphi_2\big(\psi(\vec{v})\big) \\ &= (\varphi_1 \circ \psi)(\vec{v}) + (\varphi_2 \circ \psi)(\vec{v}) = (\varphi_1 \circ \psi + \varphi_2 \circ \psi)(\vec{v}) . \end{aligned}$$

Da in der obigen Formel für die Matrixmultiplikation alle $b_{\ell j}$ linear in den Ausdrücken für c_{ij} vorkommen *usw.,* hätten sich die beiden letzen Rechenregeln auch einfach direkt nachrechnen lassen. Ebenfalls durch direktes Nachrechnen überzeugt man sich von der Formel

$$(\lambda A)B = \lambda(AB) = A(\lambda B) \quad \text{für alle } \lambda \in k .$$

folgt wohl am einfachsten durch direktes Nachrechnen und für die *Einheitsmatrix*

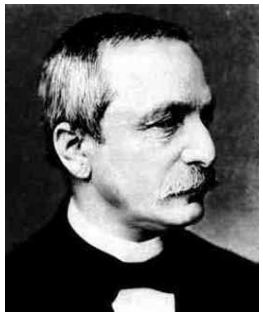
$$E = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \in k^{n\times n}$$

folgt ebenfalls sofort durch Nachrechnen wie auch durch Interpretation von E als Abbildungsmatrix der identischen Abbildung $k^n \to k^n$, die jeden Vektor auf sich selbst abbildet, daß

$$A \cdot E = A \quad \text{und} \quad E \cdot A = A \qquad \text{für alle } A \in k^{m\times n}$$

ist. Den Eintrag an der Stelle ij der Einheitsmatrix bezeichnet man als das KRONECKER-δ:

$$\delta_{ij} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases} .$$



LEOPOLD KRONECKER (1823–1891) ist heute zwar Vielen nur im Zusammenhang mit dem KRONECKER-δ bekannt, er war aber einer der bedeutendsten deutschen Mathematiker seiner Zeit. Seine Arbeiten befaßten sich mit Algebra, Zahlentheorie und Analysis, wobei er insbesondere die Verbindungen zwischen der Analysis und den beiden anderen Gebieten erforschte. Bekannt ist auch seine Ablehnung jeglicher mathematischer Methoden, die, wie die Mengenlehre oder Teile der Analysis, unendliche Konstruktionen verwenden. Er war deshalb mit vielen anderen bedeutenden Mathematikern seiner Zeit verfeindet, z.B. mit CANTOR und mit WEIERSTRASS

Bei den reellen Zahlen und auch sonst in jedem Körper gibt es zu jedem Element $a \neq 0$ ein inverses Element a^{-1}, so daß $aa^{-1} = a^{-1}a = 1$ ist. Bei Matrizen muß es das selbst für quadratische Matrizen nicht geben: Für eine beliebige Matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ ist

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$$

und

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} a & 0 \\ c & 0 \end{pmatrix} ,$$

was beides offensichtlich nie die Einheitsmatrix sein kann.

Definition: Eine $n \times n$-Matrix $A \in k^{n\times n}$ heißt *invertierbar,* wenn es eine Matrix $B \in k^{n\times n}$ gibt, so daß $AB = BA = E$ ist. B heißt inverse Matrix von A; in Zeichen $B = A^{-1}$.

(Es wäre theoretisch möglich, Invertierbarkeit auch für nicht-quadratische Matrizen zu definieren, aber das hat keinen sonderlichen Nutzen.)

Um zu sehen, wann eine Matrix $A \in k^{n\times n}$ invertierbar ist, betrachten wir wieder die Situation bei den linearen Abbildungen: Zu einer linearen Abbildung $\varphi\colon k^n \to k^n$ gibt es genau dann eine Umkehrabbildung $\psi\colon k^n \to k^n$, so daß $\varphi \circ \psi$ und $\psi \circ \varphi$ beide die identische Abbildung sind, wenn φ bijektiv ist. Nach dem Korollar am Ende von §1*i)* ist dies genau dann der Fall, wenn φ surjektiv ist, wenn also das Bild von φ Dimension n hat. Dieses Bild wird aber erzeugt von den Bildern der Einheitsvektoren, und das sind gerade die Spalten der Abbildungsmatrix. Diese n Vektoren erzeugen genau dann ganz k^n, wenn sie linear unabhängig sind.

Definition: Der (Spalten-)Rang einer Matrix $A \in k^{n\times m}$ ist die maximale Anzahl linear unabhängiger Spaltenvektoren von A.

Nach obiger Diskussion gilt daher

Lemma: Eine Matrix $A \in k^{n\times n}$ ist genau dann invertierbar, wenn sie Rang n hat. Die inverse Matrix $B = A^{-1}$ ist sowohl durch die Bedingung $AB = E$ als auch durch die Bedingung $BA = E$ eindeutig bestimmt.

Die Eindeutigkeit der inversen Matrix folgt dabei natürlich aus der Eindeutigkeit der Umkehrabbildung. Ebenfalls klar ist das folgende

Lemma: Sind $A, B \in k^{n\times n}$ invertierbar, so auch ihr Produkt, und $(AB)^{-1} = B^{-1}A^{-1}$.

Beweis: $(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AEA^{-1} = AA^{-1} = E$; also ist AB invertierbar und $B^{-1}A^{-1}$ ist die inverse Matrix. ∎

Man beachte, daß im allgemeinen $A^{-1}B^{-1}$ *nicht* invers zu AB ist: Für

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \quad \text{und} \quad AB = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$$

sind

$$A^{-1} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad B^{-1} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}$$

die inversen Matrizen, und Multiplikation zeigt, daß

$$(AB)^{-1} = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \neq \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix} = A^{-1}B^{-1}$$

ist. Insbesondere unterscheidet sich auch

$$ABA^{-1}B^{-1} = \begin{pmatrix} 21 & -8 \\ 8 & -3 \end{pmatrix}$$

deutlich von der Einheitsmatrix.

c) Matrixdarstellung der komplexen Zahlen

Die komplexen Zahlen bilden mit ihrer üblichen Addition und der Einschränkung der üblichen Multiplikation zu einer Abbildung

$$\cdot : \begin{cases} \mathbb{R} \times \mathbb{C} \to \mathbb{C} \\ (r, z) \mapsto rz \end{cases}$$

einen $\mathbb{R}$-Vektorraum und die Abbildung

$$\mathbb{C} \to \mathbb{C}; \quad z \mapsto cz$$

ist für jede komplexe Zahl $c = a + ib \in \mathbb{C}$ insbesondere eine lineare Abbildung von $\mathbb{R}$-Vektorräumen. Wählen wir $\{1, i\}$ als $\mathbb{R}$-Basis von $\mathbb{C}$, so bildet sie die beiden Basisvektoren 1 und i ab auf

$$c \cdot 1 = a + bi \quad \text{und} \quad c \cdot i = ai + bi^2 = -b + ai\,;$$

sie hat also die Abbildungsmatrix

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

Die Hintereinanderausführung zweier solcher Abbildungen entspricht der Multiplikation der entsprechenden komplexen Zahlen; insbesondere

gehört also das Produkt $(a + ib)(a' + ib')$ zweier komplexer Zahlen zur Produktmatrix

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} a' & -b' \\ b' & a' \end{pmatrix}.$$

Der Körper der komplexen Zahlen kann damit auch identifiziert werden mit der Menge aller reeller 2×2-Matrizen der obigen Form mit der Matrixaddition und dem Matrixprodukt. Man beachte, daß das Produkt zweier Matrizen dieser speziellen Form kommutativ ist, denn die Multiplikation komplexer Zahlen ist kommutativ.

Betrachten wir speziell den Fall, daß c den Betrag eins hat. Dann ist $|cz| = |c| \cdot |z| = |z|$, und allgemeiner ist für zwei beliebige komplexe Zahlen z, w auch

$$|cz - cw| = |c(z - w)| = |c| \cdot |z - w| = |z - w|\,,$$

der EUKLIDische Abstand zwischen den Bildpunkten cz und zw ist also gleich dem Abstand zwischen z und w. Die Abbildung $z \mapsto cz$ ist somit eine Kongruenzabbildung. Da sie für $c \neq 1$ den Nullpunkt als einzign Fixpunkt hat, ist sie entweder eine Drehung oder eine Drehspiegelung. Letzteres kommt nicht in Frage, da man kontinuierlich auf dem Einheitskreis von eins zu c gehen kann, also haben wir eine Drehung um einen Winkel φ.

Zur Bestimmung von φ können wir ausnutzen, daß diese Drehung den Punkt Eins in den Punkt c überführt: Betrachten wir dazu das rechtwinklige Dreieck mit Ecken $0, 1$ und $c = a + ib$! Seine Hypothenuse hat die Länge eins, seine Katheten sind a und b. Nach Definition der Winkelfunktionen als Ankathete *bzw.* Gegenkathete durch Hypothenuse ist offensichtlich $a = \cos\varphi$ und $b = \sin\varphi$. Somit ist $c = \cos\varphi + i\sin\varphi$.

Schreiben wir für einen beliebigen Winkel φ

$$c_\varphi = a_\varphi + ib_\varphi \quad \text{mit} \quad a_\varphi = \cos\varphi \quad \text{und} \quad b_\varphi = \sin\varphi\,,$$

so ist offensichtlich $c_\varphi c_\psi = c_{\varphi+\psi}$, denn die Hintereinanderausführung zweier Drehungen um den Nullpunkt ist eine Drehung mit der Summe der beiden Drehwinkel. Ausmultipliziert ergibt dies

$$c_{\varphi+\psi} = (a_\varphi + ib_\varphi)(a_\psi + ib_\psi) = (a_\varphi a_\psi - b_\varphi b_\psi) + i(a_\varphi b_\psi + a_\psi b_\varphi)\,,$$

d.h.

$$\cos(\varphi+\psi) = \cos\varphi\cos\psi - \sin\varphi\sin\psi \quad \text{und}$$
$$\sin(\varphi+\psi) = \sin\varphi\cos\psi + \cos\varphi\sin\psi\ ;$$

wir haben also die Additionstheoreme für Sinus und Kosinus gefunden.

Außerdem zeigt die Beziehung $c_\varphi c_\psi = c_{\varphi+\psi}$, daß sich c_φ als Funktion von φ wie eine Potenz verhält; wir schreiben deshalb, im Augenblick noch formal,

$$c_\varphi = e^{i\varphi} \underset{\text{def}}{=} \cos\varphi + i\sin\varphi\ .$$

Offensichtlich läßt sich jede komplexe Zahl in der Form $z = re^{i\varphi}$ schreiben, wobei $r = |z|$ der Betrag von z ist. Für $z = 0$ können wir für φ jeden beliebigen Wert nehmen; für $z \neq 0$ ist φ modulo 2π eindeutig bestimmt. In diesem Fall bezeichnen wir $\varphi = \arg z$ als das *Argument* von z und (r,φ) als die *Polarkoordinaten* von z.

In Polarkoordinaten wird die Multiplikation komplexer Zahlen deutlich einfacher als in den bislang benutzten kartesischen Koordinaten:

$$(re^{i\varphi})\cdot(se^{i\psi}) = rs\cdot e^{i(\varphi+\psi)}\ .$$

Dies läßt sich auch benutzen, um Wurzeln zu ziehen: Offensichtlich ist $\sqrt[n]{r}\cdot e^{i\varphi/n}$ eine n-te Wurzel aus $re^{i\varphi}$; weitere Wurzeln sind die Zahlen $\sqrt[n]{r}\cdot e^{i(\varphi+2k\pi)/n}$ für $k = 1,\dots,n-1$. (Für $k = n$ bekommen wir wieder dieselbe Zahl wie für $k = 0$.)

Beispielsweise ist in Polarkoordinaten $i = e^{\pi i/2}$, da i aus der Eins durch Drehung um 90° oder $\frac{\pi}{2}$ entsteht. Also ist $e^{\pi i/4}$ eine Quadratwurzel aus i. Der Winkel $\frac{\pi}{4}$ oder 45° tritt in gleichschenkligen rechtwinkligen Dreiecken auf, die man auch auffassen kann als entlang der Diagonale halbierte Quadrate; da die Diagonale eines Quadrats das $\sqrt{2}$-fache der Seite ist, sind also Sinus und Cosinus gleich $\frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$ und somit ist

$$\cos\frac{\pi}{4} + i\sin\frac{\pi}{4} = \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}$$

eine Quadratwurzel aus i ist. Die andere ist natürlich einfach das Negative davon.

Nicht nur komplexe Zahlen lassen sich mit Matrizen identifizieren; Ganz entsprechend kann man auch die Elemente der Körper $\mathbb{F}_{2^n}$ mit $n\times n$-Matrizen über $\mathbb{F}_2$ identifizieren. Nimmt man etwa $1,\alpha$ als Basisvektoren des $\mathbb{F}_2$-Vektorraums $\mathbb{F}_4$, so entsprechen die vier Elemente $0,1,\alpha$ und $\alpha+1$ des Körpers $\mathbb{F}_4$ den Matrizen

$$\begin{pmatrix}0&0\\0&0\end{pmatrix},\quad \begin{pmatrix}1&0\\0&1\end{pmatrix},\quad \begin{pmatrix}0&1\\1&1\end{pmatrix}\quad\text{und}\quad \begin{pmatrix}1&1\\1&0\end{pmatrix},$$

und für $\mathbb{F}_{32}$ mit $\mathbb{F}_2$-Basis $1,\alpha,\alpha^2,\alpha^3,\alpha^4$ und Relation $\alpha^5 = \alpha^2+1$ entspricht α der Matrix

$$\begin{pmatrix}0&0&0&0&1\\1&0&0&0&0\\0&1&0&0&1\\0&0&1&0&0\\0&0&0&1&0\end{pmatrix}.$$

d) Das Gaußsche Eliminationsverfahren

Die meisten kennen wohl aus der Schule zumindest Verfahren zur Lösung linearer Gleichungssysteme in bis zu drei Unbekannten, teilweise vielleicht auch für Systeme aus beliebig vielen Gleichungen in beliebig vielen Unbekannten.

Der GAUSS-Algorithmus, mit dem wir uns hier beschäftigen wollen, bestimmt die Lösungsmenge eines beliebigen linearen Gleichungssystems, und falls das Gleichungssystem nicht gerade eine sehr spezielle Gestalt hat, liefert er sie im allgemeinen auf die effizienteste Art und Weise.

Seine Grundidee ist sehr einfach: Im Falle einer einzigen Gleichung mit einer einzigen Unbekannten x können wir das „Gleichungssystem“

$$ax = b$$

sofort lösen: Für $a \neq 0$ ist $x = -b/a$, d.h. $\mathcal{L} = \{-b/a\}$; ansonsten gibt es für $b \neq 0$ keine Lösung, d.h. $\mathcal{L} = \emptyset$, und für $a = b = 0$ ist jedes x aus k eine Lösung, d.h. $\mathcal{L} = k$.

Das GAUSSsche Eliminationsverfahren führt ein allgemeines lineares Gleichungssystem sukzessive zurück auf solche lineare Gleichungen in einer Unbekannten, ausgehend von zwei trivialen Beobachtungen:

1. Die Lösungsmenge eines linearen Gleichungssystems ändert sich nicht, falls wir zwei Gleichungen miteinander vertauschen.
2. Die Lösungsmenge ändert sich auch dann nicht, wenn wir ein Vielfaches einer Gleichung zu einer anderen addieren, d.h. wenn wir die Gleichung

   $$\ell_j(x_1,\ldots,x_m) \underset{\text{def}}{=} a_{j1}x_1 + \ldots + a_{jm}x_m = b_j$$

   ersetzen durch

   $$\ell_j(x_1,\ldots,x_m) + \lambda\ell_i(x_1,\ldots,x_m) = b_j + \lambda b_i\,, \qquad (*)$$

   denn unter der Nebenbedingung

   $$\ell_i(x_1,\ldots,x_m) = b_i$$

   ist $(*)$ äquivalent zu $\ell_j(x_1,\ldots,x_m) = 0$.

Mit Hilfe dieser beiden Beobachtungen läßt sich nun die Variablenanzahl wie folgt sukzessive reduzieren: Beginnen wir mit der Elimination von x_1. Falls x_1 im Gleichungssystem überhaupt nicht vorkommt, falls also alle $a_{i1} = 0$ sind, gibt es nichts zu tun: Wir haben ein Gleichungssystem in $x_2,\ldots,x_m$, und für jede Lösung $(x_2,\ldots,x_m)$ dieses Systems sowie jedes beliebige $x_1 \in k$ ist $(x_1,x_2,\ldots,x_m)$ eine Lösung des ursprünglichen Systems.

Ansonsten können wir, indem wir nötigenfalls zwei Gleichungen miteinander vertauschen, annehmen, daß $a_{11} \neq 0$ ist. Dann lassen wir die erste Gleichung so stehen, wie sie ist, und ersetzen jede weitere Gleichung $\ell_j(x_1,\ldots x_m) = b_j$ durch

$$\ell_j(x_1,\ldots x_m) - \frac{a_{j1}}{a_{11}}\ell_1(x_1,\ldots x_m) = b_j - \frac{a_{j1}}{a_{11}}b_1\,;$$

in diesen Gleichungen kommt x_1 offenbar nicht mehr vor. Wir haben somit ein Gleichungssystem in einer Variablen weniger, plus der Gleichung

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = b_1\,.$$

Sobald wir das Gleichungssytem für $x_2,\ldots,x_m$ gelöst haben, wird diese Gleichung nach Einsetzen einer Lösung $(x_2,\ldots,x_m)$ zu einer linearen Gleichung für x_1, die wir lösen können:

$$x_1 = \frac{b_1 - (a_{12}x_2 + \cdots + a_{1m}x_m)}{a_{11}}\,.$$

Das Gleichungssystem für $x_2,\ldots,x_m$ wird nun, falls $m > 2$ ist, nach genau derselben Methode weiterreduziert: Falls x_2 in keiner der Gleichungen vorkommt, haben wir tatsächlich ein Gleichungssystem in $x_3,\ldots,x_m$, andernfalls können wir durch Vertauschen zweier Gleichungen annehmen, daß x_2 in der ersten Gleichung mit einem von Null verschiedenen Vorfaktor auftritt, und wir können wie oben x_2 aus allen weiteren Gleichungen eliminieren *usw.*

Da in jedem Eliminationsschritt ein Nenner auftritt, können die Nenner vor allem bei großem m gelegentlich schnell unübersichtlich groß werden. Obwohl es *im Prinzip* nicht notwendig ist, kann man um dies zu vermeiden noch als dritte Operation die Multiplikation einer Gleichung mit einem Körperelement (z.B. einem Hauptnenner der Koeffizienten) zulassen. Dies ist gleichbedeutend damit, daß man anstelle der Operation $(*)$ allgemeiner die Ersetzung von $\ell_j(x_1,\ldots x_m)$ durch

$$\mu\ell_j(x_1,\ldots x_m) + \lambda\ell_i(x_1,\ldots x_m)$$

mit beliebigem $\mu \neq 0$ aus k zuläßt. ($\mu = 0$ muß hier natürlich unbedingt ausgeschlossen werden, denn sonst läßt sich die Gleichung $\ell_j(x_1,\ldots,x_m) = 0$ aus dem neuen Gleichungssystem nicht mehr herleiten, d.h. wir erhalten auch „Lösungen“, die diese Gleichung nicht erfüllen.) Ein spezieller Fall der Multiplikation einer Gleichung mit einem Körperelement ist das Kürzen durch einen gemeinsamen Teiler der Koeffizienten, wodurch ein Gleichungssystem (egal ob das ursprüngliche oder ein Zwischenresultat) und vor allem der weitere Rechengang oft erheblich übersichtlicher wird.

e) Erste Beispiele

Betrachten wir dazu einige Beispiele; der Grundkörper k sei dabei jeweils der Körper der reellen Zahlen oder einer seiner Teilkörper, etwa $k = \mathbb{Q}$.

Sei zunächst

$$\begin{array}{rcrcrcr} 3x_1 & + & 2x_2 & + & x_3 & = & 5 \\ x_1 & + & 2x_2 & + & 4x_3 & = & 3 \\ 5x_1 & - & 3x_2 & + & 7x_2 & = & 19 \end{array}$$

das zu lösende Gleichungssystem. Da x_1 in der ersten Gleichung tatsächlich vorkommt, müssen wir nichts vertauschen; allerdings müssen wir ein Drittel der ersten Gleichung von der zweiten und fünf Drittel der ersten Gleichung von der dritten subtrahieren, um x_2 aus diesen beiden Gleichungen zu eliminieren, was auf das etwas unangenehme Gleichungssystem

$$\begin{array}{rcrcrcr} 3x_1 & + & 2x_2 & + & x_3 & = & 5 \\ & & \frac{4}{3}x_2 & + & \frac{11}{3}x_3 & = & \frac{4}{3} \\ & & -\frac{19}{3}x_2 & + & \frac{16}{3}x_3 & = & \frac{32}{3} \end{array}$$

führt. Solche Gleichungssysteme sind zwar nicht immer vermeidbar, aber hier hätten wir es auch einfacher haben können: Wenn wir im ursprünglichen Gleichungssystem die ersten beiden Gleichungen vertauschen, wird es zu

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & 4x_3 & = & 3 \\ 3x_1 & + & 2x_2 & + & x_3 & = & 5 \\ 5x_1 & - & 3x_2 & + & 7x_2 & = & 19\,, \end{array}$$

und hier müssen wir stattdessen das Dreifache *bzw.* Fünffache der ersten Gleichung von der zweiten *bzw.* dritten subtrahieren, was auf das deutlich angenehmere Gleichungssystem

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & 4x_3 & = & 3 \\ & & -4x_2 & - & 11x_3 & = & -4 \\ & & -13x_2 & - & 13x_3 & = & 4 \end{array}$$

führt. Etwas ähnliches hätten wir auch bekommen, wenn wir die letzten beiden Gleichungen des anderen Systems einfach mit drei multipliziert hätten, aber grundsätzlich ist es meistens günstiger, die Gleichung mit dem einfachsten führenden Koeffizienten an erster Stelle zu haben. Auch hier wird das Gleichungssystem zumindest optisch etwas angenehmer, wenn wir die zweite und die dritte Gleichung mit (-1) multiplizieren:

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & 4x_3 & = & 3 \\ & & 4x_2 & & 11x_3 & = & 4 \\ & & 13x_2 & + & 13x_3 & = & -4 \end{array}$$

Uns interessieren zunächst nur die letzten beiden Zeilen. Diese bilden ein lineares Gleichungssystem in x_2 und x_3, aus dem wir x_2 in einer der beiden Gleichungen eliminieren möchten.

Da x_2 in der zweiten Gleichung wirklich vorkommt, subtrahieren wir $13/4$ mal diese Gleichung von der dritten und erhalten als neues System

$$\begin{array}{rcrcrcr} 3x_1 & + & 2x_2 & + & x_3 & = & 5 \\ & & 4x_2 & + & 11x_3 & = & 4 \\ & & & & -\frac{91}{4}x_3 & = & -17\,. \end{array}$$

Hier ist die letzte Gleichung eine gewöhnliche lineare Gleichung für x_3, aus der wir sofort ablesen, daß

$$x_3 = \frac{68}{91}$$

ist. Dies setzen wir in die vorletzte Gleichung ein:

$$4x_2 + \frac{748}{91} = 4$$

ist eine lineare Gleichung für x_2 mit Lösung

$$x_2 = -\frac{96}{91}\,.$$

Dies sowie x_3 setzen wir schließlich in die erste Gleichung ein:

$$x_1 + \frac{80}{91} = 3$$

hat die Lösung

$$x_1 = \frac{193}{91}\,,$$

so daß ingesamt

$$\left(\frac{193}{91}, -\frac{96}{91}, \frac{68}{91}\right)$$

die (einzige) Lösung des Gleichungssystems ist.

Als nächstes wollen wir ein Beispiel betrachten, bei dem nicht mit einer eindeutigen Lösung zu rechnen ist: Wir nehmen ein System aus drei Gleichungen in sieben Unbekannten. Zumindest intuitiv schein klar, daß

man mit nur drei Bedingungen sieben Variable nicht eindeutig festlegen kann; wir wollen sehen, was der GAUSS-Algorithmus in so einer Situation liefert.

Das Gleichungssystem ist

$$\begin{array}{rcrcrcrcr} & & 5x_3 & & & - & 2x_6 & & = & 3 \\ & 4x_2 & & + & 2x_4 & - & 3x_6 & - & 7x_7 & = & -2 \\ x_1 & & & - & 3x_4 & + & x_5 & & & = & 0 \end{array}$$

Hier kommt x_1 in der ersten Gleichung nicht vor, wohl aber in der dritten. Wir vertauschen daher die erste Gleichung mit der dritten und erhalten

$$\begin{array}{rcrcrcrcr} x_1 & & & - & 3x_4 & + & x_5 & & & = & 0 \\ & 4x_2 & & + & 2x_4 & - & 3x_6 & - & 7x_7 & = & -2 \\ & & 5x_3 & & & - & 2x_6 & & & = & 3\,. \end{array}$$

In diesem Gleichungssystem kommt x_1 nur in der ersten Gleichung vor, x_2 nicht mehr hinter der zweiten und x_3 (natürlich) nicht mehr hinter der dritten, wir haben also bereits die Treppengestalt erreicht.

Die dritte Gleichung $5x_3 - 2x_6 = 3$ hat offensichtlich keine eindeutig bestimmte Lösung: Wir können entweder nach x_3 oder nach x_6 auflösen, so daß eine der beiden Unbekannten beliebig bewählt werden kann. Wählen wir etwa für x_6 irgendein Element $\alpha \in k$, so wird

$$x_6 = \alpha, \quad x_3 = \frac{2}{5}\alpha + \frac{3}{5}\,.$$

Die zweite Gleichung wird nach Einsetzen von $x_6 = \alpha$ zu

$$4x_2 + 2x_4 - 3\alpha - 7x_7 = -2\,,$$

also zu einer Beziehung zwischen x_2, x_4 und x_7, denn α ist ein bereits fest gewähltes und somit bekanntes Element des Grundkörpers k. Wieder können wir eine der beiden Variablen willkürlich festlegen, etwa x_7 auf ein $\beta \in k$ setzen und x_4 auf ein $\gamma \in k$; wir erhalten

$$x_7 = \beta, \quad x_4 = \gamma, \quad x_2 = \frac{3}{4}\alpha + \frac{7}{4}\beta - \frac{\gamma}{2} - \frac{1}{2}\,.$$

In der ersten Gleichung schließlich müssen wir alle bereits berechneten oder festgelegten Unbekannten einsetzen und erhalten dann die Gleichung

$$x_1 - 3\gamma + x_5 = 0\,.$$

Hier können wir noch $x_5 = \delta$ beliebig festlegen und erhalten dann für die letzte noch verbleibende Variable

$$x_1 = 3\gamma - \delta\,.$$

Insgesamt hängt die Lösung dieses Gleichungssystems also ab von vier willkürlich wählbaren Körperelementen $\alpha, \beta, \gamma, \delta \in k$ oder, wie wir auch sagen werden, von vier *Parametern*. Die *Lösungsmenge* $\mathcal{L}$ ist daher die Menge

$$\left\{ \left(3\gamma - \delta, \frac{3}{4}\alpha + \frac{7}{4}\beta - \frac{\gamma}{2} - \frac{1}{2}, \frac{2}{5}\alpha + \frac{3}{5}, \gamma, \delta, \alpha, \beta \right) \;\middle|\; \alpha, \beta, \gamma, \delta \in k \right\}\,.$$

Im ersten Beispiel brauchten wir von den beiden Operationen des GAUSS-Algorithmus nur den Eliminationsschritt, im zweiten nur den Vertauschungsschritt. Als nächstes betrachten wir ein Beispiel, in dem beide notwendig sind, das lineare Gleichungssystem

$$\begin{array}{rcrcrcrcrcr} & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ 2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ 6x_1 & - & 3x_2 & + & x_3 & + & 5x_4 & & & - & 3x_6 & = & 15 \\ 4x_1 & + & x_2 & & & - & 3x_4 & & & + & x_6 & = & 0 \\ 6x_1 & - & 6x_2 & + & x_3 & + & 10x_4 & + & 12x_5 & - & 6x_6 & = & -8 \end{array}$$

Hier kommt x_1 ausgerechnet in der ersten Gleichung nicht vor, dafür aber in allen folgenden; wir müssen also die erste Gleichung mit einer der anderen vertauschen, etwa der zweiten:

$$\begin{array}{rcrcrcrcrcr} 2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ 6x_1 & - & 3x_2 & + & x_3 & + & 5x_4 & & & - & 3x_6 & = & 15 \\ 4x_1 & - & 7x_2 & & & - & 3x_4 & & & + & x_6 & = & 0 \\ 2x_1 & - & 7x_2 & + & x_3 & + & 13x_4 & + & 12x_5 & - & 6x_6 & = & -8 \end{array}$$

Nun kommt x_1 in der zweiten Gleichung nicht mehr vor; aus der dritten, vierten *bzw.* fünften Gleichung kann x_1 eliminiert werden, indem man dreimal, zweimal *bzw.* einmal die erste Gleichung subtrahiert:

$$\begin{array}{rcrcrcrcrcr} 2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ & & -9x_2 & + & x_3 & + & 17x_4 & + & 18x_5 & - & 9x_6 & = & -21 \\ & & -3x_2 & & & + & 5x_4 & + & 12x_5 & - & 3x_6 & = & -24 \\ & & -12x_2 & + & x_3 & + & 22x_4 & + & 30x_5 & - & 12x_6 & = & -44 \end{array}$$

Als nächstes muß x_2 aus der dritten bis fünften Gleichung eliminiert werden; da x_2 in der zweiten Gleichung mit Koeffizient eins vorkommt, müssen wir einfach jenes Vielfache der zweiten Gleichung subtrahieren, das dem jeweiligen x_2-Koeffizienten entspricht. Da hier alle diese Koeffizienten negativ sind, bedeutet dies, daß wir dasjenige Vielfache der zweiten Gleichung *addieren,* das dem Betrag des Koeffizienten entspricht:

$$\begin{array}{rcrcrcrcrcr} 2x_1 & + & 2x_2 & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ & & x_2 & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ & & & x_3 & + & 62x_4 & & & & & = & 69 \\ & & & & & 20x_4 & + & 6x_5 & & & = & 6 \\ & & & x_3 & + & 82x_4 & + & 6x_5 & & & = & 76\,. \end{array}$$

Hier muß x_3 aus der letzten Gleichung eliminiert werden; dazu muß offenbar einfach die dritte Gleichung subtrahiert werden:

$$\begin{array}{rcrcrcrcrcr} 2x_1 & + & 2x_2 & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ & & x_2 & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ & & & x_3 & + & 62x_4 & & & & & = & 69 \\ & & & & & 20x_4 & + & 6x_5 & & & = & 6 \\ & & & & & 20x_4 & + & 6x_5 & & & = & 7\,. \end{array}$$

Eigentlich sollte man hier schon sehen, was los ist, aber wir rechnen zur Veranschaulichung des GAUSS-Algorithmus trotzdem stur weiter nach Schema F: Danach muß x_4 aus der letzten Gleichung eliminiert werden durch Subtraktion der vorletzten:

$$\begin{array}{rcrcrcrcrcr} 2x_1 & + & 2x_2 & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ & & x_2 & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ & & & x_3 & + & 62x_4 & & & & & = & 69 \\ & & & & & 20x_4 & + & 6x_5 & & & = & 6 \\ & & & & & & & & & 0 & = & 1\,. \end{array}$$

Damit wird endgültig klar, daß jede Lösung $(x_1, \ldots, x_6)$ des gegebenen Gleichungssystems insbesondere auch die Gleichung $0 = 1$ erfüllen muß, d.h. die Lösungsmenge ist leer.

Nachdem wir soviel Arbeit in dieses Beispiel investiert haben, sollten wir zumindest einen Teil der Rechnungen recyclen zu einem Beispiel, in dem es Lösungen gibt. Dazu muß nur die letzte der fünf ursprünglichen Gleichungen auf der rechten Seite leicht abgeändert werden: Wir

betrachten nun das System

$$\begin{array}{rcrcrcrcrcr} & & x_2 & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ 2x_1 & + & 2x_2 & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ 6x_1 & - & 3x_2 & + & x_3 & + & 5x_4 & & & - 3x_6 & = & 15 \\ 4x_1 & + & x_2 & & - & 3x_4 & & & + & x_6 & = & 0 \\ 6x_1 & - & 6x_2 & + & x_3 & + & 10x_4 & + & 12x_5 & - x_6 & = & \mathbf{-9}\,. \end{array}$$

Hierauf lassen sich genau dieselben Umformungen anwenden wie oben, anstelle des Systems mit fünfter Gleichung $0 = 1$ führen diese nun aber auf

$$\begin{array}{rcrcrcrcrcr} 2x_1 & + & 2x_2 & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\ & & x_2 & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\ & & & x_3 & + & 62x_4 & & & & & = & 69 \\ & & & & & 20x_4 & + & 6x_5 & & & = & 6 \\ & & & & & & & & & 0 & = & 0\,. \end{array}$$

Diese letzte Gleichung ist natürlich für jedes Tupel $(x_1, \ldots, x_6)$ erfüllt. Die vorletzte gibt eine Beziehung zwischen x_4 und x_5, wir können also

$$x_5 = \alpha \in k$$

beliebig wählen und erhalten dann

$$x_4 = -\frac{3}{10}\alpha + \frac{3}{10}\,.$$

Wenn wir dies in die dritte Gleichung einsetzen, bleibt dort nur noch x_3 als Variable stehen, und aus

$$x_3 - \frac{93}{5}\alpha + \frac{93}{5} = 69$$

lesen wir sofort ab, daß

$$x_3 = \frac{93}{5}\alpha + \frac{252}{5}$$

ist. Damit gehen wir in die zweite Gleichung:

$$x_2 - \frac{7}{5}\alpha + x_6 + \frac{3}{2} = 10\,.$$

Hier können wir wieder eine der beiden noch verbliebenen Variablen auf einen beliebigen Wert setzen, etwa

$$x_6 = \beta \in k\,.$$

Dann wird

$$x_2 = \frac{7}{2}\alpha - \beta + \frac{17}{2}\,,$$

was wir schließlich zusammen mit all den anderen bereits berechneten x_i in die erste Gleichung einsetzen können:

$$2x_1 + \frac{11}{5}\alpha + \frac{79}{5} = 12$$

hat die Lösung

$$x_1 = -\frac{11}{10}\alpha - \frac{19}{10}\,.$$

Damit hängt also die allgemeine Lösung dieses linearen Gleichungssystems von zwei frei wählbaren Parametern α und β ab. Sie wird geringfügig übersichtlicher, wenn wir α als $\alpha = 10\gamma$ schreiben; dann ist $\mathcal{L}$ gleich der Menge

$$\left\{ \left(-11\gamma - \tfrac{19}{10}, 35\gamma - \beta + \tfrac{17}{2}, 186\gamma + \tfrac{252}{5}, -3\gamma + \tfrac{3}{10}, 10\gamma, \beta \right) \;\middle|\; \beta, \gamma \in k \right\}.$$

Als nächstes Beispiel wollen wir ein System betrachten, daß von zwar festen, aber nicht numerisch gegebenen Parametern abhängt: Wir betrachten das Gleichstromnetzwerk aus Abbildung zwölf mit bekannten Widerständen $R_1, \ldots, R_5$ und bekanntem Eingangs- und Ausgangsstrom I; gesucht sind die Ströme $I_1, \ldots, I_5$.

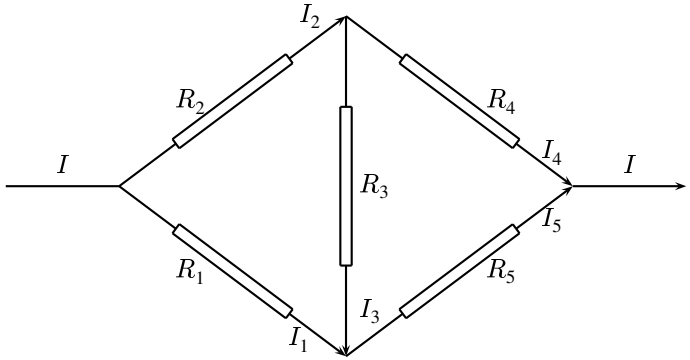


Abb. 12: Ein Gleichstromnetzwerk

Nach den KIRCHHOFFschen Gesetzen müssen sich zunächst an allen Knoten die eingehenden und die ausgehenden Ströme zu Null ergänzen, d.h. wir erhalten die Gleichungen

$$I_1 + I_2 = I = I_4 + I_5$$

für Anfang und Ende sowie

$$I_2 = I_3 + I_4 \quad \text{und} \quad I_1 + I_3 = I_5$$

für oben und unten. Außerdem müssen sich die Spannungen in den beiden Dreiecken zu Null summieren; nach dem OHMschen Gesetz führt dies auf die beiden Gleichungen

$$R_2 I_2 + R_3 I_3 - R_1 I_1 = 0 \quad \text{und} \quad R_3 I_3 + R_5 I_5 - R_4 I_4 = 0\,.$$

Ordnen wir dies nach den Strömen I_ν, erhalten wir also das lineare Gleichungssystem

$$\begin{array}{rcrcrcrcl}
I_1 & + & I_2 & & & & & = & I \\
 & & & & I_4 & + & I_5 & = & I \\
 & & I_2 & - & I_3 & - & I_4 & = & 0 \\
I_1 & & & + & I_3 & & - I_5 & = & 0 \\
-R_1 I_1 & + & R_2 I_2 & + & R_3 I_3 & & & = & 0 \\
 & & & & -R_3 I_3 & + & R_4 I_4 - R_5 I_5 & = & 0\,.
\end{array}$$

Die Elimination von I_1 aus allen Gleichungen ab der zweiten ist einfach: Wir müssen die erste Gleichung von der vierten subtrahieren und ihr R_1-faches zur fünften addieren; das neue System ist

$$\begin{array}{rcrcrcrcl}
I_1 & + & I_2 & & & & & = & I \\
 & & & & I_4 & + & I_5 & = & I \\
 & & I_2 & - & I_3 & - & I_4 & = & 0 \\
 & & -I_2 & + & I_3 & & - I_5 & = & -I \\
 & & (R_1 + R_2) I_2 & + & R_3 I_3 & & & = & R_1 I \\
 & & & & -R_3 I_3 & + & R_4 I_4 - R_5 I_5 & = & 0\,,
\end{array}$$

Da I_2 in der zweiten Gleichung nicht vorkommt, vertauschen wir diese mit der dritten; danach können wir letztere zur vierten addieren und ihr $(R_1 + R_2)$-faches von der fünften subtrahieren, um I_2 aus allen weiteren Gleichungen zu eliminieren.

Damit weiterhin jede Gleichung in eine Zeile paßt, setzen wir zur Abkürzung

$$R_{12} \underset{\text{def}}{=} R_1 + R_2 \quad \text{und} \quad R_{123} \underset{\text{def}}{=} R_1 + R_2 + R_3$$

und erhalten

$$\begin{array}{rcrcrcrcr}
I_1 & + & I_2 & & & & & = & I \\
 & & I_2 & - & I_3 & - & I_4 & = & 0 \\
 & & & & & & I_4 + I_5 & = & I \\
 & & & & & & -I_4 + -I_5 & = & -I \\
 & & & & R_{123}I_3 & + & R_{12}I_4 & = & R_1 I \\
 & & & & -R_3 I_3 & + & R_4 I_4 - R_5 I_5 & = & 0\,.
\end{array}$$

Zur Elimination von I_3 bietet sich an, die dritte Gleichung, in der I_3 nicht vorkommt, mit der sechsten zu vertauschen (diese hat einen einfacheren I_3-Koeffizienten als die fünfte), und dann das R_{123}/R_3-fache dieser Gleichung zur fünften zu addieren. Dazu müssen wir uns allerdings zunächst überlegen, ob das überhaupt zulässig ist: Falls $R_3 = 0$ ist, dürfen wir natürlich nicht durch R_3 dividieren; wenn dieser Fall nicht ausgeschlossen werden kann, muß er also ab hier getrennt behandelt werden.

Im vorliegenden Beispiel wollen wir aber davon ausgehen, daß alle fünf Widerstände tatsächlich vorhanden sind, so daß alle R_i positive reelle Zahlen sind. Dann können wir durch R_3 dividieren und das System wird zu

$$\begin{array}{rcrcrcrcr}
I_1 & + & I_2 & & & & & = & I \\
 & & I_2 & - & I_3 & - & I_4 & = & 0 \\
 & & & & -R_3 I_3 & + & R_4 I_4 - R_5 I_5 & = & 0 \\
 & & & & & & -I_4 + -I_5 & = & -I \\
 & & & & & & \alpha I_4 + \beta I_5 & = & R_1 I \\
 & & & & & & I_4 + I_5 & = & I
\end{array}$$

mit

$$\alpha = R_{12} + \frac{R_4 R_{123}}{R_3} \quad \text{und} \quad \beta = \frac{-R_5 R_{123}}{R_3}\,.$$

Schließlich muß noch I_4 aus den letzten beiden Gleichungen eliminiert werden; wir addieren also die vierte Gleichung unverändert zur letzten

und ihr α-faches zur vorletzten:

$$\begin{array}{rcrcrcrcr}
I_1 & + & I_2 & & & & & = & I \\
 & & I_2 & - & I_3 & - & I_4 & = & 0 \\
 & & & & -R_3 I_3 & + & R_4 I_4 - R_5 I_5 & = & 0 \\
 & & & & & & -I_4 + -I_5 & = & -I \\
 & & & & & & (\beta-\alpha) I_5 & = & (R_1 - \alpha) I \\
 & & & & & & 0 & = & 0
\end{array}$$

Damit können wir nacheinander

$$I_5 = \frac{R_1 - \alpha}{\beta - \alpha} \cdot I\,, \qquad I_4 = I - I_5\,, \qquad I_3 = \frac{R_4 I_4 - R_5 I_5}{R_3}\,,$$
$$I_2 = I_3 + I_4 \qquad \text{und} \qquad I_1 = I - I_2$$

bestimmen, wobei sich der Leser noch überlegen sollte, warum die Division durch $\beta - \alpha$ unproblematisch ist.

Zum Berechnen der Ströme in konkreten Beispielen reicht diese Lösungsformel aus; ist man allerdings an einem symbolischen Ausdruck interessiert, muß man die Definitionen von $\alpha, \beta, R_{12}, R_{123}$ einsetzen und nacheinander alle rechte Seiten auf Ausdrücke nur in I und den R_i reduzieren. Dies ist eine *im Prinzip* einfache Übungsaufgabe in Bruchrechnung, die man allerdings im vorliegenden Fall besser einem Computeralgebrasystem überläßt. Als Ergebnis erhält man die expliziten Formeln

$$I_1 = \frac{\big(R_2R_5 + R_2R_3 + R_2R_4 + R_3R_4\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$

$$I_2 = \frac{\big(R_1R_4 + R_1R_5 + R_3R_5 + R_1R_3\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$

$$I_3 = \frac{\big(R_1R_4 - R_2R_5\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$

$$I_4 = \frac{\big(R_1R_3 + R_1R_5 + R_2R_5 + R_3R_5\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$

$$I_5 = \frac{\big(R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}\,,$$

die für die meisten *konkreten* Anwendungen erheblich weniger nützlich sind als die obige Form des Ergebnisses.

Als letztes Beispiel schließlich betrachten wir eines, das auch von einem Parameter abhängt, bei dem man aber *nicht* wie im obigen Beispiel einfach durch Ausdrücke im Parameter dividieren darf. (Eigentlich hätten wir es da auch nicht immer dürfen, aber wir haben einfach angenommen, daß alle Widerstände wirklich vorhanden und damit positiv sind; da Kurzschlüsse immer wieder vorkommen, ist diese Annahme nicht hundertprozentig realistisch.) Das Gleichungssystem hänge ab von einem Paramter $a \in k$ und habe die Form

$$\begin{aligned} x_1 + ax_2 + x_3 &= 1 \\ x_1 + x_2 &= 1 \\ -2x_1 - 2ax_2 - ax_3 &= 1\,. \end{aligned}$$

Ein Computeralgebrasystem findet unschwer die Lösung

$$x_1 = \frac{a^2-3a-1}{a^2-3a+2}, \qquad x_2 = \frac{3}{a^2-3a+2}, \qquad x_3 = \frac{-3}{a-2}\,.$$

Diese „Lösung“ hat aber für $a = 2$ und auch für $a = 1$ Nullen im Nenner, ist dort also nicht erklärt. Für ein Computeralgebrasystem ist das kein Problem: Wie der Name sagt, rechnet es *algebraisch,* und da ist a keine reelle Zahl, sondern ein Symbol, das nichts mit irgendwelchen Zahlen zu tun hat. Damit ist $a - 2$ ein formaler Ausdruck, der nie Null sein kann, denn das *Symbol* a ist schließlich verschieden von der *Zahl* zwei.

Dieses Rechnen in sogenannten Funktionenkörpern ist mathematisch problemlos, ist aber nicht das, was in den meisten Anwendungen gefragt ist: Dort steht a im allgemeinen für eine variable Größe, in Abhängigkeit von der das Gleichungssystem gelöst werden soll. Man kann sich beispielsweise vorstellen, daß das Gleichungssystem ein lineares Regelungsproblem beschreibt in Abhängigkeit von steuerbaren Größen x_1, x_2 und x_3, wobei die zu steuernden Größen zu

$$x_1 + ax_2 + x_3, \quad x_1 + x_2 \quad \text{und} \quad -2x_1 - 2ax_2 - ax_3$$

werden. Der Paramter a wäre dann zu interpretieren als eine von außen vorgegebene Umgebungsbedingung (z.B. die Temperatur), und das lineare Gleichungssystem besagt, daß wir das System so regeln wollen,

daß die drei steuerbaren Größen allesamt eins werden – unabhängig von der Außentemperatur.

Bei einer solchen Interpretation können wir natürlich *nicht* einfach durch $a - 2$ dividieren; ein Ergebnis, wie $x_3 = -3/(a-2)$ besagt in so einem Fall, daß das Ziel im Falle $a = 2$ nicht erreichbar ist, und das ist ein sehr wichtiges Ergebnis. Bei einem System, das einen Stromkreis beschreibt, könnte das zum Beisiel bedeuten, daß beim Parameterwert $a = 2$ ein Kurzschluß entsteht, so daß dieser Parameterwert unbedingt verhindert werden muß. Deshalb muß man bei einem Gleichungssystem, das ein reales Problem beschreibt, vor jeden Division durch einen parameterabhängigen Ausdruck garantieren, daß dieser Ausdruck von Null verschieden ist, und man muß die Fälle, in denen er Null wird, gesondert diskutieren.

Im vorliegenden Beispiel (einer Vordiplomsaufgabe vom April 1999) führt dies auf folgende Lösung:

Subtraktion der ersten Gleichung von der zweiten sowie Addition der zweifachen ersten Gleichung zur dritten ergibt

$$\begin{aligned} (a-1)x_2 + x_3 &= 0 \\ (2-a)x_3 &= 3\,. \end{aligned}$$

Für $a = 2$ ist die letzte dieser beiden Gleichungen unlösbar, ansonsten ist

$$x_3 = \frac{3}{2-a} \quad \text{falls} \quad a \neq 2\,.$$

Für $a = 1$ wird die vorletzte Gleichung zu $x_3 = 0$, was der schon gefundenen Lösung

$$x_3 = \frac{3}{2-a} = 1$$

widerspricht; auch dann ist also das Gleichungssystem unlösbar. In allen anderen Fällen erhalten wir

$$x_2 = \frac{3}{(a-1)(a-2)} \quad \text{falls} \quad a \neq 1, 2\,.$$

Schließlich läßt sich noch, beispielsweise aus der zweiten Gleichung des ursprünglichen Systems, x_1 berechnen und wir erhalten als Ergebnis:

Die Lösungsmenge ist

$$\mathcal{L} = \left\{ \left(1 - \frac{3}{(a-1)(a-2)},\ \frac{3}{(a-1)(a-2)},\ \frac{3}{a-2} \right) \right\} ,$$

falls $a \neq 1, 2$, und

$$\mathcal{L} = \emptyset , \quad \text{falls } a = 1 \text{ oder } a = 2$$

ist.

In einem solchen Fall wird man durch die Nullstellen des Nenners gewarnt, daß hier etwas schiefgehen muß; es gibt aber auch Beispiele, in denen ein Computeralgebrasystem Lösungen schlichtweg „übersieht“: Bertrachten wir etwa das lineare Gleichungssystem

$$\begin{aligned} x_1 + \quad & 2x_3 = 9 \\ 2x_1 + 3x_2 + \ & x_3 = 9 \\ - \ x_2 + a & x_3 = a+2 . \end{aligned}$$

Ein Computer findet leicht die Lösung

$$x = 7, \quad y = -2 \quad \text{und} \quad z = 1 .$$

Der GAUSSalgorithmus führt uns aber über das Gleichungssystem

$$\begin{aligned} x_1 + \quad & 2x_3 = 9 \\ 3x_2 - & 3x_3 = -9 \\ - \ x_2 + a & x_3 = a+2 \end{aligned}$$

oder

$$\begin{aligned} x_1 + \quad & 2x_3 = 9 \\ x_2 - \ & x_3 = -3 \\ - \ x_2 + a & x_3 = a+2 \end{aligned}$$

auf die Endgestalt

$$\begin{aligned} x_1 + \quad & 2x_3 = 9 \\ x_2 - \ & x_3 = -3 \\ (a-1) & x_3 = a-1 , \end{aligned}$$

in der man nur für $a \neq 1$ aus der letzten Gleichung schließen darf, daß $z = 1$ ist; für $a = 1$ haben wir die immer erfüllte Gleichung $0z = 0$. Die Lösungsmenge ist hier also

$$\mathcal{L} = \{(7, -2, 1)\} \quad \text{für } a \neq 1$$

und

$$\mathcal{L} = \{(-2\lambda + 9, \lambda - 3, \lambda) \mid \lambda \in \mathbb{R}\} \quad \text{für } a = 1 .$$

Jemand, der Maple hinreichend gut kennt, hätte natürlich auch diese vollständige Lösung damit ermitteln können, aber der einfachstmögliche Befehl reicht definitiv nicht aus – zumindest ein Grund, warum man auch heute noch lernen muß, lineare Gleichungssysteme von Hand zu lösen.

Ein anderer Grund, warum speziell Technische Informatiker das lernen müssen, liegt in der Natur vieler Anwendungen: Lineare Gleichungssysteme müssen beispielsweise gelöst werden bei Steuerungs- und Regelungsproblemen. In vielen Fällen wird diese Steuerung nicht von einem leistungsfähigen Universalrechner durchgeführt, sondern von einer eigens dafür entwickelten Schaltung, die mit möglichst wenig Aufwand arbeiten soll – sei es aus Kostengründen oder wegen des Raumbedarfs oder der Wärmeentwicklung. In solchen Fällen geht es dann darum, die Lösung möglichst effizient zu ermitteln, und bei der Definition des Wortes „effizient“ können hier durchaus auch nichtmathematische Gesichtspunkte eine Rolle spielen. Daher ist es wichtig, das volle Instrumentarium der Lösungstheorie linearer Gleichungssysteme zu beherrschen, um die jeweils beste Methode implementieren zu können. Deshalb werden wir auch noch Alternativen zum GAUSS-Algorithmus betrachten, und in der *Numerik I* werden weitere Verfahren folgen.

f) Die Struktur der Lösungsmenge

Nach diesen Beispielen ist es an der Zeit, wieder zu den theoretischen Grundlagen zurückzukehren. Sei also wieder

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m &= b_2 \\ \vdots \qquad\qquad\qquad\quad &= \ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m &= b_n , \end{aligned}$$

ein allgemeines lineares Gleichungssystem. Wenn wir die a_{ij} zu einer

Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}$$

zusammenfassen und die Unbekannten und rechten Seiten zu Vektoren

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \quad \text{und} \quad \vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

läßt sich das Gleichungssystem auch kurz als $A\vec{x} = \vec{b}$ schreiben, wobei wir das Produkt der Matrix A mit dem Vektor $\vec{x}$ so definieren, daß es jener Vektor aus k^n sein soll, dessen i-te Komponente die Summe

$$\sum_{j=1}^{m} a_{ij} x_j$$

sein soll. Identifiziert man Vektoren aus k^m mit $m \times 1$-Matrizen, ist das gerade das Produkt der $n \times m$-Matrix A mit der $m \times 1$-Matrix der Variablen.

Wir bezeichen A als die *Matrix des Gleichungssystems* und

$$\left(A \mid \vec{b}\right) \underset{\text{def}}{=} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & b_1 \\ a_{21} & a_{22} & \dots & a_{2m} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} & b_m \end{pmatrix}$$

als die *erweiterte Matrix.*

Damit können wir das Gleichungssystem in die Sprache der Vektorräume und linearen Abbildungen übersetzen: Wir betrachten die lineare Abbildung

$$\varphi: k^m \to k^n; \quad \vec{v} \mapsto A\vec{v},$$

und die Lösungsmenge des linearen Gleichungssystems ist

$$\mathcal{L} = \left\{ \vec{v} \in k^m \;\middle|\; \varphi(\vec{v}) = \vec{b} \right\}.$$

Für zwei Lösungsvektoren $\vec{v}, \vec{w} \in k^m$ ist

$$\varphi(\vec{v} - \vec{w}) = \varphi(\vec{v}) - \varphi(\vec{w}) = \vec{b} - \vec{b} = \vec{0},$$

die Differenz zweier Lösungen ist also eine Lösung des entsprechenden linearen Gleichungssystems mit lauter Nullen auf der rechter Seite.

Definition: Ein lineares Gleichungssystem $A\vec{x} = \vec{b}$ heißt *homogen,* wenn $\vec{b} = \vec{0}$ ist; ansonsten heißt es *inhomogen.* $A\vec{x} = \vec{0}$ heißt das zu $A\vec{x} = \vec{b}$ gehörige homogene Gleichungssystem.

Damit wissen wir also

Lemma: Zwei Lösungen des linearen Gleichungssystems $A\vec{x} = \vec{b}$ unterscheiden sich durch eine Lösung des homogenen Gleichungssystems. ∎

Die Lösung eines homogenen Gleichungssystems besteht aus allen Vektoren aus k^m, die von der oben definierten linearen Abbildung φ auf den Nullvektor abgebildet werden, ist also gerade der Kern von φ und damit ein Untervektorraum von k^n. Insbesondere ist die Lösungsmenge eines homogenen linearen Gleichungssystems also nie leer: Wie man sofort auch direkt sieht, gibt es immer die sogenannte *triviale* Lösung, bei der alle Variablen gleich Null sind.

Nach der Dimensionsformel aus §1*i)* ist

$$\dim \text{Kern}\, \varphi = m - \dim \text{Bild}\, \varphi,$$

und das Bild wird erzeugt von den m Vektoren $\varphi(\vec{e}_i) = A\vec{e}_i$. Das sind aber gerade die Spaltenvektoren der Matrix A; die Dimension des Bildes ist also gleich der Anzahl linear unabhängiger Spaltenvektoren von A, und diese Zahl hatten wir oben als den (Spalten-)Rang der Matrix definiert. Damit wissen wir also

Lemma: Die Lösungsmenge des homogenen linearen Gleichungssystems $A\vec{x} = \vec{0}$ in m Variablen ist ein Untervektorraum der Dimension $m - \text{Rang}\, A$. ∎

Im Falle eines inhomogenen Gleichungssystems ist die Sache nicht ganz so einfach: Schließlich hatten wir bei den Beispielen schon gesehen,

daß von zwei Gleichungssystemen, die sich nur in ihrer rechten Seite unterscheiden, das eine unlösbar sein kann, während das andere eine oder mehrere Lösungen hat.

Der Grund dafür wird klar, wenn wir das Gleichungssystem wieder über die lineare Abbildung φ interpretieren: Das Gleichungssystem $A\vec{x} = \vec{b}$ ist genau dann lösbar, wenn die rechte Seite $\vec{b}$ im Bild von φ liegt. Damit muß sie linear abhängig sein von den Spaltenvektoren von A, die ja den Bildraum erzeugen, d.h. der Rang der *erweiterten* Matrix, die außer den Spaltenvektoren von A noch den Vektor $\vec{b}$ enthält, darf nicht größer sein, als der von A. Kleiner kann er natürlich nicht sein, also haben wir gezeigt

Lemma: Ein inhomogenes Gleichungssystem $A\vec{x} = \vec{b}$ ist genau dann lösbar, wenn der Rang r seiner erweiterten Matrix gleich dem von A ist. Es ist genau dann eindeutig lösbar, wenn dieser gemeinsame Rang gleich der Anzahl m der Variablen ist. ∎

Ist der gemeinsame Rang von Matrix und erweiterter Matrix kleiner als m, so ist das Gleichungssystem nicht eindeutig lösbar. Da sich zwei Lösungen um eine Lösung des zugehörigen homogenen Gleichungssystems unterscheiden und diese wiederum einen Vektorraum der Dimension $m -$ Rang A bilden, hängt die Lösung dann von $m -$ Rang A Parametern ab, ist aber für ein echt inhomogenes Gleichungssystem kein Vektorraum: Ein Vektorraum enthält stets den Nullvektor, und wenn es eine Lösung gibt, bei der *alle* Variablen verschwinden, stehen auf der rechten Seite des Gleichungssystems lauter Nullen, d.h. das Gleichungssystem ist homogen.

Das obige Lemma wird nur selten von praktischem Nutzen sein, denn wenn Matrix und erweiterte Matrix keine sehr spezielle Gestalt haben, wird der GAUSS-Algorithmus im allgemeinen die einfachste Möglichkeit sein, um die Ränge von Matrix und erweiterter Matrix zu bestimmen. Man muß ihn allerdings nicht ganz zu Ende durchführen, denn die Ränge sind schon klar, wenn auf der linken Seite Trepengestalt erreicht ist. Der Vollständigkeit halber sei hier kurz angegeben, wie diese im allerallgemeinsten Fall aussieht:

Die verschiedenen Schritte des GAUSS-Algorithmus wirken sich auf die Matrix wie auch auf die erweiterte Matrix so aus, daß entweder zwei Zeilen miteinander vertauscht werden, oder aber ein Vielfaches einer Zeile zu einer anderen Zeile addiert wird – Addition bedeutet dabei die gewöhnliche komponentenweise Addition, d.h. die Addition von Zeilenvektoren. Am Ende entsteht eine erweiterte Matrix der Form

$$\left(\begin{array}{cccccccccc}
[0] & \bullet & [\star] & \star & [\star] & \star & \dots & [\star] & \star & [\star] & \star \\
[0] & 0 & [0] & \bullet & [\star] & \star & \dots & [\star] & \star & [\star] & \star \\
[0] & 0 & [0] & 0 & [0] & \bullet & \dots & [\star] & \star & [\star] & \star \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & \bullet & [\star] & \star \\
\hline
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & \bullet \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & \bullet \\
\hline
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & 0
\end{array}\right),$$

wobei $\bullet$ für ein von Null verschiedenes und $\star$ für ein beliebiges Element des Körpers k steht; [0] und [$\star$] stehen für keine, eine oder mehrere Nullen *bzw.* beliebige Elemente, deren Anzahl in untereinanderstehenden Termen jeweils die gleiche sein soll.

Die Zeilen zwischen den beiden eingezeichneten Linien, in denen alle Einträge der Matrix des Gleichungssystems verschwinden, müssen natürlich nicht wirklich auftreten, genauso wenig die Zeilen unterhalb der zweiten Linie, wo sogar alle Einträge der erweiterten Matrix verschwinden, so daß auch die rechten Seiten der entsprechenden Gleichungen in der Endgestalt nach Anwendung des GAUSS-Algorithmus Null sind.

Falls zwischen den beiden eingezeichneten Linien wirklich eine oder mehrere Zeilen stehen, ist das Gleichungssystem *unlösbar,* denn dann

gibt es ja in der Endform des Gleichungssystems Gleichungen, in denen links alle Koeffizienten Null sind, während rechts eine von Null verschiedene Zahl steht. Indem man gegebenenfalls ein Vielfaches der ersten solchen Gleichung von den etwa vorhandenen weiteren Gleichungen subtrahiert, kann man erreichen, daß alle weiteren Gleichungen zu $0 = 0$ werden, d.h. wir können erreichen, daß zwischen den beiden Linien *höchstens eine* Zeile steht, und das lineare Gleichungssystem ist genau dann lösbar, wenn keine dort steht.

Damit ist klar, daß die Anzahl der Zeilen oberhalb der ersten Linie der *Rang der Matrix A* ist und die Anzahl der Zeilen oberhalb der zweiten Linie (nachdem man dafür Sorge getragen hat, daß höchstens eine Zeile zwischen den beiden Linien steht) der *Rang der erweiterten Matrix.*

Noch etwas läßt sich diesem Diagramm ansehen: Der Rang der Matrix, die maximale Anzahl linear unabhängiger Spalten also, ist offenbar gerade gleich der Anzahl der Zeichen • oberhalb des ersten Strichs. Genau das ist aber auch die maximale Anzahl linear unabhängiger Zeilen der Matrix, d.h. wir können am Rande noch notieren, daß der *Zeilrang* einer Matrix gleich dem *Spaltenrang* ist, was die Kurzbezeichnung *Rang* rechtfertigt.

Betrachten wir zum Abschluß noch ein Beispiel für das Rangkriterium zur Lösbarkeit eines linearen Gleichungssystems: Im vorigen Abschnitt hatten wir das lineare Gleichungssystem

$$\begin{array}{rcrcrcrcrcrcr}
 & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\
6x_1 & - & 3x_2 & + & x_3 & + & 5x_4 & & & - & 3x_6 & = & 15 \\
4x_1 & - & 7x_2 & & & - & 3x_4 & & & + & x_6 & = & 0 \\
2x_1 & - & 7x_2 & + & x_3 & + & 13x_4 & + & 12x_5 & - & 7x_6 & = & -8
\end{array}$$

als unlösbar erkannt. Seine erweiterte Matrix ist

$$\left(\begin{array}{rrrrrrr}
0 & 1 & 0 & 5 & -2 & 1 & 10 \\
2 & 2 & 0 & -4 & -6 & 2 & 12 \\
6 & -3 & 1 & 5 & 0 & -3 & 15 \\
4 & -7 & 0 & -3 & 0 & 1 & 0 \\
2 & -7 & 1 & 13 & 12 & -7 & -8
\end{array}\right),$$

und nach Anwendung des GAUSS-Algorithmus kommen wir auf die Endgestalt

$$\left(\begin{array}{rrrrrrr}
2 & 2 & 0 & -4 & -6 & 2 & 12 \\
0 & 1 & 0 & 5 & -2 & 1 & 10 \\
0 & 0 & 1 & 62 & 0 & 0 & 69 \\
0 & 0 & 0 & 20 & 6 & 0 & 6 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right).$$

Somit hat die Matrix des linearen Gleichungssystems den Rang vier, wohingegen die erweiterte Matrix den Rang fünf hat.

Ändern wir im Gleichungssystem die rechte Seite der letzten Gleichung von -8 zu -9, so wird es lösbar. In der Endgestalt ändert sich, wie wir oben gesehen haben, auch nur die rechte Seite der letzten Gleichung; die erweiterte Matrix wird also schließlich zu

$$\left(\begin{array}{rrrrrrr}
2 & 2 & 0 & -4 & -6 & 2 & 12 \\
0 & 1 & 0 & 5 & -2 & 1 & 10 \\
0 & 0 & 1 & 62 & 0 & 0 & 69 \\
0 & 0 & 0 & 20 & 6 & 0 & 6 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right).$$

Hier haben Matrix und erweiterte Matrix denselben Rang vier.

g) Affine Räume

Für inhomogene Gleichungssysteme kennen wir bislang noch nicht die Struktur des Lösungsraums selbst; wir wissen nur, daß die *Differenzen* von Lösungen (so es welche gibt) einen Vektorraum bilden.

Dies sollte etwas an die Situation in §1*a)* erinnern, wo wir zwei Punkten deren *Verbindungsvektor* zugeordnet hatten und zwei Vektoren unabhängig von ihren Anfangs- und Endpunkten als gleich bezeichnet hatten, wenn sie nur die gleiche Länge und Richtung hatten. Eine ähnliche Situation haben wir auch hier: Lösungen eines inhomogenen Gleichungssystems (so es welche gibt) verhalten sich wie Punkte, Lösungen eines homogenen Gleichungssystems wie Vektoren.

Beispielsweise gibt es keine sinnvolle Addition zweier Punkte, genauso wie auch die Summe zweier Lösungen einen inhomogenen Gleichungssystems keine Lösung mehr ist. Andererseits kann man einen Vektor an

einem Punkt abtragen, um einen weiteren Punkt zu bekommen, und genauso kann man auch eine Lösung des homogenen Gleichungssystems zu einer Lösung des inhomogenen addieren, um eine weitere Lösung des inhomogenen zu erhalten.

Die Mathematik sucht hinter analogen Sachverhalten gemeinsame Strukturen; in diesem Fall bezeichnet sie diese als *affine Räume:*

Definition: Ein affiner Raum über dem Körper k ist gegeben durch ein Paar (A, V) bestehend aus einer nichtleeren Menge A, deren Elemente wir als Punkte bezeichnen, und einem Vektorraum V, so daß gilt:

1.) Es gibt eine Abbildung $\varphi\colon \begin{cases} A \times A \to V \\ (\boldsymbol{x}, \boldsymbol{y}) \mapsto \overrightarrow{\boldsymbol{xy}} \end{cases}$, mit der Eigenschaft, daß

$$\overrightarrow{\boldsymbol{xy}} + \overrightarrow{\boldsymbol{yz}} = \overrightarrow{\boldsymbol{xz}} \quad \text{für alle } \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in A\,.$$

2.) Für jeden Punkt $O \in A$ ist die Abbildung

$$\varphi_O\colon A \to V; \quad \boldsymbol{x} \mapsto \varphi(O, \boldsymbol{x}) = \overrightarrow{O\boldsymbol{x}}$$

bijektiv; ihre Umkehrabbildung sei mit ψ_O bezeichnet.

Die Dimension des Vektorraums V bezeichnen wir auch als Dimension des affinen Raums.

Anschaulich gesehen ordnet die Abbildung φ zwei Punkten $\boldsymbol{x}, \boldsymbol{y} \in A$ den Verbindungsvektor $\overrightarrow{\boldsymbol{xy}}$ zu, und ψ_O beschreibt das *Abtragen* eines Vektors: In §1 hatten wir Vektoren eingeführt als Äquivalenzklassen von Pfeilen gleicher Länge und gleicher Richtung; nimmt man zum Vektor $\vec{v}$ den Pfeil, der im Punkt O beginnt, so hat dieser den Endpunkt $\psi_O(\vec{v})$.

Standardbeispiel eines affinen Raums ist der $\mathbb{R}^n$; hier ist A gleich der *Menge* $\mathbb{R}^n$, deren Elemente wir als Punkte $\boldsymbol{x} = (x_1, \ldots, x_n)$ schreiben, und V ist gleich dem *Vektorraum* $\mathbb{R}^n$, dessen Elemente wir als Spaltenvektoren schreiben. φ ist die Abbildung, die den Punkten

$$\boldsymbol{x} = (x_1, \ldots, x_n) \quad \text{und} \quad \boldsymbol{y} = (y_1, \ldots, y_n)$$

deren Verbindungsvektor

$$\varphi(\boldsymbol{x}, \boldsymbol{y}) = \overrightarrow{\boldsymbol{xy}} = \begin{pmatrix} y_1 - x_1 \\ \vdots \\ y_n - x_n \end{pmatrix}$$

zuordnet, und für $O = (a_1, \ldots, a_n)$ ist

$$\psi_O\left(\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right) = (a_1 + x_1, \ldots, a_n + x_n)\,.$$

Ganz entsprechend läßt sich für jeden beliebigen Körper k die Menge k^n zu einem affinen Raum machen; der zugehörige Vektorraum ist natürlich der *Vektorraum* k^n.

Weitere interessante Beispiele sind vor allem die Geraden, Ebenen *usw.* in diesem Raum; diese fassen wir zusammen unter der

Definition: Ein affiner Unterraum des affinen Raums (A, V) ist ein affiner Raum (B, U) mit $B \subseteq A$ und $U \leq V$, wobei auch die Abbildungen φ und ψ_O Einschränkungen der entsprechenden Abbildungen von (A, V) sind.

Betrachten wir etwa die Gerade

$$\left\{(1 + \lambda, \lambda) \;\middle|\; \lambda \in \mathbb{R}\right\} \subseteq \mathbb{R}^2$$

durch den Punkt $(1, 0)$ mit Steigung 1. Hier ist B natürlich gleich der Menge aller Punkte auf der Geraden, also die gerade angegebene Menge selbst, und als Vektorraum nehmen wir den vom Richtungsvektor $\binom{1}{1}$ der Geraden aufgespannten Untervektorraum des $\mathbb{R}^2$.

Dieselbe Gerade können wir auch durch die Gleichung

$$y = x + 1 \quad \text{oder} \quad -x + y = 1$$

beschreiben. Allgemein können wir, wie wir uns als nächstes überlegen wollen, jeden affinen Unterraum eines k^n als Lösungsmenge eines linearen Gleichungssystems in n Unbekannten interpretieren, und umgekehrt ist auch jede solche Lösungsmenge, falls sie nicht leer ist, affiner

Unterraum von k^n. (Mit k^n ist hier natürlich das Paar (k^n, k^n) gemeint, das wir kurz, wenn auch schlampig, als affinen Raum k^n bezeichnen.)

Beginnen wir mit der Lösungsmenge L eines linearen Gleichungssystems

$$A\vec{x} = \vec{b} \quad \text{mit} \quad A \in k^{m\times n}, \quad \vec{b} \in k^m$$

für einen Vektor $\vec{x} \in k^n$; mit anderen Worten mit der Lösungsmenge von m linearen Gleichungen in n Unbekannten. (Hier schreiben wir, aus alter Gewohnheit, die Punkte eines affinen Raums ausnahmsweise als Vektoren.) *Falls* es Lösungen gibt, ist die Differenz $\vec{y} = \vec{x}^{(1)} - \vec{x}^{(2)}$ zweier (nicht notwendigerweise verschiedener) Lösungen $\vec{x}^{(1)}$ und $\vec{x}^{(2)}$ eine Lösung des homogenen Gleichungssystems $A\vec{y} = \vec{0}$.

Die Lösungsmenge eines *homogenen* linearen Gleichungssystems in n Variablen ist aber Kern einer linearen Abbildung und daher (und auch aus vielen anderen Gründen) ein Untervektorraum U des k^n. Zusammen mit diesem Untervektorraum wird L zu einem affinen Raum, wobei die Abbildung $\varphi\colon L \times L \to U$ zwei Lösungen $\vec{x}^{(1)}$ und $\vec{x}^{(2)}$ deren Differenz $\vec{y} = \vec{x}^{(1)} - \vec{x}^{(2)}$ als „Verbindungsvektor“ zuordnet; entsprechend macht für jede Lösung $\vec{x}$ des inhomogenen Gleichungssystems die Abbildung $\psi_{\vec{x}}\colon U \to L$ eine Lösung $\vec{y}$ des homogenen Gleichungssystems durch Addition von $\vec{x}$ zu einer Lösung des inhomogenen Gleichungssystems.

Ist umgekehrt (B, U) ein affiner Unterraum von (k^n, k^n), so können wir eine Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ von U wählen und diese zu einer Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ des k^n ergänzen. Bezüglich dieser neuen Basis des k^n ist dann U die Lösungsmenge des homogenen linearen Gleichungssystems

$$x_{r+1} = 0, \quad x_{r+2} = 0, \quad \ldots, \quad x_n = 0\,.$$

Sind nun $(c_1, \ldots, c_n)$ und $(d_1, \ldots, d_n)$ zwei Punkte aus B, so liegt ihre Differenz (d.h. ihr „Verbindungsvektor“) in U, und somit muß für alle $i > r$ notwendigerweise $c_i = d_i$ sein. Damit ist B bezüglich der Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ die Lösungsmenge des linearen Gleichungssystems

$$x_{r+1} = d_{r+1}, \quad x_{r+2} = d_{r+2}, \quad \ldots, \quad x_n = d_n\,.$$

Durch Rücktransformation auf die Standardbasis des k^n kann man dieses lineare Gleichungssystem in ein lineares Gleichungssystem bezüglich der ursprünglichen Basis umformen.

Um den Kontakt an die Schulgeometrie zu verdeutlichen, sei für Interessenten noch kurz dargestellt, was man mit affinen Unterräumen *geometrisch* anstellen kann.

Spätestens die Interpretation als Lösungsmenge eines linearen Gleichungssystems sollte jedem klargemacht haben, daß der Durchschnitt zweier affiner Unterräume entweder leer ist oder wieder ein affiner Unterraum; in beiden Fällen kann er durch Lösen eines linearen Gleichungssystems ermittelt werden. Seine Dimension hängt offensichtlich von einer ganzen Reihe von Faktoren ab: Schon in der Ebene kann der Durchschnitt zweier Geraden entweder leer sein (parallele Geraden) oder nulldimensional (ein Schnittpunkt) oder eindimensional (zusammenfallende Geraden).

Definition: Für zwei affine Unterräume (A, U) und (B, V) eines affinen Raums (C, W) ist der von A und B aufgespannte affine Unterraum von (C, W) der kleinste affine Unterraum (S, T) von (C, W), der beide enthält.

Bei dieser Definition stellt sich als erstes die Frage, ob sie überhaupt sinnvoll ist, ob es einen solchen kleinsten Raum also auch wirklich gibt. Dazu bilden wir den Durschnitt aller affiner Unterräume, die (A, U) und (B, V) als Unterräume enthalten. Der Durchschnitt einer beliebigen Menge affiner Unteräume ist entweder leer oder wieder ein affiner Unterraum. Leer kann er hier nicht sein, da wir nur Unterräume bertrachten, die die beiden gegebenen Räume enthalten, also ist er ein affiner Unterraum und offensichtlich der kleinste unter denen, die (A, U) und (B, V) als Unterräume enthalten.

Seine Dimension kann allerdings größer werden als die Summe der Dimensionen der beiden Ausgangsräume; ein wahrscheinlich aus der Schule bekanntes Beispiel dafür sind *windschiefe Geraden* im $\mathbb{R}^3$, d.h zwei Geraden, die zwar nicht parallel zueinander sind, die aber in zwei voneinander verschiedenen parallelen Ebenen liegen.

Ein Beispiel hierfür sind etwa die x-Achse eines kartesischen Koordiantensystems im $\mathbb{R}^3$ und die um eins in z-Richtung verschobene y-Achse. Diese beiden Geraden schneiden sich nicht, denn alle Punkte der ersten

haben z-Koordinate Null, während auf der zweiten $z = 1$ ist, und sie sind auch offensichtlich nicht parallel.

Wir wollen uns überlegen, daß es keine Ebene gibt, die beide Geraden enthält. Dazu gehen wir aus von der allgemeinsten Gleichung

$$ax + by + cz = d$$

für eine Ebene im $\mathbb{R}^3$. Wenn diese Ebene die x-Achse enthalten soll, also die Menge aller Punkte der Form $(x, 0, 0)$, muß $a = d = 0$ sein. Die um eins in z-Richtung verschobene y-Achse besteht aus allen Punkten der Form $(0, y, 1)$; sie liegt in der Ebenen, wenn $b = 0$ und $c = d$ ist. Für eine Ebene, die beide Geraden enthält, müßte also $a = b = c = d = 0$ sein, und dann haben wir keine Ebene mehr. Da es zwischen einer Ebenen und dem gesamten $\mathbb{R}^3$ keine weiteren affinen Unterräume mehr gibt, ist der kleinste affine Unterraum, der die beiden Geraden enthält, der gesamte $\mathbb{R}^3$.

Damit drängt sich die Frage auf, ob es möglicherweise auch im $\mathbb{R}^4$ oder in Räumen noch höherer Dimension zwei Geraden geben kann, die den gesamten Raum aufspannen. Da die Anschauung hier leider nicht weiterhilft, müssen wir abstrakt mathematisch argumentieren. Das wird am einfachsten, wenn wir rechnen können, und dazu führen wir Koordinatensysteme ein:

Definition: Ein Koordinatensystem des affinen Raums (A, V) besteht aus einem Punkt $O \in A$, genannt der Ursprung oder auch Nullpunkt des Koordinatensystems, und einer Basis von V. Die Geraden durch O mit einem Basisvektor als Richtungsvektor heißen *Koordinatenachsen.*

Bezüglich eines solchen Koordinatensystems hat dann in der Tat jeder Punkt $\boldsymbol{x}$ eines n-dimensionalen affinen Raums A ein n-tupel von Koordinaten, nämlich die Komponenten des Vektors $\overrightarrow{O\boldsymbol{x}}$. Bevor wir aber zeigen können, daß der affine Raum dadurch isomorph wird zu k^n, müssen wir zunächst definieren, was strukturerhaltende Abbildungen zwischen affinen Räumen überhaupt sein sollen:

Definition: Eine affine Abbildung zwischen den affinen Räumen (A, V) und (B, W) besteht aus einer Abbildung $\varphi\colon A \to B$ und einer linearen

Abbildung $\psi\colon V \to W$ derart, daß für alle Punkte $\boldsymbol{x}, \boldsymbol{y} \in A$ gilt:

$$\overrightarrow{\varphi(\boldsymbol{x})\,\varphi(\boldsymbol{y})} = \psi\left(\overrightarrow{\boldsymbol{x}\boldsymbol{y}}\right) .$$

Die Abbildung heißt *Isomorphismus* oder *Affinität,* falls φ bijektiv ist.

Offensichtlich ist φ genau dann bijektiv, wenn ψ ein Isomorphismus von Vektorräumen ist, so daß wir dies nicht zusätzlich fordern müssen. Es ist auch klar, daß nach Wahl eines Koordinatensystems eines n-dimensionalen affinen Raums die Koordinaten einen Isomorphismus mit dem affinen Raum k^n definieren.

Da wir wissen, wie lineare Abbildungen zwischen Vektorräumen bezüglich zweier Basen aussehen, können wir auch sofort hinschreiben, wie die Abbildung $\varphi\colon A \to B$ bezüglich zweier Koordinatensysteme aussieht:

$$(x_1, \ldots, x_n) \mapsto \Big(\sum_{j=1}^{n} a_{1j}x_j + b_1, \ldots, \sum_{j=1}^{n} a_{mj}x_j + b_m\Big) ,$$

im Gegensatz zu den homogenen linearen Abbildungen zwischen Vektorräumen sind hier also auch inhomogene lineare Abbildungen möglich. In Matrixschreibweise betrachtet man am besten die Spaltenvektoren: Ist

$$\varphi\big((x_1, \ldots, x_n)\big) = (y_1, \ldots, y_m) ,$$

so ist

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \vec{b} \quad \text{mit} \quad \vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} .$$

Geometrisch betrachtet beschreibt der konstante Vektor $\vec{b}$ also eine Verschiebung.

Vor allem in der Computergraphik faßt man die Matrix A und den Vektor $\vec{b}$ oft auch zusammen zu einer Matrix A^* mit $m + 1$ Zeilen und $n + 1$ Spalten, indem man zunächst den Vektor $\vec{b}$ als $(n + 1)$-te Spalte an die Matrix anhängt und dann als $(m + 1)$-te Zeile lauter Nullen einsetzt mit Ausnahme einer Eins an der letzten Stelle:

$$A^* = \begin{pmatrix} A & \vec{b} \\ \mathbf{0} & 1 \end{pmatrix} ,$$

wobei die fette Null für n gewöhnliche Nullen steht. Mit dieser Bezeichnung ist dann

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \\ 1 \end{pmatrix} = A^* \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ 1 \end{pmatrix} ,$$

eine Formel, die zwar rechnerisch aufwendiger ist als die obige, dafür aber etwas kompakter. Sie hat auch den Vorteil, daß sich die Hintereinanderausführung affiner Abbildungen so durch ein Matrixprodukt beschreiben läßt.

Als Anwendung von Koordinatensystemem wollen wir, wie bereits erwähnt, die Dimension des Erzeugnisses zweier affiner Unterräume berechnen.

Betrachten wir zunächst die Dimension des Erzeugnisses zweier Untervektorräume:

Definition: Die Summe $U + V$ zweier Untervektorräume eines festen Vektorraums ist der kleinste Untervektorraum, der beide enthält.

Mit der Notation aus §1*f)* können wir dies auch als $U + V = [U \cup V]$ schreiben, denn für jede Teilmenge M eines Vektorraums hatten wir $[M]$ als kleinsten Untervektorraum definiert, der M enthält.

Lemma: Sind U und V endlichdimensional, so ist

$$\dim(U + V) = \dim U + \dim V - \dim(U \cap V) .$$

Beweis: Wir starten mit einer Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ des Durchschnitts $U \cap V$ und ergänzen diese zu einer Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ von U und zu einer Basis $\{\vec{b}_1, \ldots, \vec{b}_r, \vec{c}_{r+1}, \ldots, \vec{c}_m\}$ von V. Dann ist $U + V$ der von den Vektoren $\vec{b}_i$ und $\vec{c}_j$ erzeugte Untervektorraum, denn einen kleineren Untervektorraum, der U und V enthält, kann es nicht geben. Außerdem sind diese Vektoren linear unabhängig: Ist nämlich

$$\lambda_1\vec{b}_1 + \cdots + \lambda_n\vec{b}_n + \mu_{r+1}\vec{c}_{r+1} + \cdots + \mu_m\vec{c}_m = \vec{0} ,$$

so können wir dies auch schreiben als

$$\vec{v} \underset{\text{def}}{=} \lambda_1\vec{b}_1 + \cdots + \lambda_n\vec{b}_n = -\left(\mu_{r+1}\vec{c}_{r+1} + \cdots + \mu_m\vec{c}_m\right) .$$

Der Vektor $\vec{v}$ läßt sich damit sowohl als Linearkombination von Vektoren aus U darstellen wie auch als Linearkombination von Vektoren aus V; er liegt also in $U \cap V$. Da dieser Durchschnitt die Vektoren $\vec{b}_1, \ldots, \vec{b}_r$ als Basis hat, müssen daher λ_{r+1} bis λ_n verschwinden. Damit ist also

$$\lambda_1\vec{b}_1 + \cdots + \lambda_{\mathbf{r}}\vec{b}_{\mathbf{r}} + \mu_{r+1}\vec{c}_{r+1} + \cdots + \mu_m\vec{c}_m = \vec{0} ,$$

und hier haben wir eine Darstellung des Nullvektors als Linearkombination von Basisvektoren des Vektorraums V. Also müssen auch alle restlichen λ_i sowie die sämtlichen μ_j verschwinden.

Somit ist $\{\vec{b}_1, \ldots, \vec{b}_n, \vec{c}_{r+1}, \ldots, \vec{c}_m\}$ eine Basis von $U + V$, d.h.

$$\dim(U + V) = n + (m - r) = \dim U + \dim V - \dim(U \cap V) .$$

■

Bei Untervektorräumen ist damit alles klar; kehren wir zurück zur den affinen Unterräumen; wir betrachten also zwei affinen Unterräume (A, U) und (B, V) eines affinen Raums (C, W) sowie den davon erzeugten affinen Unterraum (S, T).

Falls A und B nichtleeren Durchschnitt haben, können wir Koordinatensysteme mit einem gemeinsamen Nullpunkt finden und W ist der Vektorraum $U + V$, d.h. in diesem Fall ist

$$\begin{aligned} \dim S = \dim T &= \dim(U + V) = \dim U + \dim V - \dim(U \cap V) \\ &= \dim A + \dim B - \dim(A \cap B) , \end{aligned}$$

denn $A \cap B$ hat ein Koordinatensystem mit demselben Nullpunkt und daher ist der zugehörige Untervektorraum $U \cap V$.

Falls A und B aber disjunkt sind, müssen wir verschiedene Punkte O_A und O_B als Nullpunkte der entsprechenden Koordinatensysteme wählen; jetzt ist der Vektorraum T des von A und B aufgespannten affinen Raums (S, T) ganz sicher nicht mehr $U + V$, denn er muß nun auch den Verbindungsvektor von O_A und O_B enthalten. Wäre dieser als Summe $\vec{u} + \vec{v}$ von Vektoren $\vec{u} \in U$ und $\vec{v} \in V$ darstellbar, könnten wir den Vektor $\vec{u}$ an O_A abtragen und $-\vec{v}$ an O_B; beide Vektoren

hätten denselben Endpunkt, der somit im (leeren) Durchschnitt von A und B liegen müßte. Also ist T mindestens gleich $U+V$ plus dem vom Verbindungsvektor aufgespannten eindimensionalen Untervektorraum. Das reicht aber auch schon, denn nimmt man etwa O_A als Ursprung des Koordinatensystems, kommt man nun durch Abtragen der Vektoren aus diesem Untervektorraum zu jedem Punkt von A oder B. Also ist in diesem Fall

$$\begin{aligned}\dim S &= \dim T = 1 + \dim(U+V)\\ &= \dim U + \dim V - \dim(U\cap V) + 1\,.\end{aligned}$$

Insgesamt haben wir damit gezeigt

Lemma: (A,U) und (B,V) seien affine Unterräume eines affinen Raum (C,W), und (S,T) sei der von beiden erzeugte affine Unterraum. Dann ist

$$\dim S = \begin{cases}\dim A + \dim B - \dim(A\cap B) & \text{falls } A\cap B \neq \emptyset\\ \dim A + \dim B - \dim(U\cap V) + 1 & \text{falls } A\cap B = \emptyset\,.\end{cases}$$

∎

Betrachten wir als Beispiel die beiden Unterräume

$$A = \left\{ (1,0,1,0,1) + \lambda \begin{pmatrix}1\\2\\3\\4\\5\end{pmatrix} + \mu \begin{pmatrix}5\\4\\3\\2\\1\end{pmatrix} \;\middle|\; \lambda,\mu\in\mathbb{R} \right\}$$

und

$$B = \left\{ (0,1,0,1,0) + \lambda \begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix} + \mu \begin{pmatrix}0\\1\\4\\9\\16\end{pmatrix} \;\middle|\; \lambda,\mu\in\mathbb{R} \right\}$$

von $\mathbb{R}^5$.

Der Untervektorraum zu A ist

$$\begin{aligned}U &= \left[\begin{pmatrix}1\\2\\3\\4\\5\end{pmatrix}, \begin{pmatrix}5\\4\\3\\2\\1\end{pmatrix} \right] = \left[\begin{pmatrix}1\\2\\3\\4\\5\end{pmatrix}, \begin{pmatrix}6\\6\\6\\6\\6\end{pmatrix} \right] = \left[\begin{pmatrix}1\\2\\3\\4\\5\end{pmatrix}, \begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix} \right]\\ &= \left[\begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix}, \begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix} \right],\end{aligned}$$

der zu B ist

$$V = \left[\begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix}, \begin{pmatrix}0\\1\\4\\9\\16\end{pmatrix} \right] = \left[\begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix}, \begin{pmatrix}0\\0\\2\\6\\12\end{pmatrix} \right].$$

Der Durchschnitt der beiden Untervektorräume ist also eindimensional und wird vom gemeinsamen Basisvektor aufgespannt.

Zur Berechnung des Durchschnitts von A und B müssen wir untersuchen, ob es reelle Zahlen $\lambda_1,\mu_1,\lambda_2,\mu_2$ gibt, so daß

$$(1,0,1,0,1)+\lambda_1 \begin{pmatrix}1\\2\\3\\4\\5\end{pmatrix} +\mu_1 \begin{pmatrix}5\\4\\3\\2\\1\end{pmatrix} = (0,1,0,1,0)+\lambda_2 \begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix} +\mu_2 \begin{pmatrix}0\\1\\4\\9\\16\end{pmatrix}$$

ist. Mit den gerade berechneten neuen Basen von U und V können wir stattdessen auch das etwas einfachere Gleichungssystem

$$(1,0,1,0,1)+\lambda_3 \begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix} +\mu_3 \begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix} = (0,1,0,1,0)+\lambda_4 \begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix} +\mu_4 \begin{pmatrix}0\\0\\2\\6\\12\end{pmatrix}$$

betrachten mit neuen Variablen $\lambda_3,\mu_3,\lambda_4$ und μ_4, denn der Durchschnitt zweier affiner Unterräume hängt natürlich nicht vom Koordinatensystem

ab. Hier können wir alle Basisvektoren auf die linke Seite bringen und die beiden Basispunkte in Gestalt ihres Verbindungsvektors auf die rechte; mit den neuen Variablen $\lambda = \lambda_3$, $\mu = \mu_3 - \lambda_4$ und $\nu = -\mu_4$ wird das Gleichungssystem dann zu

$$\lambda \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \mu \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} + \nu \begin{pmatrix} 0 \\ 0 \\ 2 \\ 6 \\ 12 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} .$$

Bei diesem System kann das für die erste Komponente nur dann gutgehen, wenn $\lambda = -1$ ist; setzen wir diesen Wert ein und betrachten dann die zweite Komponente, folgt, daß $\mu = 2$ sein muß. Für ν bleibt noch das Gleichungssystem

$$\nu \begin{pmatrix} 0 \\ 0 \\ 2 \\ 6 \\ 12 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -4 \\ -4 \\ -8 \end{pmatrix} ,$$

das offensichtlich unlösbar ist. Damit war auch das ursprüngliche Gleichungssystem unlösbar, wir sind hier also im Fall $A \cap B = \emptyset$.

Nach obiger Formel hat das Erzeugnis von A und B daher die Dimension

$$\dim A + \dim B - \dim(U \cap V) + 1 = 2 + 2 - 1 + 1 = 4 ;$$

wie wir oben gesehen haben, enthält der zugehörige Untervektorraum außer $U + V$ auch noch den Verbindungsvektor

$$\begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$$

der Nullpunkte $(0, 1, 0, 1, 0)$ von A und $(1, 0, 1, 0, 1)$ von B.

Da wir $U + V$ und auch den Verbindungsvektor kennen, läßt sich das Erzeugnis von A und B nun leicht explizit angeben: Es ist der affine

Unterraum

$$\left\{ (1,0,1,0,1) + \lambda \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \mu \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} + \nu \begin{pmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \end{pmatrix} + \rho \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \;\middle|\; \begin{matrix} \lambda, \mu, \nu, \\ \rho \in \mathbb{R} \end{matrix} \right\} .$$

h) Ausblick: Numerische Lösung linearer Gleichungssysteme

Mancher Leser mag sich in die Mittelstufe zurückversetzt gefühlt haben, als bei den Beispielen dieses Paragraphen plötzlich Brüche auftauchten anstelle der Dezimalzahlen, mit denen der „echte Praktiker“ seinen Taschenrechner füttert.

Tatsächlich sollte der GAUSS-Algorithmus, so wie er hier vorgestellt wurde, nur für Körper benutzt werden, in denen man *exakt* rechnen kann; beim Rechnen mit reellen Zahlen, vor allem wenn es per Computer oder Taschenrechner geschieht, muß man sich meist mit Näherungslösungen begnügen, und leider können in ungünstigen Fällen selbst kleine Rundungsfehler große Auswirkungen auf das Ergebnis haben.

Der Umgang mit Gleitkommazahlen ist Gegenstand der Numerik (und teilweise auch der Praktischen Informatik); hier sei nur anhand weniger Beispiele gezeigt, daß Probleme auftreten *können:*

Beginnen wir mit dem Gleichungssystem

$$\begin{aligned} 1{,}35x &- 2{,}768y &&= -10 \\ 4{,}241x &- 8{,}69y &&= -31{,}4 + \varepsilon . \end{aligned}$$

Seine Lösung liegt für $\varepsilon = 0$ bei

$$x \approx 41{,}527 \quad \text{und} \quad y \approx 23{,}866 .$$

Stört man aber die rechte Seite nur im 0,01 in der zweiten Gleichung, ersetzt man dort also die 31,4 durch 31,41, wird die Lösung zu

$$x \approx 74{,}558 \quad \text{und} \quad y \approx 39{,}976 ,$$

und ersetzt man sie gar durch 31,5, erhält man

$$x \approx 371{,}838 \quad \text{und} \quad y \approx 184{,}964 .$$

Schon kleine Störungen der Konstanten, wie sie durch allfällige Rundungsfehler immer wieder auftreten, können also das Ergebnis stark verfälschen. Im vorliegenden Fall läßt sich dieser Effekt leicht quantifizieren: Für allgemeines ε hat das obige Gleichungssystem die Lösung

$$x \approx 41{,}527 + 3303{,}10\varepsilon \quad \text{und} \quad y \approx 23{,}866 + 1610{,}98\varepsilon\,,$$

jede Störung der rechten Seite der zweite Gleichung führt also zu einer mehr als 3000-fachen Störung des Werts von x.

In diesem einfachen Beispiel kann man relativ schnell sehen, was passiert; außerdem deuten die im Vergleich zur rechten Seite etwas kleinen Koeffizienten auf der linken Seite darauf hin, daß es Probleme geben könnte. Leider ist die Situation nicht immer so klar: Beim linearen Gleichungssystem

$$\begin{aligned}
\frac{x_1}{2} + \frac{x_2}{3} + \frac{x_3}{4} + \frac{x_4}{5} + \frac{x_5}{6} + \frac{x_6}{7} &= 1\\
\frac{x_1}{3} + \frac{x_2}{4} + \frac{x_3}{5} + \frac{x_4}{6} + \frac{x_5}{7} + \frac{x_6}{8} &= 1\\
\frac{x_1}{4} + \frac{x_2}{5} + \frac{x_3}{6} + \frac{x_4}{7} + \frac{x_5}{8} + \frac{x_6}{9} &= 1\\
\frac{x_1}{5} + \frac{x_2}{6} + \frac{x_3}{7} + \frac{x_4}{8} + \frac{x_5}{9} + \frac{x_6}{10} &= 1\\
\frac{x_1}{6} + \frac{x_2}{7} + \frac{x_3}{8} + \frac{x_4}{9} + \frac{x_5}{10} + \frac{x_6}{11} &= 1\\
\frac{x_1}{7} + \frac{x_2}{8} + \frac{x_3}{9} + \frac{x_4}{10} + \frac{x_5}{11} + \frac{x_6}{12} &= 1
\end{aligned}$$

sind die Koeffizienten im Vergleich zur rechten Seite durchaus in vernünftigen Größenordnungen, und die Lösung

$$\begin{aligned}
x_1 &= \text{-}42, & x_2 &= 840, & x_3 &= \text{-}5040,\\
x_4 &= 12600, & x_5 &= \text{-}13860, & x_6 &= 5544
\end{aligned}$$

ist auch nicht sonderlich exotisch.

Berechnen wir die Lösung allerdings numerisch mit nur drei geltenden Ziffern, so erhalten wir die numerische Lösung

$$\begin{aligned}
x_1 &= \text{-}19{,}0, & x_2 &= 171 & x_3 &= \text{-}359,\\
x_4 &= 216 & x_5 &= \text{-}83{,}8, & x_6 &= 98{,}2,
\end{aligned}$$

die mit der korrekten Lösung offensichtlich nicht viel zu tun hat.

Bei diesem wie bei allen folgenden Beispielen von Gleitkommarechnungen wird ein Leser, der die Ergebnisse überprüfen möchte, mit hoher Wahrscheinlichkeit andere als die angegebenen Zahlenwerte erhalten. Der Grund liegt darin, daß das Ergebnis einer Rechnung mit Gleitkommazahlen nicht nur von der Anzahl der mitgeführten Dezimalstellen abhängt, sondern auch von der Reihenfolge der Rechenschritte. Das Assoziativgesetz gilt bei Gleitkommazahlen weder für die Addition noch für die Multiplikation: Beispielsweise ist bei drei geltenden Ziffern

$$(4{,}21 + 6{,}82) - 2{,}13 = 11{,}0 - 2{,}13 = 8{,}87\,,$$

aber

$$4{,}21 + (6{,}82 - 2{,}13) = 4{,}21 + 4{,}69 = 8{,}90\,,$$

wobei hier so gerechnet wurde, daß jedes Ergebnis *einer* Rechenoperation zunächst exakt bestimmt wurde und dann auf drei geltende Ziffern gerundet wurde. Die meisten heutigen Computer runden bei Gleitkommarechnungen auf diese Weise, auch wenn sie natürlich im allgemeinen nicht wirklich das exakte Zwischenergebnis berechnen. Es gibt aber Algorithmen, mit denen sich zu einer vorgegebenen Rechengenauigkeit eine Zahl berechnen läßt, die dasselbe Rundungsergebnis hat wie das exakte Zwischenergebnis. Taschenrechnern treiben nur selten einen so hohen Aufwand; hier können noch zusätzliche Fehler entstehen.

Die Nichtassoziativität von Gleitkommaoperationen hat noch eine weitere unerwartete Konsequenz: Zumindest bei Lösungsmethoden, die bei der Wahl des nächsten Eliminationsschritts auch Zufallsentscheidungen einfließen lassen (wie dies beispielsweise bei Maple der Fall ist), kann auch die mehrfache Lösung desselben Gleichungssystems mit demselben Algorithmus auf demselben Computer zu verschiedenen Ergebnissen führen, das Ergebnis ist also nicht nur nicht richtig, sondern sogar nicht einmal konsistent falsch.

Hier sind beispielsweise für das gerade betrachtete Beispiel die Ergebnisse aus zehn Lösungsversuchen mit Maple, gerechnet mit jeweils vier geltenden Ziffern:

x_1	x_2	x_3	x_4	x_5	x_6
−33,900	136,700	−120,800	510,500	−1455,000	1001,000
−20,400	235,900	−956,200	2039,000	−2304,000	1035,000
−33,900	136,700	−120,800	510,500	−1455,000	1001,000
−34,000	136,700	−120,000	509,700	−1456,000	1002,000
−37,200	172,500	−226,300	613,100	−1461,000	977,300
−34,000	136,700	−120,000	509,700	−1456,000	1002,000
17,600	−208,200	689,700	−636,100	−306,300	472,300
19,500	−249,300	874,000	−936,800	−120,000	442,500
8,000	−143,900	587,400	−692,700	−80,040	349,400
−20,400	235,900	−956,200	2039,000	−2304,000	1035,000

Egal wie stark die Zahlenwerte, die ein Leser beim Rechnen mit drei oder vier geltenden Ziffern bekommt, von den obigen abweichen, werden sie also mit ziemlicher Sicherheit eines gemeinsam haben: Sie haben nicht einmal entfernt mit den korrekten Zahlen zu tun und sind somit völlig unbrauchbar.

Eine Erhöhung der Ziffernzahl führt nur langsam zu besseren Ergebnissen: Bei einer Genauigkeit von vier bis fünfzehn Ziffern erhalten wir nacheinander die folgenden „Lösungen“:

Ziffern	x_1	x_2	x_3
4	-21,100	236,600	-940,200
5	8,550	-29,800	-458,970
6	19,806	-159,680	2,252
7	1,683	137,057	-1502,936
8	-33,359	699,124	-4325,108
9	-41,766	835,783	-5017,165
10	-42,004	840,010	-5039,853
11	-41,998	839,974	-5039,864
12	-41,999	839,988	-5039,936
13	-42,000	839,998	-5039,992
14	-42,000	840,000	-5039,999
15	-42,000	840,000	-5040,000
korrekt	-42	840	-5040

und

Ziffern	x_4	x_5	x_6
4	613,100	-1461,000	977,300
5	-312,910	-702,410	663,730
6	1427,120	-2813,060	1530,740
7	4960,798	-6457,628	2897,409
8	11062,780	-12367,780	5009,726
9	12418,887	-13683,876	5480,857
10	12599,672	-13859,628	5543,852
11	12600,048	-13860,032	5544,007
12	12599,855	-13859,858	5543,949
13	12599,981	-13859,982	5543,994
14	12599,998	-13859,998	5543,999
15	12600,000	-13860,000	5544,000
korrekt	12600	-13860	5544

Wie man sieht, werden die Ergebnisse erst ab zehnstelliger Genauigkeit halbwegs korrekt, h.h. trotz der kleinen Koeffizienten müßte man im vorliegenden Fall mit doppeltgenauen reellen Zahlen rechnen. (Bei einer Gleitkommaarithmetik nach IEEE Standard 754, wie sie in den meisten heutigen Prozessoren realisiert ist, sind Gleitkommazahlen einfacher Genauigkeit mit einem relativen Fehlerrn von bis zu etwa $5{,}96 \cdot 10^{-8}$ behaftet, was zwischen sieben und acht geltenden Ziffern entspricht; bei doppelter Genauigkeit liegt die Schranke bei etwa $1{,}11 \cdot 10^{-16}$; das entspricht nicht ganz sechzehn geltenden Ziffern.)

Daß selbst doppelte Genauigkeit nicht immer ausreicht, zeigt die Vergrößerung des obigen Gleichungssystems auf 15 Variable:

$$\begin{aligned}
\tfrac{x_1}{2}+\tfrac{x_2}{3}+\tfrac{x_3}{4}+\tfrac{x_4}{5}+\tfrac{x_5}{6}+\tfrac{x_6}{7}+\tfrac{x_7}{8}+\tfrac{x_8}{9}+\tfrac{x_9}{10}+\tfrac{x_{10}}{11}+\tfrac{x_{11}}{12}+\tfrac{x_{12}}{13}+\tfrac{x_{13}}{14}+\tfrac{x_{14}}{15}+\tfrac{x_{15}}{16} &= 1\\
\tfrac{x_1}{3}+\tfrac{x_2}{4}+\tfrac{x_3}{5}+\tfrac{x_4}{6}+\tfrac{x_5}{7}+\tfrac{x_6}{8}+\tfrac{x_7}{9}+\tfrac{x_8}{10}+\tfrac{x_9}{11}+\tfrac{x_{10}}{12}+\tfrac{x_{11}}{13}+\tfrac{x_{12}}{14}+\tfrac{x_{13}}{15}+\tfrac{x_{14}}{16}+\tfrac{x_{15}}{17} &= 1\\
\tfrac{x_1}{4}+\tfrac{x_2}{5}+\tfrac{x_3}{6}+\tfrac{x_4}{7}+\tfrac{x_5}{8}+\tfrac{x_6}{9}+\tfrac{x_7}{10}+\tfrac{x_8}{11}+\tfrac{x_9}{12}+\tfrac{x_{10}}{13}+\tfrac{x_{11}}{14}+\tfrac{x_{12}}{15}+\tfrac{x_{13}}{16}+\tfrac{x_{14}}{17}+\tfrac{x_{15}}{18} &= 1\\
\tfrac{x_1}{5}+\tfrac{x_2}{6}+\tfrac{x_3}{7}+\tfrac{x_4}{8}+\tfrac{x_5}{9}+\tfrac{x_6}{10}+\tfrac{x_7}{11}+\tfrac{x_8}{12}+\tfrac{x_9}{13}+\tfrac{x_{10}}{14}+\tfrac{x_{11}}{15}+\tfrac{x_{12}}{16}+\tfrac{x_{13}}{17}+\tfrac{x_{14}}{18}+\tfrac{x_{15}}{19} &= 1\\
\tfrac{x_1}{6}+\tfrac{x_2}{7}+\tfrac{x_3}{8}+\tfrac{x_4}{9}+\tfrac{x_5}{10}+\tfrac{x_6}{11}+\tfrac{x_7}{12}+\tfrac{x_8}{13}+\tfrac{x_9}{14}+\tfrac{x_{10}}{15}+\tfrac{x_{11}}{16}+\tfrac{x_{12}}{17}+\tfrac{x_{13}}{18}+\tfrac{x_{14}}{19}+\tfrac{x_{15}}{20} &= 1\\
\tfrac{x_1}{7}+\tfrac{x_2}{8}+\tfrac{x_3}{9}+\tfrac{x_4}{10}+\tfrac{x_5}{11}+\tfrac{x_6}{12}+\tfrac{x_7}{13}+\tfrac{x_8}{14}+\tfrac{x_9}{15}+\tfrac{x_{10}}{16}+\tfrac{x_{11}}{17}+\tfrac{x_{12}}{18}+\tfrac{x_{13}}{19}+\tfrac{x_{14}}{20}+\tfrac{x_{15}}{21} &= 1\\
\tfrac{x_1}{8}+\tfrac{x_2}{9}+\tfrac{x_3}{10}+\tfrac{x_4}{11}+\tfrac{x_5}{12}+\tfrac{x_6}{13}+\tfrac{x_7}{14}+\tfrac{x_8}{15}+\tfrac{x_9}{16}+\tfrac{x_{10}}{17}+\tfrac{x_{11}}{18}+\tfrac{x_{12}}{19}+\tfrac{x_{13}}{20}+\tfrac{x_{14}}{21}+\tfrac{x_{15}}{22} &= 1
\end{aligned}$$

$$
\begin{aligned}
\tfrac{x_1}{9}+\tfrac{x_2}{10}+\tfrac{x_3}{11}+\tfrac{x_4}{12}+\tfrac{x_5}{13}+\tfrac{x_6}{14}+\tfrac{x_7}{15}+\tfrac{x_8}{16}+\tfrac{x_9}{17}+\tfrac{x_{10}}{18}+\tfrac{x_{11}}{19}+\tfrac{x_{12}}{20}+\tfrac{x_{13}}{21}+\tfrac{x_{14}}{22}+\tfrac{x_{15}}{23} &= 1\\
\tfrac{x_1}{10}+\tfrac{x_2}{11}+\tfrac{x_3}{12}+\tfrac{x_4}{13}+\tfrac{x_5}{14}+\tfrac{x_6}{15}+\tfrac{x_7}{16}+\tfrac{x_8}{17}+\tfrac{x_9}{18}+\tfrac{x_{10}}{19}+\tfrac{x_{11}}{20}+\tfrac{x_{12}}{21}+\tfrac{x_{13}}{22}+\tfrac{x_{14}}{23}+\tfrac{x_{15}}{24} &= 1\\
\tfrac{x_1}{11}+\tfrac{x_2}{12}+\tfrac{x_3}{13}+\tfrac{x_4}{14}+\tfrac{x_5}{15}+\tfrac{x_6}{16}+\tfrac{x_7}{17}+\tfrac{x_8}{18}+\tfrac{x_9}{19}+\tfrac{x_{10}}{20}+\tfrac{x_{11}}{21}+\tfrac{x_{12}}{22}+\tfrac{x_{13}}{23}+\tfrac{x_{14}}{24}+\tfrac{x_{15}}{25} &= 1\\
\tfrac{x_1}{12}+\tfrac{x_2}{13}+\tfrac{x_3}{14}+\tfrac{x_4}{15}+\tfrac{x_5}{16}+\tfrac{x_6}{17}+\tfrac{x_7}{18}+\tfrac{x_8}{19}+\tfrac{x_9}{20}+\tfrac{x_{10}}{21}+\tfrac{x_{11}}{22}+\tfrac{x_{12}}{23}+\tfrac{x_{13}}{24}+\tfrac{x_{14}}{25}+\tfrac{x_{15}}{26} &= 1\\
\tfrac{x_1}{13}+\tfrac{x_2}{14}+\tfrac{x_3}{15}+\tfrac{x_4}{16}+\tfrac{x_5}{17}+\tfrac{x_6}{18}+\tfrac{x_7}{19}+\tfrac{x_8}{20}+\tfrac{x_9}{21}+\tfrac{x_{10}}{22}+\tfrac{x_{11}}{23}+\tfrac{x_{12}}{24}+\tfrac{x_{13}}{25}+\tfrac{x_{14}}{26}+\tfrac{x_{15}}{27} &= 1\\
\tfrac{x_1}{14}+\tfrac{x_2}{15}+\tfrac{x_3}{16}+\tfrac{x_4}{17}+\tfrac{x_5}{18}+\tfrac{x_6}{19}+\tfrac{x_7}{20}+\tfrac{x_8}{21}+\tfrac{x_9}{22}+\tfrac{x_{10}}{23}+\tfrac{x_{11}}{24}+\tfrac{x_{12}}{25}+\tfrac{x_{13}}{26}+\tfrac{x_{14}}{27}+\tfrac{x_{15}}{28} &= 1\\
\tfrac{x_1}{15}+\tfrac{x_2}{16}+\tfrac{x_3}{17}+\tfrac{x_4}{18}+\tfrac{x_5}{19}+\tfrac{x_6}{20}+\tfrac{x_7}{21}+\tfrac{x_8}{22}+\tfrac{x_9}{23}+\tfrac{x_{10}}{24}+\tfrac{x_{11}}{25}+\tfrac{x_{12}}{26}+\tfrac{x_{13}}{27}+\tfrac{x_{14}}{28}+\tfrac{x_{15}}{29} &= 1\\
\tfrac{x_1}{16}+\tfrac{x_2}{17}+\tfrac{x_3}{18}+\tfrac{x_4}{19}+\tfrac{x_5}{20}+\tfrac{x_6}{21}+\tfrac{x_7}{22}+\tfrac{x_8}{23}+\tfrac{x_9}{24}+\tfrac{x_{10}}{25}+\tfrac{x_{11}}{26}+\tfrac{x_{12}}{27}+\tfrac{x_{13}}{28}+\tfrac{x_{14}}{29}+\tfrac{x_{15}}{30} &= 1
\end{aligned}
$$

Die Koeffizienten der linken und der rechten Seiten unterscheiden sich hier höchstens um den (moderaten) Faktor dreißig, und auch die Anzahl der Gleichungen ist im Vergleich zu den Systemen mit mehreren Tausend Variablen, die in vielen Anwendungen auftreten, eher gering. Daß es trotzdem numerische Probleme gibt, sieht man an der folgenden Tabelle; sie zeigt die korrekten und die mit fünfzehnstelliger Genauigkeit berechneten Werte der Variablen x_i:

x_1	240	-3173,68185135567
x_2	-28560	228151,786322960
x_3	1113840	-5303344,76558434
x_4	-21162960	58391350,3363271
x_5	232792560	-353947335,741599
x_6	-1629547920	1247733630,21701
x_7	7682154480	-2469397672,05011
x_8	-25241364720	1965727944,18240
x_9	58896517680	2105462770,85717
x_{10}	-98160862800	-6442117291,59057
x_{11}	116008292400	4902607321,68726
x_{12}	-94915875600	1334651674,35581
x_{13}	51108548400	-4588357739,33491
x_{14}	-16287339600	2863533974,79321
x_{15}	2326762800	-619210289,163625

Offensichtlich hat das berechnete Ergebnis nichts mit dem tatsächlichen zu tun, und auch hier werden verschiedene Computer oder sogar (im Falle etwa von Maple) derselbe Computer bei verschiedenen Versuchen völlig verschiedene Ergebnisse liefern.

Beim Lösen von linearen Gleichungssystemen muß man also unbedingt

darauf achten, daß Rundungsfehler auf das absolut notwendige Minimum beschränkt werden, und selbst dann muß man noch sehr vorsichtig sein. In der Numerik werden daher oft Iterationsverfahren bevorzugt, da diese ihre Rundungsfehler (zumindest tendenziell) selbst korrigieren.

Hier seien nur zwei Faustregeln für den GAUSS-Algorithmus erwähnt:

1.) Die verschiedenen Gleichungen des Systems sollten Koeffizienten in ähnlichen Größenordnungen enthalten. Da sich nichts an der Lösungsmenge ändert, wenn man eine der Gleichungen mit einer von Null verschiedenen Konstanten multipliziert, kann man dies etwa dadurch erreichen, daß man alle Zeilenvektoren der Matrix des Gleichungssystems auf dieselbe Länge (z.B. Eins) normiert.

2.) Falls man ein Vielfaches einer Gleichung zu einer anderen addiert, sollte der Koeffizient, mit dem die erste Gleichung multipliziert wird, möglichst klein sein, damit die zweite Gleichung nicht zu stark gestört wird. Dies läßt sich etwa dadurch erreichen, daß man bei der Elimination einer Variablen genau die Gleichung unverändert läßt, in der diese Variable mit dem betragsgrößten Koeffizienten vorkommt.

Für alles weitere sei auf die *Numerik* verwiesen; wer nicht so lange warten möchte, findet sehr viel mehr zu diesem Thema, als in jede Numerikvorlesung paßt, in der einschlägigen Spezialliteratur, z.B. im klassischen Buch

J.H. WILKINSON: Rounding errors in algebraic processes, *Prentice Hall,* 1963 oder *Dover,* 1994

oder in neueren Büchern wie

MICHAEL L. OVERTON: Numerical Computing with IEEE Floating Point Arithmetic – *Including One Theorem, One Rule of Thumb and One Hundred and One Exercises,* SIAM, 2001,

wo die Probleme des numerischen Rechnens in sehr elementarer und praxisnaher Darstellung behandelt werdet; mehr theoretischen Hintergrund findet man bei

FRANÇOISE CHAITIN-CHATELIN, VALÉRIE FRAYSSÉ: Lectures on finite precision computations, SIAM, 1996

und in dem sehr ausführlichen Buch

NICHOLAS J. HIGHAM: Accuracy and stability of numerical algorithms, SIAM, 1996.

i) Matrixgleichungen und die Berechnung der Inversen

Wir haben inzwischen recht gut verstanden, wann eine Matrix invertierbar ist und können dies auch leicht rechnerisch überprüfen; wir kennen aber noch keine Methode, mit dem wir die inverse Matrix wirklich berechnen könnten. Wie wir gleich sehen werden, stellt und der GAUSS-Algorithmus eine entsprechendes Verfahren zur Verfügung, das uns nicht nur die Bestimmung der Inversen erlaubt, sondern allgemeiner die Berechnung der Lösungsmenge einer beliebigen (linearen) Matrixgleichung.

Wir gehen aus von einer $n \times m$-Matrix $A = (a_{ij}) \in k^{n\times m}$ und einer $p \times q$-Matrix $B = (b_{i\ell}) \in k^{n\times p}$; gesucht ist eine Matrix $X = (x_{j\ell}) \in k^{m\times p}$; gesucht ist eine Lösung der Matrixgleichung $AX = B$. Besonders interessant ist hier der Fall $n = m = p$ und $B = E$ die Einheitsmatrix; in diesem Fall ist die Lösung offensichtlich gerade die Matrix $X = A^{-1}$.

Sind

$$\vec{x}^{\,(\ell)} = \begin{pmatrix} x_{1\ell} \\ \vdots \\ x_{m\ell} \end{pmatrix} \quad \text{und} \quad \vec{b}^{\,(\ell)} = \begin{pmatrix} b_{1\ell} \\ \vdots \\ b_{n\ell} \end{pmatrix}$$

die Spaltenvektoren von X und B, so ist die Matrixgleichung $AX = B$ äquivalent zu den p linearen Gleichungssystemen

$$A\vec{x}^{\,(\ell)} = \vec{b}^{\,(\ell)} \quad \text{für} \quad \ell = 1, \ldots, p\,,$$

die allesamt dieselbe Matrix A haben. Falls diese Gleichungen alle lösbar sind, gibt es also eine Matrix X mit $AX = B$.

Bei der Lösung dieser Gleichungssysteme wird es im allgemeinen vorteilhaft sein, auch bei der Rücksubstitution alle rechten Seiten simultan zu behandeln, d.h. anstelle des Einsetzens werden links durch Zeilenoperationen so lange Koeffizienten eliminiert, bis in jeder Zeile möglichst nur noch ein einziger von Null verschiedener Eintrag steht. (Dies ist

natürlich nur dann erreichbar, wenn das Gleichungssystem höchstens eine Lösung hat.) Durch Division läßt sich der verbliebene Eintrag noch auf Eins normieren, so daß sich dann rechts leicht die Lösung ablesen läßt.

Speziell im Falle quadratisher Matrizen kann man bei eindeutig lösbaren Gleichungen durch Zeilenvertauschungen sogar erreichen, daß links die Einheitsmatrix steht; da alle angewandten Operationen die Matrixgleichung „links $\times$ X = rechts“ erhalten, steht dann rechts die Lösungsmatrix X.

Betrachten wir als Beispiel die Inversion der Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 7 \\ 8 & 9 & 12 \end{pmatrix}.$$

Wir müssen das lineare Gleichungssystem $A\vec{x} = \vec{b}$ lösen für die drei rechten Seiten

$$\vec{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Dazu wenden wir den GAUSS-Algorithmus simultan auf alle drei Gleichungssysteme an, indem wir die drei rechten Seiten nebeneinander schreiben, d.h. wir gehen aus von einer um alle drei rechten Seiten erweiterten Matrix

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 7 & 0 & 1 & 0 \\ 8 & 9 & 12 & 0 & 0 & 1 \end{array}\right)$$

und wenden darauf die üblichen Eliminationsschritte an:

Als erstes müssen wir in der ersten Spalte alle Koeffizienten bis auf einen zu Null machen; dazu subtrahieren wir das vier- *bzw.* achtfache der ersten Zeile von der zweiten *bzw.* dritten und erhalten

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -5 & -4 & 1 & 0 \\ 0 & -7 & -12 & -8 & 0 & 1 \end{array}\right).$$

Als nächstes sollte der zweite Eintrag der letzten Zeile eliminiert werden; dazu können wir $7/3$ der mittleren Zeile von der letzten subtrahieren

oder aber, um Nenner zu vermeiden, sieben mal die zweite Gleichung von dreimal der dritten subtrahieren; dies führt auf

$$\begin{pmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -5 & -4 & 1 & 0 \\ 0 & 0 & -1 & 4 & -7 & 3 \end{pmatrix}.$$

In dieser Gestalt lassen sich die drei Gleichungssysteme nun leicht durch Rücksubstitution lösen; wir wollen aber stattdessen lieber weiter mit Zeilenoperationen arbeiten. (Kurzes Nachdenken zeigt, daß dies tatsächlich nur eine andere Betrachtungsweise für genau dieselben Rechenoperationen ist.)

Subtrahieren wir fünfmal die dritte Zeile von der zweiten und addieren sie dreimal zur ersten, erhalten wir

$$\begin{pmatrix} 1 & 2 & 0 & 13 & -21 & 9 \\ 0 & -3 & 0 & -24 & 36 & -15 \\ 0 & 0 & -1 & 4 & -7 & 3 \end{pmatrix}.$$

Diese Matrix wird deutlich übersichtlicher, wenn wir die zweite Zeile durch -3 dividieren und die dritte durch -1 und schließlich noch zweimal die neue zweite Zeile von der ersten subtrahieren; sie bekommt dann die Form

$$\begin{pmatrix} 1 & 0 & 0 & -3 & 3 & -1 \\ 0 & 1 & 0 & 8 & -12 & 5 \\ 0 & 0 & 1 & -4 & 7 & -3 \end{pmatrix}.$$

Damit haben wir linke eine Einheitsmatrix und können die Lösungsvektoren sofort ablesen: Es sind gerade die rechten Seiten, d.h. die rechte Matrix

$$X = \begin{pmatrix} -3 & 3 & -1 \\ 8 & -12 & 5 \\ -4 & 7 & -3 \end{pmatrix}$$

ist die Lösung von $AX = E$, wovon man sich zur Vorsicht vielleicht noch einmal durch Nachrechnen überzeugen sollte.

Daß so etwas auch beim Rechnen mit Computern kein Schaden sein kann, zeigt das folgende Beispiel einer Matrixinversion in MATLAB (*Mat*rix *Lab*oratory), einen häufig verwendeten System für numerische Berechnungen mit Matrizen:


```
seiler@hilbert:~ $ matlab

                        < M A T L A B >
          Copyright 1984-2000 The MathWorks, Inc.
                Version 6.0.0.88 Release 12
                          Sep 21 2000

To get started, type one of these: helpwin, helpdesk, or
demo.
For Product information, visit www.mathworks.com.

>> A = [1,2,3;4,5,6;7,8,9]

A =
        1     2     3
        4     5     6
        7     8     9

>> B = inv(A)

Warning: Matrix is close to singular or badly scaled.
         Results may be inaccurate. RCOND = 1.541976e-18.
B =
     1.0e+16 *
      -0.4504    0.9007   -0.4504
       0.9007   -1.8014    0.9007
      -0.4504    0.9007   -0.4504
>> A*B
ans =
        2     0     0
        8     0    -4
       16     0     8
>> B*A
ans =
        0    -4     0
        0     8     0
        4     0     0
```

Wie man auch hier wieder einmal sieht, sind Computer zwar sehr nützliche Werkzeuge, aber eben nur für Anwender, die mitdenken und wissen, was sie tun. Computer können uns das Rechnen abnehmen, nicht aber das Denken. Wenn es um reelle Zahlen geht, kommt hinzu, daß Computer mit dieser überabzählbaren Menge genauso wenig umgehen können wie wir. Der beim numerischen Rechnen übliche Ausweg besteht darin, statt im Körper der reellen Zahlen in der Menge der Gleitkommazahlen zu rechnen. Dies ist aber grob fahrlässig, sofern man nicht sich nicht *vorher* genau überlegt, welche Genauigkeit man vom Ergebnis erwarten kann – was im allgemeinen durchaus nichttriviale Mathematik erfordert. Theoretisch gibt es einen Ausweg: Mit der sogenannten Intervallarithmetik lassen sich Berechnungen so durchführen, daß das Ergebnis ein Intervall ist, in dem das theoretisch richtige Resultat *mit Sicherheit* liegt. Leider werden dessen Schranken aber sehr schnell sehr pessimistisch, so daß das Intervall nach umfangreichen Rechnungen oft kaum noch praktisch verwertbare Information liefert.

j) Spezielle Matrizen

Matrizen mit vielen Einträgen werden schnell unhandlich, insbesondere wenn viele der Einträge von Null verschieden sind. In diesem Abschnitt wollen wir einige Matrizen spezieller Form betrachten und uns überlegen, ob und gegebenenfalls wie wir deren Gestalt beim Rechnen ausnutzen können.

