MafIA:

Mathematik für Informatik–Anfänger

Stand:

24. Juli 2008

Vorwort

Dieses Manuskript ist für die TeilnehmerInnen der Vorlesung

MafIA: "Mathematik für Informatik(-Anfäng)er"

an der Universität Göttingen gedacht. Es basiert auf einem älteren und unvollständigen Text aus dem Studienjahr 2003/2004, und es wird im Studienjahr 2007/2008 an dieser Stelle gründlich durchgearbeitet, ergänzt und korrigiert. Insbesondere sollen mehr Beispiele und Bilder eingebaut werden (Dank an Anna Eggers, die etliche angefertigt hat), und viele Links und die Querverbindungen zu MuPAD und MATLAB nachgeliefert werden. Ferner ist eine gewisse Aufteilung zwischen dem Stoff der Vorlesung und dem der parallelen Saalübung geplant.

Das Ganze ist, wie jede website, "*under construction*". Ich bitte alle Studierenden, mich per e–mail auf Fehler, Ungenauigkeiten und Unvollständigkeiten hinzuweisen, und Tim Rohlfs danke ich für sein sehr gründliches Korrekturlesen.

Soweit zur Entstehung und zum Hörerkreis. Aber es sollte auch noch etwas zum Inhalt der Vorlesung und zur den Auswahlprinzipien für den Stoff gesagt werden.

Mathematik ist zwar auch für viele andere Disziplinen wichtig, aber für die Informatik ist sie unerläßlich. Dazu gibt es vom Altmeister Prof. Dr. Dr. h.c. mult. F.L Bauer einen schönen Artikel[1], aber es sind noch einige Argumente hinzuzufügen.

Mathematik ist die einzige Wissenschaft, in der man seiner Aussagen in einem gewissen Sinn **sicher** sein kann, weil man sie unwiderleglich **bewiesen** hat. Deshalb sind alle **Sicherheitsaspekte** und **Korrektheitsfragen** in der Informatik notwendig mit Mathematik verbunden. Das betrifft diverse Teildisziplinen der Informatik, u.a. die **Kryptographie** und das **Software–Engineering**.

Und das in den Übungen zu Mathematikvorlesungen erlernte unwiderlegliche **Beweisen** ist eine oft unterschätzte **Schlüsselqualifikation**, denn man lernt, das Wesentliche vom Unwesentlichen zu unterscheiden, eine Argumentationslinie sauber aufzubauen und alle Einwände unmöglich zu machen. Das ist auch außerhalb der Mathematik extrem nützlich, z.B. wenn man einen

[1] `http://www.num.math.uni-goettingen.de/schaback/teaching/texte/MafIA/bauer_kr.html`

Kunden, den Unternehmensvorstand oder ein Gericht von etwas überzeugen will. Deshalb wird in Mathematikveranstaltungen darauf bestanden, daß die Studierenden mündlich und schriftlich in der Lage sind, einwandfreie Beweise zu formulieren. Daß dies nebenbei die Studierenden der Informatik dazu erzieht, möglichst fehlerfreie Algorithmen zu entwerfen, dürfte klar sein.

Weniger klar ist hingegen, daß man das Erlernen von Mathematik nicht durch reine Faktenvermittlung erreichen kann. Wie beim Lernen des Klavierspiels, einer Fremdsprache oder einer Programmiersprache reicht es nicht, eine noch so gut geschriebene Anleitung zu lesen. Man muss eine Fremdsprache regelmäßig sprechen, in einer Programmiersprache eine Folge immer komplizierterer Programme schreiben und am Klavier täglich üben, sonst wird es nichts. In diesem Sinne muß man auch Mathematik immer wieder üben, und das geschieht im begleitenden **Übungsbetrieb**. Dieser ist mindestens so wichtig wie die Vorlesung oder ein begleitendes Buch, und das Erfolgskriterium einer Mathematikveranstaltung ist aus gutem Grund nicht ein gutes Faktenwissen allein, sondern der Nachweis, mit der Mathematik praktisch umgehen zu können.

Deshalb wird dieser Text in seiner Endform auch viel Material zum Üben enthalten. Dieses ist an bestimmten Stellen eingestreut. **Fragen** sollten gleich beim Lesen schon beantwortet werden können, **Aufgaben** erfordern etwas Nachdenken und in der Regel auch Papier und Bleistift, sind aber trotzdem im laufenden Text enthalten. **Übungen** werden in späteren Textversionen getrennt aufgelistet sein, und an verschiedenen Stellen wird es praktische **Anleitungen** geben, etwa zum sauberen Formulieren von Beweisen oder zum Umgang mit begleitender Software, z.B. MATLAB© oder MuPad©.

Insgesamt ist der Inhalt durch die begrenzte Vorlesungszeit auf das unbedingt Nötige eingeschränkt, und es wird an verschiedenen Stellen darauf hingewiesen, welche Gebiete der Mathematik in welchen Gebieten der Informatik zur Anwendung kommen. Die Tabelle 1 bringt eine Liste mit keineswegs vollständigen Beispielen. Differenzial– und Integralrechnung sind nicht in dieser Liste, aber sie sind unerläßliche Hilfsmittel für verschiedene der explizit genannten Gebiete, z.B. für die Fouriertransformation und die digitale Signalverarbeitung. Ebenso ist die Lineare Algebra nicht nur wichtig in direkten Anwendungen, sondern sie liefert Methoden, Geometrie algorithmisch zu betreiben und in der Computergraphik anzuwenden. Viele Informatikdisziplinen setzen Kenntnisse aus **mehreren** mathematischen Gebieten auf einmal voraus. Beispielsweise erfordert das zur Zeit sehr modische **maschinelle**

Mathematik	Informatik–Anwendung
Relationen	relationale Datenbanken
Logik	Schaltlogik
	regelbasierte Verfahren
	maschinelles Beweisen
Zahlen	Rechnen mit Gleitkommazahlen
	Kryptosysteme
Lineare Algebra	Modellierung
	Data Mining
Geometrie	Computergraphik
	Computer–Aided Design
Folgen und Reihen	Komplexitätstheorie
	Analyse von Algorithmen
Vektoranalysis	Modellierung von Strömungsvorgängen
Fouriertransformation	Signalverarbeitung
	Kompressionsverfahren wie JPEG und MPEG

Tabelle 1: Mathematikdisziplinen und ihre Informatik–Anwendungen

Lernen[1] nicht nur die Differenzial– und Integralrechnung, sondern auch die Lineare Algebra und ein gerütteltes Maß an Stochastik.

Deshalb darf man nicht erwarten, daß diese Vorlesung die in der Informatik nötige Mathematik komplett abdeckt. Dazu wäre ein Vielfaches an Aufwand nötig. Es ist aber möglich, den Studierenden die wichtigsten Anfangsgründe beizubringen und sie in die Lage zu versetzen, von hier aus andere mathematische Disziplinen, soweit sie in späteren Studienrichtungen nötig werden, sich ohne grundlegende Probleme zu erarbeiten.

Die **Diskrete Mathematik** und die **Stochastik** werden parallel bzw in einer nachfolgenden Vorlesung gelehrt. Deshalb werden hier die Querverbindungen zu diesen Vorlesungen und zur Grundausbildung in Informatik nur in Form von Verweisen behandelt.

Gegenüber der älteren Version des Skriptes wurden einige neue Lehrbücher [2, 6, 7, 9, 8, 3] in das Literaturverzeichnis aufgenommen. Meinen eigenen Vorstellungen kommen die Bände [6, 7] von Gerald und Susanne Teschl am nächsten. Das Buch [4] von P. Hartmann enthält deutlich weniger Stoff (es hat ja auch nur einen Band), zeichnet sich aber durch viele Beispiele

[1] `http://www.kernel-machines.org`

und eine besondere Leserfreundlichkeit aus. In beiden Werken sind Diskrete Mathematik und Stochastik mit enthalten.

R. Schaback, 24. Juli 2008

Inhaltsverzeichnis

1 Mengen und Abbildungen

Wo beginnt die Mathematik? Sie setzt strukturiertes Denken voraus, und diese Disziplin nennt man **Logik**. Die Aufstellung oder Beschreibung der Struktur der Logik erfordert aber selbst wieder ein strukturiertes Denken. Analysiert man dieses Dilemma etwas genauer, so stellt sich heraus, daß man für eine saubere Darstellung der mathematischen Logik die logischen Begriffe der Mengenlehre braucht, für die Mengenlehre aber wiederum die Logik.

Aus diesem rekursiven Dilemma kommt man nur heraus, wenn man erst einmal ganz naiv und unstrukturiert sowohl Mengenlehre als auch Logik behandelt, um danach den Boden des naiven Wissens zu verlassen und in einem zweiten Durchgang sowohl die Logik als auch die Mengenlehre sauber zu strukturieren. Der naive Zugang kann sowohl mit Mengenlehre als auch mit Logik begonnen werden (vgl. [4] und [1] für zwei verschiedene Zugänge). Wir beginnen hier mit Mengenlehre, lassen die Logik folgen und holen die saubere Strukturierung im Abschnitt 2.4 nach.

1.1 Mengenlehre

1.1.1 Grundbegriffe

Definition 1.1 *Eine* **Menge** *(im Sinne der "naiven" Mengenlehre[1]) ist eine beliebige Zusammenfassung von bestimmten wohlunterschiedenen Objekten unserer Anschauung oder unseres Denkens zu einem Ganzen (nach* **Cantor**[2]*. Die Objekte heißen* **Elemente** *der Menge.*

Man kann Mengen durch **Aufzählung** konstruieren:

$$\{1, 3, 7\} \text{ hat die Elemente 1, 3 und 7} \tag{1.2}$$

oder durch Angabe einer **Eigenschaft**, die alle Elemente haben sollen:

$$\{x \ : \ x \text{ hat die Eigenschaft } E\}. \tag{1.3}$$

Diese Methode ist sehr naiv und muß später etwas genauer formuliert werden (z.B. ist "x hat die Eigenschaft E" eine Aussage und setzt deshalb die Aussagenlogik voraus). Obendrein führt sie auf Widersprüche, aber auch das werden wir jetzt noch nicht untersuchen.

[1] `http://de.wikipedia.org/wiki/Mengenlehre`

[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Cantor.html`

Definition 1.4 *Ist x ein Element einer Menge M, so schreibt man $x \in M$ und sagt auch "x liegt in M".*
Ist x nicht Element einer Menge M, so schreibt man $x \notin M$.
Die **leere Menge** *wird mit $\emptyset$ bezeichnet. Sie hat keine Elemente.*

Genaugenommen haben die Definitionen (1.2) und (1.3) nur gemeinsam mit Definition 1.4 einen Sinn, denn die Mengendefinition (1.2) für eine Menge

$$M := \{a, b, c, \ldots\} \tag{1.5}$$

durch Aufzählung ist eine Kurzform für

$$\text{es gilt } a \in M \text{ und } b \in M, \text{ und } c \in M \text{ usw.}$$

während (1.3) die Bedeutung

$$\text{für alle } x \text{ gilt } x \in M \text{ genau dann, wenn } x \text{ die Eigenschaft } E \text{ hat.}$$

Man mache sich klar, daß die umgangssprachliche Verwendung von "ist Element von" und "hat Elemente" zu vage ist und durch die formalere Schreibweise $x \in M$ abstrahiert wird. Das werden wir uns noch genauer ansehen, wenn wir $\in$ als Relation verstehen, aber wir werden auf unserem Weg hin zu einer saubereren Formulierung jetzt weitgehend das Wort "Element" vermeiden und von $\in$ reden.

Wir verwenden die unsymmetrische Notation $:=$ (sprich: "ist definiert als") für Definitionen wie in (1.5), wenn wir etwa durch $x :=$`A` die Bedeutung des Symbols x durch einen Ausdruck `A` festlegen wollen. Das hat natürlich gar nichts mit der **Wertzuweisung** in **PASCAL** zu tun.

Aus der Interpretation der Klammerschreibweise in (1.5) folgt sofort, daß die Mengen $\{1, 3, 7\}$ und $\{7, 1, 3\}$ gleich sind.

Weitere Beispiele:

$\{0, 1, 2, 3, 5, 7, 11\}$

$\{x \ : \ x \text{ ist eine ganze Zahl und durch 2 teilbar }\}$

$\{x \ : \ x \text{ ist ein roter Hering}\}$

Hier noch ein paar Klarstellungen. Die leere Menge $\emptyset$ hat keine Elemente, und das kann man als

für alle x ist die Aussage $x \in \emptyset$ immer falsch

beschreiben. Dagegen hat die Menge $\{\emptyset\}$ per Definition ein Element, nämlich $\emptyset$. Es können also durchaus Mengen gleichzeitig Element von etwas sein.

Frage: Was sind die Elemente der Menge $\{a, \{b, c\}, \emptyset\}$?

Definition 1.6 *Es seien M und N Mengen.*

1. *Man sagt, M sei in N* **enthalten** *oder sei eine* **Teilmenge** *von N und schreibt $M \subseteq N$ oder $N \supseteq M$, wenn jedes Element x von M auch Element von N ist, d.h. wenn für alle x aus der Aussage $x \in M$ immer die Aussage $x \in N$ folgt.*

2. *Man sagt, M und N seien* **gleich** *und schreibt $M = N$, wenn sie dieselben Elemente haben. Dies ist gleichbedeutend mit der Aussage, daß die Aussagen $M \subseteq N$ und $N \subseteq M$ beide zutreffen.*

In manchen Büchern wird statt $\subseteq$ auch $\subset$ geschrieben. Wegen der Analogie zu dem Symbol $\leq$ für "kleiner oder gleich" zwischen Zahlen ziehe ich $\subseteq$ vor.

Die Definition 1.6 verwendet den Begriff "aus Aussage A folgt Aussage B". Das ist natürlich wieder Aussagenlogik, aber die haben wir noch nicht behandelt.

Wichtige Mengen sind

$$\begin{array}{lcl} \mathbb{N} & := & \{0, 1, 2, 3, \ldots\} \\ \mathbb{Z} & := & \{0, +1, -1, 2, -2, \ldots\} \\ \mathbb{R} & := & \text{reelle Zahlen } = \{\text{infinite Dezimalbrüche mit Vorzeichen}\} \end{array}$$

aber man sollte nach Möglichkeit die "Pünktchen–Notation" und unklare Begriffe wie "infiniter Dezimalbruch" vermeiden. Wir werden das später besser machen. Die reellen Zahlen sollen hier erst einmal so wie in der Schule verstanden werden.

Frage: Wieviel Elemente hat die Menge $\{\mathbb{N}, \mathbb{Z}\}$?

In der Informatik gibt es den Begriff der (einfachen) **Datentypen**. Sie sind definiert als Mengen, nämlich als Mengen von Werten. Aber das soll in der Informatikvorlesung gelehrt werden.

1.1.2 Exkurs: Was heißt "gleich"?

Bei Definition 1.6 liegt ein erster Fall einer "**Gleichheit**" vor, die durch das Zeichen "=" ausgedrückt wird. Damit muß man sehr vorsichtig sein, besonders als Informatiker, denn viele Programmiersprachen benutzen Zeichenfolgen wie

$$x = c * (a + b);$$

in ganz anderer Bedeutung als einer "Gleichheit" der linken und rechten Seite. Immerhin gibt es in der **Informatik** inzwischen auch "==" als Symbol für den Test auf Gleichheit der linken und rechten Seite, aber es ist z.B. fraglich, ob damit **Referenzgleichheit** oder **Wertgleichheit** gemeint ist (diese Begriffe werden in der Informatikvorlesung erklärt). Alle Studierenden der Informatik sollten nervös werden, wenn man ohne klare Definition von "Gleichheit" redet. Auch im deutschen Sprachgebrauch sind "das gleiche" und "dasselbe" eben nicht dasselbe, und man würde in Teufels Küche kommen, wenn man für beides dieselbe Notation verwenden würde.

1.1.3 Potenzmenge

Es gibt auch Mengen von Mengen:

Definition 1.7 *Die* **Potenzmenge**[1] *einer Menge M besteht aus allen Teilmengen von M und wird mit $P(M)$ oder manchmal auch mit $Pot(M)$ bezeichnet.*

Man mache sich klar, daß für alle N die Aussagen $N \in P(M)$ und $N \subseteq M$ gleichbedeutend sind. Außerdem mag es Anfänger verwirren, daß $P(\emptyset) \neq \emptyset$ gilt.

Frage: Warum ist das so?

1.1.4 Grundregeln des formellen Beweisens

Dies ist ein Exkurs, der in die parallele Saalübung gehört. Der Vorlesungstext geht mit Abschnitt 1.8 weiter.

Einen Beweis für $M \subseteq N$ führt man im allgemeinen so:

1. Man nimmt sich ein beliebiges Element von M und nennt es x. Wenn es kein solches gibt, ist M gleich der leeren Menge $\emptyset$ und man hat nichts mehr zu beweisen. Dies bedarf keiner besonderen Erwähnung, denn die Definition von $M \subseteq N$ ist so gemacht, daß man nur für Elemente von M etwas beweisen muß. Es gilt also immer $\emptyset \subseteq N$ für alle Mengen N.

[1] `http://de.wikipedia.org/wiki/Potenzmenge`

2. Dann argumentiert man für dieses x so lange, bis man die angestrebte Aussage $x \in N$ bekommt. Dazu verwendet man alles, was man über M und N weiß.

Beispiel: Man beweise die Behauptung, daß für drei Mengen L, M und N aus $L \subseteq M$ und $M \subseteq N$ immer auch $L \subseteq N$ folgt. Bei naiver Argumentation würde man einfach folgendes sagen:

Alles, was in L liegt, liegt in M.
Alles, was in M liegt, liegt in N.
Also liegt alles, was in L liegt, auch in N.

Das ist inhaltlich richtig, aber eher der Sprechweise eines Juristen und nicht der einer Mathematikerin oder eines Informatikers oder eines Computers angemessen.
Wir müssen früh üben, so etwas ganz formal aufzuschreiben. Wie geht man vor? Man schreibt sich erst hin, was man weiß:

1. $L \subseteq M$, d.h. für alle x gilt, daß aus $x \in L$ immer $x \in M$ folgt.

2. $M \subseteq N$, d.h. für alle x gilt, daß aus $x \in M$ immer $x \in N$ folgt.

Was will man zeigen?
Für alle x gilt, daß aus $x \in L$ immer $x \in N$ folgt. Man kann sich also ein beliebiges x mit $x \in L$ hernehmen. Dann kann man 1. benutzen, um auf $x \in M$ zu schließen. Danach benutzt man 2. um auf $x \in N$ zu kommen. Fertig.

Wie schreibt man so einen Beweis sauber auf?
Man beginnt mit der genauen Formulierung der

> **Behauptung**: Sind L, M, N beliebige Mengen und gilt $L \subseteq M$ und $M \subseteq N$, so gilt auch $L \subseteq N$.

Dabei darf man keine unerklärten Symbole verwenden, d.h. man darf nicht weglassen, daß L, M und N Mengen sein sollen.
Dann schreibt man hin, was man weiß:

> **Voraussetzungen**: L, M, N sind Mengen. Ferner gilt
>
> 1. $L \subseteq M$, d.h. für alle x gilt, daß aus $x \in L$ immer $x \in M$ folgt.
> 2. $M \subseteq N$, d.h. für alle x gilt, daß aus $x \in M$ immer $x \in N$ folgt.

Das haben wir oben schon gut gemacht. Dann formuliert man das Ziel genauer:

Zu zeigen ist: $L \subseteq N$, d.h.

zu zeigen ist: Für alle x gilt, daß aus $x \in L$ immer $x \in N$ folgt.

Jetzt kann man die Argumentation durchführen:

Sei $x \in L$ beliebig.

Dann gilt nach 1. auch $x \in M$.

Dann gilt nach 2. auch $x \in N$.

Also folgt für beliebige $x \in L$ immer auch $x \in N$,
quod erat demonstrandum, was zu beweisen war. □

Man mache sich klar, woraus ein formaler **Beweis** besteht:

1. Eine genau formulierte Behauptung.
2. Eine Aufstellung der Voraussetzungen, unter Heranziehung des bisher vorliegenden Wissens (Definitionen und bekannte "Sätze").
3. Eine passende Umformulierung des Ziels ("... zu zeigen ist ..."), wieder unter Benutzung des Vorwissens.
4. Eine Aufstellung der Beweisschritte, dabei immer von zutreffenden Aussagen ausgehend und neue zutreffende Aussagen erschließend.

Bevor man so einen Beweis aufschreibt, mus man ihn gedanklich erarbeiten. Dazu kann man alle möglichen Hilfsmittel benutzen, auch "raten" oder einen Hellseher befragen, es kommt auf Korrektheit nicht an. Aber dann muß der Beweis sauber und schlüssig nach den obigen Regeln hingeschrieben werden. Im Extremfall, in der Disziplin "**Maschinelles Beweisen**" der "**Künstlichen Intelligenz**", müssen die Schritte in einer für Computer verständlichen Form sequentiell eingegeben werden.

Ein besonders übler Anfängerfehler ist, von ungesicherten und zu beweisenden Aussagen auszugehen, dann gesicherte Aussagen zu erschließen und dann zu behaupten, die zu Anfang formulierten Aussagen seien damit bewiesen.

**Aus Unsinn kann man etwas Sinnvolles folgern,
ohne daß dadurch der Unsinn sinnvoll wird!**

Beispiel: Aus der unsinnigen Gleichung $3 = 7$ für natürliche Zahlen folgt durch die legale Multiplikation mit 0 auf beiden Seiten die korrekte Aussage $0 = 0$, aber das beweist nicht, daß $3 = 7$ korrekt war.

Also noch einmal: ein korrekter Beweis erfordert u.a.

eine Aufstellung der Beweisschritte, dabei immer von **zutreffenden** Aussagen ausgehend und neue zutreffende Aussagen erschließend.

In der Abfolge der Beweisschritte dürfen natürlich auch keine Lücken sein. Wir werden das im folgenden üben.

Es sollte bis hierher schon klar sein, daß ein Beweisgang eine **Richtung** hat, und zwar immer von einer schon bewiesenen Aussage zu einer neuen, die dann auch bewiesen ist. Umkehrungen dieser **Schlußrichtung** sind im allgemeinen unzulässig, wenn nicht exakt bewiesen wird, daß auch die Umkehrung der Schlußrichtung korrekt ist. Wenn aus einer Aussage A eine Aussage B folgt, sagt man auch, B sei **notwendig** für A oder eine **notwendige Bedingung** für A.

Beispiel: *Ist p eine Primzahl größer als 2, so ist p ungerade.* Hier ist die Aussage *p ist ungerade* eine notwendige Bedingung dafür, daß p eine Primzahl größer als 2 ist. Diese Schlußrichtung läßt sich nicht umkehren, denn man sieht am Beispiel der 9, daß nicht alle ungeraden Zahlen Primzahlen sind.

Wenn aus einer Aussage A eine Aussage B folgt, sagt man auch, A sei **hinreichend** für B oder eine **hinreichende Bedingung** für B.

Aufgabe: Für beliebige Mengen M und N folgt aus $M \subseteq N$ immer $P(M) \subseteq P(N)$.

Deshalb ist bei der obigen Aufgabe die Bedingung $M \subseteq N$ hinreichend für $P(M) \subseteq P(N)$.

Wenn sich die Schlußrichtung zwischen zwei Aussagen A und B umkehren läßt, wenn also A hinreichend für B und B hinreichend für A ist, so heißen die Aussagen A und B **logisch äquivalent**. Man sagt auch, A sei notwendig und hinreichend für B (oder umgekehrt).

Die beim Publikum beliebtesten logischen Fehler entsteht beim Lösen von Gleichungen. Wenn eine Gleichung, etwa $x^3 - 1 = 0$ zu "lösen" ist, so hat man zunächst die Problemstellung sauberer zu formulieren:

1. Man finde eine reelle Zahl x, so daß $x^3 - 1 = 0$ gilt. Oder:
2. Man gebe alle reellen Zahlen x mit $x^3 - 1 = 0$ an. Oder:
3. Man gebe alle komplexen Zahlen x mit $x^3 - 1 = 0$ an (es gibt 3).

Im ersten Fall reicht es, ein Beispiel anzugeben und die "Probe" zu machen, indem man z.B. für die reelle Zahl $x = 1$ zeigt, dass $x^3 - 1 = 0$ gilt. Es ist ja gar nicht danach gefragt, ob es noch andere Lösungen gibt. Im zweiten und dritten Fall muss man erstens eine oder mehrere Zahlen angeben, zweitens beweisen, dass diese Zahlen Lösungen sind ("Probe") und drittens den Beweis führen, dass es keine anderen Lösungen gibt.

Der allererste Fehler besteht oft darin, Gleichungen wie $x + y = 2,\ x - y = 0$ einfach hinzuschreiben, ohne eine klare Problemstellung damit zu verbinden.

Gleichungen an sich sind sinnlos.

Es sollte z.B. heißen:
Gesucht sind alle reellen Zahlen x, y mit $x + y = 2,\ x - y = 0$.

Der nächste Fehler schließt sich an, wenn die Gleichungen hingeschrieben werden, dann so lange gerechnet wird, bis man zu $x = y = 1$ kommt, und dann "Schluß gemacht" wird. Wenn man die Gleichungen hinschreibt und losrechnet, muss man **vorher** die Annahme machen, es gäbe Zahlen x, y, die die Gleichungen erfüllen.

Denn mit etwas, was nicht existiert, kann man nicht rechnen.

Macht man die Annahme, es gäbe reelle Zahlen x, y mit $x + y = 2,\ x - y = 0$ und bekommt dann nach einiger Rechnung $x = y = 1$ heraus, so hat man folgendes bewiesen: *Wenn es Lösungen der Gleichungen gibt, so sind sie alle gleich, und zwar $x = y = 1$.* Das beweist keineswegs, daß $x = y = 1$ die Gleichungen löst, sondern nur die **Eindeutigkeit** der Lösung unter der Voraussetzung der **Existenz** der Lösung. Ohne den zusätzlichen Existenzbeweis (die "**Probe**") hängt aber auch der Eindeutigkeitsbeweis in der Luft, weil er nur unter der Voraussetzung der Existenz einer Lösung gilt.

Beim "Lösen" von Gleichungen ist die "Probe" unerläßlich.

Der letzte Standardfehler betrifft das Rechnen von "Proben". Die immer wieder anzutreffende Rechenkette

$$\begin{array}{rcl} x+y & = & 2 \\ 1+1 & = & 2 \\ 0 & = & 0 \end{array} \qquad \begin{array}{rcl} x-y & = & 0 \\ 1-1 & = & 0 \\ 0 & = & 0 \end{array}$$

hat mehrere Fehler: Erstens darf man nie Gleichungen hinschreiben, ohne eine Annahme der Existenz der vorkommenden Größen zu machen und das Erfülltsein der Gleichungen anzunehmen (das soll die "Probe" aber gerade beweisen!). Zweitens hilft es nicht, die korrekte Aussage $0 = 0$ herzuleiten, denn das beweist gar nichts, weil man aus falschen Aussagen richtige erschliessen kann.

Man sollte Gleichungsproben immer so hinschreiben, daß man die Gleichungen selber nicht verwendet, sondern sie aus gesicherten Aussagen herleitet.

Man kann das durch getrenntes Ausrechnen der rechten und linken Seiten machen, mit einem Vergleich am Schluß.

Beispiel:
Behauptung:
Die Zahlen $x = y = 1$ erfüllen die Gleichungen $x + y = 2,\ x - y = 0$.
Beweis: Beim Einsetzen von $x = y = 1$ ergeben sich die linken Seiten der Gleichungen als $2 = 1+1$ bzw. $0 = 1-1$. Weil diese getrennt berechneten linken Seiten mit den entsprechenden rechten Seiten der gegebenen Gleichungen übereinstimmen, sind die gegebenen Gleichungen erfüllt.

Oder:
Behauptung: Die Zahl $x = 1$ genügt der Gleichung $x^2 - 2x + 1 = x - 1$.
Beweis: Die Zahl $x = 1$ erfüllt

$$x^2 - 2x + 1 = 1 - 2 + 1 = 0 \text{ und } x - 1 = 0.$$

Deshalb ist die Gleichung $x^2 - 2x + 1 = x - 1$ für $x = 1$ erfüllt.

Oder: Behauptung: Die Zahl $x = 1$ genügt der Gleichung $x^2 - 2x + 1 = x - 1$.
Beweis: Die Gleichung ist äquivalent zu $x^2 - 3x + 2 = 0$.
Die Zahl $x = 1$ erfüllt

$$x^2 - 3x + 2 = 1 - 3 + 2 = 0.$$

Hier spart man sich das getrennte Ausrechnen der rechten Seite. Aber das setzt voraus, daß auf der rechten Seite nichts mehr zu rechnen ist, weil dort nur noch eine Konstante steht.

1.1.5 Mengenoperationen

Definition 1.8 *Seien M und N beliebige Mengen. Dann sind*

$$\begin{aligned} M \cap N &:= \{x \mid x \in M \text{ und } x \in N\} \\ M \cup N &:= \{x \mid x \in M \text{ oder } x \in N\} \end{aligned}$$

als **Durchschnitt** *und* **Vereinigung** *von M und N definiert. Etwas anders formuliert: Es gilt für alle x die Aussage*

$$x \in \left\{ \begin{array}{c} M \cap N \\ M \cup N \end{array} \right\} \text{ genau dann, wenn } x \in M \left\{ \begin{array}{c} \text{und} \\ \text{oder} \end{array} \right\} x \in N$$

gilt. Zwei Mengen M und N mit $M \cap N = \emptyset$ heißen **disjunkt**. *Eine Menge P ist die* **disjunkte Vereinigung** *zweier Mengen M und N, wenn gilt*

$$\begin{aligned} M \cap N &= \emptyset \\ M \cup N &= P. \end{aligned}$$

Dabei ist "oder" nicht als "ausschließendes oder" gemeint, und es gilt deshalb

$$M \cap N \subseteq M \subseteq M \cup N$$

für alle Mengen M und N (Beweis?). Die offensichtliche Parallelität zwischen den Mengenoperationen $\cap$ und $\cup$ und den logischen Operationen "und" und "oder" wird uns noch beschäftigen. Die Studierenden der Informatik sollten zumindestens ahnen, daß Rechner aus "**Schaltlogik**" bestehen, und deshalb sind die Gesetze der Logik und allgemeiner der nach **Boole**[1] benannten **Booleschen Algebra**[2] ein unabdingbares Grundwissen für angehende Informatiker. Die einfachsten solchen Gesetze, hier in der "Verkleidung" als Regeln für Mengenoperationen, bringt

Theorem 1.9 *Für beliebige Mengen M, N und S sowie die obigen Mengenoperationen gelten die Regeln*

$$\begin{aligned}
M \cup N &= N \cup M && \textit{Kommutativität von } \cup \\
M \cap N &= N \cap M && \textit{Kommutativität von } \cap \\
(M \cup N) \cup S &= M \cup (N \cup S) && \textit{Assoziativität von } \cup \\
(M \cap N) \cap S &= M \cap (N \cap S) && \textit{Assoziativität von } \cap \\
(M \cup N) \cap S &= (M \cap S) \cup (N \cap S) && \textit{Distributivität von } \cup \textit{ und } \cap \\
(M \cap N) \cup S &= (M \cup S) \cap (N \cup S) && \textit{Distributivität von } \cap \textit{ und } \cup \\
M \cup \emptyset &= M && \textit{Absorptionsgesetz für } \cup \\
M \cap \emptyset &= \emptyset && \textit{Absorptionsgesetz für } \cap
\end{aligned}$$

Aufgabe: Man übe das saubere Aufschreiben von Beweisen an

$$\text{Aus } A \subseteq M \text{ und } B \subseteq M \text{ folgt } A \cup B \subseteq M. \tag{1.10}$$

Definition 1.11 *Sind M und N Mengen, so ist*

$$M \setminus N := \{x \mid x \in M \textit{ und nicht } x \in N\}$$

die **Differenzmenge**, *bestehend aus allen Elementen von M, die nicht in N sind.*

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Boole.html`
[2] `http://de.wikipedia.org/wiki/Boolesche_Algebra`

Man mache sich klar, daß immer $M \setminus N \subseteq M$ gilt, und deshalb sind Aussagen wie $M \setminus (M \setminus N) = N$ falsch. Aus diesem Grund ist es auch schlecht, die Differenzmenge mit $M - N$ statt mit $M \setminus N$ zu bezeichnen, denn für Zahlen macht die Formel $M - (M - N) = N$ durchaus Sinn und wirkt verführerisch. Das Ganze ist ein Beispiel für verschiedene denkbare Interpretationen der formalen Zeichenkette $M - (M - N) = N$. Auch in dieser Hinsicht sind wir hier mitten in der Informatik.

Frage: Wie läßt sich $M \setminus (M \setminus N)$ einfacher schreiben?

Definition 1.12 *Sind Mengen M und N Teilmengen einer gemeinsamen Obermenge G, so kann man das* **Komplement** *von M bzw. N bezüglich G durch $\overline{M} := G \setminus M$ bzw. $\overline{N} := G \setminus N$ bezeichnen.*

Diese Bezeichnungsweise macht nur Sinn, wenn die Obermenge klar definiert und für die auftretenden Mengen **gemeinsam** ist. Eigentlich müßte man das Symbol G in die Notation aufnehmen, z.B. durch $\overline{M}^G := G \setminus M$.

Theorem 1.13 *Sind Mengen M und N Teilmengen einer gemeinsamen Obermenge G, so gelten für die Komplementbildung bezüglich G die Regeln*

$$\begin{array}{rcl} M \setminus N & = & M \cap \overline{N} \\ \overline{M \cup N} & = & \overline{M} \cap \overline{N} \\ \overline{M \cap N} & = & \overline{M} \cup \overline{N} \\ \textit{Aus } M \subseteq N & \textit{folgt} & \overline{N} \subseteq \overline{M} \\ \overline{\overline{M}} & = & M. \end{array}$$

Alle Rechenregeln dieses Abschnitts eignen sich zum Üben von sauberen Beweisen, aber man sieht dabei, daß man eigentlich schon die Regeln der Logik kennen muß, um diese Beweise zu führen. Wir werden das also etwas später nachholen. Man sollte an dieser Stelle aber schon ahnen, daß die Komplementbildung mit der logischen Negation, der "nicht"–Operation, zusammenhängt.

1.1.6 Cartesische Produkte

Definition 1.14 *Sind M und N Mengen, so ist nach Renee* **Descartes**[1] *das* **cartesische Produkt**[2] *$M \times N$ die Menge*

$$\{(x, y) \mid x \in M \text{ und } y \in N\}.$$

von **geordneten Paaren** *von Elementen von M und N.*

[1] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Descartes.html
[2] http://de.wikipedia.org/wiki/Kartesisches_Produkt

Man mache sich klar, daß hier eine Absprache über die Verwendung runder Klammern und Kommata getroffen wird, die sich von der Verwendung geschweifter Klammern und Kommata bei der Mengendefinition (1.2) durch Aufzählung auf Seite 9 wesentlich unterscheidet. Die Mengen $\{1, 2\}$ und $\{2, 1\}$ sind gleich, aber die Paare $(1, 2)$ und $(2, 1)$ sind es nicht.

Man sehe sich unbedingt die Beispiele aus [4], Seite 12–13 an, wobei klar werden sollte, was cartesische Produkte mit cartesischen Koordinaten zu tun haben.

Natürlich kann man auch mehrfache cartesische Produkte bilden, etwa

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

als Menge von Tripeln.

Frage: Das ist nicht dasselbe wie $(L \times M) \times N$ bzw. $L \times (M \times N)$, oder?

Hat man n Mengen $M_1, \ldots, M_n$, so definiert man entsprechend

$$M_1 \times \ldots \times M_n := \{(x_1, \ldots, x_n) \mid x_i \in M_i \text{ für alle } i \text{ von 1 bis } n\}$$

als Menge der n–**Tupel** aus $M_1, \ldots, M_n$. Dieser etwas seltsame Begriff verallgemeinert die Tripel, Quadrupel, Quintupel usw. zu n–Tupeln. Man nennt die einzelnen Mengen M_i dann **Komponenten** oder **Faktoren** des cartesischen Produkts. Sind alle Komponenten M_i eines cartesischen Produkts gleich einer einzigen Menge M, so vereinfacht man das Ganze zu

$$M^n := \{(x_1, \ldots, x_n) \mid x_i \in M \text{ für alle } i \text{ von 1 bis } n\}. \tag{1.15}$$

Man mache sich klar, daß zwischen den Mengen M^{m+n} und $M^m \times M^n$ für beliebige positive m und n zwar ein feiner Unterschied besteht (welcher?), der aber nicht wesentlich ist, so daß die Potenznotation nicht ganz unsinnig gewählt ist.