1) Diagonalmatrizen: Nach der Nullmatrix und der Einheitsmatrix am einfachsten sind die Diagonalmatrizen:

Definition: Eine quadratische Matrix $D = (d_{ij}) \in k^{n\times n}$ heißt Diagonalmatrix, wenn sämtliche Einträge außerhalb der Diagonale verschwinden, d.h. $d_{ij} = 0$ für $i \neq j$.

Eine $n \times n$-Diagonalmatrix ist somit gegeben durch ihre n Diagonaleinträge d_{ii}, anstelle von n^2 Werten müssen also nur n gespeichert werden. Addition und Multiplikation von Diagonalmatrizen sind denkbar einfach: Wir müssen nur die einander entsprechenden Einträge addieren *bzw.* multiplizieren.

Auch mit Inversen gibt es keine Probleme: Offensichtlich ist der Rang einer Diagonalmatrix gleich der Anzahl nichtverschwindender Diagonaleinträge, die Matrix ist also genau dann invertierbar, wenn keiner der Diagonaleinträge verschwindet. Die inverse Matrix ist dann einfach die Diagonalmatrix mit den Inversen der Diagonaleinträge der gegebenen Matrix.

Die Matrix einer linearen Abbildung $\varphi: V \to V$ bezüglich einer Basis $\mathcal{B} = (\vec{b}_1, \ldots, \vec{b}_n)$ ist genau dann eine Diagonalmatrix, wenn jeder der Basisvektoren $\vec{b}_i$ von φ auf ein Vielfaches $\lambda_i \vec{b}_i$ von sich selbst abgebildet wird; alsdann ist die Abbildungsmatrix gleich der Diagonalmatrix mit Einträgen $\lambda_1, \ldots, \lambda_n$. Somit hängt es sehr von der Basis ab, ob die Abbildungsmatrix Diagonalgestalt hat oder nicht.

Die Matrix

$$A = \begin{pmatrix} 1 & 2 & 1 & -2 \\ 2 & 1 & -2 & 1 \\ 1 & -2 & 1 & 2 \\ -2 & 1 & 2 & 1 \end{pmatrix} \in \mathbb{R}^{4\times 4}$$

etwa ist ganz sicher keine Diagonalmatrix. Betrachten wir die lineare Abbildung

$$\varphi: \mathbb{R}^4 \to \mathbb{R}^4; \quad \vec{v} \mapsto A\vec{v}$$

aber bezüglich der Basis $\mathcal{B} = (\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4)$ mit

$$\vec{b}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \vec{b}_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \quad \vec{b}_3 = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \quad \text{und} \quad \vec{b}_4 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

so rechnet man leicht nach, daß

$$\varphi(\vec{b}_1) = A\vec{b}_1 = 2\vec{b}_1, \quad \varphi(\vec{b}_2) = 2\vec{b}_2, \quad \varphi(\vec{b}_3) = -4\vec{b}_3 \quad \text{und} \quad \varphi(\vec{b}_4) = 4\vec{b}_4$$

ist, bezüglich $\mathcal{B}$ hat φ also die Diagonalmatrix

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

als Abbildungsmatrix. Wenn keine schwerwiegenden anderen Gründe dagegensprechen, wird es bei umfangreichen Rechnungen mit der Matrix A meist eine gute Idee sein, statt mit der Standardbasis von $\mathbb{R}^4$ mit der Basis $\mathcal{B}$ zu rechnen.

Wir werden im weiteren Verlauf dieses Paragraphen noch lernen, wie man das Rechnen mit der Matrix A auf das Rechnen mit der sehr viel angenehmeren Diagonalmatrix D zurückführen kann und wie man, wo dies möglich ist, eine Basis $\mathcal{B}$ wie im obigen Beispiel finden kann. Im Augenblick wollen wir uns damit begnügen, die dabei auftretenden Begriffe zu definieren:

Definition: Ein Vektor $\vec{v} \in k^n \setminus \{\vec{0}\}$ heißt *Eigenvektor* der Matrix $A \in k^{n\times n}$, wenn es eine Zahl $\lambda \in k$ gibt, so daß $A\vec{v} = \lambda\vec{v}$ ist. Dieses λ bezeichnen wir als einen *Eigenwert* von A.

Falls es also zu einer Matrix A eine Basis aus Eigenvektoren gibt, können wir sie bezüglich dieser Basis als Diagonalmatrix darstellen; wie wir später sehen werden, gibt es solche Basen häufig, aber nicht immer.

Diagonalmatrizen sind nicht nur einfach, wenn man sie untereinander verknüpft, auch die Multiplikation einer Diagonalmatrix mit einer beliebigen anderen Matrix ist problemlos:

Sei etwa $A \in k^{n\times m}$ eine beliebige Matrix und $D \in k^{m\times m}$ eine Diagonalmatrix mit Diagonaleinträgen $d_1, \ldots, d_m$. Wir wollen das Produkt AD berechnen. Seine i-te Spalte ist das Produkt von AD mit dem i-ten Einheitsvektor $\vec{e}_i$ von k^m. Das Produkt $D\vec{e}_i$ ist entsprechend der i-te Spaltenvektor von D, also $d_i\vec{e}_i$, und $A\vec{e}_i$ ist der i-te Spaltenvektor $\vec{a}_i$ von A. Also ist

$$(AD)\vec{e}_i = A(D\vec{e}_i) = A(d_i\vec{e}_i) = d_i(A\vec{e}_i) = d_i\vec{a}_i\,,$$

die Matrix AD entsteht somit aus A einfach daraus, daß der i-te Spaltenvektor von A mit dem i-ten Diagonaleintrag d_i von D multipliziert wird für alle i.

Ist D stattdessen eine $n \times n$-Matrix, ist das Produkt DA definiert und wir können es auf dieselbe Weise berechnen: Zunächst ist $(DA)\vec{e}_i = D(A\vec{e}_i) = D\vec{a}_i$. Multiplikation einer Diagonalmatrix mit einem Vektor multipliziert die j-te Komponente dieses Vektors mit dem j-ten Diagonaleintrag von D, also wird der j-te Eintrag von $\vec{a}_i$ mit d_j multipliziert. Dies passiert für jeden Spaltenvektor $\vec{a}_i$, also entsteht die Matrix DA aus A, indem man für jedes j ihre j-te Zeile mit d_j multipliziert.

Als *Regel* können wir damit festhalten: Multipliziert man eine $n \times m$-Matrix A von *links* mit einer $n \times n$-Diagonalmatrix D mit Einträgen $d_1, \ldots, d_n$, so wird die i-te *Zeile* von A mit d_i multipliziert. Multipliziert man A von *rechts* mit einer $m \times m$-Diagonalmatrix D mit Einträgen $d_1, \ldots, d_m$, so wird die i-te *Spalte* von A mit d_i multipliziert.

2) Dreiecksmatrizen: Nachdem wir das Rechnen mit Diagonalmatrizen gut im Griff haben, können wir zu etwas komplizierteren Matrizen übergehen, z.B. den Dreiecksmatrizen:

Definition: Eine Matrix $A = (a_{ij}) \in k^{n\times n}$ heißt *untere Dreiecksmatrix,* falls $a_{ij} = 0$ für $i < j$; sie heißt *obere Dreiecksmatrix,* falls $a_{ij} = 0$ wann immer $i > j$.

Eine 2×2-Matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ ist also genau dann eine obere Dreiecksmatrix, wenn $a_{21} = 0$ ist, und genau dann eine untere Dreiecksmatrix, wenn $a_{12} = 0$ ist, d.h.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{pmatrix} \quad \textit{bzw.} \quad A = \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix}\,.$$

Zählen wir zunächst, wie viele Einträge einer $n \times n$-Dreiecksmatrix nach Definition verschwinden müssen: Für eine untere Dreiecksmatrix ist das in der ersten Spalte keiner, in der zweiten einer, *usw.;* in der n-ten Spalte schließlich sind es alle bis auf einen, also $n - 1$. Insgesamt sind also

$$0 + 1 + \cdots + (n-1) = \frac{(n-1)n}{2}$$

Einträge notwendigerweise gleich Null, so daß nur

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

Einträge gespeichert werden müssen. Für eine obere Dreiecksmatrix kommen wir auf dieselben Zahlen, wir müssen nur die Spalten in umgekehrter Reihenfolge betrachten.

Es ist klar, daß die Summen von oberen *bzw.* unteren Dreiecksmatrizen wieder obere *bzw.* untere Dreiecksmatrizen sind; für Produkte und – so sie existieren – inverse Matrizen müssen wir uns allerdings etwas mehr Gedanken machen.

Wie bei Diagonalmatrizen sieht man sofort, daß eine Dreiecksmatrix genau dann invertierbar ist, wenn keiner ihrer Diagonaleinträge verschwindet: Falls etwa der i-te Diagonaleintrag einer oberen Dreiecksmatrix verschwindet, liegen die ersten i Spaltenvektoren im von den ersten $i-1$ Koordinateneinheitsvektoren erzeugten $(i-1)$-dimensionalen Untervektorraum von k^n und können somit unmöglich zusammen mit den restlichen $n-i$ Spaltenvektoren einen n-dimensionalen Vektorraum aufspannen. Bei unteren Dreiecksmatrizen argumentiert man im wesentlichen genauso mit dem vom i-ten bis zum n-ten Spaltenvektor aufgespannten Untervektorraum.

Am einfachsten geht das, wenn wir wieder die lineare Abbildung

$$\varphi\colon k^n \to k^n; \quad \vec{v} \mapsto A\vec{v}$$

betrachten. Da in den Spalten der Abbildungsmatrix die Bilder der Koordinateneinheitsvektoren $\vec{e}_i$ stehen, muß im Fall einer oberen Dreiecksmatrix A der erste dieser Vektoren auf ein Vielfaches von sich selbst abgebildet werden, der zweite auf eine Linearkombination von $\vec{e}_1$ und $\vec{e}_2$ *usw.;* allgemein ist $A\vec{e}_i$ eine Linearkombination der Vektoren $\vec{e}_1, \ldots, \vec{e}_i$. Die lineare Abbildung φ bildet also jeden der Untervektorräume $U_i = [\vec{e}_1, \ldots, \vec{e}_i]$ auf sich selbst ab, und diese Bedingung reicht auch aus um sicherzustellen, daß die Abbildungsmatrix von φ bezüglich der Basis $(\vec{e}_1, \ldots, \vec{e}_n)$ eine obere Dreiecksmatrix ist.

Genau dann, wenn die Matrix invertierbar ist, haben wir eine bijektive Abbildung φ; diese ist auch bijektiv, wenn man sie auf die Untervektorräume $U_i = [\vec{e}_1, \ldots, \vec{e}_i]$ einschränkt, denn natürlich bleibt sie injektiv, und wie wir aus §2*i)* wissen, ist jede injektive Abbildung eines endlichdimensionalen Vektorraums auf sich selbst auch surjektiv.

Im Falle der unteren Dreiecksmatrizen haben wir im wesentlichen dasselbe Ergebnis, nur daß wir jetzt die Basisvektoren von hinten her betrachten müssen: A ist also genau dann eine untere Dreiecksmatrix, wenn φ jeden der Untervektorräume $V_i = [\vec{e}_i, \ldots, \vec{e}_n]$ auf sich selbst abbildet.

Als Charakterisierung einer Dreiecksmatrix ist das Kriterium, das wir gerade hergeleitet haben, sicherlich nicht sehr nuützlich: Die direkte Definition ist sehr viel einfacher nachzuprüfen. Dafür bekommen wir aber mit diesem Kriterium fast gratis das folgende

Lemma: *a)* Das Produkt zweier $\left\{\begin{matrix}\text{unterer}\\ \text{oberer}\end{matrix}\right\}$ Dreiecksmatrizen ist wieder eine $\left\{\begin{matrix}\text{untere}\\ \text{obere}\end{matrix}\right\}$ Dreiecksmatrix; deren Diagonaleinträge sind die Produkte der Diagonaleinträge der beiden Faktoren.

b) Eine $\left\{\begin{matrix}\text{untere}\\ \text{obere}\end{matrix}\right\}$ Dreiecksmatrix ist genau dann invertierbar, wenn keines ihrer Diagonalelemente verschwindet. Alsdann ist auch ihre inverse Matrix wieder eine $\left\{\begin{matrix}\text{untere}\\ \text{obere}\end{matrix}\right\}$ Dreiecksmatrix.

Beweis: a) A und B seien die beiden Matrizen, und

$$\varphi\colon k^n \to k^n; \quad \vec{v} \mapsto A\vec{v} \qquad \text{und} \qquad \psi\colon k^n \to k^n; \quad \vec{v} \mapsto A\vec{v}$$

seien die zugehörigen linearen Abbildungen.

Wie wir gerade gesehen haben, lassen sich die Eigenschaften „obere Dreiecksmatrix“ *bzw.* „untere Dreiecksmatrix“ dadurch charakterisieren, daß gewisse Untervektorräume U_i *bzw.* V_i von k^n auf sich sebst abgebildet werden. Ist aber für zwei Abbildungen φ und ψ sowohl $\varphi(U) \subseteq U$ als auch $\psi(U) \subseteq U$, so ist auch $(\varphi \circ \psi)(U) \subseteq U$, d.h. auch die Abbildungsmatrix AB von $\varphi \circ \psi$ bildet U auf sich selbst ab. Somit ist auch AB eine $\left\{\begin{matrix}\text{untere}\\ \text{obere}\end{matrix}\right\}$ Dreiecksmatrix.

Den i-ten Diagonaleintrag erhalten wir als $\vec{e}_i$-Komponente des Bilds von $\vec{e}_i$. Im Falle zweier oberer Dreiecksmatrix A, B mit i-tem Diagonaleintrag a_i *bzw.* b_i ist $B\vec{e}_i = b_i\vec{e}_i + \vec{u}_{i-1}$ mit $\vec{u}_{i-1} \in U_{i-1}$; da auch A sowohl U_i als auch U_{i-1} auf sich selbst abbildet, ist also auch

$(AB)\vec{e}_i = a_i b_i + \vec{w}_{i-1}$ mit $\vec{w}_{i-1} \in U_{i-1}$, d.h. die Diagonaleinträge werden miteinander multipliziert. Ganz entsprechend argumentiert man für untere Dreiecksmatrizen: Hier ist $A\vec{e}_i = a_i\vec{e}_i + \vec{v}_{i+1}$ mit $\vec{v}_{i+1} \in V_{i+1}$.

b) Eine Dreiecksmatrix A ist, wie wir oben gesehen haben, genau dann invertierbar, wenn keiner ihrer Diagonaleinträge verschwindet; alsdann ist auch jede der Abbildungen $U \to U;\ \vec{v} \mapsto A\vec{v}$, die eine $\left\{\begin{matrix}\text{untere}\\ \text{obere}\end{matrix}\right\}$ Dreiecksmatrix charakterisieren, bijektiv, also invertierbar. Somit bildet auch die lineare Abbildung $\psi: U \to k^n;\ \vec{v} \mapsto A^{-1}\vec{v}$ jeden der Untervektorräume U auf sich selbst ab, d.h, auch A^{-1} ist eine $\left\{\begin{matrix}\text{untere}\\ \text{obere}\end{matrix}\right\}$ Dreiecksmatrix. ∎

Um nur Teil *a)* dieses Lemmas zu beweisen, wäre der Umweg über lineare Abbildungen nicht notwendig gewesen; das hätten wir direkt auch billiger haben können. Beispielsweise ist für zwei untere Dreiecksmatrizen $A = (a_{i\ell})$ und $B = (b_{\ell j})$ mit Produkt $C = (c_{ij})$ nach Definition der Matrixmultiplikation

$$c_{ij} = \sum_{\ell=1}^{n} a_{i\ell} b_{\ell j}\,.$$

Für $i > j$ ist für jedes $\ell < i$ der Eintrag $a_{i\ell} = 0$, da A eine untere Dreiecksmatrix ist, und für $\ell \geq i > j$ ist $b_{\ell j} = 0$, da B eine untere Dreiecksmatrix ist. Also ist auch c_{ij} als Summe von lauter Nullen gleich Null. Für die Diagonalelemente erhalten wir

$$c_{ii} = \sum_{\ell=1}^{n} a_{i\ell} b_{\ell i}\,,$$

wobei für $\ell < i$ der Eintrag $a_{i\ell}$ verschwindet und für $\ell > i$ der Eintrag $b_{\ell i}$. Somit bleibt nur der Summand $a_{ii}b_{ii}$ übrig.

Ähnlich könnte man auch die entsprechende Aussage für obere Dreiecksmatrizen beweisen.

Für inverse Matrizen kennen wir jedoch keine brauchbaren Formeln; hier ist der Umweg über die linearen Abbildungen der einzige Beweis, der mit unseren Kenntnissen möglich ist, und selbst wenn man die

(ziemlich unangenehmen) Formeln kennt, nach denen man die Einträge von A^{-1} durch die von A ausdrücken kann, ist der obige Beweis kürzer und verständlicher.

3) Matrizen mit nur einem Eintrag: Bei der numerischen Behandlung partieller Differentialgleichungen oder auch in der Kontrolltheorie hat man es oft mit riesigen Matrizen zu tun, in denen dann aber nur wenige, unregelmäßig verteilte Einträge von Null verschieden sind. Man spricht hier von *spärlich besetzten Matrizen,* im Englischen *sparse matrices.* Natürlich speichert man eine solche Matrix nicht als ein Feld von $n \times m$ Einträgen, sondern als eine Liste von Tripeln (i, j, a_{ij}) zu solchen Indizes, für die $a_{ij} \neq 0$ ist. Ein eigenes Teilgebiet der numerischen Mathematik beschäftigt sich mit der Suche nach effizienten Algorithmen für solche Matrizen.

Hier wollen wir uns beschränken auf den Umgang mit den nach der Nullmatrix am spärlichsten besetzten Matrizen, also denen mit nur einem nichtverschwindenden Eintrag. Der Einfachheit halber setzen wir diesen Eintrag auf Eins; um einen anderen Wert a zu erhalten, müssen wir einfach die gesamte Matrix mit a multiplizieren.

Definition: Für eine vorgegebene Größe $n \times m$ bezeichnen wir mit E_{ij} jene $n \times m$-Matrix, die an der Stelle (i, j) den Eintrag Eins hat und sonst lauter Nullen:

$$E_{ij} = \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\!\!\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\!\!\begin{matrix} \\ \\ \\ j \\ \downarrow \\ \\ \\ \end{matrix}\!\!\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\!\!\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\!\!\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\!\!\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\!\!\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}$$

$$\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\!\!\begin{pmatrix} 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \end{pmatrix}\!\!\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{mat

des j-ten; dieser wird auf den i-ten Koordinateneinheitsvektor des k^n abgebildet. Für einen beliebigen Vektor $\vec{v} \in k^m$ ist somit $\varphi(\vec{v}) = v_j \vec{e}_i$ jener Vektor aus k^n, der an der i-ten Stelle den j-ten Eintrag von $\vec{v}$ stehen hat und sonst überall Nullen.

Für eine $n \times p$-Matrix A sei $\psi\colon k^p \to k^n$ die lineare Abbildung zu A; dann bildet ψ den ℓ-ten Koordinateneinheitsvektor von k^p ab auf den ℓ-ten Spaltenvektor von A, und dieser wird durch φ abgebildet auf $a_{j\ell}\vec{e}_i$. Die Abbildungsmatrix $E_{ij}A$ von $\varphi \circ \psi$ hat also als i-te Zeile die j-te Zeile von A; alle anderen Zeilen sind Null.

Für eine $q \times m$-Matrix B sei entsprechend $\omega\colon k^m \to k^q$ die lineare Abbildung; dann bildet $\omega \circ \varphi$ alle Koordinateneinheitsvektoren von k^n mit Ausnahme des j-ten auf den Nullvektor ab, da bereits φ diese Eigenschaft hat. Der j-te Koordinateneinheitsvektor wird von φ abgebildet auf den i-ten Koordinateneinheitsvektor von k^m, der wiederum von ω abgebildet wird auf den i-ten Spaltenvektor von B. Die Abbildungsmatrix BE_{ij} von $\omega \circ \varphi$ hat also als j-te Spalte die i-te Spalte von A; alle anderen Spalten sind Null.

Kurz können wir diese beiden Resultate zusammenfassen zu folgender Regel:

- Multiplikation von links mit E_{ij} führt zu einer Matrix, deren i-te Zeile die j-te Zeile der Ausgangsmatrix ist; alle anderen Zeilen sind Null.
- Multiplikation von rechts mit E_{ij} führt zu einer Matrix, deren i-te Spalte die j-te Spalte der Ausgangsmatrix ist; alle anderen Zeilen sind Null.

4) Permutationen und Permutationsmatrizen: Zu den einfachsten linearen Abbildungen gehören jene, die einfach die Reihenfolge der Basisvektoren ändern; um ihre Abbildungsmatrizen soll es hier gehen.

Definition: *a)* Eine Permutation ist eine bijektive Abbildung

$$\pi\colon \{1, \dots, n\} \to \{1, \dots, n\}\,.$$

Sie wird auch kurz als $\pi = \begin{pmatrix} 1 & 2 & \dots & n \\ \pi(1) & \pi(2) & \dots & \pi(n) \end{pmatrix}$ geschrieben.

b) Eine Transposition ist eine Permutation τ, die zwei Elemente von $\{1, \dots, n\}$ vertauscht und den Rest festläßt. Sind i, j die beiden Elemente, die von τ vertauscht werden, so schreiben wir kurz $\tau = (i\ j)$.

c) Die Menge aller Permutationen $\pi\colon \{1, \dots, n\} \to \{1, \dots, n\}$ wird als *symmetrische Gruppe* $\mathfrak{S}_n$ bezeichnet.

Die Matrixdarstellung einer Permutation hat natürlich nichts mit Abbildungsmatrizen von linearen Abbildungen zu tun; hier verwenden wir die Matrix nur als Darstellung der Wertetabelle von π.

Das Wort *Gruppe* tauchte bereits in §1 auf im Zusammenhang mit der Definition eines Vektorraum; es wird langsam Zeit, es wirklich zu definieren. Intuitiv versteht man unter einer Gruppe eine Struktur, deren Elemente sich so miteinander verknüpfen lassen, daß die „üblichen“ Rechenregeln gelten – mit Ausnahme, eventuell, des Kommutativgesetzes, denn dieses gilt ja beispielsweise schon bei der Matrixmultiplikation nicht mehr. Die exaktes Definition ist die folgende:

Definition: Eine *Gruppe* $(G, \circ)$ ist eine Menge G zusammen mit einer inneren Verknüpfung $\circ\colon G \times G \to G$, für die gilt:

1. $g \circ (h \circ k) = (g \circ h) \circ k$ für alle $g, h, k \in G$
2. Es gibt ein *Neutralelement* $e \in G$, so daß für alle $g \in G$ gilt $e \circ g = g \circ e = g$.
3. Zu jedem Element $g \in G$ gibt es ein *inverses Element* $g' \in g$, so daß $g \circ g' = g' \circ g = e$ ist.

Die Gruppe heißt *kommutativ* oder *abelsch,* wenn zusätzlich gilt

4. $g \circ h = h \circ g$ für alle $g, h \in G$.

Der norwegische Mathematiker NILS HENRIK ABEL (1802–1829) ist trotz seinen frühen Todes (an Tuberkulose) Initiator vieler Entwicklungen der Mathematik des neunzehnten Jahrhunderts; Begriffe wie abelsche Gruppen, abelsche Integrale, abelsche Funktionen, abelsche Varietäten, die auch in der heutigen Mathematik noch allgegenwärtig sind, verdeutlichen seinen Einfluß. Zu seinem 200. Geburtstag stiftete die norwegische Regierung zu seinen Ehren einen ABEL-Preis für Mathematik, der in ähnlicher Ausstattung und ähnlicher Weise wie die Nobelpreise verliehen wird. Erster Preisträger war 2003 JEAN-PIERRE SERRE (*1926).

Im Sinne dieser Definition bilden beispielsweise die ganzen Zahlen mit der Addition als Verknüpfung eine Gruppe, sogar eine abelsche Gruppe, nicht aber die natürlichen Zahlen, denn es gibt beispielsweise keine natürliche Zahl n, so daß $n + 2 = 0$ ist, und auch die Null

liegt zumindest nach der in dieser Vorlesung zugrundegelegten Konvention nicht in $\mathbb{N}$. Entsprechend bilden die ganzen Zahlen bezüglich der *Multiplikation* keine Gruppe, da die Gleichung $5x = 1$ in $\mathbb{Z}$ nicht lösbar ist, aber die positiven rationalen Zahlen bilden eine multiplikative Gruppe. Die rationalen Zahlen bilden keine, da $0x = 1$ unlösbar ist, aber die rationalen Zahlen ohne Null sind eine multiplikative Gruppe, genauso auch die *invertierbaren* $n \times n$-Matrizen.

Da wir die Menge aller Permutationen als *symmetrische Gruppe* bezeichnen, sollten auch diese eine Gruppe bilden.

Die Gruppenoperation ist natürlich die Hintereinanderausführung von Abbildungen, das Produkt zweier Permutationen π und ω ist also die Abbildung $\pi \circ \omega$, die eine Zahl i abbildet auf $(\pi \circ \omega)(i) = \pi(\omega(i))$.

Die Assoziativität der Verknüpfung ist, wie stets bei der Hintereinanderausführung von Abbildungen, klar; Einselement ist die identische Permutation, und Inverse gibt es, da eine Permutation nach Definition bijektiv ist und somit eine Umkehrabbildung hat. Speziell für Transpositionen ist die Bestimmung dieser Umkehrabbildung besonders einfach: Da eine Transposition nicht anderes tut, als zwei Elemente miteinander zu vertauschen, macht sie sich selbst rückgängig und ist somit ihr eigenes Inverses.

Wie meist bei der Hintereinanderausführung von Abbildungen ist auch bei Permutationen das Kommutativgesetz nicht erfüllt. Als Beispiel betrachten wir die beiden Transpositionen (1 2) und (1; 3).

Beim Produkt (1 2) ∘ (1 3) wird zuerst die Transposition (1 3) ausgeführt, die die Zahlen 1 und 3 vertauscht; sodann vertauscht (1 2) die Zahlen 1 und 2:

$$\begin{array}{ccc} 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ 3 & 2 & 1 \\ \downarrow & \downarrow & \downarrow \\ 3 & 1 & 2 \end{array}$$

Somit ist

$$(1\ 2) \circ (1\ 3) = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} .$$

Beim Produkt (1 3) ∘ (1 2) dagegen wird zuerst die Transposition (1 3) ausgeführt und dann erst (1 2); hier ist der Gang der Vertauschungen

also

$$\begin{array}{ccc} 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ 2 & 1 & 3 \\ \downarrow & \downarrow & \downarrow \\ 2 & 3 & 1 , \end{array}$$

d.h.

$$(1\ 3) \circ (1\ 2) = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} .$$

Aus der Darstellung

$$\begin{pmatrix} 1 & 2 & \dots & n \\ \pi(1) & \pi(2) & \dots & \pi(n) \end{pmatrix}$$

einer Permutation läßt sich leicht die Anzahl möglicher Permutationen von n Elementen ablesen: Füllen wir die untere Zeile von links aus systematisch auf, so gibt es für $\pi(1)$ noch die volle Auswahl unter allen n Elementen, für $\pi(2)$ kommen alle Elemente in Frage *außer* $\pi(1)$, also nur noch $n - 1$ Stück, für $\pi(3)$ gibt es entsprechend nur noch $n - 2$, bis es schließlich bei $\pi(n)$ überhaupt nichts mehr zu wählen gibt, denn $\pi(n)$ ist die einzige Zahl, die bis dahin noch nicht in der zweiten Zeile der Matrix steht. Insgesamt gibt es also

$$n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1 = n!$$

Permutationen von $\{1, \dots, n\}$, die symmetrische Gruppe $\mathfrak{S}_n$ enthält somit $n!$ Elemente.

Uns interessieren derzeit keine Abbildungen der Menge $\{1, \dots, n\}$ auf sich selbst, sondern lineare Abbildungen zwischen Vektorräumen; die bekommen wir, indem wir Permutationen auf die Basisvektoren anwenden. Wir betrachten also zu einer Permutation $\pi \in \mathfrak{S}_n$ die lineare Abbildung

$$\varphi_\pi : k^n \to k^n; \quad \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \mapsto \begin{pmatrix} v_{\pi(1)} \\ \vdots \\ v_{\pi(n)} \end{pmatrix} ;$$

ihre Abbildungsmatrix (bezüglich der Standardbasis von k^n) bezeichnen wir mit P_π. Da φ_π den i-ten Koordinateneinheitsvektor von k^n auf den

$\pi(i)$-ten abbildet, hat P_π in jeder Zeile und jeder Spalte genau einen nichtverschwindenden Eintrag: In der i-ten Spalte steht an der Position $\pi(i)$ eine Eins, überall sonst stehen Nullen. Der Eintrag $p_{k\ell}$ ist also genau dann gleich Eins, wenn $k = \pi(l)$ ist, ansonsten ist er Null.

Die Hintereinanderausführung der Abbildungen φ_π und damit die Multiplikation der Matrizen P_π ist vollständig bestimmt durch die Hintereinanderausführung der zugehörigen Permutationen: Wegen

$$(\varphi_\pi \circ \varphi_\omega)(\vec{e}_i) = \varphi_\pi\big(\varphi_\omega(\vec{e}_i)\big) = \varphi_\pi\big(\vec{e}_{\omega(i)}\big) = \vec{e}_{\pi(\omega(i))} = \vec{e}_{(\pi\circ\omega)(i)}$$

ist $\varphi_\pi \circ \varphi_\omega = \varphi_{\pi\circ\omega}$ und $P_\pi \cdot P_\omega = P_{\pi\circ\omega}$. Damit ist auch klar, daß die Matrizen P_π stets invertierbar sind: Da Permutationen als bijektive Abbildungen invertierbar sind, ist einfach $P_\pi{}^{-1} = P_{\pi^{-1}}$.

Auch die Multiplikation der Permutationsmatrix P_π mit einer beliebigen Matrix (passender Größe) läßt sich leicht ausführen: Für $\pi \in \mathfrak{S}_n$ und $A \in k^{n\times m}$ betrachten wir zu A die lineare Abbildung $\psi\colon k^m \to k^n$, die einen Vektor $\vec{v} \in k^m$ auf $A\vec{v}$ abbildet; dann ist $P_\pi A$ die Abbildungsmatrix von $\varphi_\pi \circ \psi$. Ihre Spalten sind die Bilder der Koordinateneinheitsvektoren von k^m. Der i-te dieser Vektoren wird von ψ auf den i-ten Spaltenvektor von A abgebildet, und auf dessen Komponenten wird die Permutation π angewandt. Dies passiert für jede Spalte von A, insgesamt entsteht $P_\pi A$ somit aus A, indem man die Permutation π auf dessen *Zeilen* anwendet. Für die Transposition $\pi = (12) \in \mathfrak{S}_3$ etwa ist

$$P_\pi = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

und

$$P_\pi \cdot \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 7 & 6 & 5 & 4 & 3 & 2 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

Entsprechend können wir für eine $p \times n$-Matrix B das Produkt BP_π ausrechnen: φ_π bildet den i-ten Koordinateneinheitsvektor von k^n ab auf den $\pi(i)$-ten, und dieser wird von der linearen Abbildung zu B abgebildet auf den $\pi(i)$-ten Spaltenvektor von B. Die Matrix BP_π entsteht also

aus B daraus, daß die Permutation π auf dessen *Spalten* angewandt wird. Beispielsweise ist für $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix}$

$$P_\pi = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

und

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 0 \end{pmatrix} \cdot P_\pi = \begin{pmatrix} 3 & 4 & 5 & 1 & 2 \\ 8 & 9 & 0 & 6 & 7 \end{pmatrix}.$$

k) Die LR-Zerlegung einer Matrix

Ein lineares Gleichungssystem $A\vec{x} = \vec{b}$ mit einer invertierbaren Matrix A läßt sich zumindest formal sehr leicht auflösen: Durch Multiplikation mit A^{-1} folgt, daß $\vec{x} = A^{-1} \cdot \vec{b}$ ist. Entsprechend einfach läßt sich eine Matrixgleichung der Form $AX = B$ auflösen: Hier ist $X = A^{-1}B$.

Für ein einziges lineares Gleichungssystem ist dieser Lösungsweg nicht sonderlich interessant, denn schon der Aufwand zur Berechnung von A^{-1} ist größer als der zur direkten Lösung des Gleichungssystems: Wir müssen dazu schließlich mehrere lineare Gleichungssysteme simultan lösen. Hat man aber viele lineare Gleichungssysteme, die sich nur durch ihre rechte Seiten unterscheiden, wie dies zum Beispiel bei linearen Steuerungsproblemen der Fall ist. so kann man sehr viel Zeit sparen, wenn man zunächst ein für alle Mal die inverse Matrix berechnet und dann die einzelnen Gleichungssysteme für den Preis einer Matrix-Vektor-Multiplikation lösen kann. Hinzu kommt, daß man beispielsweise bei eingebauten Steuerungen die Matrixinversion vorab auf einem leistungsfähigen Compter durchführen kann und dann für die eigentliche Steuerung nur die Matrix-Vektor-Multiplikation implementieren muß.

Falls A nicht invertierbar ist, falls man also beispielsweise eine Steuerung mit Redundanz hat, kann man nicht so vorgehen, aber eine fast

genauso effiziente Modifikation führt auch hier ans Ziel. Das entsprechende Verfahren wird je nach Lehrbuch als LR-Zerlegung oder LU-Zerlegung bezeichnet, wobei LR für *links/rechts* und LU für *lower/upper* steht.

Die Grundidee der LR-Zerlegung ist dieselbe wie die zur Berechnung der inversen Matrix: Wie wenden den GAUSS-Algorithmus simultan an auf mehrere rechte Seiten.

Um die Struktur der entstehenden Zerlegung besser zu verstehen, betrachten wir den wesentlichen Schritt des GAUSS-Algorithmus, die Addition von Vielfachen einer Gleichung zu einer anderen, als Multiplikation mit einer geeigneten Matrix.

Konkret sei $AX = B$ mit $A \in k^{n\times m}$ und $B \in k^{n\times p}$ eine Gleichung für die unbekannte Matrix $X \in k^{m\times p}$ über dem Körper k. Zur Lösung dieser Gleichung arbeiten wir mit Zeilenumformungen der erweiterten Matrix $M = (A \mid B) \in k^{n\times(m+p)}$, d.h. wir addieren ein Vielfaches einer Zeile zu einer anderen.

Angenommen, wir möchten das c-fache der j-ten Zeile zur i-ten addieren. Dazu betrachten wir die aus aus dem letzten Abschnitt bekannte Matrix $E_{ij} \in k^{n\times n}$, d.h. jene $n \times n$-Matrix, die an der Stelle (i, j) den Eintrag Eins hat und sonst lauter Nullen. Das Produkt $E_{ij}M$ hat, wie wir uns dort überlegt haben, als i-te Zeile die j-te Zeile von M und alle anderen Zeilen sind Null. Das Produkt $cE_{ij}M$ hat somit als i-te Zeile das c-fache der j-ten Zeile von M (und sonst lauter Nullzeilen). Setzen wir

$$Z_{ij}(c) = E + cE_{ij} \in k^{n\times n} ,$$

so transformiert die Addition des c-fachen der j-ten Zeile zur i-ten also die Gleichung $AX = B$ in die neue Gleichung

$$\big(Z_{ij}(c)A\big)X = Z_{ij}(c)B .$$

Da wir beim GAUSS-Algorithmus jeweils Vielfache von weiter oben stehenden Zeilen zu weiter unten stehenden addieren, ist hierbei stets $j < i$, d.h. $Z_{ij}(c)$ ist eine untere Dreiecksmatrix mit lauter Einsen in der Hauptdiagonale.

Führen wir nacheinander mehrere solche Eliminationsschritte aus, so wird M insgesamt mit einem Produkt von mehreren Matrizen der Form $Z_{ij}(c)$ multipliziert; da das Produkt von unteren Dreiecksmatrizen mit Einsen in der Hauptdiagonale wieder eine untere Dreiecksmatrix mit Einsen in der Hauptdiagonale ist, wird M also insgesamt mit einer solchen Matrix Z multipliziert: $(ZA)X = ZB$. Falls wir allein durch Zeilenumformungen die Endgestalt des GAUSS-Algorithmus erreichen können, gibt es also eine untere Dreiecksmatrix Z, die den Gesamteffekt dieser Zeilenumformungen beschreibt, und $ZA = R$ ist eine obere Dreiecksmatrix.

Diese Matrix Z ist bei einem vollbesetzten System aus n Gleichungen in m Unbekannten mit $m \geq n$ im allgemeinen ein Produkt von $\frac{1}{2}(n-1)(n-2)$ Dreiecksmatrizen, denn so viele Koeffizienten müssen wir eliminieren. Es ist klar, daß wir selbst für moderat große Werte von n eine alternative Berechnungsweise finden sollten.

Dazu betrachten wir den Spezialfall $B = E \in k^{n\times n}$, d.h. wir nehmen als rechte Seite eine $n \times n$-Einheitsmatrix. Dann führen wir, ohne uns weiter um die Matrizen $Z_{ij}(c)$ zu kümmern, Zeilenumformungen durch, wie wir es von linearen Gleichungssystemen und von der Matrixinversion her gewohnt sind, bis links eine obere Dreiecksmatrix erscheint.

Der Gesamteffekt dieser Umformungen kann auch so beschrieben werden, daß wir die Ausgangsgleichung $AX = E$ mit einer unteren Dreiecksmatrix Z mit Einsen in der Hauptdiagonale multipliziert haben, wobei als Koeffizientenmatrix eine obere Dreiecksmatrix $R = ZA$ entstand. Die umgeformte Gleichung ist also

$$(ZA)X = ZE \quad \text{oder} \quad RX = Z ,$$

d.h. wir können die Matrizen R und Z direkt ablesen: R steht dort, wo am Anfang A stand, und Z steht an der Stelle der Einheitsmatrix.

Die Gleichung $R = ZA$ läßt sich nach A auflösen, denn als untere Dreiecksmatrix mit Einsen in der Hauptdiagonalen ist Z insbesondere invertierbar, wobei auch die inverse Matrix $L = Z^{-1}$ eine untere Dreiecksmatrix mit Einsen in der Hauptdiagonalen ist. Damit haben wir

$$A = Z^{-1}R = LR$$

als Produkt einer unteren und einer oberen Dreiecksmatrix dargestellt; dies bezeichnet man als die LR-Zerlegung von A.

Im allgemeinen erreichen wir die Endgestalt beim GAUSS-Algorithmus allerdings nur, wenn wir zusätzlich zu den Eliminationsschritten auch noch Zeilenvertauschungen vornehmen. Diese entsprechen, wie wir im letzten Abschnitt gesehen haben, der Multiplikation mit Permutationsmatrizen, jetzt wird also der Gesamteffekt der Umformungen ein gemischtes Produkt aus Dreiecksmatrizen und Permutationsmatrizen beschrieben, und über solche Produkte können wir im allgemeinen nichts aussagen.

Wir können aber versuchen, die Zeilen von A gleich am Anfang so zu permutieren, daß anschließend keine Zeilenvertauschungen mehr nötig sind. In diesem Fall haben wir also A mit einer Permutationsmatrix P multipliziert und finden dann Dreiecksmatrizen Z, R und $L = Z^{-1}$, so daß

$$ZPA = R \quad \text{oder} \quad A = P^{-1}LR$$

ist. Strategien, um eine geeignete Permutation P zu finden, werden in der Numerik unter dem Stichwort *Pivotsuche* behandelt; bei kleinen Matrizen wird man sie im allgemeinen ohne Schwierigkeiten auch so finden.

Betrachten wir dazu ein konkretes Beispiel:

$$A = \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{pmatrix} .$$

Hier müssen wir vor der Anwendung des GAUSS-Algorithmus offensichtlich Zeilen vertauschen, z.B. die erste und die zweite. Dies entspricht einer Multiplikation mit

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

und führt auf

$$A' = PA = \begin{pmatrix} 3 & 4 & 5 \\ 0 & 1 & 2 \\ 6 & 7 & 8 \end{pmatrix} .$$

Wir schreiben die Einheitsmatrix daneben:

$$\left(\begin{array}{ccc|ccc} 3 & 4 & 5 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 6 & 7 & 8 & 0 & 0 & 1 \end{array}\right) .$$

Die Sechs links unten wird eliminiert durch Subtraktion der zweifachen ersten Zeile von der letzten:

$$\left(\begin{array}{ccc|ccc} 3 & 4 & 5 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & -1 & -2 & -2 & 0 & 1 \end{array}\right)$$

Die Endgestalt entsteht daraus, wenn man nun noch die zweite Zeile zur dritten addiert:

$$\left(\begin{array}{ccc|ccc} 3 & 4 & 5 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & -2 & 1 & 1 \end{array}\right)$$

Hier steht links die Matrix R, rechts steht Z, d.h.

$$R = \begin{pmatrix} 3 & 4 & 5 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{und} \quad Z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 1 & 1 \end{pmatrix} .$$

In der Tat rechnet man leicht nach, daß

$$\begin{aligned} ZPA &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{pmatrix} \\ &= \begin{pmatrix} 3 & 4 & 5 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} = R \end{aligned}$$

ist. Die Inversion von Z geht hier sehr schnell: Im Rechenschema

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ -2 & 1 & 1 & 0 & 0 & 1 \end{array}\right)$$

erhalten wir links die Einheitsmatrix, indem wir von der dritten Zeile zweimal die erste subtrahieren und und einmal die zweite addieren:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 2 & -1 & 0 \end{array}\right)$$

Somit ist

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 0 \end{pmatrix}$$

und $A = PLR$, denn als Permutationsmatrix zu einer Transposition ist P zu sich selbst invers.

Falls man die LR-Zerlegung einer quadratischen Matrix A kennt, folgt beispielsweise, daß A genau dann invertierbar ist, wenn R invertierbar ist, denn P und L sind immer invertierbar. Alsdann ist

$$A = P^{-1}LR \Longrightarrow A^{-1} = R^{-1}L^{-1}P = R^{-1}ZP\,.$$

Auch für unser Ausgangsproblem, die Lösung von Matrixgleichungen $AX = B$ ist die Kenntnis der LR-Zerlegung nützlich: Durch Multiplikation mit Z erhalten wir die neue Gleichung $RX = ZB$, die man wegen der Treppengestalt der linken Seite leicht durch sukzessives Einsetzen lösen kann.

l) Beispiel eines Basiswechsels

In den meisten (endlichdimensionalen) Vektorräumen, die wir bislang betrachtet hatten, gab es offensichtliche Basen wie etwa die Standardbasis aus den Koordinateneinheitsvektoren in $\mathbb{R}^n$ oder die reinen Potenzen bei Vektorräumen von Polynomen. Diese Basen sind allerdings für rechnerische Zwecke nicht immer optimal: Sowohl in vielen Anwendungen der Mathematik als auch innerhalb der Mathematik ist es oft günstiger, mit anderen Basen zu rechnen, in denen beispielsweise wichtige Matrizen zu Diagonal- oder Dreiecksmatrizen werden. Das letzte Thema diesen Paragraphen wird sich mit dem Problem befassen, Matrizen und Vektoren bezüglich einer Basis in eine andere Basis umzurechnen.

In diesem Abschnitt betrachten wir zur Einstimmung einen Vektorraum, in dem es keine irgendwie ausgezeichnete Basis gibt.

Wir gehen aus von dem Problem, Farben quantitativ zu charakterisieren: Physikalisch gesehen hängen Farben mit den verschiedenen Wellenlängen des sichtbaren Lichts zusammen, eine Farbe müßte also eigentlich durch eine Funktion auf dem Intervall von etwa 380 bis etwa

780 nm beschrieben werden, wobei keineswegs nur stetige Funktionen in Betracht kommen.

Nun werden Farben aber nur selten mit dem Spektrometer betrachtet; was wirklich interessiert ist der Eindruck auf das menschliche Auge. Dieses hat vier Arten von Photorezeptoren: Die sehr empfindlichen Stäbchen, die nur bei schwachem Licht eine Rolle spielen und die auch nur Helligkeitsinformationen liefern können, sowie drei Arten von weniger lichtempfindlichen Zäpfchen, k, ℓ und m, die bei gutem Licht ein Bild mit Farbinformation liefern, da jede Art ihre eigene spektrale Empfindlichkeitskurve hat. Klassische wie auch digitale Photographie sowie Farbdarstellungen auf Fernseh- und Computermonitoren beruhen allesamt darauf, daß dem Auge etwas vorgesetzt wird, was die k, ℓ und m-Zäpfchen zur gleichen Reaktion veranlaßt wie das „echte“ Farbsignal.

Damit reicht es aus, Farben als Elemente eines dreidimensionalen Raums zu beschreiben. Da sich die Empfindlichkeitsbereiche insbesondere der ℓ- und der m-Zäpfchen stark überlappen, wäre es allerdings weder sinnvoll noch sonderlich praktikabel, eine Basis über die Ausgabewerte der drei Arten von Zäpfchen zu definieren. Stattdessen werden je nach Hauptanwendungsziel mehrere Farbmodelle betrachtet, die ausgehend von den unterschiedlichsten Ansätzen allesamt denselben dreidimensionalen $\mathbb{R}$-Vektorraum beschreiben. Wir wollen uns die beiden einfachsten etwas genauer anschauen.

Am bekanntesten ist wohl das RGB-Modell, das Farben aus den drei Grundfarben Rot, Grün und Blau kombiniert. Basis des Vektorraums sind hierbei also drei Vektoren $\vec{r}, \vec{g}$ und $\vec{b}$, die einem Rot, Grün *bzw.* Blau einer vorgegebenen Frequenz und Intensität entsprechen, und alle anderen Farben werden als Linearkombinationen

$$R\vec{r} + G\vec{g} + B\vec{b}$$

dargestellt. Diese Darstellung ist insbesondere gut geeignet für die Farbdarstellung auf einem Monitor oder auch Fernsehschirm, wo Farben in genau dieser Weise erzeugt werden. Bei digitalen Bildern betrachtet man i.a. nur Koeffizienten R, G, B aus dem Intervall [0, 1], so daß gewisse Grenzhelligkeiten nicht überschritten werden können. Bei der

Darstellung mit einer Genauigkeit von acht Bit betrachtet man oft auch das 255-fache der entsprechenden Werte, gerundet zur nächsten ganzen Zahl.

Bei anderen Farbmodellen sieht man nicht so sehr auf die *Erzeugung* der Farben am Bildschirm, sondern auf deren Eigenschaften, wie sie ein menschlicher Betrachter wahrnimmt: Die Ausgabeimpulse der Zäpfchen werden noch im Sehapparat sofort weiterverarbeitet, und was wir bewußt zur Kenntnis nehmen, sind definitiv nicht die R-, G- und B-Anteile der Farben, sondern eher Helligkeiten, Farbsättigungen und ähnliche Eigenschaften.

Eine für die digitale Speicherung und Übermittlung interessante Tatsache ist, daß wir Helligkeiten viel feiner unterscheiden und viel höher auflösen können als Farbnuancen. Es liegt daher nahe, eine Basis zu wählen, in der auch die Gesamthelligkeit als Komponente auftritt, wobei diese Komponente entweder mit höherer Geauigkeit digitalisiert wird als die beiden anderen oder (häufiger) nur diese Komponente für *jedes* Pixel gespeichert wird, die beiden anderen aber nur für jedes zweite Pixel oder gar nur einmal pro Quadrat aus 2×2 Pixel. Die Wahl der Helligkeit als Basiskomponente hat auch den Vorteil, daß man dann in einfacher Weise diese Komponente für Schwarz-Weiß-Versionen der Bilder verwenden kann, was zum Beispiel beim Fernsehen eine wichtige Rolle spielt.

Wenn wir die Helligkeit als eine Basiskomponente wählen, stehen für die eigentliche chromatische Information nur noch zwei Basisvektoren zur Verfügung; diese können entweder aus zwei Farbanteilen des RGB-Systems abgeleitet werden oder (was für die Gestaltung photorealistischer Bilder gern angewandt wird) aus einer weiteren achromatischen Komponente wie etwa der Farbsättigung und nur *einer* chromatischen Komponente.

Bei den drei derzeit gebräuchlichen Fernsehstandards PAL, SECAM und NTSC sowie auch beim HDTV und beim JPEG-Format für digitale Bilder geht man den ersten Weg: Der erste Basisvektor $\vec{y}$ beschreibt die Helligkeit oder *Luminanz*, die beiden weiteren Vektoren sind gewichtete Differenzen zwischen dieser Helligkeit und dem Rot- oder Blauanteil

des Farbvektors. Die Gewichte unterscheiden sich dabei in den einzelnen Fällen; da für die Hörer dieser Vorlesung das JPEG-Format interessanter sein dürfte als Fernsehstandards, betrachten wir das dort verwendete YCbCr-Modell.

Die Helligkeit ist, wie in allen diesen Systemem, als gewichtetes Mittel der drei Farbanteile definiert; die Gewichte L_R, L_G und $L_B \in (0,\ 1)$ bezeichnet man als *Lumared, Lumagreen* und *Lumablue*. Mit diesen Bezeichnungen ist die Helligkeit

$$Y = L_R R + L_G G + L_B B \quad \text{mit} \quad L_R + L_G + L_B = 1\,;$$

Standardwerte sind

$$L_R = \frac{299}{1000}, \quad L_G = \frac{587}{1000} \quad \text{und} \quad L_B = \frac{114}{1000}\,.$$

Die beiden Chrominanzen sind festgelegt durch

$$C_b = \frac{B-Y}{2-2L_B} \quad \text{und} \quad C_r = \frac{R-Y}{2-2L_R}\,.$$

Damit haben wir eine neue Basis $\{\vec{y}, \vec{c}_b, \vec{c}_r\}$, mittels derer wir dieselben Farben beschreiben wie bezüglich der Basis $\{\vec{r}, \vec{g}, \vec{b}\}$.

Ganz offensichtlich ist es wichtig, zwischen den beiden Basen hin- und herrechnen zu können, denn schließlich werden auch JPEG-Bilder auf RGB-Monitoren betrachtet, und CCD Chips in Digitalkameras messen zunächst einmal RGB-Komponenten, aus denen dann oft ein JPEG-Bild erzeugt wird.

Im vorliegenden Fall ist es einfach, konkrete Formeln zu finden:

$$\begin{aligned}
Y &= L_R R + L_G G + L_B B\,,\\
C_b &= \frac{B-Y}{2-2L_B} = -\frac{L_R}{2-2L_B}\cdot R - \frac{L_G}{2-2L_B}\cdot G + \frac{1}{2}\cdot B\,,\\
C_r &= \frac{R-Y}{2-2L_R} = \frac{1}{2}\cdot R - \frac{L_G}{2-2L_R}\cdot G - \frac{L_B}{2-2L_R}\cdot B\,.
\end{aligned}$$

Auch die Umkehrung läßt sich leicht ausrechnen:

$$\begin{aligned}
R &= Y + (2 - 2L_R)C_r\,,\\
B &= Y + (2 - 2L_B)C_b \qquad \text{und}\\
G &= \frac{Y - L_R R - L_B B}{L_G}\\
&= \frac{(1 - L_R - L_B)Y - L_B(2 - 2L_B)C_b - L_R(2 - 2L_R)C_r}{L_G}\\
&= Y - \frac{L_B(2 - 2L_B)}{L_G}C_b - \frac{L_R(2 - 2L_R)}{L_G}C_r\,,
\end{aligned}$$

denn $1 - L_R - L_B = L_G$. Dies können wir auch mit Matrizen formulieren:

$$\begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} = \begin{pmatrix} L_R & L_G & L_B \\ \frac{-L_R}{2-2L_B} & \frac{-L_G}{2-2L_B} & \frac{1}{2} \\ \frac{1}{2} & \frac{-L_G}{2-2L_R} & \frac{-L_B}{2-2L_R} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$$

und

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} 1 & 0 & 2 - 2L_R \\ 1 & -\frac{L_B(2-2L_B)}{L_G} & -\frac{L_R(2-2L_R)}{L_G} \\ 1 & 2 - 2L_B & 0 \end{pmatrix} \begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix}\,.$$

Der Übergang zwischen den beiden Basen kann also jeweils als Multiplikation mit einer Matrix interpretiert werden, und natürlich sind die beiden zugehörigen Matrizen invers zueinander.

Wenn wir die erste der beiden Matrizen mit A bezeichnen, so sind die Spalten von A die Bilder der Einheitsvektoren des RGB-Systems, umgerechnet ins YCbCr-System; wir erhalten also die Beziehungen

$$\begin{aligned}
\vec{r} &= L_R\vec{y} - \frac{L_R}{2 - 2L_B}\vec{c}_b + \frac{1}{2}\vec{c}_r\\
\vec{g} &= L_G\vec{y} - \frac{L_G}{2 - 2L_B}\vec{c}_b - \frac{L_G}{2 - 2L_R}\vec{c}_r \quad \text{und}\\
\vec{b} &= L_B\vec{y} + \frac{1}{2}\vec{c}_b - \frac{L_B}{2 - 2L_R}\vec{c}_r\,.
\end{aligned}$$

Entsprechend sind die Spalten von A^{-1} die Bilder der Einheitsvektoren des YCbCr-Systems, umgerechnet ins RGB-System, d.h.

$$\begin{aligned}
\vec{y} &= \vec{r} + \vec{g} + \vec{b},\\
\vec{c}_b &= -\frac{2 - 2L_B}{L_G}\vec{g} + (2 - 2L_B)\vec{b} \quad \textit{und}\\
\vec{c}_r &= (2 - 2L_R)\vec{r} - \frac{2 - 2L_B}{L_G}\vec{g}\,.
\end{aligned}$$

Man beachte die Unterschiede zwischen der Darstellung der Basisvektoren in der jeweils anderen Basis und den Umrechnungsformeln für die Koeffizienten: Beim Umrechnen der RGB-Werte in YCbCr-Werte haben wir als Koeffizienten der einzelnen Gleichungen die *Zeilen* von A; die Basisvektoren $\vec{r}, \vec{g}, \vec{b}$ selbst sind aber im YCbCr-System ausgedrückt durch die *Spalten* der *inversen* Matrix A^{-1}.

Zur Verdeutlichung seien die Gleichungen nochmals angegeben mit numerischen Koeffizienten, näherungsweise berechnet für die Standardwerte von L_R, L_G und L_B: Damit ist

$$A = \begin{pmatrix} 0{,}2990 & 0{,}5870 & 0{,}1140 \\ -0{,}1687 & -0{,}3313 & 0{,}5000 \\ 0{,}5000 & -0{,}4187 & -0{,}0813 \end{pmatrix}$$

und

$$A^{-1} = \begin{pmatrix} 1 & 0 & 1{,}4020 \\ 1 & -0{,}3441 & -0{,}7141 \\ 1 & 1{,}7720 & 0 \end{pmatrix}\,,$$

also erhalten wir die Gleichungen

$$\begin{aligned}
Y &= 0{,}2990\,R + 0{,}5870\,G + 0{,}1140\,B\\
Cb &= -0{,}1687\,R - 0{,}3313\,G + 0{,}5000\,B\\
Cr &= 0{,}5000\,R - 0{,}4187\,G - 0{,}08131\,B
\end{aligned}$$

und

$$\begin{aligned}
\vec{r} &= 0{,}2990\,\vec{y} - 0{,}1687\,\vec{c}_b + 0{,}5000\,\vec{c}_r\\
\vec{g} &= 0{,}5870\,\vec{y} - 0{,}3313\,\vec{c}_b - 0{,}4187\,\vec{c}_r\\
\vec{b} &= 0{,}1140\,\vec{y} + 0{,}5000\,\vec{c}_b - 0{,}08131\,\vec{c}_r\,.
\end{aligned}$$

Entsprechend liefern Zeilen und Spalten von A^{-1} die Beziehungen

$$R = Y + 1{,}402\,Cr$$
$$G = Y - 0{,}3441\,Cb - 0{,}7141\,Cr$$
$$B = Y + 1{,}772\,Cb$$

und

$$\vec{y} = \vec{r} + \vec{g} + \vec{b}$$
$$\vec{c}_b = -0{,}3441\,\vec{g} + 10{,}772\,\vec{b}$$
$$\vec{c}_r = 1{,}402\,\vec{r} - 0{,}7141\,\vec{g}\,.$$

m) Basiswechsel im allgemeinen Fall

Gehen wir über zur allgemeinen Situation!

Wir gehen aus von einen n-dimensionalen k-Vektorraum V und betrachten darin beiden Basen

$$\mathcal{B} = \left(\vec{b}_1, \ldots, \vec{b}_n\right) \quad \text{und} \quad \mathcal{C} = \left(\vec{c}_1, \ldots, \vec{c}_n\right).$$

Da $\mathcal{B}$ eine Basis ist, lassen sich die Vektoren $\vec{c}_i$ als Linearkombinationen

$$\vec{c}_i = a_{i1}\vec{b}_1 + \cdots + a_{in}\vec{b}_n = \sum_{j=1}^{n} a_{ij}\vec{b}_j$$

der $\vec{b}_j$ schreiben; entsprechen lassen sich natürlich auch die $\vec{b}_j$ als Linearkombinationen

$$\vec{b}_j = m_{j1}\vec{c}_1 + \cdots + m_{jn}\vec{c}_n = \sum_{\ell=1}^{n} m_{j\ell}\vec{c}_\ell$$

schreiben. Dies können wir in die Darstellung von $\vec{c}_i$ einsetzen und erhalten

$$\vec{c}_i = \sum_{j=1}^{n} a_{ij}\vec{b}_j = \sum_{j=1}^{n} a_{ij} \sum_{\ell=1}^{n} m_{j\ell}\vec{c}_\ell = \sum_{\ell=1}^{n} \left(\sum_{j=1}^{n} a_{ij} m_{j\ell} \right) \vec{c}_\ell\,.$$

Da $\mathcal{C}$ eine Basis ist, bezüglich derer $\vec{c}_i$ die eindeutige Basisdarstellung

$$\vec{c}_i = 1 \cdot \vec{c}_i$$

hat, folgt

$$\sum_{j=1}^{n} a_{ij} m_{j\ell} = \begin{cases} 1 & \text{falls } i = \ell \\ 0 & \text{falls } i \neq \ell \end{cases}.$$

Fassen wir die a_{ij} zu einer Matrix A zusammen und die m_{ij} zu einer Matrix M, ist also $AM = E$, d.h. die beiden Matrizen sind invers zueinander (was eigentlich niemanden erstaunen sollte). Formal können wir dies schreiben als

$$\begin{pmatrix} \vec{c}_1 \\ \vdots \\ \vec{c}_n \end{pmatrix} = A \begin{pmatrix} \vec{b}_1 \\ \vdots \\ \vec{b}_n \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} \vec{b}_1 \\ \vdots \\ \vec{b}_n \end{pmatrix} = A^{-1} \begin{pmatrix} \vec{c}_1 \\ \vdots \\ \vec{c}_n \end{pmatrix},$$

wobei die $\vec{b}_j$ und $\vec{c}_i$ bei der Auswertung dieser Formel nur als Symbole zu betrachten sind: Wir wollen nicht wirklich einen Vektor von Vektoren definieren, sondern diese Formel einfach als kompakte Merkregel für den Zusammenhang zwischen den beiden Basen betrachten.

Beim konkreten Rechnen in Vektorräumen geht es nicht so sehr um die Basisvektoren selbst, sondern um die Koeffizienten der Basisdarstellung; sei also der Vektor

$$\vec{v} = v_1\vec{b}_1 + \cdots + v_n\vec{b}_n = w_1\vec{c}_1 + \cdots + w_n\vec{c}_n$$

bezüglich beider Basis dargestellt; wir suchen einen Zusammenhang zwischen den v_j und den w_i.

Mit der obigen Matrix $M = (m_{ij})$ ausgedrückt ist

$$\vec{v} = \sum_{j=1}^{n} v_j\vec{b}_j = \sum_{j=1}^{n} v_j \sum_{i=1}^{n} m_{ji}\vec{c}_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} v_j m_{ji} \right) \vec{c}_i\,,$$

also folgt wegen der Eindeutigkeit der Basisdarstellung

$$w_i = \sum_{j=1}^{n} v_j m_{ji} = \sum_{i=1}^{n} m_{ji} v_j\,.$$

Letzteres läßt sich leider nicht als Produkt von M mit dem Spaltenvektor der v_j schreiben: Die Indizes von m_{ji} haben die falsche Reihenfolge. Da die Formel trotzdem richtig und nützlich ist, definieren wir eine neue

Matrix, die aus der alten durch Vertauschung der Indizes, d.h. also von Zeilen und Spalten, entsteht:

Definition: Die *transponierte* Matrix $N = {}^tM$ zur einer Matrix $M = (m_{ij})_{\substack{i=1\dots n\\ j=1\dots m}} \in k^{n\times m}$ ist jene Matrix $N = (n_{ij})_{\substack{i=1\dots m\\ j=1\dots n}} \in k^{m\times n}$ mit $n_{ij} = m_{ji}$ für alle i, j.

Somit braucht man zur Umrechnung der Koeffizienten ineinander also die transponierte Matrix zu $M = A^{-1}$. Diese Matrix bezeichnen wir als die *Matrix des Basiswechsels;* sie wird auch gelegentlich als die zu A *kontragrediente* Matrix bezeichnet. Für sie gilt:

Satz: Sind $\mathcal{B} = (\vec{b}_1, \dots, \vec{b}_n)$ und $\mathcal{C} = (\vec{c}_1, \dots, \vec{c}_n)$ zwei Basen eines n-dimensionalen k-Vektorraums und ist

$$\vec{c}_i = \sum_{j=1}^{n} a_{ij}\vec{b}_j \quad \text{mit} \quad A = (a_{ij}) \in k^{n\times n}\,,$$

so berechnen sich für

$$\vec{v} = v_1\vec{b}_1 + \dots + v_n\vec{b}_n = w_1\vec{c}_1 + \dots + w_n\vec{c}_n$$

die Koeffizienten w_i der Basisdarstellung bezüglich $\mathcal{C}$ aus den v_j nach der Formel

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = {}^t\!\left(A^{-1}\right) \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}\,;$$

die Matrix des Basiswechsels ist also ${}^t\!\left(A^{-1}\right)$.

∎

Betrachten wir dazu ein Beispiel: Im Vektorraum $\mathcal{C}^0(\mathbb{R},\,\mathbb{R})$ aller stetiger Funktionen von $\mathbb{R}$ nach $\mathbb{R}$ sei der Untervektorraum U erzeugt von den beiden Funktionen e^x und e^{-x}, die wir zur Basis $\mathcal{B}$ von U zusammenfassen. Eine weitere Basis $\mathcal{C}$ von U bestehe aus den Funktionen $\sinh x$ und $\cosh x$. Dann ist

$$\sinh x = \frac{1}{2}e^x - \frac{1}{2}e^{-x} \quad \text{und} \quad \cosh x = \frac{1}{2}e^x + \frac{1}{2}e^{-x}\,,$$

die Matrix A, mittels derer wir $\mathcal{C}$ durch $\mathcal{B}$ ausdrücken, ist also

$$A = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}\,.$$

Die Berechnung der inversen Matrix ist hier nicht schwierig; man überzeugt sich leicht, daß gilt

$$A^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad \text{und somit} \quad {}^t\!\left(A^{-1}\right) = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\,.$$

Damit kennen wir die Matrix des Basiswechsels; ihr Produkt mit einem Koeffizientenvektor $\binom{a}{b} \in \mathbb{R}^2$ ist

$${}^t\!\left(A^{-1}\right)\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a-b \\ a+b \end{pmatrix}\,.$$

Also sollte gelten

$$ae^x + ae^{-x} = (a-b)\sinh x + (a+b)\cosh x\,,$$

was man in der Tat leicht nachrechnet.

Als nächstes wollen wir uns überlegen, daß oben in ${}^t\!\left(A^{-1}\right)$ die Klammern überflüssig waren; genauer gilt

Lemma: Für $A \in k^{n\times m}$ und $B \in k^{m\times p}$ ist ${}^t(AB) = {}^tB\,{}^tA$ und ${}^t\!\left(A^{-1}\right) = \left({}^tA\right)^{-1}$.

Beweis: Mit $A = (a_{ij})$ und $B = (b_{j\ell})$ hat AB an der Stelle $i\ell$ den Eintrag $\sum_{j=1}^{m} a_{ij}b_{j\ell}$, und ihre Transponierte hat denselben Eintrag an der Stelle ℓi.

${}^tB$ hat an der Stelle ℓj den Eintrag $b_{j\ell}$ und ${}^tA$ hat an der Stelle ji den Eintrag a_{ij}, das Produkt ${}^tB\,{}^tA$ hat somit an der Stelle ℓi den Eintrag $\sum_{j=1}^{m} b_{j\ell}a_{ij}$, was wegen der Kommutativität der Multiplikation in k mit dem oben berechneten Ausdruck übereinstimmt.

Damit ist die erste Formel bewiesen. Wenden wir sie an auf $B = A^{-1}$, so folgt die Beziehung

$${}^t\!\left(A^{-1}\right){}^tA = {}^t\!\left(AA^{-1}\right) = {}^tE = E\,,$$

die beiden Matrizen sind also invers zueinander. ∎

Als letztes Thema im Zusammenhang mit Basiswechseln wollen wir uns überlegen, wie sich die Abbildungsmatrix einer linearen Abbildung unter Basiswechseln verhält.

Wir betrachten also eine lineare Abbildung $\varphi: V \to W$; dabei sei V ein n-dimensionaler k-Vektorraum, und W sei ein m-dimensionaler. In jedem der beiden Vektorräume seien zwei Basen gegeben; in V seien dies

$$\mathcal{B} = (\vec{b}_1, \dots, \vec{b}_n) \quad \text{und} \quad \mathcal{C} = (\vec{c}_1, \dots, \vec{c}_n)\,,$$

in W seien es

$$\mathcal{D} = (\vec{d}_1, \dots, \vec{d}_m) \quad \text{und} \quad \mathcal{E} = (\vec{e}_1, \dots, \vec{e}_m)\,.$$

Um nicht ganz die Übersicht zu verlieren, ordnen wir jedem Vektor

$$\vec{v} = v_1\vec{b}_1 + \dots + v_n\vec{b}_n = w_1\vec{c}_1 + \dots + w_n\vec{c}_n$$

aus V die beiden Spaltenvektoren

$$\vec{v}_{\mathcal{B}} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \quad \text{und} \quad \vec{v}_{\mathcal{C}} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

aus k^n zu; entsprechend haben wir für jeden Vektor $\vec{w} \in W$ zwei Vektoren $\vec{w}_{\mathcal{D}}$ und $\vec{w}_{\mathcal{E}}$ aus k^m.

Die Abbildungsmatrix $M_{\mathcal{B},\mathcal{D}}$ von φ bezüglich der Basen $\mathcal{B}$ von V und $\mathcal{D}$ von W hat dann die Eigenschaft, daß

$$\varphi(\vec{v})_{\mathcal{D}} = M_{\mathcal{B},\mathcal{D}}\vec{v}_{\mathcal{B}}$$

ist; für die Abbildungsmatrix $M_{\mathcal{C},\mathcal{E}}$ von φ bezüglich der Basen $\mathcal{C}$ von V und $\mathcal{E}$ von W gilt entsprechend

$$\varphi(\vec{v})_{\mathcal{E}} = M_{\mathcal{C},\mathcal{E}}\vec{v}_{\mathcal{C}}\,.$$

Für den Übergang von einer Basis zur anderen rechnen wir der Einfachheit halber nicht mit der zu Beginn dieses Abschnitts betrachteten Matrix, die die Basisvektoren durcheinander ausdrückt, sondern gleich mit

der Matrix des Basiswechsels; die Matrizen $A \in k^{n\times n}$ und $B \in k^{m\times m}$ seien also so gewählt, daß für Vektoren $\vec{v} \in V$ und $\vec{w} \in W$ gilt

$$\vec{v}_{\mathcal{C}} = A\vec{v}_{\mathcal{B}} \quad \text{und} \quad \vec{w}_{\mathcal{E}} = B\vec{w}_{\mathcal{D}}\,.$$

Dann ist

$$\varphi(\vec{v})_{\mathcal{E}} = B\varphi(\vec{v})_{\mathcal{D}} = BM_{\mathcal{B},\mathcal{D}}\vec{v}_{\mathcal{B}} = BM_{\mathcal{B},\mathcal{D}}A^{-1}\vec{v}_{\mathcal{C}}\,,$$

d.h.

$$M_{\mathcal{C},\mathcal{E}} = BM_{\mathcal{B},\mathcal{D}}A^{-1}\,.$$

Im nächsten Semester werden wir vor allen den Fall $V = W$ oft benötigen; hier wählt man natürlich fast immer $\mathcal{D} = \mathcal{B}$ und $\mathcal{C} = \mathcal{E}$, so daß jede Abbildungsmatrix von nur *einer* Basis von V abhängt.

Dann ist auch $A = B$ und die obige Formel vereinfacht sich zu

$$M_{\mathcal{C}} = AM_{\mathcal{B}}A^{-1}\,,$$

wobei wir hier in $M_{\mathcal{B}} = M_{\mathcal{B},\mathcal{B}}$ den zweiten Index natürlich weglassen.

Betrachten wir als Beispiel die Differentiation im von e^x und e^{-x} erzeugten Vektorraum; ihre Abbildungsmatrix bezüglich dieser Basis $\mathcal{B}$ ist

$$M_{\mathcal{B}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\,.$$

Die Matrix des Basiswechsels Basis $\mathcal{C}$ aus $\sinh x$ und $\cosh x$ ist, wie wir oben gesehen haben,

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\,,$$

und

$$AM_{\mathcal{B}}A^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\,,$$

ganz in Übereinstimmung mit der Tatsache, daß die Differentation $\sinh x$ zu $\cosh x$ macht und umgekehrt.

§5: Euklidische und Hermitesche Vektorräume

In §2*a)* hatten wir, zur Definition von Vektoren im $\mathbb{R}^3$, Pfeile betrachtet und vereinbart, daß zwei Pfeile genau dann den gleichen Vektor darstellen sollen, wenn sie dieselbe Länge und (so diese Länge von Null verschieden ist) dieselbe Richtung haben. In der anschließenden Definition des Vektorraums allerdings kamen Längen und Richtungen nicht mehr vor – aus gutem Grund, denn es fällt in der Tat schwer, sich etwas vorzustellen unter der Länge eines Vektors über dem Körper $\mathbb{F}_{256}$. Über den reellen (und, mit Modifikationen) den komplexen Zahlen aber führen Längen und Richtungen zu interessanten Strukturen, deren Bedeutung weit über die Geometrie hinausgeht: Beispielsweise lassen sich Energie oder Leistung eines Signals oft interpretieren als eine Art Länge in einem unendlichdimensionalen Vektorraum; außerdem spielen solche verallgemeinerte Längen und Richtungen eine wichtige Rolle in der Fehler- und Ausgleichsrechnung sowie in der Statistik. Seit einiger Zeit werden werden Winkel auch bei der Informationssuche angewandt: Wenn etwa eine Suchmaschine nach den besten Dokumenten zu einer Anfrage sucht, berechnet sie dabei unter anderem die Winkel zwischen einem Dokumentenvektor und einem Anfragevektor – die beide in einem reellen Vektorraum liegen, dessen Dimension einige Millionen betragen kann.

a) Längen und Winkel in $\mathbb{R}^2$ und $\mathbb{R}^3$

Beginnen wir mit leichter vorstellbaren Längen und Winkeln.

Die Länge eines Vektors $\vec{v}$ im $\mathbb{R}^2$ läßt sich leicht nach dem Satz des PYTHAGORAS berechnen: Wir nehmen die beiden Koordinateneinheitsvektoren als Basis und schreiben bezüglich dieser Basis

$$\vec{v} = \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix} .$$

Wenn wir den Koordinateneinheitsvektoren, ihrem Namen entsprechend, die Länge eins zuordnen, haben die Vektoren auf der rechten Seite offensichtlich die Längen a und b. Außerdem bilden diese Vektoren zusammen mit dem Vektor $\vec{v}$ ein rechtwinkliges Dreieck; nach

PYTHAGORAS erfüllt die Länge c von $\vec{v}$ daher die Gleichung

$$c^2 = a^2 + b^2 \quad \text{oder} \quad c = \sqrt{a^2 + b^2} .$$

(Hier sieht man einen der Gründe, warum wir über beliebigen Körpern nicht von Längen sprechen: In Körpern wie $\mathbb{Q}$ existieren Quadratwurzeln nur ausnahmsweise, und in Körpern wie $\mathbb{C}$, in denen sie immer existieren, sind sie nicht eindeutig, da mit c stets auch $-c$ eine Wurzel aus c^2 ist. (Lediglich in Körpern, in denen (wie in $\mathbb{F}_{2^n}$) jedes Element gleich seinem Negativen ist, ist die Wurzel eindeutig.) In $\mathbb{R}$ gibt es zwar auch zwei Quadratwurzeln, aber da $\mathbb{R}$ im Gegensatz etwa zu $\mathbb{C}$ ein angeordneter Körper ist, können wir in konsistenter Weise eine der beiden auszeichnen, nämlich die nichtnegative.)

Für Vektoren im $\mathbb{R}^3$ müssen wir den Satz des PYTHAGORAS zweimal anwenden: Wir schreiben

$$\vec{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ b \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix} = \vec{u} + \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}$$

mit

$$\vec{u} = \begin{pmatrix} a \\ b \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ b \\ 0 \end{pmatrix} .$$

Dann bildet $\vec{u}$ zusammen mit den ersten beiden Vielfachen von Einheitsvektoren ein rechtwinkliges Dreieck, seine Länge d erfüllt also die Gleichung $d^2 = a^2 + b^2$. Da der Vektor $\vec{u}$ in der x, y-Ebene liegt, auf der der Vektor mit Komponenten $0, 0, c$ senkrecht steht, bilden auch $\vec{u}$, $\vec{v}$ und dieser Vektor ein rechtwinkliges Dreieck, so daß für die Länge e von $\vec{v}$ gilt

$$e^2 = d^2 + c^2 = a^2 + b^2 + c^2 \quad \text{oder} \quad e = \sqrt{a^2 + b^2 + c^2} .$$

Längen sind also problemlos berechenbar.

Zur Berechnung von Winkeln verwenden wir das (den meisten wohl bereits aus der Schule bekannte) Skalarprodukt. Es hat seinen Namen daher, daß es zwei Vektoren zu einem *Skalar* verknüpft; dieser wird mit $\vec{v} \cdot \vec{w}$ oder auch kurz $\vec{v}\vec{w}$ bezeichnet. Nach Definition ist

$$\vec{v} \cdot \vec{w} \underset{\text{def}}{=} |\vec{v}|\,|\vec{w}| \cos \measuredangle(\vec{v}, \vec{w}) ,$$

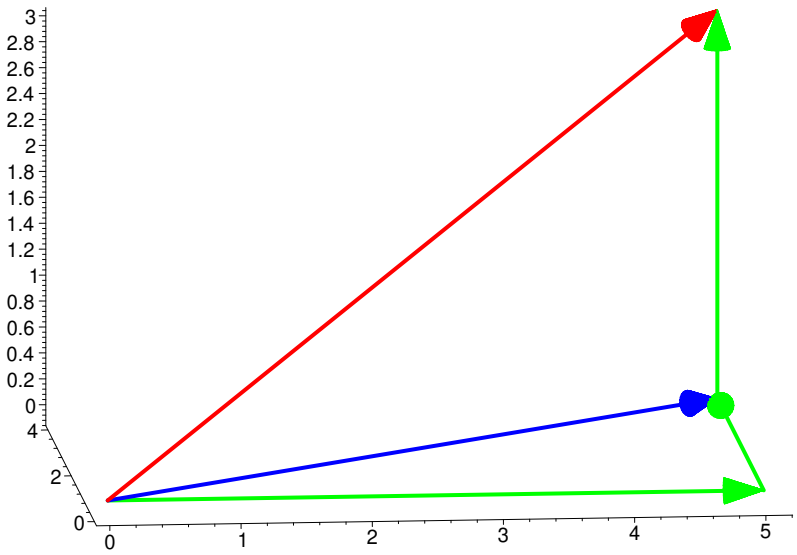


Abb. 13: Längenberechnung im $\mathbb{R}^3$

wobei Betragsstriche die Länge eines Vektors bezeichnen sollen und $\measuredangle(\vec{v},\vec{w})$ für den Winkel zwischen $\vec{v}$ und $\vec{w}$ steht. Da

$$\cos(2\pi-\varphi)=\cos(-\varphi)=\cos\varphi$$

ist, folgt sofort das *Kommutativgesetz*

$$\vec{v}\cdot\vec{w}=\vec{w}\cdot\vec{v}$$

für das Skalarprodukt; außerdem folgt wegen $\cos 0 = 1$, daß

$$\vec{v}\cdot\vec{v}=|\vec{v}|^2 \quad \text{oder} \quad |\vec{v}|=\sqrt{\vec{v}\cdot\vec{v}}$$

ist. Das Skalarprodukt erlaubt also auch die Berechnung von Längen.

Weiter ist $\cos(\pi/2)=\cos(3\pi/2)=0$ und damit $\vec{v}\cdot\vec{w}=0$, falls $\vec{v}$ und $\vec{w}$ aufeinander senkrecht stehen; falls keiner der beiden Vektoren der Nullvektor ist, stehen $\vec{v}$ und $\vec{w}$ *genau dann* senkrecht aufeinander, wenn $\vec{v}\cdot\vec{w}=0$ ist.

Das Skalarprodukt kann auch geometrisch interpretiert werden: Da der Cosinus eines Winkels gleich Ankathete durch Hypothenuse ist, zeigt

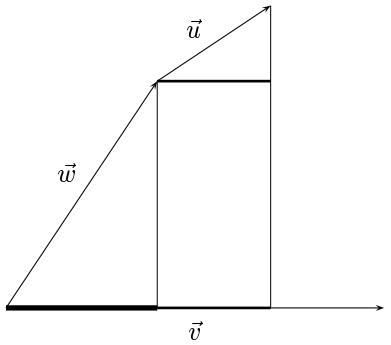


Abb. 14: Geometrische Interpretation des Skalarprodukts

Abbildung 14, daß $|\vec{w}|\cos\measuredangle(\vec{v},\vec{w})$ gerade die Länge des senkrecht auf die von $\vec{v}$ erzeugte Gerade projizierten Vektors $\vec{w}$ ist; in der Zeichnung ist dies die fett eingezeichnete Strecke.

Addieren wir zu $\vec{w}$ einen weiteren Vektor $\vec{u}$, so ist auch hier wieder $|\vec{u}|\cos\measuredangle(\vec{v},\vec{u})$ die (in Abbildung 14 halbfett eingezeichnete) Länge des auf die von $\vec{v}$ erzeugte Gerade projizierten Vektors $\vec{u}$ und entsprechendes gilt für $\vec{w}+\vec{u}$; wir erhalten somit die Regel

$$\vec{v}\cdot(\vec{w}+\vec{u})=\vec{v}\cdot\vec{w}+\vec{v}\cdot\vec{u}\,,$$

und damit wegen des Kommutativgesetzes auch

$$(\vec{v}+\vec{w})\cdot\vec{u}=\vec{v}\cdot\vec{u}+\vec{w}\cdot\vec{u}\,.$$

Zum Berechnen des Skalarprodukts ist die bisherige Definition für viele Fälle recht unhandlich, da Vektoren oft in einer solchen Weise gegeben sind, daß man den Winkel zwischen ihnen *nicht* ohne weiteres kennt. Um gut rechnen zu können, wählen wir eine Basis aus drei paarweise aufeinander senkrecht stehenden Vektoren $\vec{e}_1,\vec{e}_2,\vec{e}_3$ der Länge eins. Für diese Vektoren läßt sich das Skalarprodukt leicht ausrechnen: Da zwei verschiedene stets aufeinander senkrecht stehen und jeder einzelne die

Länge eins hat, ist

$$\vec{e}_i \cdot \vec{e}_j = \delta_{ij} = \begin{cases} 0 & \text{falls } i \neq j \\ 1 & \text{falls } i = j \end{cases} .$$

(δ_{ij} ist das aus §2 bekannte KRONECKER-δ)

Für zwei Vektoren

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad \text{und} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

ist dann

$$\vec{v} \cdot \vec{w} = (v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3) \cdot (w_1\vec{e}_1 + w_2\vec{e}_2 + w_3\vec{e}_3) ,$$

und nach obigen Rechenregeln läßt sich dies berechnen als

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{3}\sum_{j=1}^{3} v_i w_j \, \vec{e}_i\vec{e}_j = \sum_{i=1}^{3}\sum_{j=1}^{3} v_i w_j \delta_{ij} = \sum_{i=1}^{3} v_i w_i .$$

In dieser Form werden Skalarprodukte meistens ausgerechnet, und wir können aus dem Ergebnis rückwärts den Winkel zwischen den beiden Faktoren bestimmen, denn nach der ursprünglichen Definition des Skalarprodukt ist

$$\cos \measuredangle(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| \cdot |\vec{w}|} = \frac{\vec{v} \cdot \vec{w}}{\sqrt{(\vec{v} \cdot \vec{v})(\vec{w} \cdot \vec{w})}} .$$

Da wir nur den Cosinus des Winkels kennen, ist dieser nur bis auf Vielfache von 180° im Gradmaß *bzw.* π im Bogenmaß bestimmt, aber mehr ist im Dreidimensionalen ohnehin nicht sinnvoll: Für Vektoren in der Ebene unterscheiden wir zwei entgegengesetzt gleiche Winkel α und $-\alpha$ dadurch, daß wir den Gegenuhrzeigersinn als mathematisch positiv auszeichnen. Im $\mathbb{R}^3$ aber können wir die Uhr auf zwei Arten in die von zwei Vektoren aufgespannte Ebene legen: mit dem Ziffernblatt nach „oben“ oder nach „unten“, wobei wir diese Begriffe nicht konsistent unterscheiden können. Damit entfällt die Unterscheidung zwischen α und $-\alpha$.

b) Euklidische Vektorräume

Im letzten Abschnitt haben wir gesehen, daß das Skalarprodukt im $\mathbb{R}^3$ eng mit Längen und Winkeln zusammenhängt; da dies zwei der Grundbegriffe der EUKLIDischen Geometrie sind, werden wir reelle Vektorräume mit Skalarprodukt allgemein als EUKLIDische Vektorräume bezeichnen.

Wir beginnen mit einem etwas schwächeren Begriff als dem des Skalarprodukts, der sich im Gegensatz zu letzterem noch für Vektorräume über beliebigen Körpern definieren läßt:

Definition: Eine *symmetrische Bilinearform* auf dem k-Vektorraum V ist eine Abbildung $\cdot : V \times V \to k$ mit folgenden Eigenschaften:
a) $\cdot$ ist *bilinear,* d.h.

$$(\lambda\vec{v}_1 + \mu\vec{v}_2) \cdot \vec{w} = \lambda(\vec{v}_1 \cdot \vec{w}) + \mu(\vec{v}_2 \cdot \vec{w}) \qquad \text{und}$$
$$\vec{v} \cdot (\lambda\vec{w}_1 + \mu\vec{w}_2) = \lambda(\vec{v} \cdot \vec{w}_1) + \mu(\vec{v} \cdot \vec{w}_2)$$

für alle $\vec{v}_1, \vec{v}_2, \vec{w}, \vec{w}_1, \vec{w}_2 \in V$ und $\lambda, \mu \in k$.
b) $\cdot$ ist *symmetrisch,* d.h. $\vec{v} \cdot \vec{w} = \vec{w} \cdot \vec{v}$ für alle $\vec{v}, \vec{w} \in V$.

Die Skalarprodukte in $\mathbb{R}^2$ und $\mathbb{R}^3$ sind natürlich symmetrische Bilinearformen im Sinne dieser Definition; allgemeiner wird für jeden $\mathbb{R}^n$ durch

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = v_1 w_1 + \cdots + v_n w_n$$

eine Bilinearform definiert.

Noch allgemeiner erklärt diese Formel auch für die Vektorräume k^n über einem beliebigen Körper k eine symmetrische Bilinearform, die auch für (nach unserem derzeitigen Kenntnisstand) eher exotische Körper durchaus nützliche Anwendungen haben kann: Ist etwa $\mathbb{F}_2 = \{0, 1\}$ der Körper mit zwei Elementen (in dem Addition und Multiplikation modulo zwei definiert sind), so ist das Produkt eines Vektors $\vec{v} \in \mathbb{F}_2^n$ mit dem Vektor, dessen sämtliche Komponenten gleich eins sind, genau dann gleich null, wenn die Anzahl der Einsen im Vektor $\vec{v}$ gerade ist;

ansonsten ist es null. Auf diese Weise läßt sich also eine Paritätsprüfung einfach und kompakt formulieren. Wenn man außer dem Vektor mit lauter Einsen als Komponenten noch weitere geeignete Vektoren wählt, lassen sich nicht nur Paritätsfehler, sondern auch noch andere Fehler erkennen und eventuell sogar korrigieren.

Verglichen mit dem Skalarprodukt, wie wir es aus $\mathbb{R}^2$ und $\mathbb{R}^3$ gewohnt sind, fehlt diesen Bilinearformen jedoch eine wesentliche Eigenschaft: In $\mathbb{R}^2$ und $\mathbb{R}^3$ ist das Skalarprodukt eines Vektors mit sich selbst gleich dem Quadrat der Länge und verschwindet somit genau dann, wenn der Vektor gleich dem Nullvektor ist. Dies gilt auch für die gerade definierte Bilinearform auf $\mathbb{R}^n$, denn

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = v_1^2 + \cdots + v_n^2$$

verschwindet als Summe von Quadraten genau dann, wenn jedes einzelne v_i verschwindet.

Über dem Körper $\mathbb{F}_2$ dagegen ist beispielsweise

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = 1^2 + 0^2 + 1^2 + 0^2 = 1 + 1 = 0\,,$$

und auch über den komplexen Zahlen ist etwa in $\mathbb{C}^2$

$$\begin{pmatrix} 1 \\ i \end{pmatrix} \cdot \begin{pmatrix} 1 \\ i \end{pmatrix} = 1^2 + i^2 = 1 - 1 = 0\,.$$

Auch über den reellen Zahlen lassen sich Bilinearformen finden, für die das Produkt eines Vektors mit sich selbst verschwinden kann, ohne daß der Vektor gleich dem Nullvektor sein müßte: Für die Bilinearform

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \star \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \underset{\text{def}}{=} v_1 w_1 - v_2 w_2$$

etwa ist

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \star \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1^2 - 1^2 = 0\,.$$

Da wir im $\mathbb{R}^n$ (und später auch in allgemeineren Räumen) Längen und Abstände definieren wollen, um so auch dort analytische Grundbegriffe wie Konvergenz und Stetigkeit einführen zu können, müssen wir solche Bilinearformen ausschließen: Abstände dürfen nicht negativ sein, und der Abstand zwischen zwei Punkten soll nur dann verschwinden, wenn beide Punkte gleich sind.

Da man in beliebigen Körpern nicht von Positivität und Negativität reden kann, beschränken wir uns ab jetzt auf den Fall $k = \mathbb{R}$ und definieren:

Definition: *a)* Eine Bilinearform $V \times V \to \mathbb{R}$ auf einem reellen Vektorraum V heißt *positiv semidefinit,* wenn

$$\vec{v} \cdot \vec{v} \geq 0 \quad \text{für alle } \vec{v} \in V\,;$$

sie heißt *positiv definit* oder *Skalarprodukt,* wenn zusätzlich gilt

$$\vec{v} \cdot \vec{v} = 0 \Rightarrow \vec{v} = \vec{0}\,.$$

b) Ein EUKLIDischer Vektorraum ist ein Paar $(V, \cdot)$ bestehend aus einem $\mathbb{R}$-Vektorraum V und einem Skalarprodukt $\cdot : V \times V \to \mathbb{R}$.

Wie bei Produkten üblich, werden wir den Malpunkt oft weglassen. Gelegentlich, wenn Verwechslungen mit anderen Produkten zu befürchten sind, werden wir anstelle von $\vec{v} \cdot \vec{w}$ auch $(\vec{v},\ \vec{w})$ schreiben; in der Literatur findet man manchmal auch die Schreibweise $\langle \vec{v},\ \vec{w} \rangle$.

Typisches Beispiel eines EUKLIDischen Vektorraums ist natürlich der $\mathbb{R}^n$ mit seinem Standardskalarprodukt

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = v_1 w_1 + \cdots + v_n w_n\,.$$

Ein anderes Beispiel ist der Raum $C^0([0,\ 1],\ \mathbb{R})$ aller stetiger Funktionen vom Einheitsintervall $[0,\ 1]$ nach $\mathbb{R}$ mit dem Skalarprodukt

$$(f,\ g) \underset{\text{def}}{=} \int_0^1 f(t) g(t)\, dt\,.$$

(Dies ist ein typischer Fall, wo die Klammerschreibweise vorzuziehen ist, denn $f \cdot g$ ist bereits vergeben für die Funktion, die jedem $t \in [0,\ 1]$ den Wert $f(t) \cdot g(t)$ zuordnet.)

Die Bilinearität des so definierten Produkts ist klar; außerdem ist es offensichtlich positiv semidefinit:

$$(f,\ f) = \int_0^1 f(t)^2\ dt \geq 0\,,$$

da der Integrand nirgends negativ wird. Falls f nicht identisch verschwindet, gibt es einen Punkt $t_0 \in (0,\ 1)$ mit $f(t_0) = h \neq 0$. Wegen der Stetigkeit von f gibt es dazu eine Umgebung $(a,\ b)$, so daß $|f(t)| > h/2$ für $t \in (a,\ b)$. Damit ist

$$(f,\ f) = \int_0^1 f(t)^2\ dt \geq \int_a^b f(t)^2\ dt > \int_a^b \frac{h^2}{4}\ dt = \frac{h^2}{4}\,(b-a) > 0\,,$$

$(f,\ f) = 0$ ist also nur möglich, wenn f überall verschwindet.

Man beachte, daß hier die Stetigkeit von f eine wesentliche Rolle spielt: Für die Funktion f, die überall verschwindet außer im Punkt $1/2$, wo sie den Wert 1 annimmt, ist das Integral über $f(t)^2$ gleich null, obwohl die Funktion nicht die Nullfunktion ist.

Da wir EUKLIDische Vektorräume eingeführt haben, um dort so etwas wie EUKLIDische Geometrie zu betreiben, sollten wir als nächstes deren Grundbegriffe definieren:

Definition: *a)* Die Länge eines Vektors $\vec{v}$ aus einem EUKLIDischen Vektorraum V ist $|\vec{v}| = \sqrt{\vec{v} \cdot \vec{v}}$.
b) Der Winkel zwischen zwei Vektoren $\vec{v}, \vec{w} \in V \setminus \{\vec{0}\}$ ist

$$\measuredangle(\vec{v}, \vec{w}) = \arccos\left(\frac{\vec{v} \cdot \vec{w}}{\sqrt{(\vec{v} \cdot \vec{v})(\vec{w} \cdot \vec{w})}}\right)\,.$$

Da der Arkuscosinus nur Werte zwischen null und π annimmt, im Gradmaß also zwischen 0° und 180°, ist der so definierte Winkel *unorientiert.*

Wir sollten uns allerdings die Frage stellen, ob hiermit *überhaupt* ein Winkel erklärt ist: Da der Cosinus nur Werte zwischen -1 und 1 annimmt, ist das offensichtlich nur dann der Fall, wenn das Argument des Arkuskosinus in obiger Definition höchstens den Betrag eins hat.

Für das Standardskalarprodukt im $\mathbb{R}^2$ oder $\mathbb{R}^3$ ist das kein Problem: Das Skalarprodukt im $\mathbb{R}^3$ hatten wir im vorigen Abschnitt *definiert* als Produkt aus den Längen der beiden Vektoren und dem Cosinus des eingeschlossenen Winkels; wir hatten dann nachgerechnet, daß dies mit der Koordinatendefinition des hier definierten Skalarprodukts übereinstimmt.

Da zwei Vektoren stets eine Ebene aufspannen, sollten wir erwarten, daß Entsprechendes auch für beliebige reelle Vektorräume gilt, was auch in der Tat der Fall ist. Mit dem Beweis wollen wir allerdings noch warten bis zum übernächsten Abschnitt, wo wir ohne nennenswerten zusätzlichen Aufwand gleich auch noch eine ähnliche Formel für komplexe Vektorräume beweisen können.

Für EUKLIDische Vektorräume wie $\mathcal{C}^0\big([0,\ 1], \mathbb{R}\big)$ ist die „Länge“ eines „Vektors“ natürlich nichts mehr, was man sich geometrisch anschaulich vorstellen könnte; trotzdem ist sie oft eine nützliche Größe. Eine stetige Funktion $I\colon [0,\ 1] \to \mathbb{R}$ könnte beispielsweise einen zeitabhängigen Strom beschreiben, der durch einen Widerstand R fließt; nach dem OHMschen Gesetz fällt dort eine Spannung $U(t) = R \cdot I(t)$ ab, so daß die elektrische Leistung gleich $U(t) \cdot I(t) = R \cdot I(t)^2$ ist. Die elektrische Arbeit, die während des Zeitraums $[0,\ 1]$ verrichtet wird, ist daher gleich

$$\int_0^1 U(t) \cdot I(t)\ dt = \int_0^1 R \cdot I(t)^2\ dt = R \cdot \int_0^1 I(t)^2\ dt = R \cdot |I|\ ,$$

d.h. die „Länge“ von I ist bis auf einen konstanten Faktor gerade gleich der Energie des Signals I.

Ähnliche physikalische Interpretationen gibt es bei fast allen Anwendungen von Vektorräumen, deren Elemente Funktionen sind.

c) Hermitesche Vektorräume

Die unmittelbare Verallgemeinerung des Skalarprodukts des $\mathbb{R}^n$ auf $\mathbb{C}^n$ ist nicht sonderlich nützlich: Dann hätten wir nämlich zum Beispiel

$$\begin{pmatrix}1\\ i\end{pmatrix}\cdot\begin{pmatrix}1\\ i\end{pmatrix} = 1\cdot 1 + i\cdot i = 0\,.$$

Solche Vektoren mit Quadrat Null sind zwar in einigen Anwendungen (wie etwa der speziellen Relativitätstheorie) durchaus sinnvoll und nützlich, meist möchte man aber das Skalarprodukt eines Vektors mit sich selbst als Quadrat seiner Länge interpretieren, und die sollte für alle Vektoren außer dem Nullvektor positiv sein.

Wir haben bereits jeder komplexen Zahl ungleich Null eine positive reelle Zahl zugeordnet, nämlich ihren Betrag

$$|z| = \sqrt{z\overline{z}}\,.$$

Entsprechend können wir zwei komplexen Vektoren

$$\vec{v} = \begin{pmatrix}v_1\\ \vdots\\ v_n\end{pmatrix} \quad \text{und} \quad \vec{w} = \begin{pmatrix}w_1\\ \vdots\\ w_n\end{pmatrix}$$

aus $\mathbb{C}^n$ das Produkt

$$\vec{v}\cdot\vec{w} = \sum_{\ell=1}^{n} v_\ell\overline{w_\ell}$$

zuordnen. Auch hier ist offensichtlich $\vec{v}\cdot\vec{v}$ für jeden Vektor außer dem Nullvektor eine positive reelle Zahl, denn wir addieren ja die Betragsquadrate der Komponenten des Vektors.

Dieses Produkt erfüllt nun allerdings nicht mehr die Forderungen aus der Definition eines Skalarprodukts: Es ist zwar noch linear im ersten Argument, nicht mehr aber im zweiten, denn schon bei Multiplikation des zweiten Vektors mit einer komplexen Zahl $\lambda\notin\mathbb{R}$ ist

$$\vec{v}\cdot(\lambda\vec{w}) = \sum_{\ell=1}^{n} v_\ell\cdot\overline{\lambda w_\ell} = \overline{\lambda}\cdot\sum_{\ell=1}^{n} v_\ell\overline{w_\ell} = \overline{\lambda}\cdot(\vec{v}\cdot\vec{w}) \neq \lambda\cdot(\vec{v}\cdot\vec{w})\,.$$

Bezüglich der Addition gibt es keine Probleme; somit wird die Linearitätsregel für das zweite Argument zu

$$\vec{u}\cdot(\lambda\vec{v}+\mu\vec{w}) = \overline{\lambda}\cdot(\vec{u}\cdot\vec{v}) + \overline{\mu}\cdot(\vec{u}\cdot\vec{w})\,.$$

Die Symmetriebedingung ist ebenfalls verletzt, denn schon für zwei komplexe Zahlen z und w ist das Produkt $z\overline{w} = \overline{w}z = \overline{w\overline{z}}$ im allgemeinen verschieden von $w\overline{z}$, und entsprechend ist auch für zwei Vektoren

$$\vec{v}\cdot\vec{w} = \overline{\vec{w}\cdot\vec{v}}\,.$$

Der Begriff des HERMITEschen Vektorraums formalisiert diese Eigenschaften:

Definition: Ein HERMITEscher Vektorraum ist ein Paar $(V,\cdot)$ bestehend aus einem $\mathbb{C}$-Vektorraum V und einer Abbildung

$$\cdot : V\times V\to\mathbb{C};\quad (\vec{v},\vec{w})\mapsto\vec{v}\cdot\vec{w}$$

mit folgenden Eigenschaften:

a) $\cdot$ ist linear im ersten Argument, d.h.

$$(\lambda\vec{u}+\mu\vec{v})\cdot\vec{w} = \lambda(\vec{u}\cdot\vec{w}) + \mu(\vec{v}\cdot\vec{w})$$

für alle $\vec{u},\vec{v},\vec{w}\in V$ und $\lambda,\mu\in\mathbb{C}$.

b) $\cdot$ ist HERMITEsch symmetrisch, d.h.

$$\vec{v}\cdot\vec{w} = \overline{\vec{w}\cdot\vec{v}}$$

für alle $\vec{v},\vec{w}\in V$.

c) $\cdot$ ist positiv definit, d.h.

$$\vec{v}\cdot\vec{v} > 0$$

ist eine positive *reelle* Zahl für alle Vektoren $\vec{v}\neq\vec{0}$ aus V.

Die bilineare Abbildung $\cdot$ heißt HERMITE*sches Skalarprodukt;* auch hier werden wir den Malpunkt nicht immer hinschreiben und bei Verwechslungsgefahr mit anderen Produkten auch gelegentlich $(\vec{v},\,\vec{w})$ anstelle von $\vec{v}\cdot\vec{w}$ schreiben.

CHARLES HERMITE (1822–1901) war einer der bedeutendsten Mathematiker des neunzehnten Jahrhunderts. Zu seinen Resultaten zählen eine Vereinfachung des ABELschen Beweises, daß Gleichungen fünften Grades im allgemeinen nicht durch Wurzelausdrücke gelöst werden können, die explizite Lösung solcher Gleichungen durch elliptische Funktionen, der Nachweis, daß e eine transzendente Zahl ist, also keiner algebraischen Gleichung über $\mathbb{Q}$ genügt, eine Interpolationsformel und vieles mehr. HERMITE galt als ein sehr guter akademischer Lehrer; er unterrichtete an der École Polytechnique, dem Collège de France, der École Normale Supérieure und der Sorbonne.

Durch Kombination der Eigenschaften *a)* und *b)* kann man leicht ausrechnen, was anstelle der Linearität für das zweite Argument gilt: Genau wie im obigen Beispiel ist

$$\vec{u}\cdot(\lambda\vec{v}+\mu\vec{w}) = \overline{(\lambda\vec{v}+\mu\vec{w})\cdot\vec{u}} = \overline{\lambda(\vec{v}\cdot\vec{u})+\mu(\vec{w}\cdot\vec{u})}$$
$$= \overline{\lambda}\,\overline{(\vec{v}\cdot\vec{u})}+\overline{\mu}\,\overline{(\vec{w}\cdot\vec{u})} = \overline{\lambda}\,(\vec{u}\cdot\vec{v})+\overline{\mu}\,(\vec{u}\cdot\vec{w})\,.$$

Diese Eigenschaft, zusammen mit der Linearität im ersten Argument, bezeichnet man gelegentlich als *Sesquilinearität,* d.h. „anderthalbfache Linearität", im Gegensatz zur echten Linearität in zwei Argumenten, der *Bilinearität.*

Typisches Beispiel eines HERMITEschen Vektorraums ist der $\mathbb{C}^n$ mit dem eingangs definierten Produkt, jedoch wird sich im nächsten Semester zeigen, daß gerade HERMITEsche Vektorräume von komplexwertigen Funktionen sehr interessante Anwendungen haben.

Hier sei nur als Beispiel für das Rechnen mit HERMITEschen Skalarprodukten in $\mathbb{C}^n$ gezeigt, mit dem wir auch nochmals das Rechnen mit Matrizen wiederholen können, und zwar geht es um ein Verfahren von G. PHILIPPE, veröffentlicht in *Quadrature,* Oct.-Déc. 2000, S. 23–34, wie man aus komplexen Vektoren reelle quadratische Matrizen A, B definieren kann, für die $AB = -BA$ ist, die also *antikommutieren.*

Dazu betrachten wir einen Vektor $\vec{v} \in \mathbb{C}^n$, wobei $\mathbb{C}^n$ sein Standard-HERMITEsches Produkt habe. Zu den Komponenten v_i von $\vec{v}$ betrachten wir die $n \times n$-Matrix W mit Einträgen $w_{ij} = v_i\overline{v_j}$ und die Matrix

$$M = a\cdot E - 2W \quad \text{mit} \quad a = \vec{v}\cdot\vec{v}\,.$$

Da aE als Diagonalmatrix mit W kommutiert, ist

$$M^2 = a^2E - 4aW + 4W^2\,.$$

Der Eintrag an der Stelle ik von W^2 ist

$$\sum_{j=1}^{n} w_{ij}w_{jk} = \sum_{j=1}^{n} v_i\overline{v_j}\cdot v_j\overline{v_k} = v_i\overline{v}_k\sum_{j=1}^{n} v_j\overline{v_j} = w_{ik}(\vec{v}\cdot\vec{v}) = aw_{ik}\,,$$

d.h. $M^2 = a^2E$.

Bezeichen A und B die reellen $n \times n$-Matrizen aus den Real- und Imaginärteilen von M, ist also $M = A + iB$, so ist demnach

$$(A+iB)^2 = A^2 - B^2 + i(AB+BA) = a^2E$$

eine reelle Matrix, der Imaginärteil $AB + BA$ muß also verschwinden. Das ist aber gleichbedeutend damit, daß $AB = -BA$ ist.

Als Beispiel betrachten wir den Vektor $\binom{1}{i}$ aus $\mathbb{C}^2$. Hier ist $a = 2$ und

$$M = \begin{pmatrix} 1 & -i \\ i & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + i\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\,,$$

und in der Tat ist

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = -E \quad \text{und} \quad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = E\,.$$

Kompliziertere Vektoren $\vec{v}$ führen zu interessanteren Beispielen.

Auch in HERMITEschen Vektorräumen werden wir $|\vec{v}| = \sqrt{\vec{v}\cdot\vec{v}}$ gelegentlich als *Länge* des Vektors $\vec{v}$ bezeichnen; auf die Definition von Winkeln hingegen wollen wir verzichten, da die Übernahme der entsprechenden Definition für EUKLIDische Vektorräume hier auf geometrisch nicht sinnvolle komplexe Winkel führen würde.

Ansonsten können und werden wir EUKLIDische und HERMITEsche Vektorräume im folgenden gleichzeitig behandeln: Ersetzt man in der Definition eines HERMITEschen Vektorraums überall $\mathbb{C}$ durch $\mathbb{R}$, so erhält man einen EUKLIDischen Vektorraum. Zwar stehen noch an vielen Stellen die Querstriche für komplexe Konjugation, aber da sich eine reelle Zahl unter komplexer Konjugation nicht ändert, sind die Konjugationsstriche nur überflüssig, nicht schädlich.

d) Die Cauchy-Schwarzsche Ungleichung

Als erstes Beispiel wollen wir das noch offene Problem aus Abschnitt *b)* lösen und zeigen, daß die Zahl, die wir dort als Cosinus eines Winkels definiert hatten, tatsächlich zwischen null und eins liegt. Natürlich zeigen wir dies gleich etwas allgemeiner so, daß wir auch eine Aussage für HERMITEsche Vektorräume bekommen, und wir verallgemeinern auch gleich noch etwas weiter in Hinblick auf die Tatsache, daß wir in Vektorräumen, die auch unstetige Funktionen enthalten, im allgemeinen kein wirkliches Skalar- *bzw.* HERMITEsches Produkt haben. Trotzdem werden Produkte in solchen Vektorräumen im nächsten Semester für die FOURIER-Theorie sehr nützlich sein.

Cauchy-Schwarzsche Ungleichung: Für $k = \mathbb{R}$ oder $\mathbb{C}$ sei auf dem k-Vektorraum V eine Abbildung $\cdot\colon V \times V \to k;\ (f,g) \mapsto f \cdot g$ gegeben mit den Eigenschaften
a) $(\lambda\vec{u} + \mu\vec{v}) \cdot \vec{w} = \lambda(\vec{u} \cdot \vec{w}) + \mu(\vec{v} \cdot \vec{w})$
b) $\vec{w} \cdot \vec{v} = \overline{\vec{v} \cdot \vec{w}}$
c) $|\vec{v}|^2 = \vec{v} \cdot \vec{v} \geq 0$ für alle $\vec{v} \in V$.
Dann ist für alle Vektoren $\vec{v}, \vec{w} \in V$

$$|\vec{v} \cdot \vec{w}| \leq |\vec{v}| \cdot |\vec{w}| \ .$$

(Man beachte, daß hier nicht gefordert wird, daß $\vec{v} \cdot \vec{v} > 0$ für $\vec{v} \neq \vec{0}$.)

Der *Beweis* beruht auf folgendem Trick: Wegen *c)* ist für $\lambda, \mu \in \mathbb{C}$

$$\begin{aligned}(\lambda\vec{v} + \mu\vec{w}) \cdot (\lambda\vec{v} + \mu\vec{w}) &= \lambda\overline{\lambda}\,\vec{v} \cdot \vec{v} + \lambda\overline{\mu}\,\vec{v} \cdot \vec{w} + \mu\overline{\lambda}\,\vec{w} \cdot \vec{v} + \mu\overline{\mu}\,\vec{w} \cdot \vec{w} \\ &= \lambda\overline{\lambda}\,\vec{v} \cdot \vec{v} + \lambda\overline{\mu}\,\vec{v} \cdot \vec{w} + \overline{\lambda}\mu\,\overline{\vec{v} \cdot \vec{w}} + \mu\overline{\mu}\,\vec{w} \cdot \vec{w}\end{aligned}$$

stets nichtnegativ. Speziell für $\lambda = (\vec{w} \cdot \vec{w}) \in \mathbb{R}$ und $\mu = -(\vec{v} \cdot \vec{w})$ erhalten wir

$$(\vec{w} \cdot \vec{w})^2(\vec{v} \cdot \vec{v}) - 2(\vec{w} \cdot \vec{w})(\overline{\vec{v} \cdot \vec{w}})(\vec{v} \cdot \vec{w}) + (\vec{v} \cdot \vec{w})(\overline{\vec{v} \cdot \vec{w}})(\vec{w} \cdot \vec{w}) \geq 0\,,$$

also

$$(\vec{w} \cdot \vec{w})\Big((\vec{w} \cdot \vec{w})(\vec{v} \cdot \vec{v}) - |\vec{v} \cdot \vec{w}|^2\Big) \geq 0\,.$$

Ist hier $\vec{w} \cdot \vec{w} \neq 0$, können wir durch diese Zahl dividieren und die Behauptung ist bewiesen. Andernfalls können wir, falls wenigstens $\vec{v} \cdot \vec{v}$ nicht null ist, im obigen Argument die Rollen von $\vec{v}$ und $\vec{w}$ vertauschen und damit die Behauptung beweisen.

Wenn schließlich $\vec{v} \cdot \vec{v}$ und $\vec{w} \cdot \vec{w}$ beide null sind, folgt aus

$$(\vec{v} \pm \vec{w}) \cdot (\vec{v} \pm \vec{w}) = \pm(\vec{v} \cdot \vec{w} + \overline{\vec{v} \cdot \vec{w}}) = \pm 2\,\mathfrak{Re}(\vec{v}\vec{w}) \geq 0\,,$$

daß der Realteil von $\vec{v} \cdot \vec{w}$ verschwindet, und genauso verschwindet auch der Imaginärteil, da $(i\vec{v} \pm \vec{w}) \cdot (i\vec{v} \pm \vec{w}) = \pm 2\,\mathfrak{Im}(\vec{v} \cdot \vec{w}) \geq 0$ ist. Somit ist $\vec{v} \cdot \vec{w} = 0$, und die Ungleichung gilt auch in diesem Fall. ∎

Baron AUGUSTIN LOUIS CAUCHY (1789–1857) stellte als erster durch die exakte Definition von Begriffen wie *Konvergenz* und *Stetigkeit* die Analysis auf ein sicheres Fundament. In insgesamt 789 Arbeiten beschäftigte er sich u.a. auch mit komplexer Analysis, Variationsrechnung, Differentialgleichungen, FOURIER-Analysis, Permutationsgruppen, der Diagonalisierung von Matrizen und der theoretischen Mechanik. Als überzeugter Royalist hatte er häufig Schwierigkeiten mit den damaligen Regierungen; er lebte daher mehrere Jahre im Exil in Turin und später in Prag, wo er (mit sehr mäßigem Erfolg) den französischen Thronfolger unterrichtete.

Der deutsche Mathematiker KARL HERMAN AMANDUS SCHWARZ (1843–1921) beschäftigte sich hauptsächlich mit konformen Abbildungen und mit sogenannten Minimalflächen, d.h. Flächen mit vorgegebenen Eigenschaften, deren Flächeninhalt minimal ist. Im Rahmen einer entsprechenden Arbeit für die WEIERSTRASS-Festschrift von 1885 (im Falle eines durch Doppelintegrale definierten Skalarprodukts) bewies er die obige Ungleichung; CAUCHY hatte sie bereits in seinem Analysislehrbuch von 1821 für endlichdimensionale Vektoren bewiesen. SCHWARZ lehrte nacheinander in Halle, Zürich, Göttingen und Berlin.

e) Orthonormalbasen

Genau wie im Falle des $\mathbb{R}^3$ wollen wir auch für beliebige EUKLIDische oder HERMITEsche Vektorräume zwei Vektoren als *orthogonal* bezeichnen, wenn ihr (HERMITEsches) Skalarprodukt verschwindet. Viele Rechnungen vereinfachen sich, wenn man von einer Basis ausgeht, deren Vek-

toren paarweise orthogonal sind und möglicherweise zusätzlich noch die Länge eins habe; um diese Art von Basen soll es hier gehen:

Definition: *a)* Eine Basis $\mathcal{B}$ eines EUKLIDischen oder HERMITEschen Vektorraums heißt *Orthogonalbasis,* wenn jedes Element von $\mathcal{B}$ orthogonal zu allen übrigen ist.
b) Eine Orthogonalbasis $\mathcal{B}$ heißt *Orthonormalbasis,* wenn zusätzlich für jeden Vektor $\vec{b} \in \mathcal{B}$ gilt: $\vec{b} \cdot \vec{b} = 1$.

Standardbeispiel sind die Einheitsvektoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \vec{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

die sowohl für $\mathbb{R}^n$ als auch für $\mathbb{C}^n$ bezüglich des jeweiligen Standardskalar- oder HERMITEschen Produkts eine Orthonormalbasis bilden.

Tatsächlich hat sogar *jeder* EUKLIDische oder HERMITEsche Vektorraum eine Orthonormalbasis; für den Beweis werden wir uns allerdings, wie immer, wenn von Basen die Rede ist, zur Vermeidung logischer Schwierigkeiten auf den endlichdimensionalen Fall beschränken:

Satz: Jeder endlichdimensionale EUKLIDische oder HERMITEsche Vektorraum V hat eine Orthonormalbasis.

Beweis: Zunächst wissen wir, daß V überhaupt eine Basis $(\vec{b}_1, \ldots, \vec{b}_n)$ hat. Daraus konstruieren wir schrittweise eine *Orthogonalbasis* nach dem sogenannten GRAM-SCHMIDTschen Orthogonalisierungsverfahren. Dieses liefert in seinem r-ten Schritt eine Orthogonalbasis $(\vec{c}_1, \ldots, \vec{c}_r)$ des von dern ersten r Basisvektoren aufgespannten Untervektorraums $[\vec{b}_1, \ldots, \vec{b}_r]$.

Der erste Schritt ist der einfachste: Da es noch keine Orthogonalitätsbedingung für $\vec{c}_1$ gibt, können wir einfach $\vec{c}_1 = \vec{b}_1$ setzen.

Nachdem wir $r \geq 1$ Schritte durchgeführt haben, haben wir r linear unabhängige Vektoren $\vec{c}_1, \ldots, \vec{c}_r$ mit $\vec{c}_i \cdot \vec{c}_j = 0$ für $i \neq j$ aus dem von

$\vec{b}_1, \ldots, \vec{b}_r$ aufgespannten Untervektorraum. Ist $r = n$, haben wir eine Orthogonalbasis; andernfalls muß ein auf den bisher konstruierten $\vec{c}_i$ senkrecht stehender Vektor $\vec{c}_{r+1}$ gefunden werden, der zusammen mit diesen den von $\vec{b}_1$ bis $\vec{b}_{r+1}$ erzeugten Untervektorraum erzeugt.

Da $\vec{c}_1, \ldots, \vec{c}_r$ und $\vec{b}_1, \ldots, \vec{b}_r$ denselben Untervektorraum erzeugen, gilt dasselbe für $\vec{c}_1, \ldots, \vec{c}_r, \vec{b}_{r+1}$ und $\vec{b}_1, \ldots, \vec{b}_{r+1}$; das Problem ist, daß $\vec{b}_{r+1}$ im allgemeinen nicht orthogonal zu den $\vec{c}_i$ sein wird. Wir dürfen $\vec{b}_{r+1}$ aber abändern um einen beliebigen Vektor aus dem von $\vec{c}_1, \ldots, \vec{c}_r$ aufgespannten Untervektorraum; also setzen wir

$$\vec{c}_{r+1} = \vec{b}_{r+1} + \lambda_1 \vec{c}_1 + \cdots + \lambda_r \vec{c}_r$$

und versuchen, die λ_i so zu bestimmen, daß dieser Vektor orthogonal zu $\vec{c}_1, \ldots, \vec{c}_r$ wird.

Wegen der Orthogonalität der $\vec{c}_i$ ist

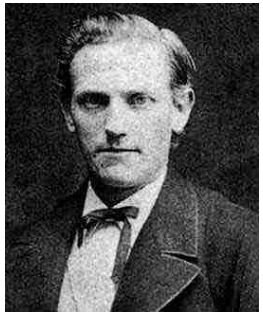
$$\vec{c}_{r+1} \cdot \vec{c}_i = \vec{b}_{r+1} \cdot \vec{c}_i + \sum_{j=1} \lambda_j (\vec{c}_j \cdot \vec{c}_i) = \vec{b}_{r+1} \cdot \vec{c}_i + \lambda_i (\vec{c}_i \cdot \vec{c}_i)\,;$$

setzen wir daher $\lambda_i = -\dfrac{\vec{b}_{r+1} \cdot \vec{c}_i}{\vec{c}_i \cdot \vec{c}_i}$, so ist $\vec{v} \cdot \vec{c}_i = 0$ für alle $i = 1, \ldots, r$.

Nach dem n-ten Schritt haben wir eine Orthogonalbasis $(\vec{c}_1, \ldots, \vec{c}_n)$ von V konstruiert. Daraus wird die gewünschte *Orthonormalbasis* $(\vec{e}_1, \ldots, \vec{e}_n)$, wenn wir jeden Vektor durch seine Länge dividieren, d.h.

$$\vec{e}_i = \frac{\vec{c}_i}{|\vec{c}_i|}\,.$$

∎



Der dänische Mathematiker JØRGAN PEDERSEN GRAM (1850–1916) lehrte an der Universität Kopenhagen, war aber gleichzeitig auch noch geschäftsführender Direktor einer Versicherungsgesellschaft und Präsident des Verbands der dänischen Versicherungsunternehmen. Er publizierte anscheinend nur eine einzige mathematische Arbeit *Sur quelque théorèmes fondamentaux de l'algèbre moderne,* die 1874 erschien. Das GRAM-SCHMIDTsche Orthogonalisierungsverfahren, durch das er heute hauptsächlich bekannt ist, stammt wohl von LAPLACE (1749–1827) und wurde auch schon 1836 von CAUCHY verwendet.

ERHARD SCHMIDT (1876–1959) wurde in Estland geboren; er studierte in Berlin bei SCHWARZ und promovierte in Göttingen bei HILBERT.

ERHARD SCHMIDT ist einer der Begründer der modernen Funktionalanalysis; insbesondere geht die Verallgemeinerung EUKLIDischer und HERMITEscher Vektorräume zu sogenannten HILBERT-Räumen, mit der wir uns im nächsten Semester im Zusammenhang mit der FOURIER-Analysis und der Theorie der Differentialgleichungen beschäftigen werden, auf ihn zurück. Er war Professor in Zürich, Erlangen, Breslau und Berlin.

Um auch den Umgang mit HERMITEschen Skalarprodukten und das Rechnen mit komplexen Zahlen zu üben, betrachten wir als Beispiel den von

$$\vec{b}_1 = \begin{pmatrix} 1 \\ i \\ -i \\ -1 \end{pmatrix}, \quad \vec{b}_2 = \begin{pmatrix} 1+2i \\ -2+i \\ -i \\ -1 \end{pmatrix} \quad \text{und} \quad \vec{b}_3 = \begin{pmatrix} 1+2i \\ -i \\ -4-i \\ 1+2i \end{pmatrix}$$

aufgespannten Untervektorraum von $\mathbb{C}^4$.

Wie oben im Beweis können wir bei der Anwendung des GRAM-SCHMIDTschen Orthogonalisierungsverfahrens $\vec{c}_1 = \vec{b}_1$ setzen; im zweiten Schritt suchen wir einen Vektor $\vec{c}_2 = \vec{b}_2 + \lambda_1 \vec{c}_1$ mit der Eigenschaft, daß

$$\vec{c}_2 \cdot \vec{c}_1 = (\vec{b}_2 + \lambda_1 \vec{c}_1) \cdot \vec{c}_1 = \vec{b}_2 \cdot \vec{c}_1 + \lambda_1 (\vec{c}_1 \cdot \vec{c}_1) = 0$$

ist.

(Wir könnten natürlich auch fordern, daß

$$\vec{c}_1 \cdot \vec{c}_2 = \vec{c}_1 \cdot (\vec{b}_2 + \lambda_1 \vec{c}_1) = \vec{c}_1 \cdot \vec{b}_2 + \overline{\lambda}_1 (\vec{c}_1 \cdot \vec{c}_1)$$

verschwindet, aber dann erhalten wir zunächst eine Formel für $\overline{\lambda}_1$ und müssen noch komplex konjugieren. Daher ist zumindest im komplexen Fall die Forderung $\vec{c}_2 \cdot \vec{c}_1 = 0$ rechnerisch etwas bequemer – auch wenn natürlich beides auf dasselbe Ergebnis führt.)

Wir brauchen also die HERMITEschen Skalarprodukte

$$\vec{b}_2 \cdot \vec{c}_1 = \begin{pmatrix} 1+2i \\ -2+i \\ -i \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ i \\ -i \\ -1 \end{pmatrix} \text{ und } \vec{c}_1 \cdot \vec{c}_1 = \begin{pmatrix} 1 \\ i \\ -i \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ i \\ -i \\ -1 \end{pmatrix}.$$

Bei deren Berechnung darf man auf keinen Fall vergessen, die Einträge des jeweils zweiten Vektors komplex zu konjugieren, d.h.

$$\begin{aligned} \vec{b}_2 \cdot \vec{c}_1 &= (1+2i) \cdot \overline{1} + (-2+i) \cdot \overline{i} + (-i) \cdot \overline{(-i)} + (-1) \cdot \overline{(-1)} \\ &= (1+2i) \cdot 1 + (-2+i) \cdot (-i) + (-i) \cdot i + (-1) \cdot (-1) \\ &= (1+2i) + (1+2i) + 1 + 1 = 4 + 4i \end{aligned}$$

Entsprechend ist

$$\begin{aligned} \vec{c}_1 \cdot \vec{c}_1 &= 1 \cdot \overline{1} + i \cdot \overline{i} + (-i) \cdot \overline{(-i)} + (-1) \cdot \overline{(-1)} \\ &= 1 \cdot 1 + i \cdot (-i) + (-i) \cdot i + (-1) \cdot (-1) = 1 + 1 + 1 + 1 = 4, \end{aligned}$$

wobei das allerdings auch mit weniger Mühe einzusehen ist: Da für einen Vektor $\vec{v} \in \mathbb{C}^n$ mit Komponenten $v_1, \ldots, v_n$

$$\vec{v} \cdot \vec{v} = \sum_{i=1}^{n} v_i \cdot \overline{v_i} = \sum_{i=1}^{n} |v_i|^2$$

ist, hätte es genügt, einfach die Betragsquadrate der vier Einträge aufzusummieren, wobei im vorliegenden Fall trivial ist, daß diese allesamt gleich eins sind.