1.2 Relationen

1.2.1 Grundbegriffe

Alle Informatik–Studierenden werden wissen oder ahnen, daß **relationale Datenbanken** im Studium und in der Praxis eine wichtige Rolle spielen. Hier ist der grundlegende mathematische Begriff dazu:

Definition 1.16 *Seien M und N beliebige Mengen. Eine* **Relation**[1] *R auf $M \times N$ ist eine Teilmenge von $M \times N$. Man schreibt für beliebige Paare $(x, y) \in M \times N$ statt $(x, y) \in R$ auch xRy.*

Etwas fortgeschrittene Informatiker sehen hier eine binäre Operation in Infixform. Darauf kommen wir später zu sprechen.

Es gibt sehr viele Beispiele zu Relationen, etwa $\leq$ auf reellen Zahlen, oder "dasselbe" und "das Gleiche" in der Umgangssprache, oder "kongruent" bei ebenen Dreiecken oder anderen ebenen Figuren. Ferner gehören dazu alle "Tabellen"[2] von Datenbanken. Auf der Menge

$$I\!N \times \{ \text{ ASCII–Strings } \} \times \{ \text{ ASCII–Strings } \}$$

ist die Tabelle

27	August	Meier
39	Berta	Lehmann
52	Carl	Schulte

eine Relation, weil sie eine Teilmenge des obigen cartesischen Produktes ist. Tabellarische Relationen sind die Grundeinheiten relationaler **Datenbanken**. Man sehe sich auch die in [4], S. 13–14 angegebenen Beispiele an. Ein typischer Fall ist auch die Relation "ist Kind von" auf der Menge $M \times M$ von Paaren von Menschen. Dabei wird klar, daß die Infixschreibweise "Hans ist Kind von Monika" im Stile von xRy besser ist als zu sagen

$$(Hans, Monika) \in istKindvon \subseteq Menschen \times Menschen.$$

Frage: In welchem Sinne und auf welchen cartesischen Produkten sind $\in$ und $\subseteq$ Relationen?

Beispiele

- Man könnte die Relation "Parabel" als
$$\{(x, y) \ : \ y = x^2\} \subseteq I\!R \times I\!R$$
definieren.
- Auf der Menge *Studenten*$\times$*Prüfungen* kann die Relation *bestanden* definiert werden.

Definition 1.17 *Gilt in obiger Definition $M = N$, so spricht man von einer (zweistelligen) Relation auf M. Allgemeiner ist eine n–stellige Relation auf einer Menge M als Teilmenge von M^n definiert.*

[1] `http://de.wikipedia.org/wiki/Relation_%28Mathematik%29`
[2] `http://de.wikipedia.org/wiki/Relation_%28Datenbank%29`

1.2.2 Äquivalenz– und Ordnungsrelationen

Definition 1.18 *Eine zweistellige Relation R auf M heißt*

reflexiv	*wenn*	*für alle $x \in M$ gilt xRx*
symmetrisch	*wenn*	*für alle $x, y \in M$ aus xRy auch yRx folgt*
transitiv	*wenn*	*für alle $x, y, z \in M$ aus xRy und yRz auch xRz folgt*

Frage: Welche dieser Eigenschaften hat die Relation $\subseteq$?

Definition 1.19 *Eine zweistellige Relation R auf M heißt* **Äquivalenzrelation**[1] *wenn sie reflexiv, symmetrisch und transitiv ist.*

Von dieser Art sollten alle Relationen sein, die irgendwie "Gleichheit" oder "Ähnlichkeit" ausdrücken, z.B. "kongruent" auf der Menge der Dreiecke. Transitiv sollten alle Relationen sein, die einen unsymmetrischen Größenvergleich anstellen, z.B. "wiegt nicht mehr als". Und Symmetrie wird u.a. gebraucht für den Unterschied zwischen "größer als" und "größer als oder gleichgroß".

Man sehe sich die Beispiele in [4], S. 15–16 an.

Definition 1.20 *Ist R eine Äquivalenzrelation auf M, so kann man zu jedem $x \in M$ die* **Äquivalenzklasse**

$$[x] := \{y \in M \mid yRx\}$$

der zu x unter R äquivalenten Elemente von M bilden. Ist A eine Äquivalenzklasse, und schreibt man A als $A = [x]$, so wird x als **Vertreter** *der Klasse A bezeichnet.*

Frage: Was würde sich ändern, wenn wir $[x] := \{y \in M \mid xRy\}$ definiert hätten?

Theorem 1.21 *Ist R eine Äquivalenzrelation auf M, so ist M die disjunkte Vereinigung der verschiedenen Äquivalenzklassen von Elementen von M.*

Wir holen den Beweis später in (1.3.4) nach. Er folgt allerdings auch leicht aus

Theorem 1.22 *Ist R eine Äquivalenzrelation auf einer Menge M, so ist jede Äquivalenzklasse durch jeden ihrer Vertreter eindeutig bestimmt.*

[1] ,http://de.wikipedia.org/wiki/%C3%84quivalenzrelation

Ein Beispiel:

Definiert man zwei Menschen als namensäquivalent, wenn sie den gleichen Nachnamen haben, so zerfällt die Menge aller Menschen in disjunkte Äquivalenzklassen, die jeweils aus den Menschen mit gleichem Familiennamen bestehen. Und jede dieser Namens-Äquivalenzklassen, z.B. die aller Menschen, die "Mayer" heißen, ist durch jedes beliebige ihrer Mitglieder eindeutig bestimmt. Jeder Mensch namens "Mayer" vertritt die Namensäquivalenzklasse aller Mayers.

Frage: Wie kann man den Satz "Vor dem Gesetz sind alle Menschen gleich" abstrahieren?

Frage: Wie kann man den Satz "Jeder Mensch ist ein Individuum und nur mit sich selbst vergleichbar" abstrahieren?

Dabei ist jeweils nach der Angabe einer geeigneten Relation gefragt, und es sollte gesagt werden, was die Äquivalenzklassen sind.

Wir sollten üben, so etwas wie die Behauptung des Theorems 1.22 ein wenig mathematischer aufzuschreiben. Nehmen wir eine beliebige Äquivalenzklasse und nennen wir sie $[x]$. Daraus nehmen wir ein beliebiges Element $y \in [x]$. Es gilt also yRx und wegen der Symmetrie auch xRy. Dann besagt unsere Behauptung, daß die Äquivalenzklasse von y gleich der von x sein muß, also muß $[x] = [y]$ bewiesen werden.

Das wiederum erfordert je einen Beweis von $[x] \subseteq [y]$ und $[y] \subseteq [x]$. Weil die Voraussetzungen xRy und yRx symmetrisch gegen Vertauschung von x und y sind, reicht es, unter diesen Voraussetzungen $[x] \subseteq [y]$ zu zeigen, denn dann gilt derselbe Beweis, unter Vertauschung von x mit y, auch für die Aussage $[y] \subseteq [x]$.

Zum Beweis von $[x] \subseteq [y]$ müssen wir beweisen, daß aus $z \in [x]$ auch $z \in [y]$ folgt. Wir setzen also $z \in [x]$ voraus, und das bedeutet, daß zRx und xRz gelten. Wegen der Transitivität der Relation R folgt aber aus zRx und xRy stets zRy, und dies ist nichts anderes als die Behauptung $z \in [y]$, die wir beweisen wollten. □

Sortieren ud Suchen sind extrem wichtige Standardaufgaben in der Informatik. Dazu braucht man noch

Definition 1.23 *Eine zweistellige Relation R auf M heißt*

- **antisymmetrisch**, *wenn für alle* $x, y \in M$ *aus* xRy *und* yRx *immer* $x = y$ *folgt,*
- **total**, *wenn für alle* $x, y \in M$ *entweder* xRy *oder* yRx *gilt.*
- *Eine reflexive, antisymmetrische und transitive Relation heißt* **Teilordnung**.
- *Eine* **Ordnungsrelation**[1] *ist total und eine Teilordnung, d.h. total, reflexiv, antisymmetrisch und transitiv.*

JAVA–Freaks sollten ahnen, daß man eine Ordnungsrelation über ein Interface *Sortable* spezifizieren sollte, das bei geeigneter Implementierung dann eine Sortierung erlaubt.

Theorem 1.24 *Eine Menge* $M = \{x_1, \ldots, x_n\}$ *mit* n *Elementen, die eine Ordnungsrelation* R *hat, kann man so umsortieren, daß* $M = \{y_1, \ldots, y_n\}$ *mit*

$$y_1 R y_2, y_2 R y_3, \ldots, y_{n-1} R y_n$$

gilt, d.h. jede endliche Menge ist **sortierbar**.

Wie man das effektiv macht, lernt man in der Informatik, und warum das immer geht, lernt man in der Mathematik.

Wir benutzen die Gelegenheit, um zu zeigen, daß ein mathematischer Beweis und eine informatischer Algorithmus sehr eng verwandt sein können, nämlich dann, wenn ein Beweis **konstruktiv** ist und aus der Angabe eines Verfahrens besteht, von dem man zeigt, daß es das Gewünschte leistet. Genau dasselbe muß man in der Informatik tun, wenn man so ein Verfahren untersucht.

Man mache sich erst einmal klar, daß einelementige Mengen mit Ordnungsrelation immer schon sortiert sind, und daß man zweielementige Mengen $\{x_1, x_2\}$ entweder als $\{x_1, x_2\}$ oder $\{x_2, x_1\}$ sortieren kann, denn es muß ja wegen der Totalität der Ordnung R immer entweder $x_1 R x_2$ oder $x_2 R x_1$ gelten.

Jetzt verwendet man ein Argument, das in der Informatik **Rekursion** und in der Mathematik **Induktion** heißt. Man reduziert das Sortieren einer endlichen Menge auf das Sortieren zweier kleinerer Teilmengen. Wenn man das immer weiter betreibt, hat man insgesamt eine Sortierung der Gesamtmenge erreicht. Die Grundidee zur Reduktion ist die von **Quicksort**[2]. Aus einer

[1] `http://de.wikipedia.org/wiki/Ordnungsrelation`
[2] `http://de.wikipedia.org/wiki/Quicksort`

gegebenen Menge $M = \{x_1, \ldots, x_n\}$ mit mindestens zwei Elementen nimmt man sich ein beliebiges Element z heraus, etwa $z = x_1$. Dann definiert man die Mengen

$$\begin{array}{lcl} M_1 & := & \{x \in M \ : \ xRz \text{ und } x \neq z\} \\ M_2 & := & \{z\} \\ M_3 & := & \{x \in M \ : \ zRx \text{ und } x \neq z\} \end{array}$$

und beweist, daß

- sie disjunkt sind,
- ihre Vereinigung ganz M ist und sie
- weniger Elemente als M enthalten

(Frage: Wie beweist man das?). Nun hat man das Problem reduziert, denn die neuen Mengen sind kleiner, und man kann annehmen, daß man sie sortieren kann, etwa in

$$M_1 = \{u_1, \ldots, u_k\}, \ M_2 = \{z\}, \ M_3 = \{v_1, \ldots, v_{n-1-k}\},$$

wobei k einen der Werte 0 bis $n-1$ annehmen kann. Es gilt also

$$u_1 R u_2, \ldots, u_{k-1} R u_k \text{ und } v_1 R v_2, \ldots, v_{n-2-k} R v_{n-1-k}$$

nach der Sortierung. Jetzt hat man in

$$u_1 R u_2, \ldots, u_{k-1} R u_k, u_k R z, z R v_1, v_1 R v_2, \ldots, v_{n-2-k} R v_{n-1-k}$$

eine Sortierung aller Elemente von $M = \{u_1, \ldots, u_k, z, v_1, \ldots, v_{n-1-k}\}$.

Anfänger, denen diese Argumentation zu schwierig erscheint, sollten sie sich später noch einmal ansehen. Sie gehört zum Kernwissen der Informatik.

Man nimmt bei jeder Menge von mehr als einem Element immer das erste Element und spaltet die Menge dann wie oben in drei Mengen auf. Klar?

Man mache sich das einmal am Beispiel $M = \{5, 3, 2, 9, 4, 7\}$ und der Relation $\leq$ deutlich. Wir nehmen das erste Element heraus, es ist die 5, und sie wird unser erstes z. Jetzt laufen wir gedanklich durch den Rest der Menge und schreiben alles, was kleiner ist als $z = 5$ nach links, und alles andere nach rechts, aber ohne es zu sortieren. Das liefert die zweite Zeite der folgenden Tabelle.

$$\begin{array}{llllllllllll} \{ & 5 & , & 3 & , & 2 & , & 9 & , & 4 & , & 7 & \} \\ \{ & 3 & , & 2 & , & 4 & \} & \{5\} & \{ & 9 & , & 7 & \} \\ \{ & 2 & \}\{ & 3 & \}\{ & 4 & \} & \{5\} & \{ & 7 & \}\{ & 9 & \} \quad \emptyset \end{array}$$

Jetzt rekurrieren wir auf die links stehende Menge $\{3, 2, 4\}$. Wir nehmen das erste Element 3 heraus und verfahren wie bisher, aber angewendet auf de Menge $\{3, 2, 4\}$. Was kleiner als 3 ist, kommt links vor die 3, was größer ist, rechts neben die 3. Und ebenso verfahren wir für die Menge $\{9, 7\}$. Das Verfahren endet, wenn nur noch einelementige oder leere Mengen da sind, und die enthalten dann die sortierten Elemente der Ausgangsmenge.

Aufgabe: Wie verläuft das Verfahren, wenn man die Buchstaben des Wortes *Vorlesung* alphabetisch sortieren will?

Aufgabe: Wie ist das Verfahren zu modifizieren, wenn es nicht auf Mengen, sondern auf Tupeln mit wiederholt vorkommenden gleichen Elementen ablaufen können soll? Beispiel: das Wort *Vorlesungsskript*.

1.2.3 Relationale Datenbanken und das Relationenkalkül

An dieser Stelle ist [1], Seite 114–117 eine passende Hintergrundliteratur. Mehrere konkrete Beispiele für **relationale Datenbanken** werden mündlich vorgeführt. Hier stellen wir die Theorie knapp zusammen. Harte praktische Anwendungen lernt man in der **Wirtschaftsinformatik**.

Der Grundgedanke ist, daß alle datenbanktechnischen Relationen aus Mengen von Tupeln bestehen, die man als Tabellen speichern kann. Sie sind also, mathematisch gesehen, immer Teilmengen cartesischer Produkte von Mengen. Um mit solchen Relationen arbeiten zu können, gibt es Verknüpfungsoperationen, die aus gegebenen Relationen neue Relationen zu konstruieren gestatten. Diese Operationen bilden das **Relationenkalkül** bzw. die **relationale Algebra**[1].

Definition 1.25 *1. Es seien $R \subseteq M_1 \times \ldots \times M_m$ und $S \subseteq N_1 \times \ldots \times N_n$ Relationen. Dann ist das* **cartesische Relationenprodukt** *$R \times S$ oder in informatiknaher Schreibweise R TIMES S die Relation in $M_1 \times \ldots \times M_m \times N_1 \times \ldots \times N_n$, die aus allen möglichen Tupeln der Form (r, s) mit $r \in R$ und $s \in S$ besteht.*

2. Eine **Projektion** *einer Relation besteht aus dem Weglassen von gewissen Komponenten der jeweiligen Tupel. Dabei entsteht eine neue Relation als Teilmenge eines cartesischen Produktes mit weniger Komponenten.*

3. Eine **Selektion** *auf einer Relation $R \subseteq M_1 \times \ldots \times M_m$ besteht aus der Auswahl von Tupeln von R, die eine bestimmte Eigenschaft haben.*

[1] `http://de.wikipedia.org/wiki/Relationale_Algebra`

Frage: Wie sehen Projektionen der Relation

$$\{(x, y) \ : \ x^2 + y^2 = 1\} \subseteq I\!R \times I\!R$$

aus?

Auf der Menge *Studenten*×*Prüfungen* sei *bestanden* eine Relation. Frage: Wie bekommt man alle Studenten heraus und wie bekommt man nur die Studenten, die eine bestimmte Prüfung bestanden haben?

Definition 1.26 *Es seien R und S Relationen in $M_1 \times \ldots \times M_m$, d.h. mit gleichen Komponenten.*

1. *Die relationale* **Vereinigung** *$R \cup S$ oder R UNION S von R und S ist die Relation in $M_1 \times \ldots \times M_m$, die durch die mengentheoretische Vereinigung $R \cup S$ der Teilmengen R und S von $M_1 \times \ldots \times M_m$ gegeben ist.*

2. *Die relationale* **Differenz** *$R \setminus S$ oder R MINUS S von R und S ist die Relation in $M_1 \times \ldots \times M_m$, die durch die mengentheoretische Differenz $R \setminus S$ der Teilmengen R und S von $M_1 \times \ldots \times M_m$ gegeben ist.*

Der relationale **Durchschnitt** von zwei Relationen R und S ist keine neue Operation, weil er sich (siehe die Frage auf Seite 19) als $R \setminus (R \setminus S)$ schreiben läßt.

Die beiden erstgenannten Operationen erlauben es, die Komponentenzahl von Relationen zu vergrößern bzw. zu verkleinern, während die beiden letzten Operationen nur auf Relationen mit gleichen Komponenten wirken. Aber die wichtigste Operation kommt noch. Sie erlaubt zwei allgemeine Relationen $R \subseteq M_1 \times \ldots \times M_m$ und $S \subseteq N_1 \times \ldots \times N_n$ und konstruiert nicht das gesamte cartesische Produkt $R \times S$, sondern nur eine Teilmenge davon. Sie setzt voraus, daß M und N eine oder mehrere Komponenten gemeinsam haben, so daß man je eine Projektion P_M auf $M_1 \times \ldots \times M_m$ und P_N auf $N_1 \times \ldots \times N_n$ mit gleichem Bildbereich

$$Q := P_M(M_1 \times \ldots \times M_m) = P_N(N_1 \times \ldots \times N_n)$$

definieren kann. Man mache sich klar, daß Q aus einer Auswahl von gemeinsamen Komponenten von $M_1 \times \ldots \times M_m$ und $N_1 \times \ldots \times N_n$ besteht.