Einsetzen der beiden Skalarprodukte in die Gleichung für λ_1 zeigt, daß

$$(4+4i) + 4\lambda_1 = 0 \quad \text{und damit} \quad \lambda_1 = \frac{-4-4i}{4} = -1-i$$

sein muß. Somit ist

$$\vec{c}_2 = \vec{b}_2 + \lambda_1 \vec{c}_1 = \begin{pmatrix} 1+2i \\ -2+i \\ -i \\ -1 \end{pmatrix} - \begin{pmatrix} 1+i \\ -1+i \\ 1-i \\ -1-i \end{pmatrix} = \begin{pmatrix} i \\ -1 \\ -1 \\ i \end{pmatrix} .$$

Im dritten Schritt muß ein Vektor

$$\vec{c}_3 = \vec{b}_3 + \mu_1 \vec{c}_1 + \mu_2 \vec{c}_2$$

berechnet werden mit

$$\vec{c}_3 \cdot \vec{c}_1 = \vec{c}_3 \cdot \vec{c}_2 = 0 .$$

Hier ist

$$\vec{c}_3 \cdot \vec{c}_1 = \vec{b}_3 \cdot \vec{c}_1 + \mu_1 \vec{c}_1 \cdot \vec{c}_1 + \mu_2 \vec{c}_2 \cdot \vec{c}_1 = \vec{b}_3 \cdot \vec{c}_1 + \mu_1 \vec{c}_1 \cdot \vec{c}_1 ,$$

da der zweite Schritt sicherstellte, daß $\vec{c}_2 \cdot \vec{c}_1 = 0$ ist. Das HERMITEsche Skalarprodukt $\vec{c}_1 \cdot \vec{c}_1 = 4$ kennen wir bereits, und eine kurze Rechnung zeigt, daß

$$\vec{b}_3 \cdot \vec{c}_1 = (1+2i) + (-1) + (1-4i) + (-1-2i) = -4i$$

ist. Also folgt

$$\mu_1 = -\frac{\vec{b}_3 \cdot \vec{c}_1}{\vec{c}_1 \cdot \vec{c}_1} = -\frac{-4i}{4} = i$$

und entsprechend folgt auch

$$\mu_2 = -\frac{\vec{b}_3 \cdot \vec{c}_2}{\vec{c}_2 \cdot \vec{c}_2} = -\frac{8}{4} = -2 ,$$

wobei diesesmal die genaue Berechnung der beiden HERMITEschen Skalarprodukte dem Leser als Übungsaufgabe überlassen sei. Der dritte Basisvektor $\vec{c}_3 = \vec{b}_3 + i\vec{c}_1 - 2\vec{c}_2$ ist daher gleich

$$\begin{pmatrix} 1+2i \\ -i \\ -4-i \\ 1+2i \end{pmatrix} + i \begin{pmatrix} 1 \\ i \\ -i \\ -1 \end{pmatrix} - 2 \begin{pmatrix} i \\ -1 \\ -1 \\ i \end{pmatrix} = \begin{pmatrix} 1+i \\ 1-i \\ -1-i \\ 1-i \end{pmatrix} .$$

Damit ist eine *Orthogonalbasis* gefunden; wenn wir eine *Orthonormalbasis* wollen, müssen wir die Vektoren noch durch ihre Längen dividieren: Wir wissen bereits, daß $\vec{c}_1 \cdot \vec{c}_1 = \vec{c}_2 \cdot \vec{c}_2 = 4$ ist, so daß die ersten

beiden Vektoren die Länge zwei haben; $\vec{c}_3$ hat vier Komponenten vom Betrag zwei, also ist

$$\vec{c}_3 \cdot \vec{c}_3 = 4 \cdot 2 = 8 \quad \text{und} \quad |\vec{c}_2| = \sqrt{8} = 2\sqrt{2} .$$

Die gesuchte Orthonormalbasis besteht daher aus den Vektoren

$$\frac{1}{2} \begin{pmatrix} 1 \\ i \\ -i \\ 1 \end{pmatrix} , \quad \frac{1}{2} \begin{pmatrix} i \\ -1 \\ -1 \\ i \end{pmatrix} \quad \text{und} \quad \frac{\sqrt{2}}{4} \begin{pmatrix} 1+i \\ 1-i \\ -1-i \\ 1-i \end{pmatrix} .$$

Die Nützlichkeit von Orthonormal- und Orthogonalbasen ergibt sich vor allem aus den beiden folgenden Sätzen:

Satz: $(\vec{e}_1, \ldots, \vec{e}_n)$ sei eine Orthonormalbasis eines EUKLIDischen oder HERMITEschen Vektorraums V.
a) Sind $\vec{v} = a_1\vec{e}_1 + \cdots + a_n\vec{e}_n$ und $\vec{w} = b_1\vec{e}_1 + \cdots + b_n\vec{e}_n$ zwei Vektoren aus V, so ist

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} a_i \overline{b_i} .$$

b) Für einen Vektor $\vec{v} \in V$ ist

$$\vec{v} = a_1\vec{e}_1 + \cdots + a_n\vec{e}_n \quad \text{mit} \quad a_i = \vec{v} \cdot \vec{e}_i .$$

Der *Beweis* ist in beiden Fällen ein Einzeiler:

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{b_j}\, \vec{e}_i \cdot \vec{e}_j = \sum_{i=1}^{n} a_i \overline{b_i} , \quad \text{denn} \quad \vec{e}_i \cdot \vec{e}_j = \delta_{ij}$$

$$\vec{v} \cdot \vec{e}_i = (a_1\vec{e}_1 + \cdots + a_n\vec{e}_n) \cdot \vec{e}_i = \sum_{j=1}^{n} a_j \vec{e}_j \cdot \vec{e}_i = a_i .$$

■

Bezüglich einer Orthonormalbasis sieht also jedes (gewöhnliche oder HERMITEsche) Skalarprodukt aus wie das entsprechende Standardskalarprodukt – und das unabhängig von der gewählten Orthonormalbasis. Sobald man Vektoren bezüglich einer Orthonormalbasis dargestellt hat,

lassen sich also alle Skalarprodukte sowie auch die Basisdarstellung eines beliebigen Vektors auf die einfachst denkbare Weise darstellen.

Da die Vektoren einer Orthonormalbasis oftmals Wurzelausdrücke als Komponenten enthalten, sind vor allem beim Rechnen ohne Hilfsmittel Orthogonalbasen oft übersichtlicher als Orthonormalbasen; wir wollen den Satz also der Vollständigkeit halber auch für Othogonalbasen formulieren:

Satz: $(\vec{e}_1, \ldots, \vec{e}_n)$ sei eine Orthogonalbasis eines EUKLIDischen oder HERMITEschen Vektorraums V.
a) Sind $\vec{v} = a_1\vec{e}_1 + \cdots + a_n\vec{e}_n$ und $\vec{w} = b_1\vec{e}_1 + \cdots + b_n\vec{e}_n$ zwei Vektoren aus V, so ist

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} a_i \overline{b_i} \, |\vec{e}_i|^2 \ .$$

b) Für einen Vektor $\vec{v} \in V$ ist

$$\vec{v} = a_1\vec{e}_1 + \cdots + a_n\vec{e}_n \quad \text{mit} \quad a_i = \frac{(\vec{v} \cdot \vec{e}_i)}{|\vec{e}_i|^2} \ .$$

Die *Beweise* sind fast dieselben Einzeiler:

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\overline{b_j} \; \vec{e}_i \cdot \vec{e}_j = \sum_{i=1}^{n} a_i\overline{b_i} \, |\vec{e}_i|^2 \ , \quad \text{denn} \quad \vec{e}_i \cdot \vec{e}_j = \delta_{ij} \, |\vec{e}_i|^2$$

$$\vec{v} \cdot \vec{e}_i = (a_1\vec{e}_1 + \cdots + a_n\vec{e}_n) \cdot \vec{e}_i = \sum_{j=1}^{n} a_j \vec{e}_j \cdot \vec{e}_i = a_i \, |\vec{e}_i|^2 \ .$$

■

f) Die QR-Zerlegung einer Matrix

Wir betrachten den Vektorraum $V = \mathbb{R}^n$ oder $V = \mathbb{C}^n$ mit seiner Standardbasis und mit seiner üblichen Struktur als EUKLIDischer *bzw.* HERMITEscher Vektorraum, außerdem noch irgendeine weitere Basis

$(\vec{b}_1, \ldots, \vec{b}_n)$. Nach dem GRAM-SCHMIDTschen Orthogonalisierungsverfahren können wir daraus eine Orthogonalbasis $(\vec{c}_1, \ldots, \vec{c}_n)$ von V konstruieren, wobei gilt

$$\vec{c}_k = \vec{b}_k + \lambda_{k,k-1}\vec{b}_{k-1} + \cdots + \lambda_{k1}\vec{c}_1 \quad \text{mit} \quad \lambda_{kj} = -\frac{\vec{b}_k \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j} \ .$$

Lösen wir auf nach $\vec{b}_k$, erhalten wir

$$\vec{b}_k = \vec{c}_k - \lambda_{k,k-1}\vec{b}_{k-1} - \cdots - \lambda_{k1}\vec{c}_1 = \sum_{j=1}^{n} r_{jk}\vec{c}_j$$

$$\text{mit } r_{jk} = \begin{cases} -\lambda_{kj} = \dfrac{\vec{b}_k \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j} & \text{für } j < k \\ 1 & \text{für } j = k \\ 0 & \text{für } j > k \end{cases} \ .$$

Fassen wir die Vektoren $\vec{b}_1, \ldots \vec{b}_n$ auf als Spaltenvektoren einer Matrix B und die $\vec{c}_j$ als Spalten von C, so gilt also für die i-ten Komponenten

$$b_{ik} = \sum_{j=1}^{n} r_{jk}c_{ij} = \sum_{j=1}^{n} c_{ij}r_{jk} \quad \text{oder} \quad B = CR \quad \text{mit} \quad R = (r_{ij}) \, .$$

Da r_{ij} für $i > j$ verschwindet und alle $r_{ii} = 1$ sind, ist R eine obere Dreiecksmatrix mit Einsen in der Hauptdiagonalen.

Ist B eine beliebige invertierbare $n \times n$-Matrix, so bilden die Spaltenvektoren $\vec{b}_i$ von B eine Basis des $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$; wie wir gerade gesehen haben, gibt es also eine obere Dreiecksmatrix R und eine Matrix C, deren Spalten eine Orthogonalbasis bilden, so daß $B = CR$ ist.

Normieren wir die Spalten von C auf Länge eins, erhalten wir eine $n \times n$-Matrix Q, deren Spalten eine Orthonormalbasis bilden; um die Beziehung $B = QR$ zu erhalten, müssen wir dann allerdings auch noch für jedes $i = 1, \ldots, n$ die i-te Zeile von R mit der Länge des i-ten Spaltenvektors $\vec{c}_i$ von C multiplizieren. Dabei bleibt R zwar eine obere Dreiecksmatrix, in der Hauptdiagonalen stehen nun aber im allgemeinen keine Einsen mehr.

Was passiert, wenn B nicht invertierbar ist? Selbst wenn wir für B eine beliebige $n \times m$-Matrix nehmen, die Anzahl m der Spaltenvektoren $\vec{b}_i$

also nicht mehr gleich der Anzahl der Zeilen ist, lassen sich die Rechenschritte des GRAM-SCHMIDTschen Orthogonalisierungsverfahrens weiterhin durchführen – mit einem wesentlichen Unterschied:

Betrachten wir eine Folge $\vec{b}_1, \ldots, \vec{b}_m$ von Vektoren aus $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$, versehen mit dem EUKLIDischen oder HERMITEschen Standardskalarprodukt. Wir konstruieren wieder im k-ten Schritt zunächst eine Orthogonalbasis des von $\vec{b}_1$ bis $\vec{b}_k$ aufgespannten Untervektorraums.

Schon im ersten Schritt kann es eine Änderung geben: Der Vektor $\vec{b}_1$ könnte der Nullvektor sein und damit definitiv nicht als erster Basisvektor in Frage kommen.

Wir können also im ersten Schritt nicht einfach den Vektor $\vec{b}_1$ als ersten Vektor der zu konstruierenden Orthogonalbasis nehmen, sondern wir müssen den ersten Vektor $\vec{b}_\ell$ nehmen, der vom Nullvektor verschieden ist.

Die eventuell davor stehenden Nullvektoren können wir allerdings nicht ganz vergessen, denn sie sind schließlich Spalten der Matrix und müssen am Ende auch im Produkt CR wieder auftauchen. Wir müssen deshalb in der Matrix R die ersten $\ell - 1$ Zeilen auf Null setzen.

Danach wird $\vec{c}_1 = \vec{b}_\ell$ gesetzt; gegebenenfalls kann man den Vektor auch gleich noch auf Länge eins normieren: Zumindest wenn man von Hand rechnet, wird das davon abhängen, wie kompliziert die Länge ist. Die ℓ-te Zeile von R beginnt mit einer Eins (oder der Länge von $\vec{b}_\ell$, falls $\vec{c}_1$ auf Länge Eins normiert wurde) und enthält ansonsten lauter Nullen.

Im $(k+1)$-ten Schritt, für $k \geq 1$, gehen wir davon aus, daß wir eine Orthogonalbasis $(\vec{c}_1, \ldots, \vec{c}_k)$ des von $\vec{b}_1$ bis $\vec{b}_\ell$ aufgespannten Untervektorraums gefunden haben für ein $\ell \leq m$. Falls $\ell < m$ ist, machen wir den Ansatz

$$\vec{c}_{k+1} = \vec{b}_{\ell+1} - r_{\ell+1,1}\vec{c}_1 - \cdots - r_{\ell+1,k}\vec{c}_k$$

und bestimmen die Koeffizienten $\vec{r}_{ij}$ so, daß $\vec{c}_{k+1}\vec{c}_j = 0$ ist für alle $j \leq k$, d.h.

$$r_{\ell+1,j} = \frac{\vec{b}_{\ell+1} \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j} .$$

Falls $\vec{b}_{\ell+1}$ linear unabhängig ist von $\vec{b}_1, \ldots, \vec{b}_\ell$, ist dann

$$\vec{c}_{k+1} = \vec{b}_{\ell+1} - r_{\ell+1,1}\vec{c}_1 - \cdots - r_{\ell+1,k}\vec{c}_k$$

linear unabhängig von $\vec{c}_1, \ldots, \vec{c}_k$, denn diese Vektoren erzeugen denselben Untervektorraum wie $\vec{b}_1, \ldots, \vec{b}_\ell$.

Wenn allerdings $\vec{b}_{\ell+1}$ linear abhängig ist von $\vec{b}_1, \ldots, \vec{b}_\ell$, ist $\vec{b}_{\ell+1}$ auch linear abhängig von $\vec{c}_1, \ldots, \vec{c}_k$; da diese Vektoren eine Orthogonalbasis des von ihnen erzeugten Vektorraums sind, ist daher

$$\vec{b}_{\ell+1} = \lambda_1\vec{c}_1 + \cdots + \lambda_k\vec{c}_k \quad \text{mit} \quad \lambda_j = \frac{\vec{b}_{\ell+1} \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j} .$$

Vergleicht man dies mit der obigen Formel für $r_{\ell+1,j}$, so sieht man, daß in diesem Fall $\vec{c}_{k+1}$ der Nullvektor ist, wir bekommen also kein neues Element der Orthogonalbasis.

In diesem Fall müssen wir daher mit dem nächsten Vektor $\vec{b}_{\ell+2}$ – so es einen gibt – einen neuen Ansatz

$$\vec{c}_{k+1} = \vec{b}_{\ell+2} - r_{\ell+1,1}\vec{c}_1 - \cdots - r_{\ell+2,k}\vec{c}_k$$

machen, und wieder die Koeffizienten $\vec{r}_{ij}$ so bestimmen, daß $\vec{c}_{k+1}\vec{c}_j = 0$ ist für alle $j \leq k$, d.h.

$$r_{\ell+2,j} = \frac{\vec{b}_{\ell+2} \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j} ,$$

wobei der entstehende Vektor $\vec{c}_{k+1}$ wieder der Nullvektor sein kann und so weiter, bis wir entweder einen vom Nullvektor verschiedenen Vektor $\vec{c}_{k+1}$ erhalten oder aber kein neuer Vektor $\vec{b}_j$ mehr existiert.

Im ersteren Fall machen wir weiter mit dem $(k+2)$-ten Schritt, andernfalls haben wir eine Basis $(\vec{c}_1, \ldots, \vec{c}_k)$ gefunden für den von $\vec{b}_1, \ldots, \vec{b}_m$ aufgespannten Untervektorraum von $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$.

Das muß allerdings noch keine Basis des gesamten $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ sein, denn offensichtlich ist die Anzahl der Vektoren $\vec{c}_k$, die wir so erhalten, gleich dem Rang von A, und der könnte auch kleiner als n sein – für $m < n$ muß das sogar so sein.

Wenn wir eine quadratische Matrix C suchen, müssen wir daher gegebenenfalls noch die Vektoren $\vec{c}_1, \ldots, \vec{c}_k$ zu einer vollen Orthogonalbasis ergänzen; dazu können wir beispielsweise so lange GRAM-SCHMIDT auf Vektoren aus der Standardbasis anwenden, bis wir n orthogonale Vektoren $\vec{c}_i$ gefunden haben.

Falls alle Vektoren $\vec{c}_i$ gleich auf Länge eins normiert wurden, bilden sie sogar eine Orthonormalbasis. Da eine solche Normierung allerdings oft Nenner mit Wurzeln produziert, ist es oft rechnerisch angenehmer, die $\vec{c}_i$ unnormiert zu lassen. In diesem Fall müssen sie zum Schluß auf die Länge eins gebracht werden; wir setzen dazu

$$\vec{q}_i = \frac{1}{|\vec{c}_i|}\vec{c}_i$$

und definieren Q als die $n \times n$-Matrix mit Spaltenvektoren $\vec{q}_1$ bis $\vec{q}_n$.

Da die Einträge der Matrix R bislang so berechnet waren, daß $CR = A$ ist, müssen wir auch diese noch modifizieren: Um von C auf Q zu kommen, haben wir die i-te Spalte durch $|\vec{c}_i|$ dividiert; zur Kompensation müssen wir daher in R die i-te *Zeile* mit $|\vec{c}_i|$ *multiplizieren;* dann haben wir mit der so modifizierten Matrix R die Produktzerlegung $A = QR$, die sogenannte QR-Zerlegung von A.

Obwohl die Matrix R hier nicht mehr quadratisch ist, wollen wir sie als (obere) Dreiecksmatrix bezeichnen, denn auch hier verschwindet r_{ij} wann immer $i > j$ ist.

Damit haben wir gezeigt:

Satz: Zu jeder reellen oder komplexen $n \times m$-Matrix A gibt es eine $n \times n$-Matrix Q, deren Spalten eine Orthonormalbasis von $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ bilden, sowie eine obere Dreiecksmatrix $R \in \mathbb{R}^{n\times m}$ *bzw.* $\mathbb{C}^{n\times m}$, so daß gilt: $A = QR$.
Hat A einen Rang $k < n$, so enthalten die letzten $n - k$ Zeilen von R nur Nullen; daher ist auch $A = Q'R'$, wobei Q' die $n \times k$-Matrix aus den ersten k Spalten von Q ist und die $k \times m$-Matrix R' aus den ersten k Zeilen von R besteht. ∎

Als Beispiel betrachten wir die Matrix

$$A = \begin{pmatrix} 0 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 3 & 3 & 0 \\ 0 & 0 & 3 & 4 & 0 \end{pmatrix} ;$$

ihre Spaltenvektoren bezeichnen wir mit $\vec{a}_1, \ldots, \vec{a}_5$.

$\vec{a}_1$ hat bereits die Länge eins, kann also als erster Vektor $\vec{q}_1$ der Orthonormalbasis genommen werden, so daß der erste Spaltenvektors $\vec{r}_1$ der Matrix R einfach der erste Koordinateneinheitsvektor ist.

Der zweite Spaltenvektor $\vec{a}_2$ von A hat mit $\vec{q}_1$ das Skalarprodukt eins; da $\vec{q}_1$ selbst bereits Länge eins hat, setzen wir also

$$\vec{q}_2 = \vec{a}_2 - \vec{q}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{r}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} ,$$

denn $\vec{a}_2 = \vec{q}_1 + \vec{q}_2$.

Im dritten Schritt suchen wir zunächst einen Vektor der Form

$$\vec{c}_3 = \vec{a}_3 + \lambda\vec{q}_1 + \mu\vec{q}_3 ,$$

der auf $\vec{q}_1$ und $\vec{q}_2$ senkrecht steht. Da $\vec{q}_1$ und $\vec{q}_2$ die Länge eins haben, ist hier einfach $\lambda = -\vec{a}_3 \cdot \vec{q}_1 = -3$ und $\mu = -\vec{a}_3 \cdot \vec{q}_2 = -2$, der gesuchte Vektor ist also der dreifache vierte Einheitsvektor, und als zugehörigen Vektor $\vec{q}_3$ der Länge eins nehmen wir den vierten Einheitsvektor selbst. Dann ist

$$\vec{a}_3 = 3\vec{q}_1 + 2\vec{q}_2 + 3\vec{q}_3, \quad \text{also} \quad \vec{r}_3 = \begin{pmatrix} 3 \\ 2 \\ 3 \\ 0 \end{pmatrix} .$$

Beim nächsten Spaltenvektor $\vec{a}_4$ sieht man eigentlich bereits ohne Rechnung, daß er im Erzeugnis von $\vec{q}_1$ bis $\vec{q}_3$ liegt, denn offensichtlich ist $\vec{a}_4 = \vec{a}_3 + \vec{q}_2 + \vec{q}_3$. Wenn wir trotzdem stur nach Schema F losrechnen mit dem Ansatz

$$\vec{c}_4 = \vec{a}_4 + \lambda_1\vec{q}_1 + \lambda_2\vec{q}_2 + \lambda_3\vec{q}_3 \quad \text{mit} \quad \lambda_i = -\vec{a}_4 \cdot \vec{q}_i ,$$

erhalten wir erwartungsgemäß $\vec{c}_4 = \vec{0}$, also keinen neuen Vektor der Orthogonalbasis. Wir bekommen aber, da wir auch $\vec{a}_4$ als Linearkombination der $\vec{q}_i$ darstellen müssen, eine neue Spalte $\vec{r}_4$ der Matrix R; wegen

$$\vec{a}_4 = 3\vec{q}_1 + 3\vec{q}_2 + 4\vec{q}_3 \quad \text{ist} \quad \vec{r}_4 = \begin{pmatrix} 3 \\ 3 \\ 4 \\ 0 \end{pmatrix} .$$

In der letzten Spalte von A stehen lauter Nullen; also brauchen wir auch hier keinen neuen Basisvektor und sehen auch ohne jede Rechnung, daß $\vec{r}_5$ gleich dem Nullvektor ist.

Damit kennen wir die Matrix R vollständig, allerdings fehlt für eine vollständige QR-Zerlegung noch eine Spalte von Q. Diese kann offensichtlich völlig unabhängig von A gewählt werden als irgendein Vektor, der auf $\vec{q}_1$ bis $\vec{q}_3$ senkrecht steht. Die kanonische Wahl ist natürlich der dritte Einheitsvektor; die einzige Alternative dazu wäre sein negatives.

Damit ist $A = QR$ mit

$$Q = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{und} \quad R = \begin{pmatrix} 1 & 1 & 3 & 3 & 4 & 0 \\ 0 & 1 & 2 & 3 & 4 & 0 \\ 0 & 0 & 3 & 4 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} .$$

Natürlich ist auch

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 3 & 3 & 4 & 0 \\ 0 & 1 & 2 & 3 & 4 & 0 \\ 0 & 0 & 3 & 4 & 5 & 0 \end{pmatrix} ,$$

denn die vierte Spalte von Q wird ja bei der Multiplikation mit R stets mit Null gewichtet.

g) Orthogonale und unitäre Matrizen

Bei der Lektüre des letzten Abschnitts hat sich sicherlich mancher die Frage gestellt, warum wir selbst bei kleinem Rang von A an einer Zerlegung interessiert sind, bei der Q eine quadratische Matrix ist, deren Spalten eine Orthonormalbasis des gesamten $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ bilden. Der

Grund liegt darin, daß solche Matrizen Q eine ganze Reihe interessanter Eigenschaften haben, die zur Lösung mehrerer wichtiger Probleme verwendet werden können. Dies wollen wir uns im folgenden etwas genauer ansehen.

Betrachten wir zunächst den reellen Fall.

Sind $\vec{q}_i$ die Spaltenvektoren von Q, so ist die Tatsache, daß die $\vec{q}_i$ eine Orthonormalbasis bilden, äquivalent dazu, daß $\vec{q}_i \cdot \vec{q}_j = \delta_{ij}$ für alle i, j. Da wir $\vec{q}_j$ nicht nur als j-te Spalte von Q, sondern auch als j-te Zeilte der transponierten Matrix ${}^tQ$ auffassen können, können wir diese n^2 Gleichungen zusammenfassen zu der einen Matrixgleichung ${}^tQQ = E$.

Auch im HERMITEschen Fall ist $\vec{q}_i \cdot \vec{q}_j = \delta_{ij}$ für alle i, j, aber nun ist das HERMITEsche Skalarprodukt nicht mehr durch dieselbe Formel definiert, die auch bei der Matrixmultiplikation Anwendung findet, sondern enthält noch eine komplexe Konjugation im zweiten Faktor. Daher sind die Gleichungen $\vec{q}_i \cdot \vec{q}_j = \delta_{ij}$ hier äquivalent dazu, daß ${}^tQ\overline{Q}$ die Einheitsmatrix ist. Transponieren macht dies zu ${}^t({}^tQ\overline{Q}) = {}^t\overline{Q}Q = {}^tE = E$.

Die Schreibweise ${}^t\overline{Q}$ ist etwas umständlich; deshalb führen wir eine Abkürzung ein:

Definition: Die *adjungierte Matrix* zu einer Matrix $A \in \mathbb{C}^{n\times m}$ ist die Matrix $A^* = {}^t\overline{A} = \overline{{}^tA}$.

Damit ist also beispielsweise

$$A^* = \begin{pmatrix} 1-i & 2-i & 3-i \\ 4-i & 5-i & 6-i \end{pmatrix} \quad \text{für} \quad A = \begin{pmatrix} 1+i & 4+i \\ 2+i & 5+i \\ 3+i & 6+i \end{pmatrix} .$$

Für eine reelle Matrix A ist natürlich $A^* = {}^tA$ einfach die transponierte Matrix.

Definition: *a)* Eine Matrix $Q \in \mathbb{R}^{n\times n}$ heißt *orthogonal,* wenn ${}^tQQ = E$ ist.
b) Eine Matrix $U \in \mathbb{C}^{n\times n}$ heißt *unitär,* wenn $U^*U = E$ ist.

Die Matrix Q aus der QR-Zerlegung einer reellen Matrix ist somit orthogonal; im Falle einer komplexen Matrix ist Q unitär.

Lemma: *a)* Eine Matrix $Q \in \mathbb{R}^{n\times n}$ ist genau dann orthogonal, wenn $Q^{-1} = {}^tQ$ ist. Insbesondere ist jede orthogonale Matrix invertierbar.
b) Eine Matrix $U \in \mathbb{C}^{n\times n}$ ist genau dann unitär, wenn $U^{-1} = U^*$ ist; insbesondere ist jede unitäre Matrix invertierbar.
c) Die inverse Matrix einer orthogonalen *bzw.* unitären Matrix ist wieder orthogonal *bzw.* unitär.
d) Das Produkt zweier orthogonaler *bzw.* unitärer Matrizen ist wieder orthogonal *bzw.* unitär.

Beweis: Da Unitarität und Orthogonalität für reelle Matrizen äquivalent sind, können wir uns auf unitäre Matrizen beschränken.

a) und *b)* sind klar, denn nach Definition einer unitären Matrix U ist U^* invers zu U. Für *c)* muß somit gezeigt werden, daß U^* wieder unitär ist. Nach Definition ist U invers zu dieser Matrix, wegen $(U^*)^* = U$ folgt die Unitarität.

Für *d)* schließlich seien U_1 und U_2 zwei unitäre Matrizen; dann ist

$$(U_1U_2)^* = U_2^*U_1^* = U_2^{-1}U_1^{-1} = (U_1U_2)^{-1}\,,$$

und damit ist auch U_1U_2 unitär. ∎

Insbesondere die Aussagen *a)* und *b)* zeigen einen großen praktischen Vorteil orthogonaler und unitärer Matrizen: Man kann sie mit minimalem Aufwand invertieren.

Ist etwa $A\vec{x} = \vec{b}$ ein lineares Gleichungssystem mit einer $n \times m$-Koeffizientenmatrix A, und ist $A = QR$ die QR-Zerlegung von A, so ist

$$A\vec{x} = \vec{b} \Longleftrightarrow QR\vec{x} = \vec{b} \Longleftrightarrow R\vec{x} = Q^{-1}\vec{b} = Q^*\vec{b}\,.$$

Da R eine obere Dreiecksmatrix ist, hat das Produkt $R\vec{x}$ ausgeschrieben die Treppengestalt, die der GAUSSS-Algorithmus produziert, und die rechte Seite läßt sich für jede neue rechte Seite zum Preis einer einzigen Matrixmultiplikation berechnen. Im Gegensatz zur LR-Zerlegung ist also keine Matrixinversion nötig, und dazu ist diese Methode auch noch numerisch stabiler, falls man mit Gleitkommazahlen rechnet und einen guten numerischen Algorithmus zur Berechnung der QR-Zerlegung verwendet. (GRAM-SCHMIDT ist numerisch eher nicht zu empfehlen.)

Der wesentliche Grund für die guten numerischen Eigenschaften orthogonaler und unitärer Matrizen besteht darin, daß sie Längen respektieren. Um das einzusehen, beweisen wir zunächst ein allgemeines Lemma (das auch der historische Grund für die Bezeichnung „adjungierte Matrix" ist):

Lemma: Für $k = \mathbb{R}$ oder $\mathbb{C}$ und $A \in k^{n\times m}$ ist

$$(A\vec{v})\cdot\vec{w} = \vec{v}\cdot(A^*\vec{w}) \quad \text{für alle } \vec{v} \in k^m \text{ und } \vec{w} \in k^n\,,$$

wobei links das (HERMITEsche) Standardskalarprodukt von k^n steht und rechts das von k^m.

Beweis: Es genügt, den HERMITEschen Fall zu betrachten. Dazu fassen wir einen Vektor $\vec{w} \in \mathbb{C}^n$ auf als eine $n \times 1$-Matrix $w_M \in \mathbb{C}^{n\times 1}$; für einen weiteren Vektor $\vec{u} \in \mathbb{C}^n$, aufgefaßt als Matrix $u_M \in \mathbb{C}^{n\times 1}$, ist dann das HERMITEsche Skalarprodukt $\vec{v}\cdot\vec{w}$ gleich dem Matrixprodukt ${}^tv_M \cdot \overline{w_M}$.

Ganz entsprechend ordnen wir dem Vektor $\vec{v} \in \mathbb{C}^m$ eine $m \times 1$-Matrix $v_M \in \mathbb{C}^{m\times 1}$ zu; die Matrix zum Vektor $A\vec{v}$ ist dann die Produktmatrix Av_M.

Somit ist

$$\begin{aligned}(A\vec{v})\cdot\vec{w} &= {}^t(Av_M)\cdot\overline{w_M} = {}^tv_M\cdot{}^tA\cdot\overline{w_M} = {}^tv_M\cdot\overline{\overline{{}^tA}\cdot w_M}\\ &= {}^tv_M\cdot\overline{A^*\cdot w_M} = \vec{v}\cdot(A^*\vec{w})\,.\end{aligned}$$

∎

Als mehr oder weniger unmittelbare Folgerung erhalten wir:

Satz: *a)* Eine Matrix $A \in \mathbb{R}^{n\times n}$ ist genau dann orthogonal, wenn für das Standardskalarprodukt des $\mathbb{R}^n$ gilt: $(A\vec{v})\cdot(A\vec{w}) = \vec{v}\cdot\vec{w}$ für alle $\vec{v},\vec{w} \in \mathbb{R}^n$.
b) Eine Matrix $A \in \mathbb{C}^{n\times n}$ ist genau dann unitär, wenn für das HERMITEsche Standardprodukt des $\mathbb{C}^n$ gilt: $(A\vec{v})\cdot(A\vec{w}) = \vec{v}\cdot\vec{w}$ für alle $\vec{v},\vec{w} \in \mathbb{C}^n$.

Beweis: Nach dem gerade bewiesenen Lemma ist

$$(A\vec{v})\cdot(A\vec{w}) = \vec{v}\cdot\left(A^*(A\vec{w})\right) = \vec{v}\cdot\left((A^*A)\vec{w}\right)\,.$$

A ist genau dann orthogonal *bzw.* unitär, wenn AA^* gleich der Einheitsmatrix ist; in diesem Fall ist $(A^*A)\vec{w} = \vec{w}$ und damit $(A\vec{v})\cdot(A\vec{w}) = \vec{v}\cdot\vec{w}$.

Falls wir umgekehrt wissen, daß $(A\vec{v})\cdot(A\vec{w}) = \vec{v}\cdot\vec{w}$ ist für alle Vektoren $\vec{v}, \vec{w}$, so ist auch $\vec{v}\cdot\big((A^*A)\vec{w}\big) = \vec{v}\cdot\vec{w}$, insbesondere also

$$\vec{e}_i\cdot\big((A^*A)\vec{e}_j\big) = \vec{e}_i\cdot\vec{e}_j = \delta_{ij}$$

für die Koordinateneinheitsvektoren.

$(A^*A)\vec{e}_j$ ist die j-te Spalte der Matrix A^*A, ihr Skalarprodukt mit $\vec{e}_i$ also der ij-Eintrag von A^*A. Da dieser gleich δ_{ij} sein muß, ist also $A^*A = E$ und A somit orthogonal *bzw.* unitär.

■

h) Orthogonale Projektionen

Ist U ein r-dimensionaler Untervektorraum eines n-dimensionalen Vektorraums V, so können wie jede Basis $\{\vec{b}_1,\ldots,\vec{b}_r\}$ von U ergänzen zu einer Basis $\{\vec{b}_1,\ldots,\vec{b}_n\}$ von V. Der von $\vec{b}_{r+1}$ bis $\vec{b}_n$ erzeugte Untervektorraum $W \le V$ hat dann die Eigenschaft, daß $U\cap W$ der Nullraum ist, während $U\cup W$ dem gesamten Vektorraum V erzeugt. Einen solchen Untervektorraum W bezeichnen wir als *Komplement* von U; es ist natürlich, genau wie seine Basisvektoren $\vec{b}_{r+1}$ bis $\vec{b}_n$, alles andere als eindeutig bestimmt.

Für EUKLIDische und HERMITEsche Vektorräume können wir allerdings jedem Untervektorraum ein wohlbestimmtes ausgezeichnetes Komplement zuordnen, das *orthogonale Komplement.*

Definition: V sei ein EUKLIDischer oder HERMITEscher Vektorraum und $U \le V$ sei ein Untervektorraum von V. Das orthogonale Komplement von U ist der Untervektorraum

$$U^\perp \underset{\text{def}}{=} \big\{\vec{v}\in V \;\big|\; \vec{u}\cdot\vec{v} = 0 \text{ für alle } \vec{u}\in U\big\}\,.$$

Wegen der Linearität des EUKLIDischen wie auch HERMITEschen Skalarprodukts im ersten Argument ist klar, daß $U^\perp$ ein Untervektorraum von V ist. Außerdem ist klar, daß es reicht die Bedingung $\vec{v}\cdot\vec{u} = 0$ für die Vektoren $\vec{u}$ aus einer Basis von U nachzurechnen, denn wenn alle

diese Produkte verschwinden, verschwindet auch jedes Produkt mit einer Linearkombination solcher Vektoren. (Es stört dabei nicht, daß wir im HERMITEschen Fall keine Linearität im zweiten Argument haben, sondern die Koeffizienten komplex konjugieren müssen.)

Lemma: U sei ein Untervektorraum des n-dimensionalen EUKLIDischen oder HERMITEschen Vektorraums V, und $(\vec{b}_1,\ldots,\vec{b}_r)$ sei eine Orthogonalbasis von U. Ergänzt man diese zu einer Orthogonalbasis $(\vec{b}_1,\ldots,\vec{b}_n)$ von V, so ist $(\vec{b}_{r+1},\ldots,\vec{b}_n)$ eine Orthogonalbasis von $U^\perp$. Insbesondere hat also das orthogonale Komplement eines r-dimensionalen Untervektorraum die Dimension $n-r$ und $U\cap U^\perp = \{\vec{0}\}$.

Beweis: Nach dem gerade Gesagten liegt ein Vektor $\vec{v} = \lambda_1\vec{b}_1+\cdots+\lambda_n\vec{b}_n$ aus V genau dann in $U^\perp$, wenn für alle $i \le r$ gilt:

$$\vec{v}\cdot\vec{b}_i = \left(\sum_{j=1}^{n}\lambda_j\vec{b}_j\right)\cdot\vec{b}_i = \sum_{j=1}^{n}\lambda_j\vec{b}_j\cdot\vec{b}_i = \lambda_i\vec{b}_i\cdot\vec{b}_i = 0\,.$$

Da $\vec{b}_i$ als Basisvektor nicht der Nullvektor sein kann, ist $\vec{b}_i\cdot\vec{b}_i \neq 0$; daher ist dies äquivalent zum Verschwinden aller λ_i mit $i \le r$, also zur Darstellbarkeit von $\vec{v}$ als Linearkombination der Vektoren $\vec{b}_{r+1},\ldots,\vec{b}_n$. Als Teil einer Basis sind diese linear unabhängig, also Basis ihres Erzeugnisses $U^\perp$.

■

Korollar: *a)* V sei ein endlichdimensionaler EUKLIDischer oder HERMITEscher Vektorraum, und $U \le V$ sei ein Untervektorraum. Dann läßt sich jedes Element $\vec{v}\in V$ eindeutig schreiben als $\vec{v} = \vec{u}+\vec{w}$ mit $\vec{u}\in U$ und $\vec{w}\in U^\perp$.
b) $U^{\perp\perp} = U$

Beweis: a) Wir wählen eine Orthogonalbasis $(\vec{b}_1,\ldots,\vec{b}_r)$ von U und ergänzen sie zu einer Orthogonalbasis $(\vec{b}_1,\ldots,\vec{b}_n)$ von V; nach dem Lemma ist dann $(\vec{b}_{r+1},\ldots,\vec{b}_n)$ eine Orthogonalbasis von $U^\perp$. Schreiben wir $\vec{v} = v_1\vec{b}_1+\cdots+v_n\vec{b}_n$, so ist also

$$\vec{u}\underset{\text{def}}{=} v_1\vec{b}_1+\cdots+v_r\vec{b}_r\in U,\quad \vec{w}\underset{\text{def}}{=} v_{r+1}\vec{b}_{r+1}+\cdots+v_n\vec{b}_n\in U^\perp$$

und $\vec{v} = \vec{u} + \vec{w}$.

Ist $\vec{v} = \vec{x} + \vec{y}$ irgendeine Darstellung von $\vec{v}$ als Summe zweier Vektoren $\vec{x} \in U$ und $\vec{y} \in V$, so ist

$$\vec{u} + \vec{w} = \vec{x} + \vec{y} \Longrightarrow \vec{u} - \vec{x} = \vec{y} - \vec{w}\,.$$

In der letzteren Gleichung steht links der Vektor $\vec{u} - \vec{x} \in U$ und rechts $\vec{y} - \vec{w} \in U^\perp$; wegen $U \cap U^\perp = \{\vec{0}\}$ ist also $\vec{u} = \vec{x}$ und $\vec{w} = \vec{y}$, was die Eindeutigkeit dieser Zerlegung zeigt.·

b) Für $\vec{u} \in U$ und $\vec{w} \in U^\perp$ verschwindet nach Definition von $U^\perp$ das Produkt $\vec{w} \cdot \vec{u}$, also wegen dessen (HERMITEscher) Symmetrie auch $\vec{u} \cdot \vec{w}$. Damit ist

$$\vec{u} \in U^{\perp\,\perp} = \left\{ \vec{v} \in V \;\middle|\; \vec{v} \cdot \vec{w} = 0 \;\text{ für alle } \vec{w} \in U^\perp \right\},$$

also liegt U in $U^{\perp\,\perp}$. Nach dem obigen Lemma ist

$$\dim U^{\perp\,\perp} = \dim V - \dim U^\perp = \dim V - (\dim V - \dim U) = \dim U\,,$$

also muß $U = U^{\perp\,\perp}$ sein ∎

Bemerkung: Tatsächlich gilt dieses Korollar auch für unendlichdimensionale Vektorräume; da die Existenz von wie auch der Umgang mit Basen im Unendlichdimensionalen etwas problematisch ist, soll aber hier, wie bereits mehrfach in diesem Skriptum, der endlichdimensionale Fall genügen.

Definition: V sei ein endlichdimensionaler EUKLIDischer oder HERMITEscher Vektorraum, und $U \leq V$ sei ein Untervektorraum. Die Abbildung $\pi_U : V \to U$, die jedem Vektor $\vec{v} = \vec{u} + \vec{w} \in V$ mit $\vec{u} \in U$ und $\vec{w} \in U^\perp$ den Vektor $\vec{u}$ zuordnet, heißt *orthogonale Projektion* von V nach U.

Wegen der eindeutigen Zerlegbarkeit eines Vektors in eine Komponente aus U und eine aus $U^\perp$ ist π_U offensichtlich wohldefiniert und linear; der Kern von π_U ist $U^\perp$.

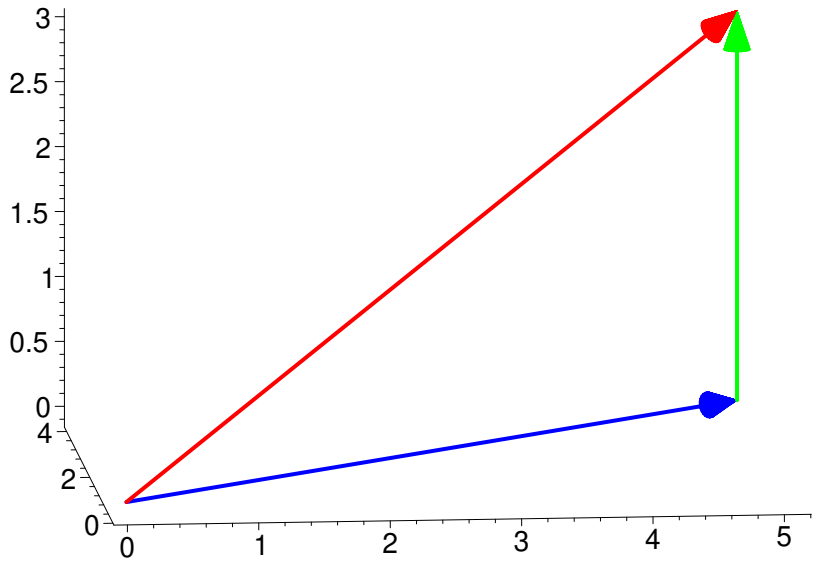


Abb. 15: Orthogonale Projektion eines Vektors

Orthogonale Projektionen sind aus der Geometrie bekannt, beispielsweise als Grundriß, Aufriß und Kreuzriß eines dreidimensionalen Körpers; uns interessiert hier vor allem ihre folgende Eigenschaft:

Lemma: V sei ein endlichdimensionaler EUKLIDischer oder HERMITEscher Vektorraum, $U \leq V$ ein Untervektorraum und $\vec{v} \in V$. Dann gilt für jeden Vektor $\vec{u} \in U$ die Ungleichung $|\vec{v} - \vec{u}| \leq |\vec{v} - \pi_U(\vec{v})|$, d.h. $\pi_U(\vec{v})$ ist derjenige Vektor aus U, dessen Differenz mit $\vec{v}$ am kürzesten ist.

Beweis: Wir schreiben $\vec{v} = \vec{p} + \vec{w}$ mit $\vec{p} = \pi_U(\vec{v}) \in U$ und $\vec{w} \in U^\perp$. Für jeden Vektor $\vec{u} \in U$ ist dann

$$\begin{aligned} |\vec{v} - \vec{u}|^2 &= |\vec{p} + \vec{w} - \vec{u}|^2 = |(\vec{p} - \vec{u}) + \vec{w}|^2 \\ &= (\vec{p} - \vec{u}) \cdot (\vec{p} - \vec{u}) + \vec{w} \cdot (\vec{p} - \vec{u}) + (\vec{p} - \vec{u}) \cdot \vec{w} + \vec{w} \cdot \vec{w} \\ &= |\vec{p} - \vec{u}|^2 + |\vec{w}|^2\,, \end{aligned}$$

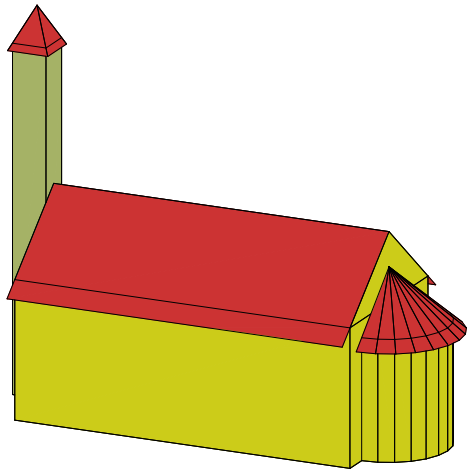


Abb. 16: Ein dreidimensionales Objekt

Abb. 17: Der Grundriß des Objekts aus Abbildung 16

denn $\vec{p} - \vec{u}$ liegt in U und $\vec{w}$ in $U^\perp$. Also ist $|\vec{v} - \vec{u}|$ nie kleiner als $|\vec{v} - \vec{p}|$, und die beiden Vektoren sind genau dann gleich lang, wenn $\vec{u} - \vec{p}$ der Nullvektor ist, also $\vec{u} = \pi_U(\vec{v})$.

■

Betrachten wir orthogonale Projektionen statt in Vektorräumen in den dazugehörigen affinen Räumen, entsprich die orthogonale Projektion auf einen Unterraum geometrisch also einfach der Konstruktion des Lotfußpunkts in diesem Unterraum.

i) Die Methode der kleinsten Quadrate

Oftmals ist zu gegebenen Beobachtungsdaten grundsätzlich bekannt, welcher *Art* von Gesetz sie genügen sollten; das Problem besteht „nur“ noch darin, die in diesem Gesetz vorkommenden *Parameter* zu bestimmen. Im einfachsten Fall könnte man etwa an einen Widerstand denken, der dadurch gemessen wird, daß man verschiedene Spannungen U_i anlegt und die zugehörigen Stromstärken I_i mißt. Nach dem Ohmschen Gesetz ist dann $U_i = R \cdot I_i$, aber aufgrund der unvermeidlichen Meßfehler werden die verschiedenen Quotienten U_i/I_i natürlich nicht alle gleich sein. Die Lösung dieses Problems ist klar: Man nimmt den Mittelwert der Quotienten. Schwieriger wird es, wenn mehrere Parameter ins Spiel kommen, wenn die Meßreihe als mehr als nur einen Parameter bestimmen soll.

Solche Fälle treten nicht nur auf in Naturwissenschaft und Technik, sondern auch in den Wirtschafts- und Sozialwissenschaften, wo es zwar selten *exakte* Gesetze gibt, man den Zusammenhang zwischen verschiedenen Größen aber trotzdem zumindest näherungsweise durch eine mathematische Formel beschreiben will – auch wenn diese in konkreten Einzelfällen gelegentlich ziemlich falsch sein kann.

Als Beispiel dieser Art können wir den Zusammenhang zwischen Korruption und Wohlstand in verschiedenen Staaten betrachten: edes Jahr veröffentlicht die Organisation *Transparancy International* ihren *corruption perceptions index (CPI),* in dem jedem Land eine Zahl zwischen null und zehn zugeordnet wird, je nachdem, wie stark Geschäftsleute, Risikospezialisten und die Bevölkerung die Korruption im betreffenden Land einschätzen: Ein Index von zehn bedeutet, daß es praktische keine Korruption gibt, während bei null nichts läuft ohne Bimbes. Die neuesten Daten stammen vom 7. Oktober 2003 und sind unter

`http://www.transparency.org/cpi/`

zu finden. Die Zahlen werden als Mittelwerte über die letzten drei Jahren berechnet, so daß singuläre Ereignisse eines Jahres nicht zu sehr ins Gewicht fallen.

Wir vergleichen diese Zahlen mit dem Bruttoinlandsprodukt pro Einwohner, wie es die Weltbank für 2001 festgestellt hat. Dies sind die

neuesten verfügbaren Zahlen; sie sind beispielsweise auf dem Server des Statistischen Bundesamtes unter

`http://www.destatis.de/ausl_prog/suche_ausland.htm`

zu finden, indem man unter „Indikatoren“ das Feld „BIP je Einwohner (real)“ auswählt. In der folgenden Tabelle sind Staaten aufgelistet, für die sowohl das Bruttosozialprodukt pro Einwohner als auch der CPI für 2003 vorliegt; das Bruttosozialprodukt pro Einwohner in US-$ ist kursiv gedruckt, der Korruptionsindex fett. Der Index ist jeweils ein Mittelwert aus verschiedenen Quellen; die angegebenen Fehlerschranken sind die Standardabweichungen der betreffenden Daten und geben an, wie groß die Unterschiede zwischen den einzelnen Quellen sind.

Ägypten	*1229*	**3,3** $\pm$ 1,3
Albanien	*1032*	**2,5** $\pm$ 0,6
Algerien	*1616*	**2,6** $\pm$ 0,5
Angola	*525*	**1,8** $\pm$ 0,3
Argentinien	*7468*	**2,5** $\pm$ 0,5
Armenien	*1068*	**3,0** $\pm$ 0,8
Aserbaidschan	*460*	**1,8** $\pm$ 0,3
Äthiopien	*121*	**2,5** $\pm$ 0,8
Australien	*24204*	**8,8** $\pm$ 0,9
Bahrain	*10609*	**6,1** $\pm$ 1,1
Bangladesch	*386*	**1,3** $\pm$ 0,7
Belgien	*31186*	**7,6** $\pm$ 0,9
Belize	*3189*	**4,5** $\pm$ 0,9
Bolivien	*944*	**2,3** $\pm$ 0,4
Bosnien und Herzegowina	*1584*	**3,3** $\pm$ 0,7
Botsuana	*4130*	**5,7** $\pm$ 0,9
Brasilien	*4633*	**3,9** $\pm$ 0,5
Bulgarien	*1670*	**3,9** $\pm$ 0,9
Chile	*5385*	**7,4** $\pm$ 0,9
China	*878*	**3,4** $\pm$ 1,0
Costa Rica	*3900*	**4,3** $\pm$ 0,7
Côte d’Ivoire	*715*	**2,1** $\pm$ 0,5
Dänemark	*39000*	**9,5** $\pm$ 0,4

Deutschland	*32855*	**7,7** $\pm$ 1,2
Dominikanische Republik	*2077*	**3,3** $\pm$ 0,4
Ecuador	*1478*	**2,2** $\pm$ 0,3
El Salvador	*1757*	**3,7** $\pm$ 1,5
Estland	*3457*	**5,5** $\pm$ 0,6
Finnland	*31853*	**9,7** $\pm$ 0,3
Frankreich	*30374*	**6,9** $\pm$ 1,1
Gambia	*382*	**2,5** $\pm$ 0,9
Georgien	*499*	**1,8** $\pm$ 0,7
Ghana	*421*	**3,3** $\pm$ 0,9
Griechenland	*13709*	**4,3** $\pm$ 0,8
Guatemala	*1554*	**2,4** $\pm$ 0,6
Haiti	*354*	**1,5** $\pm$ 0,6
Honduras	*711*	**2,3** $\pm$ 0,6
Indien	*477*	**2,8** $\pm$ 0,4
Indonesien	*1034*	**1,9** $\pm$ 0,5
Iran	*1714*	**3,0** $\pm$ 1,0
Irland	*29352*	**7,5** $\pm$ 0,7
Island	*32161*	**9,6** $\pm$ 0,3
Israel	*16576*	**7,0** $\pm$ 1,2
Italien	*21363*	**5,3** $\pm$ 1,1
Jamaika	*2171*	**3,8** $\pm$ 0,4
Japan	*44353*	**7,0** $\pm$ 1,1
Jemen	*316*	**2,6** $\pm$ 0,7
Jordanien	*1639*	**4,6** $\pm$ 1,1
Kamerun	*696*	**1,8** $\pm$ 0,2
Kanada	*23138*	**8,7** $\pm$ 0,9
Kasachstan	*1712*	**2,4** $\pm$ 0,9
Kenia	*325*	**1,9** $\pm$ 0,3
Kirgisistan	*417*	**2,1** $\pm$ 0,4
Kolumbien	*2277*	**3,7** $\pm$ 0,5
Kongo	*792*	**2,2** $\pm$ 0,5
Korea, Republik	*13581*	**4,3** $\pm$ 1,0
Kroatien	*5355*	**3,7** $\pm$ 0,6
Kuba	*1480*	**4,6** $\pm$ 1,0
Kuwait	*13345*	**5,3** $\pm$ 1,7

Lettland	*2560*	**3,8** ± 0,4
Libanon	*2890*	**3,0** ± 0,8
Litauen	*2195*	**4,7** ± 1,6
Luxemburg	*57720*	**8,7** ± 0,4
Madagaskar	*253*	**2,6** ± 1,8
Malawi	*163*	**2,8** ± 1,2
Malaysia	*4708*	**5,2** ± 1,1
Mali	*292*	**3,0** ± 1,8
Marokko	*1436*	**3,3** ± 1,3
Mauritius	*4352*	**4,4** ± 0,7
Mazedonien	*2417*	**2,3** ± 0,3
Mexiko	*3705*	**3,6** ± 0,6
Moldau, Republik	*678*	**2,4** ± 0,8
Mosambik	*213*	**2,7** ± 0,7
Myanmar	*2875*	**1,6** ± 0,3
Namibia	*2383*	**4,7** ± 1,3
Neuseeland	*18684*	**9,5** ± 0,2
Niederlande	*31577*	**8,9** ± 0,3
Nigeria	*257*	**1,4** ± 0,4
Norwegen	*39749*	**8,8** ± 0,5
Oman	*7110*	**6,3** ± 0,9
Österreich	*33492*	**8,0** ± 0,7
Pakistan	*517*	**2,5** ± 0,9
Panama	*3243*	**3,4** ± 0,8
Papua-Neuguinea	*897*	**2,1** ± 0,6
Paraguay	*1703*	**1,6** ± 0,3
Peru	*2311*	**3,7** ± 0,6
Philippinen	*1165*	**2,5** ± 0,5
Polen	*4758*	**3,6** ± 1,1
Portugal	*13104*	**6,6** ± 1,2
Rumänien	*1573*	**2,8** ± 1,0
Russische Föderation	*2609*	**2,7** ± 0,8
Sambia	*405*	**2,5** ± 0,6
Saudi-Arabien	*6614*	**4,5** ± 2,0
Schweden	*33242*	**9,3** ± 0,2
Schweiz	*47273*	**8,8** ± 0,8

Senegal	*629*	**3,2** ± 1,2
Serbien und Montenegro	*1240*	**2,3** ± 0,5
Sierra Leone	*158*	**2,2** ± 0,5
Simbabwe	*559*	**2,3** ± 0,3
Singapur	*27118*	**9,4** ± 0,1
Slowakei	*4406*	**3,7** ± 0,7
Slowenien	*12010*	**5,9** ± 1,2
Spanien	*18342*	**6,9** ± 0,8
Sri Lanka	*876*	**3,4** ± 0,7
Südafrika	*4068*	**4,4** ± 0,6
Sudan	*328*	**2,3** ± 0,3
Syrien, Arabische Republik	*796*	**3,4** ± 1,3
Tadschikistan	*420*	**1,8** ± 0,3
Taiwan (China)	*15283*	**5,7** ± 1,0
Tansania, Vereinigte Republik	*197*	**2,5** ± 0,6
Thailand	*2853*	**3,3** ± 0,9
Trinidad und Tobago	*5553*	**4,6** ± 1,3
Tschechische Republik	*5555*	**3,9** ± 0,9
Tunesien	*2562*	**4,9** ± 0,7
Türkei	*2810*	**3,1** ± 0,9
Uganda	*355*	**2,2** ± 0,7
Ukraine	*986*	**2,3** ± 0,6
Ungarn	*5694*	**4,8** ± 0,6
Uruguay	*5870*	**5,5** ± 1,1
Usbekistan	*512*	**2,4** ± 0,5
Venezuela	*3326*	**2,4** ± 0,5
Vereinigte Arabische Emirate	*18110*	**5,2** ± 0,5
Vereinigte Staaten	*31399*	**7,5** ± 1,2
Vereinigtes Königreich	*22112*	**8,7** ± 0,5
Vietnam	*390*	**2,4** ± 0,8
Weißrussland	*1493*	**4,2** ± 1,8
Zypern	*14592*	**6,1** ± 1,6

Abbildung 18 zeigt die 127 Datenpunkte zu dieser Liste graphisch, wobei der Punkt für Deutschland etwas heller eingezeichnet ist.

Der erste Augenschein zeigt, daß korruptionsärmere Länder oftmals

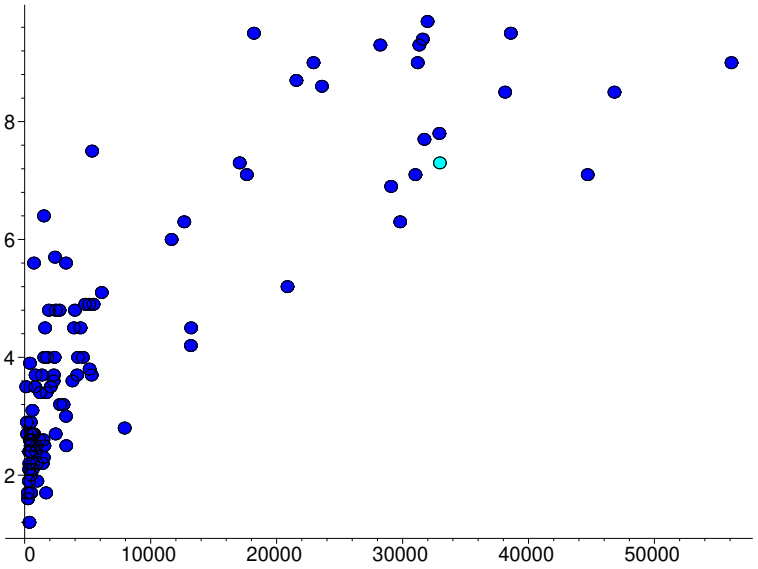


Abb. 18: Zusammenhang zwischen Korruption und Bruttoinlandsprodukt je Einwohner

reicher sind: Das weitgehend korruptionsfreie Dänemark hat ein Bruttosozialprodukt von 39 000 $ pro Einwohner, das deutlich korruptere Deutschland nur 32 855 $ und ein stark korruptes Land wie Bangladesch nur 386 $. Allerdings gibt es auch Ausnahmen, denn Chile hat, obwohl kaum korrupter als Deutschland, nur ein Bruttosozialprodukt von 5 385 $ pro Einwohner. Es gibt also sicherlich keinen deterministischen Zusammenhang zwischen Korruption und Wohlstand, aber doch eine Tendenz.

Falls wir nun versuchen, beispielsweise einen linearen Zusammenhang der Form

$$CPI = a + b \cdot BIP$$

zu finden, so haben wir 127 Gleichungen für die beiden unbekannten Koeffizienten a und b, und ein kurzer Blick auf Abbildung 18 zeigt, daß dieses lineare Gleichungssystem keine Lösung haben kann.

Wir suchen also keine Lösung, sondern zwei Zahlen a und b derart, daß

die 127 Gleichungen „möglichst gut“ gelten. Was das bedeuten soll läßt sich mathematisch auf verschiedene, nicht äquivalente Weisen definieren; da wir uns im Augenblick mit Skalarprodukten beschäftigen, bietet sich an, die 127 Bruttoinlandsprodukte pro Einwohner und die 127 Korruptionsindizes zu zwei Vektoren $\vec{x}, \vec{y} \in \mathbb{R}^{127}$ zusammenzufassen, und nach Zahlen a, b zu suchen, so daß die Länge des Differenzvektors $\vec{y} - a\vec{x} - b$ möglichst klein wird. Ausgeschrieben bedeutet dies, wenn wir die Komponenten von $\vec{x}$ mit x_i und die von $\vec{y}$ mit y_i bezeichnen, daß die Summe

$$\sum_{i=1}^{127} (y_i - ax_i - b)^2$$

der Abweichungsquadrate möglichst klein sein soll – von daher der Name „Methode der kleinsten Quadrate“ für diesen Ansatz, mit dessen Hilfe sein Schöpfer GAUSS sowohl die Position des Planetoiden Ceres vorhersagte als auch die Vermessung und Kartierung des Königreichs Hannover durchführte.

Derselbe Ansatz läßt sich natürlich auf jedes lineare Gleichungssystem über den reellen oder komplexen Zahlen anwenden: Wir haben ein möglicherweise unlösbares lineares Gleichungssystem $A\vec{x} = \vec{b}$ und wollen einen Vektor $\vec{x}$ so bestimmen, daß der Vektor $A\vec{x} - \vec{b}$ minimale Länge hat.

Falls das lineare Gleichungssystem lösbar ist, gibt es damit kein Problem: Wir bestimmen irgendeine Lösung $\vec{x}$ und haben damit einen Vektor gefunden, für den $A\vec{x} - \vec{b}$ die Länge null hat – kürzer geht es nicht.

Im allgemeinen ist aber für den gesuchten Vektor $\vec{x}$ das Produkt $A\vec{x}$ von $\vec{b}$ verschieden; es sei etwa gleich $\vec{c}$. Dann ist $\vec{c}$ ein Vektor, der sich in der Form $A\vec{x}$ darstellen läßt, und unter allen solchen Vektoren ist es derjenige, für den die Länge des Differenzvektors zu $\vec{b}$ minimal ist. Dies erinnert an die orthogonalen Projektionen aus dem vorigen Abschnitt, und in der Tat läßt sich das Problem damit lösen:

Nehmen wir an, wir haben n Gleichungen in m Unbekannten mit Koeffizienten aus $k = \mathbb{R}$ oder $k = \mathbb{C}$. Dann definiert die Matrix $A \in k^{n \times m}$

des Gleichungssystems eine lineare Abbildung

$$\varphi\colon k^m \to k^n; \quad \vec{v} \mapsto A\vec{v};$$

deren Bildraum sei U. Falls die rechte Seite $\vec{b}$ in U liegt, ist das Gleichungssystem lösbar; andernfalls suchen wir einen Vektor $\vec{x} \in k^m$, für den die Länge des Vektors $A\vec{x} - \vec{b}$ minimal wird. Da die Vektoren, die sich in der Form $A\vec{x}$ darstellen lassen, genau die Vektoren aus U sind, ist somit $A\vec{x} = \pi_U(\vec{b})$ die orthogonale Projektion von $\vec{b}$ nach U. Diese könnten wir *im Prinzip* bestimmen, indem wir die QR-Zerlegung von A berechnen, denn dann sind die ersten Spalten von Q eine Basis von U, die durch die weiteren Spalten zu einer Basis von ganz k^n ergänzt wird; danach haben wir ein lösbares lineares Gleichungssystem.

Wir wollen uns überlegen, wie wir $\vec{x}$ auch ohne die rechnerisch aufwendige QR-Zerlegung bestimmen können.

Für den gesuchten Vektor $\vec{x}$ (oder für die gesuchten Vektoren $\vec{x}$) ist $A\vec{x} = \varphi_U(\vec{b})$. Da $A\vec{x}$ bereits in U liegt, ist $\pi_U(A\vec{x}) = A\vec{x}$, also ist die Gleichung $A\vec{x} = \pi_U(\vec{b})$ äquivalent zu

$$\pi_U(A\vec{x}) = \pi_U(\vec{b}) \quad \text{oder} \quad A\vec{x} - \vec{b} \in \operatorname{Kern} \pi_U = U^\perp .$$

Das orthogonale Komplement $U^\perp$ von U besteht aus allen Vektoren $\vec{y} \in k^n$, die senkrecht stehen auf U, für die also gilt

$$(A\vec{x}) \cdot \vec{y} = 0 \quad \text{für alle } \vec{x} \in k^m .$$

Wie wir im vorletzten Abschnitt gesehen haben, ist

$$(A\vec{x}) \cdot \vec{y} = \vec{x} \cdot A^*\vec{y} \quad \text{für alle } \vec{x} \in k^m, \vec{y} \in k^n ,$$

$\vec{y}$ liegt also genau dann in $U^\perp$, wenn $A^*\vec{y}$ senkrecht steht auf allen Vektoren $\vec{x} \in k^m$. Ein solcher Vektor aus k^m ist insbesondere $A^*\vec{y}$ selbst; wegen der positiven Definitheit des (HERMITEschen) Skalarprodukts ist also $A^*\vec{y} = \vec{0}$ und somit

$$U^\perp = \left\{ \vec{y} \in k^n \;\middle|\; A^*\vec{y} = \vec{0} \right\} .$$

$A\vec{x} - \vec{b}$ liegt also genau dann im Kern von π_U, wenn $A^*(A\vec{x} - \vec{b}) = \vec{0}$ ist oder, anders ausgedrückt, wenn $\vec{x}$ eine Lösung des linearen Gleichungssystems

$$(A^*A)\,\vec{x} = A^*\vec{b}$$

ist. Da die adjungierte Matrix A^* einfach die transponierte Matrix zur komplex konjugierten Matrix zu A ist, wobei die komplexe Konjugation über $\mathbb{R}$ natürlich entfällt, läßt sich dieses Gleichungssystem schnell aufstellen und dann nach GAUSS lösen.

Betrachten wir dies konkret im eingangs diskutierten Fall eines linearen Zusammenhangs $y = ax + b$ zu N Wertepaaren $(x_i, y_i) \in \mathbb{R}^2$, wobei N sinnvollerweise größer als zwei sein sollte. Wir haben dann N Gleichungen

$$y_i = ax_i + b \quad \text{oder} \quad x_i a + b = y_i ,$$

wobei hier im Gegensatz zu unserer sonstigen Gewohnheit die Parameter a und b unbekannt sind, während die x_i und die y_i bekannt sind. Wir haben also ein lineares Geichungssystem von N Gleichungen in den beiden Variablen a und b.

Fassen wir die Werte x_i zusammen zu einem Vektor $\vec{x} \in \mathbb{R}^N$ und die y_i zu einem Vektor $\vec{y} \in \mathbb{R}^n$, so läßt sich dieses Gleichungssystem kurz schreiben als

$$\vec{x} \cdot a + \vec{1} \cdot b = \vec{y} ,$$

wobei $\vec{1} \in \mathbb{R}^N$ jenen Vektor bezeichnen soll, dessen sämtliche Komponenten eins sind.

Die Matrix des Gleichungssystems ist somit die $N \times 2$-Matrix A mit Spalten $\vec{x}$ und $\vec{1}$. Da wir mit reellen Zahlen rechnen, ist A^* einfach die transponierte Matrix dazu, also die $2 \times N$-Matrix, in deren erster Zeile die x_i stehen, während in der zweiten lauter Einser stehen. Somit ist

$${}^tAA = \begin{pmatrix} \vec{x} \cdot \vec{x} & \vec{x} \cdot \vec{1} \\ \vec{x} \cdot \vec{1} & \vec{1} \cdot \vec{1} \end{pmatrix} \quad \text{und} \quad {}^tA\vec{b} = \begin{pmatrix} \vec{x} \cdot \vec{y} \\ \vec{1} \cdot \vec{y} \end{pmatrix} ,$$

das Gleichungssystem wird also zu

$$(\vec{x} \cdot \vec{x})a + (\vec{x} \cdot \vec{1})b = \vec{x} \cdot \vec{y} \quad \text{und} \quad (\vec{x} \cdot \vec{1}) + Nb = \vec{y} \cdot \vec{1} .$$

Seine Matrix ist genau dann singulär, wenn die beiden Spalten proportional zueinander sind, wenn also $N(\vec{x} \cdot \vec{x}) = (\vec{x} \cdot \vec{1})^2$ ist. Nach der CAUCHY-SCHWARZschen Ungleichung ist

$$\left|\vec{1} \cdot \vec{x}\right| \le \left|\vec{1}\right| \cdot |\vec{x}| = \sqrt{N}\,|\vec{x}| , \quad \text{also} \quad \left|\vec{1} \cdot \vec{x}\right|^2 \le N(\vec{x} \cdot \vec{x})$$

mit Gleichheit nur dann, wenn die Vektoren $\vec{x}$ und $\vec{1}$ linear abhängig sind, wenn also alle x_i gleich sind. In diesem Fall ist die erste Gleichung ein Vielfaches der zweiten, es gibt also unendlich viele Lösungen.

Andernfalls ist die Matrix invertierbar, die Lösung also eindeutig.

Führen wir die (in der Ausgleichsrechnung ziemlich verbreiteten) Abkürzungen

$$[x^r] = \sum_{i=1}^{N} x_i^r, \quad [y^r] = \sum_{i=1}^{N} x_i^r \quad \text{und} \quad [x^r y^s] = \sum_{i=1}^{N} x_i^r y_i^s$$

ein, so erhält das Gleichungssystem die übersichtlichere Gestalt

$$[x^2]a + [x]b = [xy] \quad \text{und} \quad [x]a + Nn = [y]\,.$$

Subtraktion von $[x]/[x^2]$ mal der ersten Gleichung von der zweiten führt auf

$$\left(N - \frac{[x]^2}{[x^2]}\right) b = [y] - \frac{[x]}{[x^2]}[xy]$$

oder $(N[x^2] - [x]^2)b = [y][x^2] - [x][xy]$, d.h.

$$b = \frac{[y][x^2] - [x][xy]}{N[x^2] - [x]^2}\,.$$

(Man beachte, daß im Falle der eindeutigen Lösbarkeit sowohl $[x^2] > 0$ als auch $N[x^2] - [x]^2 > 0$ ist.)

Einsetzen von b in die erste Gleichung ergibt dann auch

$$a = \frac{[xy] - [x]b}{[x^2]}\,.$$

Im Falle des Zusammenhangs zwischen Korruptionsindex CPI und Bruttoinlandsprodukt pro Einwohner BIP erhalten wir nach diesen Formeln die Ausgleichsgerade

$$CPI = 2{,}917 + 0{,}000163\,BIP\,,$$

die Steigung ist also erwartungsgemäß positiv. Der relativ große konstante Term zeigt, daß *im Mittel* Korruption selbst bei sehr armen Ländern

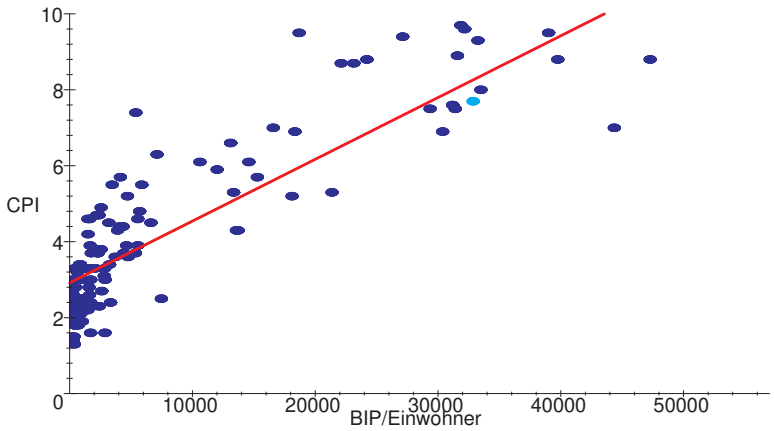


Abb. 19: Ausgleichsgerade zu Abbildung 18

deutlich über dem unteren Ende der Skala liegt. Abbildung 19 zeigt die Ausgleichsgerade zusammen mit den Daten.

Natürlich sind die Datenpunkte relativ breit gestreut um die Ausgleichsgerade; der Zusammenhang zwischen Korruption und Wohlstand ist schließlich zum Glück kein unausweichliches deterministisches Gesetz, sondern nur eine empirische Beobachtung.

Auch bei Messungen physikalischer Größen, wo die verschiedenen Meßgrößen meist durch wohlbekannte Naturgesetze miteinander verbunden sind, gibt es praktisch immer eine Streuung der Daten um die theoretisch richtige Meßkurve; absolut fehlerfreie Messungen sind, trotz aller Mühe der Experimentatoren, fast nie möglich, da es praktisch immer ein Grundrauschen der Meßgeräte und/oder nicht in ihrer Gesamtheit erfaßbare Umgebungseinflüsse *usw.* gibt. Vor allem bei Messungen, mit denen Konstanten für Naturgesetze ermittelt werden sollen oder gar ein Experiment zwischen zwei oder mehr Hypothesen entscheiden soll, ist es daher wichtig zu wissen, wie gut die Übereinstimmung zwischen den Daten und der berechneten Kurve (oder Fläche *usw.*) wirklich ist.

Solche Maße stellt die Statistik zur Verfügung; für ihr Verständnis sind

daher meist zumindest Grundlagenkenntnisse der Statistik notwendig, wie wir sie (wenn auch nur kurz) im nächsten Semester behandeln werden. Im einfachsten und zugleich wichtigsten Fall eines linearen Zusammenhangs zwischen zwei Größen allerdings reicht die lineare Algebra, um das sowohl in der Theorie wie auch den Anwendungen wichtigste Qualitätsmaß zu definieren, den Korrelationskoeffizienten.

Angenommen, wir haben N Datenpaare (x_i, y_i), zwischen denen ein perfekter linearer Zusammenhang besteht, d.h.

$$y_i = ax_i + b \quad \text{für alle } i = 1, \ldots, N\,.$$

Wir wollen den Datenvektoren $\vec{x} \in \mathbb{R}^n$ mit Komponenten x_i und $\vec{y} \in \mathbb{R}^N$ mit Komponenten y_i Vektoren zuordnen, die nicht nur in einem linearen Zusammenhang stehen, sondern sogar gleich sind; mit anderen Worten, wir wollen die Parameter a und b aus obiger Gleichung eliminieren.

Dazu betrachten wir als erstes die Mittelwerte

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{und} \quad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i\,.$$

Da $y_i = ax_i + b$ ist für alle i, folgt sofort, daß auch

$$\overline{y} = a\overline{x} + b$$

ist, und damit

$$(y_i - \overline{y}) = a(x_i - \overline{x}) \quad \text{für alle } i = 1, \ldots, N\,.$$

Damit haben wir den Parameter b eliminiert. Bezeichnen wir wieder mit $\vec{1} \in \mathbb{R}^N$ den Vektor, dessen sämtliche Komponenten Einser sind, ist nun also $(\vec{y} - \overline{y}\vec{1}) = a(\vec{x} - \overline{x}\vec{1})$. Aus dieser Gleichung können wir nun leicht a bis auf sein Vorzeichen eliminieren, indem wir die beiden Vektoren durch ihre Länge dividieren. Dies ist natürlich nur möglich, wenn keiner der beiden Vektoren gleich dem Nullvektor ist, wenn also nicht alle $x_i = \overline{x}$ oder alle $y_i = \overline{y}$ sind. Bei nicht getürkten Messungen ist dies allerdings *praktisch* nie der Fall, so daß die Nützlichkeit der folgenden Diskussion und Definition nicht darunter leidet, daß wir diesen Fall ausschließen müssen.

Falls also weder $\vec{y} - \overline{y}\vec{1}$ noch $\vec{x} - \overline{x}\vec{1}$ der Nullvektor ist, betrachten wir die beiden auf Länge eins normierten Vektoren

$$\frac{\vec{y} - \overline{y}\vec{1}}{\left|\vec{y} - \overline{y}\vec{1}\right|} \quad \text{und} \quad \frac{\vec{x} - \overline{x}\vec{1}}{\left|\vec{x} - \overline{x}\vec{1}\right|}\,.$$

Diese sind nun offensichtlich entweder gleich (für $a > 0$) oder entgegengesetzt gleich (für $a < 0$).

Wenn (wie in der Realität meist der Fall) *kein* perfekter linearer Zusammenhang zwischen den x_i und den y_i besteht, können wir trotzdem – falls weder $\vec{y} - \overline{y}\vec{1}$ noch $\vec{x} - \overline{x}\vec{1}$ der Nullvektor ist – die beiden Vektoren

$$\frac{\vec{y} - \overline{y}\vec{1}}{\left|\vec{y} - \overline{y}\vec{1}\right|} \quad \text{und} \quad \frac{\vec{x} - \overline{x}\vec{1}}{\left|\vec{x} - \overline{x}\vec{1}\right|}$$

betrachten. Da beides Einheitsvektoren sind, unterscheiden sie sich nur in der Richtung; als Maß für ihren Unterschied bietet sich daher den Winkel zwischen $\vec{x}$ und $\vec{y}$ an. Rechnerisch einfacher ist der Cosinus dieses Winkels, denn der ist bei Einheitsvektoren einfach gleich dem Skalarprodukt.

Definition: Der Korrelationskoeffizient zwischen zwei Datenvektoren $\vec{x}$ und $\vec{y} \in \mathbb{R}^n$, die keine Vielfachen des Vektors $\vec{1} \in \mathbb{R}^N$ sind, ist

$$\rho = \frac{(\vec{x} - \overline{x}\cdot\vec{1})\cdot(\vec{y} - \overline{y}\cdot\vec{1})}{\left|\vec{x} - \overline{x}\cdot\vec{1}\right| \cdot \left|\vec{y} - \overline{y}\cdot\vec{1}\right|}\,.$$

Damit ist also $\rho = \pm 1$ genau dann, wenn es einen perfekten linearen Zusammenhang $y_i = ax_i + b$ zwischen den beiden Größen gibt, mit $\rho = 1$ für $a > 0$ und $\rho = -1$ für $a < 0$. Ansonsten ist der Zusammenhang umso besser, je größer der Betrag von ρ ist. Für $\rho = 0$ stehen die beiden Vektoren $\vec{x} - \overline{x}\vec{1}$ und $\vec{y} - \overline{y}\vec{1}$ senkrecht aufeinander, d.h. wenn x_i größer ist als der Mittelwert $\overline{x}$, kann y_i im Mittel genauso gut größer wie auch kleiner als der Mittelwert $\overline{y}$ sein. (in der Statistik ist dies die *Definition* für die Unabhängigkeit von Daten.)

Definition: Zwei Größen x und y heißen $\left\{\begin{matrix}\text{positiv}\\\text{negativ}\end{matrix}\right\}$ korreliert, wenn $\rho \left\{\begin{matrix}>\\<\end{matrix}\right\} 0$ ist. Sie heißen unkorreliert oder voneinander unabhängig, wenn $\rho = 0$ ist.

Im Beispiel der Korruption erhalten wir einen Korrelationskoeffizienten von

$$\rho \approx 0{,}8627227\,;$$

dies entspricht einem Winkel von etwa $30{,}376^\circ$ zwischen den oben definierten Vektoren.

Um ein Gefühl für Korrelationskoeffizienten zu bekommen, wollen wir zwei Beispiele betrachten, die sich zumindest visuell sehr unterscheiden:

Der CPI für Deutschland hatte in den letzten Jahren folgende Werte:

Jahr:	1980–1985	1988–1992	1995	1996	1997	1998
CPI:	8,14	8,13	8,14	8,27	8,23	7,9
Jahr:	1999	2000	2001	2002	2003	
CPI:	8,0	7,6	7,4	7,3	7,7	

Wie Abbildung 20 zeigt, sieht der Zusammenhang zwischen Jahr und CPI nicht sonderlich linear aus: Der Bimbesknick ist unverkennbar, jedoch scheint die Talsohle inzwischen durchschritten. Der Korrelationskoeffizient $\kappa \approx -0{,}634$ sieht auf den ersten Blick recht gut aus; bedenkt man aber, daß dies der Kosinus eines Winkels von rund 130° ist, sieht man schon diesem Wert an, daß der Zusammenhang nicht sonderlich gut durch eine lineare Funktion beschrieben werden kann. In der Tat ist die eingezeichnete Ausgleichsgerade nicht einmal eine grobe qualitative Beschreibung des Zusammenhangs.

Vergleichen wir dagegen die Ergebnisse von Europawahl und Gemeinderatswahl am 13. Juni 2004 miteinander, so sind die Stimmenanteile der vier Parteien, die an beiden teilgenommen haben, zwar teilweise deutlich verschieden, der Zusammenhang ist aber, wie Abbildung 21 zeigt, doch in recht guter Näherung linear. In der Tat erhalten wir hier $\kappa \approx 0{,}9900614$, was einem Winkel von etwa acht Grad entspricht.

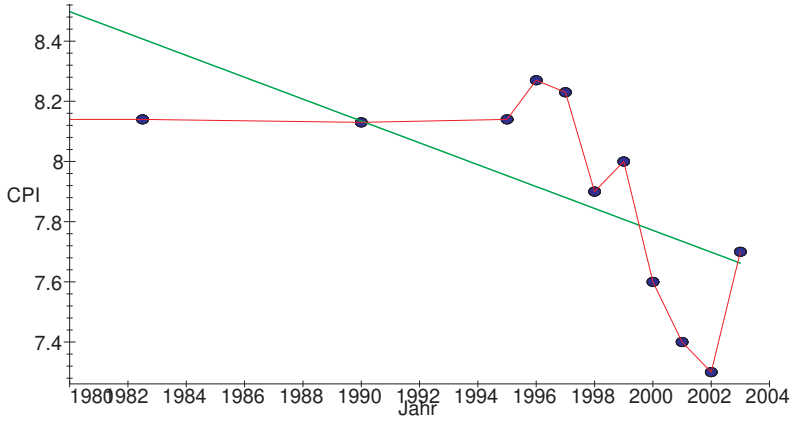


Abb. 20: Zeitabhängigkeit des CPI für Deutschland

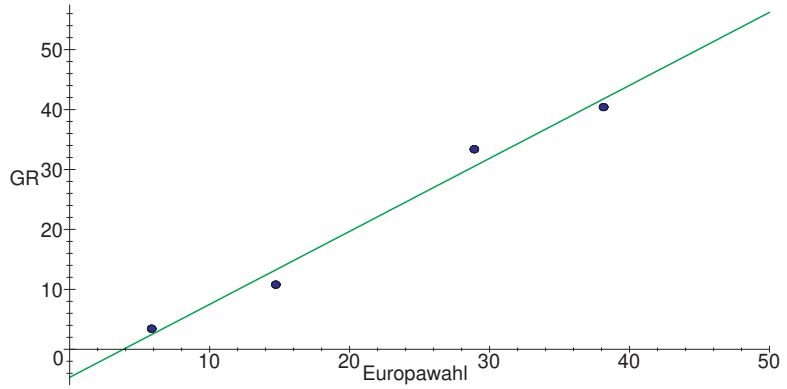


Abb. 21: Zusammenhang Europawahl/Gemeinderatswahl

	Europawahl	*Gemeinderatswahl*
CDU	38,14	40,41
SPD	28,91	33,38
Grüne	14,72	10,19
FDP	5,86	3,43

Korrelationskoeffizienten mit kleinem Betrag müssen nicht unbedingt bedeuten, daß kein deterministischer Zusammenhang zwischen den Daten besteht:

In Fällen schlechter Korrelation liegt es nahe, zu Alternativen nach einem linearen Zusammenhang zu fragen.

j) Euklidische Vektorräume in der Informationssuche

Internetsuchmaschinen finden innerhalb von Sekundenbruchteilen aus mehreren Millionen untersuchten Dokumenten mehr oder weniger gut passende zu einer Suchanfrage; genauso haben wissenschaftliche Literaturdatenbanken zumindest Tausende (meist deutlich mehr) wissenschaftlicher Arbeiten gespeichert, aus denen ein Anwender die für seine Forschung relevanten finden möchte. Am einfachsten geht das, wenn entweder der Autor oder ein Berichterstatter die Arbeiten nach Themengebieten ordnet: In der Mathematik etwa gibt es ein umfangreiches Klassifikationsschema der beiden westlichen Referatorgane *Zentralblatt für Mathematik und ihre Grenzgebiete* und *Mathematical Reviews,* das auch die meisten Fachzeitschriften verwenden; in anderen Wissenschaften ist es ähnlich.

Solche Zuordnungen sind meistens recht genau, da sie von Experten der jeweiligen *Teil*gebiete vorgenommen werden; andererseits birgt natürlich auch gerade das die Gefahr in sich, daß eine Arbeit, die für mehrere Gebiete relevant ist, möglicherweise nur denen zugeordnet wird, für die sich der Autor oder Berichterstatter interessiert. Außerdem ist selbst eine sehr detailierte Einteilung, die im Falle der Mathematik immerhin 35 Seiten Kleingedrucktes benötigt, immer noch zu grob, um genau *die* drei Arbeiten zu finden, in denen ein sehr spezielles Problem behandelt wird. Im Internet mit seiner Vielzahl von teilweise sehr schnell variierenden Informationsangeboten ist ein solcher Ansatz von vornherein chancenlos.

Zusätzlich zur Klassifikation durch menschliche Experten braucht man daher bei der Informationssuche auch noch Algorithmen, die gelesene Informationen automatisch klassifizieren und bezüglich ihrer Relevanz zu einer konkreten Suchanfrage beurteilen können.

Große Internetsuchmaschinen verwenden dazu eine Vielzahl von Algorithmen; mit Ausnahme von `google.com`, die ihr Rangbildungsverfahren unter

`http://www.google.com/technology/pigeonrank.html`

mehr oder weniger ausführlich beschreiben, schweigen sie sich allerdings aus über die genauen Einzelheiten und Parameter: Schließlich sollen die vielen unseriösen Anbieter, die mit allen Tricks Besucher auf ihre Webseiten locken wollen, nicht auch noch unterstützt werden.

Wir müssen uns daher auf die grundlegenden mathematischen Algorithmen beschränken, die wohl in der einen oder andereren Form in praktisch jeer Suchmaschine zu finden sind und die, als Gegenstand wissenschaftlicher Forschung, natürlich öffentlich bekannt sind.

Die ersten Systeme arbeiteten mit den üblichen Suchalgorithmen aus der Textverarbeitung, durchsuchten also alle gespeicherten Dokumente nach dem Vorkommen einer oder mehrerer vorgegebener Zeichenketten. Auch wenn es dafür sehr effiziente Algorithmen gibt, ist dieses Verfahren bei wirklich großen Datenmengen nicht mehr mit realistischem Aufwand durchführbar, so daß nun meist Verfahren aus der linearen Algebra verwendet werden.

Dazu wird eine Liste von Suchbegriffen s_i, $i = 1, \cdots, n$ festgelegt – beispielsweise die Wörter aus einem Wörterbuch der Dokumentsprache. Oftmals werden darauf noch geeignete Operationen angewandt wie *stemming,* d.h. Wörter mit gleichem Stamm werden miteinander identifiziert, oder *latent semantic indexing,* wo durch Clusterbildung bei den vorhandenen Dokumenten Begriffspaare identifiziert werden, die im allgemeinen im gleichen Kontext auftreten und die dann auch bei Suchanfragen als äquivalent betrachtet werden; außerdem werden sogenannte „Nullwörter“, die für Suchanfragen typischerweise ohne Bedeutung sind, eliminiert. Dabei handelt es sich beispielsweise um Artikel und Praepositionen, gelegentlich aber auch um spezifische Wörter aus dem Kontext des jeweiligen Systems: Bei Boeing, die ein solches System zur Verwaltung ihrer Wartungshandbücher aufbauten, ist etwa das Wort „aeroplane“ ein Nullwort – die Gesellschaft verkauft schließlich keine Rasenmäher.

Sind nun m Dokumente zu betrachten, so bildet man eine $n \times m$-Matrix A, deren Eintrag a_{ij} etwas über das Vorkommen des i-ten Suchbegriffs im j-ten Dokument aussagt. Im einfachsten Fall setzt man einfach $a_{ij} = 1$, falls der Begriff vorkommt und null sonst, alternativ kann a_{ij} auch die Häufigkeit des Begriff im Dokument sein, wobei diese Häufigkeit oft noch gewichtet wird, indem beispielsweise Vorkommen im vorderen Teil des Dokuments höher gewichtet wird oder aber die Suchmaschine ohnehin nur den Anfangsteil des Dokuments bis zu einer gewissen Maximallänge berücksichtigt. Auch das Vorkommen in Überschriften oder zwischen <META>-tags kann eventuell gesondert behandelt werden, indem man beispielsweise Inhalte, die im Browserfenster nicht sichtbar werden, wegen der damit verbundenen Mißbrauchsmöglichkeit ignoriert. Gelegentlich wird auf das Ergebnis noch eine Skalierungsfunktion wie etwa $\log(1 + x)$ angewendet.

Die entstehende Matrix ist natürlich riesig; schon 1998 wurde geschätzt, daß allein für englischsprachige Dokumente bis zu 300 000 Suchbegriffe notwendig sind, die in etwa 300 Millionen Dokumenten gesucht werden müssen; die Matrix hat also knapp hundert Billionen Einträge. Bei nur einem Byte pro Einträg hätte man also bei der Speicherung als Feld einen Platzbedarf von etwa 90 Terabyte.

Nun kommt allerdings in fast jedem Dokument nur ein verschwindend geringer Bruchteil der Suchbegriffe vor, so daß die meisten Einträge von A Nullen sind. Die Matrix läßt sich daher kompakter erheblich kompakter speichern, wenn man beispielsweise nur die Tripel (i, j, a_{ij}) notiert, für die $a_{ij} \neq 0$ ist. Die numerische Mathematik kennt eine ganze Reihe von Algorithmen, mit denen man auch solche sogenannte „spärlich besetzte“ Matrizen effizient behandeln kann.

Der Inhalt des j-ten Dokuments wird nun also kodiert durch den j-ten Spaltenvektor der Matrix A, einen Vektor aus $\mathbb{R}^n$. Auch eine Suchanfrage läßt sich durch einen solchen Vektor kodieren, indem man die j-te Komponente auf Eins setzt, falls der j-te Suchbegriff in der Anfrage vorkommt, und auf Null sonst. (Man kann natürlich auch andere Werte wählen und beispielsweise seltene Wörter höher gewichten als häufige *usw.*)

Ein Dokument sollte nun umso besser zu einer Suchanfrage passen, je weniger sich die dazu gehörigen Vektoren voneinander unterscheiden. Als Maß für den Unterschied zweier Vektoren haben wir im vorigen Abschnitt den Cosinus des eingeschlossenen Winkels kennengelernt; falls man die Spaltenvektoren der Matrix auf Länge Eins normiert, läst sich dieser durch eine einziges Skalarprodukt berechnen. Ein Dokument wird dann als relevant für die Suchanfrage betrachtet, wenn dieser Wert über einer festzulegenden Schranke liegt, und die so gefundenen Dateien können dann eventuell noch mit anderen Methoden (Volltextsuche, Links von anderern Seiten, ...) weiter untersucht werden zur Festlegung der endgültigen Reihenfolge, in der sie dem Benutzer gezeigt werden.

Für sehr große Datenmengen ist allerdings die Matrix A trotz ihrer spärlichen Besetzung immer noch zu groß; wie bei der Komprimierung von Bilddaten sucht man daher nach einer Art und Weise, sie bei möglichst geringem Informationsverlust deutlich zu komprimieren. Ein angenehmer Nebeneffekt dabei ist, wie experimentelle Untersuchungen zeigen, auch eine gewisse „Rauschunterdrückung“: Es ist zwar schwierig, exakt zu definieren, was „Rauschen“ in einer Term-Dokument-Matrix sein soll, aber jeder wird wohl damit übereinstimmen, daß etwa dieses Skriptum nicht die ideale Referenz zum Thema „Rasenmäher“ ist, obwohl dieses Wort hier nun schon zum zweiten Mal vorkommt.

Einen Ansatz zur Datenreduktion liefert die QR-Zerlegung: Ist $A = QR$ und $\vec{a} \in \mathbb{R}^n$ eine Suchanfrage, so ist für die j-ten Spalten $\vec{a}_j$ von A und $\vec{r}_j$ von R

$$\vec{a} \cdot \vec{a}_j = \vec{a} \cdot (Q\vec{r}_j) = ({}^tQ\vec{a}) \cdot \vec{r}_j \,,$$

und die Matrix R wird im allgemeinen deutlich mehr Nullen enthalten als A, da der Rang von A wohl deutlich unter n liegen dürfte. Eine weitere Komprimierung wird dadurch erreicht, daß Einträge von R, die unterhalb einer gewissen Schranke liegen, auf Null gesetzt werden; dadurch ändert sich bei hinreichend kleiner Schranke an den meisten Skalarprodukten nicht viel, dafur verringert sich aber der Speicherbedarf noch einmal beträchtlich.

Oft verwendet man anstelle der QR-Zerlegung auch die hier nicht be-

handelte Singulärwertzerlegung von A: Danach läßt sich A schreiben als Produkt UDV mit orthogonalen Matrizen $U \in \mathbb{R}^{n\times n}$ und $V \in \mathbb{R}^{m\times m}$ sowie einer Diagonalmatrix $D \in \mathbb{R}^{n\times m}$. U und V können so gewählt werden, daß die Diagonaleinträge der Größe nach angeordnet sind, und man erhält die gewünschte Rangreduktion, indem man alle Einträge unterhalb einer gewissen Größe auf Null setzt.

Eine ausführlichere Darstellung der Verfahren zur Textsuche, die keine über den Inhalt dieses Skriptums hinausgehende Mathematikkenntnisse voraussetzt, findet man beispielsweise in

MICHAEL W. BERRY, MURRAY BROWNE: Understanding Search Engines: Mathematical Modeling and Text Retrieval, SIAM, 1999,

Fallstudien im Tagungsband

MICHAEL W. BERRY [Hrsg.]: Computational Information Retrieval, SIAM, 2001.

§6: Volumina und Determinanten

Im vorigen Paragraphen haben wir Längen und Winkel eingeführt, aber noch keine Flächeninhalte und Volumina. Darum soll es nun in diesem Paragraphen gehen.

a) Das Vektorprodukt im $\mathbb{R}^3$

Beginnen wir mit dem Flächeninhalt eines Parallelogramms in $\mathbb{R}^2$, aufgespannt von zwei Vektoren $\vec{v}$ und $\vec{w}$.

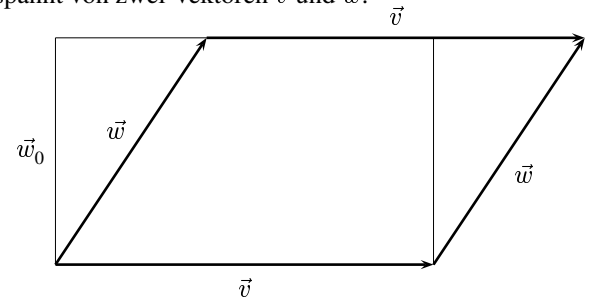


Abb. 22: Flächeninhalt eines Parallelogramms

Durch Scherung können wir aus dem Parallelogramm ein Rechteck machen, dessen eine Seite der Vektor $\vec{v}$ ist, die andere ist jener Vektor $\vec{w}_0$ der aus $\vec{w}$ entsteht durch Projektion auf die Senkrechte zu $\vec{v}$. Die beiden Vierecke haben denselben Flächeninhalt, da die beiden Dreiecke um die sie sich links und rechts unterscheiden kongruent sind. Der Flächeninhalt ist somit gleich dem Produkt der Längen von $\vec{v}$ und $\vec{w}_0$.

Da der Cosinus eines Winkels im rechtwinkligen Dreieck gleich Ankathete durch Hypothenuse ist, und im übrigen auch gleich der Sinus des Komplementärwinkels, ist der Flächeninhalt somit gleich $|\vec{v}|\,|\sin\measuredangle(\vec{v},\vec{w})|$.

Im Dreidimensionalen (und nur dort) können wie ein Produkt von Vektoren definieren, daß diesen Betrag hat: das sogenannte *Vektorprodukt* oder *Kreuzprodukt*. Wie die beiden Namen sagen, liefert es als Ergebnis einen Vektor und wird mit $\vec{v}\times\vec{w}$ bezeichnet.

Der Vektor $\vec{v}\times\vec{w}$ soll die Länge

$$|\vec{v}\times\vec{w}| = |\vec{v}|\,|\vec{w}|\,|\sin\measuredangle(\vec{v},\vec{w})|$$

haben und senkrecht stehen sowohl auf $\vec{v}$ als auch auf $\vec{w}$. Insbesondere ist damit $\vec{v}\times\vec{w} = \vec{0}$ genau dann, falls $\vec{v}$ und $\vec{w}$ dieselbe (oder entgegengesetzte) Richtung haben, oder (mindestens) einer der beiden Vektoren der Nullvektor ist, d.h. also genau dann, wenn $\vec{v}$ und $\vec{w}$ linear abhängig sind.

In allen anderen Fällen spannen $\vec{v}$ und $\vec{w}$ eine Ebene auf, und $\vec{v}\times\vec{w}$ muß auf dieser Ebenen senkrecht stehen. Dazu gibt es zwei Möglichkeiten: Entweder er steht ab nach oben oder nach unten (wie immer man auch oben und unten festlegen möchte; für eine beliebige Ebene ist das natürlich völlig willkürlich).

Um diese beiden Fälle voneinander zu unterscheiden, benötigen wir einen neuen Begriff: Wir sagen, daß drei Vektoren $\vec{v}_1$, $\vec{v}_2$ und $\vec{v}_3$, die nicht in einer Ebenen liegen, ein *Rechtssystem* bilden, wenn sich die Finger der *rechten* Hand so ausrichten lassen, daß der Daumen in Richtung von $\vec{v}_1$ zeigt, der Zeigefinger in Richtung von $\vec{v}_2$ und der Mittelfinger in Richtung von $\vec{v}_3$. Alternativ kann man es, wenn der dritte Vektor so wie hier senkrecht auf den beiden ersten steht, auch so definieren, daß sich ein

von $\vec{v}_1$ nach $\vec{v}_2$ gedrehter Korkenzieher in Richtung $\vec{v}_3$ in den Kork bohrt. Ähnlich kann man das Rechtssystem auch mit Schrauben definieren; da es allerdings neben den (üblichen) Rechtsschrauben auch die (seltenen) Linksschrauben gibt, ist diese Definition eventuell zirkulär: Alles hängt davon ab, wie man Rechtsschrauben definiert.

Aus jeder dieser Regeln folgt sofort die *Antikommutativität* des Vektorprodukts:

$$\vec{v} \times \vec{w} = -\vec{w} \times \vec{v}\,.$$

Die Linearität des Vektorprodukts im ersten Argument folgt sofort aus der Interpretation als Flächeninhalt eines Parallelogramms; wir haben also

$$\vec{v} \times (\vec{w} + \vec{u}) = \vec{v} \times \vec{w} + \vec{v} \times \vec{u}$$

und wegen der Antikommutativität entsprechend auch

$$(\vec{u} + \vec{v}) \times \vec{w} = \vec{u} \times \vec{w} + \vec{v} \times \vec{w}\,.$$

Um das Vektorprodukt in Koordinaten ausrechnen zu können, müssen wir zunächst die Produkte der $\vec{e}_i$ kennen; diese können wir ausrechnen, sobald wir wissen, ob $\vec{e}_1, \vec{e}_2$ und $\vec{e}_3$ ein Rechtssystem oder ein (analog zu definierendes) Linkssystem bilden. Wir nehmen an, daß sie ein Rechtssystem bilden; dann folgt sofort, daß $\vec{e}_1 \times \vec{e}_2 = \vec{e}_3$ ist, und nach einigen Fingerübungen auch, daß

$$\vec{e}_2 \times \vec{e}_3 = \vec{e}_1 \quad \text{und} \quad \vec{e}_1 \times \vec{e}_3 = -\vec{e}_2$$

ist. Die Produkte mit vertauschten Faktoren sind natürlich gerade das negative davon, und $\vec{e}_i \times \vec{e}_i = 0$. Für

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad \text{und} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

ist also

$$\vec{v} \times \vec{w} = (v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3) \times (w_1\vec{e}_1 + w_2\vec{e}_2 + w_3\vec{e}_3)\,,$$

und nach obigen Rechenregeln ist dies gleich

$$\sum_{i=1}^{3}\sum_{j=1}^{3} v_i w_j\, \vec{e}_i \times \vec{e}_j$$
$$= (v_2w_3 - v_3w_2)\vec{e}_1 + (v_3w_1 - v_1w_3)\vec{e}_2 + (v_1w_2 - v_2w_1)\vec{e}_3\,,$$

d.h.

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \times \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} v_2w_3 - v_3w_2 \\ v_3w_1 - v_1w_3 \\ v_1w_2 - v_2w_1 \end{pmatrix}\,.$$

Dies läßt sich dadurch merken, daß man im Schema

$$\begin{array}{ccccccccc} \vec{e}_1 & & \vec{e}_2 & & \vec{e}_3 & & \vec{e}_1 & & \vec{e}_2 \\ & \searrow & & \diagdown\!\!\!\diagdown & & \diagdown\!\!\!\diagdown & & \swarrow & \\ v_1 & & v_2 & & v_3 & & v_1 & & v_2 \\ & \swarrow & & \diagdown\!\!\!\diagdown & & \diagdown\!\!\!\diagdown & & \searrow & \\ w_1 & & w_2 & & w_3 & & w_1 & & w_2 \end{array}$$

von $\vec{e}_i$ ausgeht und als dessen Koeffizient das Zweierprodukt entlang der schrägen Linie nach rechts unten *positiv* und das entlang der schrägen Linie nach links unten *negativ* nimmt.

In weiteren Verlauf dieses Paragraphen geht es uns um Volumina n-dimensionaler Körper in $\mathbb{R}^n$; wenn wir für $n = 2$ das „Volumen“, also den Flächeninhalt des von $\vec{v} = \binom{v_1}{v_2}$ und $\vec{w} = \binom{w_1}{w_2}$ aufgespannten Parallelogramms berechnen wollen, können wir also diese beiden Vektoren in den $\mathbb{R}^3$ einbetten durch Anfügen einer dritten Komponente null, und erhalen als Volumen dann den Betrag des Vektorprodukts der so entstehenden Vektoren. Da das Vektorprodukt selbst auf $\vec{v}$ und vw senkrecht steht, hat der Vektor nur eine einzige Komponenten, die bis aufs Vorzeichen gleich der Länge ist:

$$\begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix} \times \begin{pmatrix} w_1 \\ w_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ v_1w_2 - v_2w_1 \end{pmatrix}$$

Wir definieren deshalb

Definition: Für zwei Vektoren $\vec{v} = \binom{v_1}{v_2}$ und $\vec{w} = \binom{w_1}{w_2}$ aus dem Vektorraum k^2 über einem Körper k bezeichnen wir die Zahl

$$\det(\vec{v}, \vec{w}) \underset{\text{def}}{=} \begin{vmatrix} v_1 & w_1 \\ v_2 & w_2 \end{vmatrix} \underset{\text{def}}{=} v_1 w_2 - v_2 w_1$$

als *Determinante* von $\vec{v}$ und $\vec{w}$.

Für $k = \mathbb{R}$ ist $\det(\vec{v}, \vec{w})$ also bis aufs Vorzeichen gerade der Flächeninhalt des von $\vec{v}$ und $\vec{w}$ aufgespannten Parallelogramms; oft bezeichnet man auch $\det(\vec{v},\vec{w})$ *mit* Vorzeichen als den *orientierten* Flächeninhalt des von $\vec{v}$ und $\vec{w}$ aufgespannten Parallelogramms.

Über sonstigen Körpern können wir nicht sinnvollerweise von Flächen reden, aber auch unabhängig von der Herleitung über das Kreuzprodukt ist klar, daß $\vec{v}$ und $\vec{w}$ genau dann linear abhängig sind, wenn $\det(\vec{v}, \vec{w})$ verschwindet. Genauer haben wir folgende drei Eigenschaften:

(D1) $\det(\vec{u}, \vec{v}) = -\det(\vec{v}, \vec{u})$

(D2) $\det(\vec{u}, \vec{v})$ ist linear in jedem ihrer beiden Argumente, d.h.

$$\det(\lambda\vec{u}_1 + \mu\vec{u}_2, \vec{v}) = \lambda \det(\vec{u}_1, \vec{v}) + \mu \det(\vec{u}_2, \vec{v})$$

und

$$\det(\lambda(\vec{u}, \lambda\vec{v}_1 + \mu\vec{v}_2) = \lambda \det(\vec{u}, \vec{v}_1) + \mu \det(\vec{u}, \vec{v}_2)\,.$$

(D3) $\det(\vec{u}, \vec{v}) = 0$ genau dann, wenn $\vec{u}$ und $\vec{v}$ linear abhängig sind.

b) Das Spatprodukt

Die Interpretation der Länge des Vektorprodukts als Fläche eines Parallelogramms wirft die Frage auf, ob man vielleicht auch Volumina auf diese Weise berechnen kann; dies würde dann auch auf ein Kriterium für die lineare Abhängigkeit liefern, denn drei Vektoren im $\mathbb{R}^3$ sind genau dann linear unabhängig, wenn sie einen Körper mit Volumen ungleich null aufspannen.

Zwei Vektoren liegen stets in einer Ebenen; kommt noch ein *dritter* Vektor dazu, entsteht ein dreidimensionales Analogon zum Parallelogramm, das in Abbildung 23 gezeigte *Parallelepiped,* das man sich als einen verzerrten Quader vorstellen kann, dessen rechte Winkel zu Winkeln einer

beliebigen Größe wurden. Es wird auch als *Spat* bezeichnet, da Kalkspat und einige Feldspäte gelegentlich Kristalle dieser Form bilden. Aus dem Physikunterricht der Schule ist vielleicht der sogenannte Isländischen Doppelspat bekannt, eine der vielen Formen des Kalkspats: Dieser bildet meist sehr schöne Parallelepipede und ist vor allem bekannt durch seine ungewöhnlich stark ausgeprägte Doppelbrechung, die deshalb sehr häufig anhand dieser Kristalle demonstriert wird. (Sprachlich ist die Bezeichnung *Spat* für das Parallelepiped nicht sonderlich sinnvoll, denn die Endung „spat“ beim Feldspat oder Kalkspat deutet nur auf die leichte Spaltbarkeit des Materials hin; sie hat nichts mit der geometrischen Form zu tun.)

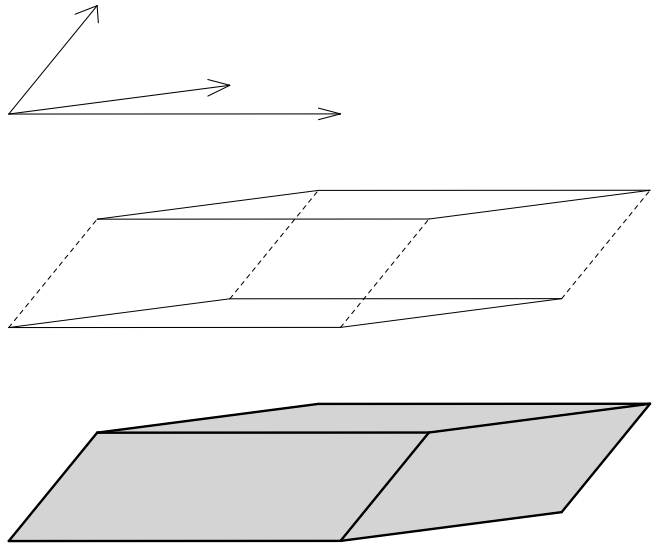


Abb. 23: Das von drei Vektoren aufgespannte Parallelepiped

Das Volumen eines Parallelepipeds läßt sich auf ähnliche Weise be-

rechnen wie das eines Parallelogramms: Durch Scherung wie in Abbildung 24 kann man erreichen, daß der dritte Vektor senkrecht auf den beiden anderen steht; das Volumen des Parallelepipeds ist also gleich dem Volumen des entsprechenden Prismas. Das Volumen eines Prismas wiederum ist gleich der Grundfläche mal der Höhe, und die Grundfläche ist eine uns bekannte Parallelogrammfläche.

Bezeichnen wir also die Vektoren der Reihe nach als $\vec{u}, \vec{v}$ und $\vec{w}$, so ist die Fläche des von $\vec{u}$ und $\vec{v}$ aufgespannten Parallelogramms gleich $|\vec{u} \times \vec{v}|$. Die Höhe des Prisma ist gleich der Länge des auf eine Achse senkrecht zum Parallelogramm projezierten Vektors $\vec{w}$.

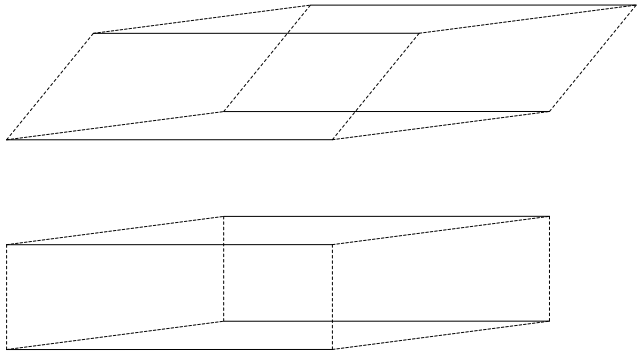


Abb. 24: Scherung eines Parallelepipeds zum Prisma

Diese Achse wird beispielsweise aufgespannt durch den Vektor $\vec{u} \times \vec{v}$; wir können die Höhe also berechnen als Betrag des Skalarprodukt von $\vec{w}$ mit einem Einheitsvektor, der dieselbe Richtung hat wie $\vec{u} \times \vec{v}$:

$$h = \left| \frac{\vec{u} \times \vec{v}}{|\vec{u} \times \vec{v}|} \cdot \vec{w} \right| = \frac{|(\vec{u} \times \vec{v}) \cdot \vec{w}|}{|\vec{u} \times \vec{v}|} .$$

Das Volumen des Prismas und damit des Parallelepipeds ist somit gleich

$$|\vec{u} \times \vec{v}| \cdot \frac{|(\vec{u} \times \vec{v}) \cdot \vec{w}|}{|\vec{u} \times \vec{v}|} = |(\vec{u} \times \vec{v}) \cdot \vec{w}| .$$

Wir bezeichnen deshalb das Produkt

$$(\vec{u} \times \vec{v}) \cdot \vec{w}$$

der drei Vektoren $\vec{u}, \vec{v}$ und $\vec{w}$ als das *Spatprodukt* dieser Vektoren; sein Betrag ist gleich dem Volumen des von den drei Vektoren aufgespannten Parallelepipeds. Da dieses Volumen nicht von der Reihenfolge der Vektoren abhängt, ist auch das Spatprodukt bis auf Vorzeichen unabhängig von der Reihenfolge der Vektoren; Nachrechnen zeigt, daß

$$(\vec{u} \times \vec{v}) \cdot \vec{w} = (\vec{v} \times \vec{w}) \cdot \vec{u} = (\vec{w} \times \vec{u}) \cdot \vec{v}$$

dasselbe Vorzeichen haben; das andere Vorzeichen haben die drei Produkte

$$(\vec{v} \times \vec{u}) \cdot \vec{w} = (\vec{w} \times \vec{v}) \cdot \vec{u} = (\vec{u} \times \vec{w}) \cdot \vec{v} ,$$

was man schon daran sieht, daß hier die Reihenfolge der Faktoren des Kreuzprodukts umgekehrt wurde.

In Koordinaten ausgedrückt ist $(\vec{u} \times \vec{v}) \cdot \vec{w}$

$$= \left(\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \right) \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

$$= u_2v_3w_1 - u_3v_2w_1 + u_3v_1w_2 - u_1v_3w_2 + u_1v_2w_e - u_2v_1w_3 .$$

Diese Formel läßt sich leicht merken nach der folgenden Regel von PIERRE SARRUS (1798–1861), von der wir bereits beim Kreuzprodukt einen Spezialfall kennengelernt haben: Schreibt man die Komponenten der Vektoren in der Form

$$\begin{array}{ccccccccc} u_1 & & v_1 & & w_1 & & u_1 & & v_1 \\ & \searrow & & \searrow & & \searrow & & \swarrow & \\ u_2 & & v_2 & & w_2 & & u_2 & & v_2 \\ & \swarrow & & \searrow & & \searrow & & \searrow & \\ u_3 & & v_3 & & w_3 & & u_3 & & v_3 \end{array} ,$$

so sind die Produkte entlang der drei schrägen Linien von links oben nach rechts unten, der Hauptdiagonalen und ihrer Parallelen also, positiv zu rechnen und die entlang der drei schrägen Linien von rechts oben nach links unten negativ.

Definition: Für drei Vektoren

$$\vec{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \quad \vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad \text{und} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

aus dem Vektorraum k^3 über einem Körper k bezeichnen wir die Zahl

$$\det(\vec{u}, \vec{v}, \vec{w}) \underset{\text{def}}{=} \begin{vmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{vmatrix}$$
$$\underset{\text{def}}{=} u_2v_3w_1 - u_3v_2w_1 + u_3v_1w_2 - u_1v_3w_2 + u_1v_2w_e - u_2v_1w_3$$

als *Determinante* von $\vec{u}, \vec{v}$ und $\vec{w}$.

Da der Betrag der Determinante für $k = \mathbb{R}$ das Volumen des von $\vec{u}, \vec{v}$ und $\vec{w}$ aufgespannten Parallelepipeds ist, ist dieser Betrag unabhängig von der Reihenfolge der Vektoren. Das Vorzeichen hängt aber natürlich von der Reihenfolge ab, denn wegen der Schiefsymmetrie des Vektorprodukts ist beispielsweise

$$\det(\vec{v}, \vec{u}, \vec{w}) = -\det(\vec{u}, \vec{v}, \vec{w})\,;$$

entsprechend ändert sich das Vorzeichen, wie man sich leicht überzeugt, bei jeder Vertauschung der Reihenfolge zweier Vektoren.

Da sowohl das Vektorprodukt als auch das Skalarprodukt linear in ihren Argumenten sind, ist auch det linear in jedem ihrer drei Argumente, wir haben also die drei Eigenschaften

(D1) $\det(\vec{u}, \vec{v}, \vec{w})$ ändert ihr Vorzeichen, nicht aber ihren Betrag, wenn irgendwelche zwei der Argumente miteinander vertauscht werden.

(D2) $\det(\vec{u}, \vec{v}, \vec{w})$ ist linear in jedem ihrer drei Argumente.

(D3) $\det(\vec{u}, \vec{v}, \vec{w}) = 0$ genau dann, wenn $\vec{u}, \vec{v}$ und $\vec{w}$ linear abhängig sind.

c) Forderungen an eine allgemeine Determinante

Wir möchten gerne für beliebige endlichdimensionale Vektorräume V eine Determinantenfunktion definieren, d.h. wir suchen für einen n-

dimensionalen Vektorraum V eine Abbildung

$$\det: \underbrace{V \times \cdots \times V}_{n \text{ mal}} \to k; \quad (\vec{v}_1, \ldots, \vec{v}_n) \mapsto \det(\vec{v}_1, \ldots, \vec{v}_n)$$

mit den Eigenschaften

(D1) $\det(\vec{v}_1, \ldots, \vec{v}_n)$ ändert ihr Vorzeichen, nicht aber ihren Betrag, wenn irgendwelche zwei ihrer Argumente miteinander vertauscht werden.

(D2) $\det(\vec{v}_1, \ldots, \vec{v}_n)$ ist linear in jedem ihrer n Argumente.

(D3) $\det(\vec{v}_1, \ldots, \vec{v}_n) = 0$ genau dann, wenn $\vec{v}_1, \ldots, \vec{v}_n$ linear abhängig sind.

Um zu zeigen, daß es eine solche Abbildung auch tatsächlich gibt, folgen wir einer auch sonst bei Existenzbeiweisen oftmals nützlichen Strategie: Wir nehmen an, wir *hätten* bereits eine Funktion det mit den Eigenschaften (D1) bis (D3), und versuchen dann, diese Funktion aufgrund dieser Eigenschaften möglichst explizit auszurechnen. Dies wird auf eine Formel für det führen, die wir dann als *Definition* nehmen können, wenn wir nachweisen, daß sie (D1) bis (D3) erfüllt.

Ausgangspunkt ist eine Basis $\{\vec{e}_1, \ldots, \vec{e}_n\}$ von V; für diese ist wegen Eigenschaft (D3) dann $\det(\vec{e}_1, \ldots, \vec{e}_n) \neq 0$. Über den genauen Wert wissen wir nichts, denn mit det erfüllt offensichtlich auch jedes skalare Vielfache $\lambda \cdot \det$ mit $\lambda \neq 0$ die Forderungen (D1) bis (D3).

Wir betrachten nun n beliebige Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ und schreiben diese bezüglich der Basis $\{\vec{e}_1, \ldots, \vec{e}_n\}$ als Linearkombinationen

$$\vec{v}_i = \sum_{j=1}^{n} a_{ij}\vec{e}_j$$

der gewählten Basisvektoren.

Sukzessive Anwendung der Linearitätsregel (D2) auf die n Argumente

von det führt auf

$$\begin{aligned}
\det(\vec{v}_1,\dots,\vec{v}_n) &= \det\Big(\sum_{j_1=1}^{n} a_{1j_1}\vec{e}_{j_1},\vec{v}_2,\dots,\vec{v}_n\Big)\\
&\underset{(D2)}{=} \sum_{j_1=1}^{n} a_{1j_1}\det(\vec{e}_{j_1},\vec{v}_2,\dots,\vec{v}_n)\\
&= \sum_{j_1=1}^{n} a_{1j_1}\det\Big(\vec{e}_{j_1},\sum_{j_2=1}^{n} a_{2j_2}\vec{e}_{j_2},\vec{v}_3,\dots,\vec{v}_n\Big)\\
&\underset{(D2)}{=} \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} a_{1j_1}a_{2j_2}\det(\vec{e}_{j_1},\vec{e}_{j_2},\vec{v}_3,\dots,\vec{v}_n)\\
&\;\;\vdots\\
&= \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\cdots\sum_{j_n=1}^{n} a_{1j_1}a_{2j_2}\cdot\dots\cdot a_{nj_n}\det(\vec{e}_{j_1},\cdots,\vec{e}_{j_n})\,.
\end{aligned}$$

Um dies noch weiter ausrechnen zu können, müssen wir die Vektoren $\vec{e}_{j_\nu}$ in jedem der Summanden in die natürliche Reihenfolge bringen; dabei hilft die Regel (D1). Eine ganze Reihe von Summanden können wir allerdings gleich von vornherein außer Acht lassen, denn nach (D3) ist

$$\det(\vec{e}_{j_1},\cdots,\vec{e}_{j_n}) = 0\,,$$

wenn (mindestens) zwei der $\vec{e}_{j_\nu}$ übereinstimmen: Falls $\vec{e}_{j_\nu}$ und $\vec{e}_{j_\mu}$ übereinstimmen, ist $\vec{e}_{j_\nu} - \vec{e}_{j_\mu} = \vec{0}$ eine nichttriviale Darstellung des Nullvektors. Ein von Null verschiedener Wert ist also nur möglich, wenn $\{j_1,\dots,j_n\} = \{1,\dots,n\}$ ist, wenn also die Abbildung

$$\pi\colon\{1,\dots,n\}\to\{1,\dots,n\};\;\; \nu\mapsto j_\nu$$

bijektiv ist.

Unsere nächste Aufgabe wird daher sein, solche Abbildungen zu studieren und sie auf Vertauschungen zweier Elemente zurückzuführen, so daß wir schließlich die Summanden mittels (D1) auf $\det(\vec{e}_1,\dots,\vec{e}_n)$ zurückführen können.

d) Gerade und ungerade Permutationen

Erinnern wir uns an §3*j4):* Dort hatten wir bijektive Abbildungen

$$\pi\colon\{1,\dots,n\}\to\{1,\dots,n\}$$

als Permutationen bezeichnet; unsere nächste Aufgabe besteht also darin für eine Permutation π nachzurechnen, für welches Vorzeichen

$$\det(\vec{e}_{\pi(1)},\dots,\vec{e}_{\pi(n)}) = \pm\det(\vec{e}_1,\dots,\vec{e}_n)$$

ist. In einem Spezialfall wissen wir das schon: Ist nämlich die Permutation π eine *Transposition,* vertauscht also genau zwei Elemente und bildet die restlichen Elemente auf sich selbst ab, so ist das Vorzeichen nach Regel (D1) ein Minuszeichen.

Der folgende Satz erlaubt es uns, den allgemeinen Fall darauf zurückzuführen:

Satz: Jede Permutation π kann als Produkt

$$\pi = \tau_1\circ\cdots\circ\tau_r$$

von Transpositionen geschrieben werden.

Beweis: Da es sich hier um eine Mathematikvorlesung für Informatiker handelt, möchte ich sowohl einen mathematischen als auch einen informatischen Beweis geben.

Der einfachste mathematische Beweis benutzt vollständige Induktion nach der Elementanzahl n der zu permutierenden Menge. Für $n = 1$ gibt es keine Permutation außer der Identität, die man als leeres Produkt von Transpositionen betrachten kann; da aber nicht jeder diese Art von logischen Taschenspielertricks liebt, betrachten wir zur Vorsicht auch noch den Fall $n = 2$ als weitere Verankerung der Induktion.

Hier gibt es genau zwei Permutationen: Die Identität, die alles festläßt, und die Vertauschung der beiden Elemente. Letzere ist eine Transposition, erstere kann wieder als leeres Produkt von Transpositionen betrachtet werden oder, falls man das nicht möchte, als Quadrat der Vertauschung, denn nach zweimaligem Vertauschen ist wieder alles beim alten.

Beim *Induktionsschritt* nehmen wir an, wir hätten den Satz für Permutationen $(n-1)$-elementiger Mengen bewiesen und betrachten nun eine Permutation

$$\pi\colon \{1,\dots,n\} \to \{1,\dots,n\}\,.$$

Für diese gibt es ein Element $i \in \{1,\dots,n\}$, so daß $\pi(i) = n$ ist, denn π ist ja bijektiv. Die Permutation

$$\pi' = \pi \circ (i\ n)$$

bildet n ab auf

$$\pi'(n) = \big(\pi \circ (i\ n)\big)(n) = \pi\big((i\ n)(n)\big) = \pi(i) = n\,,$$

d.h. n bleibt fest. Da π' bijektiv ist, muß es daher auch die Menge $\{1,\dots,n-1\}$ auf sich selbst abbilden; wenn wir n ignorieren, können wir π' daher auch als eine Permutation von $\{1,\dots,n-1\}$ auffassen. Von der wissen wir, daß sie als Produkt

$$\pi' = \tau_1 \circ \cdots \circ \tau_r$$

von Transpositionen darstellbar ist. Wegen $\pi' = \pi \circ (i\ n)$ und weil Transpositionen zu sich selbst invers sind, ist damit auch

$$\pi = \pi' \circ (i\ n) = \tau_1 \circ \cdots \circ \tau_r \circ (i\ n)$$

als Produkt von Transpositionen darstellbar, und damit ist der Satz einmal bewiesen.

Zum zweiten Beweis, auf Grundlage der Informatik, führt die Beobachtung, daß jeder gängige Sortieralgorithmus, der auf Vertauschungen von Elementen beruht (und das tut fast jeder) ein konstruktives Verfahren liefert, um Permutationen als Produkte von Transpositionen zu schreiben: Soll π zerlegt werden, so sortiere man durch fortgesetztes Vertauschungen zweier Zahlen die Folge

$$\pi(1),\ \pi(2),\ \pi(3),\ \dots,\pi(n-1),\ \pi(n)$$

der Größe nach; die erste angewandte Transposition sei τ_1, die letzte τ_n.

Die sortierte Folge der $\pi(i)$ ist natürlich $1,\dots,n$ ist; daher für jedes i

$$(\tau_r \circ \cdots \circ \tau_1 \circ \pi)(i) = i\,,$$

d.h.

$$\tau_r \circ \cdots \circ \tau_1 \circ \pi = \text{Identität}\,.$$

Multipliziert man beide Seiten von links mit $\tau_1 \circ \cdots \circ \tau_r$, so folgt

$$\pi = \tau_1 \circ \cdots \circ \tau_r\,,$$

denn

$$\begin{aligned}
&(\tau_1 \circ \cdots \circ \tau_r) \circ (\tau_r \circ \cdots \circ \tau_1)\\
&= \tau_1 \circ \cdots \circ \tau_{r-1} \circ (\tau_r \circ \tau_r) \circ \tau_{r-1} \circ \cdots \circ \tau_1\\
&= \tau_1 \circ \cdots \circ \tau_{r-2} \circ (\tau_{r-1} \circ \tau_{r-1}) \circ \tau_{r-2} \circ \cdots \circ \tau_1\\
&= \tau_1 \circ \cdots \circ \tau_{r-3} \circ (\tau_{r-2} \circ \tau_{r-2}) \circ \tau_{r-3} \cdots \circ \tau_1\\
&= \cdots = \tau_1 \circ (\tau_2 \circ \tau_2) \circ \tau_1 = \tau_1 \circ \tau_1 = \text{Identität}\,.
\end{aligned}$$

Also ist π als Produkt der Transpositionen τ_1 bis τ_r darstellbar. ∎

Ein Leser, der mit den gebräuchlichen Sortierverfahren vertraut ist, wird unschwer erkennen, daß der „mathematische“ Beweis ein Spezialfall des informatischen ist, in dem als Sortierverfahren ein relativ einfaches $O(n^2)$-Verfahren benutzt wurde (das allerdings zumindest für einstellige n den meisten „besseren“ Sortierverfahren überlegen sein dürfte). Die beiden Beweise unterscheiden sich also nicht sonderlich.

Die Anzahl n der benötigten Vertauschungen hängt natürlich von der Vorgehensweise und der Effiziens des gewählten Sortierverfahrens ab; die Darstellung von π als Produkt von Transpositionen ist also alles andere als eindeutig. Da wir für die Determinantenberechnung vor allem wissen müssen, ob die Abzahl der Vertauschungen gerade oder ungerade ist, wollen wir uns als nächstes davon überzeugen, daß wenigstens dies, d.h. also die Parität der Anzahl r der benötigten Transpositionen, eindeutig bestimmt ist: r ist für gegebenes π entweder *immer* gerade oder *immer* ungerade.

Dazu definieren wir für jede Permutation $\pi \in \mathfrak{S}_n$ ein Vorzeichen:

Definition: Das *Vorzeichen* der Permutation $\pi \in \mathfrak{S}_n$ ist die Zahl

$$\varepsilon(\pi) \underset{\text{def}}{=} \prod_{\{i,j\}\subset\{1,\dots,n\}} \frac{\pi(j)-\pi(i)}{j-i}\,,$$

wobei das Produkt über die sämtlichen zweielementigen Teilmengen von $\{1,\dots,n\}$ genommen wird.

$\varepsilon(\pi)$ ist wirklich ein Vorzeichen, das heißt gleich ± 1, denn

$$|\varepsilon(\pi)| = \frac{\prod\limits_{\{i,j\}\subset\{1,\dots,n\}} |\pi(j)-\pi(i)|}{\prod\limits_{\{i,j\}\subset\{1,\dots,n\}} |j-i|} = 1\,,$$

weil mit $\{i,j\}$ wegen der Bijektivität von π auch $\{\pi(i),\pi(j)\}$ die sämtlichen zweielementigen Teilmengen von $\{1,\dots,n\}$ durchläuft, so daß im Zähler und Nenner bis auf Reihenfolge genau dieselben Faktoren stehen.

Für eine Transposition $\tau = (k\ \ell)$ ist

$$\frac{\tau(j)-\tau(i)}{j-i} = \frac{j-i}{j-i} = 1 \quad \text{falls} \quad \{i,j\}\cap\{k,l\} = \emptyset\,,$$

da $(k\ \ell)$ alle Zahlen außer k und ℓ festläßt. Die Zahlen k und ℓ werden vertauscht, daher ist

$$\frac{\tau(k)-\tau(\ell)}{k-\ell} = \frac{\ell-k}{k-\ell} = -1\,.$$

Bleibt noch der Fall, daß eine der beiden Zahlen i oder j gleich k oder ℓ ist; da wir zweielementige *Mengen* betrachten und $\{i,j\} = \{j,i\}$ ist, können wir vereinbaren, daß diese Zahl i sein soll.

Ist $i = k$, so ist das Produkt der Terme für $\{k,j\}$ und $\{\ell,j\}$ gleich

$$\frac{\tau(k)-\tau(j)}{k-j}\cdot\frac{\tau(\ell)-\tau(j)}{\ell-j} = \frac{\ell-j}{k-j}\cdot\frac{k-j}{\ell-j} = +1\,;$$

und genau entsprechend kann man auch für $i = \ell$ argumentieren.

Bildet man daher das Produkt über alle zweielementigen Teilmengen von $\{1,\dots,n\}$, so erhält man $\varepsilon(\tau) = -1$ und damit das

Lemma: Das Vorzeichen einer Transposition ist -1. ∎

Der Zusammenhang zwischen dem Vorzeichen und der Transpositionsdarstellung einer Permutation ergibt sich aus

Lemma: Für zwei Permutationen $\pi, \pi' \in \mathfrak{S}_n$ ist

$$\varepsilon(\pi\circ\pi') = \varepsilon(\pi)\cdot\varepsilon(\pi')\,.$$

Beweis: Nach Definition ist

$$\begin{aligned}\varepsilon(\pi\circ\pi') &= \prod_{\{i,j\}\subset\{1,\dots,n\}} \frac{(\pi\circ\pi')(j)-(\pi\circ\pi')(i)}{j-i}\\ &= \prod_{\{i,j\}\subset\{1,\dots,n\}} \frac{\pi\big(\pi'(j)\big)-\pi\big(\pi'(i)\big)}{j-i}\,.\end{aligned}$$

Nach Erweiterung mit

$$\prod_{\{i,j\}\subset\{1,\dots,n\}} \frac{\pi'(j)-\pi'(i)}{\pi'(j)-\pi'(i)}$$

und Umordnung wird dies zu

$$\varepsilon(\pi\circ\pi') = \prod_{\{i,j\}} \frac{\pi\big(\pi'(j)\big)-\pi\big(\pi'(i)\big)}{\pi'(j)-\pi'(i)}\cdot\prod_{\{i,j\}}\frac{\pi'(j)-\pi'(i)}{j-i}\,.$$

Das zweite dieser Produkte ist natürlich $\varepsilon(\pi')$, und das erste ist $\varepsilon(\pi)$, da mit $\{i,j\}$ auch $\{\pi'(i),\pi'(j)\}$ die sämtlichen zweielementigen Teilmengen von $\{1,\dots,n\}$ durchläuft. ∎

Sind nun

$$\pi = \tau_1\circ\cdots\circ\tau_r = \sigma_1\circ\cdots\circ\sigma_s$$

zwei Darstellungen einer Permutation π als Produkt von Transpositionen, so ist nach den gerade bewiesenen Lemmata einerseits

$$\varepsilon(\pi) = \prod_{i=1}^{r}\varepsilon(\tau_i) = (-1)^r,$$

andererseits aber auch

$$\varepsilon(\pi) = \prod_{i=1}^{s}\varepsilon(\sigma_i) = (-1)^s,$$

also muß $(-1)^r = (-1)^s$ sein und damit r, s entweder beide gerade oder beide ungerade.

Definition: Eine Permutation $\pi \in \mathfrak{S}_n$ heißt *gerade,* wenn sie als Produkt einer geraden Anzahl von Transpositionen geschrieben werden kann, ansonsten heißt sie *ungerade.*

Das Vorzeichen $\varepsilon(\pi)$ von π ist also +1 für gerades π und -1 für ungerades.

e) Existenz von Determinanten

Nach diesem Einschub über Permutationen und Transpositionen haben wir das nötige Rüstzeug, um die Rechnung aus Abschnitt *c)* zu Ende zu führen:

Da die Determinante bei jeder Vertauschung zweier Elemente ihr Vorzeichen wechselt, folgt sofort aus den obigen Betrachtungen, daß für jede Permutation $\pi \in \mathfrak{S}_n$ gilt

$$\det\bigl(\vec{e}_{\pi(1)}, \ldots, \vec{e}_{\pi(n)}\bigr) = \begin{cases} +\det(\vec{e}_1, \ldots, \vec{e}_n) & \text{falls } \pi \text{ gerade} \\ -\det(\vec{e}_1, \ldots, \vec{e}_n) & \text{falls } \pi \text{ ungerade} \end{cases},$$

d.h. $\det\bigl(\vec{e}_{\pi(1)}, \ldots, \vec{e}_{\pi(n)}\bigr) = \varepsilon(\pi) \cdot \det(\vec{e}_1, \ldots, \vec{e}_n)$. Wenn wir die Rechnung aus Abschnitt *b)* fortführen, erhalten wir daher

$$\begin{aligned} \det\bigl(\vec{v}_1, \ldots, \vec{v}_n\bigr) &= \sum_{\pi \in \mathfrak{S}_n} a_{1\pi(1)} \cdots a_{n\pi(n)} \cdot \det\bigl(\vec{e}_{\pi(1)}, \ldots, \vec{e}_{\pi(n)}\bigr) \\ &= \left(\sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)} \right) \cdot \det\bigl(\vec{e}_1, \ldots, \vec{e}_n\bigr) \,. \end{aligned}$$

Somit ist $\det\bigl(\vec{v}_1, \ldots, \vec{v}_n\bigr)$ eindeutig bestimmt bis auf den Wert von $\det\bigl(\vec{e}_1, \ldots, \vec{e}_n\bigr)$; falls es überhaupt Determinanten gibt, müssen sie also so aussehen. Wir definieren daher

Definition: Die Determinante von n Vektoren

$$\vec{v}_1 = a_{11}\vec{e}_1 + \cdots + a_{1n}\vec{e}_n, \quad \ldots, \quad \vec{v}_n = a_{n1}\vec{e}_1 + \cdots + a_{nn}\vec{e}_n$$

des n-dimensionalen Vektorraums V bezüglich der Basis $\vec{e}_1, \ldots, \vec{e}_n$ ist

$$\det\bigl(\vec{v}_1, \ldots, \vec{v}_n\bigr) = \sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)} \,.$$

Insbesondere ist also $\det(\vec{e}_1, \ldots, \vec{e}_n) = 1$.

Satz: Die so definierte Determinante hat die Eigenschaften (D1) bis (D3).

Beweis: Beginnen wir mit (D1), d.h. die Determinante wechselt ihr Vorzeichen, wenn zwei der Argumente vertauscht werden. Vertauscht man etwa die Argumente $\vec{v}_i$ und $\vec{v}_j$, so werden anstelle der Summanden

$$\varepsilon(\pi)\, a_{1\pi(1)} \cdots a_{i\pi(i)} \cdots a_{j\pi(j)} \cdots a_{n\pi(n)}$$

die Summanden

$$\varepsilon(\pi)\, b_{1\pi(1)} \cdots b_{j\pi(j)} \cdots b_{i\pi(i)} \cdots b_{n\pi(n)}$$

aufaddiert, wobei

$$b_{k\ell} = \begin{cases} a_{k\ell} & \text{falls } k \neq i, j \\ a_{j\ell} & \text{falls } k = i \\ a_{i\ell} & \text{falls } k = j \end{cases}$$

ist. Die Summanden sind also

$$\begin{aligned} &\varepsilon(\pi)\, a_{1\pi(1)} \cdots a_{j\pi(i)} \cdots a_{i\pi(j)} \cdots a_{n\pi(n)} \\ &= \varepsilon(\pi)\, a_{1\pi(1)} \cdots a_{i\pi(j)} \cdots a_{j\pi(i)} \cdots a_{n\pi(n)} \\ &= \varepsilon(\pi)\, a_{1\pi'(1)} \cdots a_{i\pi'(i)} \cdots a_{j\pi'(j)} \cdots a_{n\pi'(n)} \end{aligned}$$

mit

$$\pi'(k) = \begin{cases} \pi(k) & \text{falls } k \neq i, j \\ \pi(j) & \text{falls } k = i \\ \pi(i) & \text{falls } k = j \,, \end{cases}$$

d.h. $\pi' = \pi \circ (i\ j)$. Als Produkt von Transpositionen geschrieben hat π' daher einen Faktor mehr als π; falls π ungerade war, ist also π' gerade, und umgekehrt. Somit ist

$$\varepsilon(\pi') = -\varepsilon(\pi) \,,$$

und damit lassen sich die Summanden auch schreiben als

$$-\varepsilon(\pi')\, a_{1\pi'(1)} \cdots a_{i\pi'(i)} \cdots a_{j\pi'(j)} \cdots a_{n\pi'(n)} .$$

Summiert man über alle $\pi' \in \mathfrak{S}_n$, ist also

$$\begin{aligned} &\det(\vec{v}_1, \ldots, \vec{v}_j, \ldots, \vec{v}_i, \ldots, \vec{v}_n) \\ &= \sum_{\pi' \in \mathfrak{S}_n} \big(-\varepsilon(\pi')\big)\, a_{1\pi'(1)} \cdots a_{i\pi'(i)} \cdots a_{j\pi'(j)} \cdots a_{n\pi'(n)} \\ &= -\sum_{\pi' \in \mathfrak{S}_n} \varepsilon(\pi')\, a_{1\pi'(1)} \cdots a_{i\pi'(i)} \cdots a_{j\pi'(j)} \cdots a_{n\pi'(n)} \\ &= -\det(\vec{v}_1, \ldots, \vec{v}_i, \ldots, \vec{v}_j, \ldots, \vec{v}_n) , \end{aligned}$$

womit (D1) bewiesen wäre.

Eigenschaft (D2), die Linearität in jedem der Argumente, ist klar, denn jedes der Monome $a_{1\pi(1)} \cdots a_{n\pi(n)}$ enthält *genau eine* Komponente des Vektors $\vec{v}_i$ und ist damit linear in $\vec{v}_i$.

Bleibt noch (D3), d.h. $\det\big(\vec{v}_1, \ldots, \vec{v}_n\big)$ soll genau dann verschwinden, wenn die $\vec{v}_i$ linear abhängig sind.

Sind zunächst die Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ linear abhängig, so läßt sich einer von ihnen als Linearkombination der anderen schreiben; durch Umordnen (wobei die Determinante nach der schon bewiesenen Eigenschaft (D1) höchstens ihr Vorzeichen ändert) können wir annehmen, daß dies etwa $\vec{v}_1$ sei; wir setzen

$$\vec{v}_1 = \sum_{i=2}^{n} \lambda_i \vec{v}_i .$$

Dann ist

$$\begin{aligned} \det\big(\vec{v}_1, \ldots, \vec{v}_n\big) &= \det\left(\sum_{i=2}^{n} \lambda_i \vec{v}_i, \vec{v}_2, \ldots, \vec{v}_n\right) \\ &\underset{(D2)}{=} \sum_{i=2}^{n} \lambda_i \det\big(\vec{v}_i, \vec{v}_2, \ldots, \vec{v}_n\big) . \end{aligned}$$

Im i-ten Summanden der letzten Summe tritt der Vektor $\vec{v}_i$ zweimal als Argument von det auf: Einmal an der ersten, und dann noch an der i-ten.

Vertauscht man diese beiden (gleichen) Argumente, ändert sich natürlich nichts; andererseits haben wir aber bereits (D1) bewiesen, wonach sich beim Vertauschen irgend zweier Argumente das Vorzeichen ändert. Der i-te Summand hat also einen Wert s_i, für den $s_i = -s_i$ ist. Falls wir mit reellen Zahlen arbeiten, folgt hieraus natürlich sofort, daß alle s_i verschwinden und somit auch ihre Summe.

Für beliebige Körper gilt das aber leider nicht: Aus $s_i = -s_i$ folgt, wenn wir auf beiden Seiten s_i addieren, zunächst nur, daß

$$s_i + s_i = (1+1)s_i = 0$$

ist, und daraus folgt genau dann, daß auch $s_i = 0$ ist, wenn $1 + 1 \neq 0$ ist. Dies ist aber beispielweise in dem für die digitale Informationsverarbeitung wichtigen Körper $\mathbb{F}_2 = \{0, 1\}$ nicht der Fall; wir müssen uns also mit Rücksicht auf diesen (und eine ganze Reihe anderer) Körper noch überlegen, daß wirklich immer

$$\det(\vec{v}_1, \ldots, \vec{v}_n) = 0$$

ist, wenn zwei der Argumente $\vec{v}_i$ übereinstimmen.

Dazu beachten wir, daß im Falle $\vec{v}_i = \vec{v}_j$ auch für alle ℓ die Koeffizienten $a_{i\ell}$ und $a_{j\ell}$ übereinstimmen, d.h. in der Summe

$$\det A = \sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{i\pi(i)} \cdots a_{j\pi(j)} \cdots a_{n\pi(n)}$$

ist stets

$$a_{1\pi(1)} \cdots a_{i\pi(i)} \cdots a_{j\pi(j)} \cdots a_{n\pi(n)} = a_{1\pi(1)} \cdots a_{i\pi(j)} \cdots a_{j\pi(i)} \cdots a_{n\pi(n)} .$$

Hier ist die rechte Seite gerade der Summand zur Permutation $\pi \circ (i\ j)$, die genau dann gerade ist, wenn π ungerade ist, und umgekehrt. Also haben beide Terme verschiedenes Vorzeichen und heben sich gegenseitig weg; die Summe über alle Permutationen ist daher Null.

Somit verschwindet die Determinante unabhängig vom Grundkörper immer dann, wenn die Vektoren linear abhängig sind.

Nun seien $\vec{v}_1, \ldots, \vec{v}_n$ linear unabhängig; für dieses Fall müssen wir zeigen, daß die Determinante *nicht* verschwindet. Wir wissen, daß für die Basisvektoren $\vec{e}_1, \ldots, \vec{e}_n$

$$\det\big(\vec{e}_1, \ldots, \vec{e}_n\big) = 1 \neq 0$$

ist; außerdem wissen wir, daß n beliebige linear unabhängige Vektoren eines n-dimensionalen Vektorraums eine Basis bilden.

Daher können wir die $\vec{e}_i$ als Linearkombinationen der $\vec{v}_j$ schreiben, etwa als

$$\vec{e}_i = \sum_{j=1}^{n} b_{ij}\vec{v}_j \; .$$

Nun können wir die Rechnungen der letzten Paragraphen wörtlich wiederholen, nur daß dieses Mal $\vec{v}_1, \ldots, \vec{v}_n$ die Basisvektoren sind und $\vec{e}_1, \ldots, \vec{e}_n$ die Vektoren, für die wir det berechnen möchten. Wir wissen zwar noch nicht, daß $\det(\vec{v}_1, \ldots, \vec{v}_n)$ nicht verschwindet, aber das haben wir in der dortigen Rechnung auch nie verwendet: Alles hing nur ab von den bereits bewiesenen Forderungen (D1) und (D2) und der ebenfalls schon bewiesenen Tatsache, daß det verschwindet, wenn zwei der Argumente gleich sind. Somit erhalten wir die Formel

$$\det(\vec{e}_1, \ldots, \vec{e}_n) = \left(\sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) b_{1\pi(1)} \cdots b_{n\pi(n)} \right) \cdot \det(\vec{v}_1, \ldots, \vec{v}_n) \; .$$

In dieser Gleichung steht links die Zahl eins, also kann keiner der beiden Faktoren auf der rechten Seite verschwinden; insbesondere ist $\det(\vec{v}_1, \ldots, \vec{v}_n) \neq 0$.

Damit ist der Satz vollständig bewiesen. ∎

Fassen wir zusammen, was wir bislang in diesem Paragraphen gemacht haben:

- Wir haben anhand des Beispiels von Determinanten im $\mathbb{R}^2$ und $\mathbb{R}^3$ drei zentrale Forderungen an eine allgemeine Determinante aufgestellt.
- Unter der Annahme, daß es Funktionen gibt, die diese drei Forderungen erfüllen, haben wir ausgerechnet, wie solche Funktionen aussehen müssen; inbsbesondere haben wir gesehen, daß sie durch ihren Wert auf den Vektoren einer Basis eindeutig bestimmt sind.
- Dann haben wir die so erhaltene explizite Formel als *Definition* einer allgemeinen Determinanten genommen und gezeigt, daß die so definierte Funktion tatsächlich die drei Forderungen erfüllt.

Insgesamt haben wir also gesehen, daß es Funktionen gibt, die den drei Forderungen (D1) bis (D3) genügen und daß solche Funktionen bis auf eine multiplikative Konstante eindeutig bestimmt sind.

f) Die Determinante einer Matrix

Definition: Unter der Determinante einer Matrix $A \in k^{n\times n}$ verstehen wir die Determinante ihrer Spaltenvektoren; ist also

$$A = (a_{ij}) \quad \text{und} \quad \vec{v}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix} \; ,$$

so ist

$$\det A \underset{\text{def}}{=} \det(\vec{v}_1, \ldots, \vec{v}_n) \; .$$

Wir schreiben auch kurz

$$\det A = |A| = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} \; .$$

Um etwas Übung im Umgang mit Determinanten zu bekommen, wollen wir dies für $n = 1$ und die beiden bereits bekannten Fälle $n = 2$ und $n = 3$ noch einmal explizit ausrechnen – auch wenn wir bereits wissen, was herauskommen muß.

Der Fall $n = 1$ einer 1×1-Matrix ist völlig uninteressant: Die einzige Permutation einer einelementigen Menge ist die Identität, die für jedes n gerade ist; für $A = (a) \in k^{1\times 1}$ ist also $\det A = a$. Die Schreibweise $\det A = |a|$ ist hier irreführend, da es zumindest für $k = \mathbb{R}$ oder $\mathbb{C}$ zu Verwechslungen mit der Betragsfunktion kommen kann. Da Determinanten von 1×1-Matrizen völlig uninteressant sind, ist dies jedoch nicht weiter schlimm.

Für $n = 2$ gibt es zwei Permutationen: Die (gerade) Identität und die (ungerade) Transposition (1 2). Nach Definition der (allgemeinen) Determinanten ist

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \; ,$$

wie erwartet.

Für $n = 3$ gibt es bereits $3! = 6$ Permutationen, darunter natürlich die (gerade) Identität.

Jede weitere Permutation kann höchstens ein Element festlassen, denn würde sie zwei festlassen, müßte wegen der Bijektivität der Abbildung auch das dritte auf sich selbst abgebildet werden, wir hätten also die Identität.

Falls ein Element festgehalten wird, bleibt einer nichtidentischen Permutation daher nichts anderes übrig, als die beiden anderen zu vertauschen; dies führt auf die drei (ungeraden) Transpositionen (2 3), (1 3) und (1 2).

Falls kein Element auf sich selbst abgebildet wird, muß die Eins entweder auf 2 oder auf 3 abgebildet werden. Im ersten Fall geht dann 2 auf 3, denn sonst müßte zwei entweder festbleiben oder wir hätten die Transposition (1 2), die drei auf sich selbst abbildet. Wegen der Bijektivität der Abbildung gibt es dann keine andere Möglichkeit als daß drei auf eins abgebildet wird; wir haben also die zyklische Vertauschung

$$1 \mapsto 2 \mapsto 3 \mapsto 1 \,.$$

Diese Permutation ist gerade, denn sie läßt sich beispielsweise als Produkt (1 3)(1 2) zweier Transpositionen schreiben: Wenden wir dieses Produkt an auf 1, so wird die Eins zunächst von (1 2) auf zwei abgebildet, und zwei bleibt fest unter (1 3), so daß insgesamt eins auf zwei abgebildet wird, wie verlangt.

Zwei wird von (1 2) auf die Eins abgebildet, und die wiederum von (1 3) auf drei, also haben wir auch hier insgesamt das richtige Ergebnis. Drei schließlich kann dann wegen der Bijektivität von (1 3)(1 2) nur auf eins abgebildet werden, was man auch schnell direkt sieht, denn (1 2) läßt die Drei unverändert, während (1 3) sie auf eins abbildet.

Die noch verbleibende Permutation bildet eins auf drei und daher drei auf zwei ab, sie ist also die zyklische Vertauschung

$$1 \mapsto 3 \mapsto 2 \mapsto 1 \quad \text{oder} \quad 3 \mapsto 2 \mapsto 1 \mapsto 3 \,,$$

d.h. die Umkehrabbildung zur gerade betrachteten Permutation. Daher ist sie als Produkt von Transpositionen gleich

$$\big((1\ 2)(1\ 3)\big)^{-1} = (1\ 3)^{-1}(1\ 2)^{-1} = (1\ 3)(1\ 2)$$

und somit auch gerade.

Zur Berechnung der Determinanten einer 3×3-Matrix (a_{ij}) beginnen wir mit den geraden Permutationen: Die Identität liefert den Term $a_{11}a_{22}a_{33}$, d.h. das Produkt der drei Diagonalelemente, die zyklische Vertauschung $1 \to 2 \to 3 \to 1$ liefert das Produkt $a_{12}a_{23}a_{31}$, und ihr Inverses $a_{13}a_{21}a_{32}$.

Die drei Transpositionen führen auf die negativ zu nehmenden Produkte $a_{11}a_{23}a_{32}$, $a_{13}a_{22}a_{31}$ und $a_{12}a_{21}a_{33}$, insgesamt ist also

$$\begin{aligned} \det A &= \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \\ &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\qquad - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} \,, \end{aligned}$$

genau wie wir es in Abschnitt *b)* über das Spatprodukt berechnet hatten und uns aufgrund der SARRUSschen Regel leicht merken können.

Für die Determinanten von 4×4-Matrizen gibt es keine entsprechende Regel mehr. Da nun bereits $4! = 24$ Summanden berücksichtigt werden müssen, eignet sich die Formel aus der Definition der Determinanten nicht mehr gut zur Berechnung, von noch größeren Matrizen ganz zu schweigen: Wie wir im nächsten Kapitel sehen werden, ist $\log n! \approx n \log n$, d.h. die Anzahl $n!$ der Summanden wächst mit größer werdendem n noch schneller als die Exponentialfunktion.

Wir werden im folgenden daher auf Verfahren hinarbeiten, die es erlauben, auch größere Determinanten mit vertretbarem Aufwand zu berechnen.

Wir kennen schon einige Rechenregeln für Determinanten als Funktionen von n Vektoren, beispielsweise die Forderungen (D1) bis (D3), aus

denen wir die Determinantendefinition hergeleitet haben oder auch die gerade angewandte Regel

$$\begin{aligned}&\det\bigl(\vec{v}_1,\ldots,\vec{v}_i+\lambda\vec{v}_j,\ldots,\vec{v}_n\bigr)\\ =&\det\bigl(\vec{v}_1,\ldots,\vec{v}_i,\ldots,\vec{v}_n\bigr)+\lambda\det\bigl(\vec{v}_1,\ldots,\vec{v}_j,\ldots,\vec{v}_n\bigr)\\ =&\det\bigl(\vec{v}_1,\ldots,\vec{v}_i,\ldots,\vec{v}_n\bigr)\,.\end{aligned}$$

In diesem Abschnitt soll es nun um Rechenregeln für Determinanten von Matritzen gehen.

Am wichtigsten ist der *Multiplikationssatz:*

Satz: *a)* Für zwei Matrizen $A, B \in k^{n\times n}$ ist $\det(AB) = \det A \cdot \det B$ *b)* Für eine invertierbare Matrix $A \in k^{n\times n}$ ist $\det(A^{-1}) = (\det A)^{-1}$.

Beweis: a) Die Determinante von AB ist die Determinante der Spaltenvektoren $\vec{v}_1,\ldots,\vec{v}_n$ von AB, die von B ist gleich der Determinante der Spaltenvektoren $\vec{b}_1,\ldots,\vec{b}_n$ von B, und nach Definition der Matrizenmultiplikation ist mit $A = (a_{ij})$

$$\vec{v}_i = \sum_{j=1}^{n} a_{ij}\vec{b}_j\,.$$

Damit sind wir in der Situation der Abschnitte *b)* und *d)* (wobei hier die $\vec{b}_j$ die Rolle der dortigen $\vec{e}_j$ übernehmen) und können aus den Eigenschaften (D1) bis (D3) folgern, daß

$$\det\bigl(\vec{v}_1,\ldots,\vec{v}_n\bigr) = \left(\sum_{\pi\in\mathfrak{S}_n}\varepsilon(\pi)a_{1\pi(1)}\cdots a_{n\pi(n)}\right)\cdot\det\bigl(\vec{b}_1,\ldots,\vec{b}_n\bigr)$$

ist. Das ist aber genau die Formel

$$\det(AB) = (\det A)\cdot(\det B)\,.$$

b) Für eine invertierbare Matrix A ist $A\cdot A^{-1} = E$ die Einheitsmatrix, und deren Determinante ist gleich der Determinante der Basisvektoren in ihrer natürlichen Reihenfolge, also 1. Daher ist nach Teil *a)*

$$\det A\cdot\det\bigl(A^{-1}\bigr) = \det E = 1\,,$$

woraus die Behauptung folgt. ∎

Der Multiplikationssatz läßt sich in ein Berechnungsverfahren für Determinanten übersetzen: Falls wir die LR-Zerlegung $A = LR$ der Matrix A kennen, sagt uns der Satz, daß

$$\det A = \det L\cdot\det R$$

ist. Die Determinanten von L und R lassen sich leicht ausrechnen nach dem folgenden

Lemma: Die Determinante einer (unteren oder oberen) Dreiecksmatrix ist gleich dem Produkt ihrer Diagonalelemente.

Beweis: $A = (a_{ij})$ sei eine Dreiecksmatrix. Wie üblich ist

$$\det A = \sum_{\pi\in\mathfrak{S}_n}\varepsilon(\pi)a_{1\pi(1)}\cdots a_{n\pi(n)}\,.$$

Nun ist aber für eine untere *bzw.* obere Dreiecksmatrix $a_{ij} = 0$ falls $i < j$ *bzw.* $i > j$ ist. Da Permutationen bijektive Abbildungen sind, muß es für jede Permutation außer der Identität mindestens ein i geben mit $\pi(i) < i$ und mindestens ein j mit $\pi(j) > j$, d.h. das Produkt $a_{1\pi(1)}\cdots a_{n\pi(n)}$ enthält für jede nichtidentische Permutation mindestens einen Faktor Null. Damit kann höchstens die Identität einen von Null verschiedenen Beitrag zur Summe liefern; da sie Vorzeichen +1 hat ist also

$$\det A = a_{11}\cdots a_{nn}$$

das Produkt der Diagonalelemente von A. ∎

So wie wir die LR-Zerlegung definiert haben, ist L eine untere Dreiecksmatrix mit lauter Einsen in der Hauptdiagonale und R eine beliebige obere Dreiecksmatrix. Also ist $\det L = 1$, die Determinante von R ist gleich dem Produkt der Diagonalelemente von R, und damit ist auch

$$\det A = \text{Produkt der Diagonalelemente von } R\,.$$

Zumindest für große n ermöglicht dies eine erheblich effizientere Berechnung von Determinanten als die definierende Formel: Der Aufwand

für eine LR-Zerlegung ist ungefähr proportional zu n^3, was für große n weitaus günstiger ist als der überexponentiell steigende Aufwand proportional $n \cdot n!$ für die Summe aus der definierenden Formel. Für die Praxis wichtig ist, daß wir die Matrix L nicht wirklich kennen müssen: Ihre Determinante ist auf jeden Fall eins. Daher reicht es, den Algorithmus für die LR-Zerlegung nur auf die Matrix A selbst anzuwenden; wir müssen sie nicht auch noch die Einheitsmatrix mitschleppen. Im wesentlichen geht es also einfach um GAUSS-Elimination. Dies wird auch klar anhand der folgenden Regeln, die man üblicherweise an Stelle der LR-Zerlegung anwendet:

Eine Determinante ändert ihren Wert nicht, wenn man ein Vielfaches einer Spalte zu einer anderen Spalte addiert: Die Determinante einer Matrix ist nach Definition die Determinante der Spaltenvektoren, und wenn man das λ-fache der j-ten Spalte zur i-ten Spalte addiert, ist

$$\begin{aligned}&\det\bigl(\vec{v}_1,\ldots,\vec{v}_i+\lambda\vec{v}_j,\ldots,\vec{v}_n\bigr)\\ =&\det\bigl(\vec{v}_1,\ldots,\vec{v}_i,\ldots,\vec{v}_n\bigr)+\lambda\det\bigl(\vec{v}_1,\ldots,\vec{v}_j,\ldots,\vec{v}_n\bigr)\\ =&\det\bigl(\vec{v}_1,\ldots,\vec{v}_i,\ldots,\vec{v}_n\bigr)\,,\end{aligned}$$

da der zweite Summand den Vektor $\vec{v}_j$ doppelt enthält und somit verschwindet.

Etwas entsprechendes gilt auch für Zeilen; das könnten wir zwar direkt nachrechnen, aber wenn wir uns daran erinnern, daß die Zeilen einer Matrix gleich den Spalten der transponierten Matrix sind, folgt das viel bequemer aus dem folgenden

Lemma: Für jede $n \times n$-Matrix A ist $\det A = \det {}^tA$.

Beweis: Nach Definition ist für $A = (a_{ij})$

$$\det A = \sum_{\pi\in\mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)}\cdots a_{n\pi(n)}$$

und

$$\det {}^tA = \sum_{\pi\in\mathfrak{S}_n} \varepsilon(\pi) a_{\pi(1)1}\cdots a_{\pi(n)n}\,.$$

Ordnet man die Faktoren des Produkts $a_{\pi(1)1}\cdots a_{\pi(n)n}$ nach dem ersten Index, so erhält man

$$a_{\pi(1)1}\cdots a_{\pi(n)n} = a_{1\pi^{-1}(1)}\cdots a_{1\pi^{-1}(n)}$$

und damit

$$\det {}^tA = \sum_{\pi\in\mathfrak{S}_n} \varepsilon(\pi) a_{1\pi^{-1}(1)}\cdots a_{1\pi^{-1}(n)}\,.$$

Da mit π auch π^{-1} die gesamte Gruppe $\mathfrak{S}_n$ durchläuft und $\varepsilon(\pi) = \varepsilon(\pi^{-1})$ ist, stimmt die Summe dieser Terme mit der aus der Formel für $\det A$ überein, die beiden Determinanten sind also gleich. ∎

Für Leser, denen nicht klar ist, warum $\varepsilon(\pi)$ und $\varepsilon(\pi^{-1})$ gleich sind, seien zur Übung des Umgangs mit Permutationen hier kurz zwei Beweise angegeben: Am einfachsten sieht man diese Formel aufgrund des obigen Lemmas, wonach für zwei beliebige Permutationen π und π' gilt: $\varepsilon(\pi\circ\pi') = \varepsilon(\pi)\cdot\varepsilon(\pi')$. Setzt man hier $\pi' = \pi^{-1}$, so folgt

$$1 = \varepsilon(\text{Identität}) = \varepsilon(\pi)\cdot\varepsilon(\pi^{-1})\,,$$

so daß die beiden Vorzeichen gleich sein müssen.

Alternativ können wir uns auch explizit überlegen, daß sich π^{-1} als Produkt von r Transpositionen schreiben läßt, falls dies für π der Fall ist; genauer gilt: Ist $\pi = \tau_1\circ\cdots\circ\tau_r$, so ist $\pi^{-1} = \tau_r\circ\cdots\circ\tau_1$, denn wie wir bereits im Abschnitt über Permutationen gesehen haben, ist

$$\pi\circ(\tau_r\circ\cdots\circ\tau_1) = (\tau_1\circ\cdots\circ\tau_r)\circ(\tau_r\circ\cdots\circ\tau_1) = \text{Identität}\,,$$

da jede Transposition zu sich selbst invers ist.

Aus obigem Lemma folgt insbesondere, daß man die Determinante einer Matrix auch als Determinante der Zeilenvektoren definieren kann und daß die Spaltenvektoren einer Matrix genau dann linear unabhängig sind, wenn es die Zeilenvektoren sind. Mit nur wenig mehr Aufwand könnte man so auch zeigen, daß der im Zusammenhang mit der Lösung linearer Gleichungssysteme definierte *Rang* einer Matrix genausogut über Zeilen- wie über Spaltenoperationen berechnet werden kann.

Eine andere Anwendung ist das folgende

Korollar: *a)* Die Determinante einer orthogonalen Matrix ist ± 1.
b) Die Determinante einer unitären Matrix hat Betrag eins.

Beweis: a) Für eine orthogonale Matrix Q ist ${}^tQQ = E$, also nach dem Multiplikationssatz $\det {}^tQ \cdot \det Q = \det E = 1$. Da nach Lemma $\det {}^tQ = \det Q$ ist, folgt die Behauptung.

b) Im unitären Fall ist ${}^tQ\overline{Q} = E$; da Determinten via Grundrechenarten aus den Einträgen einer Matrix bestimmt werden, ist dabei offensichtlich $\det \overline{Q} = \overline{\det Q}$. Damit folgt, daß $\det Q \cdot \overline{\det Q} = |\det Q|^2 = 1$ ist, also auch $|\det Q| = 1$.

∎

Es gibt also zwei Arten von orthogonalen Matrizen: Solche mit Determinante eins und solche mit Determinante minus eins. Im Zweidimensionalen sieht man leicht, daß die orthogonalen Matrizen mit Determinante eins Drehungen beschreiben, während die mit Determinante minus eins zusätzlich noch eine Spiegelung enthalten. Allgemein beschreiben orthogonale Matrizen genau dann eine orientierungserhaltende Abbildung, wenn sie Determinante eins haben; dies kann man entweder als Definition des Begriffs „orientierungserhaltend“ ansehen, oder aber man kann Lehrbücher der linearen Algebra konsultieren, in denen orientierte Vektorräume definiert werden durch Verallgemeinerung der hier betrachteten Rechts- und Linkssysteme im $\mathbb{R}^3$ und die betreffende Aussage dann leicht aus den Rechenregeln für Determinanten folgt.

Wenden wir das Korollar an auf die QR-Zerlegung $A = QR$ einer reellen $n \times n$-Matrix A, ist $\det A = \det Q \det R = \pm \det R$, wobei die Determinante der Dreiecksmatrix R wieder einfach das Produkt der Diagonaleinträge ist. Falls es auf das Vorzeichen nicht ankommt oder wir auf einfache Weise entscheiden können, onb die Matrix Q orientierungserhaltend ist oder nicht, läßt sich auch das zu einer effizienten Berechnung der Determinante verwenden, die im allgemeinen numerisch stabiler ist als die Berechnung via LR-Zerlegng.

Die QR-Zerlegung zeigt uns auch, warum Determinanten im Reellen für beliebige Dimensionen als Volumina aufgefaßt werden können: Wir betrachten ein Parallelepiped, daß im $\mathbb{R}^n$ von den n-Vektoren $\vec{a}_1, \ldots, \vec{a}_n$ aufgespannt werden, und bilden dazu die QR-Zerlegung der

Matrix $A \in \mathbb{R}^{n \times n}$ mit Spalten $\vec{a}_1, \ldots, \vec{a}_n$. Da die orthogonale Matrix Q, wie auch ihr Inverses, Skalarprodukte respektiert, ändert die Anwendung von Q^{-1} nichts am Volumen, das Parallelepiped hat also dasselbe Volumen wie das von den Spalten der Dreiecksmatrix R aufgespannte. Daß wir dieses durch Scherung zu einem Quader gleichen Volumens machen können, entspricht algebraisch der Tatsache, daß die Dreiecksmatrix R dieselbe Determinante hat wie die Matrix, die dieselben Diagonaleinträge hat wie R und ansonsten nur Nullen. Die Spalten dieser Diagonalmatrix spannen einen Quader auf, dessen unorientiertes Volumen gleich dem Produkt der Seitenlängen, also dem Produkt der Beträge der Diagonaleinträge ist. Falls man sich für das orientierte Volumen interessiert, muß man das Produkt der Diagonaleinträge von R mit der Determinante ± 1 von Q multiplizieren.

Zurück zu den Rechenregeln für Determinanten! Aus dem gerade bewiesenen Lemma folgt sofort:

Eine Determinante ändert ihren Wert nicht, wenn man ein Vielfaches einer Zeile zu einer anderen Zeile addiert, denn die Zeilen einer Matrix sind die Spalten der transponierten Matrix.

Ebenso zeigt man: *Eine Determinante wird mit -1 multipliziert, wenn man entweder zwei Zeilen oder zwei Spalten miteinander vertauscht,* denn für Spalten ist das gerade die definierende Eigenschaft (D1) der Determinanten.

Als letzte Regel sei noch aufgeführt: *Eine Determinante wird mit λ multipliziert, wenn* eine *ihrer Zeilen oder* eine *ihrer Spalten mit λ multipliziert werden.* Für Spalten folgt das sofort aus der Linearität in jedem Argument, für Zeilen folgt es daraus nach Transponieren.

g) Der Entwicklungssatz von Laplace

Hier geht es um ein Berechnungsverfahren, das die definierenden Eigenschaften der Determinante ausnutzt, die „Entwicklung nach einer Zeile oder Spalte“.

Für die Entwicklung einer Matrix beispielsweise nach der j-ten Spalte schreiben wir den j-ten Spaltenvektor

$$\vec{v}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}$$

als Linearkombination der Einheitsvektoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots \quad \vec{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

d.h. als

$$\vec{v}_j = a_{1j}\vec{e}_1 + \cdots + a_{nj}\vec{e}_n \,.$$

Dann ist wegen der Linearität der Determinante in ihrem j-ten Argument

$$\begin{aligned} \det A &= \det\bigl(\vec{v}_1, \dots, \vec{v}_j, \dots, \vec{v}_n\bigr) = \det\bigl(\vec{v}_1, \dots, \sum_{i=1}^{n} a_{ij}\vec{e}_i, \dots, \vec{v}_n\bigr) \\ &= \sum_{i=1}^{n} a_{ij} \det\bigl(\vec{v}_1, \dots, \vec{e}_i, \dots, \vec{v}_n\bigr) \,. \end{aligned}$$

Um daraus die Determinante von A zu berechnen, müssen wir die Determinanten

$$D_{ij} \underset{\text{def}}{=} \det\bigl(\vec{v}_1, \dots, \vec{e}_i, \dots, \vec{v}_n\bigr)$$

kennen, wobei $\vec{e}_i$ an der j-ten Stelle steht.

Dazu wenden wir die definierende Formel an: In den Produkten $a_{1\pi(1)} \cdots a_{n\pi(n)}$ ist für die gewählte Stelle j, an der wir den Vektor $\vec{e}_i$ eingefügt haben, $a_{j\ell} = 0$ für $\ell \neq i$ und $a_{ji} = 1$. Also verschwindet $a_{1\pi(1)} \cdots a_{n\pi(n)}$ für jede Permutation π mit $\pi(j) \neq i$.

Am einfachsten läßt sich dies im Fall $i = j = n$ weiter ausrechnen: Dann müssen wir nur Permutationen mit $\pi(n) = n$ betrachten, und die entsprechen eineindeutig den Permutationen auf der Menge aller Zahlen von 1 bis $n-1$; was übrigbleibt ist also (wegen $a_{nn} = 1$) gerade die Determinante jener $(n-1) \times (n-1)$-Matrix, die durch Streichung der letzten Spalte und der letzten Zeile aus A entsteht.

Der allgemeine Fall läßt sich durch Vertauschungen darauf zurückführen: Wir können zwar nicht einfach die j-te Spalte mit der n-ten vertauschen, denn dann steht ja die n-te Spalte an j-ter Stelle, aber wir können die j-te Spalte durch sukzessive Vertauschung mit ihrem rechten Nachbarn immer weiter nach außen bringen: Wir vertauschen zunächst die j-te Spalte mit der $(j+1)$-ten, dann die neue $(j+1)$-te (=alte j-te) Spalte mit der $(j+2)$-ten *usw.,* bis schließlich die alte j-te Spalte an n-ter Stelle steht, ohne daß sich vor oder nach der Stelle j irgendetwas an der relativen Reihenfolge der Spalten geändert hätte. Die Anzahl der dazu notwendigen Vertauschungen ist $(n-j)$, die Determinante wurde bei der gesamten Prozedur also mit $(-1)^{n-j}$ multipliziert.

Als nächstes müssen wir auch noch die i-te Zeile nach unten bringen; dazu verwenden wir die gerade gezeigte Formel $\det A = \det {}^tA$. Die i-te Zeile von A ist gleich der i-ten Spalte von ${}^tA$; wie wir oben gesehen haben, läßt sie sich durch $(n-i)$ Spaltenvertauschungen zur letzten Spalte machen, wobei die Determinante mit $(-1)^{n-i}$ multipliziert wird, und durch nochmaliges Transponieren erhalten wir schließlich jene Matrix, in der die i-te Zeile zur n-ten geworden ist, ohne daß sich vor oder nach der Stelle i irgendetwas an der relativen Reihenfolge der Zeilen geändert hätte.

Die Determinante der so entstehenden Matrix ist, wie wir oben beim Spezialfall $i = j = n$ gesehen haben, gleich der Determinante jener $(n-1) \times (n-1)$-Matrix A_{ij}, die durch Streichen der letzten Zeile und Spalte entsteht; bezogen auf die ursprüngliche Matrix A entsteht sie durch Streichen der i-ten Zeile und der j-ten Spalte. Also ist, wenn wir noch die Vertauschungen berücksichtigen,

$$D_{ij} = (-1)^{n-i} \cdot (-1)^{n-j} \cdot \det A_{ij} = (-1)^{i+j} \det A_{ij} \,,$$

denn $(n-i) + (n-j) = 2n - (i+j)$ hat dieselbe Parität wie $(i+j)$. Fassen wir alles zusammen, erhalten wir die Formel

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} \,.$$

Die Entwicklung nach einer Zeile geht ganz entsprechend: Da die i-te Zeile von A gleich der i-ten Spalte von ${}^tA$ ist, führt obige Rechnung für

die transponierte Matrix auf die Formel

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} \, .$$

Damit haben wir den folgenden Satz bewiesen, der trotz seines Namens bereits LEIBNIZ bekannt war:

Entwicklungssatz von Laplace: A sei eine $n \times n$-Matrix und A_{ij} sei jene $(n-1) \times (n-1)$-Matrix, die durch Streichung der i-ten Zeile und der j-ten Spalte von A entsteht. Dann ist für jedes feste j

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij}$$

und für jedes feste i

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} \, ,$$

■

Die erste Formel aus diesem Satz bezeichnet man als die *Entwicklung der Determinante nach ihrer j-ten Spalte,* die zweite (die durch Anwendung der ersten auf die transponierte Matrix entsteht) entsprechend als die *Entwicklung nach der i-ten Zeile.*

PIERRE-SIMON DE LAPLACE (1749–1827) war einer der bedeutendsten französischen Wissenschaftler seiner Zeit; berühmt sind vor allem seine Anwendungen der Analysis auf die Wahrscheinlichkeitstheorie, die Himmelsmechanik, die Potentialtheorie und sein Vergleich des Universums mit einem Uhrwerk. Bekannt wurde er auch durch die nach ihm und KANT benannte Theorie zur Entstehung des Universums. Als politischer Opportunist kam er gut durch die Wirren seiner Zeit; er saß unter anderem in dem Kommitee, das Maßeinheiten nach dem Dezimalsystem einführte, war kurze Zeit Innenminister und wurde schließlich sogar Graf und Marquis.

BARON GOTTFRIED WILHELM VON LEIBNIZ (1646–1716) gilt als der letzte Universalgelehrte, der das gesamte Wissen seiner Zeit überblickte. In der Mathematik ist er vor allem berühmt durch die Entwicklung der Infinitesimalrechnung (bezüglich derer es einen langen Prioritätsstreit mit NEWTON gab); Bezeichnungen wie $\frac{dy}{dx}$ und $\int f(x)\,dx$ gehen auf ihn zurück. Durch seine Begründung der symbolischen Logik legte er auch einen wesentlichen Grundstein der späteren Informatik. Weitere Arbeiten befassen sich mit den Naturwissenschaften und der Technik, der Philosophie, Theologie und der Geschichte.

Als erste Anwendung wollen wir noch einmal die SARRUSsche Regel für Determinanten von 3×3-Matrizen beweisen: Entwicklung nach der ersten Spalte zeigt, daß

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}$$

$$= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{21}(a_{12}a_{33} - a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{13}a_{22})$$

ist, was ausmultipliziert natürlich genau die alte Formel ergibt.

Rein theoretisch ist es gleichgültig, nach welcher Zeile oder Spalte man eine Determinante entwickelt, aber man wird natürlich versuchen, eine zu finden, in der möglichst viele Nullen stehen, so daß in obiger Summe möglichst wenige Summanden übrigbleiben. In der Tat überlegt man sich leicht, daß die Anwendung des LAPLACEschen Entwicklungssatzes auf eine vollbesetzte Determinante wie bei der definierenden Formel rund $n \cdot n!$ Recheenoperationen erfordert und man somit gegenüber der direkten Anwendung dieser Formel nur den Vorteil hat, daß die Rechenoperationen klarer gegliedert werden.

Interessant zur praktischen Berechnung von Determinanten, vorzugsweise für nicht allzu große n, ist die Formel daher nur, wenn man sie mit den Rechenregeln aus dem vorigen Abschnitt kombiniert, um sich Nullen und einfache Einträge zu verschaffen. Da dieselbe Art von Operationen auch zur LR-Zerlegung führen, wird dann der Übergang zwischen

LAPLACEschem Entwicklungssatz und Berechnung via LR-Zerlegung fließend.

Betrachten wir als Beispiel die Determinante der Matrix

$$A = \begin{pmatrix} 1 & 2 & -2 & 1 \\ -1 & -3 & 2 & -2 \\ -4 & -2 & -1 & 0 \\ 3 & 2 & -1 & 1 \end{pmatrix}.$$

Sie hat zwar einen Eintrag Null, aber da in der dritten und vierten Spalte gleich in zwei Zeilen die Einträge entgegengesetzt gleich sind, ist es besser, zunächst die dritte Spalte durch die Summe von dritter und vierter Spalte zu ersetzen:

$$\begin{vmatrix} 1 & 2 & -2 & 1 \\ -1 & -3 & 2 & -2 \\ -4 & -2 & -1 & 0 \\ 3 & 2 & -1 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 2 & -1 & 1 \\ -1 & -3 & 0 & -2 \\ -4 & -2 & -1 & 0 \\ 3 & 2 & 0 & 1 \end{vmatrix}.$$

Wenn wir jetzt nach der dritten Spalte entwickeln, kommen nur zwei der Summanden wirklich vor: Der erste ($i = 1$ und $j = 3$) und der dritte ($i = j = 3$). In beiden Fällen ist $i + j$ gerade, der Vorfaktor $(-1)^{i+j}$ ist also jeweils +1. Die 3×3-Determinanten, die durch Streichung der dritten Spalte und der ersten *bzw.* dritten Zeile entstehen, können nach der Regel von SARRUS berechnet werden:

$$\begin{vmatrix} -1 & -3 & -2 \\ -4 & -2 & 0 \\ 3 & 2 & 1 \end{vmatrix} = (2 + 0 + 16) - (12 + 0 + 12) = -6$$

und

$$\begin{vmatrix} 1 & 2 & 1 \\ -1 & -3 & -2 \\ 3 & 2 & 1 \end{vmatrix} = (-3 - 12 - 2) - (-9 - 4 - 2) = -2\,.$$

Also ist $\det A = (-1)(-6) + (-1)(-2) = 8$.

Auch als Hilfmittel zur Berechnung von abstrakt gegebenen Determinanten ist der LAPLACEsche Entwicklungssatz oft nützlich, da er es bei geschickter Anwendung gelegentlich erlaubt, eine Rekursionsformel zu finden und damit eine allgemeine Formel für einen gewisse Klasse von Determinanten.

Als Beispiel betrachten wir die unter anderem in der Numerik wichtige VANDERMONDEsche Determinante

$$V(a_1, \ldots, a_n) = \begin{vmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \ldots & a_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_n & a_n^2 & \ldots & a_n^{n-1} \end{vmatrix}.$$

Der Franzose ALEXANDRE THÉOPHILE VANDERMONDE (1735–1796) war zunächst Musiker; erst im Alter von 35 Jahren begann er sich für Mathematik zu interessieren und publizierte in den Jahren 1771 und 1772 vier Arbeiten über Gleichungen, Determinanten und über das Problem, einen Springer so über ein Schachbrett zu bewegen, daß er jedes Feld genau einmal betritt. Die VANDERMONDEsche Determinante ist nirgends in seinem publizierten Werk zu finden; sie wurde erst um 1935 von HENRI LEBESGUE (1875–1941) nach ihm benannt.

Für die Anwendung des LAPLACEschen Entwicklungssatzes auf diese Determinante bietet sich an, nach der ersten Spalte zu entwickeln, denn diese besteht aus lauter Einsen, und die $(n-1) \times (n-1)$-Determinanten, die nach dem Entwicklungssatz auftreten, sind im wesentlichen wieder VANDERMONDEsche Determinanten. Allerdings entstehen dabei n Summanden, und für jeden von diesen muß die ganz Prozedur wiederholt werden *usw.*– die Berechnung der Determinante auf diese Weise ist also zumindest ziemlich aufwendig.

Ein besserer Ansatz ergibt sich, wenn wir die Einsen in der ersten Spalte dadurch ausnutzen, daß wir etwa die erste Zeile von jeder anderen Zeile subtrahieren. Der Wert der Determinante ändert sich dadurch nicht, d.h.

$$V(a_1, \ldots, a_n) = \begin{vmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^{n-1} \\ 0 & a_2 - a_1 & a_2^2 - a_1^2 & \ldots & a_2^{n-1} - a_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_n - a_1 & a_n^2 - a_1^2 & \ldots & a_n^{n-1} - a_1^{n-1} \end{vmatrix}.$$

Wenn wir hier nach der ersten Spalte entwickeln, muß nur eine einzige $(n-1) \times (n-1)$-Determinante berücksichtigt werden, alle anderen

haben den Vorfaktor null. Also ist

$$V(a_1,\dots,a_n)=\begin{vmatrix} a_2-a_1 & a_2^2-a_1^2 & \dots & a_2^{n-1}-a_1^{n-1}\\ a_3-a_1 & a_3^2-a_1^2 & \dots & a_3^{n-1}-a_1^{n-1}\\ \vdots & \vdots & \ddots & \vdots\\ a_n-a_1 & a_n^2-a_1^2 & \dots & a_n^{n-1}-a_1^{n-1}\end{vmatrix}$$

gleich jeder Determinanten, die durch Streichung der ersten Spalte und der ersten Zeile entsteht.

Hier können wir in jeder Zeile die jeweils vorne stehende Differenz ausklammern, denn genau wie

$$x^k-1=(x-1)(x^{k-1}+x^{k-2}+\cdots+x+1)$$

durch $(x-1)$ teilbar ist, ist auch

$$a_i^k-a_1^k=(a_i-a_1)(a_i^{k-1}+a_i^{k-2}a_1+a_i^{k-3}a_1^2+\cdots+a_ia_1^{k-2}+a_1^{k-1})$$

durch (a_i-a_1) teilbar; den Quotienten schreiben wir kurz als $q_{i,k-1}$:

$$q_{i,k-1}\underset{\text{def}}{=}a_i^{k-1}+a_i^{k-2}a_1+a_i^{k-3}a_1^2+\cdots+a_ia_1^{k-2}+a_1^{k-1}\,.$$

Wegen der Linearität der Determinante können wir jeden Faktor, den wir aus einer Zeile (oder Spalte) ausklammern, vor die Determinante ziehen und erhalten für $V(a_1,\dots,a_n)$ somit den Wert

$$(a_2-a_1)(a_3-a_1)\cdots(a_n-a_1)\begin{vmatrix}1 & q_{21} & q_{22} & \cdots & q_{2,n-2}\\ 1 & q_{31} & q_{32} & \cdots & q_{3,n-2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & q_{n1} & q_{n2} & \cdots & q_{n,n-2}\end{vmatrix}\,.$$

Die Nützlichkeit dieser Formel steht und fällt damit, daß wir die q_{ij} gut miteinander in Verbindung bringen können. Für verschiedene Indizes i haben die entsprechenden Ausdrücke offensichtlich wenig miteinander zu tun; sie enthalten nicht einmal dieselben Variablen. Schreiben wir allerdings

$$\begin{aligned}q_{ij}&=a_i^j+a_i^{j-1}a_1+a_i^{j-2}a_1^2+\cdots a_ia_1^{j-1}+a_1^j\\&=a_i^j+a_1(a_i^{j-1}+a_i^{j-2}a_1+\cdots a_ia_1^{j-2}+a_1^{j-1})\,,\end{aligned}$$

so sehen wir, daß

$$q_{ij}=a_i^j+a_1q_{i,j-1}\quad\text{oder}\quad q_{ij}-a_1q_{i,j-1}=a_i^j$$

ist. Subtrahieren wir also zuerst a_1 mal die vorletzte Spalte von der letzten, so werden die Einträge der letzten Spalte zu a_i^{n-2}. Entsprechend subtrahieren wir a_1 mal die $(n-2)$-te Zeile von der $(n-1)$-ten und erhalten lauter Einträge a_i^{n-3} und so weiter, bis schließlich die Subtraktion des a_1-fachen der ersten Spalte von der zweiten die Einträge der letzteren zu

$$q_{i1}-a_1=(a_i+a_1)-a_1=a_i$$

macht. Somit ist $V(a_1,\dots.a_n)$ gleich

$$(a_2-a_1)(a_3-a_1)\cdots(a_n-a_1)\begin{vmatrix}1 & a_2 & a_2^2 & \dots & a_2^{n-2}\\ 1 & a_3 & a_3^2 & \dots & a_3^{n-2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & a_n & a_n^2 & \dots & a_n^{n-2}\end{vmatrix}\,.$$

Die Determinante rechts ist offensichtlich wieder eine VANDERMONDEsche Determinante, allerdings mit um eins verminderter Zeilen- und Spaltenzahl und mit einer Variablen weniger.

Damit haben wir die Rekursionsformel

$$V(a_1,\dots,a_n)=(a_2-a_1)(a_3-a_1)\cdots(a_n-a_1)\,V(a_2,\dots,a_n)\,,$$

die es erlaubt, die Berechnung von $V(a_1,\dots,a_n)$ auf eine einzige VANDERMONDEsche Determinante der Größe $(n-1)\times(n-1)$ zurückzuführen.

Zur vollständigen Berechnung von $V(a_1,\dots,a_n)$ fehlt uns jetzt nur noch ein Induktionsanfang; direktes Nachrechnen zeigt sofort, daß

$$V(a_n)=\det(1)\quad\text{und}\quad V(a_{n-1},a_n)=\begin{vmatrix}1 & a_{n-1}\\ 1 & a_n\end{vmatrix}=a_n-a_{n-1}$$

ist, also folgt induktiv

$$V(a_1,\dots,a_n)=\prod_{j<i}(a_i-a_j)\,.$$

h) Die Cramersche Regel

Determinanten können auch angewandt werden, um die Lösungen eines linearen Gleichungssystems vom Rang n aus n Gleichungen in n Unbekannten in geschlossener Form als Funktion der Koeffizienten darzustellen. Dazu schreiben wir das Gleichungssystem $A\vec{x} = \vec{b}$ in der Form $x_1\vec{a}_1 + \cdots + x_n\vec{a}_n = \vec{b}$, wobei die $\vec{a}_i$ die Spaltenvektoren der Matrix A seien und $(x_1, \ldots, x_n)$ eine Lösung des linearen Gleichungssystems.

Nun ersetzen wir in $\det A = \det(\vec{a}_1, \ldots, \vec{a}_n)$ rechts den Vektor $\vec{a}_i$ durch die rechte Seite $\vec{b}$ des Gleichungssystems. (Es gibt eigentlich keinen vernünftigen Grund, warum wir das tun sollten; auch dieser Trick wird, wie so viele, erst nachträglich durch das Ergebnis gerechtfertigt.) Die so entstehende Determinante ist

$$\begin{aligned}
&\det(\vec{a}_1, \ldots, \vec{b}, \ldots, \vec{a}_n) \\
&= \det(\vec{a}_1, \ldots, \sum_{j=1}^{n} x_j\vec{a}_j, \ldots, \vec{a}_n) \\
&= \sum_{j=1}^{n} x_j \det(\vec{a}_1, \ldots, \vec{a}_j, \ldots, \vec{a}_n) \\
&= x_i \det(\vec{a}_1, \ldots, \vec{a}_i, \ldots, \vec{a}_n) = x_i \det A\,,
\end{aligned}$$

da für jeden Index $j \neq i$ der Vektor $\vec{a}_j$ zweimal als Argument der Determinante auftritt, so daß alle Summanden bis auf den i-ten verschwinden.

Falls $\det A = 0$ ist, nützt uns diese Formel überhaupt nichts; ist allerdings $\det A \neq 0$, wissen wir bereits, daß das Gleichungssystem eindeutig lösbar ist, und wir können die Komponenten dieser eindeutig bestimmten Lösung explizit ausdrücken durch die

CRAMERsche Regel: $$x_i = \frac{\det(\vec{a}_1, \ldots, \vec{b}, \ldots, \vec{a}_n)}{\det A}$$

Der Schweizer Mathematiker GABRIEL CRAMER (1704–1752) lehrte an der Universität Genf. Bekannt wurde er vor allem durch seine Arbeiten über Determinanten, er beschäftigte sich aber auch viel mit Analysis und Geometrie, insbesondere ist er Autor eines Buchs über algebraische Kurven. Weitere Arbeitsgebiete sind mathematische Methoden der Physik sowie die Geschichte der Mathematik. Die CRAMERsche Regel im Spezialfall $n = 2$ ist bereits 1545 in der *Ars magna* des italienischen Mathematikers GIROLAMO CARDANO (1501–1576) zu finden.

Die CRAMERsche Regel ist sicherlich kein Verfahren, das man oft anwendet zur Lösung eines einzelnen Gleichungssystems: Abgesehen von einigen Fällen mit sehr spärlich besetzter Matrix ist der Aufwand für die Berechnung von $n + 1$ Determinanten weit größer als eine Lösung nach dem GAUSS-Algorithmus. Falls man allerdings eine ganze Familie ähnlicher Gleichungssysteme hat, in der sich nur wenige Parameter ändern und bei denen die Determinanten auf Grund einer speziellen Form des Gleichungssystems gut berechenbar sind, kann es sich lohnen, mit Hilfe der CRAMERschen Regel eine (von den Parametern abhängige) Lösungsformel zu berechnen und dann diese anzuwenden.

i) Eigenwerte und Eigenvektoren

Aus §3*j1)* kennen wir die Begriffe *Eigenwert* und *Eigenvektor:* Ein vom Nullvektor verschiedener Vektor $\vec{v} \in k^n$ heißt Eigenvektor der Matrix $A \in k^{n\times n}$ zum Eigenwert $\lambda \in k$, wenn $A\vec{v} = \lambda\vec{v}$ ist, wenn also die Matrix A den Vektor $\vec{v}$ einfach auf eines seiner skalaren Vielfachen abbildet. Falls es eine Basis aus Eigenvektoren gibt, bekommt die Matrix A bezüglich dieser Basis Diagonalgestalt, wobei in der Diagonalen gerade die zugehörigen Eigenwerte stehen.

Wir werden uns erst im nächsten Semester (im Zusammenhang mit Systemen linearer Differentialgleichungen) genauer mit Eigenwerten und Eigenvektoren befassen und dann auch Kriterien kennenlernen, wann eine Matrix diagonalisierbar ist. In diesem Semester möchte ich, als Abschluß des Kapitels über lineare Algebra, nur kurz zeigen werden, wie man die Eigenwerte und Eigenvektoren einer Matrix finden kann.

Die Gleichung $A\vec{v} = \lambda\vec{v}$ läßt sich umschreiben als

$$(A - \lambda E)\vec{v} = \vec{0}\,,$$

das heißt, als ein homogenes lineares Gleichungssystem für den Vektor $\vec{v}$. Ein solches homogenes Gleichungssystem hat bekanntlich stets den Nullvektor als Lösung, der aber nach Definition genau aus diesem Grund kein Eigenvektor ist. Weitere Lösungen gibt es genau dann, wenn die Matrix $A - \lambda E$ des Gleichungssystems nicht den maximal möglichen Rang hat, wenn ihre Spaltenvektoren also linear abhängig sind oder wenn $\det(A - \lambda E)$ verschwindet.

Somit ist $\lambda \in k$ genau dann ein Eigenwert, wenn $\det(A - \lambda E) = 0$ ist; die zugehörigen Eigenvektoren sind die nichttrivialen Lösungen des homogenen linearen Gleichungssystems $(A - \lambda E)\vec{v} = \vec{0}$.

Damit ist klar, wie man Eigenwerte und Eigenvektoren berechnen kann: Man löse die Gleichung $\det(A - \lambda E) = 0$ und dann für jede Nullstelle λ_i dieser Gleichung das lineare Gleichungssystem $(A - \lambda E)\vec{v} = \vec{0}$. Dieses homogene lineare Gleichungssystem hat *nie* maximalen Rang, da es nach Definition eines Eigenwerts nichttriviale Lösungen geben muß; kommt man also auf ein eindeutig lösbares Gleichungssystem (und damit auf den Nullvektor als einzige Lösung), ist das immer ein Zeichen für einen Rechenfehler.

Dies zeigt auch, daß die numerische Bestimmung von Eigenvektoren nicht nach obigem Schema vorgehen kann: Falls man die Nullstellen der charakterischen Gleichung nur näherungsweise kennt, kann die Matrix des linearen Gleichungssystems eine leicht von null verschiedene Determinante haben, so daß das Gleichungssystem nur die triviale Lösung hat. In der Numerik werden Eigenwerte und Eigenvektoren daher simultan berechnet – sofern man beides braucht. Es gibt auch numerische Algorithmen, die nur die (angenäherten) Eigenwerte bestimmen ohne $\det(A - \lambda E)$ zu berechnen; dies ist vor allem nützlich für große n, wo die Berechnung einer Determinanten sehr aufwendig wäre. Für Genaueres sei auf die Numerikvorlesung verwiesen.

Als Beispiel für die Berechnung von Eigenwerten und Eigenvektoren

nach obiger Methode betrachten wir die Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 8 & 7 & 6 & 5 \\ 4 & 3 & 2 & 1 \end{pmatrix}.$$

Hier ist

$$A - \lambda E = \begin{pmatrix} 1-\lambda & 2 & 3 & 4 \\ 5 & 6-\lambda & 7 & 8 \\ 8 & 7 & 6-\lambda & 5 \\ 4 & 3 & 2 & 1-\lambda \end{pmatrix},$$

und nach dem Entwicklungssatz ist (bei Entwicklung nach der ersten Zeile)

$$\begin{aligned}
\det(A - \lambda E) &= (1-\lambda)\begin{vmatrix} 6-\lambda & 7 & 8 \\ 7 & 6-\lambda & 5 \\ 3 & 2 & 1-\lambda \end{vmatrix} \\
&\quad - 2\begin{vmatrix} 5 & 7 & 8 \\ 8 & 6-\lambda & 5 \\ 4 & 2 & 1-\lambda \end{vmatrix} \\
&\quad + 3\begin{vmatrix} 5 & 6-\lambda & 8 \\ 8 & 7 & 5 \\ 4 & 3 & 1-\lambda \end{vmatrix} - 4\begin{vmatrix} 5 & 6-\lambda & 7 \\ 8 & 7 & 6-\lambda \\ 4 & 3 & 2 \end{vmatrix} \\
&= (1-\lambda)(-\lambda^3 + 13\lambda^2 + 35\lambda) - 2(5\lambda^2 + 53\lambda) \\
&\quad + 3(-8\lambda^2 + \lambda) - 4(4\lambda^2 - 17\lambda) \\
&= \lambda^4 - 14\lambda^3 - 72\lambda^2 = \lambda^2(\lambda^2 - 14\lambda - 72)\,.
\end{aligned}$$

Dieser Ausdruck verschwindet genau dann, wenn entweder λ verschwindet oder der Klammerausdruck ganz hinten. Letzterer ist eine quadratische Gleichung für λ, die man entweder nach der üblichen Formel lösen kann, oder aber man beachtet den Wurzelsatz von VIÈTE, wonach die Summe der beiden Lösungen 14 und ihr Produkt -72 sein muß; da $72 = 4 \times 18$ ist, können die beiden Lösungen daher nur -4 und 18 sein. Die Nullstellen des Polynoms sind also $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = -4$ und $\lambda_4 = 18$.

Die Eigenwerte von A sind damit 0, -4 und 18; für jeden dieser Werte müssen wir über ein lineares Gleichungssystem die Eigenvektoren bestimmen.

Beginnen wir mit $\lambda_1 = \lambda_2 = 0$. Hier ist das Gleichungssystem einfach $A\vec{x} = \vec{0}$, d.h.

$$\begin{aligned} x_1 + 2x_2 + 3x_3 + 4x_4 &= 0 \\ 5x_1 + 6x_2 + 7x_3 + 8x_4 &= 0 \\ 8x_1 + 7x_2 + 6x_3 + 5x_4 &= 0 \\ 4x_1 + 3x_2 + 2x_3 + x_4 &= 0\,. \end{aligned}$$

Aufgrund der speziellen Struktur der Matrix A ergänzen sich die erste und die vierte Gleichung zu

$$5(x_1 + x_2 + x_3 + x_4) = 0\,;$$

entsprechend addieren sich die beiden mittleren Gleichungen zu

$$13(x_1 + x_2 + x_3 + x_4) = 0\,.$$

Da außerdem noch die Differenz zwischen den ersten beiden Gleichungen auf

$$4(x_1 + x_2 + x_3 + x_4) = 0$$

führt, genügt es, außer der ersten Gleichung noch als einzige weitere Gleichung

$$x_1 + x_2 + x_3 + x_4 = 0 \quad \text{oder} \quad x_1 = -(x_2 + x_3 + x_4)$$

zu betrachten. Subtrahiert man diese Gleichung von der ersten oder, was dasselbe ist, setzt man sie ein, so bleibt

$$x_2 + 2x_3 + 3x_4 = 0$$

als einzige Relation zwischen x_2, x_3 und x_4. Das gegebene Gleichungssystem ist also äquivalent zu

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 0 \\ x_2 + 2x_3 + 3x_4 &= 0\,; \end{aligned}$$

insbesondere hat es den Rang zwei und damit einen Lösungsraum der Dimension $4 - 2 = 2$.

(Es ist klar, daß der Rang des Gleichungssystems kleiner als vier sein muß, denn sont gäbe es keine nichttriviale Lösungen. Bei der Berechnung von Eigenvektoren über lineare Gleichungssysteme gibt es also immer Abhängigkeiten zwischen den einzelnen Gleichungen.)

Im noch verbliebenen Gleichungssystem kann beispielsweise x_4 beliebig gewählt werden. Ist $x_4 = 0$, so bleibt noch die Relation $x_2 + 2x_3 = 0$ übrig, in der etwa x_3 beliebig gewählt werden kann; danach sind $x_2 = -2x_3$ und $x_1 = x_3$ eindeutig festgelegt; die entsprechenden Lösungen sind also die Vielfachen der Lösung $(1, -2, 1, 0)$.

Genausogut könnte etwa $x_3 = 0$ gesetzt werden; dann wäre $x_2 = -3x_4$ und $x_1 = 2x_4$; dies führt auf die Vielfachen der Lösung $(2, -3, 0, 1)$.

Damit haben wir zwei linear unabhängige Lösungen gefunden; da der Lösungsraum zweidimensional ist, sind die sämtlichen Lösungen des Gleichungssystems die Linearkombinationen

$$x_1 = -\lambda - 2\mu, \quad x_2 = 2\lambda - 3\mu, \quad x_3 = \lambda \quad \text{und} \quad x_4 = \mu$$

dieser beiden Lösungen; die beiden Vektoren

$$\begin{pmatrix} -1 \\ 2 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} -2 \\ -3 \\ 0 \\ 1 \end{pmatrix}$$

bilden also eine Basis des Eigenraums von A zum Eigenwert Null.

Genauso lassen sich auch die Eigenwerte -4 und 18 behandeln: -4 führt auf das lineare Gleichungssystem

$$\begin{aligned} 5x_1 + 2x_2 + 3x_3 + 4x_4 &= 0 \\ 5x_1 + 10x_2 + 7x_3 + 8x_4 &= 0 \\ 8x_1 + 7x_2 + 10x_3 + 5x_4 &= 0 \\ 4x_1 + 3x_2 + 2x_3 + 5x_1 &= 0\,, \end{aligned}$$

dessen sämtliche Lösungen die Vielfachen von $(-1, -1, 1, 1)$ sind, und für den Eigenwert 18 schließlich erhalten wir das lineare Gleichungssy-

stem

$$\begin{aligned} -17x_1 + 2x_2 + 3x_3 + 4x_4 &= 0 \\ 5x_1 - 12x_2 + 7x_3 + 8x_4 &= 0 \\ 8x_1 + 7x_2 - 12x_3 + 5x_4 &= 0 \\ 4x_1 + 3x_2 + 2x_3 - 17x_4 &= 0 \end{aligned}$$

mit den Vielfachen von $(5, 13, 13, 5)$ als Lösungsraum.

Im nächsten Semester werden wir (durch ein sehr einfaches Argument) allgemein sehen, daß Eigenvektoren zu verschiedenen Eigenwerten stets linear unabhängig sind. Hier bei diesem Beispiel allerdings soll die lineare Unabhängigkeit zum weiteren Üben des Umgangs mit Determinanten direkt nachgerechnet werden:

$$\det(\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4) = \begin{vmatrix} 1 & 2 & -1 & 5 \\ -2 & -3 & 1 & 13 \\ 1 & 0 & 1 & 13 \\ 0 & 1 & 1 & 5 \end{vmatrix} .$$

Dieses Mal wollen wir die Determinante zur Abwechslung à la GAUSS ausrechnen, d.h. indem wir die Matrix zut Zeilenumformungen auf Dreiecksgestalt bringen: Addition von zweimal der ersten Zeile zur zweiten und Subtraktion der ersten Gleichung von der dritten zeigt, daß

$$\det(\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4) = \begin{vmatrix} 1 & 2 & -1 & 5 \\ 0 & 1 & -3 & 23 \\ 0 & -2 & 2 & 8 \\ 0 & 1 & 1 & 5 \end{vmatrix} = \begin{vmatrix} 1 & -3 & 23 \\ -2 & 2 & 8 \\ 1 & 1 & 5 \end{vmatrix}$$

ist. Addition von zweimal der ersten Zeile zur zweiten und Subtraktion der ersten Zeile von der dritten vereinfacht die rechtsstehende Determinante weiter zu

$$\det(\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4) = \begin{vmatrix} 1 & -3 & 23 \\ 0 & -4 & 54 \\ 0 & 4 & -18 \end{vmatrix} = \begin{vmatrix} -4 & 54 \\ 4 & -18 \end{vmatrix} .$$

Addiert man schließlich noch in der verbleibenden 2×2-Matrix die erste Zeile zur zweiten, erhält man das Endergebnis

$$\det(\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4) = \begin{vmatrix} -4 & 54 \\ 0 & 36 \end{vmatrix} = -4 \cdot 36 = -144 \neq 0 .$$

Also sind die Vektoren $\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4$ linear unabhängig und bilden eine Basis des $\mathbb{R}^4$.

Was haben wir nun erreicht? Betrachten wir die lineare Abbildung

$$\varphi: \mathbb{R}^4 \to \mathbb{R}^4; \quad \vec{v} \mapsto A \cdot \vec{v} .$$

Für die Vektoren

$$\vec{b}_1 = \lambda \begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \end{pmatrix}, \ \vec{b}_2 = \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \end{pmatrix}, \ \vec{b}_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \text{ und } \vec{b}_4 = \begin{pmatrix} 5 \\ 13 \\ 13 \\ 5 \end{pmatrix}$$

ist

$$\varphi(\vec{b}_1) = 0\,\vec{b}_1, \quad \varphi(\vec{b}_2) = 0\,\vec{b}_2, \quad \varphi(\vec{b}_3) = -4\,\vec{b}_3 \quad \text{und} \quad \varphi(\vec{b}_4) = 18\,\vec{b}_4 ;$$

bezüglich der neuen Basis $\{\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4\}$ hat φ daher die Abbildungsmatrix

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 18 \end{pmatrix} ,$$

bei der die Eigenwerte von A in der Hauptdiagonalen stehen und alle sonstigen Einträge verschwinden, wir konnten die Matrix A also diagonalisieren.

j) Geschichte und Anwendungen von Determinanten

Determinanten erblickten das Licht der Welt im Jahr 1683, und zwar gleich zweimal: Der japanische Mathematiker SEKI benutzte sie, ohne Ihnen einen Namen zu geben, zur Lösung von Gleichungen höheren Grades und zeigte anhand von Beispielen, wie man sie für 2×2- bis 5×5-Matrizen berechnet. Im gleichen Jahr schrieb auch LEIBNIZ einen Brief an DE L'HÔPITAL, in dem er erwähnte, daß ein gewisses homogenes lineares Gleichungssystem in drei Variablen nichttrivial lösbar sei, da (in heutiger Terminologie) seine Determinante verschwinde.

TAKAKAZU SEKI KOWA (1642–1708) war Sohn eines Samurai, wurde aber schon sehr jung von einem Adeligen namens SEKI GOROZAYEMON adoptiert. Einer von dessen Dienern weckte das Interesse des neunjährigen SEKI an der Mathematik, woraufhin dieser eine große Bibliothek japanischer und chinesischer Mathematikbücher anschaffte, anhand derer er sich selbst in das Gebiet einarbeitete. Als Staatsbeamter und ab 1704 Zeremonienmeister des Shogun befaßte er sich weiterhin viel mit Mathematik und entdeckte außer Determinanten beispielsweise auch das NEWTON-Verfahren und (vor JAKOB BERNOULLI) die BERNOULLI-Zahlen.

LEIBNIZ sprach noch nicht von Determinanten, sondern von *Resultanten;* ein Begriff, den unabhängig davon 1772 auch LAPLACE benutzte, als er damit die Bahnen der inneren Planeten berechnete. Das *Wort* Determinante erschien erstmal 1801 in den *Disquisitiones arithmeticae* von GAUSS, der damit die Eigenschaften quadratischer Formen untersuchte. Auch CAUCHY, der 1812 den Multiplikationssatz bewies, sprach von Determinanten.

Heute bezeichnet man als *Resultanten* spezielle Determinaten, die zur Lösung nichtlinearer Gleichungssysteme benutzt werden und auf den englischen Mathematiker JAMES JOSEPH SYLVESTER (1814–1897) zurückgehen. Mit Hilfe solcher Resultanten gelang es beispielsweise 1985 zwei Mathematikern bei *General Motors* das inverse kinematische Problem für einen Roboterarm mit sechs Freiheitsgraden zu lösen, d.h. also ein Verfahren zu entwickeln, mit dem der Manipulator am Ende des Arms *automatisch* auf eine vorgegebene Position und Ausrichtung gebracht werden kann.

Weitere wichtige Anwendungen haben Determinanten auch in der Numerik, wo sie sowohl *Konditionszahlen* definieren, die etwas über die Stabilität und Robustheit eines Verfahrens aussagen, als auch beispielsweise (wie die VANDERMONDEsche Determinante) bei der Approximation von Funktionen oder Datenpunkten durch Polynomfunktionen verwendet werden. Natürlich sind sie auch wichtig für Volumenberechnungen; dieser Aspekt wird uns auch im nächsten Kapitel wieder begegnen, wenn wir Mehrfachintegrale von einem Koordinatensystem in ein anderes transformieren.

Kapitel 2
Mehrdimensionale Analysis

Im letzten Kapitel hatten wir *lineare* Funktionen zwischen Vektorräumen betrachtet, insbesondere also auch Funktionen von $\mathbb{R}^n$ nach $\mathbb{R}^m$. Um solche Funktionen soll es auch in diesem Kapitel gehen, allerdings lassen wir nun die Forderung der Linearität fallen und verlangen nur noch deutlich schwächere Eigenschaften wie etwa Stetigkeit und/oder Differenzierbarkeit.

§1: Funktionen und ihre Eigenschaften

a) Darstellungsmöglichkeiten für Funktionen

Bei den linearen Funktionen im vorigen Kapitel hatten wir sowohl die Argumente als auch die Bilder als *Vektoren* aufgefaßt. Zumindest bei den Argumenten entspricht dies definitiv nicht der Betrachtungsweise der Analysis oder auch der Geometrie: Wir wollen Funktionen in *Punkten* auswerten, nicht in Vektoren. Wir fassen daher bei einer Funktion $f\colon \mathbb{R}^n \to \mathbb{R}^m$ den $\mathbb{R}^n$ nicht als Vektorraum auf, sondern als Punktmenge, wie wir es schon aus Kapitel I, §3*g)* von den affinen Räumen her gewohnt sind. Wie dort wollen wir aber Punkte mit Vektoren verknüpfen, wobei der Punkt $\boldsymbol{x} + \vec{h}$ der Endpunkt des im Punkt $\boldsymbol{x}$ abgetragenen Vektors $\vec{h}$ sein soll. Seine i-te Koordinate ist also $x_i + h_i$, wobei x_i die i-te Koordinate von $\boldsymbol{x}$ und h_i die i-te Komponente von $\vec{h}$ bezeichnet.

Für Funktionen $f\colon \mathbb{R} \to \mathbb{R}$ sind wir gewohnt, deren Graphen

$$\Gamma_f \underset{\text{def}}{=} \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; y = f(x)\right\}$$

zu zeichnen; entsprechend können wir natürlich auch für eine Funktion $f: D \to \mathbb{R}^m$ mit $D \subseteq \mathbb{R}^n$ den Graphen

$$\Gamma_f \underset{\text{def}}{=} \left\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^m \;\middle|\; \boldsymbol{y} = f(\boldsymbol{x})\right\}$$

definieren; da dieser in $\mathbb{R}^{n+m}$ liegt, ist er allerdings nur für $n + m \leq 3$ wirklich anschaulich, und auch da kann es bei komplizierten Funktionen stark von der gewählten Perspektive abhängen, was man sieht. Für einfache reellwertige Funktionen zweier Veränderlicher jedoch ist der Graph sicherlich die einfachste Methode der Veranschaulichung. Beim Graphen der Funktionen

$$f: \begin{cases} [-1,1] \times [-1,1] & \to \mathbb{R} \\ (x, y) & \mapsto \sqrt{4 - x^2 - y^2} \end{cases}$$

in Abbildung 25 etwa sieht man recht gut, daß Γ_f Teil einer Kugeloberfläche ist.

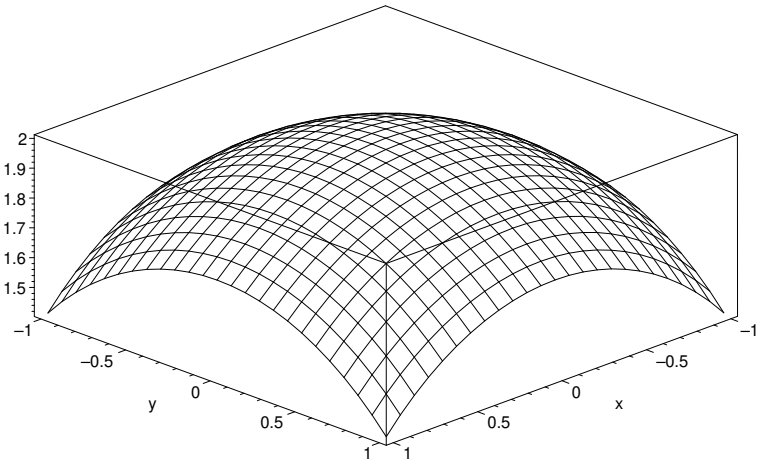


Abb. 25: Graph der Funktion $f(x, y) = \sqrt{4 - x^2 - y^2}$

Eine andere Möglichkeit zur Veranschaulichung von Funktionen zweier Veränderlicher ist von topographischen Karten her bekannt: Dort wird die Höhe über dem Meeresspiegel, eine Funktion der beiden Ebenenkoordinaten, dargestellt durch *Höhenlinien.* Entsprechend können wir für

eine beliebige Funktion $f: D \to \mathbb{R}$ mit $D \subseteq \mathbb{R}^2$ und jeden Wert $c \in \mathbb{R}$ die *Niveaulinie*

$$N_c(f) \underset{\text{def}}{=} \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; f(x, y) = c\right\}$$

definieren; sie muß natürlich keine „Linie“ sein, sondern kann auch nur aus einigen Punkten bestehen, leer sein oder – im Falle einer konstanten Funktion – für einen Wert c aus dem gesamten Definitionsbereich D der Funktion bestehen.

Im Falle des obigen Beispiels etwa ist $N_c(f)$ für $c > 2$ und für $c < \sqrt{2}$ die leere Menge; für $c = 2$ besteht sie nur aus dem Nullpunkt, und für $c = \sqrt{2}$ aus den vier Punkten $(0, \pm 1)$ und $(\pm 1, 0)$. Für $\sqrt{2} < c < 2$ erhalten wir die in Abbildung 26 für $c = 1{,}5$ bis $c = 2$ in Schritten von 0,05 dargestellten Kreislinien

$$\sqrt{4 - x^2 - y^2} = c \quad \text{oder} \quad x^2 + y^2 = 4 - c^2\,,$$

eingeschränkt natürlich auf das Einheitsquadrat als dem Definitionsbereich von f.

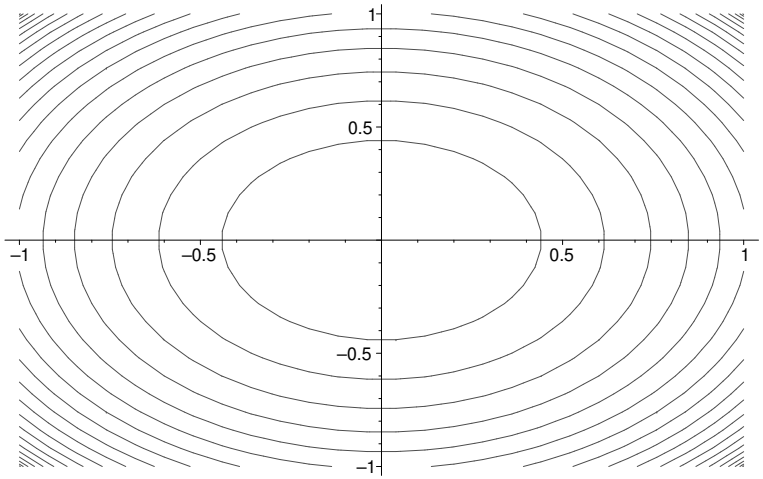


Abb. 26: Niveaulinien der Funktion $f(x, y) = \sqrt{4 - x^2 - y^2}$

Für Funktionen von mehr als zwei Veränderlichen ist die Visualisierung naturgemäß schwieriger; wir können Graphen und Niveaulinien, -flächen *usw.* zwar problemlos definieren, aber nicht mehr zeichnen – es sei denn, es handelt sich um sehr einfache Niveauflächen im $\mathbb{R}^3$. Bei Funktionen in einem mehrdimensionalen Raum kommt hinzu, daß die Niveauflächen dann nicht mehr nur von einem, sondern von mehreren Parametern abhängen.

Ansonsten gibt es für Funktionen von mehr als zwei Veränderlichen beispielsweise die Möglichkeit, einen Teil der Variablen auf interessanten Werten festzuhalten und die so eingeschränkte Funktion darzustellen. Dies gibt natürlich kein vollständiges Bild der Funktion, aber mehrere geschickt ausgewählte solche Bilder können doch einen recht guten Eindruck von der Funktion vermitteln.

Eine weitere Möglichkeit besteht darin, auf einem zwei- oder dreidimensionalen Graphen durch Farbe, Textur *usw.* weitere Dimensionen darzustellen; allgemein bekannt ist die Kodierung der Höhe durch von Grün nach Braun laufende Farben in Atlanten oder auch die Darstellung der Temperatur durch Farbverläufe von Blau über Rot nach Weiß. Grundsätzlich kann man mit Farben auch mehr als eine Dimension darstellen, da wir ja in Kapitel I, §4*d)* gesehen haben, daß Farben durch Punkte eines dreidimensionalen Raums beschrieben werden können. Zwar wird eine RGB-Darstellung von drei Dimensionen die meisten Betrachter überfordern, aber die Farbdarstellung zweier Dimensionen etwa durch eine Luminanz- und eine Chrominanzkoordinate ist durchaus anschaulich.

Ein eigenes Forschungsgebiet der Mathematik und Informatik, die Visualisierung, beschäftigt sich mit den Problem, die für eine bestimmte Fragestellung interessanten Aspekte einer (analytisch oder empirisch gegebenen) Funktion mehrerer Veränderlichen graphisch herauszuarbeiten.

b) Normierte Vektorräume

Mit Hilfe von klassischen wie auch HERMITEschen Skalarprodukten konnten wir die Länge $|\vec{v}| = \sqrt{\vec{v}\cdot\vec{v}}$ eines Vektors definieren und damit

beispielsweise auch Abstände zwischen Punkten eines affinen Raums –sofern der zugehörige Vektorraum EUKLIDisch oder HERMITEsch ist. Manchmal konnt es nur auf diese Längen an, nicht auf die Produkte; daher wollen wir diese hier für sich betrachten:

Sie haben folgende Eigenschaften:

a) $|\lambda\vec{v}| = |\lambda|\,|\vec{v}|$ für alle $\lambda \in \mathbb{R}$ und $\vec{v} \in V$

b) $|\vec{v}| \geq 0$ für alle $\vec{v} \in V$, und $|\vec{v}| = 0$ genau dann, wenn $\vec{v} = \vec{0}$

c) $|\vec{v}+\vec{w}| \leq |\vec{v}| + |\vec{w}|$.

Abgesehen von *c),* der *Dreiecksungleichung,* sind diese Eigenschaften klar; wegen

$$|\vec{v}+\vec{w}|^2 = |\vec{v}|^2 + |\vec{w}|^2 + 2\vec{v}\vec{w} \qquad \text{und}$$
$$(|\vec{v}| + |\vec{w}|)^2 = |\vec{v}|^2 + |\vec{w}|^2 + 2\,|\vec{v}|\,|\vec{w}|$$

folgt letztere im Falle $\vec{v}\cdot\vec{w} \geq 0$ sofort aus der CAUCHY-SCHWARZschen Ungleichung und für $\vec{v}\cdot\vec{w} < 0$ aus der Nichtnegativität der Norm.

Definition: Ein normierter Vektorraum ist ein $\mathbb{R}$- oder $\mathbb{C}$-Vektorraum V zusammen mit einer Abbildung $\|\cdot\| : V \to \mathbb{R}$ mit den Eigenschaften *a)* bis *c).*

Damit ist also jeder EUKLIDische oder HERMITEsche Vektorraum insbesondere auch ein normierter Vektorraum, es gibt aber auch normierte Vektorräume, deren Norm nichts mit einem Skalarprodukt zu tun hat:

Beispielsweise erfüllt die *Maximumsnorm*

$$\|\cdot\|_\infty : \mathbb{R}^n \to \mathbb{R}; \quad \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \mapsto \max_{i=1}^{n} |v_i|$$

offensichtlich die Bedingungen *a)* bis *c).* Sie kommt aber nicht von einem Skalarprodukt, denn gäbe es ein Skalarprodukt $\star$ mit $\|\vec{v}\|_\infty = \sqrt{\vec{v}\star\vec{v}}$; so wäre etwa im $\mathbb{R}^2$

$$\begin{pmatrix}1\\0\end{pmatrix} \star \begin{pmatrix}1\\0\end{pmatrix} = \left\|\begin{pmatrix}1\\0\end{pmatrix}\right\|_\infty^2 = 1 \quad \text{und} \quad \begin{pmatrix}0\\1\end{pmatrix} \star \begin{pmatrix}0\\1\end{pmatrix} = \left\|\begin{pmatrix}0\\1\end{pmatrix}\right\|_\infty^2 = 1\,,$$

also

$$1 = \left\| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\|_\infty^2 = \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) \star \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = 1+1+2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \star \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

und somit $\begin{pmatrix} 1 \\ 0 \end{pmatrix} \star \begin{pmatrix} 0 \\ 1 \end{pmatrix} = -\frac{1}{2}$ und $\begin{pmatrix} 2 \\ 0 \end{pmatrix} \star \begin{pmatrix} 0 \\ 1 \end{pmatrix} = -1$. Mithin wäre

$$4 = \left\| \begin{pmatrix} 2 \\ 1 \end{pmatrix} \right\|_\infty^2 = \left(\begin{pmatrix} 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) \star \left(\begin{pmatrix} 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = 4 + 1 - 2 = 3 \,,$$

ein offensichtlicher Widerspruch.

Maximumsnormen lassen sich nicht nur für $\mathbb{R}^n$ oder $\mathbb{C}^n$ definieren, sondern auch für Funktionenräume: Eine stetige Funktion $f\colon [a,\, b] \to \mathbb{R}$ auf einem *abgeschlossenen* Intervall $[a,\, b]$ nimmt ihr Maximum wirklich an, d.h die Abbildung

$$\|\cdot\|_\infty : \mathcal{C}^0\big([a,\, b],\, \mathbb{R}\big) \to \mathbb{R}; \quad f \mapsto \max_{x \in [a,\, b]} |f(x)|$$

ist wohldefiniert (für eine stetige Funktion f ist auch $|f|$ eine stetige Funktion) und sie hat die Eigenschaften *a)* bis *c): a)* und *b)* sind, wie in den meisten Fällen, trivial, und sind $f, g \in \mathcal{C}^0\big([a,\, b],\, \mathbb{R}\big)$ zwei Funktionen, von denen f ihr Maximum in x_1 annimmt, g in x_2 und $f+g$ in x^*, so ist

$$\begin{aligned} \|f+g\|_\infty &= \big|(f+g)(x^*)\big| = \big|f(x^*) + g(x^*)\big| \le \big|f(x^*)\big| + \big|g(x^*)\big| \\ &\le \max_{x \in [a,\, b]} |f(x)| + \max_{x \in [a,\, b]} |g(x)| = f(x_1) + g(x_2) \\ &= \|f\|_\infty + \|g\|_\infty \,. \end{aligned}$$

Maximumsnormen spielen unter anderem in der Numerik eine wichtige Rolle; sie gestatten es beispielsweise, Fehlerschranken für numerische Rechnungen herzuleiten:

Führen wir beispielsweise auf dem Vektorraum $\mathbb{R}^{n \times m}$ aller $n \times m$-Matrizen die Maximumsnorm ein, so ist natürlich

$$\|A\|_\infty = \max_{i=1}^{n} \max_{j=1}^{m} \big|a_{ij}\big| \,.$$

Betrachten wir auch $\mathbb{R}^m$ mit der Maximumsnorm, so folgt für ein Produkt $A\vec{v} = \vec{b}$ sofort aus der Multiplikationsregel

$$b_i = \sum_{j=1}^{m} a_{ij} v_j$$

und der gewöhnlichen Dreiecksungleichung aus $\mathbb{R}$

$$|b_i| \le \sum_{j=1}^{m} \big|a_{ij}\big| \cdot \big|v_j\big| \,,$$

daß $\left\|\vec{b}\right\|_\infty \le \|A\|_\infty \cdot \|\vec{v}\|_\infty$ ist.

Wird also der Vektor $\vec{v}$ durch einen Fehlervektor $\vec{\epsilon}$ gestört, so ist

$$A(\vec{v} + \vec{\epsilon}) = A\vec{v} + A\vec{\epsilon} = \vec{b} + A\vec{\epsilon}$$

mit einem Fehler behaftet, dessen Komponenten *mit Sicherheit* kleiner sind als

$$\|A\|_\infty \cdot \|\vec{\epsilon}\|_\infty \,.$$

(Tatsächlich wird diese Schranke in den meisten Fällen viel zu groß sein, aber realistische Schranken sind in der Numerik oft – wenn überhaupt – nur mit sehr großem Aufwand zu finden.)

Ebenfalls eine sehr wichtige Rolle spielen Normen bei Funktionenräumen; dafür werden wir im nächsten Semester zahlreiche Beispiele sehen. Der wesentliche Punkt ist, daß man bezüglich einer Norm in offensichtlicher Verallgemeinerung der klassischen Definitionen Begriffe wie Konvergenz und Stetigkeit definieren kann:

Definition: *a)* $\big(V, \|\cdot\|\big)$ sei ein normierter $\mathbb{R}$-Vektorraum. Eine Folge $\vec{v}_1, \vec{v}_2, \ldots$ von Vektoren aus V konvergiert gegen den Vektor $\vec{v} \in V$, wenn es zu jedem $\varepsilon > 0$ eine natürliche Zahl $N \in \mathbb{N}$ gibt, so daß $\|\vec{v} - \vec{v}_n\| < \varepsilon$ für alle $n > N$.
b) Eine Abbildung $f\colon D \to W$ von der Teilmenge $D \subseteq V$ eines normierten Vektorraums $\big(V, \|\cdot\|_1\big)$ in einen normierten Vektorraum $\big(W, \|\cdot\|_2\big)$ heißt stetig in $\vec{v}_0 \in D$, wenn es für jedes $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für alle $\vec{v} \in D$ gilt: Ist $\|\vec{v} - \vec{v}_0\| < \delta$, so ist $\|f(\vec{v}) - f(\vec{v}_0)\| < \varepsilon$.
c) f heißt stetig, wenn f in jedem Punkt $\vec{v}_0 \in D$ stetig ist.

Anschaulich betrachtet bedeutet Konvergenz, daß ich die Punkte immer näher an einen Grenzwert annähern, und Stetigkeit bedeutet, daß die Funktion keine Sprünge macht. Warum brauchen wir für so einfache und anschauliche Dingen so komplizierte Definitionen wie die obigen?

Ursprünglich, bei NEWTON und bei LEIBNIZ geht tatsächlich noch alles ohne δ und ϵ mit Definitionen, die der anschaulichen Vorstellung entsprechen, und beide hatten damit großen Erfolg: Insbesondere wurde die Physik auf eine völlig neue Grundlage gestellt und die naturwissenschaftliche Erklärung der Welt machte rasante Fortschritte.

SIR ISAAC NEWTON (1643–1727) war ab 1661 Student, ab 1669 Professor an der Universität Cambridge. Dort entwickelte er die Infinitesimalrechnung, die er 1671 in seinem Buch *De Methodis Serierum et Fluxionum* beschrieb, arbeitete über Optik, wo er unter anderem dünne Schichten und Beugungsphänomene untersuchte (NEWTONsche Ringe), entdeckte seine Bewegungsgesetze und das Gravitationsgesetz, veröffentlicht 1687 in seinem Buch *Philosophiae naturalis principia mathematica,* das von vielen als bedeutendstes wissenschaftliches Buch aller Zeiten angesehen wird. Nach zwei Nervenzusammenbrüchen ging er 1693 nach London, wo er die königliche Münze leitete.

Die neue Weltsicht einiger Naturwissenschaftler führte zu Spannungen mit der Theologie; mehrere Theologen veröffentlichten ihrerseits mehr oder weniger fundierte Kritiken an den Naturwissenschaften. Heute noch bekannt sind beispielsweise *Gulliver's travels* von JONATHAN SWIFT (1667–1745), wo im dritten Buch vor allem die Mathematiker und Naturwissenschaftler aufs Korn genommen werden.

Am berechtigtsten war die Kritik in BERKELEYs Buch „The Analyst: or a discourse addressed to an infidel mathematician" (gemeint war wahrscheinlich EDMOND HALLEY (1656?–1743), heute vor allem bekannt durch den nach ihm benannten Kometen), in dem er zeigte, wie unsicher die Grundlagen der Infinitesimalrechnung sind und wie leicht man im Umgang damit zu unsinnigen Ergebnissen kommen kann – ganz im Gegensatz, so seine Meinung, zur wissenschaftlich erheblich fundierteren und logischer aufgebauten Theologie. Kleiner Auszug: *And what are these fluxions? The velocities of evanescent increments. And what are these same evanescent increments? They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them ghosts of departed quantities?*

Es dauerte geraume Zeit, bis die Mathematik dieser Kritik etwas Substantielles entgegensetzen konnte: Erst um 1800 gelang es CAUCHY die Analysis logisch zweifelsfrei zu verankern in der Theorie der algebraischen Ungleichungen; seit dieser Zeit arbeiten wir mit ε und δ. Mittlerweile gibt es mit der sogenannten *nonstandard analysis* auch eine Alternative, die in der Prädikatenlogik erster Stufe verankert ist; da die Prädikatenlogik erster Stufe aber zum Beispiel keine vollständige Induktion gestattet, ist diese Alternative jedoch keineswegs einfacher, sondern hängt an recht diffizilen logischen Sätzen.

GEORGE BERKELEY (1685–1753) studierte und lehrte Theologie am Trinity College in Dublin. In der Philosophie gilt er als einer der Begründer des Empirizismus, den er als Gegenposition zum mechanistischen Materialismus der Naturwissenschaften seiner Zeit aufbaute. Als sehr streitbarer Gelehrter war er bald mit fast allen Kollegen am Trinity College verkracht; daher wurde er 1734 zum Bischof von Cloyne, weit weg von Dublin, ernannt, wo er sowohl praktisch als auch durch eine dreibändige Streitschrift für die Rechte der Landbevölkerung kämpfte. Anläßlich einer Ruhr-Epedemie 1741 begann er auch mit pharmazeutischen Studien.

Mit den obigen Definitionen haben wir somit ein zwar solides, aber doch gelegentlich unhandliches Werkzeug, mit dem wir in jedem normierten Vektorraum Analysis betreiben können. Wir müssen aber damit rechnen, daß die Ergebnisse im allgemeinen stark von der Norm abhängen werden, und in der Tat werden wir dies im nächsten Semester eindrucksvoll sehen können. In diesem Semester allerdings interessieren wir uns vor allem für die Vektorräume $\mathbb{R}^n$, und dort ist die Situation deutlich harmloser.

Offensichtlich definieren zwei Normen denselben Konvergenzbegriff und denselben Stetigkeitsbegriff, wenn sie äquivalent sind im Sinne der folgenden Definition:

Definition: Zwei Normen $\|\cdot\|_1$ und $\|\cdot\|_2$ auf einen Vektorraum V heißen äquivalent, wenn es reelle Konstanten $c_1, c_2 > 0$ gibt, so daß

$$c_1 \|\vec{v}\|_1 \leq \|\vec{v}\|_2 \leq c_2 \|\vec{v}\|_1 \ .$$

Offensichtlich ist dann auch

$$\frac{1}{c_2} \|\vec{v}\|_2 \leq \|\vec{v}\|_1 \leq \frac{1}{c_1} \|\vec{v}\|_2 \ ,$$

die Äquivalenz ist also, wie es sein muß, symmetrisch.

Beispielsweise sind auf jedem $\mathbb{R}^n$ die EUKLIDische Norm $\|\cdot\|_2$ und die Maximumsnorm $\|\cdot\|_\infty$ äquivalent, denn

$$\|\vec{v}\|_2 = \sqrt{v_1^2 + \cdots + v_n^2} \le \sqrt{n\,\|\vec{v}\|_\infty^2} = \sqrt{n}\,\|\vec{v}\|_\infty$$

und

$$\|\vec{v}\|_\infty = \sqrt{\|\vec{v}\|_\infty^2} \le \sqrt{v_1^2 + \cdots + v_n^2} = \|\vec{v}\|_2 \ ,$$

d.h.

$$\|\vec{v}\|_\infty \le \|\vec{v}\|_2 \le \sqrt{n}\,\|\vec{v}\|_\infty \ .$$

Anschaulich bedeutet dies, daß jeder Würfel in eine Kugel eingebettet werden kann und umgekehrt, denn

$$\left\{\boldsymbol{x} \in \mathbb{R}_n \;\middle|\; \|\boldsymbol{x}\|_\infty \le a\right\}$$

ist ein Würfel mit Kantenlänge $2a$ und

$$\left\{\boldsymbol{x} \in \mathbb{R}_n \;\middle|\; \|\boldsymbol{x}\|_2 \le r\right\}$$

eine Kugel mit Radius r.

Tatsächlich kann man zeigen, daß in $\mathbb{R}^n$ alle Normen äquivalent sind, so daß es dort nur einen Konvergenz- und Stetigkeitsbegriff gibt.

Wir werden in $\mathbb{R}^n$ praktisch immer mit der EUKLIDischen Norm oder der Maximumsnorm arbeiten, wobei die erstere anschaulicher ist, die letztere aber meist einfacher für konkretes Nachrechnen.

Als Beispiel sei noch einmal angegeben, wie die obige Stetigkeitsdefinition aussieht für eine Funktion von $\mathbb{R}^n$ nach $\mathbb{R}$, ausgedrückt in der Maximumsnorm:

Definition: Eine Funktion $f\colon D \to \mathbb{R}$ auf $D \subseteq \mathbb{R}^n$ heißt stetig im Punkt $\boldsymbol{x} = (x_1, \ldots, x_n) \in D$, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für jeden Punkt $\boldsymbol{y} = (y_1, \ldots, y_n) \in D$ gilt: Falls $|y_i - x_i| < \delta$ ist für alle i, dann ist $|f(\boldsymbol{y}) - f(\boldsymbol{x})| < \varepsilon$.

Im Eindimensionalen kann man der graphischen Darstellung einer Funktion leicht ansehen, ob sie stetig ist oder nicht; wir wollen schauen, wie dies im Mehrdimensionalen ist.

Der Einfachheit halber beschränken wir uns auf ein Funktion zweier Veränderlicher, z.B. die Funktion

$$f\colon \begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x,y) & \mapsto \begin{cases} \frac{2xy^2}{x^2+y^4} & \text{falls } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases} \end{cases} .$$

Auf $\mathbb{R}^2 \setminus \{(0,0)\}$ ist die Funktion offensichtlich stetig, denn dort ist sie nur durch Grundrechenarten definiert, wobei Division durch Null ausgeschlossen ist, da $x^2 + y^4$ nur im Nullpunkt verschwindet. Bleibt also der Punkt $(0,0)$ zu untersuchen.

Der Graph von f in einer kleinen Umgebung von $(0,0)$ ist in Abbildung 27 zu sehen. Er zeigt zwar einen relativ steilen Sprung entlang der Geraden $x = 0$, aber nur für die etwas weiter vom Nullpunkt entfernten x-Werte; bei $y = 0$ sieht alles harmlos und ziemlich eben aus.

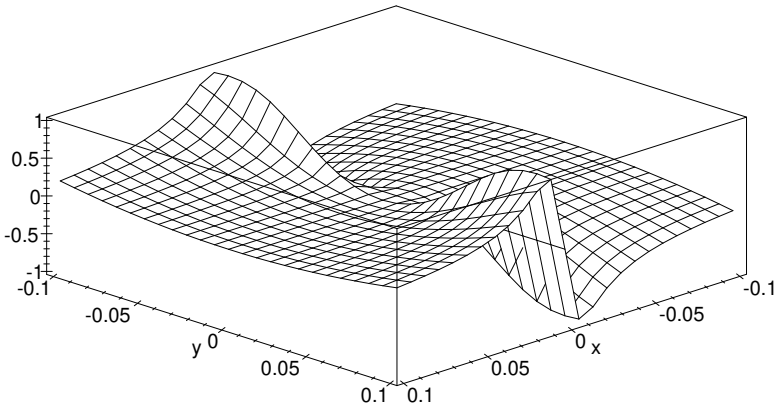


Abb. 27: Ist diese Funktion stetig?

Nun ist der abgebildete Graph natürlich von einem Computer anhand von nur endlich vielen Stützpunkten konstruiert; um wirklich zu entscheiden, ob f im Nullpunkt stetig ist, können wir uns nicht auf diese Approximation verlassen, sondern müssen die Funktion etwas genauer untersuchen.

Da wir uns im Eindimensionalen recht gut auskennen, können wir beispielsweise die Einschränkungen von f auf die verschiedenen Geraden durch den Nullpunkt betrachten. Abgesehen von der y-Achse haben diese alle die Form $y = ax$ mit $a \in \mathbb{R}$, und

$$f(x, ax) = \frac{2x \cdot (ax)^2}{x^2 + (ax)^4} = \frac{2ax^3}{x^2(1 + ax^2)} = \frac{2ax}{1 + ax^2}$$

für $x \neq 0$. Für $x \to 0$ geht beim rechtsstehenden Ausdruck der Zähler gegen Null und der Nenner gegen eins; der Grenzwert existiert also und ist gleich $f(0, 0) = 0$, d.h. die Einschränkung von f ist stetig auf der Geraden $y = ax$.

Auf der y-Achse verschwindet f für jeden Wert von x, ist also ebenfalls stetig, d.h. die Einschränkung von f auf jede Gerade durch den Nullpunkt ist stetig.

Betrachten wir zur Vorsicht auch noch die Einschränkung von f auf die Parabel $x = ay^2$! Dort ist

$$f(ay^2, y) = \frac{2ay^2 \cdot y^2}{a^2y^4 + y^4} = \frac{2ay^4}{(1 + a^2)y^4} = \frac{2a}{1 + a^2}$$

für alle $y \neq 0$, wohingegen $f(0, 0) = 0$ ist. Für $a \neq 0$ ist die Einschränkung von f auf diese Parabel also nicht stetig, und damit kann auch f nicht stetig sein, denn für eine stetige Funktion muß schließlich

$$\lim_{(x,y)\to(0,0)} f(x, y) = f(0, 0)$$

unabhängig davon sein, auf welchem Weg der Punkt (x, y) gegen $(0, 0)$ wandert.

Zusammen mit den Koordinatenachsen überdecken die Parabeln $x = ay^2$ mit $a \in \mathbb{R} \setminus \{0\}$ den gesamten $\mathbb{R}^2$, die Niveaulinie $N_0(f)$ von f besteht also aus den beiden Koordinatenachsen, während die anderen Niveaulinien aus Parabeln $x = ay^2$ jeweils ohne den Nullpunkt bestehen; siehe Abbildung 28. Da die Gleichung

$$\frac{2a}{1 + a^2} = c$$

für $c \neq 0$ die beiden Lösungen

$$a = \frac{1 \pm \sqrt{1 - c^2}}{c}$$

hat, besteht jede Niveaulinie für $0 < |c| < 1$ aus zwei dieser Parabeln, für $|c| = 1$ aus einer, und für $|c| > 1$ ist $N_c(f) = \emptyset$.

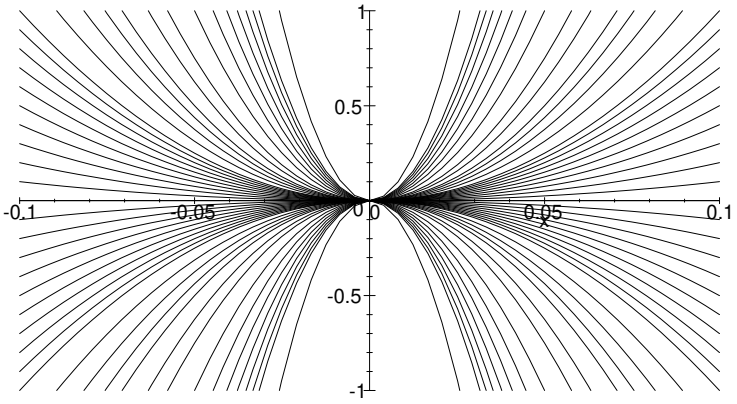


Abb. 28: Die Niveaulinien illustrieren die Unstetigkeit im Nullpunkt

Abbildung 28 zeigt, daß f im Nullpunkt unstetig ist, denn alle Niveaulinien kommen dem Nullpunkt beliebig nahe, obwohl dieser nur auf $N_0(f)$ liegt. Daran sehen wir übrigens auch, daß der Graph in Abbildung 27 falsch ist: In jeder Umgebung von $(0, 0)$ wird jeder Wert c zwischen -1 und 1 angenommen, der Abschluß des Graphen enthält also die Strecke $[-1, 1]$ auf der z-Achse. In Abbildung 27 ist dies nicht zu sehen, da Maple, wie die meisten Computergraphikprogramme, zum Zeichnen nur die Linien $x =$ konstant und $y =$ konstant berücksichtigt.

Bei Funktionen, in deren Definition Fallunterscheidungen eingehen, ist also größere Vorsicht geboten als im Eindimensionalen; bei in der Praxis auftretenden Funktionen wird es allerdings wohl meist so sein, daß die Unstetigkeitsstellen genau dort auftreten, wo man Sprünge definiert hat. Schwierig wird es nur, wenn man wie im obigen Beispiel in einem Punkt eine Situation der Art „$0/0$“ hat und entscheiden muß, ob

man stetig ergänzen kann: Hier hilft im Mehrdimensionalen keine DE L'HOSPITALsche Regel, es hilft auch nicht, die Annäherung der Funktion an den problematischen Punkt aus allen Richtungen zu untersuchen, sondern man muß wirklich auf die Definition der Stetigkeit zurückgehen.

c) Die Ableitung einer Funktion

Als nächstes möchte ich auf die *Differenzierbarkeit* von Funktionen eingehen, als erstes im wohlbekannten Fall einer Funktion auf einem Intervall $(a, b) \subset \mathbb{R}$.

Rein formal betrachtet ist $f\colon (a, b) \to \mathbb{R}$ differenzierbar in einem Punkt $x \in (a, b)$, wenn der Grenzwert

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

existiert; sein Wert wird dann als Ableitung von f im Punkt x bezeichnet.

Hier und im folgenden wird oft die LANDAUsche o-Notation nützlich sein: Wir schreiben $o(h)$, sobald wir *irgendeine* uns nicht weiter interessierende Funktion von h haben, die für $h \to 0$ schneller gegen null geht als h selbst, d.h.

$$\varphi(h) = o(h) \iff \lim_{h\to 0} \frac{\varphi(h)}{h} = 0\,.$$

$o(h)$ ist hier also keine Funktion, sondern steht für eine ganze Klasse von Funktionen; beispielsweise ist

$$h^2 = o(h), \quad h^5 = o(h) \quad \text{und} \quad h \cdot \sin h = o(h)\,,$$

aber $\sin h$ können wir nicht als $o(h)$ schreiben, denn

$$\lim_{h\to 0} \frac{\sin h}{h} = 1\,.$$

Entsprechend schreiben wir auch

$$\varphi(h) = o\big(\psi(h)\big), \quad \text{wenn} \quad \lim_{h\to 0} \frac{\varphi(h)}{\psi(h)} = 0$$

ist.

EDMUND GEORG HERMANN LANDAU (1877–1938) wurde in Berlin geboren und studierte an der dortigen Universität, wo er auch von 1899 bis 1909 lehrte. Dann bekam er einen Ruf an die damals führende deutsche Mathematikfakultät in Göttingen. 1933 verlor er seinen dortigen Lehrstuhl, denn die Studenten boykotierten seine Vorlesungen, da sie meinten, sie könnten Mathematik nur von einem Professor ihrer eigenen Rasse lernen. LANDAUS zahlreiche Publikationen beschäftigen sich vor allem mit der Zahlentheorie, über die er auch ein bedeutendes Lehrbuch schrieb; sehr bekannt sind seine Arbeiten über die Verteilung von Primzahlen.

Mit LANDAUS o-Notation können wir kurz sagen, die Funktion f sei genau dann differenzierbar in x mit Ableitung $f'(x)$, wenn

$$f(x+h) = f(x) + hf'(x) + o(h)$$

ist, denn das bedeutet gerade, daß

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

Anschaulich können wir das auch so interpretieren, daß für kleine h

$$f(x+h) \approx f(x) + hf'(x)\,,$$

ist, d.h. die Funktion f sieht in einer hinreichend kleinen Umgebung von x aus wie eine lineare Funktion. Die Abbildungen 29 bis 32 zeigen dies für $f(x) = \sin x$ in der Umgebung von $x = 1$, wobei diese Umgebung von einer Abbildung zur nächsten jeweils um den Faktor fünf vergrößert wird.

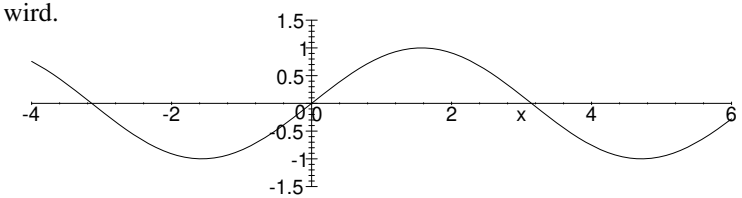


Abb. 29: Graph der Funktion $y = \sin x$

Für eine Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ können wir ganz entsprechend vorgehen: Wenn wir etwa die Funktion $f(x, y) = \sin x \cos y$ auf $\mathbb{R}^2$ in der Umgebung des Punktes $(1, 1)$ betrachten und diese wieder sukzessive um

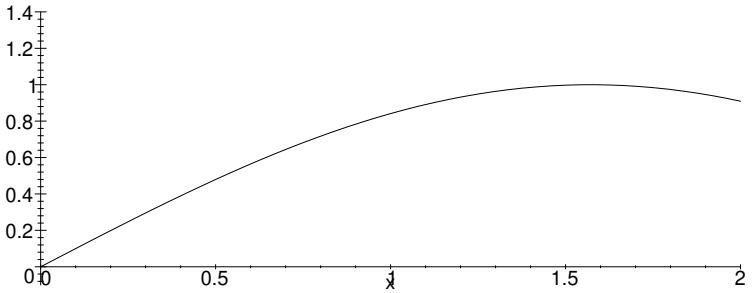


Abb. 30: Vergrößerung um den Faktor fünf um $\big(1, \sin(1)\big)$

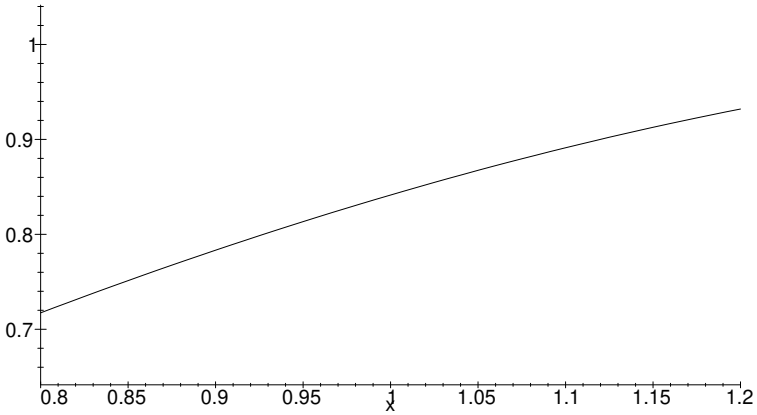


Abb. 31: Nochmalige Vergrößerung um den Faktor fünf um $\big(1, \sin(1)\big)$

den Faktor fünf vergrößern, erhalten wir die Abbildungen 33 bis 36, die zeigen, daß sich der Graph der Funktion in einer hinreichend kleinen Umgebung des betrachteten Punktes nur wenig von einer Ebenen unterscheidet, d.h. die Funktion ist dort annähernd linear.