Dann ist der **Verbund** oder **join** von R und S **über** Q definiert als die Menge aller Paare $(r, s) \in R \times S \subseteq M_1 \times \ldots \times M_m \times N_1 \times \ldots \times N_n$, für die $P_M(r) = P_N(s)$ gilt, d.h. die in den Komponenten von Q übereinstimmen.

Beispiel: Eine Firma hat für ihre Arbeiter zwei Relationen in Tabellenform gespeichert:

- die (fast konstanten) allgemeinen Personaldaten wie Name, Adresse, Telefonnummer usw. in einer Relation auf

$$Personalnummer \times Name \times Adresse \times Telefonnummer \times \ldots$$

- den jeweiligen Wochenlohn für eine bestimmte Woche als Relation auf

$$Personalnummer \times Wochenlohn$$

Um allen Arbeitern einen Brief zu schreiben, in dem ihnen ihr Wochenlohn mitgeteilt wird, macht man einen JOIN der beiden Relationen über die $Personalnummer$. Man bekommt Tupel aus dem cartesischen Produkt

$$Personalnummer \times Wochenlohn \times Name \times Adresse \times \ldots$$

bei denen der Wochenlohn und die Adresse zum jeweiligen Mitarbeiter passen. Mit diesen Daten baut man dann einen Serienbrief.

Wenn man sich über Datenbanken genauere Gedanken macht, stellt sich heraus:

> **Alles, was man mit relationalen Datenbanken machen kann, kann man mit den fünf oben zuerst angegebenen Operationen des Relationenkalküls ausdrücken**.

Das ist natürlich kein sauber formulierter mathematischer Satz, denn was soll *"Alles, was man mit relationalen Datenbanken machen kann"* heißen? Aber es bleibt ja im Informatikstudium noch genug Zeit, dieser Frage nachzugehen. Ein einfaches Beispiel ist die Reduktion des JOIN: Man kann zuerst das gesamte cartesische Produkt bilden und dann mit einer Selektion diejenigen Paare herausfiltern, die der JOIN–Bedingung entsprechen.

Man formuliert Standardoperationen auf Datenbanken schon seit langem in **SQL**[1], der "Structured Query Language", und das funktioniert, wenn es über hinreichend abstrakt formulierte Zugangsschnittstellen wie ODBC[2] und JDBC[3] realisiert wird, sogar unabhängig von der jeweiligen Datenbank und ihrer Implementierung.

Es sollte hier noch ein weiteres mündliches Beispiel für eine logische Datenbankabfrage angegeben werden, die sich auf Operationen des Kalküls stützt. In [1] steht ein Beispiel, Seite 115–117.

[1] `http://de.wikipedia.org/wiki/SQL`
[2] `http://de.wikipedia.org/wiki/ODBC`
[3] `http://de.wikipedia.org/wiki/JDBC`

1.3 Abbildungen

1.3.1 Grundbegriffe

Definition 1.27 *Es seien M und N Mengen. Eine* **Abbildung**[1] *f von M in N, geschrieben als*

$$f \ : \ M \to N \ \textit{oder} \ M \xrightarrow{f} N \ \textit{mit} \ x \mapsto f(x)$$

ist dann eine Vorschrift, die zu jedem $x \in M$ genau ein mit $f(x)$ bezeichnetes Element von N angibt. Man nennt dann $f(x)$ den **Wert** *von f auf x, und x ist* **Urbild** *oder* **Argument** *von $f(x)$. Die Menge M heißt* **Urbildmenge** *oder* **Definitionsbereich** *von f, die Menge N heißt* **Zielmenge** *von f. Die* **Bildmenge** *einer Teilmenge $U \subseteq M$ unter f ist*

$$f(U) := \{y \ : \ y \in N, \ \textit{es gibt ein} \ x \in U \ \textit{mit} \ y = f(x)\} \subseteq N.$$

Analog ist

$$f^{-1}(V) := \{x \ : \ x \in M, f(x) \in V\} \subseteq M.$$

die **Urbildmenge** *einer Teilmenge $V \subseteq N$. Schließlich ist die Teilmenge*

$$\{(x, f(x) \ : \ x \in M\}$$

des cartesischen Produkts $M \times N$ der **Funktionsgraph**[2] *von f. Der Funktionsgraph ist somit eine Relation auf $M \times N$.*

Abbildungen zwischen Mengen aus Zahlen werden oft auch als **Funktionen** bezeichnet. Wenn man die Schreibweisen

$$f \ : \ M \to N \text{ oder } M \xrightarrow{f} N$$

benutzt, ist immer klar, daß M und N Mengen sind und f eine Abbildung zwischen diesen ist. Man braucht dies nicht besonders zu erwähnen

Man sehe sich **unbedingt** die Beispiele von [4], S. 17–19 an!

Wichtig ist, daß eine Abbildung erst dann sauber definiert ist, wenn man Urbild– und Zielbereich exakt angibt. Die drei Symbole f, M, N einer Abbildung $f \ : \ M \to N$ gehören zusammen.

In der Informatik ist alles, was aus einem **Input** einen **Output** produziert, mathematisch gesehen eine Abbildung. Darunter fallen in der **Programmierung** alle **Funktionen**, **Prozeduren** und **Methoden**, und zwar auch

[1] `http://de.wikipedia.org/wiki/Funktion_%28Mathematik%29`
[2] `http://de.wikipedia.org/wiki/Funktionsgraph`

bei funktionaler oder objektorientierter Programmierung, nicht nur bei der prozeduralen. In diesem Sinne ist es das tägliche Brot von Informatikern, Abbildungen zu definieren. Auch in der Informatik ist eine Spezifikation einer Abbildung (d.h. eines Programms oder einer Methode) ohne saubere Angabe des In– und Outputs unvollständig und inexakt.

In der objektorientierten Programmierung (**OOP**) deklariert man Klassen. Das entspricht einer abstrakten Mengendefinition durch Eigenschaften. Elemente dieser Mengen existieren zunächst nicht, sondern müssen durch Instanzierung bzw. duch Aufruf eines Konstruktors erst erzeugt werden. Aber es kommt zur Klassendeklaration hinzu, daß man auch Methoden spezifiziert, und das sind Abbildungen, deren Definitionsbereich i.A. die zu deklarierende Klasse ist. Man deklariert also in der OOP (bei mathematischer Sichtweise) gleichzeitig Mengen und Abbildungen.

Man mache sich den Unterschied zwischen $\to$ und $\mapsto$ klar:

- $\to$ steht zwischen Urbildmenge und Zielmenge, also zwischen Mengen,
- $\mapsto$ steht zwischen Urbild und Wert, also zwischen Elementen.

Die Bezeichnungsweisen für $f(U)$ und $f^{-1}(V)$ sind etwas fragwürdig, weil man eigentlich die durch f indirekt definierten Abbildungen

$$\begin{array}{rcl} P(M) & \to & P(N), \quad U \mapsto f(U) \\ P(N) & \to & P(M), \quad V \mapsto f^{-1}(V) \end{array}$$

mit anderen Symbolen bezeichnen müßte, denn die Abbildung f bildet Elemente auf Elemente ab, und kann nicht identisch sein mit einer Abbildung, die Teilmengen in Teilmengen abbildet. Aber die Bezeichnungen sind praktisch und haben sich gut bewährt.

Aufgabe (zum sauberen Aufschreiben): Ist $f : M \to N$ eine Abbildung und ist $L \subseteq M$ eine Teilmenge des Definitionsbereichs, so gilt $f(L) \subseteq f(M)$.

Besonders wichtige Abbildungen in der Informatik sind die **Codes** und die **Speicherabbildungen**, aber das sollte mündlich vertieft werden. Typische Beispiele sind Morse– und ASCII–Code sowie die beim Hashing oder beim *memory management* verwendeten dynamisch veränderlichen Speicherabbildungen.

Abbildungen sind nichts Neues, weil sie spezielle Relationen sind. Man kann nämlich die oben angegebene und etwas fragwürdige Definition (was heißt "Zuordnung"?) ersetzen durch

Definition 1.28 *Es seien M und N Mengen. Eine* **Abbildung** *f von M in N ist gegeben durch eine Relation R_f auf $M \times N$ mit den Eigenschaften*

1. *Zu jedem $x \in M$ gibt es genau ein $y =: f(x) \in N$, so daß das Paar $(x, y) = (x, f(x))$ in $R_f \subseteq M \times N$ liegt.*
2. *R_f hat* **nur** *die dadurch definierten Elemente, keine anderen.*

Natürlich ist dann die Menge R_f gleich dem Funktionsgraphen von f.

Aufgabe: Das Ausfüllen eines Lottoscheins kann man als Abbildung zwischen den Mengen $M = \{1, 2, \ldots, 6\}$ und $N = \{1, 2, \ldots, 49\}$ (in der Theorie) oder als Abbildung zwischen $N = \{1, 2, \ldots, 49\}$ und $P = \{\square, \times\}$ (in der Praxis) beschreiben. Sind die Abbildungen beliebig? Worauf hat man zu achten? Man schreibe eine saubere Spezifikation der Anforderungen hin.

1.3.2 Exkurs zum Rechnen mit Gleichungen

Dies ist wieder etwas für die parallele Saalübung. Der Vorlesungsstoff geht mit Abschnitt 1.3.3 weiter.

Wenn man zwischen Elementen x, y einer Menge M eine Gleichung $x = y$ hat, so folgt daraus bei Anwendung einer Abbildung $f : M \to N$ immer auch die Gleichung $f(x) = f(y)$ in N. Dies ist die banale Grundlage allen Rechnens mit Gleichungen.

> Auf die beiden Seiten einer gültigen Gleichung kann man eine beliebige Abbildung simultan anwenden und erhält wieder eine gültige Gleichung.

In der Schule lernt man, daß man an einer zwischen Zahlen oder zahlenwertigen Ausdrücken bestehenden Gleichung verschiedene legale Operationen auf beiden Seiten simultan ausführen darf, z.B.

- Addition einer Zahl c
- Multiplikation mit einer Zahl z
- Quadrieren.

Das ist nichts als die simultane Anwendung der Abbildungen

- $f(x) = x + c$
- $f(x) = x * z$

- $f(x) = x^2$

auf die beiden Seiten. Natürlich kann man aber auch ganz beliebige Abbildungen anwenden. Das Rechnen mit Ungleichungen ist schwieriger und wird uns noch beschäftigen.

1.3.3 Eigenschaften

Definition 1.29 *Sei $f : M \to N$ eine Abbildung. Dann heißt f*

- **injektiv**[1]*, wenn für alle $x_1, x_2 \in M$ mit $x_1 \neq x_2$ gilt $f(x_1) \neq f(x_2)$*
- **surjektiv**[2]*, wenn es für alle $y \in N$ ein $x \in M$ gibt mit $y = f(x)$*
- **bijektiv**[3]*, wenn f surjektiv und injektiv ist.*

Man sehe sich die Beispiele aus [4] S. 20 **unbedingt** an!

In der Vorlesung werden diverse Veranschaulichungen von injektive, surjektiven und bijektiven Abbildungen vorgestellt, unter anderem $f(x) = x^2$ und $f(x) = x^3$ als Abbildungen $\mathbb{R} \to \mathbb{R}$. Um Bijektivität zu haben, muß man z.B. bei der Definition der Exponentialfunktion und des Logarithmus darauf achten, das die Definitionsbereiche korrekt sind, z.B.

$$\begin{array}{rrcl} \exp & : & \mathbb{R} & \to & (0,\infty), \\ \log & : & (0,\infty) & \to & \mathbb{R}. \end{array}$$

Will man einen Kreis in $\mathbb{R}^2$ darstellen, so kann man das zunächst nur für Halbkreise, weil Abbildungen immer nur einen Wert haben dürfen. das führt z.B. auf

$$\begin{array}{rcl} f(x) & := & +\sqrt{r^2 - x^2},\ -r \leq x \leq r \\ g(x) & := & -\sqrt{r^2 - x^2},\ -r \leq x \leq r \end{array}$$

für die beiden Halbkreise. Etwas eleganter ist es, die Abbildung

$$t \mapsto r * (\cos(t), \sin(t)) \in \mathbb{R}^2$$

zu benutzen, aber man muß mit dem Definitionsbereich aufpassen, wenn die Abbildung injektiv sein soll. Zum Beispiel kann man $-\pi \leq t < \pi$ oder $0 \leq t < 2\pi$ nehmen. Warum?

[1] `http://de.wikipedia.org/wiki/Injektivit%C3%A4t`
[2] `http://de.wikipedia.org/wiki/Surjektiv`
[3] `http://de.wikipedia.org/wiki/Bijektiv`

Definition 1.30 *Sind $f : L \to M$ und $g : M \to N$ Abbildungen, so ist die* **Komposition** *oder* **Hintereinanderanwendung** *oder* **Verkettung** *"g nach f" von f und g definiert durch die Abbildung*

$$g \circ f : L \to N, \quad x \mapsto g(f(x)) \text{ für alle } x \in L.$$

Eine typische Veranschaulichung ist das Diagramm

$$\begin{array}{ccccc} L & & \xrightarrow{g\circ f} & & N \\ f & \searrow & & \nearrow & g \\ & & M & & \end{array}$$

Es ist klar, daß für drei Abbildungen

$$K \xrightarrow{f} L \xrightarrow{g} M \xrightarrow{h} N$$

die Beziehung

$$(h \circ g) \circ f = h \circ (g \circ f) =: h \circ g \circ f$$

gilt. Oder?

Theorem 1.31 *Es seien $f : L \to M$ und $g : M \to N$ Abbildungen.*

- *Sind f und g injektiv, so auch $g \circ f$.*
- *Sind f und g surjektiv, so auch $g \circ f$.*
- *Sind f und g bijektiv, so auch $g \circ f$.*
- *Ist $g \circ f$ bijektiv, so ist g surjektiv und f injektiv.*

1.3.4 Direkte und indirekte Beweise

Auch dieser Abschnitt gehört in die Saalübung.

In [4] S. 21 steht ein indirekter Beweis für den zweiten Teil der letzten Aussage. Wir beweisen hier übungshalber zunächst den ersten Teil, und zwar mit einem direkten Beweis:
Voraussetzung: Es seien $f : L \to M$ und $g : M \to N$ Abbildungen, und $g \circ f$ sei bijektiv.
Behauptung: g ist surjektiv.
Zu zeigen ist: $g(M) = N$.
Es gilt:

$$g \circ f \text{ surjektiv} \Rightarrow (g \circ f)(L) = N = g(f(L))$$

$f(L) \subseteq M \Rightarrow g(f(L)) \subseteq g(M)$ (siehe Seite 30)

und zusammen:

$$N = g(f(L)) \subseteq g(M) \subseteq N$$

und damit folgt die Behauptung $g(M) = N$. □

Jetzt beweisen wir den zweiten Teil mit einem **indirekten Beweis**. Dabei geht man von der Annahme aus, das Gegenteil der Behauptung sei richtig und leitet eine falsche Aussage her. Dann kann die Annahme nicht korrekt sein, denn aus korrekten Aussagen folgen immer nur korrekte Aussagen, keine falschen.

Beginnen wir also erst einmal mit der
Voraussetzung: Es seien $f : L \to M$ und $g : M \to N$ Abbildungen, und $g \circ f$ sei bijektiv.
Behauptung: f ist injektiv.
Indirekter Beweis: Voraussetzung: f ist nicht injektiv.
Schlußkette:

f ist nicht injektiv.

$\Rightarrow$ Es gibt zwei Elemente $x_1 \neq x_2$ in L, so daß $f(x_1) = f(x_2)$ gilt.

$\Rightarrow$ $g(f(x_1)) = g(f(x_2))$

$\Rightarrow$ $(g \circ f)(x_1) = (g \circ f)(x_2)$

$\Rightarrow$ $(g \circ f)$ nicht injektiv

$\Rightarrow$ Widerspruch! □

Das obige Vorgehen besteht in abstrakte Sichtweise darin, eine Aussage C dadurch zu beweisen, daß man unter der Voraussetzung, daß C falsch sei, einen Widerspruch herleitet. Oft aber hat man eine Aussage C der Form

aus A folgt B

zu beweisen. Ein indirekter Beweis einer solchen Aussage verläuft so, daß man beweist

wenn B falsch ist, muss auch A falsch sein.

Wir werden in der Logik noch genauer analysieren, warum diese beiden Aussagen äquivalent sind, aber sie sind zumindestens für den naiven gesunden Menschenverstand dasselbe. Denn wenn aus A die Aussage B folgt, kann es nicht sein, daß B nicht zutrifft und A zutrifft. Und umgekehrt: Wenn man weiß, daß A immer falsch ist, sobald B falsch ist, kann man aus dem Zutreffen von A immer auf das Zutreffen von B schließen.

Wir holen damit hier den indirekten Beweis von Theorem 1.21 nach:

> Ist R eine Äquivalenzrelation auf M, so ist M die disjunkte Vereinigung der verschiedenen Äquivalenzklassen von Elementen von M.

Beweis: Es ist klar, daß jedes Element $x \in I\!R$ der Klasse $[x]$ angehört, also liegt M in der Vereinigung der Äquivalenzklassen. Umgekehrt ist die Vereinigung der Äquivalenzklassen eine Vereinigung von Teilmengen von M, also selber eine Teilmenge von M, siehe (1.10). Zu zeigen bleibt, daß zwei verschiedene Äquivalenzklassen disjunkt sind. Das kann man dadurch zeigen, daß man beweist, daß zwei Äquivalenzklassen gleich sind, wenn sie ein gemeinsames Element haben. Das ist ein indirekter Beweis.

Es wird also angenommen, die Äquivalenzklassen $[x]$ und $[y]$ hätten ein gemeinsames Element z. Dann gelten die Aussagen zRx und zRy, und wegen Symmetrie und Transitivität muß dann auch xRy gelten, d.h. x und y sind selber äquivalent. Dann sind aber auch die Klassen $[x]$ und $[y]$ gleich. Denn aus $u \in [x]$ folgt uRx, und wegen xRy folgt mit der Transitivität uRy und $u \in [y]$. das beweist $[x] \subseteq [y]$, und analog beweist man $[y] \subseteq [x]$. □

1.3.5 Identität und Umkehrabbildung

Definition 1.32

- *Ist M eine beliebige Menge, so wird die Abbildung von M in M, die jedes Element $x \in M$ auf sich selbst abbildet, die* **Identität** *oder* **identische Abbildung** *genannt und mit Id oder id oder Id_M bezeichnet. Sie ist bijektiv.*

- *Zu einer bijektiven Abbildung $f : M \to N$ gibt es eine eindeutig bestimmte* **Umkehrabbildung**[1] *oder* **inverse Abbildung**

$$f^{-1} : N \to M \text{ mit } f^{-1} \circ f = Id_M \text{ und } f \circ f^{-1} = Id_N,$$

d.h. $f^{-1}(f(x)) = x$ für alle $x \in M$ und $f(f^{-1}(y)) = y$ für alle $y \in N$.

[1] `http://de.wikipedia.org/wiki/Umkehrabbildung`

Achtung: Die Umkehrabbildung im obigen Sinne existiert nur zu einer **bijektiven** Abbildung. Die Abbildung $V \mapsto f^{-1}(V)$, die zu einer beliebigen Abbildung $f : M \to N$ die Urbildmengen $f^{-1}(V)$ von Bildmengen $V \subseteq N$ liefert, existiert immer, kann aber nicht auf Elemente von N, sondern nur auf Teilmengen angewendet werden. Manche Autoren unterscheiden diese beiden Abbildungen durch die Notationen f^{-1} und $\overset{-1}{f}$.

1.3.6 Exkurs zum Aufgabenlösen

Aufgabe: Man beweise: Die Umkehrabbildung einer bijektiven Abbildung ist bijektiv.

Das kann man "zu Fuß" machen, indem man auf die Definition von "bijektiv" und "Umkehrabbildung" zurückgeht und Schritt für Schritt den Beweis zusammenbaut. Oder man benutzt bereits bekannte Tatsachen und macht sich das Leben etwas leichter. Wie?
In solchen Fällen behaupten viele Studierende, sie hätten "keine Idee dazu". So etwas kann man aber lernen. Man macht sich wie bisher erst einmal klar, was man weiß und was man beweisen will:
Voraussetzung: $f : M \to N$ ist bijektiv.
Zu zeigen: $f^{-1} : N \to M$ ist bijektiv.
Man sieht in der Definition von "bijektiv" nach, was das bedeutet:
Zu zeigen: $f^{-1} : N \to M$ ist injektiv und surjektiv.
Man kramt auch die Definition der Umkehrabbildung heraus:
Voraussetzung: $f^{-1} \circ f = Id_M$ und $f \circ f^{-1} = Id_N$.
Jetzt sollte man in seinem Gedächtnis und in seinen Unterlagen nachsehen, wo man eine Aussage der Form "*... dann ist die Abbildung injektiv*" oder "*... dann ist die Abbildung surjektiv*" findet. Solche Aussagen kommen als Werkzeug in Frage, wenn man auf das Ziel sieht. Schaut man auf die Voraussetzungen, so muß man nach Aussagen suchen, die etwas über zusammengesetzte Abbildungen wie $f^{-1} \circ f$ und $f \circ f^{-1}$ voraussetzen und damit irgendetwas anstellen.

Blättern wir zurück, so finden wir in Theorem 1.31 auf Seite 33 etwas in dieser Art. Die vierte Aussage

- Ist $g \circ f$ bijektiv, so ist g surjektiv und f injektiv

genügt sogar **beiden** Anforderungen. Damit ist die Idee schon gefunden, denn wir können die Abbildung f^{-1} als g einsetzen und benutzen, daß $Id_M = f^{-1} \circ f = g \circ f$ bijektiv ist. Wir bekommen, daß f^{-1} surjektiv und f injektiv ist. Das ist nur die halbe Miete. Aber wenn man steckenbleibt, muß man sich

immer fragen, ob man auch schon alles, was man weiß, auch benutzt hat. Wir haben aber nur $Id_M = f^{-1} \circ f$ und nicht $Id_N = f \circ f^{-1}$ benutzt. Also werden wir die Aussage nochmal anwenden, jetzt aber unter Vertauschung von f und f^{-1}. Wir bekommen dann, daß f surjektiv und f^{-1} injektiv ist, und sind fertig, aber wir müssen alles noch sauber aufschreiben. □

Fazit: Man sollte, **bevor** man einen Beweis aufschreibt, zum "Finden" eines Beweises die folgenden Strategien anwenden:

1. Man schreibe alle Voraussetzungen hin, die man hat.

2. Man schreibe die Behauptung hin.

3. Man sehe sich zu allen darin vorkommenden Begriffen die Definitionen noch einmal an und formuliere damit die Voraussetzungen und die Behauptung um.

4. Dann sucht man nach Aussagen, die von Voraussetzungen ausgehen, die den gegebenen Voraussetzungen ähnlich sind. Man schreibe sie sich hin, und achte insbesondere auf die Konsequenzen, denn sie sollten ja in die Richtung der Behauptung gehen.

5. Dann sucht man nach Aussagen, die Konsequenzen haben, die der zu beweisenden Behauptung ähnlich sind. Man schreibe sie sich hin und achte auf die Voraussetzungen. Sie sollten möglichst ähnlich zu den bekannten Voraussetzungen sein. Sie kommen für einen direkten Beweis in Frage.

6. Dann sucht man nach Aussagen, die von Voraussetzungen ausgehen, die dem Gegenteil der zu beweisenden Behauptung ähnlich sind. Man schreibe sie sich hin. Man kann sie vielleicht für einen indirekten Beweis gebrauchen, und man sollte jetzt darauf achten, ob sie zu Unsinn führen, denn das ist bei einem indirekten Beweis das Gesuchte.

7. Man muß sich jetzt zwischen einem direkten und einem indirekten Beweis entscheiden. Das hängt davon ab, was man im vorigen Schritt gefunden hat. Man kann beide Möglichkeiten probieren, weil man ja noch lange nicht den formalen Beweis aufschreibt, sondern immer noch "sucht".

8. Jetzt beginnt ein Puzzlespiel. Man versucht, aus den aufgeschriebenen Bausteinen einen Beweis zusammenzubasteln. Das gelingt in der Regel nicht auf Anhieb. Aber man sollte, wenn es nicht funktioniert, auf folgendes achten:

- Habe ich schon alle Voraussetzungen benutzt?
- Wo stehe ich? Welche Aussagen folgen aus den Voraussetzungen, welche Aussagen, wenn sie denn schon bewiesen wären, würden die Behauptung liefern?
- Was fehlt mir? Kann ich das irgendwoher bekommen?
- Gibt es einfachere Zwischenziele?

Es ist wie bei einem Brückenbau, weil man von den Voraussetzungen bis hin zur Behauptung eine Kette von tragfähigen Schlüssen finden muß. Man braucht zum erfolgreichen Bau einer Brücke

- einen Überblick über das gesamte zur Verfügung stehende Baumaterial,
- eine gute Kenntnis der beiden Ufer und je einen soliden Pfeiler dort.
- Bei einer halbfertigen Brücke muß man genau wissen, welche Stücke noch fehlen, und
- wenn die Brücke lang werden soll, und man nicht sieht, wie man die beiden Ufer mit einem Stück überspannen kann, sollte man erst einmal ein paar freistehende Pfeiler hinsetzen, um eine Kette kleinerer Brücken zu bauen.

Bei der Lösung der mathematischen Probleme im Standard–Studium reichen diese Strategieschritte aus. "*Ich habe keine Idee dazu...*" ist eine faule Ausrede.

1.3.7 Gleichmächtigkeit

Definition 1.33 *Zwei Mengen M und N heißen* **gleichmächtig**[1]*, wenn es eine bijektive Abbildung $f : M \to N$ gibt.*

Theorem 1.34

- *Gleichmächtigkeit ist eine Äquivalenzrelation auf der Menge aller Mengen.*
- *Endliche Mengen mit gleicher Anzahl von Elementen sind gleichmächtig.*

[1] `http://de.wikipedia.org/wiki/Gleichm%C3%A4chtig`

Man lese sich hierzu durch, was in [4] auf S. 22–23 steht. Und man kann sich dazu selbst einen Beweis überlegen.

Gleichmächtigkeit ist bei endlichen Mengen dieselbe Äquivalenzrelation wie "hat gleiche Elementzahl". Bei unendlichen Mengen kann es aber vorkommen, daß eine Menge zu einer echten Untermenge gleichmächtig ist. Beispielsweise sind $\mathbb{N} := \{0, 1, 2, 3, \ldots\}$ und $\mathbb{N} \setminus \{0\} := \{1, 2, 3, \ldots\}$ durch die bijektive Abbildung $f(x) := x + 1$ gleichmächtig. Ebenso $\mathbb{N}$ und $2\mathbb{N} := \{0, 2, 4, 6, 8, \ldots\}$ durch $f(x) = 2x$.

Definition 1.35 *Eine Menge M heißt* **abzählbar unendlich**[1]*, wenn sie gleichmächtig zu $\mathbb{N}$ ist.*

Es ist klar, daß jede Menge M, die man in der Form $M = \{x_1, x_2, x_3, \ldots\}$ mit paarweise verschiedenen Elementen x_j schreiben kann, abzählbar ist. Deshalb ist $\mathbb{Z} := \{0, 1, -1, 2, -2, 3, -3, \ldots\}$ abzählbar, d.h. die unendlichen Mengen $\mathbb{N}$ und $\mathbb{Z}$ sind gleichmächtig. Ebenso werden wir später sehen, daß die reellen Zahlen $\mathbb{R}$ und die Potenzmenge $P(\mathbb{N})$ gleichmächtig sind. Aber: $\mathbb{N}$ und $\mathbb{R}$ sind **nicht** gleichmächtig, wie ein schönes Argument von **Cantor**[2] zeigt, das wir noch ansehen werden. Es gibt also mindestens zwei "Arten" von "Unendlich", nämlich "abzählbar" und "überabzählbar".

Die Frage ob es eine Menge M gibt, die weder zu $\mathbb{N}$ noch zu $\mathbb{R}$ gleichmächtig ist, aber zu einer Teilmenge von $\mathbb{R}$, war lange Zeit offen und ist es in gewissem Sinne immer noch. Denn Kurt **Gödel**[3] hat 1940 bewiesen, daß diese Aussage (die spezielle **Kontinuumshypothese**)[4] aus den üblichen Axiomen der Mathematik nicht widerlegt werden kann. Paul Cohen[5] bewies 1963, dass sie weder beweisbar noch widerlegbar ist.

Ist $f : M \to N$ eine Abbildung, so kann man eine Äquivalenzrelation A_f auf $M \times M$ definieren durch

$$uA_fv \leftrightarrow f(u) = f(v) \text{ für alle } u, v \in M.$$

Frage: Ist das wirklich eine Äquivalenzrelation?

Theorem 1.36 *Ist $f : M \to N$ eine Abbildung, so ist die Menge der Äquivalenzklassen von A_f gleichmächtig zur Menge $f(M) \subseteq N$.*

[1] http://de.wikipedia.org/wiki/Abz%C3%A4hlbar
[2] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Cantor.html
[3] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Godel.html
[4] http://de.wikipedia.org/wiki/Kontinuumshypothese
[5] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Cohen.html

Wie sieht die zugehörige Abbildung aus? Wir wollen natürlich

$$F([x]) := f(x) \text{ für alle } x \in M \tag{1.37}$$

definieren, und wenn wir die Menge aller Äquivalenzklassen von M unter der Relation A_f als M/A_f definieren, soll das eine Abbildung

$$F : M/A_f \to F(M)$$

werden. Aber in (1.37) ist die Abbildung nicht direkt durch das Urbild $[x]$ ausgedrückt, sondern durch $f(x)$. Dann muß man noch **Wohldefiniertheit** der Abbildung nachweisen, denn das Bild von $[x]$ unter F darf nur von $[x]$ abhängen, nicht von x. Wenn man $[x] = [y]$ hat, muß auch $f(x) = f(y)$ folgen, sonst hat man ein Problem. Aber das ist ja gerade der Inhalt unserer Äquivalenzrelation, so daß F wohldefiniert ist.

Aufgabe: Man schreibe einen formal korrekten Beweis für die Bijektivität dieser Abbildung hin.

Um Mißverständnissen vorzubeugen: Man muß für eine Abbildung $F : M \to N$ immer dann Wohldefiniertheit nachweisen, wenn man den Wert $F(m)$ für ein $m \in M$ nicht durch m selbst, sondern über einen irgendwie von m abhängigen Ausdruck $A(m)$ definiert. Dann muß man zeigen, daß aus $m = n$ für Elemente $m, n \in M$ auch immer $A(m) = A(n)$ folgt, denn sonst ist F nicht sauber definiert. In unserem Beispiel ist $m = [x]$, aber wir arbeiten mit $A(m) = f(x)$, und deshalb müssen wir zeigen, daß aus $[x] = [y]$ immer $f(x) = f(y)$ folgt.

Das "Strickmuster" von Satz 1.36 kommt in der Mathematik an verschiedenen Stellen wieder vor. Wenn man eine beliebige Abbildung $f : M \to N$ mit Brachialgewalt umkehren will, so kann man das zunächst nur auf $f(M) \subseteq N$, weil die Elemente von $N \setminus F(M)$ gar keine Urbilder haben. Dort aber kann man zu einem Element $z \in f(M) \subseteq N$, das sich eventuell als $z = f(x) = f(y)$ mit verschiedenen $x, y \in M$ schreiben läßt, nicht klar sagen, ob man z auf x oder y abbilden soll. Aber man kann auf die Äquivalenzklasse $[x] := \{y \in M : f(x) = f(y)\}$ gefahrlos abbilden.

Wenn man z.B. Personen auf ihre Nachnamen abbildet (vgl. das Beispiel auf Seite 23), so hat man natürlich keine surjektive Abbildung, weil es Personen mit gleichem Nachnamen gibt, und die Abbildung ist nicht umkehrbar. Aber man kann zu jedem Nachnamen die Äquivalenzklasse der Personen mit diesem Namen bilden, und das definiert ganz sauber eine Abbildung.

Frage: Wie sehen die Äquivalenzklassen aus, wenn man die Abbildung $f : \mathbb{R} \to [0, \infty)$ mit $f(x) = x^2$ betrachtet?.

Frage: Wie sehen die Äquivalenzklassen aus, wennn man die Abbildung $f : \mathbb{R} \to \mathbb{R}^2$ mit $f(x) = (\cos(x), \sin(x))$ betrachtet?.

2 Sprache und Logik

Wir haben bisher ja schon "logisch" argumentiert, ohne genauer definiert zu haben, was eigentlich "logisch" ist. Um das genauer zu fassen, braucht man etwas Mengenlehre, aber für die Mengenlehre braucht man etwas Logik. Aus diesem Teufelskreis kommt man nur heraus, wenn man erst auf einer naiv–intuitiven Ebene die Logik und die Mengenlehre einführt, um beide dann später auf einer streng formalen Ebene klarer zu fassen. Das haben wir schon zu Beginn des ersten Kapitels angedeutet, als wir die Mengenlehre einführten. Wir sehen uns jetzt die Logik an, und es kommt nebenbei heraus, daß die formal–logischen Grundlagen von Mathematik und Informatik kaum unterscheidbar sind. Wir berühren dabei fundamentale Fragen der Philosophie, können uns aber nicht auf Seitenwege einlassen.

Es geht hier natürlich vor allem darum, mathematische Aussagen zu machen, die in einem gewissen Sinne "wahr" sind. In der Informatik redet man eher davon, daß ein Programm "korrekt" ist, aber das ist nichts wesentlich anderes, weil man z.B. wissen will, ob die Aussage "Das Programm P löst die Aufgabe A" wahr ist. Weil man eine solche Aussage mit mathematischer Exaktheit formulieren kann, ist sie nicht von einer mathematischen Aussage verschieden. Es ist also nötig, klar zu sagen, wie man einer Aussage "Wahrheit" beimißt, was eine "Aussage" ist und wovon wir überhaupt sagen können, daß wir es "wissen". Die Kulturgeschichte zeigt, daß so etwas nicht einfach ist und mit dem Verstehen von Sprache zusammenhängt:

- "Was ist Wahrheit?" (P. Pilatus, Joh. 18,38)
- "Was sich überhaupt (aus–)sagen läßt, läßt sich klar sagen; und wovon man nicht reden kann, darüber muß man schweigen" (Ludwig Wittgenstein[1], 1859 - 1951)

2.1 Aussagen und Aussagenlogik

2.1.1 Zeichen, Alphabete, Worte und Sprachen

Wir beginnen mit Begriffen, die für Informatik und mathematische Logik grundlegend sind:

Definition 2.1 *1. Ein* **Zeichen**[2] *ist ein nicht näher erklärtes Symbol wie* `a` *oder* x *oder* $\in$.

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Biographies/Wittgenstein.html`
[2] `http://de.wikipedia.org/wiki/Zeichen`

2. *Zeichen lassen sich zu geordneten* **Zeichenketten**[1] *hintereinandersetzen:*

$$x \in M \quad \textit{oder} \quad \texttt{diesisteineZeichenkette}$$

3. *Ein* **Alphabet**[2] *ist eine endliche Menge von Zeichen.*

4. *Ein* **Wort** *oder* **Satz** *der Länge n über dem Alphabet A ist eine Verkettung von n Zeichen aus A. Die Menge dieser Worte wird mit A^n bezeichnet. Mit ϵ wird in der Informatik das* **leere Wort** *bezeichnet, das keine Zeichen hat.*

5. *Die* **freie Sprache** *oder* **Kleene'sche Hülle**[3] *A^* über einem Alphabet A besteht aus der Menge aller Worte aller Längen.*

6. *Eine (formale)* **Sprache**[4] *S über einem Alphabet A ist eine Teilmenge von A^*.*

Es sollte klar sein, daß es nicht sehr schadet, daß die obige Definition von A^n nicht mit der des cartesischen Produkts A^n aus (1.15) übereinstimmt. In der Regel haben Sprachen ein **Leerzeichen** oder **Trennzeichen**, und deshalb muß man nicht zwischen einzelnen Worten und ganzen Sätzen unterscheiden. In der **theoretischen Informatik** werden Sprachen im Sinne der obigen Definition genauer untersucht. **Programmiersprachen** sind die wichtigsten Beispiele.