Eine lineare Funktion zweier Veränderlicher hat die Form

$$L(x, y) = a + bx + cy;$$

also ist

$$L(x + h, y + k) = a + b(x + h) + c(y + k) = (a + bx + cy) + ah + bk$$

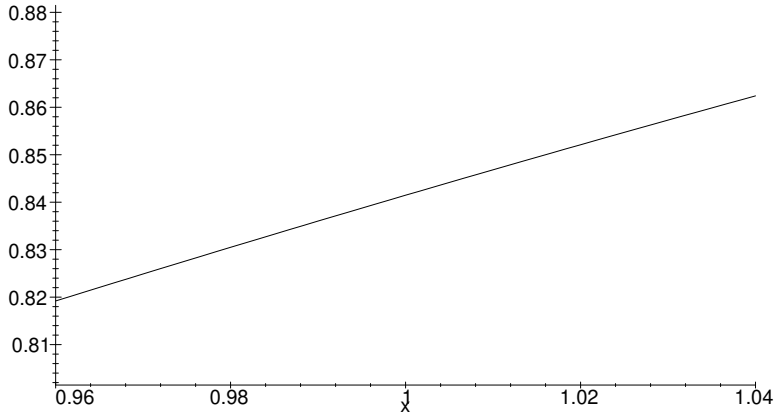


Abb. 32: Nach noch einer Vergrößerung sieht die Funktion praktisch linear aus

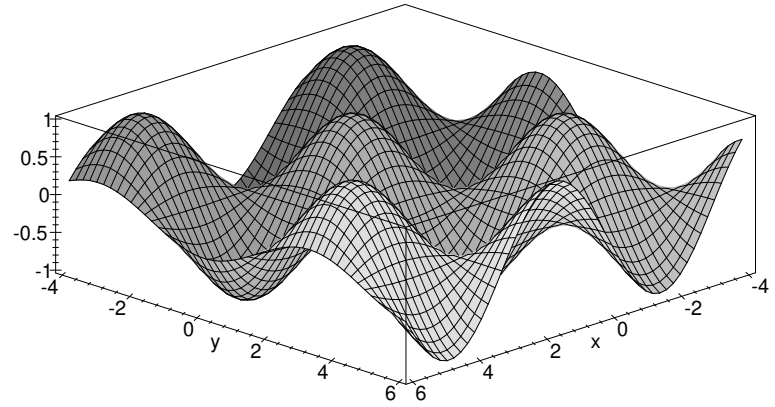


Abb. 33: Graph der Funktion $z = \sin x \cos y$

eine Approximation für f in der Umgebung des betrachteten Punkts (x, y), hier gleich $(1, 1)$. Im Punkt (x, y) sollte

$$L(x, y) = a + bx + cy$$

natürlich mit $f(x, y)$ übereinstimmen, und für $(h, k) \neq (0, 0)$ sollte der Unterschied zwischen f und L schneller gegen Null gehen als der Abstand zwischen $(x+h, y+k)$ und (x, y), d.h. schneller als $\sqrt{h^2 + k^2}$. Wir

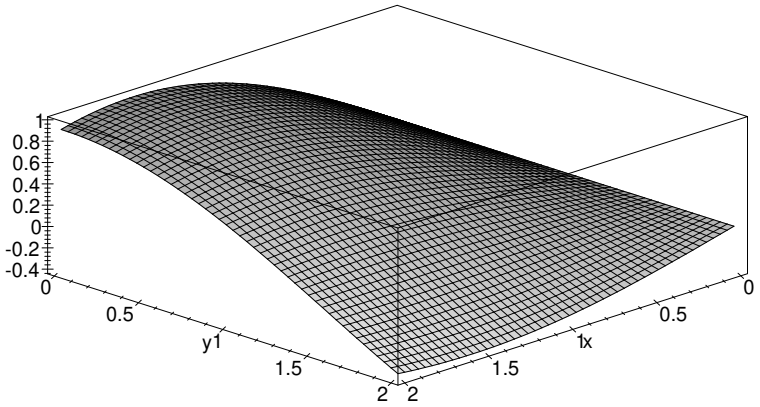


Abb. 34: Vergrößerung um den Faktor fünf um $\big(1, 1, \sin(1)\cos(1)\big)$

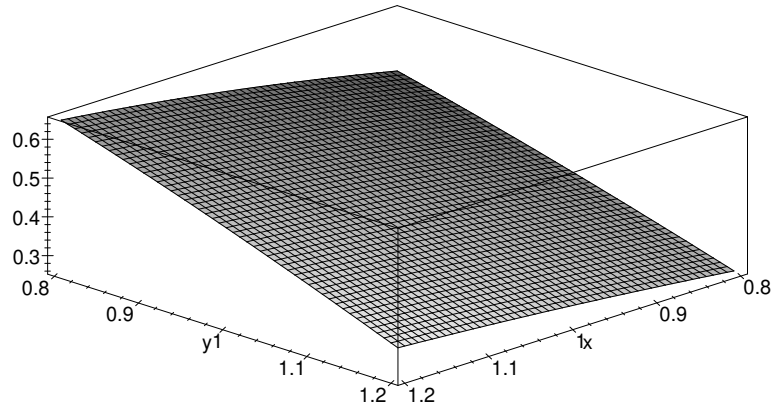


Abb. 35: Nochmalige Vergrößerung um den Faktor fünf um $\big(1, 1, \sin(1)\cos(1)\big)$

erwarten also, daß

$$f(x+h, y+k) = f(x,y) + ah + bk + o(\sqrt{h^2+k^2})$$

ist, und genau so läßt sich Differenzierbarkeit allgemein definieren:

Definition: *a)* Die Funktion $f\colon D \to \mathbb{R}^m$ mit $D \subseteq \mathbb{R}^n$ heißt *differenzierbar* im Punkt $\boldsymbol{x} \in D$, wenn es eine lineare Funktion $L\colon \mathbb{R}^n \to \mathbb{R}^m$

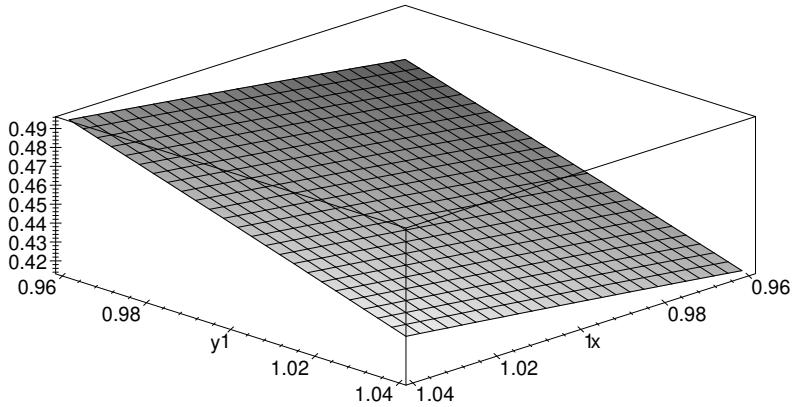


Abb. 36: Nach noch einer Vergrößerung sieht die Funktion praktisch linear aus

gibt, so daß für $\vec{h} \to 0$ gilt

$$f(\boldsymbol{x}+\vec{h}) = f(\boldsymbol{x}) + L\big(\vec{h}\big) + o\big(\|\vec{h}\|\big)\,.$$

b) Die Abbildungsmatrix $J_f(\boldsymbol{x})$ von L bezüglich der Standardbasen von $\mathbb{R}^n$ und $\mathbb{R}^m$ heißt dann JACOBI-Matrix von f im Punkt $\boldsymbol{x}$.

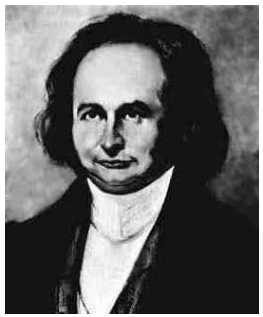
c) Für eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ kann die dann einzeilige JACOBI-Matrix mit einem Vektor aus $\mathbb{R}^n$ identifiziert werden; dieser Vektor

$$\operatorname{grad} f(\boldsymbol{x}) \underset{\text{def}}{=} \nabla f(\boldsymbol{x}) \underset{\text{def}}{=} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}$$

heißt *Gradient* von f im Punkt $\boldsymbol{x} \in D$.

Das Symbol ∇ sieht zwar aus wie ein griechischer Buchstabe, ist aber keiner; es ist ein auf den Kopf gestelltes großes Delta (Δ). ∇f wird „Nabla f“ ausgesprochen nach dem griechischen Wort $\nu\alpha\beta\lambda\alpha$ = Leier; die Bezeichnung wurde eingeführt von dem irischen Mathematiker WILLIAM ROWEN HAMILTON, den die Form von ∇ an eine Leier erinnerte.

WILLIAM ROWEN HAMILTON (1805–1865) wurde in Dublin geboren; bereits mit fünf Jahren sprach er Latein, Griechisch und Hebräisch. Mit dreizehn begann er, mathematische Literatur zu lesen, mit 21 wurde er, noch als Student, Professor der Astronomie am Trinity College in Dublin. Er verlor allerdings schon bald sein Interesse für Astronomie und arbeitete weiterhin auf dem Gebiet der Mathematik und Physik. Am bekanntesten ist seine Entdeckung der Quaternionen 1843, vorher publizierte er aber auch bedeutende Arbeiten über Optik, Dynamik und Algebra.



CARL GUSTAV JACOB JACOBI (1804–51) wurde in Potsdam als Sohn eines jüdischen Bankiers geboren und erhielt den Vornamen Jacques Simon. Im Alter von zwölf Jahren bestand er sein Abitur, mußte aber noch vier Jahre in Abschlußklasse des Gymnasiums bleiben, da die Berliner Universität nur Studenten mit mindestens 16 Jahren aufnahm. 1824 beendete er seine Studien mit dem Staatsexamen für Mathematik, Griechisch und Latein und wurde Lehrer. Außerdem promovierte er 1825 und begann mit seiner Habilitation. Etwa gleichzeitig konvertierte er zum Christentum, so daß er ab 1825 an der Universität Berlin und ab 1826 in Königsberg lehren konnte. 1832 wurde er dort Professor. Zehn Jahre später mußte er aus gesundheitlichen Gründen das rauhe Klima Königsbergs verlassen und war zunächst in Italien, danach für den Rest seines Lebens in Berlin. Er ist vor allem berühmt durch seine Arbeiten zur Zahlentheorie und über elliptische Integrale.

Mit obiger Definition haben wir etwas ähnliches wie die Definition

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

für Funktionen einer Veränderlicher: Auf diese Weise ist die Ableitung zwar *definiert,* aber sie wird – außerhalb von Anfängervorlesungen – praktisch nie so ausgerechnet. Entsprechend sollten wir auch für Funktionen mehrerer Veränderlicher eine effizientere Methode der Differentiation finden.

Am einfachsten wäre es, wenn wir den wohlbekannten Kalkül der Differentialrechnung für Funktionen einer Veränderlicher benutzen könnten, also versuchen wir, aus einer Funktion $f\colon D \to \mathbb{R}^m$ mit $D \subseteq \mathbb{R}^n$ Funktionen einer Veränderlichen zu machen.

Dabei können wir uns sofort auf den Fall $m = 1$ beschränken, denn jede Funktion $f\colon \mathbb{R}^n \to \mathbb{R}^m$ ist zusammengesetzt aus m Komponentenfunktionen $f_i\colon \mathbb{R}^n \to \mathbb{R}$; wir beschränken uns daher zunächst auf Funktionen $f\colon \mathbb{R}^n \to \mathbb{R}$.

Schwieriger ist die Reduktion von n auf eins; hier bietet sich trotz der schlechten Erfahrungen beim obigen Beispiel einer unstetigen Funktion an, die Funktion auf eine Gerade einzuschränken.

Eine Gerade durch einen gegebenen Punkt $\boldsymbol{x}$ ist eindeutig festgelegt durch einen Richtungsvektor $\vec{e}$, wobei umgekehrt der Vektor $\vec{e}$ durch die Gerade natürlich *nicht* eindeutig festgelegt ist: Jedes Vielfache von $\vec{e}$ (außer dem Nullvektor) definiert genau dieselbe Gerade.

Wenn wir die Einschränkung von f auf eine solche Gerade mit Richtungsvektor $\vec{e}$ betrachten, betrachten wir konkret die Funktion

$$g(t) = f(\boldsymbol{x} + t\vec{e})\,,$$

die überall dort definiert ist, wo $\boldsymbol{x} + t\vec{e}$ im Definitionsbereich D von f liegt, für eine offene Menge D also zumindest in einem gewissen offenen Intervall um den Nullpunkt der reellen Geraden.

Damit können wir nach der Differenzierbarkeit dieser Funktion für $t = 0$ fragen; falls sie differenzierbar ist, bezeichnen wir die Ableitung

$$g'(0) = \lim_{h\to 0} \frac{g(h) - g(0)}{h} = \lim_{h\to 0} \frac{f(\boldsymbol{x} + h\vec{e}) - f(\boldsymbol{x})}{h}$$

als *Richtungsableitung* von f in Richtung $\vec{e}$. Eine einfache Anwendung der Kettenregel, die jeder Leser am Rand des Skriptums kurz durchführen sollte, zeigt, daß diese „Richtungsableitung“ nicht nur von der *Richtung* des Vektors $\vec{e}$ abhängt, sondern auch von dessen *Länge:* Beispielsweise ist für $h(t) = f(\boldsymbol{x} + 2t\vec{e})$

$$h'(0) = 2g'(0)\,.$$

Speziell können wir diese Richtungsableitungen betrachten für den Fall, daß $\vec{e}$ ein *Einheitsvektor* ist (genau ist der Grund für die Bezeichnung $\vec{e}$), beispielsweise einer der Koordinateneinheitsvektoren

$$\vec{e}_i = {}^t(0, \ldots, 1, \ldots, 0)\,,$$

bei dem in der i-ten Zeile eine Eins steht und sonst lauter Nullen.

Alsdann ist für $g_i(t) = f(\boldsymbol{x} + t\vec{e}_i)$

$$g_i'(0) = \lim_{h\to 0} \frac{f(\boldsymbol{x} + h\vec{e}_i) - f(\boldsymbol{x})}{h}$$
$$= \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}$$

die Ableitung jener Funktion, die nur von x_i abhängt, während alle anderen Koordinaten x_j festgehalten werden. Diese Ableitung, so sie existiert, bezeichnen wir als *partielle Ableitung*

$$f_{x_i}(\boldsymbol{x}) = \frac{\partial f}{\partial x_i}(\boldsymbol{x})$$

von f nach x_i; das Symbol ∂ wird, wenn überhaupt, als „del“ ausgesprochen, wobei *del* natürlich eine Abkürzung für *delta* ist. Partielle Ableitungen, so sie existieren, lassen sich nach den üblichen Regeln der Differentialrechnung für Funktionen einer Veränderlichen berechnen, sind also für „gutartige“ Funktionen problemlos.

Falls die Funktion f in $\boldsymbol{x} \in D$ differenzierbar ist und D eine Umgebung von $\vec{x}$ enthält (was für offenes D immer der Fall ist), existiert auch jede Richtungsableitung, denn da dann für jeden Vektor $\vec{h}$ gilt

$$f\left(\boldsymbol{x} + \vec{h}\right) = f(\boldsymbol{x}) + L\left(\vec{h}\right) + o\left(\|\vec{h}\|\right),$$

ist insbesondere auch

$$f\left(\boldsymbol{x} + t\vec{e}\right) = f(\boldsymbol{x}) + L\left(t\vec{e}\right) + o\left(\|t\vec{e}\|\right) = f(\boldsymbol{x}) + tL\left(\vec{e}\right) + o(t);$$

denn $L\left(\vec{e}\right)$ und $\left|\vec{h}\right|$ sind schließlich Konstanten. Damit existiert

$$\frac{d}{dt} f\left(\boldsymbol{x} + t\vec{e}\right)_{|t=0} = L\left(\vec{e}\right)$$

für jeden Richtungsvektor $\vec{e}$; insbesondere existieren natürlich alle partiellen Ableitungen.

Die Umkehrung gilt leider nicht immer: Bei der (offensichtlich in $(0,0)$ unstetigen) Funktion

$$f: \begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x,y) & \mapsto \begin{cases} 1 & \text{falls } xy \neq 0 \\ 0 & \text{falls } xy = 0 \end{cases} \end{cases}$$

ist Null auf der x-Achse und der y-Achse, und eins überall sonst. Damit existieren im Punkt $(0,0)$ beide partielle Ableitungen und sind identisch Null. Trotzdem ist f natürlich nicht differenzierbar in $(0,0)$, denn da $f(h,k)$ für jeden Punkt, der nicht auf einer der beiden Koordinatenachsen liegt, gleich eins ist, kann es keine Linearform $L(h,k) = bh+ck$ geben, so daß

$$f(h,k) = f(0,0) + L(h,k) + o\left(\sqrt{h^2+k^2}\right) = o\left(\sqrt{h^2+k^2}\right)$$

ist, denn $L(0,0) = 0$ und $f(h,k) = 1$ für $hk \neq 0$.

Nun wird natürlich jeder vernünftige Mensch einwenden, daß dieses Beispiel sehr künstlich ist, und in der Tat verhalten sich „gutartige“ Funktionen nicht so:

Lemma: Falls $f: D \to \mathbb{R}$ auf der offenen Teilmenge $D \subseteq \mathbb{R}^n$ stetig ist und auch alle partiellen Ableitungen von f dort existieren und stetig sind, ist f in D differenzierbar und

$$\operatorname{grad} f(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}.$$

Beweis: Sei $\boldsymbol{x} \in D$. Da D offen ist, gibt es eine Kugel um $\boldsymbol{x}$, die ganz in D liegt; wir betrachten im folgenden nur Vektoren $\vec{h}$, deren Länge höchstens gleich dem Radius dieser Kugel ist, so daß $\boldsymbol{x} + \vec{h}$ stets in D liegt.

Wir betrachten f zunächst nur als Funktion der ersten Variablen; da deren Ableitung f_{x_1} in ganz D existiert, ist

$$f(\boldsymbol{x} + \vec{h}) = f(x_1 + h_1, \ldots, x_n + h_n)$$
$$= f(x_1, x_2 + h_2, \ldots, x_n + h_n) + f_{x_1}(x_1, x_2 + h_2, \ldots, x_n + h_n)h_1 + o(h_1).$$

Genauso ist, da die partielle Ableitung nach x_2 in ganz D existiert,

$$f(x_1, x_2 + h_2, \ldots, x_n + h_n) =$$
$$f(x_1, x_2, x_3, \ldots, x_n + h_n) + f_{x_2}(x_1, x_2, x_3 + h_3 \ldots, x_n + h_n)h_2 + o(h_2)$$

und so weiter. Insgesamt erhalten wir

$$\begin{aligned} f(\boldsymbol{x} + \vec{h}) = f(\boldsymbol{x}) &+ f_{x_1}(x_1, x_2 + h_2, x_3 + h_3, \ldots x_n + h_n)h_1 + o(h_1) \\ &+ f_{x_2}(x_1, x_2, x_3 + h_3, \ldots, x_n + h_n)h_2 + o(h_2) \\ &\vdots \\ &+ f_{x_{n-1}}(x_1, \ldots, x_{n-1}, x_n + h_n)h_{n-1} + o(h_{n-1}) \\ &+ f_{x_n}(x_1, \ldots, x_n)h_n + o(h_n). \end{aligned}$$

Die LANDAU-Symbole $o(h_1), \dots, o(h_n)$ können wir zu $o\left(\|\vec{h}\|\right)$ zusammenfassen, denn da $\left|h_i\right| \leq \left|\vec{h}\right|$ für jedes i, kann keines der h_i langsamer gegen Null gehen als $\left|\vec{h}\right|$.

Damit sind wir schon ziemlich nahe an dem, was wir für die Differenzierbarkeit brauchen; allerdings hängen die partiellen Ableitungen noch von den h_i ab, so daß die Differenz zwischen $f(\boldsymbol{x}+\vec{h})$ und $f(\boldsymbol{x})$ nicht durch eine lineare Funktion angenähert ist.

Hier kommt nun die Stetigkeit der partiellen Ableitungen ins Spiel: Diese impliziert, daß

$$\lim_{\vec{h}\to 0}\Big(f_{x_1}(x_1, x_2+h_2, \dots, x_n+h_n) - f_{x_1}(x_1, x_2, \dots, x_n)\Big) = 0$$

ist. Damit ist

$$\Big(f_{x_1}(x_1, x_2+h_2, \dots, x_n+h_n) - f_{x_1}(x_1, x_2, \dots, x_n)\Big)h_1 = o\left(\|\vec{h}\|\right),$$

denn wenn h_1 mit einem Ausdruck multipliziert wird, der gegen Null geht, strebt das Produkt für $\vec{h} \to 0$ schneller gegen Null als h_1 allein, und $o(h_1)$ kann durch $o\left(\|\vec{h}\|\right)$ abgeschätzt werden. Also ist

$$\begin{aligned} &f_{x_1}(x_1, x_2+h_2, \dots, x_n+h_n)h_1 \\ =&f_{x_1}(x_1, x_2, \dots, x_n)h_1 + o\left(\|\vec{h}\|\right) = f_{x_1}(\boldsymbol{x})h_1 + o\left(\|\vec{h}\|\right). \end{aligned}$$

Entsprechend können wir auch bei den übrigen partiellen Ableitungen argumentieren und erhalten insgesamt, daß

$$\begin{aligned} f(\boldsymbol{x}+\vec{h}) &= f(\boldsymbol{x}) + f_{x_1}(\boldsymbol{x})h_1 + \cdots + f_{x_n}(\boldsymbol{x})h_n + o\left(\|\vec{h}\|\right) \\ &= f(\boldsymbol{x}) + \begin{pmatrix} \dfrac{\partial f}{\partial x_1} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{pmatrix} \cdot \vec{h} + o\left(\|\vec{h}\|\right) \end{aligned}$$

ist. Damit ist das Lemma bewiesen. ∎

Für Funktionen mit stetigen partiellen Ableitungen ist der Gradient also gerade der Vektor der partiellen Ableitungen; er kann damit über die bekannten Ableitungsregeln für Funktionen einer Veränderlichen berechnet werden.

NB: Häufig wird der Gradient durch diese Formel *definiert;* in diesem Fall folgt natürlich aus der Existenz des Gradienten nicht die Differenzierbarkeit der Funktion; siehe obiges Beispiel einer unstetigen Funktion, für die alle partiellen Ableitungen in $(0,0)$ existieren.

Nächstes Ziel dieses Abschnitts sind höhere Ableitungen von Funktionen mehrerer Veränderlicher. Wir lassen uns bei der Definition wieder vom eindimensionalen Fall leiten: Die zweite Ableitung ist die Ableitung der Ableitung.

Die Ableitung einer differenzierbaren Funktion $f\colon D \to \mathbb{R}$ mit $D \subseteq \mathbb{R}^n$ ist der Zeilenvektor zum Gradienten, also die Abbildung

$$D \to \mathbb{R}^n; \quad \boldsymbol{x} \mapsto \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right);$$

die Ableitung oder JACOBI-Matrix davon ordnet jedem Punkt aus D eine $n \times n$-Matrix zu, die wir als HESSE-Matrix $H_f(\boldsymbol{x})$ von f im Punkt $\boldsymbol{x}$ bezeichnen.

LUDWIG OTTO HESSE (1811–1874) wurde in Königsberg geboren und unterrichtete zunächst Physik und Chemie am dortigen Gymnasium. 1840 bekam er eine Stelle als Mathematiker an der dortigen Universität, von 1856 bis 1868 war er Professor in Heidelberg, danach in München. Aus der Schule ist er wohl vor allem durch die HESSEsche Normalenform der Ebenengleichung bekannt; der Schwerpunkt seiner Forschungen lag allerdings auf dem Gebiet der Invariantentheorie und der algebraischen Funktionen. Auch die HESSE-Matrix führte er 1842 in einer Arbeit über Invarianten von kubischen und biquadratischen Kurven ein.

Genau wie der Gradient bei gutartigen Funktionen durch die partiellen Ableitungen berechnet werden kann, sollte auch die HESSE-Matrix durch Differentiationsverfahren aus der Analysis einer Veränderlichen berechenbar sein. Das Hilfsmittel dazu sind die zweiten partiellen Ableitungen:

Für eine in ganz D partiell differenzierbare Funktion $f\colon D \to \mathbb{R}$ ist auch jede partielle Ableitung f_{x_i} wieder eine Funktion von D nach $\mathbb{R}$, und auch diese kann wieder partiell differenzierbar sein. Falls ja, bezeichnen wir die partielle Ableitung von f_{x_i} nach x_j als zweite partielle Ableitung

$$f_{x_i x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \underset{\text{def}}{=} \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)$$

von f nach x_i und x_j. Im Fall $i = j$ schreiben wir kurz

$$f_{x_i x_i} = \frac{\partial^2 f}{\partial x_i^2}.$$

Analog lassen sich auch höhere partielle Ableitungen einführen durch die Definition

$$f_{x_{i_1}x_{i_2}\dots x_{i_k}} \underset{\text{def}}{=} \frac{\partial^k f}{\partial x_{i_k}\dots\partial x_{i_2}\partial x_{i_1}} \underset{\text{def}}{=} \frac{\partial}{\partial x_{i_k}}\frac{\partial}{\partial x_{i_{k-1}}}\cdots\frac{\partial}{\partial x_{i_2}}\frac{\partial f}{\partial x_{i_1}} .$$

Wie wir oben gesehen haben, ist Gradient einer Funktion f gleich dem Vektor der partiellen Ableitungen, falls diese allesamt existieren und stetig sind; die JACOBI-Matrix ist der entsprechende Zeilenvektor. Falls auch die zweiten partiellen Ableitungen allesamt existieren und stetig sind, zeigt dasselbe Lemma, daß deren Ableitungen die Zeilenvektoren

$$\left(\frac{\partial^2 f}{\partial x_i\partial x_1},\dots,\frac{\partial^2 f}{\partial x_i\partial x_n}\right.$$

sind, d.h.

Lemma: Falls alle ersten und zweiten partiellen Ableitungen von $f: D \to \mathbb{R}$ existieren und stetig sind, ist die HESSE-Matrix von f gleich der $n \times n$-Matrix mit Einträgen $\frac{\partial^2 f}{\partial x_i \partial x_j}$.

∎

Als erstes Beispiel können wir etwa die zweiten partiellen Ableitungen der Funktion

$$f(x,y) = x^4 + 2x^3y + 3x^2y^2 + 4xy^3 + 5y^4$$

berechnen: Die partielle Ableitung nach x ist

$$f_x(x,y) = 4x^3 + 6x^2y + 6xy^2 + 4y^3 ,$$

also ist

$$f_{xx}(x,y) = \frac{\partial^2 f}{\partial x^2}(x,y) = 12x^2 + 12xy + 6y^2 \quad \text{und}$$

$$f_{xy}(x,y) = \frac{\partial^2 f}{\partial y\partial x}(x,y) = 6x^2 + 12xy + 12y^2 .$$

Entsprechend ist

$$f_y(x,y) = 2x^3 + 6x^2y + 12xy^2 + 20y^3 ,$$

also

$$f_{yx}(x,y) = \frac{\partial^2 f}{\partial x\partial y}(x,y) = 6x^2 + 12xy + 12y^2 \quad \text{und}$$

$$f_{yy}(x,y) = \frac{\partial^2 f}{\partial y^2}(x,y) = 6x^2 + 24xy + 60y^2 .$$

Damit ist

$$H_f(x,y) = \begin{pmatrix} 12x^2 + 12xy + 6y^2 & 6x^2 + 12xy + 12y^2 \\ 6x^2 + 12xy + 12y^2 & 6x^2 + 24xy + 60y^2 \end{pmatrix} .$$

Zumindest in diesem Fall ist dies eine symmetrische Matrix, d.h.

$$f_{xy} = f_{yx} .$$

Diese Formel gilt, wie wir gleich sehen werden, *fast* immer; in der Tat galt sie für die Mathematiker des 18. Jahrhunderts wie NICOLAUS I. BERNOULLI, der 1719 darüber schrieb, LEONARD EULER (1730), JOSEPH-LOUIS LAGRANGE (1772) und viele andere als selbstverständlich. Erst im 19. Jahrhundert, als sich ein präziser Funktionsbegriff durchzusetzen begann, wurde erkannt, daß Voraussetzungen notwendig sind. Diese waren zu Beginn des Jahrhunderts zunächst unnötig stark; 1873 untersuchte dann HERMANN AMANDUS SCHWARZ, den wir bereits von der CAUCHY-SCHWARZschen Ungleichung kennen, das Problem genauer in seiner Arbeit *Über ein System voneinander unabhängiger Voraussetzungen zum Beweis des Satzes* $\frac{\partial}{\partial y}\left(\frac{\partial f(x,y)}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial f(x,y)}{\partial y}\right)$.

Als Gegenbeispiel betrachtet er die in Abbildung 37 dargestellte Funktion

$$f(x,y) = \begin{cases} y^2 \arctan\frac{x}{y} - x^2 \arctan\frac{y}{x} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases} .$$

Ihre ersten partiellen Ableitungen sind

$$f_x(x,y) = \begin{cases} y - 2x \arctan\frac{y}{x} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases}$$

und

$$f_y(x,y) = \begin{cases} -x + 2y \arctan\frac{x}{y} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases} .$$

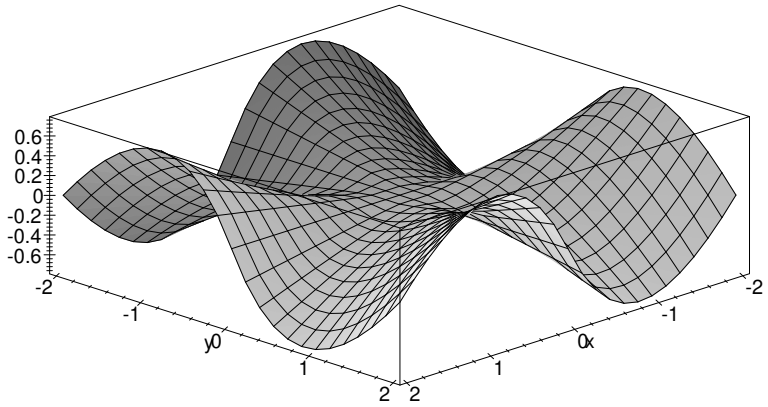


Abb. 37: Ein Gegenbeispiel zum Vertauschungssatz

Alle diese Funktionen sind stetig auch im Nullpunkt, denn da der Arkustangens nur Werte zwischen -1 und $+1$ annimmt, sorgt der jeweilige Vorfaktor dafür, daß das Produkt bei jeder Annäherung an den Nullpunkt gegen null geht. Daher können wir trotz der Nenner im Argument des Arkustangens $x = 0$ und $y = 0$ einsetzen und erhalten, daß

$$f_x(0, y) = y \quad \text{und} \quad f_y(x, 0) = -x$$

ist, also

$$f_{xy}(0, y) = +1 \quad \text{und} \quad f_{yx}(x, 0) = -1\,.$$

Im Nullpunkt ist somit $f_{xy}(0, 0) = +1 \neq -1 = f_{yx}(0, 0)$.

Für Punkte $(x, y) \neq (0, 0)$ rechnet man leicht nach, daß

$$f_{xy}(x, y) = f_{yx}(x, y) = \frac{y^2 - x^2}{y^2 + x^2}$$

ist. Insbesondere ist

$$f_{xy}(x, 0) = -1, \quad f_{xy}(0, y) = +1 \quad \text{und} \quad f_{xy}(x, x) = 0\,;$$

f_{xy} nimmt also in jeder noch so kleinen Umgebung des Nullpunkts jeden der drei Werte $0, 1$ und -1 (und viele andere) an. Damit kann f_{xy} in $(0, 0)$ nicht stetig sein, genausowenig wie f_{yx}. Wie SCHWARZ erkannte, ist genau das die fehlende Voraussetzung für die Vertauschbarkeit der partiellen Ableitungen:

Schwarzsches Lemma: $f\colon D \to \mathbb{R}$ sei auf der offenen Menge $D \subseteq \mathbb{R}^n$ erklärt und sowohl die ersten partiellen Ableitungen f_{x_i} als auch die gemischten partiellen Ableitungen $f_{x_i x_j}$ seien stetig auf D. Dann ist

$$f_{x_i x_j}(\boldsymbol{x}) = f_{x_j x_i}(\boldsymbol{x})$$

für alle $\boldsymbol{x} \in D$ und alle i, j mit $1 \leq i, j \leq n$.

Beweis: Da bei der partiellen Differentiation alle Variablen außer einer als konstant betrachtet werden, können wir uns auf den Fall $n = 2$ beschränken: Wir interessieren uns nur für die beiden Variablen x_i und x_j, die wir als x und y bezeichnen (wenn sie verschieden sind – andernfalls gibt es aber ohnehin nichts zu beweisen), und betrachten alle sonstigen x_k als konstant.

Für den Punkt (x, y) aus D wählen wir dann $h, k \in \mathbb{R}$ so, daß das Quadrat mit den vier Ecken

$$(x, y),\ (x + h, y),\ (x, y + k) \quad \text{und} \quad (x + h, y + k)$$

vollständig in D liegt; dies ist möglich, da wir D als offene Menge vorausgesetzt haben. Nach Voraussetzung existieren die partiellen Ableitungen f_x, f_y, f_{xy} sowie f_{yx} und sind stetig.

Nach Definition ist

$$f_{xy}(x, y) = \lim_{k \to 0} \frac{f_x(x, y + k) - f_x(x, y)}{k}$$

$$= \lim_{k \to 0} \frac{\lim\limits_{h \to 0} \dfrac{f(x + h, y + k) - f(x, y + k)}{h} - \lim\limits_{h \to 0} \dfrac{f(x + h, y) - f(x, y)}{h}}{k}\,.$$

Falls alle Grenzübergänge miteinander vertauschbar sind (was, wie wir im obigen Gegenbeispiel gesehen haben, keineswegs selbstverständlich ist), ist das ein Limes über den Ausdruck

$$\frac{f(x + h, y + k) - f(x, y + k) - f(x + h, y) + f(x, y)}{hk}$$

für $h, k \to 0$; es liegt also nahe, sich diesen Ausdruck genauer anzuschauen.

Für den Beweis wird es genügen, wenn wir uns auf den Fall $h = k$ beschränken; wir werden den Ausdruck

$$D(h) = \frac{f(x+h, y+h) - f(x, y+h) - f(x+h, y) + f(x, y)}{h^2}$$

auf zwei Arten ausrechnen:

Zunächst fassen wir, wie oben, die beiden ersten und die beiden letzten Summanden zusammen: Mit der Abkürzung

$$g(y) = \frac{f(x+h, y) - f(x, y)}{h}$$

ist dann

$$D(h) = \frac{g(y+h) - g(y)}{h}\,.$$

Nach dem Mittelwertsatz der Differentialrechnung ist dieser Differenzenquotient gleich dem Differentialquotient $g'(\eta)$ für eine (von h abhängige) Zahl η zwischen y und $y+h$. Somit ist

$$D(h) = g'(\eta) = \frac{f_y(x+h, \eta) - f_y(x, \eta)}{h} = f_{yx}(\xi, \eta)$$

für ein η zwischen x und $x+h$, denn natürlich können wir auch auf diesen Differenzenquotienten den Mittelwertsatz anwenden.

Für die zweite Berechnung fassen wir in $D(h)$ den ersten und den dritten sowie den zweiten und den vierten Term zusammen. Mit der Abkürzung

$$\widetilde{g}(x) = \frac{f(x, y+h) - f(x, y)}{h}$$

ist dann dieses Mal

$$D(h) = \frac{\widetilde{g}(x+h) - \widetilde{g}(y)}{h}\,,$$

und nach dem Mittelwertsatz der Differentialrechnung gibt es dazu ein $\widetilde{\xi}$ zwischen x und $x+h$, so daß dies gleich $\widetilde{g}'(\widetilde{\xi})$ ist. Also ist

$$D(h) = \widetilde{g}'(\widetilde{\xi}) = \frac{f_x(\widetilde{\xi}, y+h) - f_x(\widetilde{\xi}, y)}{h} = f_{xy}(\widetilde{\xi}, \widetilde{\eta})$$

für eine Zahl $\widetilde{\eta}$ zwischen y und $y+h$. Somit haben wir gezeigt, daß

$$D(h) = f_{yx}(\xi, \eta) = f_{xy}(\widetilde{\xi}, \widetilde{\eta})$$

ist.

Lassen wir nun h gegen Null gehen, konvergieren ξ und $\widetilde{\xi}$ gegen x und η wie auch $\widetilde{\eta}$ gegen y. Wegen der vorausgesetzten Stetigkeit der zweiten partiellen Ableitungen konvergiert daher $f_{yx}(\xi, \eta)$ gegen $f_{yx}(x, y)$ und $f_{xy}(\widetilde{\xi}, \widetilde{\eta})$ gegen $f_{xy}(x, y)$, d.h. der Grenzwert existiert und

$$D(0) = f_{yx}(x, y) = f_{xy}(x, y)\,.$$

Damit ist das Lemma bewiesen. ∎

Tatsächlich bewies SCHWARZ diesen Satz (für $n = 2$) unter einer etwas schwächeren Voraussetzung: Es reicht, wenn *eine* der partiellen Ableitungen f_{xy} oder f_{yx} existiert und stetig ist. Am Beweis ändert sich wenig; falls etwa über die Ableitung f_{yx} nichts vorausgesetzt ist, muß man die Existenz aller damit zusammenhängenden Grenzwerte explizit durch Abschätzungen nachweisen und daraus nachträglich die Existenz und Stetigkeit von f_{yx} folgern. Ein Leser, der seine *Analysis I* noch nicht ganz vergessen hat, sollte dies auf etwa einer Seite tun können. Für Anwendungen ist die SCHWARZsche Formulierung etwas nützlicher als die obige, denn wenn man beispielsweise f_{xy} berechnet und seine Stetigkeit nachgewiesen hat, folgt automatisch, daß auch f_{yx} existiert und gleich f_{xy} ist. Für uns wird das keine sehr große Rolle spielen, denn bei den meisten uns interessierenden Funktionen wird die *Existenz* und Stetigkeit der partiellen Ableitungen klar sein; lediglich ihre Berechnung wird im allgemeinen mit Arbeit verbunden sein.

Ein analoger Satz zum SCHWARZschen Lemma gilt auch für höhere partielle Ableitungen; für k-fache Ableitungen müssen wir entsprechend voraussetzen, daß die partiellen Ableitungen bis zur k-fachen existieren und stetig sind (wobei diese Voraussetzung wieder strenggenommen nicht für alle k-fachen wirklich notwendig ist).

Definition: Für eine offene Teilmenge $D \subseteq \mathbb{R}^n$ bezeichne $\mathcal{C}^k(D, \mathbb{R})$ die Menge aller Funktionen $f\colon D \to \mathbb{R}$, deren sämtliche partielle Ableitungen bis zu den k-ten existieren und stetig sind. Für $k = 0$ sei $\mathcal{C}^0(D, \mathbb{R})$ einfach die Menge aller stetiger Funktionen $D \to \mathbb{R}$.

Man überlegt sich sofort, daß $\mathcal{C}^k(D, \mathbb{R})$ ein $\mathbb{R}$-Vektorraum ist, und es ist auch nicht schwer einzusehen, daß die Funktionen aus $\mathcal{C}^k(D, \mathbb{R})$ alle die Eigenschaften haben, die man bei der Betrachtung von k-ten Ableitungen wünscht:

Erstens ist die Berechnung einer k-ten partiellen Ableitung von der Reihenfolge der partiellen Differentiationen unabhängig: Wie wir aus Kapitel I, §5*d)* wissen, kann jede Permutation als Produkt von Transpositionen geschrieben werden; es genügt also zu zeigen, daß man die Reihenfolge zweiter partieller Differentiationen vertauschen kann, und die können auch noch als benachbart angenommen werden, denn eine beliebige Transposition läßt sich immer als Produkt von Transpositionen benachbarter Elemente schreiben. Für eine solche Transposition ist

$$\frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_r}} \frac{\partial}{\partial x_{i_{r+1}}} \cdots \frac{\partial f}{\partial x_{i_k}} = \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_{r+1}}} \frac{\partial}{\partial x_{i_r}} \cdots \frac{\partial f}{\partial x_{i_k}} .$$

Für $f \in \mathcal{C}^k(D, \mathbb{R})$ ist sichergestellt, daß

$$\frac{\partial}{\partial x_{i_{r+2}}} \cdots \frac{\partial f}{\partial x_{i_k}} \in \mathcal{C}^2(D, \mathbb{R}) ;$$

nach obigem Lemma ist daher

$$\frac{\partial}{\partial x_{i_r}} \frac{\partial}{\partial x_{i_{r+1}}} \left(\frac{\partial}{\partial x_{i_{r+2}}} \cdots \frac{\partial f}{\partial x_{i_k}} \right) = \frac{\partial}{\partial x_{i_{r+1}}} \frac{\partial}{\partial x_{i_r}} \left(\frac{\partial}{\partial x_{i_{r+2}}} \cdots \frac{\partial f}{\partial x_{i_k}} \right) .$$

Differenziert man hier beide Seiten noch partiell nach x_1 bis x_{r-1}, ändert dies natürlich nichts an der Gleichheit.

Zweitens besagt das vorletzte Lemma, daß eine Funktion $f \in \mathcal{C}^1(D, \mathbb{R})$ differenzierbar ist. Eine naheliegende Verallgemeinerung des dortigen Beweises, bei der man anstelle von linearen Approximationen solche höherer Ordnung betrachtet, zeigt, daß eine Funktion $f \in \mathcal{C}^2(D, \mathbb{R})$ zweifach differenzierbar ist, und daß entsprechend eine Funktion f aus $\mathcal{C}^k(D, \mathbb{R})$ bis auf einen Fehler der Größenordnung $o\big(|\vec{h}|^k\big)$ durch ein Polynom k-ten Grades approximiert werden kann. Wie das im einzelnen aussieht, wollen wir uns im nächsten Abschnitt genauer anschauen.

d) Taylor-Reihen

Für Funktionen einer reellen Veränderlichen mit Werten in $\mathbb{R}$ kennen wir TAYLOR-Reihen aus der Analysis I. Hier gilt

Satz: $f\colon [a, b] \to \mathbb{R}$ sei stetig und mindestens $(n+1)$-fach stetig differenzierbar auf dem Intervall $(a, b) \subseteq \mathbb{R}$. Dann gilt für jedes x aus (a, b)

und jedes $h \in \mathbb{R}$ mit $x + h \in (a, b)$ die Formel

$$\begin{aligned} f(x+h) &= f(x) + h f'(x) + \frac{h^2}{2} f''(x) + \cdots + \frac{h^n}{n!} f^{(n)}(x) + R_{n+1}(x, h) \\ &= \sum_{i=0}^{n} \frac{h^i}{i!} f^{(i)}(x) + R_{n+1}(x, h) \end{aligned}$$

mit einem Restglied $R_{n+1} = O(h^{n+1})$. Dieses kann beispielsweise dargestellt werden als

$$R_{n+1}(x, h) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(x + \eta)$$

mit einer reellen Zahl η zwischen 0 und h.

Definition: *a)* $T_{f,x,n}(h) = \sum_{i=0}^{n} \frac{h^i}{i!} f^{(i)}(x)$ heißt TAYLOR-Polynom n-ten Grades von f um den Punkt x.
b) $R_{n+1}(x, h) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(x + \eta)$ ist die LAGRANGEsche Form des Restglieds.
c) Die Reihe $T_{f,x}(h) = \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(x) h^i$ heißt TAYLOR-Reihe von f um x.
d) Die Funktion f heißt *analytisch* an der Stelle x, wenn es ein $\delta > 0$ gibt, so daß die TAYLOR-Reihe $T_{f,x}(h)$ existiert und für alle h mit $|h| < \delta$ gegen $f(x + h)$ konvergiert.

BROOK TAYLOR (1685–1731) war Sohn wohlhabender Eltern und wurde daher, bevor er 1703 an die Universität Cambridge ging, nur von privaten Hauslehrern ausgebildet. In Cambridge beschäftigte er sich hauptsächlich mit Mathematik, woran sich auch nach seinem Studienabschluß nichts änderte. Sein 1715 erschienenes Buch *Methodus incrementorum directa et inversa* enthält unter anderem TAYLOR-Polynome (die in Spezialfällen bereits LEIBNIZ, NEWTON und anderen bekannt waren), sowie die Methode der partiellen Integration. Weitere Bücher und Arbeiten beschäftigen sich unter anderem mit der Perspektive sowie mit Fragen aus der Physik.

Zur Verallgemeinerung der TAYLOR-Reihen auf Funktionen mehrerer Veränderlicher verwenden wir Richtungsableitungen. Da wir eine Funk-

tion $f:\mathbb{R}^n \to \mathbb{R}^m$ problemlos komponentenweise behandeln können, genügt es, den Fall $m = 1$ zu betrachten.

Die Richtungsableitung einer Funktion $f:\mathbb{R}^n \to \mathbb{R}$ in Richtung $\vec{v}$ ist nach Definition die Ableitung der Funktion einer Veränderlichen

$$g: \begin{cases} (-a,a) & \to \mathbb{R} \\ t & \mapsto f(\boldsymbol{x} + t\vec{v}) \end{cases}$$

mit geeignet gewähltem $a \in \mathbb{R}_+$ nach t für $t = 0$; wir können sie berechnen als Skalarprodukt des Gradienten mit dem Vektor $\vec{v}$.

Natürlich können wir g nicht nur einmal ableiten; für $f \in \mathcal{C}^k(D,\mathbb{R})$ existiert das TAYLOR-Polynom k-ten Grades

$$g(h) = g(0) + h \cdot g'(0) + \frac{h^2}{2} g''(0) + \cdots + \frac{h^k}{k!} g^{(k)}(0) + o(h^k)\,.$$

Führen wir nun für die Richtungsableitung in Richtung $\vec{v}$ einen Operator $\partial_{\vec{v}}$ ein durch

$$\partial_{\vec{v}} f(x) = \operatorname{grad} f(x) \cdot \vec{v} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x}) \cdot v_i\,,$$

so wird dies zu

$$f(\boldsymbol{x} + h\vec{v}) = f(\boldsymbol{x}) + h\partial_{\vec{v}} f(\boldsymbol{x}) + \cdots + \frac{h^k}{k!} \partial_{\vec{v}}^k f(\boldsymbol{x}) + o(h^k)\,.$$

Speziell für $h = 1$ erhalten wir die kompakte Schreibweise der TAYLOR-Formel, nämlich

$$f(\boldsymbol{x} + \vec{v}) = f(\boldsymbol{x}) + \partial_{\vec{v}} f(\boldsymbol{x}) + \cdots + \frac{1}{k!} \partial_{\vec{v}}^k f(\boldsymbol{x}) + o\left(\left|\vec{v}\right|^k\right)\,.$$

$\partial_{\vec{v}}^k$ steht hierbei natürlich für die k-fache Anwendung des Operators $\partial_{\vec{v}}$; was das explizit bedeutet, sollte man sich anhand obiger Summendarstellung von $\partial_{\vec{v}}$ klarmachen. Insbesondere beachte man, daß alle Potenzprodukte der v_i in den Potenzen des Operatiors $\partial_{\vec{v}}$ stecken und daß diese nach Ausmultiplizieren schnell ziemlich unübersichtlich werden. Einige Leser werden dieses Phänomen wohl aus der Physik kennen: Die *Multipolentwicklung* eines elektrischen Feldes um einen Punkt ist nichts anderes als die TAYLOR-Entwicklung; der konstante Term hängt

ab von der elektrischen Ladung, der lineare Term vom Dipolmoment, der quadratische vom Quadrupolmoment *usw.*

Wie das für beliebiges n im einzelnen aussieht, wollen wir uns lieber nicht so genau anschauen: Eine Form vom Grad k ist gegeben durch eine k-fache Summation von 1 bis n, hat also n^k Summanden, was sehr schnell sehr unübersichtlich wird.

Der quadratische Term der TAYLOR-Reihe ist allerdings noch einigermaßen gut handhabbar: Dafür brauchen wir nur die HESSE-Matrix

$$H_f(\boldsymbol{x}) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{pmatrix}\,,$$

und nach dem Lemma von SCHWARZ ist diese für $f \in \mathcal{C}^2(D,\mathbb{R})$ eine symmetrische Matrix, was den Rechenaufwand noch einmal fast um die Hälfte reduziert.

e) Der Satz über implizite Funktionen

Der Zusammenhang zwischen zwei Größen x und y ist nicht immer explizit in der Form $y = f(x)$ gegeben; gelegentlich hat man auch nur einen impliziten Zusammenhang $F(x,y) = 0$; entsprechend auch für mehr als zwei Variablen. In diesem Abschnitt soll untersucht werden, wann eine Gleichung der Form $F(\boldsymbol{x}) = 0$ nach einer der Variablen x_i aufgelöst werden kann.

In einfachen Fällen ist dies trivial möglich, beispielsweise läßt sich

$$F(x,y,z) = ax + by + cz = 0$$

für $c \neq 0$ durch

$$z = \frac{-ax - by}{c}$$

nach z auflösen. In etwas komplizierteren Fällen, wie etwa bei

$$F(x,y) = x^2 + y^2 - 1 = 0\,,$$

kann man für die Punkte, die nicht auf der x-Achse $y = 0$ liegen, zumindest lokal eindeutig explizit auflösen durch

$$y = \pm\sqrt{1 - x^2}\,,$$

wobei das Vorzeichen gleich dem von y im betrachteten Intervall ist.

Im allgemeinen gibt es jedoch keine Möglichkeit für eine explizite Auflösung mit den „üblichen“ mathematischen Funktionen, d.h. man kann höchstens dann auflösen, wenn man neue Funktionen einführt.

Wie das Beispiel der Kreislinie zeigt, ist auch das nicht immer möglich: Für die beiden Punkte auf der x-Achse gibt es offensichtlich keine *eindeutige* Auflösung, da sowohl die positive wie auch die negative Wurzel Teilauflösungen sind.

Diese Existenz mehrerer Teilauflösungen hängt mit dem Verschwinden der partiellen Ableitung nach y zusammen: Falls diese partielle Ableitung ungleich Null ist, gibt sie an, wie sich F verändert, wenn man y ändert, und sie gibt damit zumindest in erster Näherung auch an, wie man y verändern muß, um bei einer Änderung von x die Bedingung $F(x,y) = 0$ zu erhalten. Der Satz über implizite Funktionen besagt, das dieses Nichtverschwinden der partiellen Ableitung bereits ausreicht um die Existenz einer eindeutigen Auflösung zu zeigen.

Um den Beweis wenigstens einigermaßen überschaubar zu halten, möchte ich mich zunächst auf Funktionen zweier Veränderlicher beschränken:

Satz: $D \subseteq \mathbb{R}^2$ sei offen und $F \in \mathcal{C}^1(D,\mathbb{R})$. Dann gibt es für jeden Punkt $(x_0,y_0) \in D$ mit $F(x_0,y_0) = 0$ und $F_y(x_0,y_0) \neq 0$ Intervallumgebungen I von x_0 und K von y_0 sowie eine eindeutig bestimmte Funktion $f\colon I \to K$, so daß für alle $x \in I$ gilt:

$$F\big(x, f(x)\big) = 0\,.$$

Die Funktion f ist stetig und differenzierbar; ihre Ableitung ist

$$f'(x) = -\frac{F_x(x,y)}{F_y(x,y)} \quad \text{mit} \quad y = f(x)\,.$$

Beweis: Wir beginnen mit einer Reduktion zwecks Vereinfachung der Schreibarbeit: Offensichtlich genügt es, wenn wir den Fall $x_0 = y_0 = 0$ zu betrachten. Gilt nämlich der Satz für die Funktion

$$G(x,y) = F(x + x_0, y + y_0)$$

im Punkt $(0,0)$, so folgt er sofort auch für F im Punkt (x_0,y_0). Außerdem können wir o.B.d.A. annehmen, daß $F_y(0,0)$ positiv ist, denn nach Voraussetzung ist dieser Wert ungleich Null, und falls er negativ sein sollte, ersetzen wir einfach F durch $-F$.

Nach Voraussetzung sind die partiellen Ableitungen von F stetig; daher ist F_y nicht nur im Nullpunkt positiv, sondern auch noch in einer gewissen Umgebung davon. In dieser Umgebung wählen wir ein Rechteck

$$\big\{(x,y) \in \mathbb{R}^2 \;\big|\; -\alpha \le x \le \alpha \quad \text{und} \quad -\beta \le y \le \beta\big\}\,.$$

Da F_y auf diesem Rechteck überall positiv ist, wächst die Funktion $y \mapsto F(x_0,y)$ für jedes $x_0 \in [-\alpha,\alpha]$ streng monoton; wegen $F(0,0) = 0$ ist insbesondere $F(0,-\beta) < 0$ und $F(0,\beta) > 0$. Aufgrund der Stetigkeit von F ist damit für x_1 aus einer gewissen Umgebung der Null auch $F(x_1,-\beta) < 0$ und $F(x_1,\beta) > 0$. Indem wir nötigenfalls α noch etwas verkleinern, können wir annehmen, daß dies für alle $x_1 \in [-\alpha,\alpha]$ gilt.

Damit gibt es nach dem Zwischenwertsatz für jedes $x_1 \in [-\alpha,\alpha]$ ein y_1, so daß $F(x_1,y_1)$ verschwindet; wegen der strengen Monotonie der Funktion $y \mapsto F(x_1,y)$ ist dieser Wert y_1 eindeutig bestimmt. Wir setzen daher $I = (-\alpha,\alpha)$, $K = (-\beta,\beta)$ und

$$f\colon \begin{cases} I & \to K \\ x_1 & \mapsto y_1 \end{cases}\,.$$

Damit ist f als Funktion festgelegt, und nach Konstruktion ist

$$F\big(x, f(x)\big) = 0 \quad \text{für alle } x \in I\,.$$

Wir müssen uns noch überlegen, daß die so konstruierte Funktion f stetig und differenzierbar ist.

Die Stetigkeit können wir etwa dadurch nachweisen, daß wir für jede gegen ein $x \in I$ konvergierende Folge (x_n) aus I zeigen, daß

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

ist.

Da x in I liegt, wissen wir, daß es dazu ein eindeutig bestimmtes $y \in K$ gibt, so daß $F(x,y) = 0$ ist, nämlich $y = f(x)$. Für jedes $\varepsilon > 0$ gibt es daher ein $\delta > 0$, so daß $\big|F(x',y')\big| < \varepsilon$ ist, falls der Abstand zwischen (x',y') und (x,y) kleiner ist als δ. Zu diesem δ wiederum gibt es ein $N \in \mathbb{N}$, so daß $|x - x_n| < \delta$ für alle $n > N$, so daß für alle solchen n gilt: $|F(x_n,y)| < \varepsilon$. Läßt man hier ε gegen Null gehen, folgt, daß

$$F(x,y) = F(\lim_{n\to\infty} x_n, y) = 0$$

ist, d.h. $F(x,y) = 0$ und damit $f(x) = y$, wie gewünscht.

Somit ist f stetig; als nächstes müssen wir noch die Differenzierbarkeit zeigen. Für ein $x \in I$ und ein hinreichend kleines $h \in \mathbb{R}$, für das auch noch $x + h$ in I liegt, ist nach Definition von f

$$F\big(x+h, f(x+h)\big) = 0\,.$$

Andererseits können wir diesen Funktionswert auch nach dem Mittelwertsatz berechnen: Mit $k = f(x+h) - f(x)$ ist

$$F\big(x+h, f(x+h)\big) = F\big(x+h, f(x)+k\big) = F\Big(\big(x, f(x)\big) + \tbinom{h}{k}\Big)\,,$$

und setzen wir

$$\varphi(t) \underset{\text{def}}{=} F\Big(\big(x, f(x)\big) + t\tbinom{h}{k}\Big)\,,$$

so ist nach dem Mittelwertsatz der Differentialrechnung

$$\varphi(1) = \varphi(0) + \dot\varphi(\tau)$$

für ein τ zwischen null und eins. Mit $\xi = x + \tau h$ und $\eta = f(x) + \tau k$ ist daher

$$F\big(x+h, f(x)+k\big) = F\big(x, f(x)\big) + hF_x(\xi,\eta) + kF_y(\xi,\eta)\,.$$

Da $F\big(x+h, f(x)+k\big)$ und $F\big(x, f(x)\big)$ beide verschwinden, folgt

$$\frac{f(x+h) - f(x)}{h} = \frac{k}{h} = -\frac{F_x(\xi,\eta)}{F_y(\xi,\eta)}\,.$$

Für $h \to 0$ geht die rechte Seite wegen der Stetigkeit der partiellen Ableitungen gegen $-F_x(x,y)/F_y(x,y)$, insbesondere existiert also der Grenzwert. (Man beachte, daß hier nochmals die Voraussetzung $F_y \neq 0$ benötigt wird). Damit existiert auch der Grenzwert des linksstehenden Differenzenquotienten für $h \to 0$, d.h. f ist differenzierbar und hat die behauptete Ableitung. ∎

Nachdem wir wissen, daß die Ableitung von f existiert, ist ihre Berechnung, unabhängig vom gerade bewiesenen Satz, eine einfache Übungsaufgabe: Die Funktion $F\big(x, f(x)\big)$ ist gleich der Nullfunktion, und damit verschwindet natürlich auch ihre Ableitung. Andererseits ist diese Ableitung nach der Kettenregel gleich

$$F_x\big(x, f(x)\big) + F_y\big(x, f(x)\big) \cdot f'(x)\,,$$

also folgt

$$f'(x) = -\frac{F_x\big(x, f(x)\big)}{F_y\big(x, f(x)\big)}\,.$$

Genauso kann auch die zweite Ableitung von f berechnet werden –falls sie existiert. Wenn F und f zweimal stetig differenzierbar sind, können wir $F\big(x, f(x)\big)$ zweimal ableiten, was natürlich immer noch null ist. Nach der Kettenregel ist aber die zweite Ableitung von $F\big(x, f(x)\big)$ (der Übersichtlichkeit halber jeweils ohne das Argument $\big(x, f(x)\big)$ geschrieben) gleich

$$\begin{aligned}&\frac{\partial}{\partial x}\big(F_x + F_y \cdot f'(x)\big) + \frac{\partial}{\partial y}\big(F_x + F_y \cdot f'(x)\big) \cdot f'(x)\\ &= F_{xx} + F_{yx} \cdot f'(x) + F_y \cdot f''(x) + \big(F_{xy} + F_{yy} \cdot f'(x)\big) \cdot f'(x)\,,\end{aligned}$$

d.h.

$$\begin{aligned}f''(x) &= -\frac{1}{F_y}\big(F_{xx} + 2F_{xy} \cdot f'(x) + F_{yy} \cdot f'(x)^2\big)\\ &= -\frac{1}{F_y^3}\big(F_{xx}^2F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2\big)\,.\end{aligned}$$

Für Extremwertbetrachtungen bei implizit definierten Funktionen braucht man, wie wir im nächsten Semester sehen werden, die zweite Ableitung vor allem in den Punkten, in denen die erste verschwindet; dort vereinfacht sich die Formel zu

$$f''(x) = -\frac{F_{xx}\big(x, f(x)\big)}{F_y\big(x, f(x)\big)} \quad \text{falls} \quad f'(x) = 0\,.$$

Genau wie Funktionen einer Veränderlichen können auch Funktionen mehrerer Veränderlicher implizit definiert sein; die entsprechende Verallgemeinerung des Satzes über implizite Funktionen folgt fast vollständig aus der koordinatenweisen Anwendung des obigen Satzes, lediglich für die Stetigkeit der Funktion muß man das entsprechende Argument aus dem gerade beendeten Beweis noch einmal anwenden. Die Aussage ist

Satz: $D \subseteq \mathbb{R}^n$ sei eine offene Menge, $F \in \mathcal{C}^1(D, \mathbb{R})$ eine stetig differenzierbare Funktion auf D, und $\boldsymbol{a} = (a_1, \ldots, a_n) \in D$ sei ein Punkt mit $f(\boldsymbol{a}) = 0$. Falls $F_{x_n}(\boldsymbol{a}) \neq 0$ ist, gibt es eine offene Umgebung U von $(a_1, \ldots, a_{n-1})$ in $\mathbb{R}^{n-1}$ und eine Funktion $f \in \mathcal{C}^1(U, \mathbb{R})$, so daß für alle Punkte $\boldsymbol{y} = (y_1, \ldots, y_{n-1}) \in U$ gilt:

$$F\big(\boldsymbol{y}, f(\boldsymbol{y})\big) = 0\,.$$

Für alle $i \leq n-1$ ist

$$f_{x_i}(\boldsymbol{y}) = -\frac{F_{x_i}\big(\boldsymbol{y}, f(\boldsymbol{y})\big)}{F_{x_n}\big(\boldsymbol{y}, f(\boldsymbol{y})\big)}\,.$$

■

§2: Vektorfelder

Eine Funktion $f\colon D \to \mathbb{R}^m$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ läßt sich für $n, m \geq 2$ nicht mehr durch einen Graphen veranschaulichen, denn der Graph

$$\Gamma_f = \big\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \boldsymbol{y} = f(\boldsymbol{x})\big\}$$

liegt dann im mindestens vierdimensionalen $\mathbb{R}^{n+m}$.

Im für uns wichtigsten Fall $n = m = 2$ oder 3 gibt es aber eine andere Möglichkeit der Veranschaulichung: Für $\boldsymbol{x} \in D$ läßt sich $f(\boldsymbol{x})$ als Vektor auffassen, und f kann veranschaulicht werden, indem man für hinreichend viele Punkte $\boldsymbol{x} \in D$ diesen Vektor (oder ein geeignetes konstantes Vielfaches davon) im Punkt $\boldsymbol{x}$ anträgt; siehe dazu die Abbildungen 38 und 39.

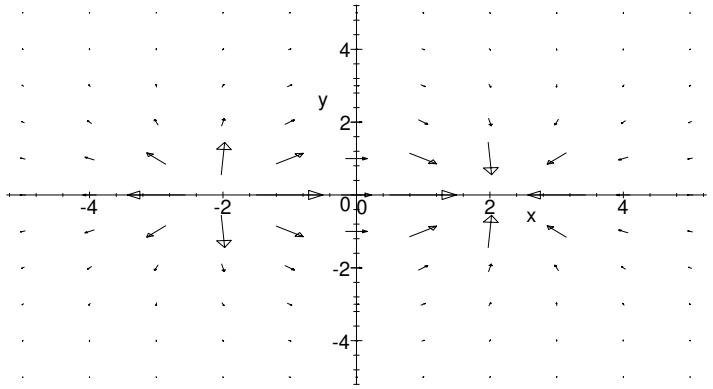


Abb. 38: Das elektrische Feld zweier entgegengesetzt gleicher Punktladungen

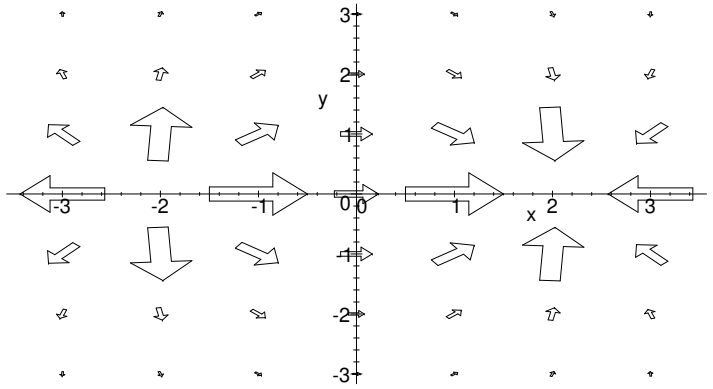


Abb. 39: Dasselbe Feld in einem kleineren Bereich mit „flächigen“ Vektoren

Wie Abbildung 38 zeigt, kann es dabei bei betragsmäßig stark variierenden Funktionswerten zu Problemen kommen: Wie man die Konstante auch wählt, mit der alle Vektoren multipliziert werden, hat man immer entweder einige extrem lange Vektoren oder aber Vektoren, die so kurz sind, daß man nichts mehr erkennen kann: Ab einer gewissen Entfernung vom Zentrum sind die Richtungen des Felds praktisch nicht mehr wahrnehmbar. Eine leichte Verbesserung ergibt sich, wenn man wie in Abbildung 39 die Vektoren „flächig“ zeichnet, wobei die Länge des Vektors proportional zur *Fläche* des eingezeichneten Pfeils ist, aber auch das hilft nicht sehr viel.

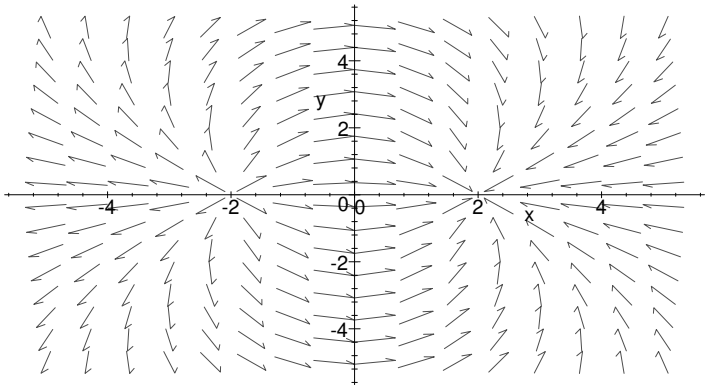


Abb. 40: Das Richtungfeld zum obigen Feld

Da gerade die Richtung sehr wesentlich für die Konstruktion von Feldlinien ist, begnügt man sich oft damit, nur die Richtungen darzustellen, ansonsten aber sämtlichen Vektoren dieselbe Länge zu geben; siehe dazu etwa Abbildung 40. Dadurch geht zwar ein Teil der Information verloren, aber das entsprechende Bild kann trotzdem oft mehr aussagen, als eines, bei dem fast alle Pfeile visuell zu Punkten degeneriert sind. Auch bei der Darstellung als Richtungsfeld müssen allerdings natürlich alle Punkte $\boldsymbol{x}$, für die $f(\boldsymbol{x}) = (0, \ldots, 0)$ ist, durch Punkte oder überhaupt nicht dargestellt werden, denn dort gibt es keine Richtung.

a) Der Begriff des Vektorfelds

Nicht nur zur Veranschaulichung kann es nützlich sein, die Werte einer Funktion als Vektoren zu betrachten: Die für uns wichtigsten Beispiele von Funktionen mehrerer Veränderlicher, deren Wert nicht in $\mathbb{R}$ sondern in einem $\mathbb{R}^n$ liegt, sind die aus der Physik bekannten Felder wie etwa ein Kraftfeld, ein elektrisches Feld oder ein Magnetfeld. Da solche Felder typischerweise als Vektoren geschrieben werden, schreiben wir deshalb in diesem Paragraphen die *Funktionswerte* vektoriell, wohingegen die Argumente weiterhin Punkte sind, die als Tupel geschrieben werden. Wir betrachten somit Funktionen

$$\vec{V}\colon D \to \mathbb{R}^m; \quad \boldsymbol{x} \mapsto \vec{V}(\boldsymbol{x}) = \begin{pmatrix} V_1(\boldsymbol{x}) \\ \vdots \\ V_n(\boldsymbol{x}) \end{pmatrix},$$

wobei die Komponentenfunktionen $V_i\colon D \to \mathbb{R}$ genau die im vorigen Paragraphen betrachteten skalarwertigen Funktionen mehrerers Veränderlicher sind.

Der mathematische Begriff, der physikalische Felder beschreibt, ist der des Vektorfelds:

Definition: Ein Vektorfeld auf der Teilmenge $D \subseteq \mathbb{R}^n$ ist eine Funktion $\vec{V}\colon D \to \mathbb{R}^n$.

(Da die typischen elektromagnetischen Felder wie $\vec{E}, \vec{D}, \vec{B}, \vec{H}$ oder auch das Kraftfeld $\vec{F}$ üblicherweise durch Großbuchstaben bezeichnet werden, verwenden wir auch für Vektorfelder meist Großbuchstaben wie $\vec{V}$ oder $\vec{W}$.)

b) Die Jacobi-Matrix

Wir werden uns in diesem Paragraphen zwar hauptsächlich mit Vektorfeldern im $\mathbb{R}^n$ beschäftigen, und uns selbst da teilweise auf den Fall $n = 3$ beschränken; zumindest die grundlegenden Definitionen sollen aber für beliebige Funktionen auf einer Teilmenge D von $\mathbb{R}^n$ mit Werten in irgendeinem $\mathbb{R}^m$ betrachtet werden.

Fast von selbst versteht sich die

Definition: Eine Funktion $\vec{f}\colon D \to \mathbb{R}^m$ auf $D \subseteq \mathbb{R}^n$ heißt stetig *bzw.* differenzierbar, wenn alle Komponentenfunktionen f_i stetig *bzw.* differenzierbar sind. Entsprechend definieren wir auch die Menge $\mathcal{C}^k(D, \mathbb{R}^m)$ aller k-fach stetig differenzierbarer Funktionen von D nach $\mathbb{R}^m$ als Menge aller Funktionen, deren Komponenten $f_i \in \mathcal{C}^k(D, \mathbb{R})$ k-mal stetig differenzierbar sind.

Wie bei skalarwertigen Funktionen überlegt sich leicht, daß $\mathcal{C}^k(D, \mathbb{R}^m)$ ein $\mathbb{R}$-Vektorraum ist. Damit wissen wir also, was Differenzierbarkeit einer Funktion $\vec{f}\colon D \to \mathbb{R}^n$ bedeutet; was aber ist ihre Ableitung?

Bei skalarwertigen Funktionen einer oder mehrerer Veränderlicher hatten wir die Ableitung jeweils als Linearisierung der Funktion in der Umgebung des betrachteten Punkts definiert; genauso gehen wir auch hier vor: Für jede der Komponentenfunktion f_i ist

$$f_i(\boldsymbol{x} + \vec{h}) = f_i(\boldsymbol{x}) + \operatorname{grad} f_i(\boldsymbol{x}) \cdot \vec{h} + o\big(\|\vec{h}\|\big)\,,$$

wobei der Punkt hier für das Skalarprodukt von Vektoren steht. Im Hinblick auf vektorwertige Funktionen ist es aber besser, anstelle des Skalarprodukt ein Matrixprodukt zu verwenden; dazu müssen wir den Gradienten als Zeilenvektor schreiben, d.h. wenn $\cdot$ das Matrixprodukt bezeichnet, ist

$$f_i(\boldsymbol{x} + \vec{h}) = f_i(\boldsymbol{x}) + {}^t\operatorname{grad} f_i(\boldsymbol{x}) \cdot \vec{h} + o\big(\|\vec{h}\|\big)\,.$$

In dieser Notation können wir die als Zeilenvektoren aufgefaßten Gradienten zu einer $n \times n$-Matrix zusammenfassen, der JACOBI-Matrix

$$J_{\vec{f}}(\boldsymbol{x}) = \begin{pmatrix} {}^t\operatorname{grad} f_1(\boldsymbol{x}) \\ \vdots \\ {}^t\operatorname{grad} f_m(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}\,.$$

Mit dieser Matrix ist dann

$$\vec{f}(\boldsymbol{x} + \vec{h}) = \vec{f}(\boldsymbol{x}) + J_{\vec{f}}(\boldsymbol{x}) \cdot \vec{h} + o\big(\|\vec{h}\|\big)\,,$$

wobei hier das Produkt als Matrixprodukt zu verstehen wird und das Symbol $o\big(\|\vec{h}\|\big)$ schlampigerweise auch für einen Vektor benutzt wird, dessen sämtliche Komponenten $o\big(\|\vec{h}\|\big)$ sind.

c) Die Divergenz eines Vektorfelds

Die JACOBI-Matrix eines Vektorfelds hängt stark ab vom gewählten Koordinatensystem. Wir wollen aus dieser Matrix Größen ableiten, von denen wir später sehen werden, daß sie *intrinsische,* d.h. also vom Koordinatensystem unabhängige Eigenschaften des Vektorfelds beschreiben.

Definition: *a)* Die *Spur* einer $n \times n$-Matrix $A = (a_{ij})$ ist die Summe

$$\operatorname{Spur} A = a_{11} + a_{22} + \cdots + a_{nn}$$

der Diagonalelemente von A.
b) Die Spur der JACOBI-Matrix $J_{\vec{V}}(\boldsymbol{x})$ eines Vektorfelds heißt *Divergenz* oder *Quellendichte* von $\vec{V}$:

$$\operatorname{div} \vec{V}(\boldsymbol{x}) = \operatorname{Spur} J_{\vec{V}}(\boldsymbol{x})\,.$$

Um ein anschauliches Verständnis von der Divergenz zu bekommen, betrachten wir den (sehr speziellen) Fall, daß die JACOBI-Matrix von $\vec{V}$ eine Diagonalmatrix ist:

$$J_{\vec{V}}(\boldsymbol{x}) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

Dann ist für jeden Einheitsvektor $\vec{e}_i$

$$\vec{V}(\boldsymbol{x} + h\vec{e}_i) = \vec{V}(\boldsymbol{x}) + h \cdot J_{\vec{V}}(\boldsymbol{x})\vec{e}_i + o(h) = \vec{V}(\boldsymbol{x}) + h\lambda_i\vec{e}_i + o(h)\,,$$

und für einen beliebigen Vektor $\vec{h} = h_1\vec{e}_1 + \cdots + h_n\vec{e}_n$ ist

$$\vec{V}(\boldsymbol{x} + \vec{h}) = \vec{V}(\boldsymbol{x}) + J_{\vec{V}}(\boldsymbol{x})\vec{h} + o\big(\|\vec{h}\|\big) = \vec{V}(\boldsymbol{x}) + \begin{pmatrix} \lambda_1 h_1 \\ \vdots \\ \lambda_n h_n \end{pmatrix} + o\big(\|\vec{h}\|\big)\,.$$

Falls alle λ_i positiv sind, hat hier jede Komponente des Vektors $\vec{V}(\boldsymbol{x}) + J_{\vec{V}}(\boldsymbol{x})\vec{h}$ dasselbe Vorzeichen wie die entsprechende Komponente von $\vec{h}$; entfernt man sich also vom Punkt $\boldsymbol{x}$, so zeigt auch die Veränderung des Vektorfelds $\vec{V}$ weg vom Punkt x und umgekehrt; die Situation

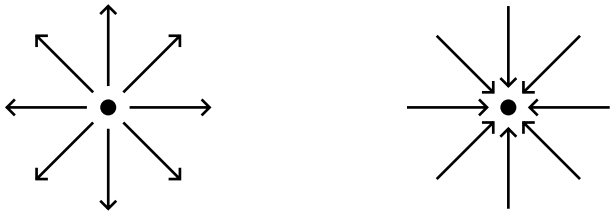


Abb. 41: Quellen und Senken

ist also ungefähr so wie auf der linken Seite von Abbildung 41. Man bezeichnet $\boldsymbol{x}$ in einem solchen Fall als eine *Quelle* des Vektorfelds $\vec{V}$.

Falls alle λ_i negativ sind, gehen die Änderungen jeweils in Gegenrichtung, es sieht also ungefähr so aus wie im rechten Teil von Abbildung 41, und man spricht von einer *Senke.*

Im allgemeinen werden allerdings weder *alle* λ_i positiv noch *alle* λ_i negativ sein; die Divergenz als Summe der λ_i sagt uns dann, ob sich der Punkt eher wie eine Quelle oder eher wie eine Senke verhält.

Nun wird es natürlich nur selten vorkommen, daß die JACOBI-Matrix eine Diagonalmatrix ist; wie wir am Ende dieses Kapitels sehen werden, ist die Divergenz trotzdem auch im allgemeinen Fall eine Maßzahl dafür, ob ein Punkt Quelle oder Senke ist, und zwar in einer viel präziseren Weise, als es die obige Diskussion für Diagonalmatrizen zeigte.

Im übrigen sind Diagonalmatrizen doch nicht so speziell, wie es zunächst den Anschein hat: Im nächsten Semester werden wir sehen, daß es zu jeder symmetrischen Matrix ein Koordinatensystem gibt, bezüglich dessen die durch die Matrix definierte lineare Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^n$ eine Diagonalmatrix als Abbildungsmatrix hat; zumindest für den symmetrischen Anteil der JACOBI-Matrix ist die obige Diskussion also richtig. Mit dem antisymmetrischen Anteil beschäftigen wir uns im nächsten Abschnitt.

In diesem Abschnitt wollen wir uns nur noch überlegen, wie man die Divergenz eines Vektorfelds tatsächlich ausrechnet. Die Divergenz ist

nach Definition gleich der Spur der JACOBI-Matrix, also ist

$$\operatorname{div}\vec{V} = \operatorname{Spur} J_{\vec{V}}(x) = \operatorname{Spur}\begin{pmatrix} \frac{\partial V_1}{\partial x_1} & \frac{\partial V_1}{\partial x_2} & \cdots & \frac{\partial V_1}{\partial x_n} \\ \frac{\partial V_2}{\partial x_1} & \frac{\partial V_2}{\partial x_2} & \cdots & \frac{\partial V_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial V_n}{\partial x_1} & \frac{\partial V_n}{\partial x_2} & \cdots & \frac{\partial V_n}{\partial x_n} \end{pmatrix}$$
$$= \frac{\partial V_1}{\partial x_1} + \frac{\partial V_2}{\partial x_2} + \cdots + \frac{\partial V_n}{x_n}\,.$$

Genauso wie wir den Gradienten einer skalarwertigen Funktion f auch als ∇f schreiben, schreiben wir die Divergenz eines Vektorfelds $\vec{V}$ auch als $\nabla\vec{V}$. Beides kann formal auch als Produkt *bzw.* Skalarprodukt mit einem vektorwertigen Operator interpretiert werden:

$$\nabla f = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \cdot f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

und

$$\nabla\vec{V} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{pmatrix} = \frac{\partial V_1}{\partial x_1} + \frac{\partial V_2}{\partial x_2} + \cdots + \frac{\partial V_n}{\partial x_n}\,.$$

Diese Interpretation darf allerdings nicht *zu* ernst genommen werden, denn obwohl Skalarprodukt und Produkt mit einem Skalar beide kommutativ sind, ist weder die Gleichung $\nabla\vec{V} = \vec{V}\nabla$ noch die Gleichung $\nabla f = f\nabla$ auch nur annähernd sinnvoll, geschweige denn richtig.

d) Die Rotation eines dreidimensionalen Vektorfelds

Erinnern wir uns zunächst an symmetrische und antisymmetrische Matrizen: Eine Matrix ist *symmetrisch,* wenn $a_{ij} = a_{ji}$ ist für alle Indexpaare (i, j); sie ist *antisymmetrisch* im Falle $a_{ij} = -a_{ji}$. Jede reelle Matrix A kann als Summe einer symmetrischen Matrix S und einer

antisymmetrischen Matrix T geschrieben werden; die Einträge von S und T sind jeweils

$$\frac{a_{ij}+a_{ji}}{2} \quad \text{und} \quad \frac{a_{ij}-a_{ji}}{2}\,.$$

Für $n = 1$ ist jede $n \times n$-„Matrix“ symmetrisch, es gibt als keinen antisymmetrischen Anteil.

Für $n = 2$ hat jede antisymmetrische Matrix A die Form

$$\begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix},$$

für den antisymmetrischen Anteil der JACOBI-Matrix ist

$$a = \frac{1}{2}\left(\frac{\partial f_1}{\partial y} - \frac{\partial f_2}{\partial x}\right).$$

Damit ist der antisymmetrische Anteil durch diese eine Zahl bestimmt.

Für $n > 3$ gibt es keine einfachere Darstellung des antisymmetrischen Anteils als die schiefsymmetrische Matrix selbst, bleibt also noch der für viele Anwendungen besonders wichtige Fall $n = 3$.

Eine antisymmetrische 3×3-Matrix A hat die Form

$$A = \begin{pmatrix} 0 & \alpha & \beta \\ -\alpha & 0 & \gamma \\ -\beta & -\gamma & 0 \end{pmatrix}.$$

Für einen Vektor $\vec{x} \in \mathbb{R}^3$ ist daher

$$A\vec{x} = \begin{pmatrix} 0 & \alpha & \beta \\ -\alpha & 0 & \gamma \\ -\beta & -\gamma & 0 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \alpha y + \beta z \\ -\alpha x + \gamma z \\ -\beta x - \gamma y \end{pmatrix}.$$

Diese spezielle Form des Matrix-Vektor-Produkts sollte Erinnerungen an das Vektorprodukt im $\mathbb{R}^3$ wecken:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} bz - cy \\ cx - az \\ ay - bx \end{pmatrix},$$

d.h.

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

und damit

$$\begin{pmatrix} 0 & \alpha & \beta \\ -\alpha & 0 & \gamma \\ -\beta & -\gamma & 0 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\gamma \\ \beta \\ -\alpha \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Somit entspricht im Dreidimensionalen die Multiplikation mit einer schiefsymmetrischen Matrix genau dem Vektorprodukt (das im übrigen auch nur im Dreidimensionalen definierbar ist).

Der Vektor $\vec{a}$ läßt sich im vorliegenden Fall leicht ausrechnen: Die JACOBI-Matrix

$$J_{\vec{V}}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial V_1}{\partial x} & \frac{\partial V_1}{\partial y} & \frac{\partial V_1}{\partial z} \\ \frac{\partial V_2}{\partial x} & \frac{\partial V_2}{\partial y} & \frac{\partial V_2}{\partial z} \\ \frac{\partial V_3}{\partial x} & \frac{\partial V_3}{\partial y} & \frac{\partial V_3}{\partial z} \end{pmatrix}$$

eines dreidimensionalen Vektorfelds $\vec{V}$ hat antisymmetrischen Anteil

$$A = \frac{1}{2}\left(A - {}^tA\right) = \frac{1}{2}\begin{pmatrix} 0 & \frac{\partial V_1}{\partial y} - \frac{\partial V_2}{\partial x} & \frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x} \\ \frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} & 0 & \frac{\partial V_2}{\partial z} - \frac{\partial V_3}{\partial y} \\ \frac{\partial V_3}{\partial x} - \frac{\partial V_1}{\partial z} & \frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z} & 0 \end{pmatrix},$$

und wie wir uns gerade überlegt haben, ist für jeden Vektor $\vec{v}$

$$A\vec{v} = \vec{a} \times \vec{v} \quad \text{mit} \quad \vec{a} = \frac{1}{2}\begin{pmatrix} \frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z} \\ \frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x} \\ \frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \end{pmatrix}.$$

Da der Faktor $\frac{1}{2}$ in konkreten Rechnungen nur hinderlich wäre, betrachten wir den Vektor $2\vec{a}$ und definieren:

Definition: $\vec{V}\colon D \to \mathbb{R}^3$ sei ein Vektorfeld auf der offenen Teilmenge $D \subseteq \mathbb{R}^3$. Dann bezeichnen wir das Vektorfeld

$$\operatorname{rot} V\colon D \to \mathbb{R}^3; \quad \boldsymbol{x} \mapsto \begin{pmatrix} \frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z} \\ \frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x} \\ \frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \end{pmatrix}$$

als Rotation von $\vec{V}$.

Für ein Vektorfeld $\vec{V} \in \mathcal{C}^1(D, \mathbb{R}^3)$, dessen JACOBI-Matrix symmetrischen Anteil S hat, ist also

$$\vec{V}(\boldsymbol{x} + \vec{h}) = \vec{V}(\boldsymbol{x}) + S\vec{h} + \tfrac{1}{2}(\operatorname{rot}\vec{V}) \times \vec{h} + o\big(\|\vec{h}\|\big) .$$

Der Name *Rotation* wird im folgenden (etwas umfangreichen) Beispiel verständlich: Wir gehen aus vom Vektor

$$\vec{a} = 2 \operatorname{rot} \vec{V}(\boldsymbol{x}) .$$

Nach Definition des Vektorprodukts steht $\vec{a} \times \vec{h}$ senkrecht sowohl auf $\vec{a}$ als auch auf $\vec{h}$. Bedenkt man noch, daß Vektoren Äquivalenzklassen von Pfeilen sind, die man sich nicht unbedingt so vorstellen muß, daß sie alle im selben Punkt beginnen, bietet sich folgende Interpretation an: Wenn wir den Vektor $\vec{h}$ um die von $\vec{a}$ aufgespannte Achse drehen, zeigt $\vec{a} \times \vec{h}$ in Richtung der Tangente der Bahn des Endpunkts von $\vec{h}$. (Damit dies stimmt, muß die Achse so orientiert werden, daß sich $\vec{h}$ im mathematisch positiven, d.h. Gegenuhrzeigersinn, um $\vec{a}$ dreht. Von der Ebene der Uhr aus gesehen zeigt dann $\vec{a}$ nach oben, $\vec{h}$ zu dem sich drehenden Punkt, und $\vec{a} \times \vec{h}$ in Richtung des Gegenuhrzeigersinns, wie man sich z.B. mit der Dreifingerregel leicht veranschaulicht.)

Der antisymmetrische Anteil der JACOBI-Matrix sollte also etwas mit einer Drehbewegung zu tun habe, genauer gesagt mit deren Geschwindigkeitsvektor, denn dieser zeigt in Richtung der Tangente der Bahnkurve.

Um dies genauer zu untersuchen, betrachten wir zunächst eine einfache Drehung eines Vektors um die z-Achse mit Drehwinkel φ. Diese ist aufgrund der Definition von Sinus und Cosinus ausdrückbar durch die lineare Abbildung

$$\begin{aligned} x &\mapsto x\cos\varphi - y\sin\varphi \\ y &\mapsto x\sin\varphi + y\cos\varphi \\ z &\mapsto z \end{aligned} .$$

Als nächstes wollen wir einen Punkt $\boldsymbol{x} \in \mathbb{R}^3$ um eine beliebige Achse drehen; der Drehwinkel sei weiterhin φ. Dazu sei $\vec{a}$ der Einheitsvektor in Richtung der Achse; er sei so orientiert, daß der Drehsinn (bei positivem Winkel φ) gleich dem Gegenuhrzeigersinn ist. Dann gelten in den auf $\vec{a}$ senkrecht stehenden Ebenen im wesentlichen dieselbe Formeln, falls wir nur die Koordinateneinheitsvektoren der x- und der y-Achse ersetzen durch zwei aufeinander und auf $\vec{a}$ senkrecht stehende gleichlange Vektoren. Als ein solcher Vektor bietet sich der Verbindungsvektor $\vec{r}$ von der Achse zum Punkt $\boldsymbol{x}$ an, d.h. jener Vektor, der den Abstand zwischen $\boldsymbol{x}$ und der Achse beschreibt. Ein sowohl auf $\vec{r}$ als auch auf $\vec{a}$ senkrecht stehender Vektor ist $\vec{s} = \vec{a} \times \vec{r}$; da $\vec{a}$ die Länge eins hat, ist $\vec{s}$ genauso lang wie $\vec{r}$.

In der von $\vec{r}$ und $\vec{s}$ aufgespannten Ebene bildet die Drehung um den Winkel φ somit den Vektor $\alpha\vec{r} + \beta\vec{s}$ ab auf

$$(\alpha\cos\varphi - \beta\sin\varphi)\,\vec{r} + (\alpha\sin\varphi + \beta\cos\varphi)\,\vec{s} .$$

Nun zerlegen wir den Ortsvektor $\vec{x}$ eines Punktes $\boldsymbol{x}$ in einen Vektor $\vec{x}_a$ auf der Achse und den radialen, d.h. senkrecht auf der Achse stehenden, Anteil $\vec{r}$.

$\vec{x}_a$ ist die Projektion von $\vec{x}$ auf die Drehachse; da wir $\vec{a}$ als Einheitsvektor vorausgesetzt haben, ist dies nach den allgemeinen Regeln der Vektorrechnung einfach

$$\vec{x}_a = (\vec{a} \cdot \vec{x})\,\vec{a} ,$$

denn die Länge des Vektors wird bei der Projektion mit dem Cosinus des Winkels zur Achse multipliziert, und die Richtung des auf die Achse projezierten Vektors ist natürlich gleich der Richtung der Achse.

Damit läßt sich auch die radiale Komponente problemlos berechnen: Da beide Komponenten zusammen den Vektor $\vec{x}$ ergeben müssen, ist

$$\vec{r} = \vec{x} - \vec{x}_a = \vec{x} - (\vec{a} \cdot \vec{x})\,\vec{a} .$$

Nun ist $\vec{x}_a$ parallel zu $\vec{a}$, und das Vektorprodukt zweier paralleler Vektoren verschwindet. Deshalb können wir auch schreiben

$$\vec{s} = \vec{a} \times \vec{r} = \vec{a} \times \Big(\vec{x} - (\vec{x} \cdot \vec{a})\vec{a}\Big) = \vec{a} \times \vec{x} .$$

Da die Axialkomponente eines Vektors bei Drehungen erhalten bleibt, wird speziell der Vektor $\vec{x} = \vec{x}_a + \vec{r}$ also abgebildet auf

$$\begin{aligned} &\vec{x}_a + \cos\varphi\ \vec{r} + \sin\varphi\ \vec{s} \\ &= (\vec{x} \cdot \vec{a})\,\vec{a} + \cos\varphi\ \Big(\boldsymbol{x} - (\vec{x} \cdot \vec{a})\,\vec{a}\Big) + \sin\varphi\ \vec{a} \times \vec{x} \\ &= (1 - \cos\varphi)(\vec{x} \cdot \vec{a})\,\vec{a} + \cos\varphi\ \vec{x} + \sin\varphi\ \vec{a} \times \vec{x} . \end{aligned}$$

Um den Geschwindigkeitsvektor einer Drehung auszurechnen, müssen wir diese zunächst zu einem dynamischen Prozeß machen, d.h. anstelle einer Drehung mit konstantem Drehwinkel φ betrachten wir eine Drehung mit Winkelgeschwindigkeit ω. Wir gehen wieder aus von einem festen Punkt $\boldsymbol{x}$ und betrachten dessen Position zum Zeitpunkt t, d.h. wir betrachten die Funktion

$$\vec{f}\colon \begin{cases} \mathbb{R} & \to \mathbb{R}^3 \\ t & \mapsto (1 - \cos\omega t)(\vec{x} \cdot \vec{a})\ \vec{a} + \cos\omega t\ \vec{x} + \sin\omega t\ \vec{a} \times \vec{x} . \end{cases}$$

Der Geschwindigkeitsvektor im Punkt $\boldsymbol{x}$ ist die Ableitung dieser Funktion im Nullpunkt, also der Vektor

$$\frac{d\vec{f}}{dt}(0) = \Big(\omega\sin\omega t\ (\vec{x} \cdot \vec{a})\vec{a} - \omega\sin\omega t\ \vec{x} + \omega\cos\omega t\ \vec{a} \times \vec{x}\Big)_{|t=0} = \omega\vec{a} \times \vec{x} .$$

Mit anderen Worten: Bei der Drehung mit Winkelgeschwindigkeit ω um die von $\vec{a}$ aufgespannte Achse ist der Geschwindigkeitsvektor im Punkt $\boldsymbol{x}$ gleich $\omega\vec{a} \times \vec{x}$.

Der Vektor $\omega\vec{a}$ läßt sich auffassen als eine vektorielle Winkelgeschwindigkeit (die in der Physik in der Tat genau so definiert wird), und für einen beliebigen Vektor $\vec{\omega} \in \mathbb{R}^3$ ist das Vektorfeld

$$\vec{V}\colon \mathbb{R}^3 \to \mathbb{R}^3; \quad \boldsymbol{x} \mapsto \vec{\omega} \times \vec{x}$$

das Feld der Geschwindigkeitsvektoren zu einer Rotation mit vektorieller Winkelgeschwindigkeit $\vec{\omega}$.

Es sei noch darauf hingewiesen, daß $\operatorname{rot}\vec{V}$ in der englischsprachigen Literatur meist als $\operatorname{curl}\vec{V}$ geschrieben wird nach dem englischen Wort *curl* = Locke.

Wie Gradient und Divergenz können wir auch die Rotation formal mit Hilfe des Operators ∇ ausdrücken: Für ein dreidimensionales Vektorfeld $\vec{V}$ ist

$$\nabla \times \vec{V} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \times \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} = \begin{pmatrix} \frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z} \\ \frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x} \\ \frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \end{pmatrix} = \operatorname{rot}\vec{V} \,.$$

e) Erste Beispiele

Zum besseren Verständnis von Divergenz und Rotation und um den Umgang mit den beiden Begriffen zu üben, wollen wir sie für zwei einfache Felder ausrechnen: Für das elektrische Feld einer Punktladung und für das Magnetfeld eines stromdurchflossenen Leiters.

1) Das elektrische Feld einer Punktladung: Eine Punktladung q im Nullpunkt erzeugt nach dem COULOMBschen Gesetz ein elektrisches Feld $\vec{E}$ der Stärke

$$\vec{E}(\boldsymbol{x}) = \frac{q}{4\pi\varepsilon_0} \frac{\vec{x}}{\left|\vec{x}\right|^3} \,,$$

denn die Feldstärke nimmt mit der Entfernung vom Nullpunkt quadratisch ab und hat die Richtung (oder – je nach Vorzeichen von q – auch Gegenrichtung) des Ortsvektors., d.h. der Einheitsvektor der Richtung ist $\pm\vec{x}/\left|\vec{x}\right|$, und er muß noch durch $\left|\vec{x}\right|^2$ dividiert werden.

Das Feld $\vec{E}$ ist nur definiert auf $\mathbb{R}^3 \setminus \{(0,0,0)\}$; im Nullpunkt haben wir einen Nenner Null, und auch physikalisch ist es nicht sinnvoll, nach dem Feld *im Innern* einer Punktladung zu fragen: Punktladungen sind nun einmal mathematische Idealisierungen, und wirklich anwendbar ist das COULOMBsche Gesetz ohnehin erst ab einem gewissen Mindestabstand: Bei zu kleinen Abständen spielen andere Wechselwirkungen eine wesentlich größere Rolle.

In Koordinaten ist also

$$\vec{E}(\boldsymbol{x}) = \begin{pmatrix} E_1(\boldsymbol{x}) \\ E_2(\boldsymbol{x}) \\ E_3(\boldsymbol{x}) \end{pmatrix} = \frac{q}{4\pi\varepsilon_0\left\|\vec{x}\right\|^3} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \,.$$

Zur Berechnung der Divergenz müssen wir die erste Komponente nach x, die zweite nach y und die dritte nach z ableiten und dann die drei Ableitungen addieren. Beachten wir, daß

$$\left\|\vec{x}\right\|^3 = (x^2+y^2+z^2)^{3/2}$$

ist, folgt sofort aus der Quotientenregel, daß

$$\begin{aligned} \frac{\partial E_1}{\partial x}(\boldsymbol{x}) &= \frac{q}{4\pi\varepsilon_0} \frac{\partial}{\partial x} \frac{x}{\left\|\vec{x}\right\|^3} = \frac{q}{4\pi\varepsilon_0} \frac{\left\|\vec{x}\right\|^3 - x\cdot\frac{3}{2}\cdot 2x\left\|\vec{x}\right\|}{\left\|\vec{x}\right\|^6} \\ &= \frac{q}{4\pi\varepsilon_0} \frac{\left\|\vec{x}\right\|^2 - 3x^2}{\left\|\vec{x}\right\|^5} = \frac{q}{4\pi\varepsilon_0\left\|\vec{x}\right\|^5}(-2x^2+y^2+z^2) \,. \end{aligned}$$

Entsprechend ist

$$\frac{\partial E_2}{\partial y}(\boldsymbol{x}) = \frac{q}{4\pi\varepsilon_0\left\|\vec{x}\right\|^5}(x^2-2y^2+z^2)$$

und

$$\frac{\partial E_3}{\partial z}(\boldsymbol{x}) = \frac{q}{4\pi\varepsilon_0\left\|\vec{x}\right\|^5}(x^2+y^2-2z^2) \,,$$

also

$$\begin{aligned} \operatorname{div}\vec{E}(\boldsymbol{x}) &= \frac{q}{4\pi\varepsilon_0\left\|\vec{x}\right\|^5}(-2x^2+y^2+z^2+x^2-2y^2+z^2+x^2+y^2-2z^2) \\ &= 0 \,. \end{aligned}$$

Zur Berechnung der Rotation benötigen wir die Einträge der JACOBI-Matrix, die nicht in der Diagonale stehen; dabei werden die Komponenten von $\vec{E}$ nach jenen Variablen abgeleitet, die *nicht* im Zähler stehen. Die Situation ist also stets dieselbe, und es genügt, wenn wir beispielsweise die Ableitung von E_1 nach y ausrechnen.

Aus den üblichen Ableitungsregeln folgt sofort, daß

$$\frac{\partial E_1}{\partial y} = \frac{q}{4\pi\varepsilon_0}\left(-\frac{3}{2}\right)\frac{x\cdot 2y}{\|\vec{x}\|^5} = -\frac{3q}{4\pi\varepsilon_0}\frac{xy}{\|\vec{x}\|^5}$$

ist, und analog ist auch

$$\frac{\partial E_2}{\partial x} = \frac{q}{4\pi\varepsilon_0}\left(-\frac{3}{2}\right)\frac{y\cdot 2x}{\|\vec{x}\|^5} = -\frac{3q}{4\pi\varepsilon_0}\frac{xy}{\|\vec{x}\|^5}\,.$$

Die x-Komponente von rot $\vec{E}$ ist somit

$$\frac{\partial E_2}{\partial x} - \frac{\partial E_1}{\partial y} = 0\,,$$

und genauso überzeugt man sich auch vom Verschwinden der anderen Komponenten.

Somit ist also das Feld einer Punktladung außerhalb des Punktes selbst sowohl quellen- als auch wirbelfrei, genau wie es nach den MAXWELLschen Gleichungen auch sein muß.

2) Das Magnetfeld eines stromdurchflossenen Leiters: Das Magnetfeld eines von einem Strom der Stärke I durchflossenen geradlinigen Drahts ist nach dem BIOT-SAVARTschen Gesetz

$$\vec{H}(\boldsymbol{x}) = \frac{I}{4\pi\|\vec{r}\|^2}\,\vec{a}\times\vec{x}\,,$$

wobei $\vec{a}$ der Einheitsvektor in Richtung des Stroms und $\vec{r}$ der Abstandsvektor zwischen dem Draht und dem Punkt $\vec{x}$ ist. $\vec{H}$ ist überall im $\mathbb{R}^3$ erklärt, außer auf der von $\vec{a}$ aufgespannten Geraden, wo $\vec{r}$ der Nullvektor ist.

Zum einfacherer Rechnen wählen wir das Koordinatensystem so, daß $\vec{a}$ der Einheitsvektor in Richtung der positiven z-Achse ist; für den Punkt mit Koordinaten (x, y, z) ist dann

$$\vec{r} = \begin{pmatrix} x\\ y\\ 0\end{pmatrix} \quad\text{und}\quad \vec{a}\times\vec{x} = \begin{pmatrix}0\\0\\1\end{pmatrix}\times\begin{pmatrix}x\\y\\z\end{pmatrix} = \begin{pmatrix}-y\\x\\0\end{pmatrix}\,,$$

d.h.

$$\vec{H}(\boldsymbol{x}) = \frac{I}{4\pi(x^2+y^2)}\begin{pmatrix}-y\\x\\0\end{pmatrix}\,.$$

Dieses Vektorfeld ist erklärt auf $\mathbb{R}^3$ minus der z-Achse; auf der z-Achse, d.h. im Innern des Leiters, bekommen wir eine Null in den Nenner.

Die drei Komponenten dieses Felds könnten wir nun wie oben nach den drei Variablen x, y, z ableiten, um so Rotation und Divergenz zu berechnen; es gibt allerdings eine einfachere Möglichkeit, die JACOBI-Matrix eines solchen Felds zu bestimmen:

Wie auch das $\vec{E}$-Feld im obigen Beispiel läßt sich $\vec{H}$ schreiben als Produkt einer skalaren Funktion mit einer *linearen* Abbildung $\mathbb{R}^3 \to \mathbb{R}^3$. Solche Funktionen, bei denen alle Nichtlinearität in einer skalaren Funktion steckt, bezeichnen wir als *quasilinear:*

Definition: Eine Funktion $\vec{f}\colon D \to \mathbb{R}^m$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ heißt *quasilinear,* wenn es eine Funktion $\varphi\colon D \to \mathbb{R}$ und eine Matrix $A \in \mathbb{R}^{n\times m}$ gibt, so daß $\vec{f}(\boldsymbol{x}) = \varphi(\boldsymbol{x})\cdot A\vec{x}$ ist.

In diesem Sinne ist $\vec{H}(x)$ quasilinear mit

$$\varphi(\boldsymbol{x}) = \frac{I}{4\pi(x^2+y^2)} \quad\text{und}\quad A = \begin{pmatrix}0 & -1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}$$

und das obige Feld $\vec{E}$ ist quasilinear mit

$$\varphi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0\|\vec{r}\|^3} \quad\text{und}\quad A = \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\,.$$

Lemma: Für eine quasilineare Funktion $\vec{f}\colon D \to \mathbb{R}^m$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ ist

$$J_{\vec{f}}(\boldsymbol{x}) = \varphi(\boldsymbol{x})\cdot A + (A\vec{x})\cdot {}^t\mathrm{grad}\,\varphi(\boldsymbol{x})\big)\,.$$

In dieser Formel steht der erste Malpunkt für die Multiplikation eines Skalars mit einer Matrix, der zweite für das Matrixmultiplikation des

Spaltenvektors $A\vec{x}$ mit dem zum *Zeilenvektor* transponierten Gradienten. Es geht hier also nicht um ein Skalarprodukt, sondern um eine Matrixmultiplikation, deren Ergebnis eine $n \times m$-Matrix ist.

Ausgeschrieben wird die Formel zu

$$\frac{\partial f_i}{\partial x_j} = \varphi(\boldsymbol{x}) \cdot a_{ij} + \left(\sum_{\nu=1}^{n} a_{i\nu} x_\nu \right) \cdot \frac{\partial \varphi}{\partial x_j} .$$

Zum *Beweis* leiten wir $f_i(\boldsymbol{x}) = \varphi(\boldsymbol{x}) \sum_{\nu=1}^{n} a_{i\nu} x_\nu$ nach x_j ab; nach der Produktregel ist, wie gewünscht,

$$\begin{aligned} \frac{\partial f_i}{\partial x_j}(\boldsymbol{x}) &= \varphi(\boldsymbol{x}) \frac{\partial}{\partial x_j} \left(\sum_{\nu=1}^{n} a_{i\nu} x_\nu \right) + \frac{\partial \varphi}{\partial x_j}(\boldsymbol{x}) \sum_{\nu=1}^{n} a_{i\nu} x_\nu \\ &= \varphi(\boldsymbol{x}) \cdot a_{ij} + \frac{\partial \varphi}{\partial x_j}(\boldsymbol{x}) \sum_{\nu=1}^{n} a_{i\nu} x_\nu . \end{aligned}$$

∎

Dieses Lemma wenden wir gleich an auf das Feld $\vec{H}(\boldsymbol{x})$; hier ist

$$\frac{\partial \varphi}{\partial x}(\boldsymbol{x}) = \frac{-2x \cdot I}{4\pi(x^2+y^2)^2} \quad \text{und} \quad \frac{\partial \varphi}{\partial y}(\boldsymbol{x}) = \frac{-2y \cdot I}{4\pi(x^2+y^2)^2} ,$$

d.h.

$$\operatorname{grad} \varphi(\boldsymbol{x}) = \frac{-2I}{4\pi(x^2+y^2)^2} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} .$$

Somit ist

$$\begin{aligned} J_{\vec{H}}(\boldsymbol{x}) &= \frac{I}{4\pi(x^2+y^2)} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \cdot \frac{-2I}{4\pi(x^2+y^2)^2} \begin{pmatrix} x & y & 0 \end{pmatrix} \\ &= \frac{I}{4\pi(x^2+y^2)^2} \left(\begin{pmatrix} 0 & -x^2-y^2 & 0 \\ x^2+y^2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right. \\ &\quad \left. + \begin{pmatrix} -y \\ x \\ 0 \end{pmatrix} \cdot \begin{pmatrix} -2x & -2y & 0 \end{pmatrix} \right) \\ &= \frac{I}{4\pi(x^2+y^2)^2} \left(\begin{pmatrix} 0 & -x^2-y^2 & 0 \\ x^2+y^2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right. \\ &\quad \left. + \begin{pmatrix} 2xy & 2y^2 & 0 \\ -2x^2 & -2xy & 0 \\ 0 & 0 & 0 \end{pmatrix} \right) \\ &= \frac{I}{4\pi(x^2+y^2)^2} \begin{pmatrix} 2xy & y^2-x^2 & 0 \\ y^2-x^2 & -2xy & 0 \\ 0 & 0 & 0 \end{pmatrix} . \end{aligned}$$

Diese Matrix ist symmetrisch und hat Spur Null, d.h auch hier ist

$$\operatorname{div} \vec{H}(\boldsymbol{x}) = 0 \quad \text{und} \quad \operatorname{rot} \vec{H}(\boldsymbol{x}) = \vec{0} ,$$

und auch dies natürlich wieder in Übereinstimmung mit den MAXWELL-schen Gleichungen.

Da die Feldlinien des hier betrachteten Felds Kreise sind, mag es auf den ersten Blick seltsam erscheinen, daß die Rotation von $\vec{H}$ verschwindet. Der Grund liegt darin, daß alle Kreise ihren Mittelpunkt auf der z-Achse haben, die wir in der obigen Rechnung ausschließen mußten: Die Rotation in einem Punkt $\boldsymbol{x}$ mißt die Rotation um $\boldsymbol{x}$, und beim betrachteten Feld rotiert eben alles um die z-Achse, so daß die Rotation überall sonst verschwindet.

f) Allgemeine Rechenregeln

Nachdem wir gerade mit relativ großem Aufwand für zwei Beispiele Divergenz und Rotation berechnet haben, wollen wir nun einige Rechenregeln herleiten, die uns unter anderem zeigen werden, daß zumindest ein Teil der obigen Rechnungen hätte vermieden werden können.

Wir wollen uns als nächstes überlegen, was passiert, wenn wir zwei der drei Operatoren grad, div und rot hintereinanderausführen – sofern dies möglich ist.

Der Gradient einer skalaren Funktion $f: D \to \mathbb{R}$ auf einer offenen Teilmenge $D \subseteq \mathbb{R}^n$ ist ein Vektorfeld; es ist also möglich, dessen Divergenz und Rotation zu berechnen. Um keine Probleme mit der Existenz und der Reihenfolge von Ableitungen zu haben, setzen wir voraus, daß f in $\mathcal{C}^2(D, \mathbb{R})$ liegt; dann ist

$$\operatorname{div}\operatorname{grad} f = \operatorname{div}\begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}\,.$$

Die rechte Seite schreiben wir auch kurz als

$$\Delta f \underset{\text{def}}{=} \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}$$

und bezeichnen

$$\Delta \underset{\text{def}}{=} \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$$

als LAPLACE-Operator.

(Der LAPLACE-Operator ist älter als Divergenz, Gradient, Rotation und so weiter; er spielt eine große Rolle sowohl bei Wellen als auch bei Potentialfeldern. HAMILTON wählte das Zeichen ∇ in Anlehnung an diesen Operator; formal kann man $\Delta = \nabla \cdot \nabla$ schreiben.)

Auch die Rotation eines Gradientenfelds läßt sich leicht berechnen, allerdings natürlich nur im $\mathbb{R}^3$, da wir für andere Dimensionen keine

Rotation definiert haben.

$$\operatorname{rot}\operatorname{grad} f = \operatorname{rot}\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f}{\partial z \partial y} - \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial x \partial z} - \frac{\partial^2 f}{\partial z \partial x} \\ \frac{\partial^2 f}{\partial y \partial x} - \frac{\partial^2 f}{\partial x \partial y} \end{pmatrix} = \vec{0}\,.$$

Die Rotation eines stetig differenzierbaren Gradientenfelds ist also stets Null; da das elektrische Feld der negative Gradient des elektrischen Potentials ist, hätten wir oben also für

$$\vec{E}(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\frac{\vec{x}}{\left\|\vec{x}\right\|^3} = -\operatorname{grad}\frac{1}{4\pi\varepsilon_0\left\|\vec{x}\right\|}$$

auf die Berechnung der Rotation verzichten können.

Die Divergenz eines Vektorfelds ist eine skalare Funktion; der einzige Operator, der darauf angewandt werden kann, ist daher der Gradient, und das Ergebnis ist völlig uninteressant: siehe Frage *5)* auf dem zehnten Übungsblatt.

Die Rotation eines Vektorfelds im Dreidimensionalen schließlich ist wieder ein Vektorfeld, wir können also Divergenz und Rotation davon berechnen:

$$\begin{aligned}\operatorname{div}\operatorname{rot}\vec{V} &= \operatorname{div}\begin{pmatrix} \frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z} \\ \frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x} \\ \frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \end{pmatrix} \\ &= \frac{\partial^2 V_3}{\partial x \partial y} - \frac{\partial^2 V_2}{\partial x \partial z} + \frac{\partial^2 V_1}{\partial y \partial z} - \frac{\partial^2 V_3}{\partial y \partial x} + \frac{\partial^2 V_2}{\partial z \partial x} - \frac{\partial^2 V_1}{\partial z \partial y} = 0\end{aligned}$$

für ein zweimal stetig differenzierbares Vektorfeld. Da ein Magnetfeld als Rotation eines Vektorpotentials geschrieben werden kann, erklärt dies, warum im obigen Beispiel $\operatorname{div}\vec{H} = 0$ war.

Bleibt noch die Rotation der Rotation; für diese ist nach Aufgabe 2 des zehnten Übungsblatts

$$\operatorname{rot}\operatorname{rot}\vec{V} = \operatorname{grad}\operatorname{div}\vec{V} - \begin{pmatrix} \Delta V_1 \\ \Delta V_2 \\ \Delta V_3 \end{pmatrix}\,.$$

Ebenfalls dort sind auch die beiden „Produktregeln“

$$\operatorname{div}(f\vec{V}) = f \operatorname{div}\vec{V} + (\operatorname{grad} f)\cdot\vec{V}$$

und

$$\operatorname{rot}(f\vec{V}) = f \operatorname{rot}\vec{V} + (\operatorname{grad} f)\times\vec{V}$$

zu finden. Mit dem Operator ∇ geschrieben werden sie völlig analog zur klassischen LEIBNIZ-Regel:

$$\nabla\cdot(f\vec{V}) = f\nabla\cdot\vec{V} + \nabla f\cdot\vec{V}$$

und

$$\nabla\times(f\vec{V}) = f\nabla\times\vec{V} + \nabla f\times\vec{V}$$

Speziell für ein lineares Vektorfeld

$$\vec{V}\colon \mathbb{R}^n \to \mathbb{R}^n;\quad \boldsymbol{x} \mapsto A\vec{x} \quad \text{mit} \quad A\in\mathbb{R}^{n\times n}$$

ist

$$\operatorname{div}\vec{V} = \sum_{i=1}^{n}\frac{\partial V_i}{\partial x_i} = \sum_{i=1}^{n} a_{ii} = \operatorname{Spur} A$$

und, im Falle $n = 3$,

$$\operatorname{rot}\vec{V} = \begin{pmatrix} a_{32}-a_{23}\\ a_{13}-a_{31}\\ a_{21}-a_{12}\end{pmatrix};$$

für eine quasilineare Funktion $\vec{f}(\boldsymbol{x}) = \varphi(\boldsymbol{x})A\vec{x}$ ist also nach den beiden Produktregeln

$$\operatorname{div}\vec{f} = \varphi \operatorname{Spur} A + (\operatorname{grad}\varphi)\cdot A\vec{x}$$

und

$$\operatorname{rot}\vec{f} = \varphi\cdot\begin{pmatrix} a_{32}-a_{23}\\ a_{13}-a_{31}\\ a_{21}-a_{12}\end{pmatrix} + (\operatorname{grad}\varphi)\times A\vec{x},$$

was uns in den Beispielen des vorigen Abschnitts manche Rechnung erspart hätte.

g) Nichtkartesische Koordinatensysteme

Auch in diesem letzten Abschnitt von §2 soll es um Möglichkeiten gehen, wie wir bei den oben betrachteten und anderen Beispielen effizienter rechnen können. Das Feld einer Punktladung beispielsweise ist kugelsymmetrisch, jedoch haben wir bei der Berechnung von Divergenz und Rotation dieses Feldes die Kugelsymmetrie mutwillig zerstört, indem wir in einem kartesischen Koordinatensystem rechneten. Hier wollen wir uns beschäftigen, wie man Koordinatensysteme finden kann, die einer solchen Symmetrie angepaßt sind.

Entsprechend der Vielzahl möglicher Symmetrien kennt die Mathematik auch eine Vielzahl solcher Koordinatensysteme; wir wollen uns auf die beiden wichtigsten un am häufigsten angewandten beschränken: die Kugel- und die Zylinderkoordinaten. Zum Einstieg in das Thema betrachten wir aber zunächst die etwas einfachere Situation der Polarkoordinaten in der Ebene.

1) Polarkoordinaten in $\mathbb{R}^2$***:*** Ein Punkt $P\in\mathbb{R}^2$ wird charakterisiert durch seien Abstand r vom Nullpunkt O sowie den Winkel φ zwischen dem Vektor $\overrightarrow{OP}$ und dem Einheitsvektor der (positiven) x-Achse, wobei dieser Winkel für $P = O$ nicht definiert ist. Ansonsten messen wir ihn im Bogenmaß und wählen den Drehsinn so, die Richtung von der positiven x-Achse zur positiven y-Achse positiv gerechnet wird.

Für einen Punkt mit kartesischen Koordinaten (x, y) ist dann

$$r = \sqrt{x^2+y^2}$$

und für $(x, y) \neq (0, 0)$ ist φ so bestimmt, daß gilt

$$x = r\cos\varphi \quad \text{und} \quad y = r\sin\varphi\,.$$

φ ist dabei, wie üblich, natürlich nur modulo 2π bestimmt. Da Sinus und Cosinus nur auf Intervallen der Länge π injektiv sind, reicht eine der beiden Umkehrfunktionen allein nicht aus, um φ direkt anzugeben, sondern man braucht noch mindestens eine Fallunterscheidung. Eine mögliche Formel für φ, auf Basis des Arkuscosinus, ist etwa

$$\varphi = \begin{cases} \arccos\frac{x}{r} & \text{falls } y\geq 0\\ -\arccos\frac{x}{r} & \text{falls } y<0\end{cases}.$$

Natürlich kann man entsprechende Formeln auch mit dem Arkussinus oder Arkustangens aufstellen.

Anstelle des Gitternetzes eines kartesischen Koordinatensystems hat man bei den Poalrkoordinaten ein Netz aus Kreisen um den Nullpunkt, auf denen der Radius r konstant ist, und aus vom Nullpunkt ausgehenden Strahlen, auf denen φ konstant ist.

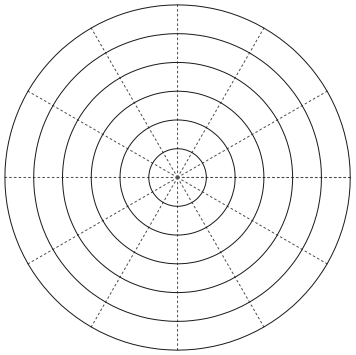


Abb. 42: Das Netz der Polarkoordinaten

Für eine Funktion $f(x,y)$ auf einer Teilmenge des $\mathbb{R}^2$ können wir die Polarkoordinatendarstellung

$$F(r,\varphi) \underset{\text{def}}{=} f(r\cos\varphi, r\sin\varphi)$$

betrachten; wir wollen sehen, wie man mit F allein immer noch den Gradienten oder beispielsweise auch die LAPLACEsche Δf berechnen kann.

Dazu müssen wir zunächst die partiellen Ableitungen von f nach x und y durch die von F nach r und φ ausdrücken. Dies ist auf direkte Weise etwas schwierig, da wir keine gute Formel für die Berechnung von φ aus x und y haben. Die umgekehrte Aufgabe dagegen ist umso

einfacher: Nach der Kettenregel ist

$$\frac{\partial F}{\partial r} = \frac{\partial}{\partial r} f(r\cos\varphi, r\sin\varphi) = \frac{\partial f}{\partial x}\cos\varphi + \frac{\partial f}{\partial y}\sin\varphi$$

und

$$\frac{\partial F}{\partial \varphi} = \frac{\partial}{\partial \varphi} f(r\cos\varphi, r\sin\varphi) = -\frac{\partial f}{\partial x} r\sin\varphi + \frac{\partial f}{\partial y} r\cos\varphi\,.$$

Da $r > 0$ ist, können wir dies auch in der Form

$$\begin{aligned} F_r &= \quad f_x\cos\varphi + f_y\sin\varphi \\ \frac{F_\varphi}{r} &= -f_x\sin\varphi + f_y\cos\varphi \end{aligned}$$

schreiben. Dies ist ein lineares Gleichungssystem für f_x und f_y; wenn wir von der mit $\cos\varphi$ multiplizierten ersten Gleichung die mit $\sin\varphi$ multiplizierte zweite subtrahieren, erhalten wir

$$f_x = F_r\cos\varphi - \frac{F_\varphi}{r}\sin\varphi\,;$$

entsprechend erhalten wir

$$f_y = F_r\sin\varphi + \frac{F_\varphi}{r}\cos\varphi\,,$$

wenn wir die mit $\sin\varphi$ multiplizierte erste Gleichung zur mit $\cos\varphi$ multiplizierten zweiten addieren.

Entsprechend können wir auch die zweiten partiellen Ableitungen von f durch partielle Ableitungen von F ausdrücken, allerdings werden hier die Ausdrücke schon deutlich komplexer, so daß hier auf die detailierte Rechnung verzichtet sei. Wichtig ist vor allem der LAPLACE-Operator, und wenn es auch sehr umständlich wäre, dessen Darstellung in Polarkoordinaten *herzuleiten* können wir die fertige Formel

$$\Delta f = f_{xx} + f_{yy} = F_{rr} + \frac{F_r}{r} + \frac{F_{\varphi\varphi}}{r^2}\,,$$

doch leicht und relativ schnell beweisen kann, indem wir

$$F_{rr} = \big(f_{xx}\cos\varphi + f_{xy}\sin\varphi\big)\cos\varphi + \big(f_{xy}\cos\varphi + f_{yy}\sin\varphi\big)\sin\varphi$$

und

$$F_{\varphi\varphi} = \big(f_{xx} r\sin\varphi - f_{xy} r\cos\varphi\big) r\sin\varphi - f_x r\cos\varphi \\ - \big(f_{xy} r\sin\varphi - f_{yy} r\cos\varphi\big) r\cos\varphi - f_y r\sin\varphi$$

nach der Kettenregel berechnen und einsetzen.

2) Zylinderkoordinaten im $\mathbb{R}^3$***:*** Für den $\mathbb{R}^3$ gibt es zwei naheliegende Verallgemeinerung der Polarkoordinaten: Einmal können wir in der (x, y)-Ebene Polarkoordinaten einführen, die z-Koordinate aber unverändert lassen, zum anderen können wir neben dem Abstand vom Nullpunkt zwei Winkelkoordinaten einführen, wie z.B. die geographische Länge und Breite. Hier soll es zunächst um den ersten Fall gehen, wir haben also drei Koordinaten (r, φ, z) mit $r > 0$, die über die Gleichungen

$$x = r\cos\varphi, \quad y = r\sin\varphi \quad \text{und} \quad z = z$$

mit den kartesischen Koordinaten verbunden sind. Diese Koordinaten heißen *Zylinderkoordinaten,* da die Flächen r = *konstant* Zylinder sind. Insbesondere ist die Koordinate r hier nicht mehr gleich dem Abstand eines Punktes vom Nullpunkt, sondern gleich dem Abstand von der z-Achse.

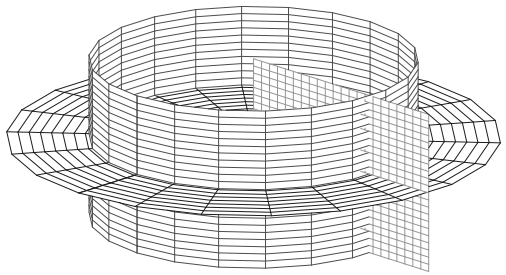


Abb. 43: Flächen mit konstantem r, φ und z für Zylinderkoordinaten

Da die z-Koordinate des kartesischen Koordinatensystems übernommen wird, brauchen wir für diese Koordinaten nicht neu zu rechnen, sondern

können die obigen Ergebnisse praktisch wörtlich übernehmen: Mit

$$F(r, \varphi, z) = f(r\cos\varphi, r\sin\varphi, z)$$

ist

$$f_x = F_r\cos\varphi - \frac{F_\varphi}{r}\sin\varphi$$
$$f_y = F_r\sin\varphi + \frac{F_\varphi}{r}\cos\varphi$$
$$f_z = F_z$$

und

$$\Delta f = f_{xx} + f_{yy} + f_{zz} = F_{rr} + \frac{F_r}{r} + \frac{F_{\varphi\varphi}}{r^2} + F_{zz}\,.$$

3) Kugelkoordinaten: Hier führen wir zusätzlich zu den beiden Polarkoordinaten der Ebenen als dritte Koordinate noch eine *Winkelkoordinate* ϑ ein, wir haben also drei Koordinaten (r, φ, ϑ), wobei r wieder positiv sein muß und den Abstand vom Nulpunkt bezeichnet.

Im Gegensatz zur Konvention der Geographen, wonach der Äquator, d.h. der Schnittkreis einer Kugel um den Nullpunkt mit der (x, y)-Ebenen, die Breite Null hat, ist es in der Mathematik, Physik und Technik üblich, dem Äquator die Koordinate $\vartheta = \pi/2$ zu geben, dem „Nordpol“, d.h. dem Schnittpunkt mit der positiven z-Achse, $\vartheta = 0$ und dem Südpol $\vartheta = \pi$. In der (x, z)-Ebene ist also

$$x = r\sin\vartheta \quad \text{und} \quad z = r\cos\vartheta\,,$$

in der (y, z)-Ebene

$$y = r\sin\vartheta \quad \text{und} \quad z = r\cos\vartheta\,,$$

und ganz allgemein ist

$$R = r\sin\vartheta \quad \text{und} \quad z = r\cos\vartheta\,,$$

wobei R den Abstand eines Punktes von der z-Achse bezeichnet.

Für einen Punkt im Abstand R von der z-Achse ist

$$x = R\cos\varphi \quad \text{und} \quad y = R\sin\varphi\,,$$

zusammen mit den obigen Formeln erhalten wir also als Zusammenhang zwischen Kugelkoordinaten und kartesischen Koordinaten

$$x = r\cos\varphi\sin\vartheta$$
$$y = r\sin\varphi\sin\vartheta$$
$$z = r\cos\vartheta\,.$$

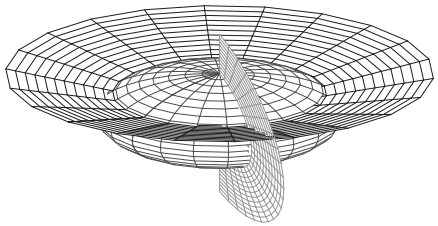


Abb. 44: Flächen mit konstantem r, φ und ϑ für Kugelkoordinaten

Die Flächen mit konstantem r sind hier natürlich Kugeln, konstantes φ führt wie bei den Zylinderkoordinaten auf Halbebenen, und die Flächen mit konstantem ϑ schließlich sind Kegel um die z-Achse.

Wieder können wir einer Funktion $f(x,y,z)$ in kartesischen Koordinaten durch

$$F(r,\varphi,\vartheta) \underset{\text{def}}{=} f(r\cos\varphi\sin\vartheta, r\sin\varphi\sin\vartheta, r\cos\vartheta)$$

eine neue Funktion zuordnen und durch Anwendung der Kettenregel sowie Lösen eines linearen Gleichungssystems die partiellen Ableitungen zueinander in Beziehung setzen; die Vorgehensweise ist im wesentlichen wie bei den ebenen Polarkoordinaten, nur daß wir jetzt ein System aus drei linearen Gleichungen bekommen:

$$F_r = f_x\cos\varphi\sin\vartheta + f_y\sin\varphi\sin\vartheta + f_z\cos\vartheta$$
$$\frac{F_\varphi}{r} = -f_x\sin\varphi\sin\vartheta + f_y\cos\varphi\sin\vartheta$$
$$\frac{F_\vartheta}{r} = f_x\cos\varphi\cos\vartheta + f_y\sin\varphi\cos\vartheta - f_z\sin\vartheta\,.$$

Der GAUSS-Algorithmus führt schnell auf die Lösung

$$f_x = F_r\cos\varphi\sin\vartheta - \frac{F_\varphi\sin\varphi}{r\sin\vartheta} + \frac{F_\vartheta\cos\varphi\cos\vartheta}{r}$$
$$f_y = F_r\sin\varphi\sin\vartheta - \frac{F_\varphi\cos\varphi}{r\sin\vartheta} + \frac{F_\vartheta\sin\varphi\cos\vartheta}{r}$$
$$f_z = F_r\cos\vartheta - \frac{F_\vartheta\sin\vartheta}{r}\,.$$

Im Prinzip genauso können auch die zweiten Ableitungen berechnet werden; eine grausame, wenn auch nicht sonderlich schwierige Rechnung führt beispielsweise auf die Formel

$$\begin{aligned}\Delta f &= f_{xx} + f_{yy} + f_{zz}\\ &= F_{rr} + \frac{2F_r}{r} + \frac{F_{\varphi\varphi}}{r^2\sin^2\vartheta} + \frac{F_{\vartheta\vartheta}}{r^2} + \frac{F_\vartheta}{r^2\tan\vartheta}\end{aligned}\,.$$

Als Beispiel wollen wir damit noch einmal die Divergenz des elektrischen Felds einer Ladung q im Nullpunkt berechnen: Diese erzeugt ein Potential

$$U(r,\varphi,\vartheta) = \frac{q}{4\pi\varepsilon_0 r}\,,$$

das Feld ist $\vec{E} = -\operatorname{grad} U$ und

$$\operatorname{div}\vec{E} = -\operatorname{div}\operatorname{grad} U = -\Delta U = -\frac{q}{4\pi\varepsilon_0}\left(\frac{2}{r^3} + \frac{2}{r}\cdot\frac{-1}{r^2}\right) = 0\,.$$

§3: Integralrechnung

Die Integralrechnung ist neben der Differentialrechnung die zweite wichtige Säule der Analysis. Wir betrachten zunächst den eindimensionalen Fall, der sich auch für die mehrdimensionale Theorie an mehreren Stellen als wichtig erweisen wird.

a) Heuristische Vorüberlegungen

Seiner großen Bedeutung entsprechend, wollen wir uns dem Integralbegriff zunächst heuristisch nähern, bevor wir ihn – mit beträchtlichem technischem Aufwand – im nächsten Paragraphen formal einführen.

Die Integration dient primär drei Zwecken:

- Sie dient zur Umkehrung der Differentiation.
- Sie dient zur Flächenberechnung.
- Sie dient zur Durchschnittsberechnung.

Wie wollen uns alle drei Aspekte kurz ansehen.

1) Integration als Umkehrung der Differentiation: Das klassische Beispiel zur Einführung der Differentiation ist die Geschwindigkeit als Ableitung des Wegs nach der Zeit. Es liegt daher nahe, zur Einführung des Integrals die Frage nach der Berechnung des Wegs aus der Geschwindigkeit zu betrachten.

Wir nehmen also an, ein Fahrzeug sei zur Zeit $t = 0$ an der Stelle $s = 0$ und beschleunige in fünf Sekunden auf die im Stadtverkehr zulässige Hchstgeschwindigkeit von 13,9 m/sec. Welchen Weg hat es bis dahin zurückgelegt?

Die Geschwindigkeit sei durch folgende Tabelle gegeben:

t [sec]	0	1	2	3	4	5
v [m/sec]	2,5	6,1	9,1	11,5	13,4	13,9

Ist $s(t)$ der bis zum Zeitpunkt t zurückgelegte Weg und $v(t)$ die dann erreichte Geschwindigkeit, so ist bekanntlich

$$v(t) = \frac{ds}{dt} \approx \frac{\Delta s}{\Delta t} = \frac{s(t + 1\,sec) - s(t)}{(t + 1\,sec) - t} = \frac{s(t + 1\,sec) - s(t)}{1\,sec}\,.$$

Also ist

$$s(t + 1\,sec) \approx s(t) + v(t) \cdot 1\,sec\,.$$

Da wir sowohl $s(0\,sec) = 0\,m$ als auch die Werte $v(t)$ für $t = 0, 1, 2, 3$ und 4 *sec* kennen, können wir $s(5\,sec)$ rekursiv berechnen als

$$\begin{aligned} s(5\,sec) &\approx v(0\,sec) \cdot 1\,sec + v(1\,sec) \cdot 1\,sec + v(2\,sec) \cdot 1\,sec \\ &\quad + v(3\,sec) \cdot 1\,sec + v(4\,sec) \cdot 1\,sec \\ &= 2{,}5\,m + 6{,}1\,m + 9{,}1\,m + 11{,}5\,m + 13{,}4\,m = 42{,}6\,m\,. \end{aligned}$$

Also hat das Fahrzeug in fünf Sekunden 42,6 m zurückgelegt?

Zumindest mit dieser Genauigkeit ist das sicherlich falsch, denn bei der Berechnung des Wegs sind wir davon ausgegangen, daß das Fahrzeug während der ersten Sekunde konstant mit einer Geschwindigkeit von 2,5 m/sec gefahren ist, exakt ab deren Ende dann aber plötzlich eine Sekunde lang mit 6,1 m/sec, *usw.* Das ist natürlich unrealistisch; tatsächlich dürfte die Geschwindigkeit etwa so verlaufen sein, wie es die Kurve in Abbildung 45 angibt, wohingegen wir mit dem ebenfalls eingezeichneten stufenförmigen Verlauf gerechnet haben. Wir haben die Geschwindigkeit somit fast durchgängig unterschätzt und damit auch einen zu kleinen Weg berechnet.

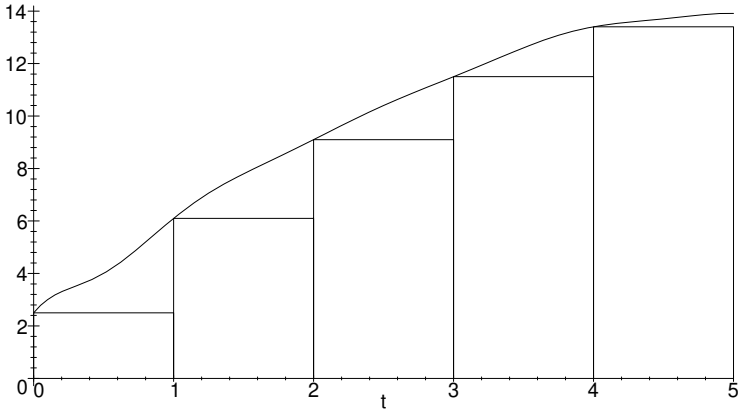


Abb. 45: Tatsächliche und unterschätzte Geschwindigkeit

Alternativ hätten wir auch einen *zu großen* Weg schätzen können, wenn wir die Geschwindigkeit während eines Sekundenintervalls jeweils auf den bekannten Wert am *Ende* dieses Intervalls gesetzt hätten; das Ergebnis wäre dann gewesen

$$\begin{aligned} s(5\,sec) &\approx v(1\,sec) \cdot 1\,sec + v(2\,sec) \cdot 1\,sec + v(3\,sec) \cdot 1\,sec \\ &\quad + v(4\,sec) \cdot 1\,sec + v(5\,sec) \cdot 1\,sec \\ &= 6{,}1\,m + 9{,}1\,m + 11{,}5\,m + 13{,}4\,m + 13{,}9\,m = 54{,}0\,m\,. \end{aligned}$$

Tatsächlich wissen wir im Augenblick also nur, daß der zurückgelegte Weg irgendwo zwischen 42,6 m und 54 m liegt.

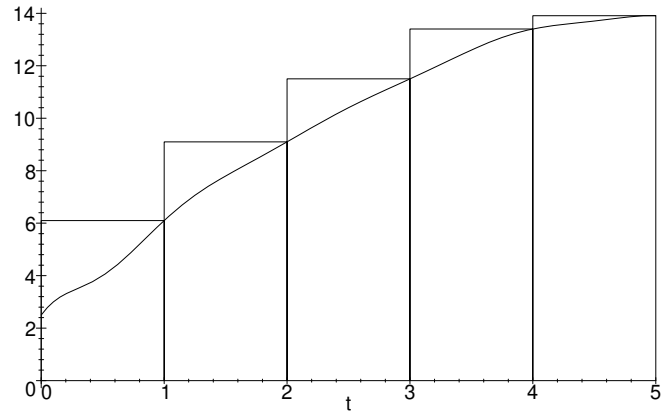


Abb. 46: Tatsächliche und überschätzte Geschwindigkeit

Eine bessere Schätzung bekommen wir, wenn wir anstelle der Sekundenintervalle Intervalle von nur einer halben Sekunde betrachtet – vorausgesetzt natürlich, wir kennen die Geschwindigkeit auch für halbzahlige Sekundenwerte. Die entsprechenden Werte seien die in der folgenden Tabelle angegebenen:

t [sec]	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
v [m/sec]	2,5	4,4	6,1	7,8	9,1	10,4	11,5	12,6	13,4	13,8	13,9

Jetzt ist die „unterschätzte“ Wegstrecke

$$\sum_{i=0}^{9} v\left(\frac{i}{2}\,sec\right)\cdot\frac{1}{2}\,sec = 45{,}6\,m\,,$$

und die „überschätzte“ ist

$$\sum_{i=1}^{10} v\left(\frac{i}{2}\,sec\right)\cdot\frac{1}{2}\,sec = 51{,}3\,m\,,$$

die Unsicherheit hat sich also etwa halbiert. Abbildung 47 zeigt den Verlauf von unterschätzter, tatsächlicher und überschätzter Geschwindigkeit für die Halbsekundenintervalle.

Beide Werte wie auch ihre Differenz lassen sich anhand von Abbildung 47 leicht geometrisch veranschaulichen: Die unterschätzte Ge-

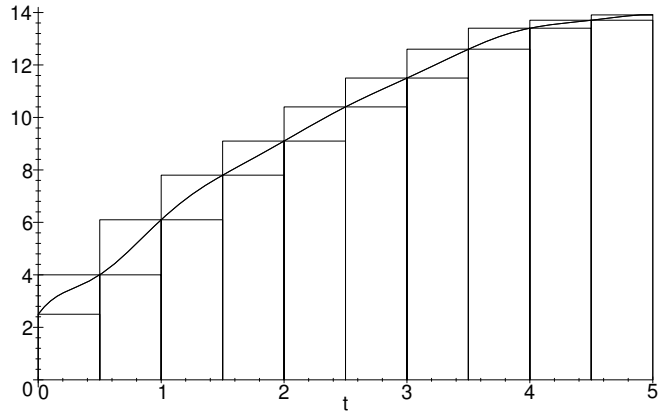


Abb. 47: Abtastung in Intervallen von einer halben Sekunde

schwindigkeit ist die Fläche der Rechtecke unterhalb der unteren Treppenfunktion, die überschätzte entsprechend die Fläche der Rechtecke unterhalb der oberen Treppenfunktion, wobei die Basis der Rechtecke jeweils auf der Zeitachse liegt. Die Differenz ist somit gleich der Fläche der Differenzrechtecke.

Diese können wir durch weitere Verfeinerung der Abtastung verkleinern. Dazu nehmen wir an, die eingezeichnete Kurve (die tatsächlich einfach eine notdürftig angepaßte Polynomfunktion darstellt) gebe den tatsächlichen Geschwindigkeitsverlauf wieder, und lassen den Computer rechnen:

Wenn wir die Geschwindigkeit viermal pro Sekunde bestimmen und den zurückgelegten Weg damit schätzen, erhalten wir als untere Schranke

$$\sum_{i=0}^{19} v\left(\frac{i}{4}\,sec\right)\cdot\frac{1}{4}\,sec \approx 47{,}14\,m$$

und als obere Schranke

$$\sum_{i=1}^{20} v\left(\frac{i}{4}\,sec\right)\cdot\frac{1}{4}\,sec \approx 49{,}99\,m\,.$$

Mit zehn Geschwindigkeitswerten pro Sekunde verbessert sich die untere Schranke auf

$$\sum_{i=0}^{49} v\left(\frac{i}{10}\,sec\right)\cdot\frac{1}{10}\,sec \approx 48{,}02\,m$$

und die obere auf

$$\sum_{i=1}^{50} v\left(\frac{i}{10}\,sec\right)\cdot\frac{1}{10}\,sec \approx 49{,}16\,m\,.$$

Wie Abbildung 48 zeigt, unterscheiden sich nun die Rechtecke der unteren Abschätzung nur noch wenig von denen der oberen, und die Fläche unter der Geschwindigkeitskurve liegt schon recht nahe bei der Fläche der Rechtecke aus jeder der beiden Abschätzungen.

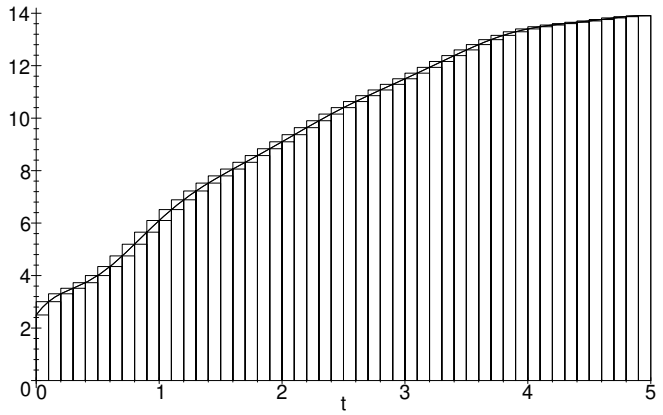


Abb. 48: Abtastung in Intervallen von einer Zehntelsekunde

Gehen wir von den Zehntelsekunden zu den Hundertsteln, sind die Rechtecke zumindest visuell in Abbildung 49 nicht mehr voneinander und von der Fläche unter der Kurve zu unterscheiden: Daß man hier keine durchweg schwarze Fläche sieht, liegt ausschließlich an sogenannten *alias*-Effekten, d.h. Diskretisierungsfehlern der Rastergraphik.

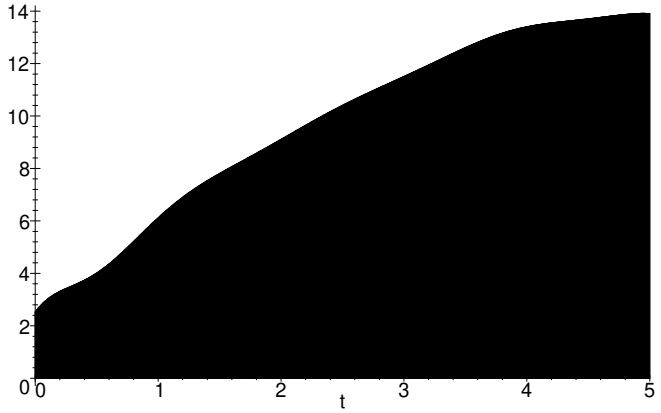


Abb. 49: Abtastung in Intervallen von einer Hundertstelsekunde

Rechnerisch gibt es noch einen kleinen Unterschied im Dezimeterbereich:

$$\sum_{i=0}^{499} v\left(\frac{i}{100}\,sec\right)\cdot\frac{1}{100}\,sec \approx 48{,}54\,m$$

und

$$\sum_{i=1}^{500} v\left(\frac{i}{100}\,sec\right)\cdot\frac{1}{100}\,sec \approx 48{,}66\,m\,.$$

Bei einer nochmaligen Verzehnfachung der Intervallanzahl ist natürlich graphisch nichts neues mehr zu sehen; die Schätzwerte verbessern sich auf

$$\sum_{i=0}^{4999} v\left(\frac{i}{1000}\,sec\right)\cdot\frac{1}{1000}\,sec \approx 48{,}595\,m$$

für die untere Schranke und

$$\sum_{i=1}^{5000} v\left(\frac{i}{1000}\,sec\right)\cdot\frac{1}{1000}\,sec \approx 48{,}607\,m$$

für die obere.

Um die eingangs gestellte Frage nach dem zurückgelegten Weg millimetergenau beantworten zu können (sofern man diese Genauigkeit wirklich als sinnvoll betrachtet), muß die Geschwindigkeitskurve 10 000 Mal pro Sekunde abgetastet werden, und wir erhalten die Schranken

$$\sum_{i=0}^{49999} v\left(\frac{i}{10000}\,sec\right)\cdot\frac{1}{10000}\,sec \approx 48{,}6006\,m$$

und

$$\sum_{i=1}^{50000} v\left(\frac{i}{10000}\right)\,sec\cdot\frac{1}{10000}\,sec \approx 48{,}6017\,m\,;$$

das Fahrzeug legte in fünf Sekunden also 48 m und 60,1 cm zurück.

2) Integration als Flächenbestimmung: Die Abbildungen 48 und vor allem 49 zeigen, daß die immer feiner werdenden Rechtecke, die wir für die obigen Abschätzungen wählten, die Fläche unter der Kurve $v = v(t)$ immer besser annähern. Wenn wir die Einheiten vergessen und t wie auch v als Längen ansehen, können wir die oben betrachteten Ausdrücke

$$\sum_{i=0}^{5n-1} v\left(\frac{i}{n}\right)\cdot\frac{1}{n} \quad\text{und}\quad \sum_{i=1}^{5n} v\left(\frac{i}{n}\right)\cdot\frac{1}{n}$$

also auch als Näherungswerte für die Fläche unter der Kurve ansehen und den Flächeninhalt als ihren gemeinsamen Grenzwert für immer größer werdendes n berechnen.

Genau genommen handelt es sich hier allerdings nicht um die *Berechnung* dieser Fläche, sondern um die *Definition* des Flächeninhalts, denn es gibt schließlich *a priori* keinen Begriff des Flächeninhalts einer krummlinig begrenzten Fläche. Die hier gewählte Definition über eine Ausschöpfung mit immer feiner werdenden Rechtecken ist zwar sehr natürlich, aber nicht zwangsläufig. Sie ist übrigens weit älter als jeder Begriff eines Integrals oder auch nur Grenzwerts: Bereits vor über zwei Jahrtausenden, um etwa 370 vor Christus, definierte der griechische Mathematiker EUDOXOS VON CNIDOS (*ca.* 408–*ca.* 355), ein Schüler und späterer Konkurrent PLATOS, Flächeninhalte auf diese Weise; später hat ARCHIMEDES VON SYRAKUS (287–212) diese Methode perfektioniert

und angewandt auf Kreise und Parabeln; insbesondere berechnete er damit seine Abschätzung $223/71 < \pi < 22/7$ für π, in Dezimalzahlen ausgedrückt $3{,}1408 < \pi < 3{,}1429$, wobei diese Abschätzung allerdings nicht auf Rechtecken beruht, sondern auf der Konstruktion des regelmäßigen 96-Ecks.

3) Integration als Durchschnittsbestimmung: Kehren wir wieder zurück zum Eingangsbeispiel eines sich beschleunigenden Fahrzeugs, und fragen wir uns, wie hoch die *Durchschnittsgeschwindigkeit* war. Bei endlich vielen Werten ist der Durchschnitt natürlich einfach das arithmetische Mittel, also z.B.

$$\begin{aligned}&\frac{1}{5}\cdot\Big(v(0\,sec)+v(1\,sec)+v(2\,sec)+v(3\,sec)+v(4\,sec)\Big)\\ =\;&\frac{1}{5}\cdot 42{,}6\,m/sec = 8{,}52\,m/sec\,,\end{aligned}$$

wenn wir die Geschwindigkeiten vom *Anfang* jedes Sekundenintervalls nehmen, und

$$\begin{aligned}&\frac{1}{5}\cdot\Big(v(1\,sec)+v(2\,sec)+v(3\,sec)+v(4\,sec)+v(5\,sec)\Big)\\ =\;&\frac{1}{5}\cdot 54{,}0\,m/sec = 10{,}8\,m/sec\,,\end{aligned}$$

wenn wir die vom Ende nehmen.

Natürlich läßt sich auch hier die Abtastung immer weiter verfeinern; die entstehenden Ausdrücke

$$\frac{1}{5n}\cdot\sum_{i=0}^{5n-1} v\left(\frac{i}{n}\,sec\right) \quad\text{und}\quad \frac{1}{5n}\cdot\sum_{i=1}^{5n} v\left(\frac{i}{n}\,sec\right)$$

entsprechen bis auf einen Faktor von $\frac{1}{5\,sec}$ genau denen aus Teil *a),* wo wir den Weg abgeschätzt haben – wie es ja auch in der Tat der Fall sein muß: Die Durchschnittsgeschwindigkeit ist schließlich nichts anderes als der zurückgelegte Weg dividiert durch die zugrundeliegende Zeitspanne von fünf Sekunden.

Trotzdem ist die Interpretation eines Integrals als Durchschnitt gelegentlich von unabhängigem Interesse, da vor allem bei Anwendungen in der

Quantenphysik oft zwar Durchschnitte existieren, aber keine „Zähler“ und „Nenner“, als deren Quotienten man sie interpretieren könnte.

b) Integration elementarer Funktionen

Abstrahieren wir vom Beispiel der Wegberechnung anhand einer Geschwindigkeitskurve und betrachten wir das allgemeine Problem, zu einer gegebenen reellwertigen Funktion f eine neue Funktion F zu finden derart, daß $F'(x) = f(x)$ ist.

Da jede konstante Funktion die Ableitung Null hat, ist mit F für jede reelle Zahl c auch $F + c$ eine solche Funktion; um ein konkretes F hinzuschreiben, müssen wir also einen Funktionswert von F festlegen; der Einfachheit halber sei dies, in Analogie zum obigen Beispiel, der Wert $F(0) = 0$.

Damit übernimmt f die Rolle der Geschwindigkeit v, dem Weg entspricht die Funktion F, die Variable ist nun x anstelle der Zeit t, und da uns F an einer beliebigen Stelle x interessiert, nicht nur wie im obigen Beispiel an der Stelle 5, empfiehlt es sich, n jetzt als Gesamtzahl der Intervalle zu nehmen, nicht wie oben als Anzahl der Intervalle pro Einheit (Sekunde). Die Länge eines jeden Intervalls wird dann zu x/n, und die Analoga zu obigen Summen sind die Ausdrücke

$$\sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} \quad \text{und} \quad \sum_{i=1}^{n} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n},$$

deren gemeinsamer Grenzwert für immer größer werdendes n der gesuchte Wert $F(x)$ sein sollte.

Wir *definieren* daher

$$F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} \qquad (*)$$

in der Hoffnung, daß dieser Grenzwert existiert und mit dem anderen übereinstimmt – zumindest in hinreichend vielen interessanten Fällen. Bevor wir uns im nächsten Paragraphen dieser Frage zuwenden, wollen wir hier zunächst etwas mit dieser vorläufigen Definition spielen und sehen, was sie bei einfachen Funktionen liefert.

1) Die Funktion $f(x) = x^2$***:*** Hier erhalten wir

$$F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} = \lim_{n\to\infty} \sum_{i=0}^{n-1} \left(\frac{ix}{n}\right)^2 \cdot \frac{x}{n}$$

$$= x^3 \lim_{n\to\infty} \frac{1}{n^3} \sum_{i=0}^{n-1} i^2 .$$

Wie man in seiner Formelsammlung nachschlägt oder sich von seinem Computeralgebrasystem sagen läßt, ist

$$\sum_{i=0}^{n-1} i^2 = \frac{n(2n-1)(n-1)}{6},$$

eine Formel, die dann nachträglich leicht durch vollständige Induktion bewiesen werden kann:

Für $n = 1$ steht links und rechts Null, die Formel ist also korrekt.

Angenommen, wir hätten sie bewiesen für $n-1$ anstelle von n, d.h. wir wüßten, daß für ein spezielles $n > 1$

$$\sum_{i=0}^{n-2} i^2 = \frac{(n-1)(2n-3)(n-2)}{6}$$

ist. Dann folgt

$$\begin{aligned}
\sum_{i=0}^{n-1} i^2 &= \frac{(n-1)(2n-3)(n-2)}{6} + (n-1)^2 \\
&= \frac{n-1}{6} \cdot \big((2n-3)(n-2) + 6(n-1)\big) \\
&= \frac{n-1}{6} \cdot (2n^2 - 7n + 6 + 6n - 6) \\
&= \frac{n-1}{6} \cdot (2n^2 - n) = \frac{(n-1)(2n-1)n}{6} \\
&= \frac{n(2n-1)(n-1)}{6},
\end{aligned}$$

d.h. die Formel für n selbst.

Nach dem Prinzip der vollständigen Induktion ist diese damit für *alle* Werte von n bewiesen; wir können sie in obige Rechnung einsetzen und erhalten

$$F(x) = x^3 \lim_{n\to\infty} \frac{1}{n^3} \cdot \frac{n(2n-1)(n-1)}{6} = x^3 \lim_{n\to\infty} \frac{n(2n-1)(n-1)}{6n^3}$$
$$= \frac{x^3}{6} \lim_{n\to\infty} \left(\frac{n}{n}\right)\left(\frac{2n-1}{n}\right)\left(\frac{n-1}{n}\right) = \frac{x^3}{6}\cdot 2 = \frac{x^3}{3}\,.$$

Diese Funktion hat tatsächlich die Ableitung $F'(x) = x^2$, zumindest in diesem Beispiel funktioniert also die Definition $(*)$.

2) Die Exponentialfunktion: Versuchen wir dasselbe nochmal für die Funktion $f(x) = e^x$. Hier führt $(*)$ (wieder mit der Festlegung $F(0) = 0$) auf den Ansatz

$$F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right)\cdot\frac{x}{n} = \lim_{n\to\infty}\sum_{i=0}^{n-1} e^{\frac{ix}{n}}\cdot\frac{x}{n}$$
$$= x \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1}\left(e^{\frac{x}{n}}\right)^i\,.$$

Die Summe ganz rechts ist eine geometrische Reihe und kann leicht ausgerechnet werden: Setzen wir zur Abkürzung $q = e^{x/n}$, so ist

$$\sum_{i=0}^{n-1} q^i = 1 + q + q^2 + q^3 + \cdots + q^{n-1}$$
$$q\sum_{i=0}^{n-1} q^i = \quad q + q^2 + q^3 + \cdots + q^{n-1} + q^n$$
$$\Rightarrow (1-q)\sum_{i=0}^{n-1} q^i = 1 \qquad\qquad\qquad\qquad - q^n\,.$$

Für $q \neq 1$ können wir durch $1-q$ dividieren und erhalten die Formel

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

Da uns nur der Fall $x \neq 0$ interessiert, ist $q = e^{\frac{x}{n}} \neq 1$; wir können die Formel also anwenden und erhalten

$$F(x) = x \lim_{n\to\infty}\frac{1}{n}\cdot\frac{1-\left(e^{\frac{x}{n}}\right)^n}{1-e^{\frac{x}{n}}} = x\lim_{n\to\infty}\frac{1}{n}\cdot\frac{1-e^x}{1-e^{\frac{x}{n}}}$$
$$= x(1-e^x)\lim_{n\to\infty}\frac{\frac{1}{n}}{1-e^{\frac{x}{n}}}\,.$$

Im Grenzwert ganz rechts gehen für $n\to\infty$ leider sowohl der Zähler als auch der Nenner gegen Null, wir müssen also die Regel von DE L'HOSPITAL anwenden. Diese gilt zwar nur für Grenzwerte von Werten stetiger Funktionen, während wir hier den Grenzwert einer diskreten Folge suchen, aber wenn für eine reelle Variable u der Grenzwert

$$\lim_{u\to\infty}\frac{\frac{1}{u}}{1-e^{\frac{x}{u}}}$$

existiert, konvergiert natürlich auch die Folge der Zahlen

$$\frac{\frac{1}{n}}{1-e^{\frac{x}{n}}} \quad \text{für } n = 1,2,3,\ldots$$

gegen eben diesen Grenzwert.

Nach DE L'HOSPITAL ist

$$\lim_{u\to\infty}\frac{\frac{1}{u}}{1-e^{\frac{x}{u}}} = \lim_{u\to\infty}\frac{\frac{d}{du}\frac{1}{u}}{\frac{d}{du}(1-e^{\frac{x}{u}})} = \lim_{u\to\infty}\frac{\frac{-1}{u^2}}{-e^{\frac{x}{u}}\left(\frac{-x}{u^2}\right)}$$
$$= \lim_{u\to\infty}\frac{-1}{xe^{\frac{x}{u}}} = \frac{-1}{x}\,.$$

Eingesetzt in den Ansatz für $F(x)$ führt dies auf

$$F(x) = x(1-e^x)\cdot\frac{-1}{x} = e^x - 1\,,$$

eine Funktion, deren Ableitung in der Tat e^x ist.

3) Die Dirichletsche Sprungfunktion: Bevor wir zu übermütig werden, möchte ich als letztes Beispiel noch eine Funktion betrachten, die sich

im Gegensatz zu Quadrat und Exponentialfunktion alles andere als gut verhält, die DIRICHLETsche Sprungfunktion

$$f(x) = \begin{cases} 1 & \text{für } x \in \mathbb{Q} \\ 0 & \text{für } x \notin \mathbb{Q} \end{cases} .$$

Mit $F(0) \underset{\text{def}}{=} 0$ ist wieder

$$F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} .$$

Dabei ist die Zahl ix/n für $i = 0$ rational, da Null; für $i \neq 0$ ist sie genau dann rational, wenn auch x rational ist. In diesem Fall ist also $f(ix/n) = 1$ für alle i, d.h.

$$F(x) = \lim_{n\to\infty} \sum_{i=0}^{n-1} \frac{x}{n} = \lim_{n\to\infty} n \cdot \frac{x}{n} = x$$

für rationales x. Für irrationales x dagegen sind alle ix/n außer der Null irrational, also ist $f(ix/n) = 0$ für alle $i \neq 0$. Damit bleibt von der Summe nur der erste Summand stehen, d.h.

$$F(x) = \lim_{n\to\infty} \frac{x}{n} = 0 .$$

Also ist

$$F(x) = \begin{cases} x & \text{für } x \in \mathbb{Q} \\ 0 & \text{für } x \notin \mathbb{Q} \end{cases} ,$$

eine offensichtlich nicht differenzierbare Funktion, und wenn $F'(x)$ nicht einmal existiert, kann es natürlich auch nicht gleich $f(x)$ sein.

Auch unter dem Gesichtspunkt der Interpretation von $F(x)$ als Fläche unterhalb der Kurve $y = f(x)$ ist das Ergebnis unbefriedigend, denn da $f(x) \geq 0$ für alle x, sollte $F(x)$ eine monoton wachsende (oder zumindest nicht fallende) Funktion von x sein, wohingegen das gerade berechnete F bei jeder rationalen Zahl auf Null zurückfällt.

Der Ansatz $(*)$ hat also auch seine Tücken, und es wird Zeit, diese heuristische Definition durch eine bessere zu ersetzen.

c) Definition des Riemann-Integrals

Tatsächlich ist der Ansatz $(*)$ nicht so schlecht: Für die beiden gutartigen Funktionen und für das Eingangsbeispiel der Streckenbestimmung führte er schließlich zu vernünftigen Ergebnissen, und wie wir bald sehen werden, kann man Integrale über stetige Funktionen *immer* nach diesem Ansatz bestimmen; zumindest als Veranschaulichung des Integralbegriffs sollte man ihn daher durchaus im Hinterkopf behalten. Problematisch wird er erst bei „schlechten“ Funktionen wie der im letzten Beispiel.

1) Warum lohnt sich ein allgemeinerer Ansatz?: Dies legt natürlich die Frage nahe, ob man solche „schlechten“ Funktionen für Anwendungen wirklich braucht, und wenn ja, ob man sie so sehr braucht, daß sich der erhebliche technische Aufwand, den wir in diesem Paragraphen treiben müssen, lohnt.

Die DIRICHLETsche Sprungfunktion ist ein rein theoretisch konstruiertes Gegenbeispiel, das wohl keinerlei praktische Anwendung haben dürfte. Eine ihrer charakteristischen Eigenschaften taucht allerdings auch bei praktisch relevanten Funktionen auf: Auch ein periodisch wiederholter Rechteckimpuls hat abzählbar unendlich viele Sprungstellen, und zumindest als Idealisierung eines realen Signals spielt diese Funktion eine Rolle – sie ist allerdings erheblich harmloser als DIRICHLETS Beispiel, da die Menge ihrer Sprungstellen diskret ist.

Wirklich kompliziert wird die Situation dagegen beispielsweise bei der mathematischen Modellierung des Rauschens in einer elektronischen Schaltung: Hier liegt eine noch einmal deutlich schwierigere Situation vor als im obigen Beispiel, und in der Tat wird auch der in diesem Paragraphen definierte Integralbegriff nicht ausreichen, um mit diesem Problem fertig zu werden. Er ist aber eine unabdingbare Voraussetzung, um die dazu bentigten Techniken zu verstehen.

Obwohl wir im folgenden fast jeden der wichtigeren Sätze aus der Analysis I noch einmal in die Erinnerung zurückrufen und anwenden müssen, ist die nun folgende Konstruktion also kein Luxus, sondern zumindest langfristig notwendig auch für praktische Anwendungen.

2) Wo sollte der bisherige Ansatz modifiziert werden?: Ein Punkt, bei dem der bisherige Ansatz nur aufgrund der Einfachheit der betrachteten Beispiele so erfolgreich war, ist die *Abschätzung* des Grenzwerts durch die endlichen Approximationen. Beim Beispiel der Geschwindigkeit eines beschleunigenden Fahrzeugs war das problemlos, denn die Geschwindigkeit war eine monoton wachsende Funktion, die für jedes Teilintervall am linken Ende ihr Minimum und am rechten ihr Maximum annimmt.

Auch $f(x) = x^2$ ist für $x \geq 0$ monoton wachsend, $f(x) = e^x$ sogar für alle $x \in \mathbb{R}$, so daß wir auch hier leicht untere und obere Schranken für $F(x)$ berechnen können.

Für nichtmonotone Funktionen dagegen ist die Situation völlig anders: Betrachten wir als Beispiel etwa die Sinuslinie zwischen Null und π mit einer Annäherung der Fläche durch sechs Rechtecke. In den Abbildungen 50 und 51 sind diese Rechtecke eingezeichnet, einmal bezogen auf den Funktionswert am linken Ende des Teilintervalls und einmal bezogen auf den am rechten. Wie man sieht, liegen die Rechtecke exakt spiegelsymmetrisch zueinander und haben damit insbesondere die gleiche Summe der Flächeninhalte, nämlich jeweils ungefähr 1,954. Die Fläche unter der Sinuslinie zwischen Null und π ist allerdings, wie wir bald sehen werden,

$$-\cos\pi - (-\cos 0) = -(-1) - (-1) = 1 + 1 = 2\,.$$

Wenn man ein (jedem Wissenschaftler zu empfehlendes) gesundes Mißtrauen gegen Computer und Taschenrechner mitbringt, könnte man versucht sein, den Wert 1,954 als eine „Taschenrechnerzwei“ zu interpretieren; da aber die Sinuswerte an den Stellen $0, \frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{4\pi}{3}, \frac{5\pi}{6}$ und 1 wohlbekannt sind (im Winkelmaß ausgedrückt sind diese Stellen gerade die Vielfachen von 30°), kann man die Summe der Rechteckflächen exakt berechnen mit dem Ergebnis

$$\frac{\pi}{6}(2+\sqrt{3})\,,$$

das schon wegen der Transzendenz von π nicht gleich zwei sein kann. Somit haben wir definitiv keine obere Schranke für den korrekten Wert, und zumindest *a priori* haben wir auch keine untere, denn daß die

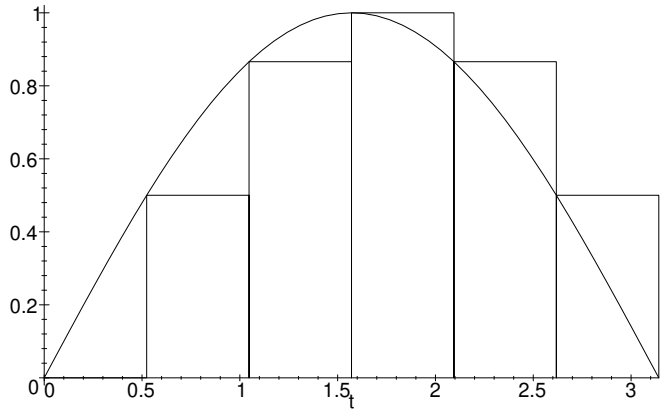


Abb. 50: Sinuslinie mit sechs Rechtecken zum jeweils linken Funktionswert

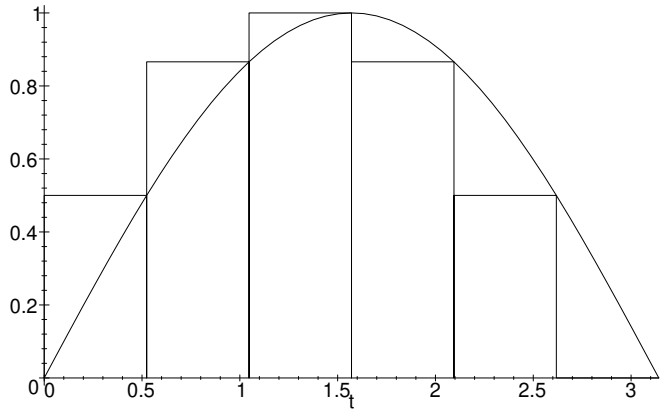


Abb. 51: Sinuslinie mit sechs Rechtecken zum jeweils rechten Funktionswert

Summe der Rechteckflächen kleiner ist als die Fläche unter der Kurve folgte je erst nachträglich durch Vergleich der beiden Werte.

Um wirklich eine untere Schranke für die Fläche zu bekommen, müßte man hier im Beispiel für die ersten drei Rechtecke den Funktionswert am linken Ende des Teilintervalls nehmen und für die letzten drei den vom rechten; für eine obere Schranke müßte man genau umgekehrt vorgehen.

Für eine beliebige Funktion gibt es aber offensichtlich keinen Grund, warum das Minimum oder Maximum überhaupt an einem Intervallende angenommen werden sollte; für einen möglichst allgemeinen Integrallbegriff sollte man also auch Punkte im Intervallinnern zur Ermittlung der Höhe des Rechtecks heranziehen: Anstelle der bislang betrachteten Unterteilung

$$a < a+\delta < a+2\delta < \cdots < a+(n-1)\delta < a+n\delta = b \quad \text{mit} \quad \delta = \frac{b-a}{n}$$

sollte also eine beliebige Unterteilung

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

treten.

3) Anwendung des Mittelwertsatzes: Einen weiteren Grund dafür liefert der wohl wichtigste Satz aus der Analysis I, der *Mittelwertsatz der Differentialrechnung:* Ist F stetig in einem Intervall $[u, v]$ und differenzierbar in (u, v), so gibt es einen Punkt $\xi \in (u, v)$, so daß

$$F'(\xi) = \frac{f(v) - f(u)}{v - u}$$

ist.

Wir suchen zu einer gegebenen Funktion f eine differenzierbare Funktion F mit $F' = f$. Falls wir annehmen, daß wir diese Funktion schon *hätten,* so wäre $F' = f$, und nach dem Mittelwertsatz gäbe es in jedem Intervall (x_i, x_{i+1}) ein Element ξ_i, so daß

$$\frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i} = F'(\xi_i) = f(\xi_i)$$

wäre, d.h.

$$F(x_{i+1}) = F(x_i) + f(\xi_i)(x_{i+1} - x_i)\,.$$

Rekursiv folgt, daß

$$\begin{aligned} F(x) = F(x_n) &= F(x_{n-1}) + f(\xi_{n-1})(x_n - x_{n-1}) \\ &= F(x_{n-2}) + f(\xi_{n-1})(x_{n-1} - x_{n-2}) + f(\xi_{n-1})(x_n - x_{n-1}) \\ &= \quad \vdots \\ &= F(x_0) + \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i)\,. \end{aligned}$$

Damit ist also die Fläche unter der Kurve $y = f(x)$ *exakt* gleich der Gesamtfläche endlich vieler geeignet gewählter Rechtecke; Abbildung 52 veranschaulicht dies anhand der ganz zu Beginn betrachteten Geschwindigkeitsfunktion. Hier ist nicht nur die Fläche unter der Kurve exakt gleich der Fläche der fünf Rechtecke, sondern auch die Fläche eines jeden Rechtecks gleich der Fläche zwischen Kurve und Grundlinie des Rechtecks, wir kommen also mit endlich vielen Rechtecken aus und brauchen nicht einmal einen Grenzwert zu berechnen.

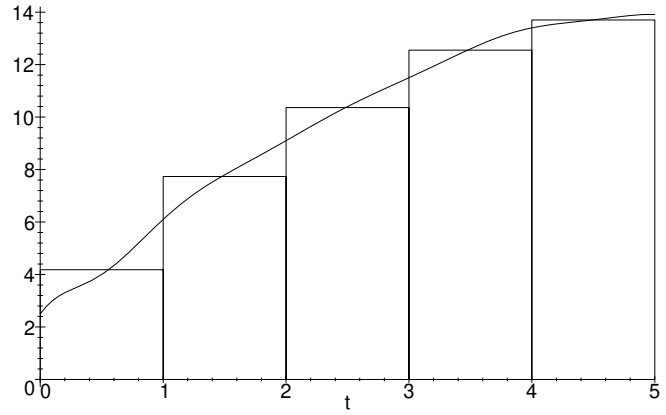


Abb. 52: Die Fläche unter der Kurve ist gleich der Fläche der fünf Rechtecke

Der Haken bei der Sache ist natürlich, daß wir die „geeignet gewählten" Rechtecke nicht kennen: Im obigen Ausdruck ist alles bekannt *außer* den Werten ξ_i, an denen die Funktion f ausgewertet wird.

Die Idee zur Definition des RIEMANN-Integrals ist nun, daß man einfach *beliebige* ξ_i zwischen x_i und x_{i+1} wählt in der Hoffnung, daß bei immer kleiner werdendem Abstand zwischen zwei aufeinanderfolgenden x_i der Grenzwert nicht mehr von der Wahl der ξ_i abhängt.

Diese Hoffnung hat natürlich nur eine Chance auf Erfüllung, wenn die Funktion f hinreichend stetig ist; ansonsten können sich die Werte von f an zwei beliebig nahe beieinanderliegenden Stellen immer noch beliebig stark unterscheiden, indem sie beispielsweise wie oben davon abhängen, ob ξ_i rational ist oder nicht.

4) Gleichmäßige Stetigkeit: Die Stetigkeit allein reicht allerdings immer noch nicht ganz aus:

Erinnern wir uns: f heißt *stetig* auf einer Teilmenge $D \subseteq \mathbb{R}$, wenn es zu jedem $\varepsilon > 0$ und zu jedem $x \in (a,b)$ ein $\delta > 0$ gibt, so daß gilt: Ist $y \in D$ und $|y - x| < \delta$, so folgt, daß $|f(y) - f(x)| < \varepsilon$ ist.

Wir möchten mehr: Wir wollen, daß sich in *jedem* Rechteck der Wert von $f(\xi)$ höchstens um $\varepsilon > 0$ ändert, falls nur die Breite des Rechtecks unter einer gewissen Schranke liegt, d.h. wir möchten, daß δ nicht von x abhängt

Dies führt auf folgende Verschärfung des Stetigkeitsbegriffs:

Definition: $D \subseteq \mathbb{R}$ sei eine Teilmenge von $\mathbb{R}$ und $f: D \to \mathbb{R}$ eine Funktion. f heißt *gleichmäßig stetig* in D, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für alle $x, y \in D$ gilt:

$$|y - x| < \delta \Longrightarrow |f(y) - f(x)| < \varepsilon \,.$$

Um zu sehen, daß dies mehr ist als die bloße Stetigkeit, betrachten wir die Funktion $f(x) = 1/x$. Diese Funktion ist stetig auf der Menge aller positiver reeller Zahlen, denn für $0 < \delta < x$ und $\varepsilon > 0$ ist

$$\left|\frac{1}{x \pm \delta} - \frac{1}{x}\right| = \left|\frac{\mp\delta}{x(x \pm \delta)}\right| = \frac{\delta}{x(x \pm \delta)} < \varepsilon$$

$$\Leftrightarrow \delta < (x^2 \pm x\delta)\varepsilon \Leftrightarrow \delta \mp x\delta\varepsilon < x^2\varepsilon \Leftrightarrow \delta < \frac{x^2\varepsilon}{1 \mp x\varepsilon} \,.$$

Der Ausdruck ganz rechts ist offensichtlich kleiner, wenn im Nenner das Pluszeichen steht, also gilt: Für gegebenes $x > 0$ und $\varepsilon > 0$ gilt für jede positive reelle Zahl y

$$|y - x| < \delta \underset{\text{def}}{=} \frac{x^2\varepsilon}{1 + x\varepsilon} \Rightarrow \left|\frac{1}{y} - \frac{1}{x}\right| < \varepsilon \,.$$

Damit ist die Stetigkeit der Funktion bewiesen. Sie ist aber nicht gleichmäßig stetig, denn es gibt aber kein von x unabhängiges $\delta > 0$ mit dieser Eigenschaft: Der angegebene Ausdruck für das größtmögliche δ geht wegen des Faktors x^2 im Zähler für $x \to 0$ selbst gegen Null, fällt also unter jede vorgegebene positive Schranke. Abbildung 53 zeigt δ in Abhängigkeit von x für den speziellen Wert $\varepsilon = \frac{1}{10}$.

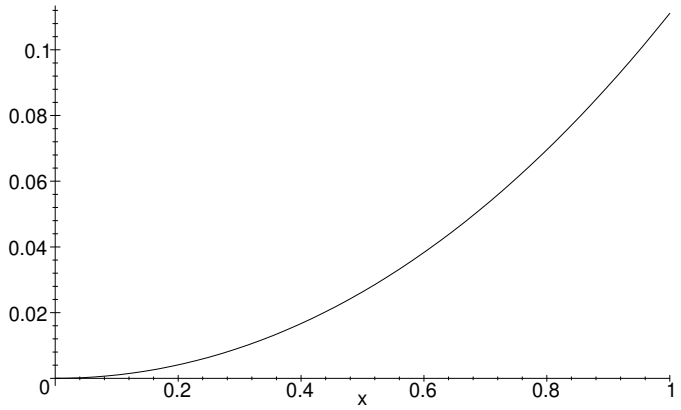


Abb. 53: δ in Abhängigkeit von x für $f(x) = 1/x$ und $\varepsilon = 0{,}1$

In einem abgeschlossenen Intervall sollte so etwas nicht möglich sein, und in der Tat gilt der (durchaus nichttriviale)

Satz: $f: [a,b] \to \mathbb{R}$ sei eine stetige Funktion auf dem abgeschlossenen Intervall $[a,b]$ mit $a, b \in \mathbb{R}$. Dann ist f auf $[a,b]$ gleichmäßig stetig.

Der *Beweis* ist technisch und indirekt:

Angenommen, f wäre *nicht* gleichmäßig stetig. Dann gäbe es für mindestens ein $\varepsilon > 0$ zu jedem $\delta > 0$ Punkte $x, y \in [a, b]$, so daß $|y - x| < \delta$, aber $|f(y) - f(x)| \geq \varepsilon$ wäre.

Aufgrund der Annahme, der Satz sei falsch, können wir ein solches ε fixieren und betrachten die speziellen Werte $\delta = \frac{1}{n}$: Nach dem gerade gesagten gibt es dazu Punkte $x_n, y_n \in [a, b]$, so daß $|y_n - x_n| < \frac{1}{n}$, aber $|f(y_n) - f(x_n)| \geq \varepsilon$ ist.

Nun kommt der nichttriviale Teil: Wir benötigen aus der Analysis I den

Satz von Bolzano-Weierstraß: Jede Folge (x_n) von Punkten aus dem abgeschlossenen Intervall $[a, b]$ hat (mindestens) eine konvergente Teilfolge x_{n_ν}.

Insbesondere muß also die hier betrachtete Folge (x_n) eine konvergente Teilfolge $\left(x_{n_\nu}\right)_{\nu \in \mathbb{N}}$ haben; deren Grenzwert sei $c \in [a, b]$.

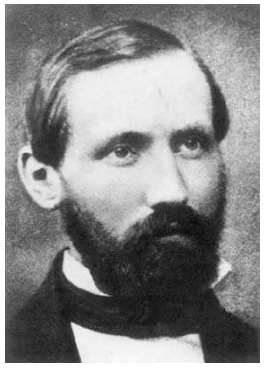
Da $\left|x_{n_\nu} - y_{n_\nu}\right| < 1/n_\nu$ ist, muß dann auch die Folge der y_{n_ν} gegen c konvergieren, und da f nach Voraussetzung eine stetige Funktion ist, gilt

$$\lim_{\nu \to \infty} f(x_{n_\nu}) = \lim_{\nu \to \infty} f(y_{n_\nu}) = f(c)\,.$$

Andererseits ist aber für jedes ν nach Konstruktion der Folgen (x_n) und (y_n) der Abstand zwischen $f(x_{n_\nu})$ und $f(y_{n_\nu})$ größer oder gleich der festgewählten Zahl ε, so daß die beiden Folgen unmöglich denselben Grenzwert haben können.

Mithin führt die Annahme, f sei *nicht* gleichmäßig stetig, auf einen Widerspruch, und damit ist der Satz bewiesen. ∎

5) Definition einer Approximation für das Integral: Nach diesen Vorarbeiten können wir ernsthaft darangehen, das Integral einer Funktion zu definieren. Es gibt in der Mathematik verschiedene Integralbegriffe; für uns reicht die Definition des deutschen Mathematikers BERNHARD RIEMANNN, das inzwischen nach ihm benannte RIEMANN-Integral.



GEORG FRIEDRICH BERNHARD RIEMANNN (1826-1866) war Sohn eines lutherischen Pastors und schrieb sich 1946 auf Anraten seines Vaters an der Universität Göttingen für das Studium der Theologie ein. Schon bald wechselte an die Philosophische Fakultät, um dort unter anderem bei GAUSS Mathematikvorlesungen zu hören. Nach Promotion 1851 und Habilitation 1854 erhielt er dort 1857 einen Lehrstuhl. Trotz seines frühen Todes initiierte er grundlegende auch noch heute fundamentale Entwicklungen in der Geometrie, der Zahlentheorie und über abelsche Funktionen. Seine Vermutung über die Nullstellen der (heute als RIEMANNsch bezeichneten) ζ-Funktion ist die berühmteste offene Vermutung der heutigen Mathematik.

Wir gehen aus von einer Funktion $f\colon [a, b] \to \mathbb{R}$ auf einem abgeschlossenen Intervall $[a, b]$; gesucht ist eine Funktion $F\colon [a, b] \to \mathbb{R}$, für die (idealerweise) $F'(x) = f(x)$ sein sollte für alle x aus dem offenen Intervall (a, b).

Wir wählen eine Unterteilung

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = x$$

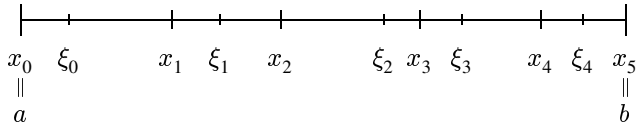
des Intervalls von a bis x. Da wir im folgenden viel mit dieser Unterteilung rechnen werden, führen wir die Abkürzung

$$\boldsymbol{x} = (x_0, x_1, \ldots, x_n)$$

dafür ein und verabreden, daß die x_i immer, wenn von $\boldsymbol{x}$ die Rede ist, die obigen Gleichungen und Ungleichungen erfüllen sollen, daß sie also tatsächlich eine Unterteilung von $[a, x]$ definieren.

Falls die gesuchte differenzierbare Funktion $F\colon [a, b] \to \mathbb{R}$ existiert, sagt uns – wie wir in Teil *c)* gesehen haben – der Mittelwertsatz der Differentialrechnung, daß für geeignete Elemente ξ_i mit

$$a = x_0 < \xi_0 < x_1 < \xi_1 < x_2 < \cdots < \xi_{n-2} < x_{n-1} < \xi_{n-1} < x_n = x$$



folgt

$$F(x) = F(a) + \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i)\,.$$

Natürlich kennen wir die Zahlen ξ_i nicht – selbst wenn wir wissen, daß F und damit die ξ_i überhaupt existieren. Deshalb betrachten wir einfach *beliebige* Zahlen ξ_i mit

$$a = x_0 < \xi_0 < x_1 < \xi_1 < x_2 < \cdots < \xi_{n-2} < x_{n-1} < \xi_{n-1} < x_n = x$$

und führen auch hier wieder die Abkürzung

$$\underline{\xi} = (\xi_0, \xi_1, \ldots, \xi_{n-1})$$

ein mit der Verabredung, daß immer, wenn ein $\underline{\xi}$ in Zusammenhang mit einem $\boldsymbol{x}$ auftritt, diese Ungleichungen erfüllt sein sollen.

Definition: Die RIEMANNsche Summe zu einem Paar $(\boldsymbol{x}, \underline{\xi})$ ist

$$I(\boldsymbol{x}, \underline{\xi}) = \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i)\,.$$

Das RIEMANN-Integral soll der Grenzwert einer Folge RIEMANNscher Summen sein, wobei die Unterteilungen $\boldsymbol{x}$ immer feiner werden. Dieses „Feinerwerden“ müssen wir natürlich auch noch erklären: Seien $\boldsymbol{x}$ und $\boldsymbol{y}$ zwei Unterteilungen des Intervalls $[a, x]$; konkret sei

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = x$$

und

$$a = y_0 < y_1 < \cdots < y_{m-1} < y_m = x\,.$$

Dann heißt $\boldsymbol{y}$ eine *Verfeinerung* von $\boldsymbol{x}$, wenn jedes Teilintervall (y_i, y_{i+1}) von $\boldsymbol{y}$ ganz in einem der Teilintervalle (x_j, x_{j+1}) von $\boldsymbol{x}$ liegt; $\boldsymbol{y}$ entsteht also aus $\boldsymbol{x}$, indem einige der Intervalle von $\boldsymbol{x}$ noch weiter unterteilt werden. Insbesondere muß dann $m \geq n$ sein.

Nun betrachten wir eine Folge $(\boldsymbol{x}^{(\nu)})$ von Unterteilungen derart, daß

- $\boldsymbol{x}^{(\nu+1)}$ stets eine Verfeinerung von $\boldsymbol{x}^{(\nu)}$ ist *und*

- die maximale Länge der Teilintervalle von $\boldsymbol{x}^{(\nu)}$ mit wachsendem ν gegen Null geht, d.h.

$$\lim_{\nu\to\infty} \max_{i=0}^{n_\nu - 1} (x_{i+1}^{(\nu)} - x_i^{(\nu)}) = 0\,.$$

Zu jeder Unterteilung $\boldsymbol{x}^{(\nu)}$ wählen wir willkürlich eine (im Sinne obiger Konvention dazu passende) Folge $\underline{\xi}^{(\nu)}$ von Zwischenpunkten und fragen nach Existenz und gegebenenfalls Wert des Grenzwerts

$$\lim_{\nu\to\infty} I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})\,.$$

Dieser Grenzwert muß natürlich im allgemeinen nicht existieren, und wenn er existiert, kann er von der Wahl der Zwischenpunkte $\xi_i^{(\nu)}$ und von der Wahl einer Folge $(\boldsymbol{x}^{(\nu)})$ von Unterteilungen abhängen. Wenn er unabhängig von all diesen Wahlen existiert und immer denselben Wert hat, bezeichnen wir diesen gemeinsamen Wert als das RIEMANN-*Integral* von f zwischen a und x, in Zeichen

$$\int_a^x f(\xi)\,d\xi = \lim_{\nu\to\infty} I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})\,.$$

Falls dieses Integral für jedes $x \in [a, b]$ existiert, sagen wir, f sei RIEMANN-*integrierbar* auf $[a, b]$.

ξ heißt *Integrationsvariable* und übernimmt die Rolle des Summationsindex in einer Summe: Genau wie beispielsweise

$$\sum_{i=0}^{n-1} i^2 = \sum_{j=0}^{n-1} j^2 = \sum_{k=0}^{n-1} k^2 = \frac{n(2n-1)(n-1)}{6}$$

davon unabhängig ist, ob der Summationsindex i, j oder k heißt, ist auch

$$\int_a^x f(\xi)\,d\xi = \int_a^x f(u)\,du = \int_a^x f(t)\,dt$$

unabhängig davon, ob die Integrationsvariable mit ξ, u oder t bezeichnet wird.

6) Existenz des Riemann-Integrals für stetige Funktionen: Damit haben wir das RIEMANN-Integral definiert; wir wissen allerdings bisher für keine einzige Funktion, daß es existiert. In diesem Abschnitt wollen wir uns überlegen, daß es zumindest für *stetige* Funktionen immer existiert:

Satz: Jede stetige Funktion $f\colon [a,b] \to \mathbb{R}$ ist RIEMANN-integrierbar auf $[a,b]$.

Zum *Beweis* müssen wir zeigen, daß das RIEMANN-Integral

$$\int_a^x f(\xi)\,d\xi$$

für jedes $x \in [a,b]$ existiert.

Dazu fixieren wir ein beliebig vorgegebenes $x \in [a,b]$ und betrachten Folgen $(\boldsymbol{x}^{(\nu)})$ von immer feiner werdenden Unterteilungen des Intervalls $[a,x]$; dazu entsprechende Folgen $(\underline{\xi}^{(\nu)})$ von Zwischenwerten und die Grenzwerte der RIEMANN-Summen $I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})$.

1. Schritt: Für eine feste Folge $(\boldsymbol{x}^{(\nu)})$ von Unterteilungen existiert der Grenzwert und ist unabhängig von der Wahl der Zwischenwerte $\xi_i^{(\nu)}$:

Um das einzusehen, betrachten wir für jede Unterteilung $\boldsymbol{x}^{(\nu)}$ die RIEMANNschen Obersummen und Untersummen: Da f nach Voraussetzung stetig ist, nimmt es in jedem abgeschlossenen Intervall sowohl sein Maximum als auch sein Minimum an; insbesondere werden also im Intervall $[x_i^{(\nu)}, x_{i+1}^{(\nu)}]$ ein Maximum $M_i^{(\nu)}$ und ein Minimum $m_i^{(\nu)}$ angenommen. Damit gilt unabhängig von der Wahl des Zwischenwerts $\xi_i^{(\nu)}$

$$m_i^{(\nu)} \le f(\xi_i^{(\nu)}) \le M_i^{(\nu)}\,.$$

Nach Definition ist

$$I(\boldsymbol{x}^{(\nu)}, \xi^{(\nu)}) = \sum_{i=0}^{n_\nu - 1} f(\xi_i^{(\nu)})\big(x_{i+1}^{(\nu)} - x_i^{(\nu)}\big)\,;$$

ersetzen wir hierin $\xi_i^{(\nu)}$ jeweils durch $m_i^{(\nu)}$, so erhalten wir eine untere Abschätzung

$$s^{(\nu)} \underset{\text{def}}{=} \sum_{i=0}^{n_\nu - 1} m_i^{(\nu)}\big(x_{i+1}^{(\nu)} - x_i^{(\nu)}\big)$$

für $I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})$; diese bezeichnen wir als RIEMANN*sche Untersumme* der Unterteilung $\boldsymbol{x}^{(\nu)}$. Entsprechend liefert die Ersetzung durch $M_i^{(\nu)}$ eine obere Abschätzung

$$S^{(\nu)} \underset{\text{def}}{=} \sum_{i=0}^{n_\nu - 1} M_i^{(\nu)}\big(x_{i+1}^{(\nu)} - x_i^{(\nu)}\big)$$

für $I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})$, die RIEMANN*sche Obersumme* der Unterteilung $\boldsymbol{x}^{(\nu)}$.

Unabhängig von der Wahl der Zwischenwerte $\xi_i^{(\nu)}$ gilt somit

$$s^{(\nu)} \le I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)}) \le S^{(\nu)}\,,$$

ein Zusammenhang, der in Abbildung 54 noch einmal graphisch dargestellt ist: Die RIEMANNsche Untersumme ist gleich der Fläche der durchgezogenen Rechtecke, für die Obersumme kommt noch der gestrichelte Anteil dazu, und für die RIEMANNsche Summe $I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ muß man bis zu den punktierten Linien gehen. Die Werte $\xi_i^{(\nu)}$ sind auf der x-Achse durch kleine Kreise gekennzeichnet, ebenso die Punkte $\big(\xi_i^{(\nu)}, f(\xi_i^{(\nu)})\big)$ auf der Kurve.

Wenn wir jetzt noch zeigen können, daß die Folge der $s^{(\nu)}$ und die der $S^{(\nu)}$ beide gegen denselben Grenzwert konvergieren, muß wegen der Einschließung

$$s^{(\nu)} \le I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)}) \le S^{(\nu)}$$

auch $I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ gegen diesen Grenzwert konvergieren, und die Behauptung im ersten Schritt ist bewiesen.

Vergleichen wir dazu zunächst die Untersummen $s^{(\nu)}$ und $s^{(\nu+1)}$: Die Unterteilung $\boldsymbol{x}^{(\nu+1)}$ entsteht aus $\boldsymbol{x}^{(\nu)}$ dadurch, daß einige der Intervalle weiter unterteilt werden. Ist aber $(x_i^{(\nu+1)}, x_{i+1}^{(\nu+1)})$ Teilintervall von $(x_j^{(\nu)}, x_{j+1}^{(\nu)})$, so kann das Minimum im Teilintervall natürlich nicht kleiner sein als

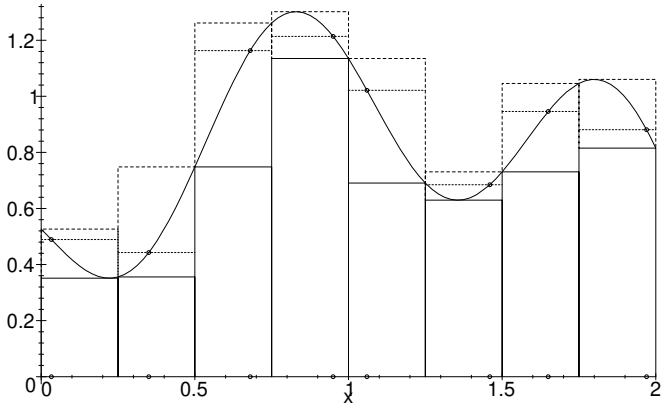


Abb. 54: Riemannsche Summen, Obersummen und Untersummen

im größeren Intervall, d.h. $s^{(\nu+1)} \geq s^{(\nu)}$. Somit ist die Folge der $s^{(\nu)}$ monoton wachsend.

Genauso folgt, daß die Folge der $S^{(\nu)}$ monoton fallend ist, und da stets $s^{(\nu)} \leq S^{(\nu)}$ ist, folgt weiter, daß

$$s^{(1)} \leq s^{(2)} \leq s^{(3)} \leq \cdots \leq S^{(3)} \leq S^{(2)} \leq S^{(1)} .$$

Insbesondere ist also $S^{(1)}$ eine obere Schranke für die Folge der $s^{(\nu)}$ und $s^{(1)}$ eine untere Schranke für die Folge der $S^{(\nu)}$.

Damit ist also die Folge $(s^{(\nu)})$ monoton wachsend und nach oben beschränkt, während $(S^{(\nu)})$ monoton fallend und nach unten beschränkt ist. Nach einem Satz aus der Analysis I (*Jede monotone und beschränkte Folge reeller Zahlen ist konvergent.*) folgt daraus die Konvergenz beider Folgen.

Nun fehlt nur noch, daß beide denselben Grenzwert haben.

Dazu betrachten wir die Differenzen

$$\begin{aligned} S^{(\nu)} - s^{(\nu)} &= \sum_{i=0}^{n_\nu - 1} M_i^{(\nu)}(x_{i+1}^{(\nu)} - x_i^{(\nu)}) - \sum_{i=0}^{n_\nu - 1} m_i^{(\nu)}(x_{i+1}^{(\nu)} - x_i^{(\nu)}) \\ &= \sum_{i=0}^{n_\nu - 1} (M_i^{(\nu)} - m_i^{(\nu)})(x_{i+1}^{(\nu)} - x_i^{(\nu)}) . \end{aligned}$$

Mit

$$\Delta^{(\nu)} \underset{\text{def}}{=} \max_i \left(M_i^{(\nu)} - m_i^{(\nu)}\right)$$

ist also

$$\begin{aligned} S^{(\nu)} - s^{(\nu)} &\leq \sum_{i=0}^{n_\nu - 1} \Delta^{(\nu)}(x_{i+1}^{(\nu)} - x_i^{(\nu)}) = \Delta^{(\nu)} \sum_{i=0}^{n_\nu - 1} (x_{i+1}^{(\nu)} - x_i^{(\nu)}) \\ &= \Delta^{(\nu)}(x_n^{(\nu)} - x_0^{(\nu)}) = \Delta^{(\nu)}(x - a) , \end{aligned}$$

und das ist eine Nullfolge, falls wir zeigen können, daß die Folge der $\Delta^{(\nu)}$ eine ist.

Hier kommt nun die in Abschnitt *c)* eingeführte gleichmäßige Stetigkeit von f zum Tragen: Danach gibt es zu jedem $\varepsilon > 0$ ein $\delta > 0$, so daß für alle Punkte $\xi_1, \xi_2 \in [a, b]$ gilt: Ist $|\xi_1 - \xi_2| < \delta$, so folgt $|f(\xi_1) - f(\xi_2)| < \varepsilon$.

Ist dabei für eine Unterteilung $\boldsymbol{x}^{(\nu)}$ jede der Differenzen $x_{i+1}^{(\nu)} - x_i^{(\nu)}$ kleiner als δ, so ist auch $M_i^{(\nu)} - m_i^{(\nu)} < \varepsilon$, denn sowohl $M_i^{(\nu)}$ als auch $m_i^{(\nu)}$ sind Funktionswerte, die irgendwo im Intervall $[x_i^{(\nu)}, x_{i+1}^{(\nu)}]$ angenommen werden.

Nun haben wir aber vorausgesetzt, daß die Unterteilungen $\boldsymbol{x}^{(\nu)}$ immer feiner werden, d.h. zu jedem $\delta > 0$ gibt es in der Tat ein ν_0, so daß $x_{i+1}^{(\nu)} - x_i^{(\nu)} < \delta$ für alle $\nu > \nu_0$ und alle i. Somit sind für $\nu > \nu_0$ alle Differenzen $M_i^{(\nu)} - m_i^{(\nu)}$ kleiner als ε, und damit ist auch $\Delta^{(\nu)} < \varepsilon$ für $\nu > \nu_0$. Also sind sowohl $(\Delta^{(\nu)})$ als auch $S^{(\nu)} - s^{(\nu)}$ Nullfolgen, d.h.

$$\lim_{\nu \to \infty} s^{(\nu)} = \lim_{\nu \to \infty} S^{(\nu)} .$$

Wie schon erwähnt, folgt daraus aufgrund der Einschließung

$$s^{(\nu)} \le I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)}) \le S^{(\nu)} ,$$

auch die Gleichung

$$\lim_{\nu\to\infty} S^{(\nu)} = \lim_{\nu\to\infty} I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)}) = \lim_{\nu\to\infty} S^{(\nu)} .$$

Da die linke und die rechte Seite obiger Ungleichung nicht von den $\xi_i^{(\nu)}$ abhängen, ist auch $\lim I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ davon unabhängig, und damit ist der erste Schritt des Beweises beendet.

Der zweite ist zum Glück erheblich einfacher:

2. Schritt: $\lim_{\nu\to\infty} I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ ist auch unabhängig von der Folge $(\boldsymbol{x}^{(\nu)})$:

Betrachten wir zwei Folgen $(\boldsymbol{x}^{(\nu)})$ und $(\boldsymbol{y}^{(\nu)})$ von Unterteilungen. Für jeden Index ν können wir zu den Unterteilungen $\boldsymbol{x}^{(\nu)}$ und $\boldsymbol{y}^{(\nu)}$ eine gemeinsame Verfeinerung $\boldsymbol{z}^{(\nu)}$ konstruieren, indem wir einfach die sämtlichen Zahlen $x_i^{(\nu)}$ und $y_j^{(\nu)}$ der Größe nach ordnen als $z_0^{(\nu)} = a, \ldots, z_r^{(\nu)} = x$.

Wie oben seien $s^{(\nu)}$ und $S^{(\nu)}$ die RIEMANNsche Unter- und Obersumme zur Unterteilung $\boldsymbol{x}^{(\nu)}$; entsprechend seien $\widetilde{s}^{(\nu)}$ und $\widetilde{S}^{(\nu)}$ die zur Unterteilung $\boldsymbol{z}^{(\nu)}$. Da $\boldsymbol{z}^{(\nu)}$ eine Verfeinerung von $\boldsymbol{x}^{(\nu)}$ ist, wissen wir aus dem ersten Schritt, daß

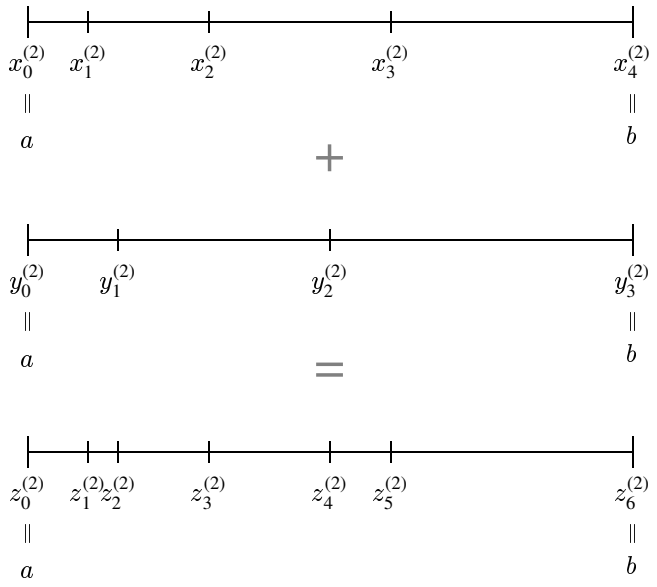
$$s^{(\nu)} \le \widetilde{s}^{(\nu)} \le \widetilde{S}^{(\nu)} \le S^{(\nu)}$$

ist; da die Folgen $(s^{(\nu)})$ und $(S^{(\nu)})$ denselben Grenzwert haben, müssen auch $(\widetilde{s}^{(\nu)})$ und $(\widetilde{S}^{(\nu)})$ gegen diesen Wert konvergieren und damit auch die Folge der $I(\boldsymbol{z}^{(\nu)}, \tilde{\underline{\xi}}^{(\nu)})$, wobei $(\tilde{\underline{\xi}}^{(\nu)})$ irgendeine Folge von Zwischenwerten bezeichnet; wie wir bereits aus dem ersten Schritt wissen, hängt der Grenzwert nicht davon ab. Also ist

$$\lim_{\nu\to\infty} I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)}) = \lim_{\nu\to\infty} I(\boldsymbol{z}^{(\nu)}, \tilde{\underline{\xi}}^{(\nu)}) .$$

Genauso folgt, daß auch

$$\lim_{\nu\to\infty} I(\boldsymbol{y}^{(\nu)}, \bar{\underline{\xi}}^{(\nu)}) = \lim_{\nu\to\infty} I(\boldsymbol{z}^{(\nu)}, \tilde{\underline{\xi}}^{(\nu)}) ,$$



wobei $(\bar{\underline{\xi}}^{(\nu)})$ die Folge der Zwischenwerte zu $\boldsymbol{y}^{(\nu)}$ bezeichnet, und dies zeigt schließlich die Behauptung

$$\lim_{\nu\to\infty} I(\boldsymbol{x}^{(\nu)}, \underline{\xi}^{(\nu)}) = \lim_{\nu\to\infty} I(\boldsymbol{y}^{(\nu)}, \bar{\underline{\xi}}^{(\nu)}) .$$

Damit ist der Satz vollständig bewiesen. ∎

7) Stückweise stetige Funktionen: Gelegentlich möchte man die Bedingung der Stetigkeit wenigstens ein bißchen lockern, um beispielweise auch einen Rechteckimpuls, der periodisch zwischen 0 und 1 (oder -1 und 1) wechselt, behandeln zu können. Die Existenz des RIEMANN-Integral wird durch solche kleineren Abweichungen nicht beeinträchtigt:

Definition: Eine Funktion $f\colon [a,b] \to \mathbb{R}$ heißt *stückweise stetig,* wenn

es eine Unterteilung

$$a = a_0 < a_1 < \cdots < a_{r-1} < a_r = b$$

des Intervalls $[a, b]$ gibt, so daß f auf jedem der offenen Intervalle (a_i, a_{i+1}) stetig ist.

f ist also überall stetig außer eventuell in endlich vielen Punkten $a_0, \ldots, a_r$. Der Wert in diesen Punkten kann, aber muß nicht mit dem linksseitigen oder rechtsseitigen Grenzwert der Funktion übereinstimmen; beispielsweise definiert man Rechteckimpulse gelegentlich auch durch die Funktion

$$f(x) = \begin{cases} 1 & \text{falls } 2n-1 < x < 2n \text{ für ein } n \in \mathbb{Z} \\ 0 & \text{falls } x \in \mathbb{Z} \\ -1 & \text{falls } 2n < x < 2n+1 \text{ für ein } n \in \mathbb{Z} \end{cases} .$$

Satz: Eine stückweise stetige Funktion $f\colon [a, b] \to \mathbb{R}$ ist RIEMANN-integrierbar.

Beweis: f sei stetig in den offenen Intervallen (a_i, a_{i+1}) mit

$$a = a_0 < a_1 < \cdots < a_{r-1} < a_r = b\,.$$

Wiederholt man den Beweis des entsprechenden Satzes für Funktionen, die auf ganz $[a, b]$ stetig sind, so funktioniert fast alles problemlos auch für f. Schwierigkeiten gibt es nur mit den Intervallen einer Unterteilung $\boldsymbol{x}$, die einen der Punkte a_i enthalten. In diesen Intervallen ist f nicht notwendigerweise stetig, so daß die Differenz zwischen dem Supremum und dem Infimum von f dort nicht mit Verkleinerung des Intervalls gegen Null gehen muß, sondern auch gegen einen von Null verschiedenen Wert, nämlich die „Sprunghöhe“ bei a_i, konvergieren kann.

Nun gibt es aber in jeder Unterteilung nur endlich viele Intervalle, die ein a_i enthalten, und auch die Sprunghöhen sind begrenzt; gehen also alle Intervallängen gegen Null, so geht auch wie im Beweis des Satzes die Differenz zwischen RIEMANNschen Ober- und Untersummen gegen Null. ∎

8) Noch einmal die Dirichletsche Sprungfunktion: In Abschnitt *b3)* waren wir nicht zufrieden mit dem Verhalten des dort heuristisch eingeführten Integrals für die DIRICHLETsche Sprungfunktion; schauen wir, was das nun eingeführte RIEMANN-Integral daraus macht. Sei also

$$f(x) = \begin{cases} 1 & \text{für } x \in \mathbb{Q} \\ 0 & \text{für } x \notin \mathbb{Q} \end{cases} ;$$

wir interessieren uns für $\int_0^1 f(x)\,dx$.

Dazu sei $\boldsymbol{x}$ eine Unterteilung des Intervalls $[0, 1]$. Unabhängig von der Wahl dieser Unterteilung können wir stets eine Folge $\underline{\xi}$ von rationalen Zwischenwerten finden, aber auch eine Folge $\underline{\widetilde{\xi}}$ von irrationalen. Dann ist, unabhängig von der Unterteilung $\boldsymbol{x}$,

$$I(\boldsymbol{x}, \underline{\xi}) = \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i) = \sum_{i=0}^{n-1} (x_{i+1} - x_i) = 1 - 0 = 1$$

und

$$I(\boldsymbol{x}, \underline{\widetilde{\xi}}) = \sum_{i=0}^{n-1} f(\widetilde{\xi}_i)(x_{i+1} - x_i) = 0\,.$$

Damit kann kein gemeinsamer Grenzwert existieren, und somit ist die DIRICHLETsche Sprungfunktion nicht RIEMANN-integrierbar – ein sehr viel überzeugenderes Ergebnis als das aus Abschnitt *b3).*

9) Ausblick: Das Lebesgue-Integral: Wir könnten allerdings auch argumentieren, daß es „nur“ abzählbar unendlich viele rationale Zahlen, aber überabzählbar viele irrationale zwischen null und eins gibt; daher sollten letztere das Geschehen dominieren, und $\int_0^1 f(x)\,dx$ sollte verschwinden.

In der Tat kann man eine Verallgemeinerung des RIEMANN-Integrals definieren, das LEBESGUE-Integral, das für stückweise stetige Funktionen mit dem RIEMANN-Integral übereinstimmt und für die DIRICHLETsche Sprungfunktion den Wert Null liefert. Dazu geht man nach dem französischen Mathematiker HENRI LÉON LEBESGUE (1875-1941) bei den Unter- und Obersummen nicht wie bei RIEMANN von *endlich* vielen

Rechtecken aus, sondern von *abzählbar unendlich* vielen. (Wie man dies im einzelnen macht, braucht uns hier nicht zu interessieren; wir werden uns im folgenden stets auf das RIEMANN-Integral beschränken.) Dieses LEBESGUE-Integral existiert *fast* immer: Man kann zwar die *Existenz* von Funktionen, für die es nicht existiert, *beweisen,* explizite Beispiele solcher Funktionen sind aber nicht bekannt.

10) Anwendung auf Flächeninhalte: Nachdem wir nun mit einen exakten Integralbegriff haben, bietet sich an, Flächen über dieses Integral zu *definieren:*

Definition: Ist $f\colon [a,b] \to \mathbb{R}$ eine nichtnegative Funktion, so bezeichnen wir die Zahl

$$\int_a^b f(x)\,dx$$

als *Fläche* zwischen der Kurve $y = f(x)$ und der x-Achse zwischen den Geraden $x = a$ und $x = b$.

Was passiert, wenn f auch negative Werte annimmt? Für $f(x) = x$ etwa ist $\int_{-1}^{1} f(x)\,dx = 0$, wie man sich leicht überlegt anhand von Unterteilungen, die symmetrisch zur Null liegen. Auch dies läßt sich als Aussage über Flächeninhalte interpretieren – wenn man davon ausgeht, daß die Formel

$$\textit{Fläche} = \textit{Länge} \times \textit{Breite}$$

für die Rechteckfläche auch bei *negativer* Länge und/oder Breite gelten soll. Da Integrale nicht nur zur Flächenbestimmung, sondern auch etwa zur Bestimmung der Ladung in einem Kondensator verwendet werden, ist dies die Interpretation, die man in der Mathematik in den meisten Fällen vorzieht; für die klassische Fläche im Sinne des Tapezierens muß man mit

$$\int_a^b |f(x)|\ dx$$

arbeiten.

Im Sinne dieser negativen Längen und Breiten ist es auch sinnvoll, Integrale für $b < a$ zu definieren durch

$$\int_a^b f(x)\,dx \underset{\text{def}}{=} -\int_b^a f(x)\,dx\,.$$

Insbesondere ist dann natürlich $\int_a^a f(x)\,dx = 0$.

d) Erste Integrationsregeln

Das RIEMANN-Integral stellt die heuristischen Überlegungen vom Beginn dieses Paragraphen insofern auf eine exaktere mathematische Grundlage, als wir nun einen Integralbegriff haben, der auch bei extrem ungewöhnlichen Funktionen wie der DIRICHLETschen Sprungfunktion keine unerwarteten Ergebnisse liefert. Was allerdings den Ausgangspunkt betrifft, die Umkehrung der Differentiation, wissen wir über das neue Integral noch gar nichts, und auch sonst kennen wir noch nicht viele Regeln über den Umgang damit und insbesondere auch über seine Berechnung *ohne* die umständlichen und sehr langsam konvergierenden RIEMANN-Summen.

In diesem Abschnitt sollen die ersten (und einfachsten) solchen Regeln zusammengestellt werden; danach erweitern wir zunächst unser Instrumentarium, um dann damit auch kompliziertere und interessantere Regeln zu beweisen.

1) Monotonieregel: Eine der einfachsten, aber trotzdem oft nützlichen Regeln für den Umgang mit Integralen übersetzt Größerbeziehungen zwischen Integranden in Größenbeziehungen zwischen Integralen:

Satz: f und g seien stückweise stetige Funktionen auf dem Intervall $[a,b]$, und für alle $x \in [a,b]$ sei $f(x) \le g(x)$. Dann ist auch

$$\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx\,.$$

Insbesondere ist

$$\left| \int_a^b f(x)\,dx \right| \leq \int_a^b |f(x)|\ dx$$

Beweis: Wir betrachten eine Unterteilung $\boldsymbol{x}$ des Intervalls $[a,b]$ und eine dazu kompatible Sequenz $\underline{\xi}$ von Zwischenwerten; $I(\boldsymbol{x},\underline{\xi})$ sei die zugehörige RIEMANN-Summe für f und $J(\boldsymbol{x},\underline{\xi})$ die für g. Da für jedes ξ_i nach Voraussetzung $f(\xi_i) \leq g(\xi_i)$ ist, folgt unmittelbar aus der Definition der RIEMANN-Summen, daß

$$I(\boldsymbol{x},\underline{\xi}) \leq J(\boldsymbol{x},\underline{\xi})$$

ist. Da sich eine solche Kleinergleichbeziehung auch auf Grenzwerte überträgt, ist daher

$$\int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx\,,$$

wie behauptet.

Insbesondere können wir dies auch anwenden auf die Ungleichungskette

$$-\,|f(x)| \leq f(x) \leq |f(x)|$$

und erhalten

$$-\int_a^b |f(x)|\ dx \leq \left| \int_a^b f(x)\,dx \right| \leq \int_a^b |f(x)|\ dx\,,$$

d.h.

$$\left| \int_a^b f(x)\,dx \right| \leq \int_a^b |f(x)|\ dx\,.$$

■

2) Linearität und Zusammensetzung: Genauso einfach ist die Linearitätsregel:

Satz: Für zwei stückweise stetige Funktionen $f,g\colon [a,b] \to \mathbb{R}$ und zwei reelle Zahlen α,β ist

$$\int_a^b \big(\alpha f(x) + \beta g(x)\big)\ dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx\,.$$

Beweis: Wie beim letzten Satz: Für jede RIEMANN-Summe zu einer festen Unterteilung $\boldsymbol{x}$ und jede damit kompatible Zwischenwertsequenz $\underline{\xi}$ gilt die zur Behauptung des Satzes analoge Gleichung, und damit gilt im Limes auch der Satz selbst. ■

Für die Zusammensetzung von Integrationsintervallen gilt erwartungsgemäß

Satz: Für $a \leq b \leq c$ und eine stückweise stetige Funktion $f\colon [a,c] \to \mathbb{R}$ ist

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx\,.$$

Beweis: $\int_a^c f(x)\,dx$ kann als Grenzwert von RIEMANN-Summen zu einer beliebigen Folge sich verfeinernder Unterteilungen von $[a,c]$ berechnet werden; insbesondere können also Unterteilungen gewählt werden, die den Zwischenpunkt b enthalten, und dafür folgt die Behauptung unmittelbar. ■

3) Der Mittelwertsatz der Integralrechnung: Der Mittelwertsatz der *Differentialrechnung* ist der wohl wichtigste Satz aus der Analysis I. Mit dieser überragenden Bedeutung kann der Mittelwertsatz der *Integralrechnung* nicht konkurrieren; trotzdem ist auch er eine sehr nützliche Aussage, die insbesondere auch die zu Beginn dieses Paragraphen postulierte Interpretation von Integralen als Mittelwerten für das RIEMANN-Integral auf eine solide Grundlage stellt.

Im Hinblick auf spätere Anwendungen sei der Satz zunächst etwas allgemeiner formuliert, als wir ihn im Augenblick bentigen; an der Schwierigkeit des Beweises ändert diese Verallgemeinerung nichts.

Satz: $f\colon [a,b] \to \mathbb{R}$ sei eine stetige Funktion; $g\colon [a,b] \to \mathbb{R}$ sei stückweise stetig und nirgends negativ. Dann gibt es ein $\xi \in [a,b]$, so daß

$$\int_a^b f(x)g(x)\,dx = f(\xi)\int_a^b g(x)\,dx\,.$$

Die Funktion g sollte man sich dabei als eine „Gewichtsfunktion“ vorstellen, die die verschiedenen x-Werte verschieden stark gewichtet. Am anschaulichsten ist der Spezialfall $g \equiv 1$, der deshalb noch einmal getrennt angegeben sei:

Korollar: Für eine stetige Funktion $f\colon [a,b] \to \mathbb{R}$ gibt es ein $\xi \in [a,b]$ mit

$$f(\xi) = \frac{1}{b-a}\int_a^b f(x)\,dx\,.$$

Mit anderen Worten: Der „Mittelwert“ der Funktion f im Intervall $[a,b]$ existiert nicht nur, sondern wird auch an (mindestens) einer Stelle im Intervall angenommen.

Zum *Beweis* des Satzes betrachten wir das Minimum m und das Maximum M von f auf $[a,b]$. Da g nirgends negativ wird, gilt dann auch für jedes $x \in [a,b]$

$$m\,g(x) \le f(x)g(x) \le M\,g(x)$$

und damit nach der Monotonieregel

$$m\int_a^b g(x)\,dx \le \int_a^b f(x)g(x)\,dx \le M\int_a^b g(x)\,dx\,.$$

Es gibt daher eine reelle Zahl μ zwischen m und M, den Mittelwert, so daß

$$\int_a^b f(x)g(x)\,dx = \mu\int_a^b g(x)\,dx$$

ist, und da f stetig ist, gibt es nach dem Zwischenwertsatz mindestens ein $\xi \in [a,b]$, für das $f(\xi) = \mu$ ist.

Damit ist der Satz bewiesen, und das Korollar ist natürlich einfach der Spezialfall $g \equiv 1$, für den das Integral über g gleich $b-a$ ist. ∎

Man beachte, daß es hier nicht genügt, daß f nur eine *stückweise* stetige Funktion ist: Für

$$f(x) = \begin{cases} 0 & \text{für } x < 0 \\ 1 & \text{für } x \ge 0 \end{cases}$$

ist

$$\frac{1}{1-(-1)}\int_{-1}^{1} f(x)\,dx = \frac{1}{2}\int_0^1 1\,dx = \frac{1}{2}$$

nicht in der Form $f(\xi)$ darstellbar.

e) Der Hauptsatz der Differential- und Integralrechnung

Eine Motivation zur Einführung eines Integrals war die Suche nach einer Umkehroperation zur Differentiation: Genau wie die Differentiation aus einem zeitabhängigen Weg eine Geschwindigkeit macht, wollten wir ausgehend von einer zeitabhängigen Geschwindigkeit den Weg zurückbekommen. Der Hauptsatz der Differential- und Integralrechnung sagt uns nun endlich, daß wir dieses Ziel erreicht haben:

Hauptsatz der Differential- und Integralrechung: $f\colon [a,b] \to \mathbb{R}$ sei eine stetige Funktion und

$$F_a(x) = \int_a^b f(\xi)\,d\xi\,.$$

Dann ist $F_a'(x) = f(x)$ für alle $x \in (a,b)$.

Ist umgekehrt F eine differenzierbare Funktion, deren Ableitung auf dem offenen Intervall (a,b) mit f übereinstimmt, so gibt es eine reelle Zahl C derart, daß $F(x) = F_a(x) + C$ ist.

Beweis: Die Ableitung von F_a ist

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

für positives h ist

$$F_a(x+h) - F_a(x) = \int_a^{x+h} f(x)\,dx - \int_a^x f(x)\,dx = \int_x^{x+h} f(x)\,dx$$

(s. Abschnitt *b)*); für negatives h ist entsprechend

$$F_a(x+h) - F_a(x) = -\int_{x+h}^{x} f(x)\,dx .$$

Nach dem Mittelwertsatz der Integralrechnung aus dem vorigen Abschnitt gibt es in jedem der beiden Fälle ein ξ zwischen x und $x+h$, so daß das Integral gleich $|h| \cdot f(\xi)$ ist, d.h.

$$\frac{F_a(x+h) - F_a(x)}{h} = f(\xi) .$$

Geht nun h gegen Null, so muß die zwischen x und $x+h$ liegende Zahl ξ gegen x gehen; wegen der Stetigkeit von f ist also $F_a'(x) = f(x)$

Ist F eine weitere Funktion mit Ableitung f, so hat die Differenz h von F und F_a die Ableitung Null. Nun erinnern wir uns an die Analysis I: Wenn die Ableitung einer Funktion $h(x)$ verschwindet, muß die Funktion konstant sein. In der Tat: Gäbe es im Definitionsbereich von h zwei Werte $x_1 \neq x_2$ mit $h(x_1) \neq h(x_2)$, so gäbe es nach dem Mittelwertsatz der Differentialrechnung ein $\xi \in (x_1, x_2)$, so daß

$$\frac{h(x_2) - h(x_1)}{x_2 - x_1} = h'(\xi)$$

wäre. Hier steht links eine von Null verschiedene Zahl, rechts aber Null, ein Widerspruch.

Also gibt es eine Konstante C, so daß $F(x) - F_a(x) = C$ oder

$$F(x) = F_a(x) + C$$

ist, wie behauptet. ∎

Als Anwendung für die Berechnung bestimmter Integrale ergibt sich das folgende

Korollar: Ist F irgendeine differenzierbare Funktion mit der Eigenschaft, daß $F'(x) = f(x)$ ist im Intervall $[a,b]$, so gilt

$$\int_a^b f(x)\,dx = F(b) - F(a) .$$

Beweis: Mit obigen Bezeichnungen ist das Integral gleich $F_a(b)$. Zur Funktion F gibt es nach dem Hauptsatz eine Konstante C, so daß $F(x) = F_a(x) + C$ ist. Damit ist $F(b) - F(a) = F_a(b)$, wie gewünscht. ∎

Für „einfache“ Integranden ist dies im allgemeinen die beste Möglichkeit zur Berechnung eines Integrals; die Funktionen F haben daher einen Namen verdient:

Definition: Eine differenzierbare Funktion F heißt *Stammfunktion* von f, wenn $F'(x) = f(x)$ ist; wir schreiben

$$F(x) = \int f(\xi)\,d\xi + C ,$$

wobei die *Integrationskonstante* C die Nichteindeutigkeit der Stammfunktion ausdrückt.

Diese Nichteindeutigkeit sorgt auch dafür, daß zwar nach obigem Hauptsatz die Differentiation die Integration rückgängig macht, die Integration umgekehrt aber die Differentiation nur bis auf auf eine Konstante:

Korollar: $\int f'(x)\,dx = f(x) + C$

Beweis: Klar, denn f ist eine Stammfunktion von f'. ∎

Falls wir eine Stammfunktion von f kennen, ist die Berechnung des Integrals

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

also ganz einfach. Eine erste Auswahl von Stammfunktionen erhalten wir durch Rückwärtslesen von Differentiationsregeln, zum Beispiel

$$\frac{d}{dx}x^n = nx^{n-1} \Rightarrow \int x^n\,dx = \frac{x^{n+1}}{n+1} \quad \text{falls } n \neq -1$$
$$\frac{d}{dx}e^{ax} = ae^{ax} \Rightarrow \int e^{ax}\,dx = \frac{e^{ax}}{a} \quad \text{falls } a \neq 0\,,$$

und auch Formeln wie

$$\int \sin\omega x\,dx = -\frac{\cos\omega x}{\omega} + C \quad \text{und} \quad \int \cos\omega x\,dx = \frac{\sin\omega x}{\omega} + C$$

sind nun problemlos. Für weitere Stammfunktionen müssen wir allerdings zunächst noch einige Funktionen kennenlernen.

f) Trigonometrische Funktionen, Hyperbelfunktionen und ihre Umkehrfunktionen

Die Ableitung einer rationalen Funktion ist wieder eine rationale Funktion; für Stammfunktionen gilt allerdings nicht dergleichen, denn wie wir gerade gesehen haben, ist schon die Stammfunktion von x^{-1} nicht mehr rational, sondern der natürliche Logarithmen. Wie wir bald sehen werden, reichen (im Reellen) auch Logarithmen noch nicht aus, um alle rationalen Funktionen integrieren zu können; wir brauchen zusätzlich noch die sogenannten Arkus- und Areafunktionen.

Die Arkusfunktionen sind die Umkehrungen der trigonometrischen Funktionen: Letztere ordnen einem Winkel eine reelle Zahl zu; die Umkehrfunktionen liefern also als Ergebnis einen Winkel oder auch Bogen, auf lateinisch *arcus* genannt.

Unter den trigonometrischen Funktionen sind zumindest der Sinus und der Cosinus aus der Analysis I bekannt; die meisten kennen wohl auch

bereits deren Quotienten

$$\tan x = \frac{\sin x}{\cos x}\,,$$

den Tangens. Im rechtwinkligen Dreieck gibt er das Verhältnis zwischen Gegenkathete und Ankathete eines Winkels an und damit das, was man üblicherweise als *Steigung* bezeichnet: das Verhältnis der Höhendifferenz zur horizontalen Entfernung.

Da $\cos x$ für $x = \pi/2 + k\pi$ mit $k \in \mathbb{Z}$ verschwindet, ist $\tan x$ an diesen Stellen nicht definiert; in ihrer Umgebung wird er links, wo Sinus und Cosinus dasselbe Vorzeichen haben (+ für gerades k, − für ungerades) beliebig groß, rechts beliebig negativ; siehe dazu auch Abbildung 55.

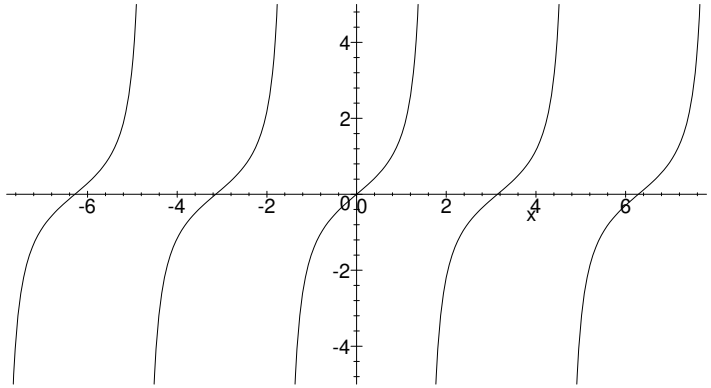


Abb. 55: Der Tangens

Aus den speziellen Werten von Sinus und Cosinus berechnet man leicht die entsprechenden Werte für Tangens:

x	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
$\tan x$	0	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	∞

Da alle trigonometrischen Funktionen periodisch sind, können sie keine global definierten Umkehrfunktionen haben; wir müssen sie also jeweils

auf ein Intervall einschränken, auf dem sie injektiv oder – besser noch – monoton sind.

Im Fall des Sinus bietet sich das Intervall von $-\pi/2$ bis $\pi/2$ an, in dem er monoton von -1 auf 1 ansteigt; wir definieren daher *den Arkussinus* oder – genauer ausgedrückt – den *Hauptwert* des Arkussinus für dieses Intervall:

$$\arcsin\colon [-1,1] \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

ist die Umkehrfunktion von

$$\sin\colon \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to [-1,1]\,.$$

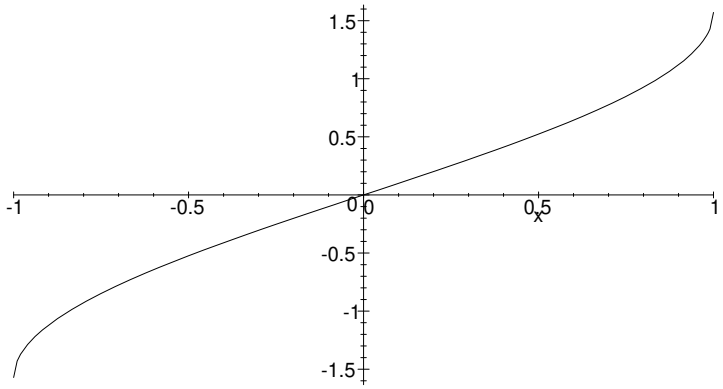


Abb. 56: Der Arkussinus

Der Cosinus fällt von seinem Maximalwert 1 an der Stelle Null monoton ab, bis er bei π den Wert -1 erreicht hat. Wir definieren den *Hauptwert* der Umkehrfunktion *Arkuscosinus* oder kurz *den* Arkuscosinus daher für dieses Intervall:

$$\arccos\colon [-1,1] \to [0,\pi]$$

ist die Umkehrfunktion von

$$\cos\colon [0,\pi] \to [-1,1]\,.$$

Aufgrund der Beziehung $\cos\varphi = \sin\left(\frac{\pi}{2} - \varphi\right)$ und der Wahl der Wertebereiche ist

$$\arccos x = \frac{\pi}{2} - \arcsin x\,.$$

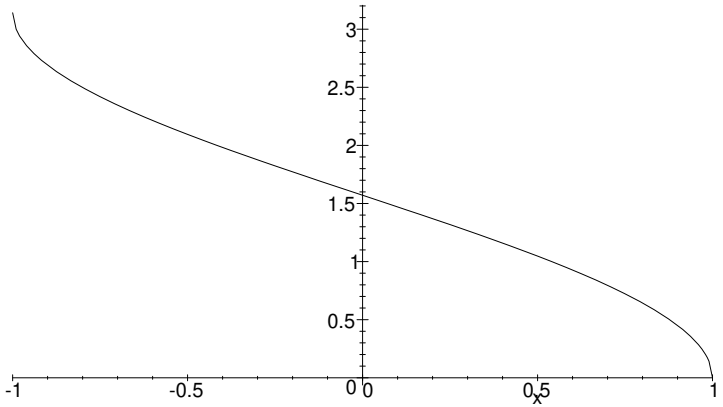


Abb. 57: Der Arkuscosinus

Der Tangens schließlich ist monoton ansteigend im offenen Intervall $(-\pi/2, \pi/2)$ und nimmt dort jeden reellen Wert an; wir definieren den *Hauptwert* der Umkehrfunktion *Arkustangens* oder kurz *den* Arkustangens daher für dieses Intervall:

$$\arctan\colon \mathbb{R} \to \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

ist die Umkehrfunktion von

$$\tan\colon \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}\,.$$

Gelegentlich ganz nützlich ist der Wert an der Stelle 1: Da nach obiger Tabelle $\tan\frac{\pi}{4}$ den Wert 1 hat, ist $\arctan(1) = \frac{\pi}{4}$ oder

$$\pi = 4\,\arctan(1)\,.$$

Falls man in einem Programm den Wert π benötigt und ihn über diese Formel als Programmkonstante definiert, kann man – eine saubere Implementierung der Arkusfunktionen vorausgesetzt – sicher sein, daß π

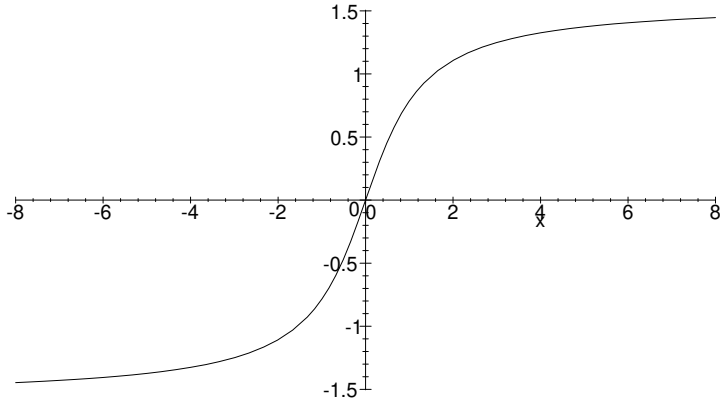


Abb. 58: Der Arkustangens

mit der maximal möglichen Stellenzahl der jeweiligen Gleitkommaarithmetik dargestellt wird – ohne daß man diese Stellenzahl oder gar π mit der entsprechenden Genauigkeit zu kennen braucht.

Abgesehen vom gerade definierten Hauptwert lassen sich natürlich noch in offensichtlicher Weise weitere Zweige definieren; wir wollen hier auf Einzelheiten verzichten. Der Leser sollte sich allerdings klarmachen, daß es beispielsweise für den Arkussinus sowohl monoton wachsende als auch monoton fallende Zweige gibt.

Mit Hilfe der Arkusfunktionen lassen sich die Polarkoordinatendarstellung und die kartesische Darstellung der komplexen Zahlen ineinander umrechnen: Da $z = re^{i\varphi}$ Realteil $x = r\cos\varphi$ und Imaginärteil $y = r\sin\varphi$ hat, sollte außer der bereits bekannten Formel

$$r = \sqrt{x^2 + y^2}$$

für $z \neq 0$ auch so etwas wie

$$\varphi = \arctan\frac{y}{x} = \arccos\frac{x}{r} = \arcsin\frac{y}{r}$$

gelten. Dies im Prinzip auch richtig, allerdings nicht immer für den Hauptwert: Da alle drei Umkehrfunktionen nur Werte in einem Intervall der Länge π annehmen, während das Argument (genauer: der

Hauptwert des Arguments) einer komplexen Zahl im doppelt so langen Intervall $(-\pi, \pi]$ liegt, kann obige Gleichung bei keiner der drei Funktionen für alle z gelten. Eine korrekte Formel mit dem Hauptwert des Arkuscosinus ist beispielsweise

$$\arg z = \begin{cases} \arccos\frac{x}{r} & \text{falls } \mathfrak{Im}\, z \geq 0 \\ -\arccos\frac{x}{r} & \text{falls } \mathfrak{Im}\, z < 0 \end{cases}.$$

Der für uns interessanteste Aspekt der Arkusfunktionen sind ihre *Ableitungen,* die uns neue, bislang unbekannte Stammfunktionen liefern sollen.

Wie in der Analysis I bei der Ableitung des Logarithmus gehen wir aus von der Formel

$$g'(x) = \frac{1}{f'\big(g(x)\big)}$$

für die Umkehrfunktion g einer Funktion f. Da die Ableitung des Sinus der Cosinus ist, erhalten wir also

$$\arcsin'(x) = \frac{1}{\cos(\arcsin x)}\,.$$

Wegen $\sin^2 y + \cos^2 y = 1$ ist

$$\cos^2(\arcsin x) = 1 - \sin^2(\arcsin x) = 1 - x^2\,;$$

da der Cosinus im Intervall $[-\pi/2, \pi/2]$, in dem der Arkussinus seine Werte annimmt, größer oder gleich Null ist, folgt

$$\cos(\arcsin x) = \sqrt{1 - x^2} \quad \text{und} \quad \arcsin'(x) = \frac{1}{\sqrt{1 - x^2}}\,.$$

Genauso ist

$$\arccos'(x) = \frac{-1}{\sin(\arccos x)} = \frac{-1}{\sqrt{1 - x^2}}\,,$$

da der Sinus zwischen 0 und π keine negativen Werte annimmt.

Für den Arkustangens schließlich müssen wir zunächst Tangens selbst ableiten; nach der Quotientenregel ist

$$\tan'(x) = \frac{d}{dx}\frac{\sin x}{\cos x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = 1 + \tan^2 x$$

und dementsprechend

$$\arctan'(x) = \frac{1}{1+\tan^2(\arctan x)} = \frac{1}{1+x^2}\,.$$

Da sich die Ableitungen von Arkussinus und Arkuscosinus nur im Vorzeichen unterscheiden (was natürlich von vornherein klar war, da sich die Werte der beiden Funktionen stets zu $\pi/2$ ergänzen) sind als Stammfunktionen vor allem der Arkussinus und der Arkustangens interessant; wir haben die beiden neuen Formeln

$$\int \frac{dx}{1+x^2}\,dx = \arctan x + C \quad \text{und} \quad \int \frac{dx}{\sqrt{1-x^2}}\,dx = \arcsin x + C\,.$$

Als nächstes wollen wir uns überlegen, wie wir die Integrale

$$\int \frac{dx}{1-x^2}\,dx \quad \text{und} \quad \int \frac{dx}{\sqrt{1+x^2}}\,dx$$

mit dem jeweils anderen Vorzeichen bekommen können. Eine offensichtliche Lösung besteht darin, einfach x durch ix zu ersetzen, und das legt es nahe, in Analogie zu den EULERschen Formeln

$$\cos x = \frac{e^{ix}+e^{-ix}}{2} \quad \text{und} \quad \sin x = \frac{e^{ix}-e^{-ix}}{2i}$$

neue Funktionen

$$\cosh x = \frac{e^{x}+e^{-x}}{2} \quad \text{und} \quad \sinh x = \frac{e^{x}-e^{-x}}{2}$$

zu definieren, den *Cosinus hyperbolicus* und den *Sinus hyperbolicus.*

Der Zusatz *hyperbolicus* läßt sich leicht verstehen: Genau wie der gewöhnliche Sinus und Cosinus als *Kreisfunktionen* bezeichnet werden, weil die Punkte $(\sin t, \cos t)$ auf dem Kreis $x^2 + y^2 = 1$ liegen, haben die Hyperbelfunktionen ihren Namen daher, daß

$$\begin{aligned}\cosh^2 t - \sinh^2 t &= \left(\frac{e^t+e^{-t}}{2}\right)^2 - \left(\frac{e^t+e^{-t}}{2}\right)^2 \\ &= \frac{e^{2t}+2+e^{-2t}}{4} - \frac{e^{2t}-2+e^{-2t}}{4} = 1\end{aligned}$$

ist, so daß die Punkte $(\cosh t, \sinh t)$ auf der Hyperbel $x^2 - y^2 = 1$ liegen, d.h

$$\cosh^2 x - \sinh^2 x = 1 \quad \text{für alle } x \in \mathbb{R}\,.$$

Damit ist insbesondere $\cosh^2 x$ stets ≥ 1; da der Cosinus hyperbolicus genau wie die beiden Exponentialfunktionen, aus denen er zusammengesetz ist, keine negativen Werte annehmen kann, folgt also

$$\cosh x \geq 1 \quad \text{für alle } x \in \mathbb{R}\,.$$

Sofort aus der Definition folgen die Ableitungsregeln

$$\cosh'(x) = \sinh x \quad \text{und} \quad \sinh' x = \cosh x\,;$$

auf Grund obiger Ungleichung ist daher der Sinus hyperbolicus monoton steigend auf ganz $\mathbb{R}$. Man sieht leicht, daß sein Vorzeichen jeweils das von x ist (Insbesondere ist also auch $\sinh 0 = 0$.), d.h. der Cosinus hyperbolicus ist monoton fallend für negative und monoton steigend für positive x.

Für große Werte von x wird e^{-x} beliebig klein, sowohl $\sinh x$ als auch $\cosh x$ unterscheiden sich für solche Werte also beliebig wenig von $\frac{1}{2}e^x$ und gehen insbesondere gegen unendlich. Für stark negative Werte wird e^x beliebig klein und das Verhalten beider Funktionen wird dominiert durch den Term $\pm e^{-x}$. Daher geht $\sinh x$ für $x \to -\infty$ gegen $-\infty$ und $\cosh x$ gegen $+\infty$. Offensichtlich ist $\sinh x$ eine ungerade, $\cosh x$ aber eine gerade Funktion. In Abbildung 59 sind gestrichelt auch noch die Funktionen $\frac{1}{2}e^x$ im positiven und $\pm\frac{1}{2}e^{-x}$ im negativen Bereich eingezeichnet; wie man sieht, ist die Übereinstimmung mit $\sinh x$ *bzw.* $\cosh x$ schon ab etwa $|x| \geq 2$ recht gut.

Der Graph des Cosinus hyperbolicus erinnert an eine durchhängende Kette, und tatsächlich kann die Variationsrechnung zeigen, daß eine Kette mit reibungsfrei gegeneinander beweglichen Gliedern genau diese Form hat.

In völliger Analogie zum Tangens definieren wir auch noch einen *Tangens hyperbolicus* durch die Vorschrift

$$\tanh x \underset{\text{def}}{=} \frac{\sinh x}{\cosh x} = \frac{e^x+e^{-x}}{e^x-e^{-x}}\,;$$

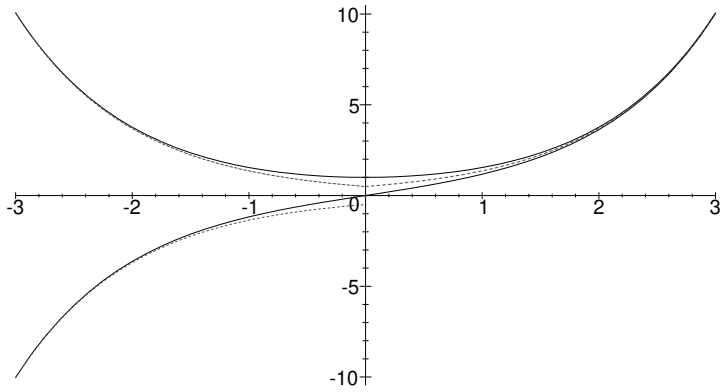


Abb. 59: Sinus hyperbolicus und Cosinus hyperbolicus

da der Cosinus hyperbolicus stets größer oder gleich eins ist, kann der Nenner nie Null werden, so daß die Funktion auf ganz $\mathbb{R}$ definiert ist. Dabei ist

$$\lim_{x\to\infty} \tanh x = \lim_{x\to\infty} \frac{e^x + e^{-x}}{e^x - e^{-x}} = \lim_{x\to\infty} \frac{e^x(1+e^{-2x})}{e^x(1-e^{-x})}$$
$$= \lim_{x\to\infty} \frac{1+e^{-2x}}{1-e^{-x}} = 1$$

und entsprechend

$$\lim_{x\to-\infty} \tanh x = \lim_{x\to-\infty} \frac{e^x + e^{-x}}{e^x - e^{-x}} = \lim_{x\to-\infty} \frac{e^{2x}+1}{e^{2x}-1} = -1\,.$$

Die Ableitung des Tangens hyperbolicus ist nach der Quotientenregel

$$\tanh' x = \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \begin{cases} \frac{1}{\cosh^2 x} & \text{nach obiger Formel} \\ 1 - \tanh^2 x & \text{durch Ausdividieren} \end{cases},$$

wobei wie beim Tangens je nach Anwendung mal die eine, mal die andere Form des Ergebnisses nützlicher ist.

Aus beiden Ausdrücken sieht man sofort, daß die Ableitung stets positiv ist, der Tangens hyperbolicus steigt also monoton von -1 nach $+1$, wobei beide Werte nur asymptotisch angenommen werden.

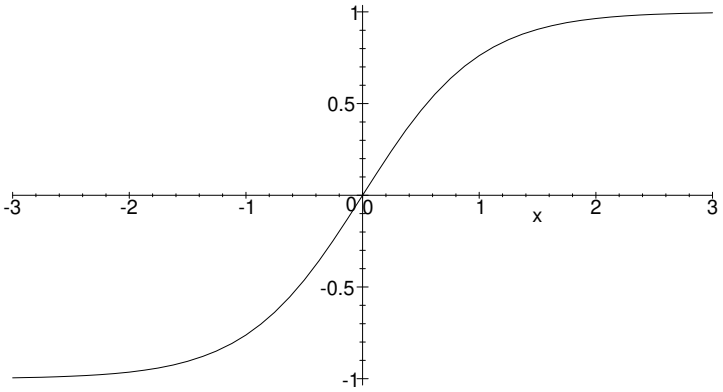


Abb. 60: Der Tangens hyperbolicus

Die Umkehrfunktionen der Hyperbelfunktionen werden als *Areafunktionen* bezeichnet und heißen *Areasinus hyperbolicus, Areacosinus hyperbolicus* und *Areatangens hyperbolicus.*

Aus obiger Diskussion folgt, daß der Areasinus hyperbolicus

$$\operatorname{arsinh}\colon \mathbb{R} \to \mathbb{R}$$

auf ganz $\mathbb{R}$ definiert ist und der Areatangens hyperbolicus

$$\operatorname{artanh}\colon (-1,1) \to \mathbb{R}$$

nur auf dem offenen Einheitsintervall. Der Cosinus hyperbolicus hat keine eindeutig bestimmte Umkehrfunktion, da er für positive und für negative Argumente jeweils denselben Wert annimmt. Wir definieren den Areacosinus hyperbolicus

$$\operatorname{arcosh}\colon \mathbb{R}_{\geq 1} \to \mathbb{R}_{\geq 0}$$

als die Umkehrfunktion des positiven Zweigs.

Da Sinus hyperbolicus und Cosinus hyperbolicus asymptotisch wie eine Exponentialfunktion ansteigend, steigen Areasinus hyperbolicus und Areacosinus hyperbolicus wie Logarithmen, also sehr langsam.

Auch diese Funktionen sind wieder vor allem wegen ihrer Ableitungen interessant; eine weitgehend zum Fall der Arkusfunktionen analoge

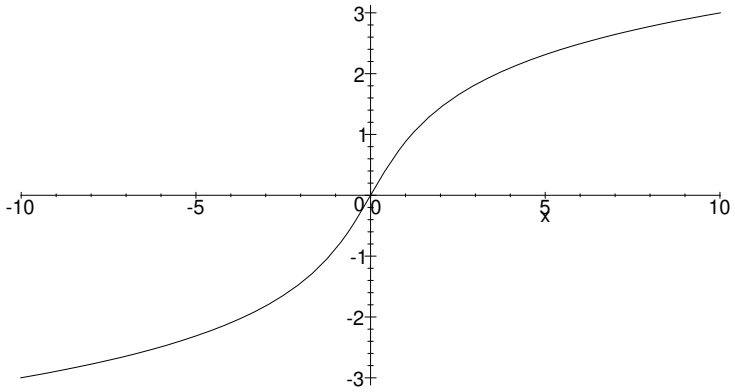


Abb. 61: Der Areasinus hyperbolicus

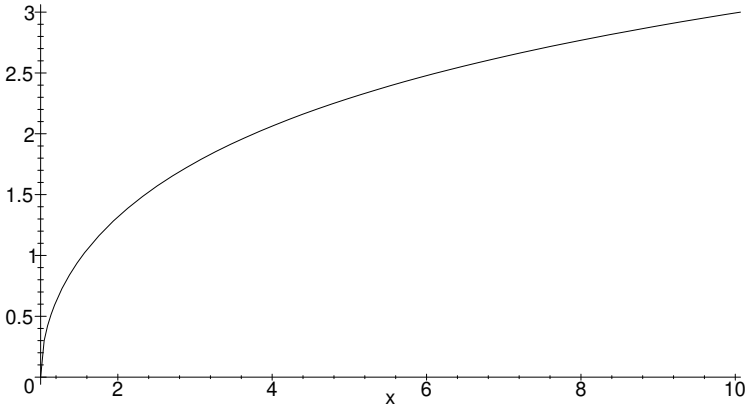


Abb. 62: Der Areacosinus hyperbolicus

Rechnung zeigt, daß

$$\operatorname{arsinh}'(x) = \frac{1}{\cosh(\operatorname{arsinh} x)} = \frac{1}{\sqrt{1+x^2}}$$

und

$$\operatorname{arcosh}'(x) = \frac{1}{\sinh(\operatorname{arcosh} x)} = \frac{1}{\sqrt{x^2-1}}$$

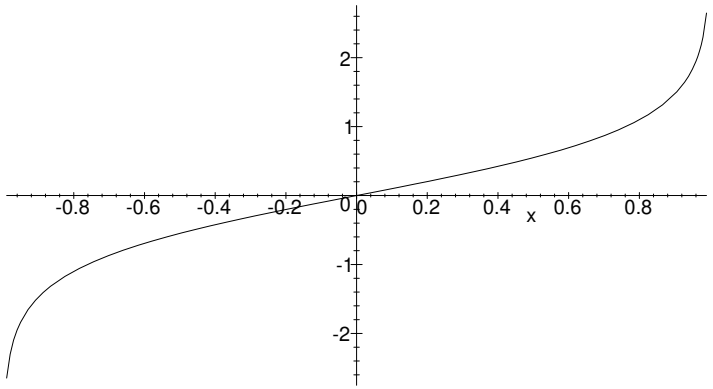


Abb. 63: Der Areatangens hyperbolicus

ist. Für den Areatangens hyperbolicus schließlich ist

$$\operatorname{artanh}'(x) = \frac{1}{1-\tanh^2(\operatorname{artanh} x)} = \frac{1}{1-x^2}\,.$$

Dies führt also auf die neuen Stammfunktionen

$$\int \frac{dx}{\sqrt{1+x^2}} = \operatorname{arsinh}(x) + C, \qquad \int \frac{dx}{\sqrt{x^2-1}} = \operatorname{arcosh}(x) + C$$

und

$$\int \frac{dx}{1-x^2} = \operatorname{artanh}(x) + C\,.$$

Die letztere wird allerdings nicht unbedingt gebraucht, denn da

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

ist, können wir dieses Integral auch ausrechnen als

$$\int \frac{dx}{1-x^2} = \frac{\ln|1-x| + \ln|1+x|}{2} + C\,,$$

wobei diese Formel noch den Vorteil eines größeren Definitionsbereichs hat.

g) Partielle Integration

Eine wichtige Regel der Differentialrechnung ist die *Produktregel*

$$\frac{d}{dx}\big(f(x)g(x)\big) = f'(x)g(x) + f(x)g'(x)\,.$$

Für die Zwecke der Integralrechnung schreiben wir sie besser als

$$f'(x)g(x) = \frac{d}{dx}\big(f(x)g(x)\big) - f(x)g'(x)\,,$$

was integriert auf

$$\int f'(x)g(x)\,dx = f(x)g(x) - \int f(x)g'(x)\,dx$$

führt – die Regel der partiellen Integration. Sie ist dann nützlich, wenn sich der Integrand als Produkt schreiben läßt, wobei einer der Faktoren eine bekannte Stammfunktion hat; gelegentlich ist dann das Produkt $f(x)g'(x)$ leichter integrierbar als $f'(x)g(x)$.