Eines der wichtigsten Probleme der Informatik besteht darin, ein effizientes Verfahren zu haben, das ein **Wortproblem**[5] löst. Das besteht darin, bei fest gegebener Sprache $S \subseteq A^*$ ein effizientes Programm zu haben, das zu jedem beliebig vorgegebenen Wort $w \in A^*$ entscheidet, ob es ein legitimes Wort der Sprache S ist. Diese Situation liegt vor, wenn ein Compiler ein Programm w darauf prüft, ob es syntaktisch korrekt ist. Natürlich ist das Wortproblem umso schwieriger, je komplexer die Sprache S ist. Deshalb lernt man in der Theoretischen Informatik, daß es Komplexitätshierarchien von Sprachen gibt, die genaue Entsprechungen in Komplexitäten von Maschinenmodellen haben, auf denen Algorithmen zur Lösung des Wortproblems ablaufen.

[1] `http://de.wikipedia.org/wiki/Zeichenkette`
[2] `http://de.wikipedia.org/wiki/Alphabet`
[3] `http://de.wikipedia.org/wiki/Kleenesche_H%C3%BClle`
[4] `http://de.wikipedia.org/wiki/Formale_Sprache`
[5] `http://de.wikipedia.org/wiki/Wortproblem`

2.1.2 Wahrheitswerte

Um mit "wahr" und "falsch" umgehen zu können, brauchen wir

Definition 2.2 *Die Menge B der formalen* **Wahrheitswerte**[1] *ist je nach Geschmack*

$$B := \{wahr, falsch\} = \{true, false\} = \{W, F\} = \{T, F\} = \{1, 0\}.$$

Dabei sollte man *wahr* und *falsch* als abstrakte Zeichen oder Objekte sehen, deren Schreibweise und "Sinn" irrelevant sind. Wir nehmen die Bezeichnung B wegen der Beziehung zur **Booleschen Algebra** und zum Datentyp `boolean` mancher Programmiersprachen, und wir wollen uns aus den Problemen der mehr als zweiwertigen Logik heraushalten.

Definition 2.3 *Es sei S eine Sprache über einem Alphabet A. Ferner sei T eine Teilmenge von S, und es gebe eine Abbildung $I : T \to B$. Dann heißen die Elemente von T logische* **Aussagen**[2]*, und die Abbildung I heißt* **Interpretation**[3].

Etwas laxer formuliert: Aussagen sind Sprachsätze, die unter einer gegebenen Interpretation einen Wahrheitswert haben.

Beispiele:

- 2 ist kleiner als 7
- 7 ist kleiner als 2
- Die Globalisierung ist ein Segen für die Menschheit
- Das Leben ist durch Schöpfung enstanden
- Das Leben ist durch physikalisch–chemisch–biologische Evolution entstanden

Die hier unterstellte Sprache ist die deutsche Umgangssprache, die Interpretationsabbildung wird durch den gesunden Menschenverstand geliefert. Die ersten beiden Beispiele verdeutlichen, daß Aussagen wahr oder falsch sein können, während die anderen ihren Wahrheitswert ändern, wenn sich die Interpretation ändert.

[1] `http://de.wikipedia.org/wiki/Wahrheitswert`
[2] `http://de.wikipedia.org/wiki/Logische_Aussage`
[3] `http://de.wikipedia.org/wiki/Interpretation_%28Logik%29`

Bilder sind auch Worte einer Sprache, weil sie als Folgen von Zeichen (Farbcodes von Pixeln) dargestellt werden.

Wir haben hier den Begriff der Interpretation sehr eng gefaßt, weil wir als Interpretationsergebnis nur *wahr* und $falsch$ zulassen, denn wir beschränken uns auf Aussagen. Jede Art von "Verstehen" eines Sprachsatzes ist aber eine Interpretation, auch wenn das "Verstehen" im menschlichen Bewußtsein abläuft oder im Sinne des "Verstehens" der Künstlichen Intelligenz aus einer Reaktion eines Computers auf den eingegebenen Sprachsatz besteht ("Das gegebene Bild zeigt ein Auto").

Es ist für Informatik–Studierende wichtig zu wissen, daß ein Computer **immer** nur Sprachsätze interpretiert. Die Sprachen können auf verschiedenen Ebenen liegen und sehr verschieden sein:

- Programmiersprachen wie C und Java,
- Beschreibungssprachen wie HTML und XML,
- Texte und Bilder,
- Maschinencode aus Bits und Bytes,

und die Interpretationsabbildungen sind dementsprechend auch sehr verschieden, aber sie bestehen immer aus einer Veränderung des Zustands des Computers.

Man kann aber auch den Zwischenschritt über die Sprache unterdrücken und etwa zu jeder ganzen Zahl x die Aussage "x ist gerade" als Abbildung von den ganzen Zahlen in B auffassen. Das ist also eine Abbildung von einer beliebigen Menge in die zweielementige Menge der Wahrheitswerte:

Definition 2.4 *Ein n–stelliges* **Prädikat**[1] *auf einer Menge M ist eine Abbildung von M^n in B.*

Die Begriffe "Prädikat" und "Relation" sind allerdings nicht wesentlich verschieden, und manche Autoren benutzen den Begriff "Relation" so, wie wir "Prädikat" verwenden. Wir haben in Definition 1.17 auf Seite 21 definiert, was eine n–stellige Relation auf einer Menge M sein soll, nämlich eine Teilmenge von M^n. Der Zusammenhang zu einem Prädikat wird aber sofort klar, wenn man sich eine Teilmenge R von M^n hernimmt und die Abbildung

$$r \ : \ M^n \to B, \ r(x) := \left\{ \begin{array}{cc} wahr & \text{wenn } x \in R \\ falsch & \text{wenn } x \notin R \end{array} \right\}$$

[1] http://de.wikipedia.org/wiki/Pr%C3%A4dikat_%28Logik%29#Das_Pr.C3.A4dikat_in_der_mathematis

betrachtet. Das ist dann ein Prädikat im Sinne der obigen Definition.

Frage: Wie kommt man umgekehrt von der Definition 2.4 eines Prädikats zur Definition 1.17 einer Relation?

Aussagen sind einstellige Prädikate auf einer speziellen Menge, nämlich einer Sprache mit einer Interpretation. Insofern können wir ab sofort ausschließlich von Prädikaten reden. Als **Aussagenvariablen** oder **Prädikatenvariablen** bezeichnen wir Symbole wie A und B, die für beliebige Aussagen oder Prädikate stehen können.

Auch hier sollte der Bezug zur Informatik klar sein: Aussagen und Prädikate sind Sprachelemente, die in Bedingungen von Programmiersprachen vorkommen, z.B.

if	Prädikat	then	Block	endif
do	Block	while	Prädikat	

wobei die jeweilige Bedingung als erfüllt gilt, wenn die Auswertung (die Interpretation zur Laufzeit) des Prädikates den Wahrheitswert "wahr" ergibt.

2.1.3 Aussagenlogische Grundoperationen

Im normalen Sprachgebrauch können wir jede Aussage **negieren**, d.h. ihr logisches Gegenteil angeben. Das erfordert in der Umgangssprache manchmal einige Verrenkungen, aber entscheidend ist, dass wir zu einer Aussage A eine andere, mit $\neg A$ bezeichnete produzieren können, deren Wahrheitswert dem der Aussage A genau entgegengesetzt ist. Beispiele werden mündlich angegeben, und das Grundschema ist in der Umgangssprache

$$\neg A = \text{ es ist nicht wahr, daß } A \text{ wahr ist}$$

mit der herkömmlichen Interpretation. Formal kann man das so fassen, daß man $\neg$ (sprich: "nicht") als eine bijektive Abbildung von B in B mit

$$\neg(wahr) := falsch,\ \neg(falsch) := wahr$$

definiert, und dann kann man zu jedem Prädikat P immer ein Prädikat $\neg P$ als **Negation**[1] von P definieren mit

$$(\neg P)(x) := \neg(P(x))$$

[1] `http://de.wikipedia.org/wiki/Negation#Logik`

für alle x aus dem Definitionsbereich von P. Statt "$P(x)$ ist wahr " oder "$P(x)$ ist falsch " schreibt man dann auch einfach $P(x)$ oder $\neg P(x)$.

Genau so kann man mit zweistelligen Verknüpfungen von Aussagen und Prädikaten verfahren. Mit $\wedge$ bzw. $\vee$ (sprich "und" bzw. "oder", **Konjunktion**[1] und **Disjunktion**[2]) bezeichnet man die Verknüpfungen zweier Aussagen A und B durch "und" bzw. (nicht–ausschließendes) "oder". Umgangssprachlich ist also $A \wedge B$ genau dann wahr, wenn A und B beide wahr sind, während $A \vee B$ genau dann wahr ist, wenn A oder B oder beide wahr sind. Zusammen mit der "wenn–dann" oder Folgerungs–Operation $\rightarrow$ und der "genau–dann–wenn"–Operation $\leftrightarrow$ ergibt sich folgende Wertetabelle mit den Abkürzungen w und f für $wahr$ und $falsch$:

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \rightarrow B$	$A \leftrightarrow B$
w	w	f	w	w	w	w
f	w	w	f	w	w	f
w	f	f	f	w	f	f
f	f	w	f	f	w	w

(2.5)

Dabei stehen links in den ersten beiden Spalten die vier möglichen Kombinationen der Wahrheitswerte von A und B, die Ergebnisse der Verknüpfungen stehen rechts in den folgenden Spalten.

Die Aussage $A \leftrightarrow B$ ist unproblematisch: sie trifft genau dann zu, wenn A und B dieselben Wahrheitswerte haben. Bei der Folgerungsoperation $A \rightarrow B$ haben Anfänger aber immer Schwierigkeiten. Der Sinn der zusammengesetzten Aussage, daß aus A immer zwangsläufig B folgt, ist der, daß es nicht sein kann, daß B falsch ist und gleichzeitig A wahr ist. Die Aussage $A \rightarrow B$ muß also wie $\neg(A \wedge (\neg B))$ definiert werden. Das kann man aber auch so ausdrücken, daß die Aussage

$$C := (\neg(A \wedge (\neg B))) \leftrightarrow (A \rightarrow B)$$

immer wahr sein muß.

Es gibt eine Standardmethode, so etwas nachzuprüfen, und die sollte man üben. Man schreibt sich zunächst alle vier Möglichkeiten der Wahrheitswerte von A und B hin:

A	B
w	w
f	w
w	f
f	f

[1] `http://de.wikipedia.org/wiki/Konjunktion_%28Logik%29`
[2] `http://de.wikipedia.org/wiki/Disjunktion`

Nach rechts baut man für jeden Zwischenausdruck eine Spalte an, und zwar so, daß man für die jeweilige nächste Spalte immer nur eine Operation auswerten muß.

A	B	$\neg B$	$A \wedge (\neg B)$	$\neg(A \wedge (\neg B))$	$A \to B$	C
w	w					
f	w					
w	f					
f	f					

Dann trägt man Spalte für Spalte die Ergebnisse ein

A	B	$\neg B$	$A \wedge (\neg B)$	$\neg(A \wedge (\neg B))$	$A \to B$	C
w	w	f	f	w	w	w
f	w	f	f	w	w	w
w	f	w	w	f	f	w
f	f	w	f	w	w	w

indem man auf die Grundtabelle (2.5) zurückgeht. Man sieht, daß C immer wahr ist, wir haben also die Folgerungsoperation korrekt definiert. An dieser Stelle merken wir uns noch einmal, daß man aus Unsinn etwas Richtiges folgern kann, denn $A \to B$ ist **immer** wahr, außer wenn A wahr ist und B falsch.

Die Operationen

- Negation $\neg$ ("nicht")
- Konjunktion $\wedge$ ("und")
- Disjunktion $\vee$ ("oder")

sind die Grundoperationen der **Aussagenlogik**[1].

In Analogie zu den Theoremen 1.9 und 1.13 gilt

Theorem 2.6 *Für die logischen Operationen gelten die Regeln*

$$
\begin{array}{rcll}
A \vee B & \leftrightarrow & B \vee A & \textit{Kommutativität von } \vee \\
A \wedge B & \leftrightarrow & B \wedge A & \textit{Kommutativität von } \wedge \\
(A \vee B) \vee C & \leftrightarrow & A \vee (B \vee C) & \textit{Assoziativität von } \vee \\
(A \wedge B) \wedge C & \leftrightarrow & A \wedge (B \wedge C) & \textit{Assoziativität von } \wedge \\
(A \vee B) \wedge C & \leftrightarrow & (A \wedge C) \vee (B \wedge C) & \textit{Distributivität von } \vee \textit{ und } \wedge \\
(A \wedge B) \vee C & \leftrightarrow & (A \vee C) \wedge (B \vee C) & \textit{Distributivität von } \wedge \textit{ und } \vee \\
\neg(A \vee B) & \leftrightarrow & (\neg A) \wedge (\neg B) & \textit{De Morgan'sche Formel} \\
\neg(A \wedge B) & \leftrightarrow & (\neg A) \vee (\neg B) & \textit{De Morgan'sche Formel} \\
\neg(\neg A) & \leftrightarrow & A. &
\end{array}
$$

[1] `http://de.wikipedia.org/wiki/Aussagenlogik`

Jede dieser Formeln kann man nach obigem Muster nachweisen.

Aufgabe: man mache dies für eine beliebige dieser Formeln..

2.1.4 Boolesche Funktionen

Definition 2.7 *Eine Abbildung f von B^n in B nennt man n–stellige* **Boolesche Funktion**[1] *oder* **Aussageformel**.

Wir haben oben schon Beispiele gesehen:

$$\begin{array}{rcl} f(A) & := & \neg A \\ f(A,B) & := & A \vee B \\ f(A,B) & := & A \wedge B \\ f(A,B) & := & (\neg(A \wedge (\neg B))) \leftrightarrow (A \rightarrow B) \end{array}$$

und alle Formeln aus Theorem 2.6.

Weil ein Rechner aus Schaltlogik besteht, kann man mit Fug und Recht sagen, daß er nichts anderes als Boolesche Funktionen ausrechnet.

Definition 2.8

- *Man nennt eine Aussageformel* **allgemeingültig**, *wenn sie für alle möglichen Wahrheitswerte ihrer Argumente immer* wahr *liefert.*
- *Man nennt eine Aussageformel* **erfüllbar**, *wenn es eine Wahl von Wahrheitswerten für ihre Argumente gibt, bei der sie den Wert* wahr *liefert.*

Man mache sich klar, daß die Aussageformel $f(A) := A \wedge (\neg A)$ nicht erfüllbar, die Aussageformel $f(A) := A \vee (\neg A)$ aber allgemeingültig ist. Wie man so etwas beweist, ist schon oben vorgeführt worden.

Beispiele für allgemeingültige Boolesche Funktionen sind die in Theorem 2.6 angegebenen Regeln.

In der theoretischen Informatik spielt das **Erfüllbarkeitsproblem**[2] eine sehr wichtige Rolle. Es besteht darin, ein Verfahren anzugeben, das zu jeder gegebenen Booleschen Funktion schnell entscheidet, ob sie erfüllbar ist oder nicht.

[1] `http://de.wikipedia.org/wiki/Boolesche_Funktion`
[2] `http://de.wikipedia.org/wiki/Erf%C3%BCllbarkeitsproblem_der_Aussagenlogik`

Wenn man die Theoreme 1.9 und 1.13 aus der Mengenlehre mit Theorem 2.6 vergleicht. stellt man fest, daß die leere Menge keine Entsprechung in der Logik zu haben scheint. Das ist aber nicht so, denn die leere Menge $\emptyset$ entspricht einer nicht erfüllbaren Aussageformel, während die in Theorem 1.13 unterstellte gemeinsame Obermenge einer allgemeingültigen Formel entspricht.

Aufgabe: Warum ist das so?

Wir werden hier die **Boolesche Algebra**[1] und die **Verbandstheorie**[2], die auf den Gesetzen aus den Theoremen 1.9, 1.13 und 2.6 basieren, nicht im Detail ausführen. das gehört in die **Diskrete Mathematik**.

Wenn man komplizierte (nicht nur Boolesche) Funktionen aus einfachen zusammensetzt, muß man in der Regel Klammern setzen, um die Anwendungsreihenfolge der Funktionen festzulegen. Mit **Präzedenzregeln**[3] kann man sich das Leben etwas erleichtern, und für die logischen Operationen gilt

1. $\neg$ hat höchste Priorität.
2. Es folgt $\wedge$ und dann
3. schließlich $\vee$.

Statt $A \wedge (\neg B)$ kan man also $A \wedge \neg B$ schreiben.

Definition 2.9 *Eine Aussageformel hat* **disjunktive Normalform**[4]*, wenn sie als Disjunktion von Aussageformeln geschrieben werden kann, die selbst aus Konjunktionen aus einzelnen Variablen oder deren Negationen bestehen. Solche Konjunktionen werden* **Term** *oder* **Klausel** *genannt.*

Nach unserer obigen Festlegung von Prioritätsregeln sind also disjunktive Normalformen völlig klammerfrei. Beispiel:

$$f(A,B,C) := A \wedge \neg B \wedge C \vee \neg A \wedge B \wedge C$$

Das ist aber leider sehr unübersichtlich, und man sollte des besseren Verständnisses wegen entweder Klammern setzen oder

$$\begin{array}{rcrcrcr} f(A,B,C) & := & & A & \wedge & \neg B & \wedge & C \\ & & \vee & \neg A & \wedge & B & \wedge & C \end{array}$$

schreiben, indem man jeden durch Konjunktionen gebildeten Term in eine neue Zeile schreibt und die Zeilen durch $\vee$ verbindet.

[1] http://de.wikipedia.org/wiki/Boolesche_Algebra
[2] http://de.wikipedia.org/wiki/Verband_%28Mathematik%29
[3] http://de.wikipedia.org/wiki/Operatorrangfolge
[4] http://de.wikipedia.org/wiki/Disjunktive_Normalform

Theorem 2.10 *Jede Boolesche Funktion läßt sich in disjunktiver Normalform schreiben.*

Das wollen wir hier nicht beweisen, sondern der Vorlesung "Diskrete Mathematik" oder der theoretischen Informatik überlassen.