Mit den wenigen Funktionen, die wir bislang kennen, lassen sich nur wenige interessante Beispiele konstruieren; die volle Kraft dieser Regel werden wir erst später kennenlernen, wenn wir unser Repertoire an Funktionen erweitert haben. Hier sei nur $\int xe^x\,dx$ betrachtet: Da die Exponentialfunktion ihre eigene Ableitung und Stammfunktion ist, empfiehlt sie sich für die Rolle der Funktion f, während $g(x) = x$ mit $g'(x) = 1$ auf Vereinfachungen auf der rechten Seite hoffen läßt. In der Tat führt partielle Integration mit

$$\int xe^x\,dx = xe^x - \int e^x\,dx = xe^x - e^x + C = (x-1)e^x + C$$

auf eine Stammfunktion. Genauso lassen sich rekursiv auch die Integrale $\int x^n e^x\,dx$ berechnen:

$$\int x^2e^x\,dx = x^2e^x - \int 2xe^x\,dx = (x^2 - 2x + 2)e^x + C$$

und

$$\int x^n e^x\,dx = x^n e^x - n\int x^{n-1}e^x\,dx\,.$$

h) Substitutionsregel

Auch die Kettenregel

$$\frac{d}{dx}f\big(g(x)\big) = f'\big(g(x)\big)\cdot g'(x)$$

läßt sich umschreiben zu einer Integrationsregel, der *Substitutionsregel*

$$\int f\big(g(x)\big)\cdot g'(x)\,dx = F\big(g(x)\big) + C\,,$$

wobei F eine Stammfunktion von f ist.

Diese Regel dürfte wohl meist der erfolgversprechendste Versuch zum Auffinden einer Stammfunktion sein – vor allem, wenn man sie von rechts nach links liest, um für eine bekannte Funktion f die unbekannte Stammfunktion F zu berechnen. Als Substitution g wählt man hier eine „geeignete“ bijektive Funktion, die zu einer Vereinfachung auf der linken Seite führt.

1) Der Spezialfall logarithmischer Ableitungen: Für $f(x) = 1/x$ führt die Substitutionsregel

$$\int f\big(g(x)\big)g'(x)\,dx = F\big(g(x)\big) + C \quad \text{mit} \quad F'(x) = f(x)$$

zur Integrationsregel

$$\int \frac{g'(x)}{g(x)}\,dx = \ln|g(x)| + C\,.$$

Als erste Anwendung betrachten wir

$$\tan x = \frac{\sin x}{\cos x} = -\frac{g'(x)}{g(x)}$$

mit $g(x) = \cos x$. Nach der gerade bewiesenen Regel ist daher

$$\int \tan x\,dx = -\ln|\cos(x)| + C\,,$$

eine Funktion, die genau wie der Tangens selbst an den Nullstellen der Cosinusfunktion nicht definiert ist.

Beim Tangens hyperbolicus gibt es keine Probleme mit dem Vorzeichen; hier ist ganz entsprechend

$$\int \tanh x \, dx = \ln|\cosh(x)| + C = \ln\cosh(x) + C \,,$$

da der Cosinus hyperbolicus nur positive Werte annimmt.

Auch Integrale wie

$$\int \frac{x}{1+x^2}\, dx$$

lassen sich nach dieser Regel ausrechnen: Da die Ableitung des Nenners $2x$ ist, folgt

$$\int \frac{x}{1+x^2}\, dx = \tfrac{1}{2}\ln\left|1+x^2\right| + C = \tfrac{1}{2}\ln(1+x^2) + C \,.$$

Damit können wir dann beispielsweise den Arkustangens integrieren: Partielle Integration zeigt, daß

$$\begin{aligned}\int 1 \cdot \arctan x \, dx &= x\arctan x - \int \frac{x}{1+x^2}\, dx + C \\ &= x\arctan x - \tfrac{1}{2}\ln(1+x^2) + C\end{aligned}$$

ist.

2) Substitutionen mit linearen Funktionen: Eine der elementarsten Anwendungen der Substitutionsregel, auf die man üblicherweise auch ohne diese Regel kommt, ist die Substitution mit linearen Funktionen $g(x) = ax + b$; hier besagt die Substitutionsregel, daß

$$\int f(ax+b)\cdot a \, dx = F(ax+b) \quad \text{mit} \quad F'(x) = f(x)$$

ist, oder besser

$$\int f(ax+b)\, dx = \frac{F(ax+b)}{a} \quad \text{mit} \quad F'(x) = f(x) \,.$$

Somit ist beispielsweise

$$\int \sin(\omega t + \varphi)\, dt = \frac{-\cos(\omega t+\varphi)}{\omega} + C$$

oder

$$\int \frac{dx}{x^2+a^2} = \int \frac{\frac{dx}{a^2}}{\left(\frac{x}{a}\right)^2+1} = \frac{1}{a}\arctan\left(\frac{x}{a}\right) + C \,.$$

Entsprechend berechnet man auch

$$\int \frac{dx}{x^2-a^2} = -\frac{1}{a}\,\mathrm{artanh}\left(\frac{x}{a}\right) + C$$

und sogar ganz allgemein $\displaystyle\int \frac{dx}{ax^2+bx+c}$:

Da der Fall $a = 0$ uninteressant ist (*bzw.* eine unbedingt notwendige Übungsaufgabe für alle, die nicht sofort einsehen warum), schließen wir dies aus und schreiben

$$ax^2+bx+c = a\left(x+\frac{b}{2a}\right)^2 + c - \frac{b^2}{4a} \,.$$

Mit

$$y = x + \frac{b}{2a} \quad \text{und} \quad d = \frac{c}{a} - \frac{b^2}{4a^2}$$

ist dann

$$\begin{aligned}\int \frac{dx}{ax^2+bx+c} &= \frac{1}{a}\int \frac{dy}{y^2+d} \\ &= \frac{1}{a}\begin{cases} \frac{1}{\sqrt{d}}\arctan\left(\frac{y}{\sqrt{d}}\right)+C & \text{falls } d>0 \\ \frac{-1}{y} & \text{falls } d=0 \\ \frac{1}{\sqrt{-d}}\,\mathrm{artanh}\left(\frac{y}{\sqrt{-d}}\right)+C & \text{falls } d<0 \,, \end{cases}\end{aligned}$$

und je nach Vorzeichen von d kann dies entweder über eine der obigen Formeln oder (für $d = 0$) elementar ausgerechnet werden. Insgesamt erhalten wir mit $\Delta = b^2 - 4ac$ (und damit $d = -\Delta/4a^2$)

$$\int \frac{dx}{ax^2+bx+c} = \begin{cases} \frac{2}{\sqrt{-\Delta}}\arctan\left(\frac{2ax+b}{\sqrt{-\Delta}}\right)+C & \text{falls } \Delta<0 \\ \frac{-2}{2ax+b}+C & \text{falls } \Delta=0 \\ \frac{-2}{\sqrt{\Delta}}\,\mathrm{artanh}\left(\frac{2ax+b}{\sqrt{\Delta}}\right)+C & \text{falls } \Delta>0 \,. \end{cases}$$

3) Substitutionen mit trigonometrischen und Hyperbelfunktionen: Wegen der Beziehungen

$$\sin^2 x + \cos^2 x = 1 \quad \text{und} \quad \cosh^2 - \sinh^2 x = 1$$

bieten sich Substitutionen mit trigonometrischen Funktionen und Hyperbelfunktionen an bei Integralen, in denen Ausdrücke der Form $\sqrt{1 \pm x^2}$ und ähnliche vorkommen.

Betrachten wir als einfachstes Beispiel die Stammfunktion von $\sqrt{1-x^2}$ selbst. Mit der Substitution $x = \sin t$ mit $-\frac{\pi}{2} \le t \le \frac{\pi}{2}$ erhalten wir (durch Rückwärtslesen der Substitutionsregel)

$$\int \sqrt{1-x^2}\,dx = \int \sqrt{1-\sin^2 t}\cos t\,dt\,,$$

und dies ist $\int \cos^2 t\,dt$, da Cosinus zwischen $-\frac{\pi}{2}$ und $\frac{\pi}{2}$ nur nichtnegative Werte annimmt.

Dieses Integral kennen wir von Aufgabe *1b)* des dritten Übungsblatts:

$$\int \cos^2 t\,dt = \tfrac{1}{2}(\sin t\cos t + t) + C\,.$$

Mit der Rücksubstitution $t = \arcsin x$ erhalten wir schließlich nach der kurzen Nebenrechnung

$$\cos(\arcsin x) = \sqrt{1-\sin^2(\arcsin x)} = \sqrt{1-x^2}$$

das Ergebnis

$$\int \sqrt{1-x^2}\,dx = \tfrac{1}{2}(x\sqrt{1-x^2} + \arcsin x) + C\,.$$

Aus ähnliche Weise läßt sich auch die Stammfunktion von $\sqrt{1+x^2}$ bestimmen: Hier bietet sich die Substitution $x = \sinh t$ an und wir erhalten

$$\int \sqrt{1+x^2}\,dx = \int \sqrt{1-\sinh^2 t}\cosh t\,dt = \int \cosh^2 t\,dt\,,$$

da $\cosh t$ nur positive Werte annimmt.

Leider kennen wir das rechtsstehende Integral noch nicht; angesichts der großen Ähnlichkeiten zwischen trigonometrischen Funktionen und

Hyperbelfunktionen lohnt es sich aber sicherlich, als ersten Ansatz vor einer partiellen Integration zu schauen, ob vielleicht etwas analoges gilt wie oben. In der Tat ist nach der Produktregel

$$\frac{d}{dt}\tfrac{1}{2}(\sinh t\cosh t + t) = \cosh^2 t + C\,,$$

also

$$\int \cosh^2 t\,dt = \tfrac{1}{2}(\sinh t\cosh t + t) + C\,.$$

Somit ist

$$\int \sqrt{1+x^2}\,dx = \tfrac{1}{2}(x\sqrt{1+x^2} + \operatorname{arsinh} x) + C\,.$$

Als letztes Beispiel, in dem x einmal nicht quadratisch vorkommt, betrachten wir noch

$$\int \sqrt{\frac{1-x}{1+x}}\,dx\,.$$

Da es in diesem Abschnitt um Substitutionen mit trigonometrischen Funktionen geht, versuchen wir es wieder mit dem Ansatz $x = \sin t$ für $-\frac{\pi}{2} \le t \le \frac{\pi}{2}$ und erhalten

$$\int \sqrt{\frac{1-x}{1+x}}\,dx = \int \sqrt{\frac{1-\sin t}{1+\sin t}}\cos t\,dt\,.$$

Hier hilft nun ein Trick weiter: Erweitern wir den Bruch unter der Quadratwurzel mit $1-\sin t$, so wird das Integral zu

$$\int \sqrt{\frac{(1-\sin t)^2}{1-\sin^2 t}}\cos t\,dt = \int \frac{1-\sin t}{\cos t}\cos t\,dt = \int (1-\sin t)\,dt$$
$$= t + \cos t + C\,.$$

Damit ist

$$\int \sqrt{\frac{1-x}{1+x}}\,dx = \arcsin x + \cos(\arcsin x) + C$$
$$= \arcsin x + \sqrt{1-x^2} + C\,.$$

Bei einem so komplizierten Integral empfiehlt es sich, das Ergebnis durch Differentiation zu überprüfen; wir erhalten

$$\frac{d}{dx}\left(\arcsin x + \sqrt{1-x^2}\right) = \frac{1}{\sqrt{1-x^2}} - \frac{x}{\sqrt{1-x^2}}\,,$$

was zunächst so aussieht, als sei irgend etwas falsch gelaufen. Beachtet man aber, daß

$$\sqrt{1+x}\cdot\sqrt{1-x} = \sqrt{1-x^2}$$

ist, so ist

$$\frac{1}{\sqrt{1-x^2}} - \frac{x}{\sqrt{1-x^2}} = \frac{1-x}{\sqrt{1-x^2}} = \frac{\sqrt{1-x}}{\sqrt{1+x}}\,,$$

was erheblich vertrauenserweckender aussieht.

4) Integrale der Form $\int h(e^{ax})\,dx$***:*** Bei solchen Integralen führt oft die Substitution $u = e^{ax}$ oder, anders ausgedrückt, $x = g(u) = \frac{1}{a}\ln u$ zu einer Vereinfachung. Die Substitutionsregel

$$\int f\big(g(u)\big)g'(u)\,du = \int f(x)\,dx$$

wird wegen $g'(u) = \dfrac{dx}{du} = \dfrac{1}{au}$ zu

$$\int h(u)\,\frac{du}{au} = \int h(e^{ax})\,dx\,;$$

insbesondere wird das Integral also bei rationalem h auf ein Integral einer rationalen Funktion zurückgeführt, und dieses kann nach dem im nächsten Abschnitt skizzierten Verfahren der Partialbruchzerlegung berechnet werden.

Im allgemeinen schreibt man eine Substitution wie die obige kurz in der Form

$$u = e^{ax}, \quad dx = \frac{du}{au}\,,$$

wobei man letztere Beziehung auch aus

$$\frac{du}{dx} = ae^{ax} = au \quad \text{oder} \quad du = audx$$

herleiten kann.

Betrachten wir als Beispiel das Integral $\displaystyle\int \frac{e^{3x}}{e^{2x}-1}\,dx$: Da sowohl e^{2x} als auch e^{3x} im Integranden vorkommen, empfiehlt sich die Substitution $u = e^x$ mit $du = u\,dx$; damit wird

$$\int \frac{e^{3x}}{e^{2x}-1}\,dx = \int \frac{u^3}{u^2-1}\,\frac{du}{u} = \int \frac{u^2}{u^2-1}\,du\,,$$

und wir sind beim Integral einer rationalen Funktion angelangt. Elementare Bruchrechnung zeigt, daß

$$\frac{u^2}{u^2-1} = 1 + \frac{1}{u^2-1} = 1 + \frac{\frac{1}{2}}{u-1} - \frac{\frac{1}{2}}{u+1}\,,$$

d.h.

$$\frac{u^2}{u^2-1}\,du = \int du + \tfrac{1}{2}\left(\int \frac{du}{u-1} - \int \frac{du}{u+1}\right) = u + \ln\left|\frac{u-1}{u+1}\right| + C\,.$$

(Für komplizierte Funktionen werden wir im nächsten Abschnitt ein Verfahren kennenlernen, daß rationale Funktionen in Summen aus einfachen Summanden zerlegt.)

Nun muß nur noch die Substitution $u = e^x$ rückgängig gemacht werden, und wir erhalten als Ergebnis

$$\int \frac{e^{3x}}{e^{2x}-1}\,dx = e^x + \tfrac{1}{2}\ln\left|\frac{e^x-1}{e^x+1}\right| + C\,.$$

Dieselbe Technik funktioniert natürlich auch bei Integralen über Hyperbelfunktionen, da man diese genausogut als Ausdrücke in Exponentialfunktionen schreiben kann. Beispielsweise ist

$$\int \frac{dx}{\sinh x} = 2\int \frac{dx}{e^x - e^{-x}} = 2\int \frac{du}{u\left(u-\frac{1}{u}\right)} = 2\int \frac{du}{u^2-1}$$
$$= \ln\left|\frac{u-1}{u+1}\right| + C = \ln\left|\frac{e^x-1}{e^x+1}\right| + C\,.$$

5) Integrale der Form $\int h(\sin x, \cos x)\,dx$***:*** Als letzte Anwendung der Substitutionsregel in diesem Abschnitt wollen wir noch Integrale betrachten, die von trigonometrischen Funktionen abhängen. Hier führt oft die Substitution

$$x = 2\arctan t, \quad dx = \frac{2\,dt}{t^2+1}$$

zum Erfolg; sie macht $\sin x$ zu $\sin(2\arctan t)$ und $\cos x$ zu $\cos(2\arctan t)$, was wir wie folgt ausrechnen können: Aus

$$\tan^2 x = \frac{\sin^2 x}{\cos^2 x} = \frac{1-\cos^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} - 1$$

folgt, daß

$$\cos(2x) = 2\cos^2(x) - 1 = \frac{2}{1+\tan^2 x} - 1 = \frac{1-\tan^2 x}{1+\tan^2 x}$$

ist und somit

$$\cos(2\arctan t) = \frac{1-t^2}{1+t^2}\,.$$

Da der Arkustangens nur Werte zwischen $-\frac{\pi}{2}$ und $\frac{\pi}{2}$ annimmt, liegt $2\arctan t$ zwischen $-\pi$ und π, ein Intervall, in dem der Sinus dasselbe Vorzeichen hat wie sein Argument. Daher ist

$$\sin(2\arctan t) = \sqrt{1-\cos^2(2\arctan t)} = \sqrt{1-\frac{(1-t^2)^2}{(1+t^2)^2}} = \frac{2t}{1+t^2}\,.$$

Als Beispiel betrachten wir

$$\int \frac{dx}{\cos x} = \int \frac{1+t^2}{1-t^2}\,\frac{2\,dt}{1+t^2} = -2\int \frac{dt}{t^2-1}\,.$$

Das letzte Integral kennen wir:

$$\int \frac{dt}{t^2-1} = \frac{1}{2}\ln\left|\frac{1-t}{1+t}\right| + C\,.$$

Also ist

$$\int \frac{dx}{\cos x} = -\frac{2}{2}\ln\left|\frac{1-t}{1+t}\right| + C = \ln\left|\frac{1+t}{1-t}\right| + C = \ln\left|\frac{1-\tan\frac{x}{2}}{1+\tan\frac{x}{2}}\right| + C\,.$$

i) Integration rationaler Funktionen

Bekanntlich ist ein Polynom in x ein Ausdruck der Form

$$f(x) = \sum_{\nu=0}^{n} a_\nu x^\nu = a_n x^n + \cdots + a_1 x + a_0\,;$$

ist $a_n \neq 0$, so bezeichnet man n als den *Grad* des Polynoms, in Zeichen

$$n = \deg f\,.$$

Polynome vom Grad eins, zwei, drei *bzw.* vier heißen linear, quadratisch, kubisch *bzw.* biquadratisch.

Unter einer *rationalen Funktion* versteht man eine Funktion der Form

$$x \mapsto \frac{f(x)}{g(x)}\,,$$

wobei $f(x)$ und $g(x)$ Polynome sind. Es ist klar, daß diese Funktion nur in jenen Punkten $x \in \mathbb{R}$ definiert ist, in denen $g(x) \neq 0$ ist.

Die Integrationstheorie rationaler Funktionen ist seit langem bekannt; im Prinzip kann man zu jeder rationalen Funktion eine Stammfunktion angeben, indem man wie folgt vorgeht: (Auf Einzelheiten und Beweise muß hier weitgehend verzichtet werden, da wir nicht über das notwendige mathematische Instrumentarium verfügen.)

1. Schritt: Man zerlege den Nenner in ein Produkt aus linearen und quadratischen Polynomen. Die ist zumindest grundsätzlich möglich, denn nach dem *Fundamentalsatz der Algebra* läßt sich ein Polynom vom Grad n (mir reellen oder komplexen Koeffizienten) stets in der Form

$$g(x) = a_n(x-z_1)(x-z_2)\cdot\cdots\cdot(x-z_n)$$

schreiben, wobei die *komplexen* Zahlen z_i die Nullstellen von g sind. Wie man sich leicht überlegt, folgt für ein reelles Polynom g aus $g(z) = 0$ sofort, daß auch $g(\overline{z}) = 0$ ist, d.h. die nichtreellen Nullstellen treten als Paare konjugiert komplexer Zahlen auf, die auch jeweils die gleiche Vielfachheit haben, da letztere über das Verschwinden von Ableitungen definiert werden kann.

Das Produkt der Linearfaktoren zu zwei konjugiert komplexen Zahlen ist

$$(x-z)(x-\overline{z}) = x^2 - (z+\overline{z})x + z\overline{z} = x^2 - 2\,\mathfrak{Re}\,z\cdot x + |z|^2 ,$$

ein quadratisches Polynom mit reellen Koeffizienten. Also läßt sich $g(x)$ als Produkt linearer und quadratischer Polynome mit reellen Koeffizienten schreiben. Faßt man gleiche Faktoren zusammen, so erhält man demnach eine Darstellung

$$g(x) = a_n \ell_1(x)^{e_1} \cdot \dots \cdot \ell_r(x)^{e_r} \cdot q_1(x)^{f_1} \cdot \dots \cdot q_s(x)^{f_s}$$

mit linearen Polynomen ℓ_i, quadratischen Polynomen q_j und natürlichen Zahlen e_i, f_j. Dabei sind sowohl die ℓ_i als auch die q_j paarweise voneinander verschieden.

Diese Zerlegung kann algorithmisch selbst bei Polynomen mit rationalen Koeffizienten algorithmisch sehr aufwendig sein; daher kommt der obige Zusatz „im Prinzip“ für die Durchführbarkeit der Integration rationaler Funktionen.

2. Schritt: Man schreibe die Funktion $f(x)/g(x)$ als Summe eines Polynoms und von *Partialbrüchen* der Form

$$\frac{\alpha_{ik}}{\ell_i(x)^k} \quad \text{und} \quad \frac{\beta_{jk}x + \gamma_{jk}}{q_j(x)^k} ,$$

wobei k von eins bis e_i *bzw.* f_j läuft und α_{ik}, β_{jk} und γ_{jk} reelle Zahlen sind, d.h.

$$\frac{f(x)}{g(x)} = Q(x) \sum_{i=1}^{r} \sum_{k=1}^{e_i} \frac{\alpha_{ik}}{\ell_i(x)^k} + \sum_{j=1}^{s} \sum_{k=1}^{f_j} \frac{\beta_{jk}x + \gamma_{jk}}{q_j(x)^k} .$$

Wir müssen uns zunächst überlegen, warum das möglich ist:

Beginnen wir mit zwei teilerfremden Polynomen $p(x)$ und $q(x)$. Der erweiterte EUKLIDische Algorithmus aus Kapitel I, §2*e)* liefert dazu neue Polynome $\alpha(x), \beta(x)$, so daß

$$\alpha(x)p(x) + \beta(x)q(x) = 1$$

ist und damit

$$\frac{1}{p(x)q(x)} = \frac{\alpha(x)p(x) + \beta(x)q(x)}{p(x)q(x)} = \frac{\alpha(x)}{q(x)} + \frac{\beta(x)}{p(x)} .$$

Ist ein drittes Polynom $r(x)$ teilerfremd sowohl zu $p(x)$ als auch zu $q(x)$, sind insbesondere auch $p(x)q(x)$ und $r(x)$ teilerfremd; es gibt also eine Darstellung

$$\frac{1}{p(x)q(x)r(x)} = \frac{\phi(x)}{p(x)q(x)} + \frac{\psi(x)}{r(x)} = \frac{\phi(x)\alpha(x)}{q(x)} + \frac{\phi(x)\beta(x)}{p(x)} + \frac{\psi(x)}{r(x)}$$

mit zwei neuen Polynomen $\phi(x)$ und $\psi(x)$, insgesamt also wieder eine Summe dreier Brüche mit den drei Faktoren $p(x), q(x)$ und $r(x)$ als Nennern.

Induktiv folgt auf diese Weise, daß es zu n paarweise teilerfremden Polynomen $p_1(x), \dots, p_n(x)$ stets Polynome $\alpha_1(x), \dots, \alpha_n(x)$ gibt, so daß

$$\frac{1}{p_1(x)\cdots p_n(x)} = \frac{\alpha_1(x)}{p_1(x)} + \dots + \frac{\alpha_n(x)}{p_n(x)}$$

ist. Für ein beliebiges weiteres Polynom ist daher auch

$$\frac{f(x)}{p_1(x)\cdots p_n(x)} = \frac{\alpha_1(x)f(x)}{p_1(x)} + \dots + \frac{\alpha_n(x)f(x)}{p_n(x)} .$$

Dies wenden wir an auf die obige Zerlegung des Nenners g: Da die $\ell_i(x)$ und die $q_j(x)$ paarweise teilerfremd sind, gilt dasselbe auch für ihre Potenzprodukte, es gibt also Polynome $\varphi_i(x), \psi_j(x)$, so daß gilt

$$\frac{f(x)}{g(x)} = \sum_{i=1}^{r} \frac{\varphi_i(x)}{\ell_i(x)^{e_i}} + \sum_{j=1}^{s} \frac{\psi_j(x)}{q_j(x)^{f_j}} .$$

Um den Grad dieser Polynome zu reduzieren, dividieren wir für jeden Bruch den Zähler mit Rest durch den Nenner und fassen die Quotientenpolynome zusammen zu einem einzigen Polynom $Q(x)$. Damit erhalten wir eine neue Darstellung

$$\frac{f(x)}{g(x)} = Q(x) + \sum_{i=1}^{r} \frac{\widetilde{\varphi}_i(x)}{\ell_i(x)^{e_i}} + \sum_{j=1}^{s} \frac{\widetilde{\psi}_j(x)}{q_j(x)^{f_j}} ,$$

in der jeder Zähler $\widetilde{\varphi}_i(x)$ und $\widetilde{\psi}_j(x)$ kleineren Grad hat als der zugehörige Nenner.

Zur weiteren Zerlegung der Zähler überlegen wir uns zunächst, daß es zu je zwei Polynomen $\varphi(x)$ und $g(x)$ eine Darstellung

$$\varphi(x) = \alpha_0(x) + \alpha_1(x)g(x) + \alpha_2 g(x)^2 + \cdots$$

gibt mit $\deg \alpha_i(x) < \deg g(x)$. (Für ein lineares Polynom $g(x) = x - a$ ist das gerade die TAYLOR-Entwicklung von φ im Punkt a.) Für $\alpha_0(x)$ nehmen wir dazu den Divisionsrest von $\varphi(x)$ durch $g(x)$; ist $q_0(x)$ der Quotient, gilt dann

$$g(x) = \alpha_0(x) + g(x)q_0(x)\,.$$

Ist nun q_1 der Quotient und α_1 der Rest bei der Division von q_0 durch g, so erhalten wir die neue Darstellung

$$g(x) = \alpha_0(x)+g(x)\big(\alpha_1(x)+g(x)q_1(x)\big) = \alpha_0(x)+\alpha_1(x)g(x)+q_1(x)g(x)^2\,.$$

$\alpha_2(x)$ ist nun entsprechend der Divisionsrest von $q_1(x)$ durch $g(x)$ und so weiter, bis ein Quotient $q_\nu(x)$ kleineren Grad als $g(x)$ hat und damit der letzte Koeffizient $\alpha_{\nu+1}(x)$ ist.

Ist $g(x)$ ein lineares Polynom, haben alle $\alpha_\nu(x)$ kleineren Grad als eins, sind also Konstanten. Da $\deg \widetilde{\varphi}(x) < \deg \ell_i^{e_i} = e_i$ ist, können wir also schreiben

$$\widetilde{\varphi}(x) = \alpha_0 + \alpha_1\ell_i(x) + \cdots + \alpha_{e_i-1}\ell_i(x)^{e_i-1}$$

mit reellen Zahlen α_ν, und

$$\begin{aligned}\frac{\widetilde{\varphi}(x)}{\ell_i(x)^{e_i}} &= \frac{\alpha_0 + \alpha_1\ell_i(x) + \cdots + \alpha_{e_i-1}\ell_i(x)^{e_i-1}}{\ell_i(x)^{e_i-1}}\\ &= \frac{\alpha_0}{\ell_i(x)^{e_i}} + \frac{\alpha_1}{\ell_i(x)^{e_i-1}} + \cdots + \frac{\alpha_{e_i-1}}{\ell_i(x)}\,.\end{aligned}$$

Entsprechend erhalten wir die Darstellungen

$$\frac{\widetilde{\psi}(x)}{q_i(x)^{f_i}} = \frac{\beta_0 x+\gamma_0}{q_i(x)^{f_i}} + \frac{\beta_1 x}{q_i(x)^{f_i-1}} + \cdots + \frac{\beta_{f_i-1}x+\gamma_{f_i-1}}{q_i(x)}\,.$$

Faßt man alles zusammen und gibt den Zählern die richtigen Indizes, erhält man die oben angegebene Zerlegung, die sogenannte *Partialbruchzerlegung.*

Der gerade durchgeführte Beweis ist zwar (bei einer gegebenen Zerlegung des Nenners) konstruktiv, das verwendete Verfahren ist allerdings nicht besonders effizient. Nachdem wir wissen, daß eine Zerlegung wie oben existiert und auch alle Nenner kennen, ist es beispielsweise erheblich einfacher, von der obigen Zerlegung auszugehen (mit unbestimmten Zählern) und die Partialbrüche nach den üblichen Regeln der Bruchrechnung zu addieren. Der Nenner des entstehenden Bruchs ist $g(x)$, der Zähler $f(x) - Q(x)g(x)$, wobei $Q(x)$ der Quotient bei der Division von $f(x)$ durch $g(x)$ ist. Der durch Addition der Partialbrüche berechnete Zähler ist eine lineare Funktion in den Koeffizienten α, β, γ; Koeffizientenvergleich mit $f(x) - Q(x)q(x)$ liefert ein lineares Gleichungssystem für die Koeffizienten, das wir schnell und einfach nach GAUSS lösen können.

Für Linearfaktoren $\ell_i(x)$, die nur in der ersten Potenz vorkommen, gibt es sogar ein noch einfacheres Verfahren: Ist x_i die (einzige) Nullstelle von $\ell_i(x)$, so überlegt man sich leicht (durch Addition der Partialbrüche und des polynomialen Anteils), daß

$$\alpha_i = \lim_{x\to x_i} \frac{(x-x_i)f(x)}{g(x)}$$

ist. Da Zähler und Nenner für $x = x_i$ verschwinden, muß dieser Grenzwert nach DE L'HOSPITAL berechnet werden; wir erhalten

$$\alpha_i = \lim_{x\to x_i} \frac{(x-x_i)f'(x) - f(x)}{g'(x)} = -\frac{f(x_i)}{g'(x_i)}\,,$$

wobei hier der Nenner nicht verschwinden kann, da x_i nur eine einfache Nullstelle von g ist.

3. Schritt: Integration der Partialbrüche: Die Integrale der Funktionen $\frac{\alpha_{ik}}{\ell_i(x)^k}$ lassen sich über die Substitution $u = \ell_i(x)$ sofort auf $\int \frac{du}{u^k}$ zurückführen und sind somit problemlos. Weiter ist

$$\int \frac{\beta x+\gamma}{ax^2+bx+c}\,dx = \frac{\beta}{2a}\int \frac{2ax+b}{ax^2+bx+c}\,dx + \int \frac{\gamma - \frac{\beta b}{2a}}{ax^2+bx+c}\,dx\,,$$

wobei das erste Integral nach der Regel über die logarithmische Ableitung aus Abschnitt *a)* berechnet werden kann, während wir das zweite in Abschnitt *b)* behandelt haben.

Bleiben schließlich noch die Integrale mit Potenzen eines quadratischen Polynoms im Nenner; hier wird k durch partielle Integration rekursiv erniedrigt.

Als ganz einfaches Beispiel, in dem alles explizit durchführbar ist, betrachten wir

$$\int \frac{dx}{x^2 - 5x + 6}\,.$$

Da hier sechs das Produkt und fünf die Summe der Nullstellen des Nenners ist, errät man leicht, daß $x^2 - 5x + 6 = (x-2)(x-3)$ ist. Die Partialbruchzerlegung bestimmen wir nun durch probieren: Da

$$\frac{1}{x-2} - \frac{1}{x-3} = \frac{(x-3)-(x-2)}{(x-2)(x-3)} = \frac{-1}{(x-2)(x-3)}$$

ist, folgt

$$\int \frac{dx}{(x-2)(x-3)} = -\int \frac{dx}{x-2} + \int \frac{dx}{x-3} = \ln\left|\frac{x-3}{x-2}\right| + C\,.$$

j) Symmetrie

Bei bestimmten Integralen kann man sich gelegentlich einen Großteil der Mühe des Ausrechnens ersparen, indem man Symmetrien berücksichtigt: Ist etwa die Funktion f spiegelsymmetrisch zur Achse $x = c$, d.h. $f(c+x) = f(c-x)$ für alle x, für die beide Seiten definiert sind, so ist auch

$$\int_c^{c+a} f(x)\,dx = \int_{c-a}^{c} f(x)\,dx\,,$$

d.h.

$$\int_{c-a}^{c+a} f(x)\,dx = 2\int_c^{c+a} f(x)\,dx\,.$$

Viel interessanter wird es, wenn f *punktsymmetrisch* zu einem Punkt $(c,0)$ auf der x-Achse ist, d.h. $f(c+x) = -f(c-x)$. Dann ist

$$\int_c^{c+a} f(x)\,dx = -\int_{c-a}^{c} f(x)\,dx \quad \text{und} \quad \int_{c-a}^{c+a} f(x)\,dx = 0\,,$$

ohne daß man irgendetwas über die Stammfunktion von f wissen müßte. So ist beispielsweise

$$\int_{-1}^{1} \frac{\sin(x)}{|x|}\,dx = 0\,,$$

aber ein Leser, der mit den uns bislang bekannten Methoden nach einer Stammfunktion suchen möchte, dürfte eine sehr harte Zeit und wenig Erfolg haben.

k) Einige nicht elementar integrierbare Funktionen

Tatsächlich ist die Stammfunktion von $\dfrac{\sin x}{x}$ nicht durch elementare transzendenten Funktionen wie die Exponentialfunktion, die trigonometrischen Funktionen sowie deren Umkehrfunktionen und/oder Wurzeln und Grundrechenarten darstellbar, und sie teilt diese Eigenschaft mit einer ganzen Reihe weiterer Stammfunktionen.

Da einige dieser Stammfunktionen trotzdem für Anwendungen wichtig sind, behilft man sich damit, daß man einigen speziellen dieser Funktionen Namen gibt und versucht, alles darauf zurückzuführen. Die so definierten Funktionen können dann, genau wie der Sinus und die Exponentialfunktion, in Tabellenwerke und Unterprogrammbibliotheken aufgenommen werden, so daß man mit den dadurch abgedeckten Integralen genauso rechnen kann wie mit den elementaren transzendenten Funktionen.

Einige wichtige dieser neuen Funktionen sind

1) Der Integralsinus: Da $\lim_{t\to 0} \dfrac{\sin t}{t} = 1$ ist, ist das bestimmte Integral

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t}\, dt$$

für alle $x \in \mathbb{R}$ wohldefiniert und eine stetige Funktion von x, der *Integralsinus.* Er ist beispielsweise wichtig für Tiefpaßfilterungen, wo Faltungsintegrale mit $\frac{\sin x}{x}$ berechnet werden können.

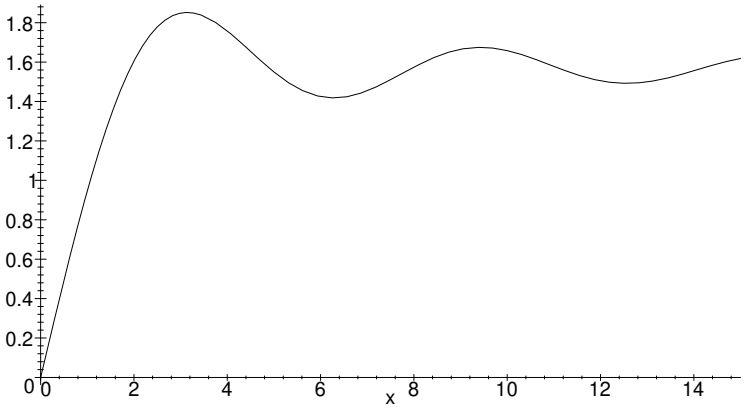


Abb. 64: Der Integralsinus

2) Die Fehlerfunktion: Auch die Funktion e^{-x^2} hat keine elementar angebbare Stammfunktion; hier definiert man als eine Stammfunktion die *Fehlerfunktion* oder *error function* durch

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt\,.$$

(Der Vorfaktor sorgt dafür, daß $\lim_{x\to\infty} \mathrm{Erf}(x) = 1$ ist, was wir im Augenblick allerdings noch nicht beweisen können.)

Sie ist wichtig für die Statistik, denn für eine normalverteilte Zufallsvariable (dazu gehören die meisten Meßgrößen) mit Mittelwert $\overline{x}$ und

Standardabweichung σ ist die Wahrscheinlichkeit dafür, daß ein Meßwert zwischen a und b liegt, gleich

$$\frac{1}{2\pi} \int_a^b e^{\frac{1}{2}\left(\frac{x-\overline{x}}{\sigma}\right)^2}\, dx\,,$$

ein Integral, von dem man sich leicht überlegt, daß es mit Hilfe von Erf berechnet werden kann als

$$\frac{1}{2}\,\mathrm{Erf}\left(\frac{\sqrt{2}}{2}\,\frac{\overline{x}-a}{\sigma}\right) - \frac{1}{2}\,\mathrm{Erf}\left(\frac{\sqrt{2}}{2}\,\frac{\overline{x}-b}{\sigma}\right)\,.$$

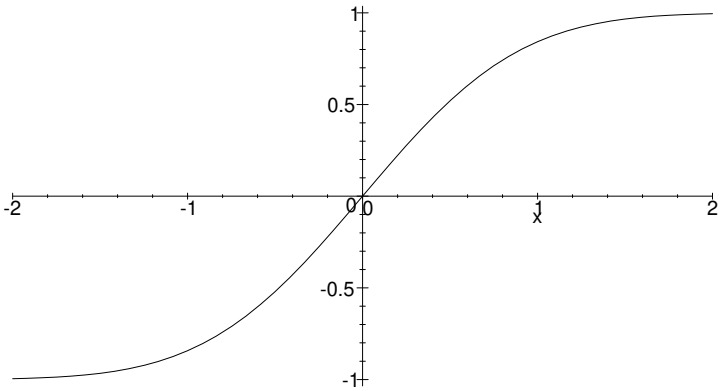


Abb. 65: Die Fehlerfunktion

3) Elliptische Integrale: Bei der Berechnung der Bogenlänge eines Ellipsensegments kommt man, je nach Art des Ansatzes und der Substitution, auf Integrale der Form

$$\int \frac{dx}{\sqrt{x^3 + ax^2 + bx + c}}\,,$$

wobei das Polynom im Nenner keine mehrfachen Nullstellen hat, oder auf Integrale der Form

$$\int \frac{dx}{\sqrt{1 - k^2 \sin^2 x}} \quad \text{oder} \quad \int \sqrt{1 - k^2 \sin^2 x}\, dx$$

mit $0 < k < 1$. Alle diese Integrale heißen *elliptische Integrale* und sind nicht elementar ausdrückbar. Eine hauptsächlich auf ADRIEN MARIE LEGENDRE (1752–1833) und KARL THEODOR WILHELM WEIERSTRASS (1815–1897) zurückgehende Theorie der elliptischen Funktionen stellt das Instrumentarium bereit, mit dem man diese vor allem in der Geodäsie, Kartographie und im Maschinenbau wichtigen Integrale berechnen kann.

4) Algebraische Integrale: Die erste Form der elliptischen Integrale ist ein Spezialfall eines sogenannen algebraischen Integrals; das sind Integrale, deren Integranden durch Wurzeln und Grundrechenarten (oder allgemeiner implizit durch Lösungen von Polynomgleichungen) gegeben sind. Die komplizierteren dieser Integrale sind fast alle nicht elementar ausdrückbar; es gibt inzwischen algorithmische Verfahren, die entscheiden, wann ein solches Integral elementar ausdrückbar ist, und die dann auch eine Stammfunktion finden können. Natürlich gibt es auch hier spezielle Funktionen, mit denen sich weitere dieser Integrale ausdrücken lassen.

l) Uneigentliche Integrale

Sei $a > 0$ und $b > a$. Dann ist für eine reelle Zahl $r \neq 1$

$$\int_a^b \frac{dx}{x^r} = \frac{-1}{(r-1)x^{r-1}} \Bigg|_a^b = \frac{1}{r-1}\left(\frac{1}{a^{r-1}} - \frac{1}{b^{r-1}}\right) .$$

Falls $r > 1$ ist, können wir hiervon den Grenzwert für b gegen unendlich betrachten und es liegt nahe, diesen als Wert des Integrals von a bis unendlich zu bezeichnen:

$$\int_a^\infty \frac{dx}{x^r} \underset{\text{def}}{=} \frac{1}{r-1}\frac{1}{a^{r-1}} \quad \text{für} \quad r > 1 .$$

Auch wenn wir nicht bis unendlich integrieren wollen, gibt es Beispiele von Integralen, denen wir via Grenzwertbetrachtung einen sinnvollen

Wert zuordnen können, ohne daß das Integral im Sinne unserer bisherigen Definitionen existieren würde: Beispielsweise ist für $a \leq c < b$ und eine relle Zahl $0 < r < 1$

$$\int_a^c \frac{dx}{(b-x)^r} = \frac{(b-x)^{1-r}}{r-1} \Bigg|_a^c = \frac{(b-a)^{1-r}}{r-1} - \frac{(b-c)^{1-r}}{r-1} ,$$

und auch hier liegt es nahe, den Grenzwert für $c \to b$ als Wert des Integrals von a bis b zu bezeichnen. Wir müssen hier allerdings vorsichtig sein mit dem Grenzübergang, denn die obige Formel gilt natürlich nur für $c < b$; für $c > b$ ist das Integral undefiniert. In einer solchen Situation können wir daher nicht den gewöhnlichen Limes betrachten, sondern brauchen einen eingeschränkten Grenzwertbegriff, der nur Folgen von einer der beiden Seiten berücksichtigt: Allgemein schreiben wir

$$y = \lim_{x \to b-} f(x) ,$$

wenn es für jedes $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß $|f(x) - y| < \varepsilon$ für alle x mit $b - \delta < x < b$. (Zur Erinnerung: Beim gewöhnlichen Grenzwert fordert man dies für alle x mit $|x - b| < \delta$, d.h. $b - \delta < x < b + \delta$.)

Völlig analog läßt sich natürlich auch ein rechtsseitiger Grenzwert

$$y = \lim_{x \to a+} f(x)$$

definieren durch die Bedingung, daß es für jedes $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß $|f(x) - y| < \varepsilon$ für alle x mit $b < x < b + \delta$.

Schließlich sollten wir, wenn wir schon beim Definieren sind, zur Bezeichnungsökonomie noch vereinbaren, daß halboffene Intervalle wie

$$[a, b) = \left\{ x \in \mathbb{R} \;\middle|\; a \leq x < b \right\}$$

auch für unendliche b definiert sein sollen durch

$$[a, \infty) \underset{\text{def}}{=} \left\{ x \in \mathbb{R} \;\middle|\; a \leq x \right\} ;$$

entsprechend auch

$$(-\infty, b] \underset{\text{def}}{=} \left\{ x \in \mathbb{R} \;\middle|\; x \leq b \right\} \quad \text{und} \quad (-\infty, \infty) = \mathbb{R} .$$

Für unendliche Intervallgrenzen sind die Definitionen für links- und rechtsseitige Grenzwerte nicht sinnvoll anwendbar; wir vereinbaren daher, daß für $b = \infty$ der Ausdruck $\lim\limits_{c\to b-} \int\limits_a^c f(x)\,dx$ für den gewöhnlichen Grenzwert $\lim\limits_{c\to\infty}$ stehen soll; entsprechend bei $\lim\limits_{x\to a+} f(x)$ für $a = -\infty$.

Mit all diesen Definitionen können wir dann eine Funktion f betrachten, die in einem halboffenen Intervall $[a, b)$ mit $b \in \mathbb{R} \cup \{\infty\}$ definiert und stückweise stetig ist; für diese definieren wir das rechtsseitig uneigentliche Integral

$$\int_a^b f(x)\,dx = \lim_{c\to b-} \int_a^c f(x)\,dx\,,$$

falls dieser Grenzwert existiert; andernfalls sagen wir, das Integral sei *divergent.*

Völlig analog definieren wir linksseitige uneigentliche Integrale: f sei stückweise stetig auf dem halboffenen Intervall $(a, b]$ mit $a \in \mathbb{R}\cup\{-\infty\}$; dann ist

$$\int_a^b f(x)\,dx = \lim_{c\to a+} \int_c^b f(x)\,dx\,,$$

falls dieser Grenzwert existiert; andernfalls sagen wir, das Integral sei *divergent.*

Diese Definitionen sind immer noch nicht allgemein genug: Eine Funktion könnte auch an *beiden* Enden eines Intervalls (a, b) undefiniert sein, wobei wir auch die Sonderfälle $a = -\infty$ und/oder $b = \infty$ zulassen wollen, und zusätzlich könnte sie auch noch Undefiniertheitsstellen $c_1 < \cdots < c_r$ im Intervallinnern haben.

In diesem Fall läßt sich das Intervall so in Teilintervalle zerlegen, daß f in jedem Teilintervall höchstens an *einem* der beiden Intervallenden uneigentlich ist: Falls es keine Undefiniertheitsstellen im Intervallinnern gibt, wählen wir willkürlich ein c_0 zwischen a und b und betrachten die beiden Intervalle $(a, c]$ und $[c, b)$. Im anderen Fall können zwei Zusatzpunkte notwendig sein: ein Punkt c_0 zwischen a und c_1 sowie ein Punkt c_{r+1} zwischen c_r und b.

Wir sagen dann, das uneigentliche Integral $\int\limits_a^b f(x)\,dx$ konvergiere, wenn *jedes* der Integrale

$$\int_a^{c_0} f(x)\,dx, \quad \int_{c_0}^{c_1} f(x)\,dx, \quad \ldots, \quad \int_{c_{r+1}}^{b} f(x)\,dx$$

konvergiert; die Summe ihrer Werte bezeichnen wir als den Wert des Integrals von a bis b.

Als Beispiel betrachten wir das an beiden Grenzen uneigentliche Integral

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2}\,.$$

Da der Integrand eine gerade Funktion ist, empfiehlt es sich, das Integrationsintervall, das hier aus ganz $\mathbb{R}$ besteht, bei Null zu unterteilen, und wir erhalten

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \int_{-\infty}^{0} \frac{dx}{1+x^2} + \int_0^{\infty} \frac{dx}{1+x^2} = 2\int_0^{\infty} \frac{dx}{1+x^2}$$

$$= 2\lim_{c\to\infty} \int_0^c \frac{dx}{1+x^2} = 2(\lim_{c\to\infty} \arctan c - \arctan 0) = \pi\,.$$

Obige Forderung, daß *jedes* der Teilintegrale einzeln konvergieren soll, ist gelegentlich etwas sehr restriktiv. Für das bei Null uneigentliche Integral

$$\int_{-2}^{4} \frac{dx}{x^3}$$

könnte man etwa argumentieren, daß der Integrand eine ungerade Funktion ist, so daß aus Symmetriegründen das Integral von -2 bis 2 verschwinden sollte und

$$\int_{-2}^{4} \frac{dx}{x^3} = \int_{-2}^{2} \frac{dx}{x^3} + \int_2^{4} \frac{dx}{x^3} = 0 + \frac{3}{32} = \frac{3}{32}$$

sein sollte. Diese Art der Argumentation ist durch das, was wir bislang gelernt haben, nicht gedeckt, und es gibt auch gute Gründe, sie zu verbieten: Schließlich geht die Stammfunktion $-\frac{1}{2}x^{-2}$ für $x \to 0$ gegen minus unendlich, und wenn eine Größe mit Relevanz in der realen Welt unendlich groß wird, hat dies im Allgemeinen zu gravierende Konsequenzen, als daß man einfach durch diesen Punkt hindurch weitergehen könnte.

Andererseits sind aber viele der mathematischen Formeln, die in den Naturwissenschaften und der Technik angewandt werden, nur näherungsweise gültig: Mathematische Modelle sind praktisch immer *vereinfachende* Modelle der Wirklichkeit, beispielsweise gilt das OHMsche Gesetz sicher nicht mehr, wenn man einen 5 Ω-Widerstand aus einer auf 5 V Spannung ausgelegten Schaltung im Hochspannungslabor mit 100 kV belastet, und es gilt auch nicht mehr ohne Korrekturterme, wenn man einen Wechselspannung mit 500 MHz anlegt.

Entsprechend gibt es durchaus Situationen, in denen das mathematische Modell einen unendlich großen Wert vorhersagt, wohingegen in der Realität limitierende Faktoren, die für Werte im „üblichen“ Größenbereich noch keine nennenswerte Rolle spielen, für eine Begrenzung sorgen. Falls man in einer solchen Situation sicher sein kann, daß auch in der realen Situation noch die Symmetrie zum Nullpunkt erhalten bleibt, kann man so wie oben argumentieren; falls allerdings die Symmetrie *nicht* erhalten bleibt, können durch die Begrenzung der Funktion beliebig große Abweichungen erzeugt werden, über die man mit dem vereinfachten mathematischen Modell nichts aussagen kann.

Da somit alles von der Anwendung abhängt, kann die Mathematik hier nicht mehr bieten als eine *Definition:* Falls für die Funktion f, die auf $[a, c) \cup (c, b]$ definiert ist, der Grenzwert

$$\lim_{h\to 0} \int_a^{c-h} f(x)\,dx + \int_{c-h}^{b} f(x)\,dx$$

existiert, bezeichnen wir ihn als CAUCHYschen Hauptwert von $\int_a^b f(x)\,dx$ nach dem aus der Analysis I bekannten französischen Mathematiker

Baron AUGUSTIN LOUIS CAUCHY (1789–1857). Entsprechend reden wir auch in komplizierteren Situationen mit mehreren Unstetigkeitsstellen vom CAUCHYschen Hauptwert, falls sich eine Aufteilung in Teilintervalle finden läßt, so daß für jedes Teilintervall der CAUCHYsche Hauptwert existiert. Im obigen Beispiel wäre also $\frac{3}{32}$ der CAUCHYsche Hauptwert des Integrals, wohingegen das Integral selbst undefiniert ist.

Die Frage, wann der CAUCHYsche Hauptwert für ein eigentlich divergentes Integral verwendet werden sollte, ist keine mathematische Frage: Unter rein mathematischen Gesichtspunkten gibt es **nie** eine Rechtfertigung für die Verwendung des CAUCHYschen Hauptwerts. Der CAUCHYsche Hauptwert ist nur dann sinnvoll anwendbar, wenn man davon ausgeht, daß ein mathematisches Modell eine Situation nur für nicht zu große Funktionswerte (ungefähr) korrekt beschreibt, und wenn man gleichzeitig sicher ist, daß die Unendlichkeitsstelle des mathematischen Modells für die Anwendung unproblematisch ist und gleichzeitig die Symmetrie, die der Berechnung des CAUCHYschen Hauptwerts zugrundeliegt, auch in der realen Anwendung gilt.

Der CAUCHYsche Hauptwert darf auch *nie* als eine Rechtfertigung dafür verstanden werden, daß man unbesonnen

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

setzt, wobei F eine zwar in den Punkten a und b, nicht aber auch für jeden Zwischenwert $a \le x \le b$ definierte Stammfunktion von f ist: So etwas kann zu Ergebnissen wie

$$\int_{-2}^{2} \frac{dx}{x^2} = -\frac{1}{x}\Big|_{-2}^{2} = \frac{-1}{2} - \frac{-1}{-2} = -1$$

führen, und natürlich ist keine Anwendung denkbar, in der eine negative Zahl in sinnvoller Weise als Integral über eine überall positive Funktion angesehen werden kann. In der Tat existiert im obigen Beispiel weder das Integral noch dessen CAUCHYscher Hauptwert, da sich die Unendlichkeiten links und rechts der Null hier nicht wegheben, sondern verstärken.

Zu einer etwas systematischeren Untersuchung uneigentlicher Integrale empfiehlt es sich, zunächst die Potenzen zu betrachten. Für positive Exponenten r ist x^r überall definiert, so daß Integrale über einen *endlichen* Bereich unproblematisch sind; für Integrale über einen *unendlichen* Bereich rechnet man leicht nach, daß sie immer divergieren. Für $r = 0$ ändert sich nichts an dieser Situation; interessant ist also nur der Fall $r < 0$, wo sowohl der Wert $x = 0$ als auch unendliche obere und/oder untere Grenzen zu Problemen führen können. Für negatives r ist $x^r = 1/x^{-r}$, wir interessieren uns also für

$$\int_a^b \frac{dx}{x^r} = \frac{1}{(1-r)x^{r-1}}\Bigg|_a^b \quad \text{für} \quad r > 0, \quad r \neq 1 \,.$$

Für $a > 0$ und $b \to \infty$ existiert ein Grenzwert genau dann, wenn $r - 1 > 0$, also $r > 1$ ist; alsdann ist

$$\int_a^\infty \frac{dx}{x^r} = \lim_{b\to\infty} \int_a^b \frac{dx}{x^r} = \frac{1}{(r-1)a^{r-1}} \,.$$

Für $b > 0$ und $a \to \infty$ existiert ein Grenzwert genau dann, wenn $r - 1 < 0$, also $r < 1$ ist; alsdann ist

$$\int_0^b \frac{dx}{x^r} = \lim_{a\to 0} \int_a^b \frac{dx}{x^r} = \frac{1}{1-r} b^{1-r} \,.$$

Zusammen mit der Monotonieeigenschaft des RIEMANN-Integrals aus §3*a)* ergeben sich hieraus zwei allgemeine Kriterien für die Konvergenz uneigentlicher Integrale:

Satz: Die Funktion f sei stetig für $x \geq a$ und g sei stetig für $0 < x \leq b$. Dann gilt:

1.) Falls es eine reelle Zahl K und eine reelle Zahl $r > 1$ gibt, so daß $|f(x)| \leq \frac{K}{x^r}$ ist, konvergiert $\int_a^\infty f(x)\,dx$.
2.) Falls es eine reelle Zahl K und eine reelle Zahl $0 < r < 1$ gibt, so daß $|g(x)| \leq \frac{K}{x^r}$ ist, konvergiert $\int_0^b g(x)\,dx$.

Beweis: 1.) Nach der Monotonieregel ist für jedes $b \geq a$

$$\int_a^b |f(x)|\,dx \leq K \int_a^b \frac{dx}{x^r} \,,$$

und letzteres Integral konvergiert, wie wir gerade nachgerechnet haben, unter den angenommenen Voraussetzungen. Das linksstehende Integral ist somit beschränkt durch eine von b unabhängige Konstante; da es wegen der Nichtnegativität des Betrags zusätzlich eine monoton wachsende Funktion von b ist, existiert daher der Grenzwert nach dem bekannten, schon für die Existenz des RIEMANN-Integrals verwendeten Satz aus der Analysis I, wonach jede monotone und beschränkte Folge reeller Zahlen konvergent ist.

Genau dasselbe gilt für die ebenfalls nichtnegative Funktion $f(x)+|f(x)|$, die durch $\frac{2K}{x^r}$ beschränkt ist; also existieren die uneigentlichen Integrale

$$\int_a^\infty \big(f(x) + |f(x)|\big)\,dx \quad \text{und} \quad \int_a^\infty |f(x)|\,dx \,,$$

und damit auch das gesuchte Integral als ihre Differenz.

2.) geht völlig analog. ∎

Als Beispiel hierfür betrachten wir die Strahlung eines schwarzen Körpers: Nach dem RAYLEIGHschen Strahlungsgesetz, wonach alle mit der Geometrie des Körpers verträglichen Wellenzahlvektoren gleichwahrscheinlich sind, wäre die Energiedichte proportional zum Quadrat der Frequenz, die Gesamtenergie also proportional zu

$$\int_0^\infty \nu^2\,d\nu \,,$$

einem offensichtlich divergenten Integral: Das ist die sogenannte „UV-Katastrophe“, die dieses Modell zum Widerspruch führt.

Die Abhilfe besteht bekanntlich darin, daß man die Gleichverteilung der Frequenzen durch eine BOSE-EINSTEIN-Statistik ersetzt und als neue

Energiedichte nun nach dem PLANCKschen Strahlungsgesetz

$$\frac{8\pi h}{c^2}\,\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}$$

erhält, wobei h das PLANCKsche Wirkungsquantum, k die BOLTZMANN-Konstante, c die Lichtgeschwindigkeit und T die absolute Temperatur bezeichnet. Hier ist die Gesamtenergie

$$\frac{8\pi h}{c^2}\int_0^\infty \frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,d\nu\,,$$

und die UV-Katastrophe wird genau dann vermieden, wenn dieses Integral konvergiert.

Mit der Substitution $x = \dfrac{h\nu}{kT}$ erhalten wir

$$\frac{8\pi h}{c^2}\int_0^\infty \frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,d\nu = \frac{8\pi h}{c^2}\left(\frac{kT}{h}\right)^4\int_0^\infty \frac{x^3}{e^x-1}\,dx\,.$$

Dieses Integral ist an beiden Grenzen uneigentlich; betrachten wir also eine positive Konstante a und getrennt die Integrale von 0 bis a und von a bis ∞.

Letzteres Integral existiert nach dem gerade bewiesenen Satz, wenn wir K, r finden können, so daß

$$\frac{x^3}{e^x-1} \le \frac{K}{x^r} \quad \text{für alle } x \ge a\,.$$

Diese Ungleichung ist äquivalent zu

$$\frac{1}{e^x-1} \le \frac{K}{x^{r+3}} \quad \text{oder} \quad e^x \ge 1 + \frac{x^{r+3}}{K}\,,$$

und dies gilt mit geeignetem a, K und (beispielsweise) $r = 2$, da die Exponentialfunktion schneller wächst als jedes Polynom.

Entsprechend gilt für $0 \le x \le 1$ die Ungleichung etwa mit $r = \frac{1}{2}$, da dort $e^x - 1$ stärker wächst als jede x-Potenz. Also konvergiert das Integral sowohl an seiner unteren als auch seiner oberen Grenze, und

dazwischen ist ohnehin alles unproblematisch, da der Integrand stetig ist.

■

Als (vorerst) letztes Beispiel eines uneigentlichen Integrals sei noch

$$\Gamma(x) \underset{\text{def}}{=} \int_0^\infty e^{-t}t^{x-1}\,dt \quad \text{für } x > 0$$

betrachtet. Dieses Integral ist natürlich immer uneigentlich an der oberen Grenze; für $x < 1$ zusätzlich auch noch an der unteren.

Diese untere Grenze ist völlig harmlos, denn für $0 < x < 1$ ist

$$e^{-t}t^{x-1} = \frac{e^{-t}}{t^{1-x}} \le \frac{1}{t^{1-x}} \quad \text{mit} \quad 0 < 1-x < 1\,,$$

so daß obiger Satz sofort die Konvergenz zeigt.

Auch die obere Grenze ist unproblematisch, denn die dazu notwendige Abschätzung

$$e^{-t}t^{x-1} \le \frac{K}{t^r} \iff e^t \ge t^{r+x-1}$$

folgt wieder, da die Exponentialfunktion stärker wächst als jede Potenz.

Die somit für alle $x > 0$ definierte Funktion $x \mapsto \Gamma(x)$ heißt EULERsche *Gamma-Funktion.*

LEONHARD EULER (1707–1783) wurde in Basel geboren und ging auch dort zur Schule und, im Alter von 14 Jahren, zur Universität. Dort legte er zwei Jahre später die Magisterprüfung in Philosophie ab und begann mit dem Studium der Theologie; daneben hatte er sich seit Beginn seines Studium unter Anleitung von JOHANN BERNOULLI mit Mathematik beschäftigt. 1726 beendete er sein Studium in Basel und bekam eine Stelle an der Petersburger Akademie der Wissenschaften, die er 1727 antrat. Auf Einladung FRIEDRICHS DES GROSSEN wechselte er 1741 an die preußische Akademie der Wissenschaften; nachdem sich das Verhältnis zwischen den beiden dramatisch verschlechtert hatte, kehrte er 1766 nach St. Petersburg zurück. Im gleichen Jahr erblindete er vollständig; trotzdem schrieb er rund die Hälfte seiner zahlreichen Arbeiten (73 Bände) danach. Sie enthalten bedeutende Beiträge zu vielen Gebieten der Mathematik, Physik, Astronomie und Kartographie.

Die wichtigste Eigenschaft der Γ-Funktion folgt durch partielle Integration:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt = e^{-t}\frac{t^x}{x}\,\Big|_0^\infty + \frac{1}{x}\int_0^\infty e^{-t} t^x\, dt = \frac{\Gamma(x+1)}{x}$$

oder

$$\Gamma(x+1) = x\Gamma(x)\,.$$

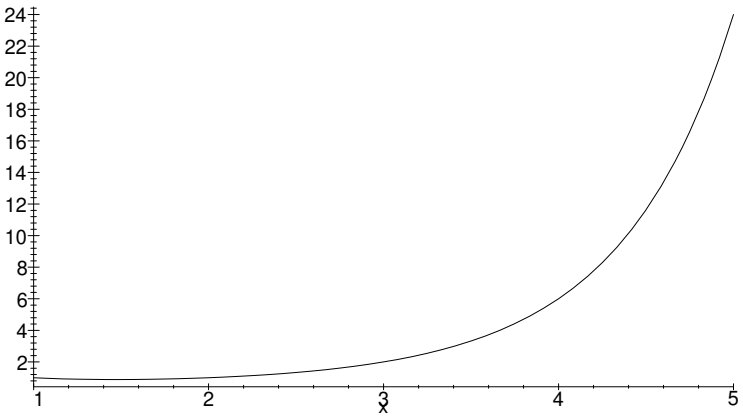


Abb. 66: Die Γ-Funktion

Aus dem elementar berechenbaren Wert

$$\Gamma(1) = \int_0^\infty e^{-t}\, dt = -e^{-t}\,\Big|_0^\infty = 1$$

ergibt sich somit, daß für alle natürliche Zahlen n gilt $\Gamma(n) = (n-1)!$; die Γ-Funktion ist also eine Art stetig gemachter Fakultätsfunktion. GAUSS definierte auf andere Weise eine stetige Funktion $\Pi(x)$, für die $\Pi(n) = n!$ ist, aber wie sich bald herausstellte, ist $\Pi(x) = \Gamma(x+1)$, so daß nur eine der beiden Funktionen wirklich gebraucht wird. Nach einigen Modewechseln im letzten Jahrhundert entscheidet man sich heute meist

für $\Gamma(x)$: Diese Funktionswerte sind in Tafelwerken tabelliert, und numerische Verfahren für ihre Berechnung stehen in den einschlägigen Unterprogrammbibliotheken und Computeralgebrasystemen zur Verfügung.

§4: Kurvenintegrale im $\mathbb{R}^n$

Bislang haben wir nur Funktionen einer Veränderlichen über ein Intervall integriert. Als ersten Schritt ins Mehrdimensionale wollen wir nun Funktionen mehrerer Veränderlicher betrachten und diese entlang einer Kurve im $\mathbb{R}^n$ integrieren. Entsprechende Integrale benötigt man einerseits, um Bogenlängen von Kurven zu berechnen, vor allem aber sind sie wichtig, um die aufzuwendende oder freiwerdende Energie bei der Bewegung eines Objekts in einem Kraftfeld (oder einem elektrisch geladenen Teilchens in einem elektrischen Feld *usw.*) zu berechnen. Wir beginnen mit der Definition von Kurven:

a) Kurven und Tangentenvektoren

Definition: *a)* Ein *Kurvenstück* γ ist eine stetig differenzierbare Abbildung

$$\gamma\colon [a,b] \to \mathbb{R}^n$$

eines Intervalls in den $\mathbb{R}^n$.
b) Eine *Kurve* γ ist eine endliche Folge von Kurvenstücken

$$\gamma_1, \ldots, \gamma_r\colon [a_i, b_i] \to \mathbb{R}^n$$

von Kurvenstücken mit der Eigenschaft, daß

$$\gamma_i(b_i) = \gamma_{i+1}(a_{i+1}) \quad \text{für } i = 1, \ldots, n-1.$$

c) Eine Kurve γ heißt geschlossen, wenn $\gamma_r(b_r) = \gamma_1(a_1)$ ist.

Die Bedingung im Teil *b)* der obigen Definition stellt sicher, daß Kurven, anschaulich betrachtet, zusammenhängend sind; es hat allein praktische Gründe, daß man nicht auch von den Intervallen verlangt, daß sie unmittelbar aneinander anschließen: Oft werden die Formeln für ein Kurvenstück einfacher, wenn man einen bestimmten Anfangswert wie etwa die Null für das Intervall nehmen kann.

Ein wesentlicher Unterschied zwischen einer Kurve und einem Kurvenstück liegt in der Differenzierbarkeit: Durch Verschiebung der Parameterintervalle könnte man jede Kurve durch eine stetige Abbildung $\gamma\colon [a,b] \to \mathbb{R}^n$ beschreiben, aber dort, wo zwei Kurvenstücke aneinanderstoßen, muß diese Abbildung nicht differenzierbar sein; anschaulich gesprochen kann die Kurve dort einen „Knick“ haben.

Die Ableitung $\dot\gamma(t)$ ist ein Vektor, dessen Komponenten die Ableitungen der Koordinatenfunktionen von $\gamma(t)$ sind; wir bezeichnen ihn als Tangentenvektor der Kurve im Punkt $\gamma(t)$. Gelegentlich wird es wichtig sein, daß dieser Vektor ungleich dem Nullvektor ist; wir definieren daher

Definition: *a)* Ein Kurvenstück $\gamma\colon [a,b] \to \mathbb{R}^n$ heißt *regulär,* wenn $\dot\gamma(t) \neq 0$ für alle $t \in (a,b)$.
b) Eine Kurve γ heißt *stückweise regulär,* wenn sie aus regulären Kurvenstücken zusammengesetzt werden kann.

Man beachte, daß für die Intervallendpunkte nichts gefordert wird: Selbst wenn die Ableitung dort existiert, muß sie nicht von Null verschieden kann. Auf diese Weise läßt sich ein Kurvenstück, bei dem $\dot\gamma(t)$ an endlich vielen Stellen gleich dem Nullvektor ist, immer noch als stückweise reguläre Kurve auffassen.

Anschaulich kann man sich ein Kurvenstück durch sein Bild im $\mathbb{R}^n$ vorstellen, allerdings muß man beachten, daß dasselbe Bild durch ganz verschiedene Funktionen parametrisiert werden kann. Als Beispiel für verschiedene Parametrisierungen einer und derselben Kurve betrachten wir den Einheitskreis $x^2 + y^2 = 1$. Seine bekannteste Darstellung als Kurvenstück ist die Parametrisierung

$$\gamma_1\colon [0,2\pi] \to \mathbb{R}^2; \quad t \mapsto (\cos t, \sin t)\,,$$

aber natürlich wäre auch

$$\gamma_2\colon [0,1] \to \mathbb{R}^2; \quad t \mapsto (\cos 2\pi t, \sin 2\pi t)$$

eine Möglichkeit. Es geht aber auch ganz anders, denn auch bei

$$\gamma_3\colon \mathbb{R} \to \mathbb{R}^2; \quad t \mapsto \left(\frac{t^2-1}{t^2+1}, \frac{2t}{t^2+1}\right)$$

liegen sämtliche Bildpunkte auf dem Einheitskreis, und zumindest für jedes endliche Teilintervall von $\mathbb{R}$ definiert auch γ_3 ein Kurvenstück. Wie eine kurze Kurvendiskussion (oder ein Blick auf Abbildung 67) zeigt, besteht das Bild von γ_3 aus allen Punkten des Einheitskreises außer $(1,0)$; letzterer Punkt ist der Grenzwert von $\gamma_3(t)$ für $t \to \pm\infty$.

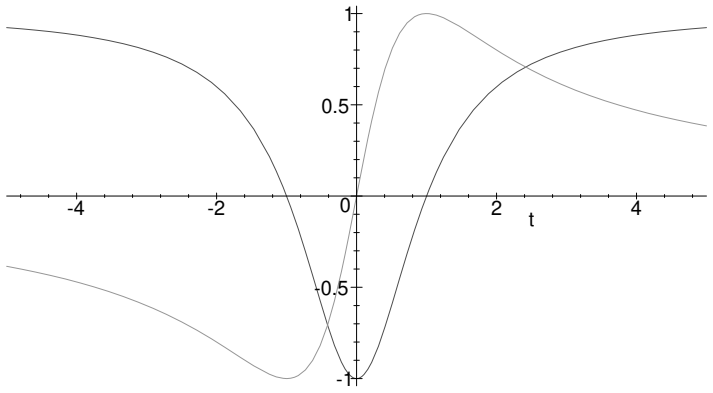


Abb. 67: Graph der Funktionen $x(t) = \frac{t^2-1}{t^2+1}$ und $y(t) = \frac{2t}{t^2+1}$

Jede dieser Parametrisierungen führt zu anderen Tangentenvektoren: Für einen festen Kurvenpunkt liegen zwar alle drei Vektoren auf ein und derselben Geraden, der Tangenten an den Kreis, aber sie haben verschiedene Länge: Für $(x,y) = \gamma_1(t)$ ist

$$\dot\gamma_1(t) = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix},$$

für $(x,y) = \gamma_2(t)$ ist

$$\dot\gamma_2(t) = \begin{pmatrix} -2\pi \sin 2\pi t \\ 2\pi \cos 2\pi t \end{pmatrix} = \begin{pmatrix} -2\pi y \\ 2\pi x \end{pmatrix},$$

und für $(x,y) = \gamma_3(t)$ schließlich zeigt eine kurze Rechnung, daß

$$\dot\gamma_3(t) = \frac{2}{(t^2+1)^2} \cdot \begin{pmatrix} 2t \\ 1-t^2 \end{pmatrix} = \frac{2}{t^2+1} \begin{pmatrix} y \\ -x \end{pmatrix}$$

ist. Bei der Parametrisierung mit γ_2 sind die Tangentenvektoren also jeweils 2π mal so lang wie bei der mit γ_1, was man anschaulich so interpretieren kann, daß der Kreis bei dieser Parametrisierung 2π mal so schnell durchlaufen wird wie bei der mit γ_1. Bei der Parametrisierung mit γ_3 ist die Länge der Vektoren variabel, und sie zeigen auch in die Gegenrichtung zu den Tangentenvektoren an γ_1 und γ_2; bei dieser Parametrisierung wird der Kreis also im Uhrzeigersinn durchlaufen.

b) Die Bogenlänge einer Kurve

In der Integralrechnung wird die Fläche unterhalb einer Kurve dadurch definiert, daß man sie durch Rechtecke angenähert; falls deren Gesamtfläche bei immer feiner werdenden Unterteilungen gegen einen festen Grenzwert konvergiert, bezeichnetet man diesen als Fläche unterhalb der Kurve oder auch als RIEMANN-Integral über die die Kurve beschreibende Funktion.

Nichts spricht dagegen, bei der Definition der Bogenlänge genauso vorzugehen: Die primitiven Bausteine, mit denen wir die Kurve annähern, sind nun natürlich keine Rechtecke mehr, sondern Strecken; am einfachsten nehmen wir dazu Tangentenvektoren. Dazu brauchen wir allerdings Differenzierbarkeit, wir müssen uns also zunächst auf ein Kurven*stück* beschränken.

Sei also $\gamma\colon [a,b] \to \mathbb{R}^n$ ein Kurvenstück; wir unterteilen das Intervall, wie wir es vom RIEMANN-Integral her gewohnt sind, durch Zwischenpunkte

$$a = t_0 < t_1 < \cdots < t_{N-1} < t_N = b$$

und wählen in jedem Teilintervall (t_i, t_{i+1}) einen Zwischenpunkt τ_i. Der Tangentenvektor im Punkt τ_i ist nach Definition $\dot\gamma(\tau_i)$, allerdings ist das nicht unbedingt der Vektor, den wir wollen: Wir wollen schließlich die Kurve durch ihre Tangente annähern, und dazu müssen wir die Länge an das Intervall anpassen, über dem wir die Kurve approximieren wollen. Indem wir den Vektor $\dot\gamma(\tau_i)$ der Differentialquotienten durch

$$\frac{1}{t_{i+1}-t_i} \cdot \overrightarrow{\gamma(t_i)\gamma(t_{i+1})} = \frac{1}{t_{i+1}-t_i} \begin{pmatrix} \gamma_1(t_{i+1}) - \gamma_1(t_i) \\ \vdots \\ \gamma_n(t_{i+1}) - \gamma_n(t_i) \end{pmatrix}$$

approximieren sehen wir, daß der Vektor $\overrightarrow{\gamma(t_i)\gamma(t_{i+1})}$ ungefähr gleich

$$\dot\gamma(\tau_i) \cdot (t_{i+1} - t_i)$$

ist, die Bogenlänge der Kurve kann also angenähert werden durch

$$\sum_{i=0}^{N-1} \left|\dot\gamma(\tau_i)\right| \cdot (t_{i+1} - t_i)\,.$$

Für die Kreislinie mit ihrer Parametrisierung $t \mapsto (\cos t, \sin t)$ ist dies in den Abbildungen 68 und 69 dargestellt für eine äquidistante Unterteilung des Intervalls $[0, 2\pi]$ in zwanzig *bzw.* fünfzig Teilintervalle; im letzteren Fall ist in der Tat kaum mehr ein Unterschied zu sehen zwischen der Kreislinie und den fünfzig Vektoren, durch die sie approximiert wird.

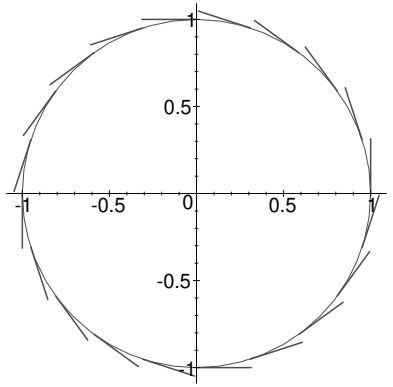


Abb. 68: Bogenlänge des Einheitskreises angenähert durch zwanzig Strecken

Da γ nach Definition eines Kurvenstücks stetig differenzierbar ist, ist auch $\left|\dot\gamma(t)\right|$ immerhin noch eine stetige und damit insbesondere RIEMANN-integrierbare Funktion; wir wissen also, daß der Grenzwert für immer enger werdende Verfeinerungen der Unterteilung existiert und gleich dem Integral über diese Funktion ist. Somit kommen wir auf ganz natürliche Weise auf die

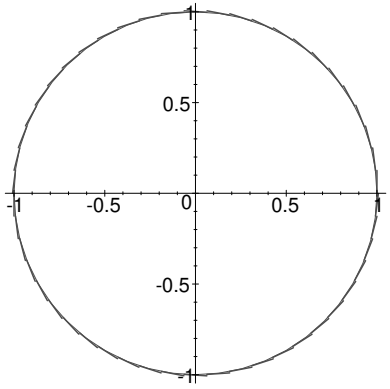


Abb. 69: Bogenlänge des Einheitskreises angenähert durch fünfzig Strecken

Definition: *a)* Die Bogenlänge eines Kurvenstücks $\gamma\colon [a,b] \to \mathbb{R}^n$ ist

$$\int_a^b \left|\dot\gamma(t)\right| dt\,.$$

b) Die Bogenlänge einer Kurve γ, bestehend aus den Kurvenstücken $\gamma_1, \ldots, \gamma_r$ ist die Summe der Bogenlängen der Kurvenstücke γ_i.

Um zu sehen, ob das alles wirklich vernünftig ist, berechnen wir die Bogenlänge der Kreislinie

$$\gamma_1\colon [0, 2\pi] \to \mathbb{R}^2; \quad t \mapsto (\cos t, \sin t)\,.$$

Hier ist

$$\dot\gamma_1(t) = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix}, \quad \text{also} \quad \left|\dot\gamma_1(t)\right| = \sqrt{\sin^2 t + \cos^2 t} = 1\,,$$

und damit ist die Bogenlänge

$$\int_0^{2\pi} \left|\dot\gamma_1(t)\right| dt = \int_0^{2\pi} dt = 2\pi\,,$$

wie erwartet; zumindest für diese Parametrisierung ist also alles vernünftig.

Alternativ hatten wir die Kreislinie auch parametrisiert durch

$$\gamma_3\colon \mathbb{R} \to \mathbb{R}^2; \quad t \mapsto \left(\frac{t^2-1}{t^2+1}, \frac{2t}{t^2+1}\right)\,;$$

hier ist die Bogenlänge im Sinne obiger Definition nicht erklärt, jedoch können wir natürlich γ_3 auf endliche Intervalle einschränken und diese immer größer werden lassen; falls ein Grenzwert existiert, bekommen wir also hier die Bogenlänge als uneigentliches Integral

$$\int_{-\infty}^{\infty} \left|\dot\gamma_3(t)\right| dt\,.$$

Berechnen wir zunächst die Ableitung von γ_3:

$$\frac{d}{dt}\left(\frac{t^2-1}{t^2+1}\right) = \frac{(t^2+1)\cdot 2t - (t^2-1)\cdot 2t}{(t^2+1)^2} = \frac{4t}{(t^2+1)^2}$$

und

$$\frac{d}{dt}\left(\frac{2t}{t^2+1}\right) = \frac{(t^2+1)\cdot 2 - 2t\cdot 2t}{(t^2+1)^2} = \frac{-2(t^2-1)}{(t^2+1)^2}\,,$$

d.h.

$$\left|\dot\gamma_3(t)\right| = \frac{\sqrt{(4t)^2 + 4(t^2-1)^2}}{(t^2+1)^2} = \frac{2\sqrt{(t^2+1)^2}}{(t^2+1)^2} = \frac{2}{t^2+1}\,.$$

Wie Abbildung 70 zeigt, ist diese Annäherung der Kreislinie durch Strecken erheblich unregelmäßiger als die obige, aber beim Übergang zu immer feiner werdenden Unterteilungen gehen natürlich trotzdem alle Streckenlängen gegen null. (Eingezeichnet sind die Strecken zu einer Unterteilung des Intervalls $[-10, 10]$ in Teilintervalle der Länge $1/2$.)

Die Bogenlänge ist somit

$$\int_{-\infty}^{\infty} \frac{2\,dt}{t^2+1} = 2\arctan t\Big|_{-\infty}^{\infty} = 2\left(\frac{\pi}{2} - \frac{-\pi}{2}\right) = 2\pi\,;$$

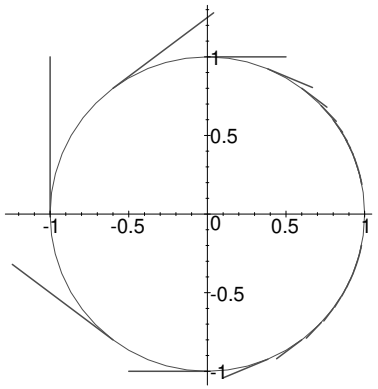


Abb. 70: Approximation durch Strecken in der rationalen Parametrisierung

auch bei dieser Parametrisierung erhalten wir also dasselbe Ergebnis, wie dies aus geometrischen Gründen auch sein muß: Die Bogenlänge ist schließlich eine Eigenschaft einer Kurve, nicht einer speziellen Parametrisierung der Kurve.

Ganz so einfach ist die Sache allerdings nun doch nicht: Schließlich parametrisiert auch

$$\gamma_4\colon [0, 20\pi] \to \mathbb{R}^2; \quad t \mapsto (\cos t, \sin t)$$

die Kreislinie, aber nun ist die Bogenlänge

$$\int_0^{20\pi} \left|\dot\gamma_4(t)\right| dt = \int_0^{20\pi} dt = 20\pi\,.$$

Auch das erscheint geometrisch durchaus sinnvoll: Wenn man nicht nur einmal, sondern zehnmal im Kreise herum geht, legt man schließlich einen zehnmal so langen Weg zurück.

Wir müssen also etwas sorgfältig sein, wenn wir präzisieren wollen, was die Unabhängigkeit der Bogenlänge von der Parametrisierung wirklich bedeuten soll; da das ganze dadurch auch etwas umfangreicher wird, sei das entsprechende Lemma in den nächsten Abschnitt verschoben, wo wir es gleich etwas allgemeiner beweisen werden.

c) Integration eines Vektorfelds längs einer Kurve

Die Hauptanwendung von Kurvenintegralen besteht nicht darin, die Längen aller möglicher Kurven zu berechnen; der Hauptgrund, warum wir solche Integrale betrachten, ist die Berechnung der aufzuwendenden oder freiwerdenden Energie bei der Bewegung eines Teilchens in einem Kraftfeld *bzw.* – im Fall eines elektrisch geladenen Teilchens – eines elektromagnetischen Felds.

Gehen wir der Einfachheit halber aus von einem Kraftfeld $\vec F(\boldsymbol{x})$ und einem Teilchen, das sich entlang eines Kurvenstücks $\gamma\colon [a, b] \to \mathbb{R}^n$ durch dieses Feld bewegt, d.h. der Vektor $\vec F\big(\gamma(t)\big)$ sei für alle t aus $[a, b]$ definiert. Im Punkt $\gamma(t)$ greift dann also der Kraftvektor $\vec F\big(\gamma(t)\big)$ an; die Arbeit, die das Teilchen verrichten muß oder gewinnt, ist das Skalarprodukt aus Kraftvektor und (Tangential-)Vektor des Wegs.

Um die Gesamtarbeit zunächst näherungsweise auszurechen, unterteilen wir wie gewohnt das Intervall $[a, b]$ durch Zwischenpunkte

$$a = t_0 < t_1 < \cdots < t_{N-1} < t_N = b$$

und wählen in jedem Teilintervall (t_i, t_{i+1}) einen Punkt τ_i. Zwischen den Punkten $\gamma(t_i)$ und $\gamma(t_{i+1})$ approximieren wir die Kurve wie im vorigen Abschnitt durch den Tangentenvektor $\dot\gamma(\tau_i) \cdot (t_{i+1} - t_i)$, die Arbeit kann dann angenähert werden durch

$$\sum_{i=0}^{N-1} \vec F\big(\gamma(\tau_i)\big) \cdot \dot\gamma(\tau_i)$$

Da γ nach Definition eines Kurvenstücks stetig differenzierbar ist, ist $\vec F\big(\gamma(t)\big) \cdot \dot\gamma(t)$ für ein stetiges Vektorfeld eine stetige und damit insbesondere RIEMANN-integrierbare Funktion; wir wissen also, daß der Grenzwert für immer enger werdende Verfeinerungen der Unterteilung existiert und gleich dem Integral über diese Funktion ist.

Definition: *a)* $\vec V \in \mathcal{C}^0(D, \mathbb{R}^n)$ sei ein stetiges Vektorfeld auf der offenen Teilmenge $D \subseteq \mathbb{R}^n$, und $\gamma\colon [a, b] \to D$ sei ein Kurvenstück. Das Integral

über $\vec{V}$ entlang γ ist

$$\int_\gamma \vec{V}(\boldsymbol{x})\,d\boldsymbol{x} \underset{\text{def}}{=} \int_a^b \vec{V}\big(\gamma(t)\big)\cdot\dot\gamma(t)\,dt\,.$$

b) Das Integral über $\vec{V}$ entlang einer Kurve γ ist die Summe der Integrale über $\vec{V}$ entlang der Kurvenstücke, aus denen γ zusammengesetzt ist.

Noch allgemeiner können wir ausgehen von einer beliebigen stetigen Funktion $f\colon C \to \mathbb{R}$ auf dem Bild $C = \gamma\big([a,b]\big)$ des Kurvenstücks $\gamma\colon [a,b] \to \mathbb{R}^n$ und diese entlang γ integrieren. Als Motivation könnte man sich etwa vorstellen, daß γ eine materielle Kurve ist und $f\big(\gamma(t)\big)$ die (lineare) Massendichte im Punkt $\gamma(t)$ angibt, oder γ könnte den Weg eines Teilchens durch eine Flüssigkeit mit räumlich variabler Viskosität beschreiben *usw.*

Wenn wir wieder von derselben Unterteilung wie oben ausgehen und das Kurvenstück durch eine Folge von Tangentenvektoren approximieren, müssen wir hier die Summen

$$\sum_{i=0}^{N-1} f\big(\gamma(\tau_i)\big)\,\big|\dot\gamma(\tau_i)\big|$$

betrachten, und diese konvergieren wegen der Stetigkeit von f gegen

$$\int_\gamma f(\boldsymbol{x})\,ds \underset{\text{def}}{=} \int_a^b f\big(\gamma(t)\big)\,\big|\dot\gamma(t)\big|\,dt\,.$$

Analog definieren wir auch das Integral über eine *Kurve* γ.

Definition: $\int_\gamma f(\boldsymbol{x})\,ds$ heißt RIEMANN-STIELTJES-Integral über f entlang γ.

(ds steht hier für das Bogenelement der Kurve, wobei der Buchstabe s wohl vom lateinischen *spatium* = Entfernung kommen dürfte.)

Der niederländische Mathematiker THOMAS JAN STIELTJES (1856–1894) studierte in Leiden, schwänzte dort aber viele Vorlesungen um stattdessen die Werke von GAUSS und JACOBI zu lesen. Obwohl er bei seinen Prüfungen dreimal durchfiel, wurde er Assistent am Observatorium von Leiden, dessen Direktor ein Freund seines Vaters war. 1883 heiratete er und wechselte unter anderem auf Betreiben seiner Frau von der Astronomie zur Mathematik. Nachdem eine Berufung nach Groningen an seinem fehlenden Hochschulabschluß gescheitert war, übersiedelte er 1885 nach Frankreich, wo er Professor an der Universität Toulouse wurde und bis an sein Lebensende blieb. Seine Arbeiten befassen sich vor allem mit Verbindungen zwischen der Zahlentheorie und der Analysis; das RIEMANN-STIELTJES-Integral wurde in einer Arbeit über Kettenbrüche eingeführt.

Dieses RIEMANN-STIELTJES-Integral wird im Spezialfall $f \equiv 1$ zur Bogenlänge, und für ein reguläres Kurvenstück γ ohne Selbstschnitte wird es für

$$f\big(\gamma(t)\big) = \frac{\vec{V}\big(\gamma(t)\big)\cdot\dot\gamma(t)}{\big|\dot\gamma(t)\big|}$$

zum Integral entlang γ über das Vektorfeld $\vec{V}$. (Falls sich die Kurve γ selbst schneidet, wie etwa eine 8, ist $\dot\gamma(t)$ im Schnittpunkt eventuell nicht nur vom Punkt $\gamma(t)$ abhängig, sondern vom Parameter t, so daß wir dann nicht in der Situation von RIEMANN-STIELTJES sind. Meist ist dies jedoch kein Problem, denn wenn so etwas nur endlich oft vorkommt, können wir die Kurve γ in endlich viele Kurvenstücke zerlegen, deren jedes die Voraussetzung erfüllt.)

Natürlich gelten auch für RIEMANN-STIELTJES-Integrale die üblichen, aus der klassischen Integralrechung bekannten Rechenregeln: Da wir Kurvenintegrale als spezielle RIEMANN-Integrale definiert haben, sind dies in der Tat einfach Spezialfälle der dortigen Regeln. Wir haben also, zum Beispiel, wieder die *Monotonieregel*

$$\int_\gamma f(\boldsymbol{x})\,ds \le \int_\gamma g(\boldsymbol{x})\,ds \quad \text{falls } f(\boldsymbol{x}) \le g(\boldsymbol{x}) \text{ für jeden Punkt auf } \gamma,$$

die *Linearität*seigenschaft

$$\int_\gamma (\alpha f + \beta g)\, ds = \alpha \int_\gamma f\, ds + \beta \int_\gamma g\, ds\,,$$

und für *Zusammensetzungen* gilt

$$\int_{\gamma+\delta} f\, ds = \int_\gamma f\, ds + \int_\delta f\, ds\,,$$

wobei $\gamma+\delta$ jene Kurve bezeichne, deren erste Kurvenstücke die Kurve γ definieren, während die restlichen die Bestandteile von δ sind.

Schließlich haben wir auch noch einen

Mittelwertsatz: Für jedes Kurvenstück $\gamma\colon [a,b] \to \mathbb{R}^n$ und jede stetige Funktion f auf $\gamma([a,b])$ gibt es einen Parameterwert $\tau \in [a,b]$, so daß gilt

$$\int_\gamma f\, ds = f(\tau) \cdot \textit{Bogenlänge von } \gamma\,.$$

Beweis: Dies ist einfach der gewöhnliche Mittelwertsatz der Integralrechnung, angewandt auf die Funktion

$$t \mapsto f\big(\gamma(t)\big) \cdot \dot\gamma(t)\,.$$

∎

Es ist hier nicht wesentlich, daß γ als Kurven*stück* vorausgesetzt war; ein einfaches Argument mit dem Zwischenwertsatz zeigt, daß der Mittelwertsatz auch für Kurven gilt. Wesentlich ist dagegen die Stetigkeit von f, denn da der der Mittelwertsatz der Integralrechnung nur für stetige Funktionen gilt, können wir ihn nur anwenden, wenn $f\big(\gamma(t)\big)$ stetig ist. Ein einfaches Gegenbeispiel zum Mittelwertsatz wäre eine Kurve aus zwei gleichlangen Komponenten, auf deren einer f konstant gleich eins ist, während es auf der anderen verschwindet. Nach dem Mittelwertsatz, angewandt auf die beiden Komponenten, folgt, daß der Mittelwert über die gesamte Kurve gleich 1/2 ist, aber dieser Wert wird von f nirgends angenommen.

Als nächstes kommen wir zu der angekündigten Unabhängigkeit des Kurvenintegrals von der Parametrisierung der Kurve. Da diese, wie wir am Beispiel der mehrfach durchlaufenenen Kreislinie gesehen haben, nicht ganz uneingeschränkt gilt, müssen wir zunächst definieren, was wir meinen:

Definition: Zwei Kurvenstücke $\gamma\colon [a,b] \to \mathbb{R}^n$ und $\delta\colon [c,d] \to \mathbb{R}^n$ heißen äquivalent, wenn es eine bijektive stetig differenzierbare Abbildung $\varphi\colon [a,b] \to [c,d]$ gibt mit $\varphi(a) = c$ und $\varphi(b) = d$ derart, daß $\delta \circ \varphi = \gamma$ ist, d.h. für alle $x \in [a,b]$ ist $\delta\big(\varphi(x)\big) = \gamma(x)$.

Dese Definition stellt insbesondere sicher, daß $\gamma\big([a,b]\big)$ und $\delta\big([c,d]\big)$ dieselbe Teilmenge von $\mathbb{R}^n$ sind, so daß die beiden Kurven als Punktmengen übereinstimmen; durch die geforderte Bijektivität von φ ist aber auch sichergestellt, daß die Kurve bei beiden Parametrisierungen gleich oft durchlaufen wird. Daher erwarten wir

Lemma: $\gamma\colon [a,b] \to D \subseteq \mathbb{R}^n$ und $\delta\colon [c,d] \to D \subseteq \mathbb{R}^n$ seien äquivalente Kurvenstücke, und $f\colon D \to \mathbb{R}$ sei eine stetige Funktion. Dann ist

$$\int_\gamma f(\boldsymbol{x})\, ds = \int_\delta f(\boldsymbol{x})\, ds\,.$$

Der *Beweis* ist eine einfache Anwendung der Kettenregel und der Substitutionsregel: Nach Definition der Äquivalenz gibt es eine bijektive stetig differenzierbare Funktion φ mit $\varphi(a) = c$ und $\varphi(b) = d$, so daß $\gamma = \delta \circ \varphi$ und damit $\dot\gamma = (\dot\delta \circ \varphi) \cdot \dot\varphi$ ist. Offensichtlich muß φ monoton wachsen, also ist $\dot\varphi(t)$ nirgends negativ und somit gleich seinem Betrag. Nach der Substitutionsregel ist

$$\int_\gamma f(\boldsymbol{x})\, ds = \int_a^b f\big(\gamma(t)\big)\, |\dot\gamma(t)|\; dt = \int_a^b f\Big(\delta\big(\varphi(t)\big)\Big)\, \Big|\dot\delta\big(\varphi(t)\big)\Big|\, \dot\varphi(t)\, dt$$

$$= \int_c^d f\big(\delta(t)\big)\, \big|\dot\delta(t)\big|\; dt = \int_\delta f(\boldsymbol{x})\, ds\,.$$

∎

d) Zirkulationsfreie und konservative Vektorfelder

Wir haben Kurvenintegrale über Vektorfelder eingeführt, um die Energie zu beschreiben, die zur Bewegung eines Teilchens durch ein Kraftfeld (oder ein elektromagnetisches Feld) aufgewendet werden muß oder dabei frei wird. Das gerade bewiesene Lemma zeigt, daß diese (Gesamt-)Energie nur vom Weg des Teilchens abhängt, nicht aber beispielsweise von seiner Geschwindigkeit. Wie aus der Physik bekannt ist, hängt aber beispielsweise bei reibungsfreier Bewegung eines Teilchens in einem mechanischen Kraftfeld der Weg nicht einmal von der Kurve ab, sondern nur von deren Anfangs- und Endpunkt oder, genauer gesagt, vom der potentiellen Energie des Anfangs- und des Endpunkts. Insbesondere verschwindet also das Integral längs einer jeden geschlossenen Kurve. In diesem Abschnitt wollen wir Vektorfelder mit dieser Eigenschaft genauer untersuchen.

Definition: *a)* Ein Vektorfeld $\vec{V}\colon D \to \mathbb{R}^n$ heißt *zirkulationsfrei,* wenn für jede geschlossene Kurve γ in D gilt:

$$\int_\gamma \vec{V}\, ds = 0\,.$$

b) Das Vektorfeld $\vec{V}$ heißt *konservativ,* wenn es eine differenzierbare Funktion $\varphi\colon D \to \mathbb{R}$ gibt, so daß $\vec{V} = \operatorname{grad}\varphi$ ist. φ heißt *Stammfunktion* von $\vec{V}$ und $-\varphi$ *Potentialfunktion.*

Zum Verständnis des Begriffs *konservativ* betrachten wir ein Beispiel aus der Physik: Ein Teilchen mit konstanter Masse m bewege sich durch das Gravitationsfeld

$$\vec{F} = -\operatorname{grad}\left(G\frac{Mm}{r}\right)$$

eines Himmelskörpers oder, allgemeiner, durch irgendein Potentialfeld $\vec{F} = -\operatorname{grad} U$. Nach dem zweiten NEWTONschen Gesetz gilt dann für die Bahn $\gamma\colon [a,b] \to \mathbb{R}^3$ des Teilchens mit der Zeit t als Parameter

$$\vec{F}\big(\gamma(t)\big) = m\ddot{\gamma}(t)\,,$$

d.h. die im Punkt $\gamma(t)$ wirkende Kraft ist gleich der Masse des Teilchens mal seiner Beschleunigung. Also ist in jedem Punkt $\gamma(t)$

$$m\ddot{\gamma}(t) + \operatorname{grad} U\big(\gamma(t)\big) = 0\,.$$

Um aus dieser Vektorgleichung eine skalare Beziehung abzuleiten, bilden wir das Skalarprodukt mit dem Geschwindigkeitsvektor $\dot{\gamma}(t)$:

$$\begin{aligned} 0 &= m\ddot{\gamma}(t)\cdot\dot{\gamma}(t) + \operatorname{grad} U\big(\gamma(t)\big)\cdot\dot{\gamma}(t) \\ &= \frac{d}{dt}\left(\frac{m}{2}\dot{\gamma}(t)\cdot\dot{\gamma}(t) + U\big(\gamma(t)\big)\right) \end{aligned}$$

nach Produkt- und Kettenregel. Also ist

$$E = \frac{m\big|\dot{\gamma}(t)\big|^2}{2} + U\big(\gamma(t)\big)$$

konstant, und das ist der klassische Energieerhaltungssatz: Die Summe aus kinetischer und potentieller Energie ist eine zeitlich unveränderliche Erhaltungsgröße.

Der Name „konservativ“ kommt vom lateinischen *conservare* = erhalten; der Grund für das negative Vorzeichen der Potentialfunktion liegt darin, daß der Gradient in Richtung des stärksten *Anstiegs* einer Funktion zeigt, wohingegen die Natur versucht, ein System zum Energie*minimum* zu bringen, so daß die Kräfte in Gegenrichtung zum Gradienten wirken.

Wie zu erwarten, sind die beiden Begriffe *konservativ* und *zirkulationsfrei* nicht unabhängig voneinander; tatsächlich werden wir gleich sehen, daß sie sogar äquivalent sind. Aus technischen Gründen wollen uns uns dabei, auch wenn es nicht unbedingt nötig wäre, auf sogenannte *zusammenhängende* Mengen beschränken:

Definition: Eine Teilmenge $D \subseteq \mathbb{R}^n$ heißt *zusammenhängend,* wenn es zu je zwei Punkten $\boldsymbol{x}, \boldsymbol{y} \in D$ eine Kurve γ gibt mit Anfangspunkt $\boldsymbol{x}$ und Endpunkt $\boldsymbol{y}$.

Diese Definition fordert zwar genau das, was wir gleich brauchen werden, sie ist aber nicht die übliche Definition einer zusammenhängenden Menge: In der Analysis wie auch in der Topologie bezeichnet man eine Menge D dann als zusammenhängend, wenn sie nicht als disjunkte Vereinigung zweier offener Teilmengen von D geschrieben werden kann; dies

ist im allgemeinen eine schwächere Bedingung als die hier geforderte. Für eine offene Teilmenge $D \subseteq \mathbb{R}^n$ sind die beiden Definitionen aber äquivalent:

Sei nämlich D_0 die Menge aller Punkte aus der offenen Menge D, die mit dem Punkt $\boldsymbol{x}_0 \in D$ durch einen Weg verbunden werden können. Dann ist D_0 offen, denn für $\boldsymbol{x}_1 \in D_0$ gibt es wegen der Offenheit von D eine ε-Umgebung von $\boldsymbol{x}_1$, die ganz in D liegt. Da jeder Punkt dieser ε-Umgebung durch eine Strecke mit dem Mittelpunkt $\boldsymbol{x}_1$ verbunden werden kann, liegt auch diese Umgebung in D_0, d.h. D_0 ist offen.

Aber auch das Komplement von D_0 in D ist offen, denn auch für einen Punkt $\boldsymbol{x}_2 \in D \setminus D_0$ enthält D eine ε-Umgebung. Läge diese nicht ganz in $D \setminus D_0$, gäbe es dort einen Punkt $\boldsymbol{x}_3$, der durch einen geeigneten Weg mit $\boldsymbol{x}_0$ verbunden werden könnte. Da aber $\boldsymbol{x}_3$ durch eine Strecke mit $\boldsymbol{x}_2$ verbunden werden kann, gäbe dann auch einen Weg von $\boldsymbol{x}_2$ nach $\boldsymbol{x}_0$, im Widerspruch zur Annahme. Also ist $D \setminus D_0$ offen.

Damit ist D also Vereinigung der offenen und disjunkten Teilmengen D_0 und $D \setminus D_0$; wenn D zusammenhängend ist, geht das nur, wenn eine der beiden Mengen leer ist. D_0 enthält den Punkt $\boldsymbol{x}_0$, also ist $D \setminus D_0 = \emptyset$ und somit $D = D_0$. Also kann jeder Punkt aus D durch einen Weg mit $\boldsymbol{x}_0$ verbunden werden, und damit ist D zusammenhängend im Sinne der obigen Definition.

Der folgdende Satz zeigt die Äquivalenz der Begriffe *konservativ* und *zirkulationsfrei;* wegen seiner zweiten Aussage kann er als das mehrdimensionale Analogon des Hauptsatzes der Differential- und Integralrechung aufgefaßt werden.

Satz: Für ein Vektorfeld $\vec{V}\colon D \to \mathbb{R}^n$ auf einer offenen zusammenhängenden Teilmenge $D \subseteq \mathbb{R}^n$ sind die folgenden Aussagen äquivalent:

1.) $\vec{V}$ ist konservativ.

2.) Es gibt eine Funktion $\varphi\colon D \to \mathbb{R}$, so daß für jede Kurve γ in D gilt:

$$\int_\gamma \vec{V} \cdot ds = \varphi\big(\gamma(b)\big) - \varphi\big(\gamma(a)\big) ;$$

insbesondere hängt das Integral also nur von den Endpunkten von γ ab.

3.) $\vec{V}$ ist zirkulationsfrei.

Beweis: 1.) $\Rightarrow$ 2.): Falls $\vec{V}$ konservativ ist, gibt es eine Stammfunktion φ auf D, so daß $\vec{V} = \operatorname{grad}\varphi$ ist. Für eine Kurve $\gamma\colon [a,b] \to D$ ist daher

nach der Kettenregel

$$\int_\gamma \vec{V} \cdot ds = \int_a^b \vec{V}\big(\gamma(t)\big) \cdot \dot\gamma(t)\, dt = \int_a^b \operatorname{grad}\varphi\big(\gamma(t)\big) \cdot \dot\gamma(t)\, dt$$

$$= \int_a^b \frac{d}{dt}\varphi\big(\gamma(t)\big)\, dt = \varphi\big(\gamma(b)\big) - \varphi\big(\gamma(a)\big) .$$

Damit ist das Integral, genau wie wir es vom Eindimensionalen gewohnt sind, einfach gleich der Differenz der Werte der Stammfunktion am Endpunkt und am Anfangspunkt der Kurve.

2.) $\Rightarrow$ 3.): Diese Implikation ist klar, denn wenn der Wert des Integrals nicht von der Kurve abhängt, ist das Integral längs einer geschlossenen Kurve gleich dem Integral längs der zu einem Punkt degenerierten Kurve, und das ist natürlich gleich Null.

3.) $\Rightarrow$ 2.): Zumindest anschaulich ist auch hier klar, was wir machen: Wir durchlaufen zunächst die Kurve γ in der üblichen Weise und dann die Kurve δ „rückwärts“. Dies ergibt eine geschlossene Kurve $\widetilde{\gamma}$, auf die wir *3.)* anwenden können.

Konkret sei für ein Kurvenstück $\delta_i\colon [c_i, d_i] \to \mathbb{R}^n$ von δ

$$\delta_i^*\colon [c_i, d_i] \to \mathbb{R}^n ; \quad t \mapsto \delta(c_i + d_i - t)$$

das rückwärts durchlaufene Kurvenstück zu δ_i; ist r die Anzahl der δ_i, so bestehe die Kurve δ^* aus den Kurvenstücken $\delta_r^*, \ldots, \delta_1^*$. Da γ und δ dieselben Anfangs- und Endpunkte haben, ist der Endpunkt von δ^* gleich dem Anfangspunkt von γ und umgekehrt; die Folge von Kurvenstücken $\gamma_1, \ldots, \gamma_s, \delta_r^*, \ldots, \delta_1^*$ ist deshalb eine geschlossene Kurve $\widetilde{\gamma}$. Nach Voraussetzung verschwindet das Integral längs einer solchen Kurve, d.h.

$$0 = \int_{\widetilde{\gamma}} \vec{V}\, ds = \int_\gamma \vec{V}\, ds + \int_{\delta^*} \vec{V}\, ds = \int_\gamma \vec{V}\, ds - \int_\delta \vec{V}\, ds ,$$

und damit sind die Integrale über γ und δ gleich.

2) $\Rightarrow$ 1): $\boldsymbol{x}_0$ sei irgendein beliebig ausgewählter Punkt von D. Da D zusammenhängend ist, gibt es dann für jeden Punkt $\boldsymbol{x} \in D$ eine Kurve $\gamma_{\boldsymbol{x}}$ mit Anfangspunkt $\boldsymbol{x}_0$ und Endpunkt $\boldsymbol{x}$.

Wir definieren nun einen Kandidaten für eine Stammfunktion durch

$$\varphi(\boldsymbol{x}) = \int\limits_{\gamma_{\boldsymbol{x}}} \vec{V}\, ds\,;$$

wegen der Voraussetzung *2)* hängt $\varphi(\boldsymbol{x})$ in der Tat nur von $\boldsymbol{x}$ ab und nicht von der Wahl des Wegs $\gamma_{\boldsymbol{x}}$.

Für einen hinreichend kleinen Vektor $\vec{h}$ liegt auch die Verbindungsstrecke von $\boldsymbol{x}$ mit $\boldsymbol{x}+\vec{h}$ in D und kann durch eine Kurve η parametrisiert werden; dann ist

$$\varphi(\boldsymbol{x}+\vec{h}) - \varphi(\boldsymbol{x}) = \int\limits_{\gamma_{\boldsymbol{x}+\vec{h}}} \vec{V}\, ds - \int\limits_{\gamma_{\boldsymbol{x}}} \vec{V}\, ds = \int\limits_{\eta} \vec{V}\, ds\,.$$

Speziell für einen Vektor $\vec{h} = h\vec{e}_i$ der Länge h in Richtung des Einheitsvektors $\vec{e}_i$ der Koordinatenachse x_i ist

$$\int\limits_{\eta} \vec{V}\, ds = \int\limits_{\eta} V_i\, dx_i = \int\limits_0^h V_i(\boldsymbol{x} + t\vec{e}_i)\, dt\,,$$

und damit

$$\lim_{h\to 0} \frac{\varphi(\boldsymbol{x}+h\vec{e}_i) - \varphi(\boldsymbol{x})}{h} = \lim_{h\to 0} \frac{1}{h} \int\limits_0^h V_i(\boldsymbol{x}+t\vec{e}_i)\, dt = V_i(\boldsymbol{x})$$

nach dem Hauptsatz der Differential- und Integralrechnung für Funktionen einer reellen Veränderlichen. Dies zeigt, daß $\operatorname{grad}\varphi = \vec{V}$ ist und das Vektorfeld somit eine Stammfunktion hat. ∎

Im $\mathbb{R}^3$ sollten wir erwarten, daß es noch eine vierte äquivalente Charakterisierung konservativer Vektorfelder gibt: Da die Zirkulationsfreiheit mit Drehungen um eine Achse zusammenhängt, sollte für solche Felder auch die Rotation verschwinden. Diese Richtung ist trivial:

Lemma: Für ein zirkulationsfreies Vektorfeld $\vec{V}\colon D \to \mathbb{R}^3$ auf $D \subseteq \mathbb{R}^3$ ist $\operatorname{rot}\vec{V} \equiv 0$ auf D.

Beweis: Nach dem gerade bewiesenen Satz ist ein zirkulationsfreies Vektorfeld konservativ, und nach den Rechenregeln aus §2*f)* verschwindet die Rotation eines Gradienten. ∎

Die umgekehrte Richtung allerdings ist zumindest in dieser Allgemeinheit falsch. Als **Gegenbeispiel** betrachten wir das Magnetfeld eines geradlinigen stromdurchflossenen Leiters, d.h. also (abgesehen von konstanten Vorfaktoren) das Vektorfeld

$$\vec{V}(x,y,z) = \frac{1}{x^2+y^2} \begin{pmatrix} -y \\ x \\ 0 \end{pmatrix}\,.$$

Wie wir in §2*e2)* gesehen haben, verschwindet seine Rotation im gesamten Definitionsbereich des Vektorfelds, d.h. überall außerhalb der z-Achse.

Trotzdem ist das Vektorfeld nicht zirkulationsfrei, denn für den Einheitskreis

$$\gamma\colon [0,2\pi] \to \mathbb{R}^3;\, t \mapsto (\cos t, \sin t, 0)$$

ist

$$\begin{aligned}\int\limits_{\gamma} \vec{V}\, ds &= \int\limits_0^{2\pi} \vec{V}\big(\gamma(t)\big)\dot{\gamma}(t)\, dt = \int\limits_0^{2\pi} \begin{pmatrix} -\sin t \\ \cos t \\ 0 \end{pmatrix} \begin{pmatrix} -\sin t \\ \cos t \\ 0 \end{pmatrix} dt \\ &= \int\limits_0^{2\pi} dt = 2\pi\,.\end{aligned}$$

Der Grund für dieses Verhalten liegt, wie wir bald sehen werden, darin, daß $\vec{V}$ auf der z-Achse nicht definiert ist: Obwohl die z-Achse im Vergleich zum gesamten $\mathbb{R}^3$ nur einen – sollte man meinen – vernachlässigbar geringen Teil ausmacht, genügt selbst diese minimale Definitionslücke, um die Umkehrung des Lemmas falsch zu machen.

§5: Mehrdimensionale Integrationstheorie

Die bisher betrachteten Kurvenintegrale waren alle zurückführbar auf gewöhnliche RIEMANN-Integrale über Funktionen einer reellen Veränderlichen. Mit ihrer Hilfe ist es möglich, die Länge von Kurven zu bestimmen, nicht aber Flächeninhalte oder Volumen von höherdimensionalen geometrischen Gebilden. Deren Berechnung, sowie die von naheliegenden Verallgemeinerungen wie dem Fluß durch die Oberfläche eines Bereichs sind Gegenstand dieses Paragraphen.

a) Flächeninhalte und Volumina

Beginnen wir mit Flächeninhalten und Volumina. Bereits das gewöhnliche RIEMANN-Integral kann als Flächeninhalt interpretiert werden, allerdings nur für die Fläche zwischen der x-Achse und einer Kurve zwischen zwei gegebenen x-Koordinaten. Hier soll es nun um beliebige Flächen im $\mathbb{R}^2$, Volumina im $\mathbb{R}^3$ und die entsprechenden höherdimensionalen Konzepte gehen.

Bei der Definition des RIEMANN-Integrals wird die Fläche unterhalb der Kurve durch Rechtecke angenähert, deren Kantenlänge in x-Richtung immer kleiner wird, während die Kantenlänge in y-Richtung durch die y-Koordinaten der Kurve $y = f(x)$ gegeben war. Im allgemeinen Fall, wo es keine ausgezeichnete Richtung mehr gibt, wird diese Unterscheidung zwischen x- und y-Richtung offensichtlich sinnlos; die einzig mögliche Verallgemeinerung des RIEMANN-Integrals besteht darin, daß man eine Fläche durch *beliebige* Rechtecke annähert und beim Grenzübergang *beide* Seiten gegen Null gehen läßt.

Ganz entsprechend müssen für dreidimensionale Bereiche Quader betrachtet werden, deren drei Seiten mit Verfeinerung der Überdeckung immer kleiner werden, *usw.*

Formal gehen wir dazu wie folgt vor: Wir wählen ein festes Koordinatensystem im $\mathbb{R}^n$ und betrachten Quader $Q_i \subset \mathbb{R}^n$, deren Kanten parallel zu denn Koordinatenachsen sind. (Für $n = 2$ sind diese „Quader“ natürlich Rechtecke, und für $n = 1$ Intervalle.) Das Volumen eines solchen Quaders soll gleich dem Produkt seiner Kantenlängen sein, genau wie wir es aus der Elementargeometrie gewohnt sind.

Definition: Eine *Elementarmenge* in $\mathbb{R}^n$ ist eine Teilmenge $E \subset \mathbb{R}^n$, die als Vereinigung endlich vieler Quader mit achsenparallelen Kanten geschrieben werden kann; dabei dürfen sich zwei Quader höchstens in gemeinsamen Randpunkten schneiden.

Im $\mathbb{R}^2$ wäre also beispielsweise jede Menge, deren Rand auf kariertem Papier so gezeichnet werden kann, daß alle Linien auf Karokanten liegen, eine Elementarmenge.

Das Volumen $\mu(E)$ einer Elementarmenge E definieren wir als die Summe der Volumina der endlich vielen Quader, aus denen die Menge besteht; man überlegt sich leicht, daß es unabhängig von der Art der Zerlegung der Menge in Quader ist.

Nun gehen wir im wesentlichen genauso vor, wie bei der Definition des RIEMANN-Integrals: Dort hatten wir die Fläche unterhalb einer Kurve $y = f(x)$ sowohl von oben als auch von unten durch Rechtecke angenähert; die Flächen der entsprechenden Elementarmengen hatten wir als RIEMANNsche Unter- *bzw.* Obersummen bezeichnet. Das Integral existierte nach Definition genau dann, wenn bei immer weiterer Verfeinerung der Überdeckung die Untersummen und die Obersummen gegen einen gemeinsamen Grenzwert konvergierten.

Entsprechend betrachten wir zur Definition des Volumens einer Teilmenge $B \subset \mathbb{R}^n$ Elementarmengen, die ganz in B enthalten sind und bezeichnen den Grenzwert bei immer feiner werdenden Quader als *unteres Volumen* $\underline{\mu}(B)$. Zur exakten Definition verwenden wir besser nicht einen Grenzwert, da es etwas umständlich wäre, hier zu definieren, über was genau wir den Grenzwert bilden, sondern wir definieren $\underline{\mu}(B)$ einfach als *Supremum* der Volumina aller Elementarmengen, die in B liegen:

$$\underline{\mu}(B) \underset{\text{def}}{=} \sup\left\{ \mu(E) \;\middle|\; E \subseteq B \text{ Elementarmenge} \right\},$$

falls dieses Supremum existiert. (Es existiert offensichtlich genau dann, wenn die Menge B beschränkt ist; für unbeschränkte Mengen wie etwa den gesamten $\mathbb{R}^n$ ist die Menge aller $\mu(E)$ unbeschränkt, so daß kein Supremum existiert.)

Genauso definieren wir ein oberes Volumen $\overline{\mu}(B)$ als *Infimum* der Volumina aller Elementarmengen, die B enthalten:

$$\overline{\mu}(B) \underset{\text{def}}{=} \inf\left\{ \mu(E) \;\middle|\; E \supseteq B \text{ Elementarmenge} \right\},$$

falls dieses Infimum existiert. (Auch hier ist die Existenz wieder an die Beschränktheit von B gekoppelt, denn für ein unbeschränkte Menge B gibt es keine Elementarmenge, die B enthält, so daß wir das Infimum über die leere Menge bilden müßten.)

Die Abbildungen 71 und 72 zeigen die Annäherung einer Kreisfläche durch Elementarmengen von innen und außen; Abbildung 73 zeigt entsprechende Elementarmengen für eine Halbkugel.

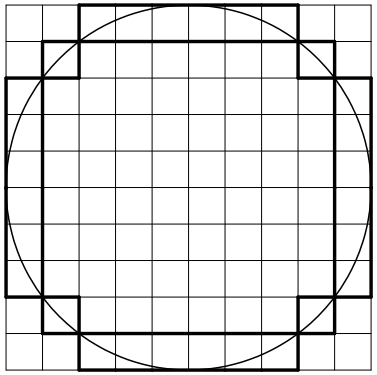


Abb. 71: Approximation einer Kreisfläche auf 10×10 Quadraten

Die Definition des Volumens ist nun fast selbstverständlich:

Definition: Wir sagen, die Menge $B \subset \mathbb{R}^n$ habe das Volumen $\mu(B)$, falls das untere Volumen $\underline{\mu}(B)$ und das obere Volumen $\overline{\mu}(B)$ beide existieren und gleich $\mu(B)$ sind.

Ganz entsprechend können wir auch Integrale über Funktionen definieren: $f\colon D \to \mathbb{R}$ sei eine Funktion auf der Teilmenge $D \subseteq \mathbb{R}^n$, und

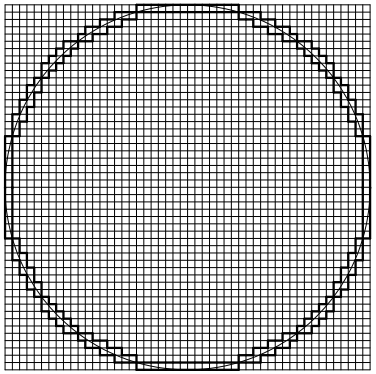


Abb. 72: Approximation einer Kreisfläche auf 50×50 Quadraten

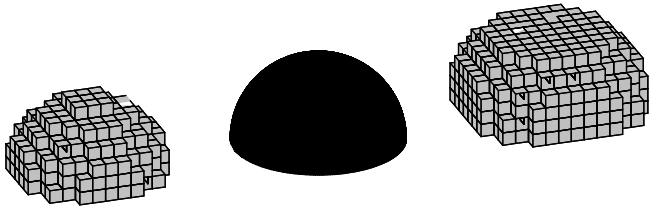


Abb. 73: Approximation einer Halbkugel durch Würfel

$B \subseteq D$ sei eine Teilmenge von D. Dann können wir wieder RIEMANNsche Untersummen definieren, indem wir eine Elementarmenge $E \subseteq D$ betrachten, bestehend etwa aus den Quadern $Q_1, \ldots, Q_N$, und dazu die RIEMANNsche Untersumme definieren als

$$\sum_{i=1}^{N} \mu(Q_i) \cdot \inf\left\{ f(\boldsymbol{x}) \;\middle|\; \boldsymbol{x} \in Q_i \right\}.$$

Die RIEMANNsche Obersumme für die Elementarmenge $E' \supseteq B$ beste-

hend aus den Quadern $Q'_1, \ldots, Q'_M$ ist entsprechend

$$\sum_{j=1}^{M} \mu(Q'_j) \cdot \sup\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in Q_j\},$$

und wir definieren

Definition: $\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \ldots dx_n$ existiert, wenn das Supremum der Menge aller RIEMANNscher Untersummen gleich dem Infimum der Menge aller RIEMANNscher Obersummen ist; dieser gemeinsame Wert ist der Wert des Integrals.

Die Schreibweise $\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \ldots dx_n$ soll dabei bedeuten, daß wir n Integralzeichen schreiben, für $B \subset \mathbb{R}^2$ also

$$\iint_B f(x, y)\, dx\, dy$$

und für $B \subset \mathbb{R}^3$

$$\iiint_B f(x, y, z)\, dx\, dy\, dz\,.$$

(Nicht alle Lehrbücher verwenden bei der mehrdimensionalen Integration mehrere Integralzeichen; einige, wie etwa **[D]**, schreiben unabhängig von der Dimension immer nur ein Integralzeichen. In Physik und Technik scheint die Schreibweise mit mehreren Integralzeichen üblicher zu sein; deshalb soll auch hier diese Konvention verwendet werden.)

Integrale und Volumina sind natürlich eng miteinander verwandt: Einerseits läßt sich das Volumen eines Bereichs B auch als

$$\mu(B) = \int_B \cdots \int 1\, dx_1 \cdots dx_n$$

schreiben, andererseits ist, völlig analog zur Flächeninterpretation des RIEMANN-Integrals, für eine auf B nichtnegative Funktion f

$$\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n =$$

$$\mu\left(\left\{(x_1, \ldots, x_n, y) \in \mathbb{R}^{n+1} \;\middle|\; \begin{array}{l}(x_1, \ldots x_n) \in B \text{ und} \\ 0 \le y \le f(x_1, \ldots, x_n)\end{array}\right\}\right).$$

Auch für das so definierte mehrdimensionale Integral gelten aus offensichtlichen Gründen die Analoga der aus aus der eindimensionalen Integralrechnung bekannten Rechenregeln wie die *Monotonieregel*

$$\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n \le \int_B \cdots \int g(\boldsymbol{x})\, dx_1 \cdots dx_n\,,$$

falls $f(\boldsymbol{x}) \le g(\boldsymbol{x})$ für alle $\boldsymbol{x} \in B$, und die *Linearität*seigenschaft

$$\int_B \cdots \int \big(\alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})\big)\, dx_1 \cdots dx_n$$
$$= \alpha \int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n + \beta \int_B \cdots \int g(\boldsymbol{x})\, dx_1 \cdots dx_n\,.$$

Außerdem haben wir wieder einen

Mittelwertsatz: Für einen zusammenhängenden abgeschlossenen und beschränkten Bereich $B \subset \mathbb{R}^n$ und eine stetige Funktion $f\colon B \to \mathbb{R}$ gibt es einen Punkt $\boldsymbol{x}_0 \in B$, so daß gilt

$$\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n = f(\boldsymbol{x}_0) \cdot \mu(B)\,.$$

Der *Beweis* geht ganz genauso wie in $\mathbb{R}$: Eine stetige Funktion nimmt auf einem abgeschlossenen Intervall sowohl ihr Maximum als auch ihr Minimum an; genauso zeigt man, daß eine stetige Funktion auf einer beschränkten abgeschlossen Teilmenge $B \subset \mathbb{R}^n$ ihr Maximum und ihr Minimum annimmt. Sei etwa $\boldsymbol{x}_1 \in B$ ein Punkt, in dem f minimal wird, und $\boldsymbol{x}_2$ einer, in dem f maximal wird. Dann ist für jedes $\boldsymbol{x} \in B$

$$f(\boldsymbol{x}_1) \le f(\boldsymbol{x}) \le f(\boldsymbol{x}_2)\,;$$

nach der Monotonieregel ist daher

$$\int_B \cdots \int f(\boldsymbol{x}_1)\, dx_1 \cdots dx_n = f(\boldsymbol{x}_1) \cdot \mu(B)$$

kleiner oder gleich

$$\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n\,,$$

was wiederum kleiner oder gleich

$$\int_B \cdots \int f(\boldsymbol{x}_2)\, dx_1 \cdots dx_n = f(\boldsymbol{x}_2) \cdot \mu(B)$$

ist. Damit gibt es eine Zahl η zwischen $f(\boldsymbol{x}_1)$ und $f(\boldsymbol{x}_2)$, so daß

$$\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n = \eta \cdot \mu(B)$$

ist. Wir müssen noch zeigen, daß dieses η ein Funktionswert von f auf B ist. Dazu verbinden wir $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ durch eine Kurve γ; indem wir nötigenfalls die Parameterintervalle der Kurvenstücke von γ verschieben, können wir annehmen, daß γ durch ein zusammenhängendes Parameterintervall $[a,\, b]$ parametrisiert wird. Wegen der Stetigkeit von f haben wir dann eine stetige Funktion

$$\varphi\colon [a,\, b] \to \mathbb{R}; \quad t \mapsto f\big(\gamma(t)\big)$$

mit $\varphi(a) = f(\boldsymbol{x}_1)$ und $\varphi(b) = f(\boldsymbol{x}_2)$. Nach dem Zwischenwertsatz gibt es dazu ein $\tau \in [a,\, b]$ mit $\varphi(c) = \eta$; mit $\boldsymbol{x}_0 = \gamma(\tau)$ ist also $f(\boldsymbol{x}_0) = \eta$. Damit ist der Satz vollständig bewiesen. ∎

Wie beim gewöhnlichen (eindimensionalen) RIEMANN-Integral ist die Existenz von mehrdimensionalen Integralen normalerweise kein Problem, allerdings kann man natürlich leicht Beispiele konstruieren, für die das Integral nicht existiert. Wir können etwa in Analogie zur DIRICHLETschen Sprungfunktion die Menge

$$B = \big\{(x, y) \in \mathbb{R}^2 \;\big|\; x, y \in \mathbb{Q} \text{ und } 0 \le x, y \le 1\big\}$$

aller Punkte mit rationalen Koordianten im Einheitsquadrat betrachten. Da $\mathbb{Q}$ keine reellen Intervalle enthält, sind alle Rechtecke, die ganz in B enthalten sind, zu Punkten degeneriert. Eine in B enthaltene Elementarmenge besteht also aus endlich vielen Punkten und hat somit den

Flächeninhalt null. Eine Elementarmenge, die ganz B enthält, muß aber das gesamte Einheitsquadrat enthalten, da die rationalen Punkte dort dicht liegen; sie hat also mindestens die Fläche eins. Somit ist eins der obere und null der untere Flächeninhalt von B; der Flächeninhalt von B existiert also nicht.