2.2 Prädikatenlogik

2.2.1 Quantoren

In Definition 2.4 hatten wir Prädikate eingeführt. Ist P ein Prädikat auf einer Menge M, so kann man P **erfüllbar** nennen, wenn es **ein** $x \in M$ **gibt**, so daß $P(x)$ wahr ist. Und P ist **allgemeingültig**, wenn $P(x)$ wahr ist **für alle** $x \in M$. In der **Prädikatenlogik**[1] führt man zu "es gibt" und "für alle" neue Bezeichnungen ein:

$$\begin{array}{ll} \exists \quad x \in M & \text{bedeutet "es gibt ein" } x \in M \\ \forall \quad x \in M & \text{bedeutet "für alle" } x \in M \end{array}$$

weil es eben ganz entscheidend ist, ob $P(x)$ nur für ein $x \in M$ oder für alle $x \in M$ wahr ist. Man nennt $\exists$ den **Existenzquantor** und $\forall$ den **Allquantor**. Manche schreiben die Quantoren anders:

$$\bigvee_{x \in M} \quad \text{statt } \exists x \in M$$

$$\bigwedge_{x \in M} \quad \text{statt } \forall x \in M$$

aber für diese "Keilschrift" kann ich mich nicht erwärmen, obwohl sie symmetrische Ästhetik besitzt. In Büchern und Artikeln vermeidet man $\exists$ und $\forall$ zur besseren Lesbarkeit, außer manchmal in Formeln.

2.2.2 Verwendung von Quantoren

Eine Zeichenkette wie

$$y = x^2$$

ist für sich genommen sinnlos. Anfänger schreiben oft so etwas hin und wundern sich über einen Punktabzug. Es ist nicht klar, was x und y sein sollen, und was "=" und die Zweierpotenz bedeuten. Diese Zeile nehmen wir als Bestandteil eines Prädikats

$$P \ : \ \mathbb{R} \times \mathbb{R} \to B \ \text{ mit } \ P(x,y) := (y = x^2) \text{ für alle } x, y \in \mathbb{R}$$

[1] `http://de.wikipedia.org/wiki/Pr%C3%A4dikatenlogik`

an und verstehen die Symbole wie in der Schule. Eine brauchbare Aussage (es ist egal, ob sie richtig oder falsch ist, es muß aber eine Aussage sein) wird daraus nur, wenn man eine der Alternativen befolgt:

- beide Argumente spezifiziert:
  $9 = 3^2$ (richtig)
  oder
  $15 = 4^2$ (falsch)

- ein Argument spezifiziert und eines quantisiert:
  $\exists x \in I\!R \ : \ 9 = x^2$ (richtig) oder
  $\exists x \in I\!R \ : \ -5 = x^2$ (falsch) oder
  $\forall y \in I\!R \ : \ y = 3^2$ (falsch)

- beide Argumente quantifiziert:
  $\forall x \in I\!R \ \exists y \in I\!R \ : \ y = x^2$ (richtig)
  $\forall y \in I\!R \ \exists x \in I\!R \ : \ y = x^2$ (falsch)
  $\forall y \in I\!R \ \forall x \in I\!R \ : \ y = x^2$ (falsch)
  $\exists y \in I\!R \ \exists x \in I\!R \ : \ y = x^2$ (richtig)

Es ist nämlich klar, daß ein Prädikat nur dann zu einer Aussage werden kann, wenn man alle Argumente entweder spezifiziert (Einsetzen eines speziellen Elementes des Definitionsbereichs) oder quantisiert. Mit anderen Worten:

> **Prädikate oder Formeln, in denen noch freie Variablen vorkommen, sind keine Aussagen.**

Sie sind im allgemeinen Aussage**formeln** im Sinne von Definition 2.7.

Besonders wichtig, weil in der Schule oft nicht mit der nötigen begrifflichen Präzision ausgeführt, ist das saubere Verwenden von Quantoren beim Lösen von Gleichungen oder Gleichungssystemen. Zwei Gleichungen wie

$$\begin{array}{rcrcr} 3x & + & 4y & = & 11 \\ x & - & y & = & -1 \end{array}$$

machen allein keinen Sinn. Normalerweise gehört dazu die Annahme, es gebe zwei reelle Zahlen x und y, so daß die Gleicgungen beide gelten. Es steht also ein doppelter Existenzquantor davor: $\exists\, x \in I\!R \, \exists\, y \in I\!R \ : \ \ldots$

Damit sind die beiden Variablen nicht mehr frei sondern quantisiert, und die Formeln machen Sinn.

Es hat sich eingebürgert, den Allquantor **nach** dem Prädikat folgen zu lassen, während der Existenzquantor dem Prädikat **vorangeht** und dann oft ein Doppelpunkt als "sodaß" vor dem Prädikat steht. Man liest dann

$$\exists\, x \in I\!R \ :\ x \leq y^2\ \forall y \in I\!R$$

als

> Es gibt ein x in $I\!R$, so daß x kleiner oder gleich y^2 ist für alle y in $I\!R$.

2.2.3 Negation von Quantoren

Nun zur Negation von Aussagen, die durch Quantifizeren von Prädikaten entstehen. Das ist einfach, denn die **Negation** von "Alle Katzen sind grau" ist die Aussage "Es gibt eine Katze, die nicht grau ist". Es gilt also für ein Prädikat P auf einer Menge M

$$\neg\,(P(x)\ \forall x \in M\ ) \leftrightarrow \exists x \in M\ :\ \neg P(x) \tag{2.11}$$

und

$$\neg\,(\exists x \in M\ :\ P(x)) \leftrightarrow \neg P(x)\ \forall x \in M. \tag{2.12}$$

2.2.4 Beweistechniken

Schon oben haben wir den direkten und den indirekten Beweis kennengelernt. Im Zusammenhang mit Quantoren kommt noch eine Variante hinzu: der Beweis durch **Gegenbeispiel**. Er kann verwendet werden, um eine mit dem Allquantor versehene Aussage $P(x)\ \forall x \in M$ zu widerlegen. Man gibt einfach ein $x \in M$ mit $\neg P(x)$ an und hat damit (2.11) ausgenutzt. Will man die Aussage nicht widerlegen, sondern beweisen, so nimmt man bei einem direkten Beweis ein ganz allgemeines $x \in M$ her und beweist $P(x)$. Ein typischer Anfängerfehler ist, dieses x nicht allgemein zu nehmen, sondern darüber irgendwelche speziellen Annahmen zu machen. Für einen indirekten Beweis von $P(x)\ \forall x \in M$ nimmt man an, es gebe ein $x \in M$ mit $\neg P(x)$ und arbeitet auf einen Widerspruch hin. Jetzt ist das x nicht mehr allgemein aus M, sondern es hat die zusätzliche Eigenschaft $\neg P(x)$, und das will man natürlich irgendwie ausnutzen.

2.2.5 Russell'sche Antinomie

Wie sind jetzt in der Lage, unsere Mengenlehre etwas genauer zu untersuchen. Die Methode zur Mengendefinition aus (1.3) besagt, daß man zu beliebigen Prädikaten auf Mengen M auch die Mengen

$$M_P := \{x\ :\ x \in M \text{ und } P(x) \text{ ist wahr}\} \tag{2.13}$$

definieren kann. Auf einer großen Menge $\mathcal{M}$, die alles enthält, was irgendwie Element oder Menge sein kann, sei $\in$ eine zweistellige Relation mit den üblichen Eigenschaften, und die obige Definitionsmethode soll natürlich so beschaffen sein, daß die Aussagen $M_P \in \mathcal{M}$ und

$$(x \in M_P) \leftrightarrow (x \in M) \wedge P(x) \quad \forall\, x \in M \tag{2.14}$$

zutreffen. Wir bilden dann das Prädikat $P(x) := \neg(x \in x)$ auf $\mathcal{M}$ und definieren

$$\mathcal{M}_P := \{x \,:\, x \in \mathcal{M} \text{ und } \neg(x \in x)\}$$

um mit Schrecken festzustellen, daß die Aussagen $\mathcal{M}_P \in \mathcal{M}_P$ und $\neg(\mathcal{M}_P \in \mathcal{M}_P)$ gleichbedeutend sind, weil $\mathcal{M}_P \in \mathcal{M}$ zutrifft und

$$(\mathcal{M}_P \in \mathcal{M}_P) \leftrightarrow (\mathcal{M}_P \in \mathcal{M}) \wedge \neg(\mathcal{M}_P \in \mathcal{M}_P)$$

durch Einsetzen von $x = \mathcal{M}_P$ und $M = \mathcal{M}$ in (2.14) gilt. Diese Katastrophe nennt man nach Bertrand **Russell**[1] die **Russellsche Antinomie**[2].

Eine Mathematik mit solch einem Widerspruch ist unzulässig, und es hat deshalb mehrere Versuche gegeben, die Mengenlehre zu sanieren. Die Reparaturmethode[3] von **Zermelo**[4] und **Fraenkel**[5] ist die einfachste. Sie führt ein zusätzliches Prädikat Mg auf $\mathcal{M}$ ein, welches anschaulich besagt, daß $Mg(x)$ wahr ist, wenn x eine "axiomatisch sauber definierte Menge" ist. Was bisher "Menge" genannt wurde, heißt jetzt "Klasse" und ist mit Vorsicht zu genießen. Man ersetzt dann die Mengenbildungsregeln (2.13) und (2.14) durch Klassenbildungsregeln

$$M_P := \{x \,:\, (x \in M) \wedge Mg(x) \wedge P(x)\}$$

und

$$(x \in M_P) \leftrightarrow (x \in M) \wedge Mg(x) \wedge P(x) \quad \forall\, x \in M.$$

Wenn man jetzt noch einmal versucht, die Russell'sche Antinomie herbeizuführen, landet man bei

$$(\mathcal{M}_P \in \mathcal{M}_P) \leftrightarrow (\mathcal{M}_P \in \mathcal{M}) \wedge Mg(\mathcal{M}_P) \wedge \neg(\mathcal{M}_P \in \mathcal{M}_P)$$

und das ist kein Widerspruch, weil daraus nur $\neg Mg(\mathcal{M}_P)$ und $\neg(\mathcal{M}_P \in \mathcal{M}_P)$ folgt. Man hat also die Existenz einer Klasse, die keine axiomatisch sauber definierte Menge ist und sich nicht selbst als Element enthält. Damit kann man leben.

[1] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Russell.html
[2] http://de.wikipedia.org/wiki/Russellsche_Antinomie
[3] http://de.wikipedia.org/wiki/Zermelo-Fraenkel-Mengenlehre
[4] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Zermelo.html
[5] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Fraenkel.html

2.3 Formales Beweisen

Man kann die Logik und ihre formalen Grundlagen auch durch eine Formalisierung der Beweisverfahren erweitern. Man formalisiert die möglichen Schlußregeln als Transformationen, die gültige Aussagen in gültige Aussagen überführen. Dann sagt man, eine Aussage A sei aus einer Menge B von Aussagen **ableitbar**, wenn es eine Folge von formalen Transformationen gibt, die A aus den Aussagen von B zu produzieren gestatten. Für die Behauptungen und die Beweise hat man also eine formale Sprache. Die Disziplin "**Maschinelles Beweisen**" der "**Künstlichen Intelligenz**" benutzt diese Technik.

Es ergibt sich die Frage, ob man dann alle Behauptungen, die man formulieren kann, auch beweisen oder widerlegen kann. Kurt **Gödel**[1] hat bewiesen, daß es in allen Sprachen, die es mindestens erlauben, von natürlichen Zahlen zu reden, immer Behauptungen gibt, die man weder beweisen noch widerlegen kann. Die Behauptung der speziellen **Kontinuumshypothese** ist ein Beispiel. Der tiefere Grund liegt darin, daß man überabzählbar viele Behauptungen aufstellen, aber nur abzählbar viele Beweise aufschreiben kann.

Es sollte deshalb nicht mehr überraschen, daß man in der Umgangssprache viele Behauptungen (z.B. religiöser Art) aufstellen kann, die keinen formalen und damit für alle Menschen und Maschinen nachvollziehbaren Beweis erlauben. Es ist allerdings auch die Rückfrage gestattet, ob es denn menschenwürdig sei, nur das als "wahr" anzuerkennen, was auch formal oder maschinell beweisbar ist. Die formal nicht beweisbaren Aussagen (z.B. über Glaube, Liebe, Hoffnung...) sind aber in der Regel für das menschliche Zusammenleben die interessanteren.

2.4 Mengen und Logik

Nun ist es Zeit für einen kleinen Rückblick. Will man in der Informatik das Rechnen mit Mengen, ihren Teilmengen und ihren Relationen implementieren, so macht man das für feste gegebene endliche Mengen. Ist $M := \{x_1, \ldots, x_m\}$ eine solche Menge, so fixiert man die Elemente in einer bestimmten Reihenfolge, zum Beispiel als $x_1, \ldots, x_m$ und beschreibt Teilmengen durch m–Tupel mit Nullen und Einsen (**Bitvektoren**). Einer Teilmenge $N \subseteq M$ entspricht dann der Bitvektor

$$b_N := (b_1, \ldots, b_m) \text{ mit } b_j = \left\{ \begin{array}{ll} 1 & \text{falls } x_j \in N \\ 0 & \text{falls } x_j \notin N \end{array} \right\}.$$

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Godel.html`

Die logischen Operationen $\vee$, $\wedge$ und $\neg$ definiert man auf der Menge $\{0,1\}$ genau wie auf $\{falsch, wahr\}$ mit der Entsprechung $1 = wahr$, $0 = falsch$. Hat man dann zwei Teilmengen K und L von M, so kann man $L \cap K$, $L \cup K$ und $\overline{L}$ durch komponentenweises $\vee$, $\wedge$ und $\neg$ auf den zugehörigen Bitvektoren ausrechnen:

$$\begin{aligned} B_{L\cap K} &= B_L \wedge B_K \\ B_{L\cup K} &= B_L \vee B_K \\ B_{\overline{L}} &= 1 - B_L \end{aligned}$$

wobei wir die Operation $1 - B_L$ als komponentenweise **Bitinversion** verstehen (aus 0 wird 1 und umgekehrt).

Die Anzahl aller möglichen Teilmengen von $M := \{x_1, \ldots, x_m\}$, also die Anzahl der Elemente von $P(M)$, ist dann gleich der Anzahl der Bitvektoren mit m Komponenten, also 2^m.

Eine Relation R auf dem cartesischen Produkt $M \times N := \{x_1, \ldots, x_m\} \times \{y_1, \ldots, y_n\}$ von endlichen Mengen kann man dann als rechteckiges Bitschema mit m horizontalen Zeilen und n vertikalen Spalten hinschreiben:

	y_1	$\ldots$	y_k	$\ldots$	y_n
x_1	b_{11}	$\ldots$	b_{1k}	$\ldots$	b_{1n}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_j	b_{j1}	$\ldots$	b_{jk}	$\ldots$	b_{jn}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_m	b_{m1}	$\ldots$	b_{mk}	$\ldots$	b_{mn}

Das Element b_{jk} an der Kreuzungsstelle von j–ter Zeile und k–ter Spalte ist gleich Eins, wenn $x_j R y_k$ gilt, sonst Null. Man kann sich das ganze Schema auch als eine Darstellung von $M \times N$ vorstellen, so daß eine Teilmenge R nach dem schon oben Gesagten durch einen Bitvektor zu dieser Menge dargestellt wird. Dieser Bitvektor ist nun aber kein $m \cdot n$–Tupel, sondern ein rechteckiges Schema (**Matrix**).

Gilt $M = N$, so ist das Schema nicht nur rechteckig, sondern sogar quadratisch. Man mache sich klar, daß es in der Diagonale (von oben links nach unten rechts) Einsen enthalten muß, wenn die Relation **reflexiv** ist. Im Falle einer **symmetrischen** Relation ist das Schema zur Diagonalen spiegelsymmetrisch, weil immer $b_{jk} = b_{kj}$ gilt.

3 Zahlen

Ein großer Teil der Mathematik handelt von Zahlen:

- den **natürlichen** Zahlen $0, 1, 2, \ldots$
- den **ganzen** Zahlen $\ldots, -2, -1, 0, 1, 2, \ldots$
- den **rationalen** Zahlen (Brüchen aus ganzen Zahlen)
- den **reellen** Zahlen, bei denen noch u.a. $\sqrt{2}$ und π hinzukommen,
- den **komplexen** Zahlen, die außerdem eine exotische "Zahl" i enthalten, deren Quadrat $i \cdot i$ auf geheimnisvolle Weise gleich -1 ist.

Aber das Wichtigste bei den Zahlen sind die üblichen Rechenoperationen $+, -, \cdot, /$ und die dafür geltenden Rechenregeln. Alles Zählen, Messen und Bewerten in anderen Wissenschaften, bis hin zum Berechnen der Kontostände in der Betriebswirtschaft, der Auswertung von Statistiken, der computergestützten Auswertung von Computertomogrammen, dem Abspielen digitaler Signale aus einem CD– oder MP3–Player, alles beruht auf der Verarbeitung von Zahlen. Die Zahlen werden hier, obwohl sie teilweise aus der Schule schon gut bekannt sein sollten, in der oben angegebenen Reihenfolge sauber definiert. Weil sich die komplexen Zahlen am besten in einem zweidimensionalen Vektorraum veranschaulichen lassen, werden sie auf Abschnitt 4.2 verschoben.

3.1 Natürliche Zahlen

3.1.1 Peano–Axiome

Definition 3.1 *([4], Def. 3.1, S. 44)*
Die Menge $\mathbb{N}$ ist nach **Peano**[1] *definiert durch die* **Peano–Axiome**[2]

1. $0 \in \mathbb{N}$.
2. *Es gibt eine* **Nachfolgerabbildung** $succ : \mathbb{N} \to \mathbb{N} \setminus \{0\}$
3. *succ ist injektiv.*
4. *Ist $M \subseteq \mathbb{N}$ eine Teilmenge von $\mathbb{N}$ mit den Eigenschaften*
   - $0 \in M$

[1] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Peano.html
[2] http://de.wikipedia.org/wiki/Nat%C3%BCrliche_Zahl

- *Aus $m \in M$ folgt $succ(m) \in M$ für alle $m \in M$*

so gilt $M = \mathbb{N}$.

Man sollte sich vorstellen, daß $succ(m) = m + 1$ für alle $m \in \mathbb{N}$ gilt, wobei hier "+" noch nicht sauber definiert ist und "naiv" genommen werden muß. Ebenso ist die "naive" Bedeutung der Zahl $n \in \mathbb{N}$ hier zu ersetzen durch

"das Ergebnis der n–maligen Anwendung von *succ* auf 0."