Die folgende Überlegung liefert ein Kriterium für die Existenz des Volumens einer Teilmenge $B \subset \mathbb{R}^n$: Die Differenzmenge zwischen einer Elementarmenge, die B enthält, und einer Elementarmenge, die in B enthalten ist, kann offenbar wieder als Elementarmenge aufgefaßt werden und enthält den Rand von B. Je besser die äußere und die innere Elementarmenge B annähern, desto weniger unterscheidet sich diese Differenzmenge vom Rand; die Existenz eines Volumens von B ist daher äquivalent dazu, daß das Volumen des Rands von B existiert und verschwindet.

Für eine *unbeschränkte* Menge B können wir bislang weder Volumen noch Integrale definieren, denn unsere Konstruktion ist nur anwendbar, wenn B in einer Elementarmenge enthalten ist. Das ist allerdings nichts neues, denn beim eindimensionalen RIEMANN-Integral tritt genau das gleiche Problem auf und wird dadurch gelöst, daß man *uneigentliche* Integrale einführt: Beispielsweise ist

$$\int_a^\infty f(x)\, dx = \lim_{b\to\infty} \int_a^b f(x)\, dx\,,$$

falls dieser Grenzwert existiert. Genauso definieren wir jetzt

Definition: $B \subseteq \mathbb{R}^n$ sei eine unbeschränkte Menge und die Funktion $f\colon D \to \mathbb{R}$ sei in einer B umfassenden Menge $D \subseteq \mathbb{R}^n$ erklärt. Wir sagen, das *uneigentliche Integral*

$$\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n$$

existiere, wenn für jede Folge

$$B_1 \subset B_2 \subset \cdots \subset B_i \subset \cdots$$

von beschränkten Mengen B_i mit $\bigcup\limits_{i\geq 1} B_i = B$ der Grenzwert

$$\lim_{i\to\infty} \int_{B_i} \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n$$

existiert, und wenn er für jede solche Folge denselben Wert hat. Diesen gemeinsamen Wert bezeichnen wir als den Wert des uneigentlichen Integrals.

Wie auch im Eindimensionalen genügt es hier nicht, nur eine einzige Folge von Mengen B_i zu betrachten; ansonsten könnte man beispielsweise dem uneigentlichen Integral

$$\iint_{\mathbb{R}^2} \sin x \sin y \, dx\, dy$$

je nachdem, ob man für die B_i Quadrate nimmt, in deren Eckpunkten der Cosinus für beide Koordinaten verschwindet, oder Quadrate, in deren Eckpunkten die beiden Cosinus einen konstanten anderen Wert haben, dem Integral die verschiedensten Werte zuordnen. Tatsächlich aber existiert dieses uneigentliche Integral natürlich genauso wenig wie das eindimensionale Integral $\int_{-\infty}^{\infty} \sin x \, dx$.

Für interessantere Beispiele sei auf den nächsten Abschnitt verwiesen, wo wir

$$\iint_{\mathbb{R}^2} e^{-(x^2+y^2)}\, dx\, dy$$

berechnen werden.

b) Integration über Normalbereiche

Der Grund dafür, daß wir bislang noch keine interessanten Beispiele für mehrdimensionale Integrale haben, liegt natürlich darin, daß man Werte von Integralen fast nie durch Anwendung der Definition bestimmt. Im Eindimensionalen wird stattdessen meist den Hauptsatz der Differential- und Integralrechnung benutzt, über den die Berechnung eines Integrals auf die (nicht immer explizit mögliche) Bestimmung einer Stammfunktion zurückgeführt wurde; hier, im Mehrdimensionalen, wollen wir die

Integration soweit wie möglich auf mehrfache eindimensionale Integration zurückführen. Zumindest in einem Fall wissen wir schon, wie das geht: Ist die Funktion f auf dem Intervall $[a, b]$ nichtnegativ, so hat

$$B = \big\{(x, y) \in \mathbb{R}^2 \;\big|\; a \leq x \leq b \text{ und } 0 \leq y \leq f(x)\big\}$$

die Fläche

$$\iint_B dx\, dy = \int_a^b f(x)\, dx\,.$$

Auch den Flächeninhalt der etwas komplizierteren Menge

$$B = \big\{(x, y) \in \mathbb{R}^2 \;\big|\; a \leq x \leq b \text{ und } g(x) \leq y \leq h(x)\big\}$$

läßt sich, wenn $g(x) \leq h(x)$ ist für alle $x \in [a, b]$, durch das eindimensionale Integral

$$\iint_B dx\, dy = \int_a^b \big(h(x) - g(x)\big)\, dx$$

ausdrücken. Ähnliche Formeln gelten auch nach Vertauschung der Rollen von x und y.

Definition: Eine Teilmenge $B \subset \mathbb{R}^2$ heißt *Normalbereich vom Typ I,* wenn es reelle Zahlen $a \leq b$ und stetig differenzierbare Funktionen $g, h\colon [a, b] \to \mathbb{R}$ gibt mit $g(x) \leq h(x)$ für alle $x \in [a, b]$, so daß

$$B = \big\{(x, y) \in \mathbb{R}^2 \;\big|\; a \leq x \leq b \text{ und } g(x) \leq y \leq h(x)\big\}$$

ist. B heißt *Normalbereich vom Typ II,* wenn es reelle Zahlen $c \leq d$ und stetig differenzierbare Funktionen $g, h\colon [c, d] \to \mathbb{R}$ gibt mit $g(y) \leq h(y)$ für alle $y \in [a, b]$, so daß

$$B = \big\{(x, y) \in \mathbb{R}^2 \;\big|\; c \leq y \leq d \text{ und } g(y) \leq x \leq h(y)\big\}\,.$$

Für diese Normalbereiche können wir nicht nur die Fläche leicht ausrechnen, sondern auch beliebige Integrale über stetige Funktionen:

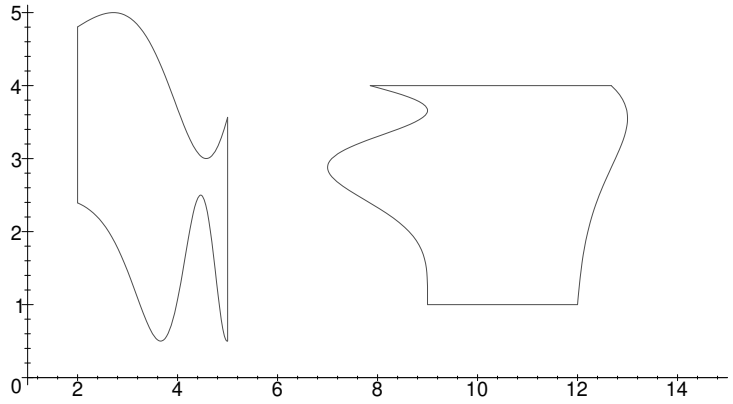


Abb. 74: Normalbereiche vom Typ I und II

Satz: $f\colon D \to \mathbb{R}$ sei eine stetige Funktion auf $D \subseteq \mathbb{R}^n$.
a) Für einen Normalbereich $B \subseteq D$ vom Typ I ist

$$\iint\limits_B f(x,y)\,dx\,dy = \int\limits_a^b \left(\int\limits_{g(x)}^{h(x)} f(x,y)\,dy \right) dx\,.$$

b) Für einen Normalbereich $B \subseteq D$ vom Typ II ist

$$\iint\limits_B f(x,y)\,dx\,dy = \int\limits_c^d \left(\int\limits_{g(y)}^{h(y)} f(x,y)\,dx \right) dy\,.$$

Beweis: Wir wissen bereits, daß der Flächeninhalt von B existiert; da f eine stetige Funktion ist, folgt dann auch ohne größere Schwierigkeiten die Existenz des Integrals von f über B. Wir wollen darauf nicht genauer eingehen, sondern nur benutzen, daß es dann ausreicht, spezielle Folgen von Elementarmengen zu betrachten. Eine solche spezielle Folge erhalten wir etwa dadurch, daß wir den ganzen $\mathbb{R}^2$ mit einem Quadrategitter überziehen; die Seitenlänge der Quadrate sei $k = \frac{b-a}{N}$ und ihr Flächeninhalt dementsprechend gleich k^2.

Da die Aussage *b)* durch Vertauschung der beiden Koordinaten in *a)* übergeht, reicht es, den Satz für Normalbereiche vom Typ I zu beweisen.

Als in B enthaltene Elementarmenge E_N wählen wir die Menge aller Quadrate, die ganz in B liegen; als B enthaltende Elementarmenge E'_N entsprechend die Menge aller Quadrate, die nichtleeren Durchschnitt mit B haben. Wir numerieren die Quadrate in E_N und E'_N mit zwei Indizes: Das Intervall $[a,b]$ wird durch das Quadrategitter in $N-1$ Teilintervalle der Länge k zerlegt; das Quadrat Q_{ij} liege über dem i-ten dieser Teilintervalle und sei, von unter her gesehen, das j-te Quadrat, das ganz in B liegt. Die Anzahl der ganz in B liegenden Quadrate im i-ten Streifen sei r_i; dann ist die RIEMANNsche Untersumme zu E_N gleich

$$\sum_{i=1}^{N-1} \sum_{j=1}^{r_i} \mu(Q_{ij}) \cdot \inf_{(x,y)\in Q_{ij}} f(x,y) = \sum_{i=1}^{N-1} \left(\sum_{j=1}^{r_i} \inf_{(x,y)\in Q_{ij}} f(x,y) \cdot k \right) \cdot k\,.$$

Für die RIEMANNsche Obersumme müssen wir alle Quadrate betrachten, die mit B nichtleeren Durchschnitt haben; dabei treten im allgemeinen auch Quadrate Q_{ij} unterhalb von Q_{i1} auf, für deren Bezeichnung wir Indizes $j \le 0$ verwenden.

Im i-ten Streifen mögen die Q_{ij} mit $s_i \le j \le t_i$ auftreten; dann ist die RIEMANNsche Obersumme gleich

$$\sum_{i=1}^{N-1} \sum_{j=s_i}^{t_i} \mu(Q_{ij}) \cdot \sup_{(x,y)\in Q_{ij}} f(x,y) = \sum_{i=1}^{N-1} \left(\sum_{j=s_i}^{t_i} \sup_{(x,y)\in Q_{ij}} f(x,y) \cdot k \right) \cdot k\,.$$

Da f als stetige Funktion RIEMANN-integrierbar ist, konvergieren für $N \to \infty$ und damit $k \to 0$

$$\sum_{j=1}^{r_i} \inf_{(x,y)\in Q_{ij}} f(x,y) \cdot k \quad \text{und} \quad \sum_{j=s_i}^{t_i} \sup_{(x,y)\in Q_{ij}} f(x,y) \cdot k$$

beide gegen

$$\int\limits_{g(x)}^{h(x)} f(x,y)\,dy\,;$$

die RIEMANNschen Unter- und Obersummen daher entsprechend gegen

$$\int\limits_a^b \left(\int\limits_{g(x)}^{h(x)} f(x,y)\,dy \right) dx\,,$$

wie behauptet. ∎

Als erstes Beispiel berechnen wir zur Kontrolle etwas Altbekanntes, die Fläche der Einheitskreisscheibe

$$B = \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; x^2 + y^2 \le 1\right\},$$

die wir hier umständlich als

$$\iint\limits_B dx\,dy$$

ausrechnen wollen.

B kann sowohl als Normalbereich vom Typ I wie auch als solcher vom Typ II geschrieben werden:

$$\begin{aligned} B &= \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; -1 \le x \le 1 \text{ und } -\sqrt{1-x^2} \le y \le \sqrt{1-x^2}\right\} \\ &= \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; -1 \le y \le 1 \text{ und } -\sqrt{1-y^2} \le x \le \sqrt{1-y^2}\right\}. \end{aligned}$$

In der ersten Darstellung ist nach Teil *a)* des gerade bewiesenen Satzes

$$\iint\limits_B dx\,dy = \int\limits_{-1}^{1} \left(\int\limits_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy \right) dx = \int\limits_{-1}^{1} 2\sqrt{1-x^2}\,dx\,.$$

Zur Berechnung dieses Integrals können wir die *Substitutionsregel* anwenden: Mit $x = \sin t$ ist

$$\int\limits_{-1}^{1} 2\sqrt{1-x^2}\,dx = \int\limits_{-\pi/2}^{\pi/2} 2\sqrt{1-\sin^2 t}\cos t\,dt\,,$$

und das rechtsstehende Integral läßt sich mit partieller Integration ausrechnen:

$$\begin{aligned} \int\limits_{-\pi/2}^{\pi/2} \cos^2 t\,dt &= \sin t \cos t \,\Big|_{-\pi/2}^{\pi/2} - \int\limits_{-\pi/2}^{\pi/2} \sin^2 t\,dt \\ &= 0 + \int\limits_{-\pi/2}^{\pi/2} (1-\cos^2 t)\,dt = \pi - \int\limits_{-\pi/2}^{\pi/2} \cos^2 t\,dt\,. \end{aligned}$$

Wenn wir das Integral ganz rechts auf die linke Seite bringen, erhalten wir den Flächeninhalt

$$\int\limits_{-\pi/2}^{\pi/2} 2\sqrt{1-x^2}\,dx = \int\limits_{-\pi/2}^{\pi/2} 2\cos^2 t\,dt = \pi\,,$$

wie erwartet.

Etwa interessanter ist die Berechnung des (hoffentlich auch bekannten) Volumens der Einheitskugel

$$K = \left\{(x,y,z) \in \mathbb{R}^3 \;\middle|\; x^2 + y^2 + z^2 \le 1\right\}.$$

Hierzu könnten wir eine dreidimensionale Integration durchführen, es geht aber zum Glück auch einfacher: Natürlich genügt es, das Volumen der Halbkugel $z \ge 0$ zu bestimmen, und dies ist die Menge aller Punkte zwischen der Einheitskreisscheibe B der (x,y)-Ebene und dem Graph der Funktion

$$f(x,y) = \sqrt{1-x^2-y^2}$$

über B. Somit ist das Volumen der Halbkugel gleich

$$\iint\limits_B \sqrt{1-x^2-y^2}\,dx\,dy = \int\limits_{-1}^{1} \left(\int\limits_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1-x^2-y^2}\,dy \right) dx\,.$$

Schreibt man mit $a = \sqrt{1-x^2}$

$$\sqrt{1-x^2-y^2} = \sqrt{a^2-y^2} = a\sqrt{1-\left(\frac{y}{a}\right)^2}\,,$$

so ist das innere Integral

$$\int\limits_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1-x^2-y^2}\,dy = a \int\limits_{-a}^{a} \sqrt{1-\left(\frac{y}{a}\right)^2}\,dy\,,$$

was durch die Substitution $y = au$ zu

$$a \int\limits_{-1}^{1} \sqrt{1-u^2}\,a\,du = a^2 \int\limits_{-1}^{1} \sqrt{1-u^2}\,du = \frac{\pi}{2}a^2$$

wird, wie wir oben gerade nachgerechnet haben. Somit ist

$$\int\limits_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1-x^2-y^2}\,dy = \frac{\pi}{2}(1-x^2)$$

und

$$\iint\limits_B \sqrt{1-x^2-y^2}\,dx\,dy = \int\limits_{-1}^{1} \frac{\pi}{2}(1-x^2)\,dx$$

$$= \frac{\pi}{2}\left(x - \frac{x^3}{3}\right)\Bigg|_{-1}^{1} = \frac{\pi}{2}\cdot\frac{4}{3} = \frac{2\pi}{3}\,;$$

das Kugelvolumen als Volumen zweier Halbkugeln ist also in der Tat, wie es sein soll, gleich $4\pi/3$.

In diesen Beispielen hatten wir nur Kreisscheiben als Normalbereiche, aber der Leser kann sich leicht überzeugen selbst davon überzeugen, daß sich die Nützlichkeit dieser Mengen keineswegs auf solche trivialen Beispiele beschränkt: Man zeichne frei irgendeneinen beschränkten zweidimensionalen Bereich auf ein Blatt Papier und überzeuge sich davon, daß sich dieser durch Einfügen von (meist sehr wenigen) waagrechten und senkrechten Strecken als Vereinigung von Normalbereichen der Typen I und II darstellen läßt.

c) Die Transformationsformel

Von den vielen Regeln zur expliziten Bestimmung einer Stammfunktion ist sicherlich die Substitutionsregel die nützlichste; es lohnt sich daher, nach einer Verallgemeinerung dieser Regel für mehrdimensionale Integrale zu suchen.

Die Idee im Eindimensionalen ist bekanntlich, daß wir die Integrationsvariable x als Funktion $x = \varphi(t)$ einer neuen Variablen t schreiben: Mit $a = \varphi(t_0)$ und $b = \varphi(t_1)$ ist

$$\int\limits_a^b f(x)\,dx = \int\limits_{t_0}^{t_1} f\big(\varphi(t)\big)\dot{\varphi}(t)\,dt\,.$$

Genauso können wir auch bei einer Funktion mehrerer Veränderlicher diese als Funktionen neuer Variabler schreiben.

Beginnen wir der Anschaulichkeit halber mit einer Funktion $f(x,y)$ zweier Veränderlicher und schreiben wir diese als Funktionen

$$x = x(u,v) \quad \text{und} \quad y = y(u,v)$$

zweier neuer Variabler u und v. Ein wichtiges Beispiel, das man zur Veranschaulichung während der folgenden Rechnungen im Kopf behalten sollte, ist die Polarkoordinatensdarstellung

$$x = x(r,\varphi) = r\sin\varphi \quad \text{und} \quad y = y(r,\varphi) = r\cos\varphi\,.$$

Zur Definition des Integrals

$$\iint\limits_B f(x,y)\,dx\,dy$$

approximierten wir den Intgegrationsbereich B durch kleine achsenparallele Rechtecke. Wenn wir statt über x und y über u und v integrieren, müssen wir entsprechend den Bereich B', in dem sich diese neuen Variablen bewegen, in achsenparallele Rechtecke zerlegen; dabei fordern wir nun natürlich Parallelität zur u- und zur v-Achse. Wir betrachten ein festes dieser Rechtecke; es habe die Eckpunkte

$$(u_0,v_0), \quad (u_0+h,v_0), \quad (u_0,v_0+k) \quad \text{und} \quad (u_0+h,v_0+k)$$

und somit den Flächeninhalt hk.

Die Menge der Punkte (x,y), die zu den Punkten (u,v) aus diesem Rechteck gehören, d.h. also die Menge

$$\Big\{\big(x(u,v),y(u,v)\big) \;\Big|\; u_0 \le u \le u_0+h \quad \text{und} \quad v_0 \le v \le v_0+k\Big\}$$

ist natürlich im allgemeinen kein Rechteck, sondern eine krummlinig begrenzte Figur; im Beispiel der Polarkoordinaten etwa wäre sie ein Winkelbereich zwischen zwei Kreisbögen. (An den bei Polarkoordinaten etwas problematischen Nullpunkt als Ecke denken wir in diesen Zusammenhang lieber nicht; es ist klar, daß sein Einfluß bei immer

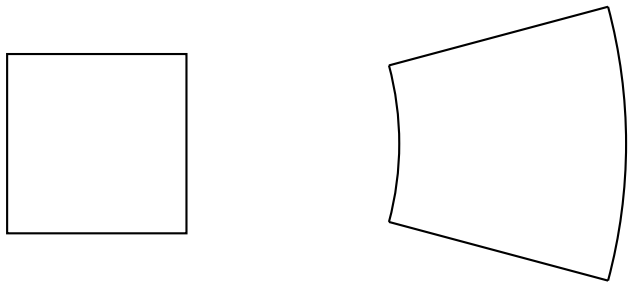


Abb. 75: Rechteck in (r, φ)-Ebene mit Bild in (x, y)-Ebene

kleiner werdenden Rechtecken für eine um den Nullpunkt beschränkte Funktion f immer kleiner wird.)

Trotzdem machen wir, wenn x und y *differenzierbare* Funktionen von u und v sind, bei kleinen Rechtecken keinen allzu großen Fehler, wenn wir die transformierte Menge als *Parallelogramm* betrachten, denn nach Definition der Differenzierbarkeit ist

$$x(u_0 + h, v_0) = x(u_0, v_0) + h\frac{\partial x}{\partial u}(u_0, v_0) + o(h)$$
$$y(u_0 + h, v_0) = y(u_0, v_0) + h\frac{\partial y}{\partial u}(u_0, v_0) + o(h)$$

$$x(u_0, v_0 + k) = x(u_0, v_0) + k\frac{\partial x}{\partial v}(u_0, v_0) + o(k)$$
$$y(u_0, v_0 + k) = y(u_0, v_0) + k\frac{\partial y}{\partial v}(u_0, v_0) + o(k)$$

und

$$\begin{aligned} x(u_0 + h, v_0 + k) = x(u_0, v_0) &+ h\frac{\partial x}{\partial u}(u_0, v_0) + o(h) \\ &+ k\frac{\partial x}{\partial v}(u_0, v_0) + o(k) \\ y(u_0 + h, v_0 + k) = y(u_0, v_0) &+ h\frac{\partial y}{\partial u}(u_0, v_0) + o(h) \\ &+ k\frac{\partial y}{\partial v}(u_0, v_0) + o(k)\,; \end{aligned}$$

wenn wir Terme der Größenordnung $o(h)$ und $o(k)$ vernachlässigen, ist die transformierte Menge also ein Parallelogramm mit Kantenvektoren

$$h\begin{pmatrix} \frac{\partial x}{\partial u}(u_0, v_0) \\ \frac{\partial y}{\partial u}(u_0, v_0) \end{pmatrix} \quad \text{und} \quad k\begin{pmatrix} \frac{\partial x}{\partial v}(u_0, v_0) \\ \frac{\partial y}{\partial v}(u_0, v_0) \end{pmatrix}.$$

Der Flächeninhalt eines Parallelogramms ist bekanntlich gleich dem Produkt der Kantenlängen mal dem Sinus des eingeschlossenen Winkels; falls wir die beiden Vektoren in den $\mathbb{R}^3$ einbetten, indem wir ihnen eine Null als dritte Komponente geben, ist das gerade gleich dem Betrag des Vektorprodukts, das hier nur in der dritten Komponenten von null verschieden ist; die Fläche des Parallelogramms ist also

$$hk \cdot \left| \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \right|.$$

Der gesamte Integrationsbereich wird approximiert durch die Vereinigung der Parallelogramme zu den sämtlichen Rechtecken, mit denen B' approximiert wurde; wenn wir also davon ausgehen, daß die Koordinatentransformation zwischen und B und B' bijektiv ist und wenn wir – was eigentlich noch durch genauere Abschätzungen zu rechtfertigen wäre – auch davon ausgehen, daß wir die oben erwähnten Fehler beim Grenzübergang $h \to 0$ und $k \to 0$ vernachlässigen können, erhalten wir die

Transformationsformel: $f\colon B \to \mathbb{R}$ sei eine integrierbare Funktion auf $B \subseteq \mathbb{R}^2$, und die Variablen x, y seien als differenzierbare Funktionen $x = x(u, v)$ und $y = y(u, v)$ neuer Variabler u, v dargestellt. Ist dann $B' \subseteq \mathbb{R}^2$ ein Integrationsbereich, für den die Abbildung

$$B' \to B; \quad (u, v) \mapsto \big(x(u, v),\ y(u, v)\big)$$

bijektiv ist, so gilt

$$\iint\limits_B f(x, y)\,dx\,dy = \iint\limits_{B'} f\big(x(u, v)\,y(u, v)\big) \left| \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \right| du\,dv\,.$$

Für unbeschränkte Integrationsbereiche stehen hier natürlich uneigentliche Integrale; da diese – so sie existieren – Grenzwerte von üblichen Integralen sind, ist dies kein Problem.

Im Fall

$$x = r\cos\varphi \quad \text{und} \quad y = r\sin\varphi$$

der Polarkoordinaten ist

$$\frac{\partial x}{\partial r}\frac{\partial y}{\partial \varphi} - \frac{\partial x}{\partial \varphi}\frac{\partial y}{\partial r} = \cos\varphi\cdot(r\cos\varphi) - (-r\sin\varphi)\cdot\sin\varphi = r\,,$$

d.h.

$$\iint\limits_B f(x,y)\,dx\,dy = \iint\limits_{B'} f(r\cos\varphi, r\sin\varphi)\,r\,dr\,d\varphi.$$

Für diese Formel hätten wir eigentlich nicht den ganzen Apparat der Transformationsformel gebrauch; Abbildung 75 zeigt uns, wie wir die Fläche des Bilds eines Parallelogramms exakt ausrechnen können: Variiert r zwischen r und $r+h$ und φ zwischen φ und $\varphi+k$, so erhalten wir in der (x,y)-Ebene als Bild die Differenz zwischen zwei Kreissektoren mit Öffnungswinkel k und Radius $r+h$ beziehungsweise r; der Flächeninhalt ist also

$$\tfrac{1}{2}k(r+h)^2 - \tfrac{1}{2}kr^2 = r\cdot kh + \tfrac{1}{2}kh^2\,.$$

Wenn h und k simultan gegen null gehen, können wir kh^2 gegenüber kh vernachlässigen, der Flächeninhalt kh des Rechtecks aus der (r,φ)-Ebene wird also im wesentlichen nur mit r multipliziert – genau wie es die obige Rechnung auch zeigt.

Mit dieser Formel können wir beispielsweise noch einmal die Fläche eines Kreises ausrechnen: Die Punkte der Kreisscheibe B mit Radius R um den Nullpunkt haben Polarkoordinaten (r,φ) im Rechteck

$$B' = \big\{(r,\varphi) \;\big|\; 0 \le r \le R \quad \text{und} \quad 0 \le \varphi < 2\pi\big\}\,;$$

die Fläche von B ist also

$$\iint\limits_B dx\,dy = \iint\limits_{B'} r\,dr\,d\varphi = \int\limits_0^{2\pi}\left(\int\limits_0^R r\,dr\right)d\varphi = \int\limits_0^{2\pi}\frac{R^2}{2}\,d\varphi = \pi R^2\,.$$

Im Vergleich zum letzten Abschnitt, wo wir das Integral (für $R=1$) in kartesischen Koordinaten ausrechneten und dazu die Funktion $\sqrt{1-x^2}$ integrieren mußten, ist diese Rechnung erheblich einfacher; die Transformationsformel leistet also genau das, was wir von einer verallgemeinerten Substitutionsregel erwarten: Bei *geschickter,* an das Problem angepaßter Substitution kann sie die Berechnung eines Integrals erheblich vereinfachen.

Als nächstes Beispiel wollen wir endlich einmal ein Integral ausrechnen, bei dem wir das Ergebnis nicht besser und viel einfacher durch elementargeometrische Überlegungen bekommen können: das Integral

$$I \underset{\text{def}}{=} \int\limits_{-\infty}^{\infty} e^{-x^2}\,dx\,.$$

Man kann zeigen, daß die Stammfunktion von e^{-x^2} nicht in geschlossener Form durch elementare Funktionen ausdrückbar ist; Integration mittels Stammfunktion ist also zwecklos. Stattdessen benutzen wir folgenden Trick:

Durch Grenzübergang können wir auch auch auf das unendliche „Rechteck“ $\mathbb{R}^2$ als Normalbereich auffassen; daher ist

$$\begin{aligned}\iint\limits_{\mathbb{R}^2} e^{-(x^2+y^2)}\,dx\,dy &= \int\limits_{-\infty}^{\infty}\left(\int\limits_{-\infty}^{\infty} e^{-(x^2+y^2)}\,dx\right)dy\\ &= \int\limits_{-\infty}^{\infty} e^{-y^2}\left(\int\limits_{-\infty}^{\infty} e^{-x^2}\,dx\right)dy = \int\limits_{-\infty}^{\infty} e^{-y^2}\cdot I\,dy\\ &= I\cdot\int\limits_{-\infty}^{\infty} e^{-y^2}\,dy = I^2\,.\end{aligned}$$

In Polarkoordinaten entspricht $\mathbb{R}^2$ dem Bereich $B' = \mathbb{R}_{\ge 0}\times[0,2\pi)$; also ist das betrachtete Integral nach der Transformationsformel auch gleich

$$\iint\limits_{\mathbb{R}^2} e^{-(x^2+y^2)}\,dx\,dy = \iint\limits_{B'} e^{-r^2}\,r\,dr\,d\varphi = \int\limits_0^{2\pi}\left(\int\limits_0^{\infty} re^{-r^2}\,dr\right)d\varphi\,.$$

Die Stammfunktion des neuen Integranden re^{-r^2} ist leicht zu finden: Aus

$$\frac{d}{dr}e^{-r^2} = -2re^{-r^2}$$

folgt sofort, daß

$$\int re^{-r^2}\,dr = -\tfrac{1}{2}e^{-r^2} + C\,.$$

Damit können wir weiterrechnen und erhalten das Ergebnis

$$\int_0^{2\pi} \left(\int_0^{\infty} r e^{-r^2}\, dr \right) d\varphi = \int_0^{2\pi} -\tfrac{1}{2} e^{-r^2} \Big|_0^{\infty} d\varphi = \int_0^{2\pi} \tfrac{1}{2}\, d\varphi = \pi.$$

Also ist $I^2 = \pi$ und, da der Integrand von I überall positiv ist, folgt

$$I = \int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}\,.$$

Zur Verallgemeinerung der Transformationsformel auf höhere Dimensionen beachten wir zunächst, daß der Term

$$\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$$

in der Formel

$$\iint_B f(x,y)\, dx\, dy = \iint_{B'} f\big(x(u,v)\, y(u,v)\big) \left| \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \right| du\, dv$$

gerade gleich der Determinanten der JACOBI-Matrix

$$\begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix}$$

des Koordinatenwechsels

$$(u,v) \mapsto \big(x(u,v),\ y(u,v)\big)$$

ist. Diese können wir uns auch geometrisch veranschaulichen, denn sie ist ja gerade der Flächeninhalt des Parallelogramms mit den Kantenvektoren

$$\begin{pmatrix} \dfrac{\partial x}{\partial u}(u_0,v_0) \\ \dfrac{\partial y}{\partial u}(u_0,v_0) \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} \dfrac{\partial x}{\partial v}(u_0,v_0) \\ \dfrac{\partial y}{\partial v}(u_0,v_0) \end{pmatrix}.$$

Den Flächeninhalt eines Parallelogramms mit Kantenvektoren $\vec{a}$ und $\vec{b}$ können wir auch wie folgt ausrechnen:

Das (kartesische) Koordinatensystem in $\mathbb{R}^2$ sei so gewählt, daß der Vektor $\vec{a}$ ein Vielfaches $a\vec{e}_1$ des ersten Koordinateneinheitsvektors ist. Falls dann auch noch $\vec{b} = b\vec{e}_2$ ein Vielfaches des zweiten sein sollte, falls also das Parallelogramm sogar ein achsenparalleles Rechteck sein sollte, ist dessen Fläche gleich

$$ab = \det \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} = \det(\vec{a}, \vec{b})\,.$$

Im allgemeinen wird das Parallelogramm natürlich kein Rechteck sein; da die Basisvektoren im Gegensatz zu $\vec{a}$ und $\vec{b}$ aufeinander senkrecht stehen, müssen wir $\vec{b}$ dann als Linearkombination $\vec{b} = c\vec{e}_1 + b\vec{e}_2$ schreiben. Dabei können wir $b\vec{e}_2$ als Projektion von $\vec{b}$ auf die von $\vec{e}_2$ aufgespannte Koordinatenachse auffassen. Da sich das Parallelogramm durch Scherung in das Rechteck mit Kantenvektoren $\vec{a}$ und $b\vec{e}_2$ überführen läßt und sich der Flächeninhalte bei Scherungen nicht ändern, hat das Parallelogramm immer noch den Flächeninhalt

$$ab = \det \begin{pmatrix} a & c \\ 0 & b \end{pmatrix} = \det(\vec{a}, \vec{b})\,.$$

Beim Übergang zu einem anderen orthonormalen Koordinatensystem werden die Koordinatenachsen gedreht, d.h. sie werden mit einer Matrix der Form

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$$

multipliziert. Da deren Determinante gleich eins ist, ändert sich dabei nach dem Multiplikationssatz für Determinanten nichts am Wert von $\det(\vec{a}, \vec{b})$. Der Flächeninhalt eines von den Vektoren $\vec{a}$ und $\vec{b}$ aufgespannten Parallelogramms ist also stets gleich $\det(\vec{a}, \vec{b})$, egal bezüglich welcher Orthonormalbasis die Kantenvektoren dargestellt werden.

Genauso können wir auch in höheren Dimensionen argumentieren: Durch mehrfache Scherung läßt sich ein beliebiges Parallelepiped in einen Quader überführen; das Volumen des von n Kantenvektoren $\vec{a}_1, \ldots, \vec{a}_n$ aufgespannten Parallelepipeds ist also gleich

$$\det(\vec{a}_1, \ldots, \vec{a}_n)\,.$$

(Für $n = 3$ kennen wir diese Formel bereits, denn wir wissen, daß dort das Volumen gleich dem Spatprodukt $(\vec{a}_1 \times \vec{a}_2) \cdot \vec{a}_3$ der drei Vektoren ist, und dieses wiederum ist gleich der Determinanten.)

Damit gilt also im $\mathbb{R}^n$ die

Transformationsformel: $f\colon B \to \mathbb{R}$ sei eine integrierbare Funktion auf $B \subseteq \mathbb{R}^n$, und die Variablen $x_1, \ldots, x_n$ seien als differenzierbare Funktionen

$$x_1 = x_1(u_1, \ldots, u_n)$$
$$\vdots$$
$$x_n = x_n(u_1, \ldots, u_n)$$

neuer Variabler $u_1, \ldots, u_n$ dargestellt. Ist dann $B' \subseteq \mathbb{R}^2$ ein Integrationsbereich, für den die Abbildung

$$B' \to B; \quad \boldsymbol{u} \mapsto \boldsymbol{x}(\boldsymbol{u})$$

bijektiv ist, so gilt

$$\int_B \cdots \int f(\boldsymbol{x})\, dx_1 \cdots dx_n = \int_{B'} \cdots \int f\big(\boldsymbol{x}(\boldsymbol{u})\big)\, |\det J_{\boldsymbol{x}}(\boldsymbol{u})|\; du_1 \cdots du_n\,.$$

■

Ausgeschrieben wird die Formel für den $\mathbb{R}^3$, wenn wir kurz

$$F(u,v,w) \underset{\text{def}}{=} f\big(x(u,v,w),\ y(u,v,w),\ z(u,v,w)\big)$$

schreiben, zu

$$\iiint_B f(x,y,z)\, dx\, dy\, dz =$$

$$\iiint_{B'} F(u,v,w) \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{pmatrix} \right| du\, dv\, dw\,.$$

Als Beispiel können wir wieder Koordinatensysteme betrachten: Für Kugelkoordinaten ist

$$x = r\cos\varphi\sin\vartheta$$
$$y = r\sin\varphi\sin\vartheta\ ,$$
$$z = r\cos\vartheta$$

die JACOBI-Matrix ist also

$$\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} & \frac{\partial x}{\partial \vartheta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} & \frac{\partial y}{\partial \vartheta} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \varphi} & \frac{\partial z}{\partial \vartheta} \end{pmatrix} = \begin{pmatrix} \cos\varphi\sin\vartheta & -r\sin\varphi\sin\vartheta & r\cos\varphi\cos\vartheta \\ \sin\varphi\sin\vartheta & r\cos\varphi\sin\vartheta & r\sin\varphi\cos\vartheta \\ \cos\vartheta & 0 & -r\sin\vartheta \end{pmatrix}.$$

Entwicklung nach der dritten Zeile ergibt für die Determinante den Wert

$$\begin{aligned} &\cos\vartheta \begin{vmatrix} -r\sin\varphi\sin\vartheta & r\cos\varphi\cos\vartheta \\ r\cos\varphi\sin\vartheta & r\sin\varphi\cos\vartheta \end{vmatrix} \\ &- r\sin\vartheta \begin{vmatrix} \cos\varphi\sin\vartheta & -r\sin\varphi\sin\vartheta \\ \sin\varphi\sin\vartheta & r\cos\varphi\sin\vartheta \end{vmatrix} \\ &= -r^2\cos\vartheta(\sin\vartheta\cos\vartheta) - r^2\sin\vartheta\cdot\sin^2\vartheta = -r^2\sin\vartheta\,. \end{aligned}$$

Der Betrag der Determinanten der JACOBI-Matrix ist also

$$r^2\,|\sin\vartheta|\,,$$

und mit

$$F(r,\varphi,\vartheta) = f(r\cos\varphi\sin\vartheta,\ r\sin\varphi\sin\vartheta,\ r\cos\vartheta)$$

wird die Transformationsformel zu

$$\iiint_B f(x,y,z)\, dx\, dy\, dz = \iiint_{B'} F(r,\varphi,\vartheta)\cdot r^2\,|\sin\vartheta|\; dr\, d\varphi\, d\vartheta\,.$$

Berechnen wir zur Kontrolle schnell noch einmal das Volumen der Kugel B um den Nullpunkt mit Radius R:

$$\iiint_B dx\, dy\, dz = \int_0^\pi \left(\int_0^{2\pi} \left(\int_0^R r^2\,|\sin\vartheta|\; dr \right) d\varphi \right) d\vartheta$$
$$= \int_0^\pi \left(\int_0^{2\pi} \frac{R^3}{3}\,|\sin\vartheta|\; d\varphi \right) d\vartheta = \frac{2\pi R^3}{3} \int_0^\pi \sin\vartheta\, d\vartheta = \frac{4\pi R^3}{3}\,.$$

Außer den Kugelkoordinaten hatten wir im $\mathbb{R}^3$ auch noch Zylinderkoordinaten betrachtet; für diese ist

$$x = r\cos\varphi$$
$$y = r\sin\varphi \qquad \text{und}$$
$$z = z\,,$$

also

$$\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} & \frac{\partial y}{\partial z} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \varphi} & \frac{\partial z}{\partial z} \end{pmatrix} = \begin{pmatrix} \cos\varphi & -r\sin\varphi & 0 \\ \sin\varphi & r\cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

und die Determinante davon ist wie bei den ebenen Polarkoordinaten gleich r. Also haben wir hier im wesentlichen dieselbe Transformationsformel wie bei den ebenen Polarkoordinaten (Zylinderkoordinaten sind schließlich im wesentlichen auch nichts anderes als ebene Polarkoordinaten), nämlich

$$\iiint_B f(x,y,z)\,dx\,dy\,dz = \iiint_{B'} F(r,\varphi,z)\cdot r\,dr\,d\varphi\,dz\,,$$

wobei wieder

$$F(r,\varphi,z) = f(r\cos\varphi, r\sin\varphi, z)$$

sein soll.

Genauso lassen sich auch beliebige andere Koordinatensysteme behandeln, beispielsweise könnte man zum Rechnen mit einem Ellipsoid auch Ellipsoidkoordinaten

$$x = a\cos\varphi\sin\vartheta$$
$$y = b\sin\varphi\sin\vartheta$$
$$z = c\cos\vartheta$$

einführen und damit das Volumen des Ellipsoids

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

bestimmen. (Die Antwort ist natürlich *(?)* $\frac{4}{3}\pi abc$.)

In Physik und Technik sind noch zahlreiche weitere bewährte Koordinatensysteme im Einsatz, auf die wir hier nicht eingehen können; dank der obigen Transformationsformel kann man immer das benutzen, das der jeweiligen Situation am besten angepaßt ist und mithin am wenigsten Rechnung verlangt.

d) Der Satz von Green und der ebene Satz von Gauß

$B \subset \mathbb{R}^2$ sei ein Bereich, dessen Rand aus endlich vielen Kurvenstücken bestehe. Dann können wir sowohl über B als auch über dessen Randkurve ∂B integrieren; die beiden Sätze in diesem Abschnitt befassen sich mit der Beziehung zwischen diesen beiden Integrationen.

Im allgemeinen kann man natürlich nicht viel erwarten: Nimmt man aus B eine Kurve γ heraus, ist es bei der Integration einer stetigen Funktion gleichgültig, ob man über B oder über $B \setminus \gamma$ integriert, das Randintegral ändert sich aber um das im allgemeinen nicht verschwindende Kurvenintegral über γ.

Für eine geschlossene Kurve γ dagegen erwarten wir, wenn wir uns von der Anschauung leiten lassen und uns die Kurve als eine Art deformierten Kreis vorstellen, daß sie die Ebene in zwei Bereiche zerlegt: einen beschränkten Bereich B und einen unbeschränkten Bereich B'. Hier sollten sich γ und B gegenseitig so stark beeinflussen, daß eigentlich auch die Integration über B etwas mit der Integration über γ zu tun haben müßte.

Nun ist es natürlich etwas gefährlich, sich *nur* von der Anschauung leiten zu lassen, denn es gibt sicherlich erheblich mehr Kurven, als man sich gemeinhin vorstellt. Beispielsweise gibt es auch Kurven, die wie eine Ziffer 8 aussehen, etwa die Lemniskate, und diese zerlegen die Ebene in *drei* Bereiche. Wir müssen also noch zusätzlich fordern, daß sich die Kurve nicht selbst überkreuzt, aber dann gilt in der Tat

Jordanscher Kurvensatz: γ sei eine geschlossene ebene Kurve ohne Überkreuzungen. Dann hat $\mathbb{R}^2 \setminus \gamma$ zwei Zusammenhangskomponenten, von denen die eine beschränkt und die andere unbeschränkt ist.

Der Beweis dieses anschaulich fast selbstverständlichen Satzes erfordert einen erstaunlich großen Aufwand: Man muß die Ebene in kleine Quadrate (oder Dreiecke oder etwas ähnliches) unterteilen, γ durch Kantenzüge aus deren Seiten annähern, den Satz durch aufwendige Rechnungen mit formalen Summen von Quadraten und Quadratseiten für die angenäherten Kurven beweisen und dann schließlich noch durch ein Limesargument zeigen, daß die Aussage auch für γ selbst gilt. Die genaue Durchführung dieses Beweises würde etwa drei Vorlesungstermine erfordern – ein Aufwand, der für uns in keinem vernünftigen Zusammenhang mit seinem Nutzen steht: Bei den meisten in der Praxis vorkommenden Kurven wird die Aussage des JORDANschen Kurvensatzes zumindest für diese speziellen Kurven ohnehin klar sein.

MARIE ENNEMOND CAMILLE JORDAN (1838–1922) leistete wesentliche Beiträge zur Entwicklung der Topologie und der Gruppentheorie. Er beschäftigte sich auch mit der Lösbarkeit nichtlinearer Gleichungen und stellte in einem vielbeachteten Buch die GALOISsche Theorie über die Nichtlösbarkeit allgemeiner Gleichungen vom Grad größer vier dar. Seine Untersuchungen über endliche Körper führten ihn zu der heute als JORDAN-Zerlegung bekannten Darstellung von Matrizen, mit der wir uns im nächsten Semester beschäftigen werden; außerdem bewies er Satze über die Konvergenz von FOURIER-Reihen.

Um den JORDANschen Kurvensatz zu umgehen, starten wir einfach mit einem Bereich $B \subset \mathbb{R}^2$, von dem wir all das verlangen, was uns nachher das Leben einfach macht; in der Praxis sollte es (hoffentlich) selten schwierig sein, diese Forderungen für die Bereiche aus konkreten Anwendungen zu verifizieren.

Satz von Green: $B \subset \mathbb{R}^2$ sei eine abgeschlossene Menge, die sowohl als Vereinigung endlich vieler Normalbereiche vom Typ I wie auch als Vereinigung endlich vieler Normalbereiche vom Typ II geschrieben werden kann; ihre Randkurve γ sei so orientiert, daß B im Gegenuhrzeigersinn umlaufen wird. Weiter sei $\vec{V} \in \mathcal{C}^1(D, \mathbb{R}^2)$ ein differenzierbares

Vektorfeld auf einer offenen Teilmenge $D \supset B$ von $\mathbb{R}^2$. Dann ist

$$\int\limits_{\gamma} \vec{V}\, ds = \iint\limits_{B} \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \right) dx\, dy\,.$$

Beweis: Wir betrachten zunächst nur ein Vektorfeld $\vec{V} = \binom{V_1}{0}$ mit verschwindender y-Komponente und setzen B als Normalbereich

$$N = \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; a \le x \le b \quad \text{und} \quad f(x) \le y \le g(x) \right\}$$

voraus. Das Integral über die Randkurve γ berechnen wir am besten komponentenweise:

$$\gamma_1 \colon [a,b] \to \mathbb{R}^2; \quad t \mapsto \big(t, f(t)\big)$$

sei die untere Begrenzungskurve,

$$\gamma_2 \colon \big[f(b), g(b)\big] \to \mathbb{R}^2; \quad t \mapsto (b,t)$$

die rechte Randstrecke,

$$\gamma_3 \colon [a,b] \to \mathbb{R}^2; \quad t \mapsto \big(t, g(t)\big)$$

die obere Begrenzungskurve, und

$$\gamma_4 \colon \big[f(a), g(a)\big] \to \mathbb{R}^2; \quad t \mapsto \big(a,t\big)$$

schließlich die linke Seitenstrecke.

Dann ist $\dot{\gamma}_1(t) = \binom{1}{f'(t)}$, also

$$\int\limits_{\gamma_1} \vec{V}\, ds = \int\limits_a^b \begin{pmatrix} V_1\big(t, f(t)\big) \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ f'(t) \end{pmatrix} dt = \int\limits_a^b V_1\big(t, f(t)\big)\, dt$$

und entsprechend

$$\int\limits_{\gamma_3} \vec{V}\, ds = \int\limits_a^b \begin{pmatrix} V_1\big(t, g(t)\big) \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ g'(t) \end{pmatrix} dt = \int\limits_a^b V_1\big(t, g(t)\big)\, dt\,.$$

Für die beiden Seitenstrecken ist $\dot\gamma_2(t) = \binom{0}{1} = \dot\gamma_4(t) = \binom{0}{1}$, also ist

$$\int\limits_{\gamma_2} \vec V\,ds = \int\limits_{f(a)}^{g(a)} \begin{pmatrix} V_1(a,t) \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} dt = 0\,,$$

und genauso folgt auch das Verschwinden des Integrals über γ_4.

Wenn wir im Gegenuhrzeigersinn um N herum integrieren, werden die Kurvenstücke γ_3 und γ_4 rückwärts durchlaufen, also ist

$$\begin{aligned}\int\limits_{\gamma} \vec V\,ds &= \int\limits_{\gamma_1} \vec V\,ds + \int\limits_{\gamma_2} \vec V\,ds - \int\limits_{\gamma_3} \vec V\,ds - \int\limits_{\gamma_4} \vec V\,ds \\ &= \int\limits_a^b V_1\big(t, f(t)\big)\,dt - \int\limits_a^b V_1\big(t, g(t)\big)\,dt \\ &= \int\limits_a^b \Big(V_1\big(t, f(t)\big) - V_1\big(t, g(t)\big)\Big)\,dt\,.\end{aligned}$$

Da N Normalbereich vom Typ I ist, können wir auch das Flächenintegral über N leicht ausrechnen:

$$\begin{aligned}&\iint\limits_N \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y}\right) dx\,dy = \iint\limits_N -\frac{\partial V_1}{\partial y}\,dx\,dy \\ &= -\int\limits_a^b \left(\int\limits_{f(x)}^{g(x)} \frac{\partial V_1}{\partial y}\,dy\right) dx = \int\limits_a^b \Big(V_1\big(x, g(x)\big) - V_1\big(x, f(x)\big)\Big)\,dx\,.\end{aligned}$$

Zumindest für das hier betrachtete spezielle Vektorfeld V ohne y-Komponente und einen Normalbereich vom Typ I ist der Satz also richtig.

Damit gilt er aber zumindest für dieses spezielle Vektorfeld auch für jeden Bereich B, der sich in Normalbereiche vom Typ I zerlegen läßt: Das Flächenintegral über B ist gleich der Summe der Flächenintegrale über die Normalbereiche vom Typ I, in die wir B zerlegt haben und somit

problemlos. Die Randkurve von B besteht einerseits aus Kurvenstücken, die auch zum Rand von B gehören, andererseits aus solchen, die durch die Zerlegung eingeführt wurden und zwei Normalbereiche voneinander trennen. Letztere werden aber als Randkurven der beiden angrenzenden Normalkurven jeweils in entgegengesetzter Richtung durchlaufen, so daß sich die Integrale längs solcher Kurvenstücke gegenseitig wegheben. Summiert man also die Integrale über die Ränder aller Normalbereiche auf, bleiben am Ende nur die Integrale längs jener Kurvenstücke übrig, die auf der Randkurve von B liegen, die Summe ist somit gleich dem Integral über die Randkurve.

Damit ist der Satz bewiesen für alle Vektorfelder $\vec V$ mit verschwindender y-Komponente.

Als nächstes beweisen wir ihn für Vektorfelder $\vec V$ mit verschwindender x-Komponente; wenn man mit Normbereichen vom Typ II argumentiert statt mit solchen vom Typ I, kann man dazu die obigen Argumente fast wörtlich wiederholen.

Der Rest des Beweises ist nun einfach: Ein beliebiges Vektorfeld $\vec V$ läßt sich als Summe

$$\vec V(x,y) = \begin{pmatrix} V_1(x,y) \\ V_2(x,y) \end{pmatrix} = \begin{pmatrix} V_1(x,y) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ V_2(x,y) \end{pmatrix}$$

schreiben, und der Satz gilt für beide Summanden. Da beide Seiten der Behauptung linear in $\vec V$ sind, folgt der Satz auch für $\vec V$ selbst. ∎

GEORGE GREEN (1793–1841) war der Sohn eines Bäckers aus Nottingham. Er besuchte nur von 1801 bis 1802 eine Schule, danach arbeitete er in der Bäckerei und später in der dazugekauften Mühle. Es ist nicht bekannt, wann und wie er Mathematik lernte. 1827 veröffentlichte er sein Buch *An Essay on the Application of Mathematical Analysis to the Theory of Electricity and Magnetism,* von dem 51 Exemplare verkauft wurden, größtenteils in Nottingham selbst. Einer der Leser stellte Kontakte zur Cambridge Philosophical Society und zur Royal Academy in Edinburgh her, so daß GREENs weitere Arbeiten (über Elektrizität und über Hydrodynamik) dort veröffentlicht wurden. 1833 begann er mit dem Studium der Mathematik an der Universität Cambridge; nach dessen Abschluß blieb er in Cambridge, bis er 1840 wegen gesundheitlicher Probleme nach Nottingham zurückkehrte. Weder GREEN selbst noch seine Zeitgenossen erkannten die Bedeutung seiner Arbeiten, die, in heutiger Sprechweise, Potentialfunktionen einführten und für die

Physik nutzbar machten. Auf dieser Grundlage bauten später JAMES CLERK MAXWELL (1831–1879) und andere die Elektrodynamik auf.

Wenn wir für $\vec{V}$ speziell ein Vektorfeld der Form

$$\vec{V}(x,y) = \begin{pmatrix} -\Phi_y(x,y) \\ \Phi_x(x,y) \end{pmatrix} \quad \text{mit} \quad \Phi \in \mathcal{C}^2(D,\mathbb{R})$$

einsetzen, erhalten wir für das Flächenintegral

$$\iint\limits_B \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \right) dx\,dy = \iint\limits_B (\Phi_{xx} + \Phi_{yy})\,dx\,dy = \iint\limits_B \Delta\Phi\,dx\,dy\,,$$

wobei Δ wie üblich den LAPLACE-Operator

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

bezeichnet.

Das Integral über die Randkurve $\gamma = (\gamma_1, \gamma_2)$ wird zu

$$\begin{aligned} &\int\limits_\gamma \begin{pmatrix} -\Phi_y\big(\gamma(t)\big) \\ \Phi_x\big(\gamma(t)\big) \end{pmatrix} \begin{pmatrix} \dot\gamma_1(t) \\ \dot\gamma_2(t) \end{pmatrix} dt \\ &= \int\limits_\gamma \Big(-\Phi_y\big(\gamma(t)\big) \cdot \dot\gamma_1(t) + \Phi_x\big(\gamma(t)\big) \cdot \dot\gamma_2(t) \Big)\, dt \\ &= \int\limits_\gamma \begin{pmatrix} \Phi_x\big(\gamma(t)\big) \\ \Phi_y\big(\gamma(t)\big) \end{pmatrix} \begin{pmatrix} \dot\gamma_2(t) \\ -\dot\gamma_1(t) \end{pmatrix} dt\,. \end{aligned}$$

Der erste Vektor im Integranden der dritten Zeile ist einfach der Gradient von Φ. Der zweite Vektor hat Skalarprodukt null mit dem Tangentenvektor $\dot\gamma(t) = \binom{\dot\gamma_1(t)}{\dot\gamma_2(t)}$, steht also auf diesem senkrecht und ist somit ein *Normalenvektor.* Das Skalarprodukt des Gradienten einer Funktion mit einem Vektor ist bekanntlich die *Richtungsableitung* der Funktion in Richtung dieses Vektors; mit

$$\vec{n} = \begin{pmatrix} \dot\gamma_2(t) \\ -\dot\gamma_1(t) \end{pmatrix} \quad \text{und} \quad \partial_{\vec{n}} = \vec{n} \cdot \operatorname{grad}$$

erhalten wir somit den

Satz von Gauß (ebener Fall)**:** Für B, D und γ wie im Satz von GREEN und $\Phi \in \mathcal{C}^2(D,\mathbb{R}^2)$ gilt

$$\iint\limits_B \Delta\Phi\ dx\,dy = \int\limits_\gamma \partial_{\vec{n}}\Phi\ ds\,.$$

■

Mit einer Verallgemeinerung sowie auch mit der anschaulichen Interpretation dieses Satzes werden wir uns in Abschnitt *f)* noch genauer befassen.

e) Oberflächenintegrale

Im dreidimensionalen Fall wird der Satz von GAUSS, wie die meisten wohl bereits aus der Physik wissen, ein Volumenintegral mit dem Fluß durch eine Oberfläche in Verbindung bringen. Was Volumenintegrale sind, haben wir inzwischen auch hier definiert; in diesem Abschnitt geht es um den Begriff des Oberflächenintegrals, um den Fluß eines Vektorfelds durch eine Oberfläche und um verwandte Themen.

Die Vorgehensweise bei der Definition eines Oberflächenintegrals entspricht genau der bei der Definition des Kurvenintegrals: Dort hatten wir Kurvenstücke als Funktionen $\gamma\colon [a,b] \to \mathbb{R}^n$ definiert; hier definieren wir entsprechend *Flächenstücke* durch Funktionen zweier Parameter. Wir müssen dabei allerdings etwas sorgfältiger vorgehen, denn im Vergleich zum Kurvenfall ist die Situation zumindest in zweierlei Hinsicht komplexer: Erstens gibt es auf der reellen Geraden im wesentlichen nur eine Art von zusammenhängenden Teilmengen, nämlich die Intervalle. In der Ebenen gibt es erheblich mehr Möglichkeiten – aber damit hatten wir uns ja schon bei der Definition von Flächen- und Volumenintegralen beschäftigt. Zweitens mußten wir bei Kurven nur fordern, daß $\dot\gamma(t)$ nirgends verschwindet, um die Existenz eines Tangentenvektors zu garantieren. Hier im Zweidimensionalen recht das nicht mehr: Ein Flächenstück hat in jedem Punkt Anspruch nicht nur auf eine Tangente, sondern auf eine ganze *Tangentialebene;* wir müssen also sicherstellen, daß es in jedem Punkt mindestens zwei linear unabhängige Tangentenvektoren gibt.

Die dazu notwendigen Definitionen kann man problemlos für Flächenstücke in jedem $\mathbb{R}^n$ hinschreiben; wir wollen uns aber, da dies für alle hier betrachteten Anwendungen ausreicht, auf den etwas anschaulicheren Fall von Flächenstücken im $\mathbb{R}^3$ beschränken.

Ein Flächenstück soll demnach gegeben sein durch eine Funktion

$$f\colon B \to \mathbb{R}^3; \quad (u,v) \mapsto f(u,v) = \begin{pmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{pmatrix}$$

auf einer Teilmenge $B \subset \mathbb{R}^2$; mit Blick auf unsere Erfahrungen bei Kurvenstücken wollen wir von vornherein fordern, daß f differenzierbar sein soll in einer offenen Umgebung von B. Für kleine reelle Zahlen h, k ist dann für einen Punkt $(u_0, v_0) \in B$

$$f(u_0+h, v_0+k) = f(u_0,v_0) + f_u(u_0,v_0)\cdot h + f_v(u_0,v_0)\cdot k + o\left(\sqrt{h^2+k^2}\right) .$$

Da hier nur h und k variabel sind, bedeutet die Forderung nach zwei linear unabhängigen Tangentialvektoren, daß die beiden Vektoren $f_u(u_0,v_0)$ und $f_v(u_0,v_0)$ in allen Punkten $(u_0,v_0) \in B$ linear unabhängig sein müssen.

An dieser Stelle führt die Beschränkung auf den $\mathbb{R}^3$ zu einer kleinen Vereinfachung: Im $\mathbb{R}^3$, und nur dort, existiert ein Vektorprodukt; zwei Vektoren sind genau dann linear unabhängig, wenn dieses Vektorprodukt ungleich dem Nullvektor ist. Sowohl der Betrag als auch die Richtung dieses Vektors werden uns schon bald in mehrfacher Hinsicht nützlich sein; wir definieren daher

Definition: Ein reguläres Flächenstück im $\mathbb{R}^3$ ist gegeben durch einen Bereich $B \subset \mathbb{R}^2$ und eine in einer offenen Menge $D \supseteq B$ differenzierbaren Funktion $f\colon D \to \mathbb{R}^3; \quad (u,v) \mapsto f(u,v)$, für die gilt:
1.) Die Einschränkung $f_{|B}$ von f auf B ist injektiv.
2.) In jedem Punkt $(u_0,v_0) \in B$ ist $f_u(u_0,v_0) \times f_v(u_0,v_0) \neq \vec{0}$.

Die wichtigste Oberfläche, mit der wir es im folgenden zu tun haben werden, ist die der Kugel. Die Oberfläche einer Kugel vom Radius R hat in Kugelkoordinaten die Gleichung $r = R$ mit beliebigen Winkelkoordinaten φ und ϑ, sie ist also (in kartesischen Kordinaten) gegeben

durch die Funktion

$$f\colon [0,2\pi) \times [0,\pi] \to \mathbb{R}^3; \quad (\varphi,\vartheta) \mapsto \begin{pmatrix} R\cos\varphi\sin\vartheta \\ R\sin\varphi\sin\vartheta \\ R\cos\vartheta \end{pmatrix} .$$

Die Differenzierbarkeit ist hier überhaupt kein Problem: Die angegebene Funktion ist sogar auf ganz $\mathbb{R}^2$ beliebig oft stetig differenzierbar.

Bei der Injektivität gibt es allerdings Probleme: Für $\vartheta = 0$ und $\vartheta = \pi$ ist $\sin\vartheta = 0$; daher werden alle Parameterpaare $(\varphi, 0)$ auf $(0,0,R)$ und aller Parameterpaare (φ,π) auf $(0,0,-R)$ abgebildet.

Das ist aber glücklicherweise auch bereits alles was passieren kann, denn falls $\sin\vartheta \neq 0$ ist, können wir aus der dritten Komponenten $R\cos\vartheta$ den Winkel $\vartheta \in [0,\pi]$ eindeutig bestimmen. Damit kennen wir auch $R\sin\vartheta$ und bekommen aus den ersten beiden Komponenten von f den Cosinus und den Sinus von φ, wodurch $\varphi \in [0,2\pi)$ eindeutig festgelegt ist.

Die partiellen Ableitungen von f sind

$$f_\varphi = \begin{pmatrix} -R\sin\varphi\sin\vartheta \\ R\cos\varphi\sin\vartheta \\ 0 \end{pmatrix} \quad \text{und} \quad f_\vartheta = \begin{pmatrix} R\cos\varphi\cos\vartheta \\ R\sin\varphi\cos\vartheta \\ -R\sin\vartheta \end{pmatrix} ,$$

also ist

$$f_\varphi \times f_\vartheta = \begin{pmatrix} R^2\cos\varphi\sin^2\vartheta \\ R^2\sin\varphi\sin^2\vartheta \\ R^2\sin\vartheta\cos\vartheta \end{pmatrix} = R\sin\vartheta \cdot f(\varphi,\vartheta) .$$

Da für eine echte Kugel $R > 0$ sein muß und $f(\varphi,\vartheta)$ die Länge R hat, ist dieser Vektor genau dann gleich dem Nullvektor, wenn $\sin\vartheta = 0$ ist; wie bei der Injektivität gibt es also wieder Probleme an den Polen, aber auch nur dort.

Somit können wir die Kugel nur dann als reguläres Flächenstück auffassen, wenn wir die beiden Pole herausnehmen, d.h. wenn wir uns auf den Parameterbereich $[0,2\pi) \times (0,\pi)$ beschränken.

Für die meisten praktischen Zwecke ist das natürlich völlig unproblematisch: Bei der Flächenbestimmung oder allgemeiner der Integration einer beschränkte Funktion sind einzelne, isoliert liegende Punkte irrelevant. Lediglich im Falle von uneigentlichen Integranden muß man

hier sehr vorsichtig sein; hier in dieser Vorlesung sollen entsprechende Integrale deshalb vorsichtshalber gleich gar nicht erst definiert werden.

Nachdem wir nun wissen, was ein reguläres Flächenstück ist, müssen wir als nächstes lernen, damit zu rechnen. Bevor wir uns an die Bestimmung von Flächeninhalten machen, ist es vielleicht ganz nützlich (wenn auch nicht unbedingt notwendig), daß wir uns ein paar Gedanken über die *Längenmessung* machen.

Längen sind Eigenschaften von Kurven; wir müssen also eine Kurve oder – da wir Längen auch stückweise aneinandersetzen können – einfacher ein Kurvenstück auf einem Flächenstück betrachten. Genau wie man etwa bei der Navigation auf der Erde nicht von einem dreidimensionalen kartesischen Koordinatensystem ausgeht, sondern von der geographischen Länge und Breite, empfiehlt es sich auch hier, Kurven über die Parameter des Flächenstücks zu definieren – und sei es auch nur, um sicher zu sein, daß die Kurve auch wirklich auf dem Flächenstück liegt.

Wir beschreiben eine Kurve auf einem Flächenstück $f\colon B \to \mathbb{R}^3$ daher durch eine Funktion

$$\delta\colon [a,b] \to B; \quad t \mapsto \big(u(t), v(t)\big)\ ;$$

die eigentliche Kurve im $\mathbb{R}^3$ wird dann beschrieben durch die zusammengesetzte Abbildung

$$\gamma = f \circ \delta\colon [a,b] \to \mathbb{R}^3; \quad t \mapsto f\big(u(t), v(t)\big)\ .$$

Wir wollen natürlich, daß diese Kurve immer dieselbe Länge hat, egal ob wir sie einfach so im $\mathbb{R}^3$ oder auf einem Flächenstück betrachten. Die Länge des Kurvenstücks γ auf dem Flächenstück f ist daher gleich der bekannten Länge des Kurvenstücks γ im $\mathbb{R}^3$, also

$$\int_a^b |\dot\gamma(t)|\ dt = \int_a^b \left| f_u\big(u(t), v(t)\big)\frac{du}{dt} + f_v\big(u(t), v(t)\big)\frac{dv}{dt}\right| dt\ .$$

Der Betrag des Vektors $f_u\big(u(t), v(t)\big)\frac{du}{dt} + f_v\big(u(t), v(t)\big)\frac{dv}{dt}$ ist gleich der Wurzel aus dem Skalarprodukt des Vektors mit sich selbst, also der

Wurzel aus

$$\begin{aligned} &f_u\big(u(t), v(t)\big)^2 \left(\frac{du}{dt}(t)\right)^2 \\ &+ 2 f_u\big(u(t), v(t)\big) \cdot f_v\big(u(t), v(t)\big)\ \left(\frac{du}{dt}(t)\right)\left(\frac{dv}{dt}(t)\right) \\ &+ f_v\big(u(t), v(t)\big)^2 \left(\frac{dv}{dt}\right)^2\ . \end{aligned}$$

In dieser Formel können wir zwei Arten von Termen unterscheiden: Rechts stehen jeweils die Ableitungen von u und v nach t; diese hängen nur von der jeweiligen Kurve ab, die ja gerade durch die Funktionen $u(t)$ und $v(t)$ definiert ist. Links dagegen stehen Ausdrücke in den partiellen Ableitungen von f nach u und v; diese sind Funktionen auf dem Flächenstück und sind insbesondere unabhängig von jeder Kurve. Wir bezeichnen daher diese drei letzteren Größen

$$\begin{aligned} E(u,v) &= f_u(u,v) \cdot f_u(u,v)\,, \\ F(u,v) &= f_u(u,v) \cdot f_v(u,v) \quad \text{und} \\ G(u,v) &= f_v(u,v) \cdot f_v(u,v) \end{aligned}$$

als *Fundamentalgrößen* des Flächenstücks f.

Um etwas Übung im Umgang mit diesen Größen zu bekommen, wollen wir sie für die Kugeloberfläche in ihrer oben angegebenen Parametrisierung berechnen. Hier ist, wie wir schon wissen,

$$f_\varphi = \begin{pmatrix} -R\sin\varphi\sin\vartheta \\ R\cos\varphi\sin\vartheta \\ 0 \end{pmatrix} \quad \text{und} \quad f_\vartheta = \begin{pmatrix} R\cos\varphi\cos\vartheta \\ R\sin\varphi\cos\vartheta \\ -R\sin\vartheta \end{pmatrix}\ ;$$

also ist

$$\begin{aligned} E(\varphi,\vartheta) &= R^2(\sin^2\varphi\sin^2\vartheta + \cos^2\varphi\sin^2\vartheta) = R^2\sin^2\vartheta \\ F(\varphi,\vartheta) &= R^2(-\sin\varphi\sin\vartheta\cos\varphi\cos\vartheta + \cos\varphi\sin\vartheta\sin\varphi\cos\vartheta) = 0 \\ G(\varphi,\vartheta) &= R^2(\cos^2\varphi\cos^2\vartheta + \sin^2\varphi\cos^2\vartheta + \sin^2\vartheta) = R^2\ . \end{aligned}$$

Dieses Ergebnis sollte man auch geometrisch interpretieren: Das Verschwinden von F, die Orthogonalität von f_φ und f_ϑ also, bedeutet,

daß die Längenkreise und die Breitenkreise in jedem Punkt der Kugel (außer den beiden Polen) aufeinander senkrecht stehen; die Konstans von G kommt daher, daß alle Längenkreise gleich lang sind, und die ϑ-Abhängigkeit von E schließlich reflektiert die Tatsache, daß die Breitenkreise verschiedene Radien haben.

Allgemein kann man die Länge einer Kurve auf einem Flächenstück somit als Integral

$$\int_a^b \sqrt{E\,\dot{u}^2 + 2F\,\dot{u}\dot{v} + G\,\dot{v}^2}\,dt$$

berechnen; dem interessierten Leser sei empfohlen, damit beispielsweise überprüfen, daß Breitenkreise i.a. nicht die kürzeste Verbindung zwischen zweien ihrer Punkte sind, sondern daß der Kreis um den Kugelmittelpunkt durch diese beiden Punkte (der sogenannte Großkreis) eine kürzere Verbindungskurve liefert.

Unser Hauptziel hier sind allerdings nicht Längenberechnungen, sondern die Berechnung von Oberflächen *bzw.* von Flüssen von Vektorfeldern durch diese Oberflächen. Dabei gehen wir genauso vor, wie bei der Transformationsformel im vorigen Abschnitt: Wir unterteilen den Parameterbereich B durch Rechtecke, und betrachten deren Bilder auf dem Flächenstück. Diese sind, da wir f als differenzierbar vorausgesetzt haben, in erster Näherung Parallelogramme, wobei die Näherung bei zunehmender Verfeinerung des Rechteckgitters auf B immer besser wird. Der Flächeninhalt eines solchen Parallelogramms mit Kantenvektoren $f_u \cdot h$ und $f_v \cdot k$ ist

$$hk \cdot |f_u \times f_v| \ ,$$

also sollte die Fläche des Flächenstücks gleich

$$\iint_B |f_u \times f_v|\ du\,dv$$

sein, und genau das definieren wir auch als den Flächeninhalt eines Flächenstücks.

Definition: Der Flächeninhalt eines Flächenstücks $f\colon B \to \mathbb{R}^3$ ist

$$\iint_f dO \underset{\text{def}}{=} \iint_B |f_u \times f_v|\ du\,dv \ .$$

(Das O in dO soll dabei an Oberfläche erinnern.)

Zur Kontrolle, ob diese Definition sinnvoll sein kann, berechnen wir die Oberfläche der Kugel: Hier ist, wie wir bereits wissen,

$$f_\varphi \times f_\vartheta = \begin{pmatrix} R^2 \cos\varphi \sin^2\vartheta \\ R^2 \sin\varphi \sin^2\vartheta \\ R^2 \sin\vartheta \cos\vartheta \end{pmatrix} = R \sin\vartheta \cdot f(\varphi,\vartheta)\,;$$

die Länge dieses Vektors ist also, da $f(\varphi,\vartheta)$ als Radiusvektor natürlich die Länge R haben muß, gleich $R^2 \sin\vartheta$. Die Oberfläche der Kugel berechnet sich daher zu

$$\begin{aligned}\iint_B R^2 \sin\vartheta &= \int_0^{2\pi} \left(\int_0^{\pi} R^2 \sin\vartheta\,d\vartheta \right) d\varphi \\ &= 2\pi \cdot R^2 \cdot (-\cos\pi + \cos 0) = 4\pi R^2 \ ,\end{aligned}$$

womit wir uns wieder einmal in hundertprozentiger Übereinstimmung mit der Schulmathematik befinden.

Genau wie Kurvenintegrale sind auch Oberflächenintegrale unabhängig von der Parametrisierung des Flächenstücks; im wesentlichen geht alles genau wie bei den Kurvenintegralen, ist aber etwas aufwendiger. Der Vollständigkeit halber sei die entsprechende Aussagen samt Beweis im Kleindruck abgedruckt:

Lemma: $f\colon B \to \mathbb{R}^3$ und $g\colon B' \to \mathbb{R}^3$ seien zwei Flächenstücke, für die es eine in einer Umgebung von B stetig differenzierbare Funktion φ gebe, die B bijektiv auf B' abbildet derart, daß $g = f \circ \varphi$ ist. Dann ist

$$\iint_f du\,dv = \iint_g du\,dv \ .$$

Beweis: Nach Definition ist

$$\iint_f dO = \iint_B |f_u \times f_v|\ du\,dv \quad \text{und} \quad \iint_g dO = \iint_B |g_u \times g_v|\ du\,dv \ .$$

Wenn wir die drei Komponenten von f *bzw.* g mit $f^{(x)}, f^{(y)}, f^{(z)}$ *bzw.* $g^{(x)}, g^{(y)}, g^{(z)}$ bezeichnen, ist

$$f_u \times f_v = \begin{pmatrix} f_u^{(y)} f_v^{(z)} - f_u^{(z)} f_v^{(y)} \\ f_u^{(z)} f_v^{(x)} - f_u^{(x)} f_v^{(z)} \\ f_u^{(x)} f_v^{(y)} - f_u^{(y)} f_v^{(x)} \end{pmatrix} = \begin{pmatrix} \det \begin{pmatrix} f_u^{(y)} & f_v^{(y)} \\ f_u^{(z)} & f_v^{(z)} \end{pmatrix} \\ \det \begin{pmatrix} f_u^{(z)} & f_v^{(z)} \\ f_u^{(x)} & f_v^{(x)} \end{pmatrix} \\ \det \begin{pmatrix} f_u^{(x)} & f_v^{(x)} \\ f_u^{(y)} & f_v^{(y)} \end{pmatrix} \end{pmatrix} ,$$

und genau entsprechend natürlich auch für g. Die links stehenden Determinanten sind offensichtlich gerade die Determinaten der JACOBI-Matrizen der verschiedenen zweidimensionalen Projektionen von f; an der ersten Stelle etwa die der Funktion

$$f^{(yz)} \underset{\text{def}}{=} \begin{pmatrix} f^{(y)} \\ f^{(z)} \end{pmatrix} : B \to \mathbb{R}^2 .$$

Analog dazu definieren wir auch Funktionen $f^{(zx)}, f^{(xy)}$ und die entsprechenden Funktionen für g.

Aus $g = f \circ \varphi$ folgt, daß auch $g^{(x)} = f^{(x)} \circ \varphi$ und entsprechend für die anderen Komponenten. Damit ist auch $f^{(yz)} = g^{(yz)} \circ \varphi$ *usw.*

Nach der zweidimensionalen Kettenregel ist dann

$$J_{g^{(yz)}} = \left(J_{f^{(yz)}} \circ \varphi \right) \cdot J_\varphi ,$$

also ausgeschrieben

$$\begin{pmatrix} g_u^{(y)} & g_v^{(y)} \\ g_u^{(z)} & g_v^{(z)} \end{pmatrix} = \begin{pmatrix} f_u^{(y)} \circ \varphi & f_v^{(y)} \circ \varphi \\ f_u^{(z)} \circ \varphi & f_v^{(z)} \circ \varphi \end{pmatrix} \cdot J_\varphi .$$

Entsprechendes gilt auch für die anderen Indexkombinationen, und nach dem Multiplikationssatz für Determinanten gilt eine entsprechende Produktbeziehung auch für die Determinaten der hier stehenden Matrizen; insgesamt erhalten wir also, daß

$$g_u \times g_v = \left((f_u \circ \varphi) \times (f_v \circ \varphi) \right) \cdot \det J_\varphi$$

ist und dementsprechend

$$|g_u \times g_v| = \left((f_u \circ \varphi) \times (f_v \circ \varphi) \right) \cdot |\det J_\varphi| .$$

Damit folgt die Behauptung aus der Transformationsformel. ∎

Die Berechnung von Vektorprodukten ist etwas umständlich und sehr anfällig für Vorzeichenfehler; wir wollen daher sehen, daß die Fundamentalgrößen einer Fläche ihrem Namen gerecht werden und uns auch diese Berechnung abnehmen können: Der Flächeninhalt eines Parallelogramms mit Kantenvektoren $\vec{a}$ und $\vec{b}$ ist gleich dem Produkt der Längen

der beiden Vektoren mal dem Sinus des eingeschlossenen Winkels α. Das Quadrat des Flächeninhalts ist also

$$\begin{aligned} &(\vec{a} \cdot \vec{a})(\vec{b} \cdot \vec{b}) \sin^2 \alpha = (\vec{a} \cdot \vec{a})(\vec{b} \cdot \vec{b})(1 - \cos^2 \alpha) \\ &= (\vec{a} \cdot \vec{a})(\vec{b} \cdot \vec{b}) - (\vec{a} \cdot \vec{a})(\vec{b} \cdot \vec{b}) \cos^2 \alpha = (\vec{a} \cdot \vec{a})(\vec{b} \cdot \vec{b}) - (\vec{a} \cdot \vec{b})^2 , \end{aligned}$$

und der Flächeninhalt selbst somit

$$\sqrt{(\vec{a} \cdot \vec{a})(\vec{b} \cdot \vec{b}) - (\vec{a} \cdot \vec{b})^2} .$$

Wir interessieren uns für das Parallelogramm mit Kantenvektoren $h \cdot f_u$ und $k \cdot f_v$; hier wird diese Formel zu

$$hk \cdot \sqrt{(f_u \cdot f_u)(f_v \cdot f_v) - (f_u \cdot f_v)^2} = hk \sqrt{EG - F^2} .$$

Die Fläche eines Flächenstücks kann also auch berechnet werden als

$$\iint_B \sqrt{EG - F^2}\, du\, dv .$$

Im konkreten Fall steht hier natürlich genau dasselbe Integral wie bei der Formel mit dem Betrag des Vektorprodukts, allerdings erfordert diese Formel, sofern man die Fundamentalgrößen einer Fläche bereits kennt, deutlich weniger Rechenaufwand. Dem Leser sei empfohlen, sich am Beispiel der Kugeloberfläche hiervon zu überzeugen!

Obwohl wir uns auf Flächen im $\mathbb{R}^3$ beschränken wollen, sei hier zumindest kurz darauf hingewiesen, daß wir bei der Herleitung dieser neuen Darstellung des Flächeninhalts nirgends benutzen mußten, daß wir im $\mathbb{R}^3$ sind; diese Formel gilt also auch für Flächenstücke im $\mathbb{R}^n$ mit $n > 3$.

Eine für uns wichtige Besonderheit, die *nur* im $\mathbb{R}^3$ gilt, ist dagegen die Tatsache, daß wir jedem Flächenstück eine eindeutig bestimmte Normalenrichtung zuordnen können: Im Dreidimensionalen gibt es nur eine Richtung, die auf der Tangentialebenen eines Flächenstücks senkrecht steht; in Dimension $n > 3$ gibt es dagegen einen ganzen $(n-2)$-dimensionalen Raum mit dieser Eigenschaft.

Wenn wir von der Parameterdarstellung $f: D \to B$ eines regulären Flächenstücks im $\mathbb{R}^3$ und einer festen Reihenfolge der Parameter ausgehen, können wir sogar unterscheiden, nach welcher Seite diese Richtung

sich von der Fläche entfernt: Die Tangentialebene wird aufgespannt von den partiellen Ableitungen f_u und f_v von f nach den beiden Parametern u, v auf $D \subseteq \mathbb{R}^2$, und das Kreuzprodukt $f_u \times f_v$ liefert einen Vektor, der auf diesen beiden Tangentialvektoren und somit der gesamten Tangentialebenen senkrecht steht.

Damit können wir jedem Punkt eines *regulären* Flächenstücks eine eindeutig bestimmte Normalenrichtung zuordnen. Man beachte, daß dies nur möglich ist aufgrund der Forderungen, die wir an ein reguläres Flächenstück gestellt haben: Die Abbildung f muß injektiv sein, so daß die Parameterwerte (u, v) zu jedem Punkt des Flächenstücks eindeutig bestimmt sind, und $f_u \times f_v$ darf nirgends verschwinden, da der Nullvektor keine wohlbestimmte Richtung hat.

Typisches Beispiel eines nichtregulären Flächenstücks ist das MÖBIUS-Band

$$f\colon D = [0,1] \times [0, 2\pi) \to \mathbb{R}^3; \quad (u,v) \mapsto \begin{pmatrix} (2 + u\cos v)\cos 2v \\ (2 + u\cos v)\sin 2v \\ u \sin v \end{pmatrix} ,$$

das durch Zusammenkleben der Enden eines einmal verdrillten rechteckigen Streifens entsteht. Hier ist f für $u = 0$ nicht injektiv, denn $f(0, v) = f(0, v + \pi)$; wir haben also kein reguläres Flächenstück. In der Tat wir für die Parameterwerte $(0, v)$ der Mittelkreis des Bandes zweimal durchlaufen. Dort ist

$$f_u(0,v) = \begin{pmatrix} \cos v \cos 2v \\ \cos v \sin 2v \\ \sin v \end{pmatrix} \quad \text{und} \quad f_v(0,v) = \begin{pmatrix} -4\sin 2v \\ 4\cos 2v \\ 0 \end{pmatrix} ,$$

also

$$f_u(0,v) \times f_v(0,v) = \begin{pmatrix} -4\sin v \cos 2v \\ -4\sin v \sin 2v \\ 4\cos v \end{pmatrix}$$

und damit

$$f_u(0,v) \times f_v(0,v) = -f_u(0, v+\pi) \times f_v(0, v+\pi) .$$

es gibt also auf dem Mittelkreis keine eindeutig bestimmte Normalenrichtung.

Für Kurven hatten wir einerseits RIEMANN-STIELTJES-Integrale definiert, die für eine beliebige Funktion auf der Kurve erklärt sind und in deren Definition die Länge des Tangentenvektors eingeht; andererseits hatten wir Integrale über Vektorfelder, in deren Definition der Tangentenvektor selbst einging. Genauso gehen wir auch hier bei den Oberflächenintegralen vor, nur daß wir jetzt den Normalenvektor anstelle des Tangentenvektors benutzen:

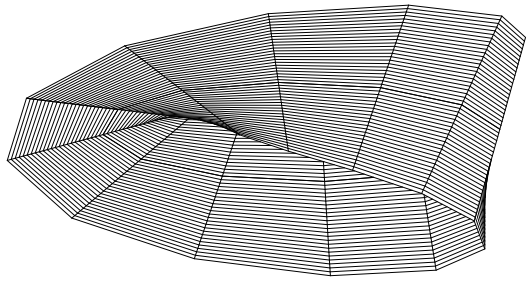


Abb. 76: Das Möbiusband

Definition: $f\colon D \to \mathbb{R}^3$ sei ein reguläres Flächenstück.
a) Ist g eine auf $f(D)$ definierte stetige Funktion, so bezeichnen wir

$$\iint\limits_f g\, dO \underset{\text{def}}{=} \iint\limits_D g\big(f(u,v)\big)\ \big|f_u(u,v) \times f_v(u,v)\big|\ du\, dv$$

als *Oberflächenintegral* von g über f.
b) Ist $\vec{V}$ ein in einer Umgebung von $f(D)$ definiertes Vektorfeld, so bezeichnen wir

$$\iint\limits_f \vec{V} \cdot d\vec{O} \underset{\text{def}}{=} \iint\limits_D \vec{V}\big(f(u,v)\big) \cdot \big(f_u(u,v) \times f_v(u,v)\big)\, du\, dv$$

als den *Fluß* des Vektorfelds $\vec{V}$ durch die Oberfläche f.

Eine unproblematische Verallgemeinerung des obigen Lemmas zeigt, daß auch diese Integrale im dort definierten Sinne unabhängig von der Parametrisierung sind.

Der Name *Fluß* für das unter *b)* definierte Integral wird klar, wenn man sich f etwa als eine Kugeloberfläche vorstellt. In jedem Punkt $\boldsymbol{x}$ auf dieser Oberfläche kann man ein kartesisches Koordinatensystem aus einem Normaleneinheitsvektor $\vec{n}_{\boldsymbol{x}}$ und zwei Tangenteneinheitsvektoren

verankern. Drückt man nun den Vektor $\vec{V}(\boldsymbol{x})$ in diesem Koordinatensystem aus, so beschreibt die $\vec{n}_{\boldsymbol{x}}$-Komponente jenen Teil des Vektors, der durch die Kugeloberfläche hindurch nach innen oder außen geht (welches von beiden hängt ab vom Vorzeichen der Komponente und von der Orientierung des Normaleneinheitsvektors), während die beiden anderen Komponenten den Teil beschreiben, der auf der Oberfläche bleibt, also sozusagen den Fluß *auf* der Oberfläche im Gegensatz zum Fluß *durch* die Oberfläche. Letzterer läßt sich berechnen als das Skalarprodukt $\vec{V}(\boldsymbol{x}) \cdot \vec{n}_x$ mit dem Normaleneinheitsvektor, und das Integral über diese Funktion ist

$$\iint_f \left(\vec{V} \cdot \vec{n}_{\boldsymbol{x}}\right) dO = \iint_f \vec{V} \cdot d\vec{O} \,,$$

denn

$$f_u \times f_v = |f_u \times f_v| \; \vec{n}_{\boldsymbol{x}} \,.$$

f) Die Sätze von Stokes und Gauß

In diesem Abschnitt geht es um drei der zentralsten Sätze der mehrdimensionalen Analysis: Außer den beiden im Titel erwähnten Sätzen soll noch das mehrdimensionale Analogon des Hauptsatzes der Differential- und Integralrechnung bewiesen werden.

Wir beginnen mit letzterem sowie dem Satz von STOKES: Bei beiden geht es darum, die Zirkulation eines Vektorfelds zu bestimmen. Beide Sätze gelten in beliebiger Dimension, jedoch wollen wir sie der Einfachheit halber nur für den $\mathbb{R}^3$ beweisen.

Ausgangspunkt ist ein reguläres Flächenstück $f\colon D \to \mathbb{R}^3$; dessen Bild $B = f(D)$, das „eigentliche“ geometrische Flächenstück also, sei beschränkt und habe eine reguläre Kurve $\gamma\colon [a,b] \to \mathbb{R}^3$ als Rand. Wie schon mehrfach erwähnt, wollen wir Kurvenintegrale längs γ mit Oberflächenintegralen über f oder – wie wir wegen der Parameterunabhängigkeit auch einigermaßen korrekt sagen können – B in Verbindung bringen.

Wir gehen also aus von einem Vektorfeld $\vec{V}\colon G \to \mathbb{R}^3$, dessen (offener) Definitionsbereich G sowohl B als auch γ enthält, und wollen Informa-

tionen über das Kurvenintegral

$$\int_\gamma \vec{V}\, ds \,.$$

Dazu approximieren wir B durch ein Flächenstück B^*, das wir aus endlich vielen ungefähr rechteckförmigen Teilmengen R_i zusammensetzen können; ein grobes Bild einer solchen Unterteilung ist in Abbildung 77 zu sehen.

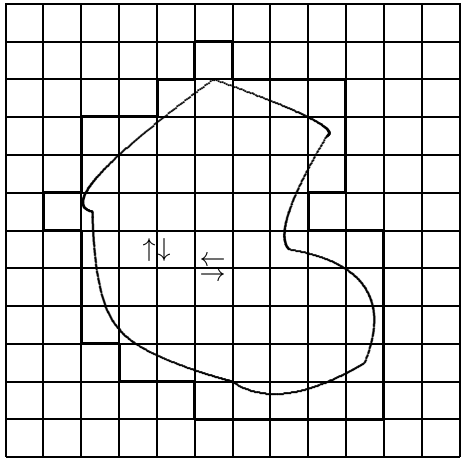


Abb. 77: Unterteilung eines Bereichs

Da f als reguläres Flächenstück vorausgesetzt war, ist überall in B ein wohldefinierter Normalenvektor bestimmt; damit können wir auch für jedes der Rechtecke einen Umlaufsinn festlegen: Dieser soll, wenn man von der Spitze des Normalenvektors aus auf des Rechteck schaut, der Gegenuhrzeigersinn sein. Da der Normalenvektor eines regulären Flächenstücks stetig von den Parametern der Fläche abhängt und nirgends verschwindet, ist es gleichgültig, über welchem Punkt eines (hinreichend kleinen) Rechtecks wir ihn betrachten.

In Abbildung 77 wurde davon ausgegangen, daß alle Normalenvektoren nach oben zeigen; der entsprechende Umlaufsinn ist für einige der

Rechteckkanten eingezeichnet. Wie man sieht, wird jede gemeinsame Kante zweier benachbarter Rechtecke von diesen beiden Rechtecken in verschiedener Weise orientiert. Falls wir daher für jedes Rechteck R_i das Wegintegral entlang seines Rands berechnen und alle diese Integrale aufaddieren, bleiben nur die Integrale längs der in Abbildung 77 fett ausgezogenen Kanten erhalten, d.h.

$$\sum_{i=1}^{M} \int\limits_{\partial R_i} \vec{V}\, ds = \int\limits_{\partial B^*} \vec{V}\, ds\,,$$

wobei M die Anzahl der „Rechtecke“ R_i bezeichnet. Natürlich ist erst recht für jede Funktion h auf B^*

$$\sum_{i=1}^{M} \iint\limits_{R_i} h\, dO = \iint\limits_{B^*} h\, dO\,,$$

wenn wir eine Formel für die einzelnen Rechtecke R_i haben, gilt diese also automatisch auch für B^* und somit, nach einem Grenzübergang, auch für B.

Betrachten wir also ein festes Rechteck R_i und das Kurvenintegral entlang seines Umfangs. Da wir die Rechtecke als klein voraussetzen, können wir $\vec{V}(\boldsymbol{x})$ auf R_i ohne großen Fehler linearisieren.

Dazu sei $\boldsymbol{x}_i \in R_i$ ein beliebiger Punkt des Rechtecks. Ein beliebiger Punkt $\boldsymbol{x} \in R_i$ hat dann die Form $\boldsymbol{x} = \boldsymbol{x}_i + \vec{h}$ mit einem nicht allzu großen Vektor $\vec{h}$. Nach Definition der Differenzierbarkeit ist dann

$$\begin{aligned}\vec{V}(\boldsymbol{x}) &= \vec{V}(\boldsymbol{x}_0 + \vec{h}) = \vec{V}(x_0) + J_{\vec{V}}(\boldsymbol{x}_0)\vec{h} + o\big(\|\vec{h}\|\big)\\ &= \vec{V}(\boldsymbol{x}_0) + S\cdot\vec{h} + \tfrac{1}{2}\operatorname{rot}\vec{V}(\boldsymbol{x}_0)\times h + o\big(\|\vec{h}\|\big)\,,\end{aligned}$$

wobei S für den symmetrischen Anteil der JACOBI-Matrix $J_{\vec{V}}(\boldsymbol{x}_0)$ steht.

Wenn wir den Term $o\big(\|\vec{h}\|\big)$ vernachlässigen, ist also

$$\begin{aligned}\int\limits_{\partial R_i} \vec{V}(\boldsymbol{x})\, ds &= \int\limits_{\partial R} \vec{V}(\boldsymbol{x}_0 + \vec{x})\, ds\\ &\approx \int\limits_{\partial R} \vec{V}(\boldsymbol{x}_0)\, ds + \int\limits_{\partial R} S\vec{x}\, ds + \tfrac{1}{2}\int\limits_{\partial R} \operatorname{rot}\vec{V}(\boldsymbol{x}_0)\times\vec{x}\, ds\,,\end{aligned} \qquad (*)$$

wobei R das um den negativ genommenen Ortsvektor von $\boldsymbol{x}_0$ verschobene Rechteck R_i sei, d.h., wenn $\boldsymbol{x}$ den Rand von R durchläuft, durchläuft $\boldsymbol{x}_0 + \boldsymbol{x}$ den Rand von R_i.

Das erste Integral in dieser Summe ist ein Kurvenintegral über das konstante Vektorfeld, das jedem Punkt den Vektor $\vec{V}(\boldsymbol{x}_0)$ zuordnet. Dieses Vektorfeld ist offensichtlich ein Potentialfeld, denn sind V_1, V_2 und V_3 die Komponenten dieses Vektors, so ist

$$\operatorname{grad}\big(\vec{V}(\boldsymbol{x}_0)\cdot\boldsymbol{x}\big) = \operatorname{grad}(V_1 x + V_2 y + V_3 z) = \vec{V}(x_0)\,.$$

Genauso ist auch $S\vec{h}$ ein Potentialfeld, denn der Gradient der quadratischen Form

$$\begin{aligned}{}^t\vec{x}S\vec{x} &= S_{11}x^2 + S_{22}y^2 + S_{33}z^2\\ &\quad + (S_{12}+S_{21})xy + (S_{13}+S_{31})xz + (S_{23}+S_{32})yz\end{aligned}$$

ist wegen $S_{k\ell} = S_{\ell k}$ gleich dem Zweifachen dieses Vektorfelds.

Damit sind also die beiden Vektorfelder $\vec{V}(x_0)$ und $S\vec{x}$ zirkulationsfrei, und die Integrale über den (als Kurve geschlossenen) Rand von R_i verschwinden.

Damit haben wir (bis auf die hier unterdrückten, für einen richtigen Beweis aber unbedingt notwendige Abschätzung der Fehlerterme) den folgenden Satz bewiesen:

Satz: Das Vektorfeld $\vec{V}\colon G \to \mathbb{R}^3$ habe eine symmetrische JACOBI-Matrix, d.h. $\operatorname{rot}\vec{V}$ sei identisch null. Dann gilt für jedes reguläre

Flächenstück $f: D \to G$, dessen Rand eine Kurve γ ist,

$$\int_\gamma \vec{V}\, ds = 0\,.$$

Denn nach obiger Rechnung verschwindet bei symmetrischer JACOBI-Matrix das Kurvenintegral entlang eines jeden Rechtecks, und die Summe all dieser Kurvenintegrale konvergiert bei immer feinerer Rechteckunterteilung des Flächenstücks gegen das Kurvenintegral längs des Rands γ von $B = f(D)$.

Dies reicht schon für den Hauptsatz der Differential- und Integralrechnung im $\mathbb{R}^3$: Wir wissen bereits, daß ein Vektorfeld genau dann eine Stammfunktion hat, wenn es zirkulationsfrei ist – die eine Richtung dieser Aussage haben wir gerade wieder angewandt. Außerdem wissen wir, daß für ein Vektorfeld mit Stammfunktion die Rotation verschwindet, denn für eine zweimal stetig differenzierbare Funktion φ ist $\operatorname{rot}\operatorname{grad}\varphi = 0$. Für ein leicht nachprüfbares Kriterium zur Existenz einer Stammfunktion fehlt also nur noch die Aussage, daß aus dem Verschwinden der Rotation die Zirkulationsfreiheit folgt.

Diese Aussage ist aber leider falsch: Das Beispiel des Magnetfelds eines stromdurchflossenen Leiters zeigte, daß die Rotation sehr wohl identisch verschwinden kann, ohne daß das Vektorfeld zirkulationsfrei ist. Dieses Magnetfeld $\vec{V}(x,y,z) = \frac{c}{x^2+y^2}\begin{pmatrix}-y\\ x\end{pmatrix}$ ist allerdings auf der z-Achse nicht definiert, und wir hatten im Beispiel Kreise um die z-Achse betrachtet.

Bei einem überall definiertes Vektorfeld hätten wir für einen solchen Kreis einfach die geschlossene Kreisscheibe B betrachten können, deren Rand der betrachtete Integrationsweg ist, und nach obigem Satz wäre das Integral längs des Randes verschwunden.

Das Problem bei diesem Gegenbeispiel liegt also offensichtlich darin, daß der betrachtete Integrationsweg nicht als Rand eines regulären Flächenstücks im Definitionsbereich des Vektorfelds geschrieben werden kann: Da das Vektorfeld auf der z-Achse nicht definiert ist, muß der Schnittpunkt mit der z-Achse aus der Kreisscheibe herausgenommen

werden, wir haben also nur noch einen punktierten Kreis, und dessen Rand besteht aus der Kreislinie *plus* dem herausgenommenen Punkt.

Um solche Fälle auszuschließen definieren wir:

Definition: Eine offene zusammenhängende Teilmenge $G \subseteq \mathbb{R}^3$ heißt *einfach zusammenhängend,* wenn jede geschlossene Kurve γ in G Rand eines regulären Flächenstücks ist. Anschaulich kann man dies auch so interpretieren, daß die Kurve innerhalb von G auf einen Punkt zusammengezogen werden kann: Man denke sich die Kurve als einen stark angespannten Gummiring; wenn man diesen auf ein Flächenstück legt, zieht er sich automatisch zusammen. Umgekehrt überstreicht der Ring beim Zusammenziehen auf einen Punkt ein Flächenstück, dessen Rand die Ausgangsposition des Gummis ist.

Mit dieser Definition gilt dann offensichtlich die folgende Verallgemeinerung des Hauptsatzes der Differential- und Integralrechnung auf den $\mathbb{R}^3$:

Satz: Ein differenzierbares Vektorfeld $\vec{V}: G \to \mathbb{R}^3$ auf einer einfach zusammenhängenden Teilmenge $G \subseteq \mathbb{R}^3$ hat genau dann eine Stammfunktion, wenn $\operatorname{rot}\vec{V}$ dort identisch verschwindet, d.h. also, wenn die JACOBI-Matrix symmetrisch ist.

■

Für ein ebenes Vektorfeld gilt entsprechend

Satz: Ein differenzierbares Vektorfeld $\vec{V}: G \to \mathbb{R}^2$ auf einer einfach zusammenhängenden Teilmenge $G \subseteq \mathbb{R}^2$ hat genau dann eine Stammfunktion, wenn

$$\frac{\partial V_1}{\partial y} - \frac{\partial V_2}{\partial x}$$

dort identisch verschwindet, d.h. also, wenn die JACOBI-Matrix symmetrisch ist.

Beweis: Dies folgt sofort aus dem Satz von GREEN.

■

Auf höhere Dimension soll nicht genauer eingegangen werden; nur soviel sei erwähnt: Die eindeutige Bestimmtheit des Normalenvektors

eines Flächenstücks im $\mathbb{R}^3$ hat in höheren Dimensionen keine Entsprechung mehr; dort muß man von einem Flächenstück explizit *fordern,* daß es orientierbar ist im folgenden Sinne: Man kann es durch kleine rechteckförmige Flächen überdecken und jedem dieser „Rechtecke" einen Umlaufsinn zuordnen derart, daß die gemeinsame Kante zweier Nachbarrechtecke von diesen beiden Rechtecken entgegengesetzt orientiert wird.

Wenn man dann den einfachen Zusammenhang einer offenen zusammenhängenden Teilmenge $G \subseteq \mathbb{R}^n$ dadurch definiert, daß jede geschlossene Kurve in G Rand eines so orientierbaren Flächenstücks sein soll, was auch wieder im wesentlichen äquivalent ist zur Zusammenziehbarkeit der Kurve, gilt auch hier der

Satz: Ein differenzierbares Vektorfeld $\vec{V}\colon G \to \mathbb{R}^n$ auf einer einfach zusammenhängenden Teilmenge $G \subseteq \mathbb{R}^n$ hat genau dann eine Stammfunktion, wenn die JACOBI-Matrix von $\vec{V}$ symmetrisch ist. ∎

Damit genug zum ersten der drei Hauptsätze dieses Paragraphen; wir machen weiter, wo wir vor der Spezialisierung auf symmetrische Vektorfelder aufgehört haben und folgern, daß nach $(*)$ für ein beliebiges differenzierbares Vektorfeld gilt

$$\int\limits_{\partial R_i} \vec{V}(\boldsymbol{x})\,ds \approx \tfrac{1}{2} \int\limits_{\partial R} \operatorname{rot} \vec{V}(\boldsymbol{x}_0) \times \vec{x}\,ds\,;$$

wir müssen uns also dieses Integral genauer ansehen. Dabei setzen wir zur Abkürzung

$$\vec{a} = \operatorname{rot} \vec{V}(x_0)\,;$$

dann ist der Integrand

$$\tfrac{1}{2}(\vec{a} \times \vec{x}) \cdot ds\,.$$

Nach Definition eines Kurvenintegrals ist das Integral hierüber Grenzwert einer Summe von Termen der Art

$$\tfrac{1}{2}(\vec{a} \times \vec{x}_j) \cdot \vec{t}_j\,,$$

wobei $\boldsymbol{x}_j$ Punkte auf der Kurve sind und t_j die Tangentenvektoren in diesen Punkten. Dieses Skalarprodukt eines Vektorprodukts mit einem

Vektor ist bekanntlich das *Spatprodukt,* und es ist gleich der Determinante mit den Spaltenvektoren $\vec{a}, \vec{x}_j$ und t_j. Diese Determinante ändert ihr Vorzeichen, wenn man zwei Spalten vertauscht; tut man dies zweimal, kehrt sie wieder zu ihrem alten Wert zurück. Daher ist

$$(\vec{a} \times \vec{x}_j) \cdot \vec{t}_j = (\vec{x}_j \times \vec{t}_j) \cdot \vec{a}\,.$$

Der Vektor $\vec{x}_j \times \vec{t}_j$ steht senkrecht auf $\vec{x}_j$ und auf $\vec{t}_j$; falls wir also von einem *flachen* Rechteck ausgehen, steht er auch senkrecht auf diesem und hat somit die Richtung des Normalenvektors. Sein Betrag ist gleich der Fläche des von $\vec{x}_j$ und $\vec{t}_j$ aufgespannten Parallelogramms, die Hälfte davon also gleich der Fläche des von diesen beiden Vektoren aufgespannten Dreiecks. Die Summe aller dieser Dreiecksflächen ist im Limes gleich der Fläche des Rechtecks, über dessen Rand wir integrieren, also ist das Integral über den Rand des Rechtecks für kleine Rechtecke näherungsweise gleich einem Skalarprodukt $\vec{n} \cdot \vec{a}$, wobei $\vec{n}$ als Länge den Flächeninhalt des Rechtecks hat und als Richtung die des Normalenvektors. Als Grenzwert einer Summe solcher Skalarprodukte ist aber gerade das Integral

$$\iint\limits_f \operatorname{rot} \vec{V}\,d\vec{O}$$

erklärt; wenn wir immer feinere Rechteckunterteilungen betrachten, konvergiert die Summe der Kurvenintegrale über die Ränder dieser Rechtecke also einerseits gegen das Kurvenintegral des Vektorfelds über den Rand γ des betrachteten Flächenstücks, andererseits aber auch gegen das Integral der Rotation von $\vec{V}$ über das Flächenstück selbst. Dies ist der

Satz von Stokes: $\vec{V}\colon G \to \mathbb{R}^3$ sei ein differenzierbares Vektorfeld auf der offenen Teilmenge $G \subseteq \mathbb{R}^3$. Weiter sei $f\colon D \to G$ ein reguläres Flächenstück mit einer stückweise regulären Kurve γ als Rand. Dann ist

$$\int\limits_\gamma \vec{V}\,ds = \iint\limits_f \operatorname{rot} \vec{V}\,d\vec{O}\,.$$

GEORGE GABRIEL STOKES (1819–1903) wurde in Irland geboren als jüngster von sechs Söhnen eines protestantischen Pfarrers. Nach dem Tod seines Vaters kam er im Alter von 16 Jahren an eine Schule nach Bristol in England und begann zwei Jahre später sein Studium an der Universität Cambridge. Er wurde vor allem bekannt durch seine Arbeiten zur mathematischen Physik, für die er viele mathematische Techniken entwickelte. Er gilt als der Begründer sowohl der Strömungslehre als auch der Geodäsie und hatte unter anderem großen Einfluß auf die Entwicklung von MAXWELL. Ab 1849 lehrte er als Professor an der Universität Cambridge.

Aus dem Satz von STOKES können wir eine neue Charakterisierung der Rotation herleiten:

Satz: $\vec{V}\colon G \to \mathbb{R}^3$ sei ein differenzierbares Vektorfeld, $\vec{e}$ ein Einheitsvektor, und $\gamma_{\vec{e},r}$ ein Kreis mit Radius r um $\boldsymbol{x}_0$ ist, der in einer Ebenen senkrecht zu $\vec{e}$ liege und von $\vec{e}$ aus gesehen im Gegenuhrzeigersinn durchlaufen werde; die gesamte von $\gamma_{\vec{e},r}$ berandete Kreisscheibe liege in G. Dann ist

$$\operatorname{rot}\vec{V}(\boldsymbol{x}_0) = \lim_{r\to 0}\frac{1}{\pi r^2}\int\limits_{\gamma_{\vec{e},r}} \vec{V}\,ds\,.$$

Beweis: $\gamma_{\vec{e},r}$ ist Rand einer Kreisscheibe D vom Radius r mit Flächeninhalt πr^2; nach dem Satz von STOKES ist

$$\int\limits_{\gamma_{\vec{e},r}} \vec{V}\,ds = \iint\limits_D \operatorname{rot}\vec{V}\,d\vec{O} = \iint\limits_D (\operatorname{rot}\vec{V})\cdot\vec{e}\,dO\,,$$

und für immer kleiner werdende Werte von r stimmt dies immer besser überein mit $\operatorname{rot}\vec{V}(\boldsymbol{x}_0)\cdot\vec{e}$ mal dem Flächeninhalt der Kreisscheibe. ∎

Falls die Voraussetzungen dieses Satzes erfüllt sind, falls also insbesondere die Kreislinie Rand einer im Definitionsbereich von $\vec{V}$ enthaltenen Fläche ist, läßt sich die Rotation somit als Limes einer Art „normierter" Zirkulation auffassen, d.h. als Integral längs einer geschlossenen Kurve dividiert durch die Länge dieser Kurve. Die Rotation gibt also auch ein Maß für die Abweichung eines Vektorfelds von der Zirkulationsfreiheit.

Damit wäre auch der zweite Hauptsatz dieses Abschnitts bewiesen; bleibt noch der

Satz von Gauß: $\vec{W}\colon G \to \mathbb{R}^3$ sei ein differenzierbares Vektorfeld und $f\colon D \to G$ sei ein reguläres Flächenstück derart, daß $B = f(D)$ Rand eines beschränkten dreidimensionalen Bereichs V sei. Dann ist

$$\iint\limits_f \vec{W}\,d\vec{O} = \iiint\limits_V \operatorname{div}\vec{W}\,dx\,dy\,dz\,.$$

Bemerkung: Der Satz gilt auch, mit den offensichtlichen Definitionen, falls der Rand von V kein Flächen*stück* ist, sondern eine Fläche, die aus endlich vielen regulären Stücken zusammengesetzt ist. Am Beweis ändert sich dabei abgesehen vom größeren Schreibaufwand praktisch nichts.

Der *Beweis* des Satzes von GAUSS beruht auf derselben Idee wie der des Satzes von STOKES: Dort hatten wir ein Flächenstück durch Rechtecke angenähert, um das Kurvenintegral längs seines Randes zu einem Integral über das Flächenstück in Beziehung zu setzen; hier unterteilen wir entsprechend das Volumen V in kleine, der Einfachheit halber als achsenparallel vorausgesetzte Quader, um das Oberflächenintegral über den Rand von V mit einem Integral über V in Beziehung zu setzen.

Ein solcher Quader sei Q_i, und $\vec{h},\vec{k},\vec{\ell}$ seien die drei Kantenvektoren von Q_i in Richtung der x-, y- und z-Achse. Dann ist das Integral von $\vec{W}$ über die Oberfläche von Q_i gleich

$$\iint\limits_{\partial Q_i} \vec{W}\,d\vec{O} = \underbrace{\iint \vec{W}\,d\vec{O}}_{\substack{\text{vordere und}\\ \text{hintere Seitenfläche}}} + \underbrace{\iint \vec{W}\,d\vec{O}}_{\substack{\text{linke und}\\ \text{rechte Seitenfläche}}} + \underbrace{\iint \vec{W}\,d\vec{O}}_{\substack{\text{obere und}\\ \text{untere Seitenfläche}}}\,,$$

und weiter ist beispielsweise

$$\begin{aligned}\iint\limits_{\substack{\text{linke und}\\ \text{rechte Seitenfläche}}} \vec{W}\,d\vec{O} &= \iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \vec{W}\,d\vec{O} + \iint\limits_{\substack{\text{rechte}\\ \text{Seitenfläche}}} \vec{W}\,d\vec{O}\\ &= \iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \vec{W}(\boldsymbol{x})\,d\vec{O} - \iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \vec{W}(\boldsymbol{x}+\vec{h})\,d\vec{O}\\ &= -\iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \big(\vec{W}(\boldsymbol{x}+\vec{h}) - \vec{W}(\boldsymbol{x})\big)\,d\vec{O}\,,\end{aligned}$$

denn da alle Normalenvektoren nach außen zeigen, sind die linke und die rechte Seite des Quaders verschieden orientiert.

Nun kommt die Differenzierbarkeit von $\vec{W}$ ins Spiel und sagt uns, daß

$$\vec{W}(\boldsymbol{x}+\vec{h}) = \vec{W}(\boldsymbol{x}) + J_{\vec{W}}(\boldsymbol{x})\cdot\vec{h} + o\big(\|\vec{h}\|\big)$$

ist. Da der Vektor $\vec{h}$ in Richtung der x-Achse zeigt, können wir das (mit $h = \left|\vec{h}\right|$) auch einfacher schreiben als

$$\vec{W}(x+h,y,z) = \vec{W}(x,y,z) + \left(\frac{\partial}{\partial x}\vec{W}\right)\cdot h + o(h)\,,$$

wobei die partielle Ableitung von $\vec{W}$ nach x für jenes Vektorfeld stehen soll, dessen Komponenten die partiellen Ableitungen der entsprechenden Komponenten von $\vec{W}$ sind. Da wir in diesem Semester noch viel zu tun haben, vernachlässigen wir den Term $o(h)$ in der Hoffnung, daß er für immer feiner werdende Unterteilungen trotz der dann ansteigenden *Anzahl* dieser Terme keine Rolle mehr spielen wird.

In der Tat sollte ein mit den Abschätzungstechniken aus der Analysis I vertrauter Leser keine Schwierigkeiten haben, dies mathematisch streng zu zeigen. Die Heuristik, nach der man dabei vorgeht, ist (hoffentlich) klar: Die Anzahl aufeinanderliegender Quader in Richtung der x-Achse ist proportional zu $1/\,|\vec{h}|$, und $1/\,|\vec{h}|$ Terme der Größenordnung $o(|\vec{h}|$ addieren sich zu einem Term der Größenordnung $o(1)$, d.h. zu einer Funktion, die gegen Null geht. (Die Aussage „geht schneller als 1 gegen Null“ ist natürlich gleichbedeutend damit, daß die Funktion überhaupt gegen Null geht.)

Der Einheitsnormalenvektor der linken Seite des Quaders zeigt in Richtung der negativen x-Achse, denn die linke Seite ist parallel zur (y,z)-Ebene und zeigt nach außen. Wenn wir den Einheitsvektor der x-Achse mit $\vec{e}_x$ bezeichnen, ist daher

$$d\vec{O} = -\vec{e}_x\,dy\,dz$$

und damit

$$\begin{aligned}\iint\limits_{\substack{\text{linke und}\\ \text{rechte Seitenfläche}}} \vec{W}\,d\vec{O} &\approx -\iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \big(\vec{W}(\boldsymbol{x}+\vec{h}) - \vec{W}(\boldsymbol{x})\big)\,d\vec{O}\,,\\ &= \iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} h\cdot\frac{\partial}{\partial x}\vec{W}\cdot\vec{e}_x = h\cdot\iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \begin{pmatrix}\frac{\partial V_1}{\partial x}\\ \frac{\partial V_1}{\partial y}\\ \frac{\partial V_1}{\partial z}\end{pmatrix}\cdot\begin{pmatrix}1\\0\\0\end{pmatrix}\,dx\,dy\\ &= h\iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \frac{\partial V_1}{\partial x}\,dy\,dz\,.\end{aligned}$$

Ganz entsprechend ist

$$\iint\limits_{\substack{\text{vordere und}\\ \text{hintere Seitenfläche}}} \vec{W}\,d\vec{O} \approx k\iint\limits_{\substack{\text{vordere}\\ \text{Seitenfläche}}} \frac{\partial V_2}{\partial y}\,dx\,dz$$

und

$$\iint\limits_{\substack{\text{obere und}\\ \text{untere Seitenfläche}}} \vec{W}\,d\vec{O} \approx \ell\iint\limits_{\substack{\text{obere}\\ \text{Seitenfläche}}} \frac{\partial V_3}{\partial z}\,dx\,dy\,.$$

Nach dem Mittelwertsatz für Flächenintegrale können wir diese Integrale weiter abschätzen: Es gibt Punkte ξ auf der linken, η auf der vorderen und ζ auf der unteren Seitenfläche, so daß

$$\begin{aligned}\iint\limits_{\substack{\text{linke und}\\ \text{rechte Seitenfläche}}} \vec{W}\,d\vec{O} &\approx h\iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} \frac{\partial V_1}{\partial x}\,dy\,dz\\ &= h\frac{\partial V_1}{\partial x}(\xi)\iint\limits_{\substack{\text{linke}\\ \text{Seitenfläche}}} dy\,dy = hkl\frac{\partial V_1}{\partial x}(\xi)\,.\end{aligned}$$

Genauso ist

$$\iint\limits_{\substack{\text{vordere und}\\ \text{hintere Seitenfläche}}} \vec{W}\, d\vec{O} \approx hkl \frac{\partial V_2}{\partial y}(\eta)$$

und

$$\iint\limits_{\substack{\text{obere und}\\ \text{untere Seitenfläche}}} \vec{W}\, d\vec{O} \approx hkl \frac{\partial V_3}{\partial z}(\zeta)\,.$$

Nun ist es wieder an der Zeit, eine etwas komplexere Abschätzung unter den Tisch fallen zu lassen: Die Punkte ξ, η und ζ liegen auf den Seitenflächen eines Quaders, der immer kleiner werden soll. Damit rücken auch diese Punkte immer weiter zusammen und wir sollten wohl keinen allzu großen Fehler machen, wenn wir einfach *irgendeinen* Punkt $\boldsymbol{x}_i$ im Quader Q_i auswählen und sowohl ξ, η als auch ζ durch diesen Punkt ersetzen; die formale Rechtfertigung hierfür geht wieder aus von der Differenzierbarkeit des Vektorfelds V und beruht im übrigen auf Abschätzungen.

Wenn wir das alles glauben, ist also

$$\iint\limits_{\substack{\text{linke und}\\ \text{rechte Seitenfläche}}} \vec{W}\, d\vec{O} \approx hkl \frac{\partial V_1}{\partial x}(\boldsymbol{x}_i)\,,$$

$$\iint\limits_{\substack{\text{vordere und}\\ \text{hintere Seitenfläche}}} \vec{W}\, d\vec{O} \approx hkl \frac{\partial V_2}{\partial y}(\boldsymbol{x}_i) \quad \text{und}$$

$$\iint\limits_{\substack{\text{obere und}\\ \text{untere Seitenfläche}}} \vec{W}\, d\vec{O} \approx hkl \frac{\partial V_3}{\partial z}(\boldsymbol{x}_i)\,.$$

Das Oberflächenintegral über die gesamte Quaderoberfläche ist die Summe dieser drei Teilintegrale, und da bekanntlich

$$\frac{\partial V_1}{\partial x}(\boldsymbol{x}_i) + \frac{\partial V_2}{\partial y}(\boldsymbol{x}_i) + \frac{\partial V_3}{\partial z}(\boldsymbol{x}_i) = \operatorname{div} \vec{W}(\boldsymbol{x}_i)$$

ist, haben wir somit gezeigt, daß

$$\iint\limits_{\partial Q_i} \vec{W}\, d\vec{O} \approx hkl\ \operatorname{div} \vec{W}(\boldsymbol{x}_i) = \operatorname{div} \vec{W}(\boldsymbol{x}_i) \cdot \operatorname{Vol}(Q_i) \qquad (*)$$

ist.

Damit haben wir eine lokale Version des Satzes von GAUSS gezeigt; zum Beweis des Satzes selbst nähern wir das Volumen V an durch die Quader Q_i. Dann ist

$$\iiint\limits_V \operatorname{div} \vec{W}\, dx\, dy\, dz \approx \sum_i \operatorname{div} \vec{W}(\boldsymbol{x}_i) \cdot \operatorname{Vol}(Q_i)\,,$$

und wenn wir die Quader immer weiter verkleinern, wird im Limes aus dem Ungefährzeichen ein Gleichheitszeichen – genau so hatten wir schließlich Volumenintegrale definiert.

Im Falle des Integrals über die Randfläche $B = f(D)$ von V ist die Situation etwas komplizierter: Wie beim Beweis des Satzes von STOKES überlegt man sich zunächst leicht, daß eine gemeinsame Seitenfläche zweier benachbarter Quader von den beiden Quadern verschieden orientiert wird, so daß sich die beiden Integrale über diese Fläche gegenseitig aufheben; die Summe über alle Oberflächenintegrale über die Q_i ist also gleich der Summe über alle Oberflächenintegrale über jene Seitenflächen von Quadern Q_i, die nur Seitenfläche eines einzigen Quaders sind, d.h. also über die Randfläche der Vereinigung aller Quader Q_i.

Die Quaderflächen, die diesen Rand ausmachen, sind allesamt parallel zu Koordinatenebenen; ihre Normalenvektoren sind also parallel zu Koordinaten*achsen,* wohingegen die Normalenvektoren von B natürlich beliebige Richtungen haben können. Wir müssen uns überlegen, warum die Integrale im Limes trotzdem gleich sein können.

Der Integrand bei einem Oberflächenintegral über ein Vektorfeld ist das Skalarprodukt aus dem Normalenvektor von B im jeweils betrachteten Punkt und dem Wert $\vec{W}(\boldsymbol{x})$ des Vektorfelds in diesem Punkt. Konkret sei etwa $\vec{n}$ der Normalenvektor im Punkt $\boldsymbol{x} \in B$, und in Koordinaten sei

$$\vec{n} = \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} \quad \text{und} \quad \vec{W}(\boldsymbol{x}) = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}\,.$$

Dann ist

$$\begin{aligned} \vec{W}(\boldsymbol{x}) \cdot \vec{n} &= v_1 n_1 + v_2 n_2 + v_3 n_3 \\ &= \vec{W}(\boldsymbol{x}) \cdot (n_1 \vec{e}_1) + \vec{W}(\boldsymbol{x}) \cdot (n_2 \vec{e}_2) + \vec{W}(\boldsymbol{x}) \cdot (n_3 \vec{e}_3)\,; \end{aligned}$$

wir erhalten also dasselbe Skalarprodukt auch, indem wir $\vec{W}(\boldsymbol{x})$ nacheinander mit geeigneten Normalenvektoren von Quaderflächen parallel zu den Koordinatenebenen multiplizieren und die Ergebnisse aufaddieren.

Falls $\vec{n}$ drei nichtverschwindende Komponenten hat, ist klar, daß die Fläche B in der Nähe von $\boldsymbol{x}$ nur so durch Quader angenähert werden kann, daß es dort freie Randflächen parallel zu allen drei Koordinatenebenen gibt; ein nicht allzu schwieriges Argument über den Satz von PYTHAGORAS zusammen mit Limesbetrachtungen zeigt, daß auch mit den Längen alles gut geht, so daß auch das Oberflächenintegral über B gleich der Summe der Integrale über die Quaderoberflächen ist.

Wie wir oben gesehen haben, ist der Satz für einen einzelnen Quader richtig; da auf beiden Seiten das Integral durch Summen entsprechender Integrale für Quader angehähert werden kann, ist also der Satz von GAUSS (modulo zahlreicher Auslassungen) bewiesen. ∎

Genau wie der Satz von STOKES zu einer alternativen Definition der Rotation führte, kann mit dem Satz von GAUSS die *Divergenz* auf andere Weise ausgedrückt werden; zumindest für Quader haben wir die entsprechende Formel im gerade beendeten Beweis als Formel $(*)$ bereits hergeleitet:

Satz: Für ein differenzierbares Vektorfeld $\vec{W}\colon G \to \mathbb{R}^3$ ist

$$\operatorname{div}\vec{W}(\boldsymbol{x}_0) = \lim_{r\to 0} \frac{3}{4\pi r^3} \iint\limits_{\partial B_r} \vec{W}\, d\vec{O}\,,$$

wobei B_r eine Kugel mit Radius r um $\boldsymbol{x}_0$ ist, deren Normalenvektoren nach außen zeigen.

Beweis: Nach dem Mittelwertsatz ist das Volumenintegral über B_r gleich dem Produkt des Volumens $\frac{4}{3}\pi r^3$ von B_r mit dem Wert des Integranden $\operatorname{div}\vec{W}$ an einem geeigneten Punkt der Kugel. Falls r gegen Null geht, muß dieser Punkt immer näher an den Mittelpunkt der Kugel rücken, bis er im Limes mit diesem zusammenfällt. ∎

Auch der Satz von GAUSS läßt sich wieder anschaulich interpretieren: Bei der Definition der Divergenz haben wir uns bereits überlegt, daß diese mißt, inwieweit ein Punkt eher eine Quelle oder eine Senke für ein Vektorfeld ist. Da das Oberflächenintegral gerade gleich dem Fluß des Vektorfelds durch die betrachtete Oberfläche ist und wir die Oberfläche so orientiert haben, daß der Normalenvektor nach außen zeigt, sagt der Satz von GAUSS also einfach, daß der gesamte Fluß einer Vektorfelds durch die Oberfläche eines Volumens V genau das ist, was im Innern von V erzeugt oder (bei negativen Vorzeichen auf beiden Seiten) vernichtet wird.

Der obige Satz zur Charakterisierung der Divergenz zeigt dementsprechend noch einmal, warum die Divergenz auch als *Quellendichte* bezeichnet wird: Nach dem Satz von GAUSS ist der Fluß durch die Oberfläche einer Kugel gleich der Summe des im Innern erzeugten oder vernichteten; dividiert man dies durch das Volumen der Kugel, entsteht die Quellen*dichte,* deren Grenzwert für immer kleiner werdende Kugeln gleich der Divergenz ist.

$\mathcal{E}\ \mathcal{N}\ \mathcal{D}\ \mathcal{E}$

$\mathcal{S}\ \mathcal{C}\ \mathcal{H}\ \ddot{\mathcal{O}}\ \mathcal{N}\ \mathcal{E}\quad \mathcal{F}\ \mathcal{E}\ \mathcal{R}\ \mathcal{I}\ \mathcal{E}\ \mathcal{N}\ !$