Das ist eine formell harte Definition, wenn man davon absieht, daß wir eigentlich die Zeichen 0,1,2,3,...,27 usw. noch definieren müßten.

Wen es stört, daß oben von "0" die Rede ist, und wer bemerkt hat, daß die Peano–Axiome zwar die formale Struktur, nicht aber die Existenz der natürlichen Zahlen klären, kann sich ein Modell der natürlichen Zahlen aus der axiomatischen Mengenlehre bzw. aus Zeichenketten bauen:

- die "0" ist die leere Menge $\emptyset$.
- Ist $x \in \mathbb{N}$, so ist $succ(x) := \{x\}$

Das ist für Informatiker nichts Besonderes, weil man damit die Zahlen auf Zeichenketten reduziert. Eine andere Möglichkeit der Einführung von Zahlen besteht aus den Äquivalenzklassen von Mengen bezüglich Gleichmächtigkeit. Aber alle diese Methoden haben den Nachteil, daß man nicht ohne weiteres zu allgemeingültigen Aussagen über **alle** natürlichen Zahlen kommen kann. Dies liefert gerade das vierte Peano–Axiom, und der folgende Abschnitt erklärt, wie das geht.

3.1.2 Induktion

Theorem 3.2 *Es sei $P : \mathbb{N} \to B$ ein Prädikat auf $\mathbb{N}$, d.h. für alle $n \in \mathbb{N}$ sei $P(n)$ wahr oder falsch. Gelten dann die beiden Aussagen*

1. **Induktionsanfang***:*
   $P(0)$ ist wahr
2. *Für alle $n \in \mathbb{N}$ gilt:*
   *Aus $P(n)$ ist wahr (***Induktionsannahme***) folgt*
   *$P(succ(n))$ ist wahr (***Induktionsschluß***)*

so ist P allgemeingültig über $\mathbb{N}$, d.h. $P(n)$ ist wahr für alle $n \in \mathbb{N}$.

Man kann das natürlich auch formal und unverständlich schreiben:

$$(P(0) \wedge (\forall\, n \in I\!N\ (P(n) \rightarrow P(succ(n))))) \rightarrow (\forall\, n \in I\!N P(n))$$

Es ist nicht schwer, dieses **Prinzip der "vollständigen" Induktion** zu beweisen, wenn man die Peano–Axiome hat. Vergleicht man die Axiome mit der Behauptung des Satzes, so wird klar, daß man die Menge

$$M := \{n \in I\!N\ :\ P(n) \text{ ist wahr}\}$$

definieren sollte. Die Voraussetzungen des Satzes 3.2 sind gerade so, daß man das vierte Peano–Axiom auf M anwenden kann, um $M = I\!N$ zu bekommen, und das ist die Behauptung des Satzes.

Man sehe sich in [4] die Beispiele auf S. 45–47 an.

Induktionsbeweise werden in der parallelen Vorlesung "Diskrete Mathematik" intensiv geübt, deshalb kommen sie hier nur am Rande vor.

3.1.3 Rekursion

In der Informatik hat man oft Abbildungen $f\ :\ I\!N \rightarrow M$ auf $I\!N$ zu definieren, die als Programme laufen sollen, um irgendwelche Resultate $f(n) \in M$ zu jedem beliebigen $n \in I\!N$ zu produzieren. Durch **Rekursion** kann man sich oft das Leben erleichtern. Sie besteht darin, die Auswertung von $f(succ(n))$ durch die Auswertung von $f(n)$ auszudrücken und den Fall der Auswertung von $f(0)$ separat zu behandeln. Das Prinzip ist

1. Man gebe eine Vorschrift zur Berechnung von $f(0) \in M$ an.
2. Zur Berechnung von $f(n+1)$ verwendet man eine möglichst einfache Abbildung $F\ :\ I\!N \times M \rightarrow M$ und setzt

$$f(n+1) := F(n+1, f(n)) \text{ für alle } n \in I\!N.$$

Rekursion kommt häufig als prozeduraler Trick in der Informatik, und als Definitionsmethode in der Mathematik vor. Es folgen einige Beispiele.

3.1.4 Operationen

Bisher kennen wir in den natürlichen Zahlen nur die Nachfolgerfunktion *succ*, aber weder Addition noch Multiplikation. Mit rekursiven Definitionen kann man sich diese aber leicht besorgen:

Definition 3.3 *Die* **Addition** *ist als Abbildung* $ADD : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ *rekursiv definierbar als*

$$\begin{array}{rcll} ADD(n,0) & := & n & \text{für alle } n \in \mathbb{N} \\ ADD(n, succ(m)) & := & succ(ADD(n,m)) & \text{für alle } m,n \in \mathbb{N}. \end{array}$$

Die Multiplikation ist als Abbildung $MULT : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ *rekursiv definierbar als*

$$\begin{array}{rcll} MULT(n,0) & := & 0 & \text{für alle } n \in \mathbb{N} \\ MULT(n, succ(m)) & := & ADD(MULT(n,m), n) & \text{für alle } m,n \in \mathbb{N}. \end{array}$$

Dabei ist das (beliebige) erste Argument n festgelassen worden, und die Rekursion erfolgt nur im zweiten Argument. Diese Unsymmetrie führt dazu, daß wir Dinge wie die **Kommutativitätsgesetze**

$$\begin{array}{rcl} ADD(n,m) & = & ADD(m,n) \text{ für alle } n,m \in \mathbb{N} \\ MULT(n,m) & = & MULT(m,n) \text{ für alle } n,m \in \mathbb{N} \end{array}$$

kunstvoll beweisen müssen.

Wir haben hier die zweistellige geklammerte Präfixschreibweise benutzt und können dann zur Infixschreibweise

$$\begin{array}{rcll} m+n & := & ADD(m,n) & \text{für alle } m,n \in \mathbb{N} \\ m*n & := & MULT(m,n) & \text{für alle } m,n \in \mathbb{N} \end{array}$$

übergehen. Man braucht dann einen Sack voll Induktionsbeweise, um die Rechenregeln

$$\begin{array}{rcll} m+n & = & n+m & \text{(Kommutativität von } +) \\ m*n & = & n*m & \text{(Kommutativität von } *) \\ m+0 & = & m & \text{(Neutralität von 0 bzgl. } +) \\ m*1 & = & m & \text{(Neutralität von 1 bzgl. } *) \\ k+(m+n) & = & (k+m)+n & \text{(Assoziativität von } +) \\ k*(m*n) & = & (k*m)*n & \text{(Assoziativität von } *) \\ k*(m+n) & = & (k*m)+(k*n) & \text{(Distributivität von } * \text{ und } +) \end{array}$$

für alle $k,m,n \in \mathbb{N}$ zu beweisen. Das wollen wir uns weitgehend ersparen, aber wir führen ein Beispiel vor, wobei wir uns der Verfremdung wegen der Präfixschreibweise bedienen. das Ganze dient auch zur Vorführung, wie man einen Induktionsbeweis sauber durchführt und aufschreibt.

Am Ende wollen wir die Kommutativität $m+n = n+m$ beweisen, aber zuerst behaupten wir nur

$$ADD(0,m) = ADD(m,0) \text{ für alle } m \in \mathbb{N}.$$

Das beweisen wir per Induktion über m.

Induktionsanfang $m = 0$: Zu zeigen ist $ADD(0,0) = ADD(0,0)$, und das ist trivial.

Induktionsschluß: Es gelte $ADD(0,m) = ADD(m,0)$ und wir behaupten

$$ADD(0, succ(m)) = ADD(succ(m), 0).$$

Wir verfahren nach der Schlußkette

$$\begin{array}{rcll} ADD(0,succ(m)) & = & succ(ADD(0,m)) & \text{(Definition von } ADD) \\ & = & succ(ADD(m,0)) & \text{(Induktionsvoraussetzung)} \\ & = & succ(m) & \text{(Definition von } ADD) \\ & = & ADD(succ(m),0) & \text{(Definition von } ADD). \end{array}$$

Als eine zweite Übung und Hilfsbehauptung beweisen wir

$$ADD(succ(n), m) = ADD(n, succ(m)) \text{ für alle } m, n \in \mathbb{N}$$

per Induktion über m bei festem n. Das ist natürlich nichts anderes als

$$(n+1) + m = n + (m+1),$$

aber auch das ist nicht unmittelbar aus der Definition der Addition ablesbar.

Induktionsanfang $m = 0$: Zu zeigen ist $ADD(succ(n),0) = ADD(n, succ(0))$.

Das klappt leicht mit der vorigen Behauptung:

$$\begin{array}{rcll} ADD(succ(n),0) & = & ADD(0,succ(n)) & \text{(vorige Behauptung)} \\ & = & succ(ADD(0,n)) & \text{(Definition von } ADD) \\ & = & succ(ADD(n,0)) & \text{(vorige Behauptung)} \\ & = & ADD(n,succ(0)) & \text{(Definition von } ADD). \end{array}$$

Induktionsschluß: Gelte $ADD(succ(n),m) = ADD(n,succ(m))$ für festes m und n, und wir wollen beweisen, daß

$$ADD(succ(n), succ(m)) = ADD(n, succ(succ(m)))$$

gilt. Das folgt aus

$$\begin{array}{rcll} & ADD(succ(n),succ(m)) & & \\ = & succ(ADD(succ(n),m)) & & \text{(Definition von } ADD) \\ = & succ(ADD(n,succ(m))) & & \text{(Induktionsvoraussetzung)} \\ = & ADD(n,succ(succ(m)))) & & \text{(Definition von } ADD). \end{array}$$

Jetzt werden wir unsere beiden Hilfsbehauptungen benutzen, um

$$ADD(n,m) = ADD(m,n) \text{ für alle } m,n \in I\!N$$

zu beweisen, und zwar mit Induktion über n.

Induktionsanfang: Zu zeigen: $ADD(0,m) = ADD(m,0)$ für alle $m \in I\!N$.

Das war die erste Hilfsbehauptung.

Induktionsschluß: Es gelte $ADD(n,m) = ADD(m,n)$ für alle $m \in I\!N$ und wir wollen zeigen, daß

$$ADD(succ(n),m) = ADD(m,succ(n)) \text{ für alle } m \in I\!N$$

gilt. Das machen wir mit

$$\begin{array}{rcll} ADD(succ(n),m) & = & ADD(n,succ(m)) & \text{(zweite Hilfsbehauptung)} \\ & = & succ(ADD(n,m)) & \text{(Def. von ADD)} \\ & = & succ(ADD(m,n)) & \text{(Induktionsvoraussetzung)} \\ & = & ADD(m,,succ(n)) & \text{(Def. von ADD).} \end{array}$$

□

3.2 Ganze Zahlen

Eine der Möglichkeiten, formell die ganzen Zahlen $Z\!\!Z := \{0,-1,1,-2,2,\ldots\}$ einzuführen ohne sich zuviel Arbeit einzuhandeln, besteht darin, auf $I\!N \times I\!N$ eine Äquivalenzrelation $\approx$ einzuführen:

$$(m,n) \approx (p,q) \text{ genau dann, wenn } m+q = n+p \text{ gilt.}$$

Alle Tupel der Form $(m,0)$ sind dann nicht äquivalent, denn aus $(m,0) \approx (p,0)$ folgt $m+0 = 0+p$. Diese Paare entsprechen den üblichen nichtnegativen Zahlen, während die (ebenfalls nicht zueinander äquivalenten) Paare $(0,n)$ mit $n \geq 1$ den üblichen negativen Zahlen entsprechen. Der Null entspricht $(0,0)$. Dahinter steht die Idee, daß (m,n) in herkömmlicher Sichtweise der Zahl $m-n$ entspricht, und $(m,n) \approx (p,q)$ bedeutet $m-n = p-q$, stellt also dieselbe herkömmliche Zahl dar. Die obige Technik erlaubt eine Definition "negativer" Zahlen ohne jede Spekulation darüber, was "Minus" bedeutet.

3.2.1 Operationen

Man definiert dann einfach die Addition durch

$$[(m,n)] + [(u,v)] := [(m+u, n+v)] \text{ für alle } m,n,u,v \in \mathbb{N},$$

und man kann leicht verifizieren, daß dem die übliche Beziehung $(m-n)+(u-v) = (m+u)-(n+v)$ entspricht und die Addition wohldefiniert ist.

Machen wir zu Übungszwecken einen Wohldefiniertheitsbeweis. Wir haben die Abbildung über Vertreter der Klassen definiert, aber wir müssen zeigen, daß die Definition nur von der Klasse, nicht vom Vertreter abhängt. Wir nehmen also an, daß

$$[(m,n)] = [(m',n')] \text{ und } [(u,v)] = [(u',v')] \tag{3.4}$$

mit $m,n,m',n',u,v,u',v' \in \mathbb{N}$ gilt, und wir müssen zeigen, daß

$$[(m+u, n+v)] = [(m'+u', n'+v')]$$

gilt. Zu zeigen ist also

$$m+u+n'+v' = n+v+m'+u'.$$

Unsere Voraussetzung (3.4) liefert

$$m+n' = n+m' \text{ und } u+v' = v+u'.$$

Wenn wir diese beiden Gleichungen addieren, folgt die Behauptung.

Die Multiplikation definiert man als

$$[(m,n)] * [(u,v)] := [(m*u+n*v, m*v+n*u)] \text{ für alle } m,n,u,v \in \mathbb{N},$$

weil im üblichen Sinne $(m-n)*(u-v) = (m*u+n*v)-(m*v+n*u)$ gilt. Auch hier muß man Wohldefiniertheit nachweisen.

Aufgabe: Man zeige die Wohldefiniertheit der Multiplikation.

Aber jetzt brauchen wir noch die Subtraktion als neue binäre Operation. Man kann aber auch erst die Vorzeichenumkehr als einstellige Operation durch

$$-[(u,v)] := [(v,u)] \text{ für alle } u,v \in \mathbb{N}$$

definieren (Frage: ist das wohldefiniert?) und dann die Subtraktion als

$$\begin{aligned}[(m,n)] - [(u,v)] &:= [(m,n)] + (-[(u,v)]) \\ &= [(m,n)] + [(v,u)] \\ &= [(m+v, n+u)]\end{aligned}$$

für alle $m, n, u, v \in \mathbb{N}$. Man kann dann zu jedem Element $[(m,n)]$ das Element $-[(m,n)]$ angeben mit $[(m,n)] + (-[(m,n)] = [(m,n)] + [(n,m)] = [(m+n, m+n)] = [(0,0)]$. Man nennt dieses Element $-[(m,n)]$ das additive Inverse zu $[(m,n)]$.

Die ganzen Zahlen $\mathbb{Z}$ bilden unter der Addition eine abelsche Gruppe.

Definition 3.5 *Eine nichtleere Menge G heißt* **Gruppe** *unter einer Abbildung $\circ : G \times G \to G$, wenn gilt*

1. *$(a \circ b) \circ c = a \circ (b \circ c)$ für alle $a, b, c \in G$*
   (Assoziativität von $\circ$)
2. *Es gibt ein* **neutrales Element** *$e \in G$ mit $a \circ e = a$ für alle $a \in G$*
3. *Jedes Element $a \in G$ hat ein* **Inverses** *a^{-1} mit $a \circ a^{-1} = e$ für alle $a \in G$.*

Man kann dann zeigen, daß e eindeutig bestimmt ist, und daß auch

$$\begin{aligned}e \circ a &= a \\ a^{-1} \circ a &= e\end{aligned}$$

für alle $a \in G$ gilt, und daß das Inverse a^{-1} zu jedem a eindeutig bestimmt ist. Gilt ferner das **Kommutativitätsgesetz**

$$a \circ b = b \circ a \text{ für alle } a, b \in G,$$

so heißt die Gruppe G abelsch oder kommutativ.

Im Falle $G = \mathbb{Z}$ ist die Abbildung $\circ$ die Addition $+$, das neutrale Element e ist die Null, und das Inverse zu a wird als $-a$ geschrieben.

Wir kommen auf Gruppen gelegentlich zurück, wollen hier aber nicht tiefer in die Gruppentheorie[1] einsteigen. Die Vorlesung "Diskrete Mathematik" bringt Weiteres, z.B. daß die ganzen Zahlen ein kommutativer **Ring** mit Einselement sind. Für uns reicht es, daß man mit ganzen Zahlen genau wie

[1] `http://de.wikipedia.org/wiki/Gruppentheorie`

in der Schule rechnen kann, und wir gehen zu der naiven Notation negativer Zahlen und der Subtraktion zurück. Der Sinn dieser ganzen Betrachtungen ist, daß die Definition negativer Zahlen gerade so gemacht werden kann, daß die üblichen Rechenregeln herauskommen.

Als Querverbindung zur Diskreten Mathematik streifen wir noch kurz die "**Restklassenarithmetik** modulo n". Zu jeder positiven natürlichen Zahl n kann man in $\mathbb{Z}$ die Äquivalenzrelation

$$xR_ny \Leftrightarrow x - y = p \cdot n \text{ mit } p \in \mathbb{Z} \tag{3.6}$$

definieren, d.h. x und y sind äquivalent, wenn $x - y$ durch n teilbar ist. Die Menge $\mathbb{Z}/R_n$ der Restklassen ist dann durch die Vertreter $0, 1, \ldots, n-1$ im Sinne von Satz 1.21 eindeutig bestimmt, und man kann auf diesen Restklassen die Addition und Multiplikation so ausführen, daß man immer vom normalen Ergebnis den Rest nach Division durch n bildet. Man bekommt die Rechenregeln eines kommutativen Rings mit Einselement. Auch in der Informatik braucht man mitunter die Restklassenarithmetik, z.B. beim Verfahren von Schönhage–Strassen[1] zur schnellen Multiplikation sehr großer Zahlen, oder bei Hashfunktionen[2] mit der Divisions–Rest–Methode[3].

3.3 Rationale Zahlen

Sind die ganzen Zahlen $\mathbb{Z}$ durch die obige Konstruktion gegeben, so kann man als nächstes die rationalen Zahlen[4] $\mathbb{Q}$ aufbauen, und zwar wieder mit einer Äquivalenzrelation. Man denkt sich einen Bruch $\frac{m}{n}$ mit $m \in \mathbb{Z}$ und $n \in \mathbb{Z} \setminus \{0\}$ als äquivalent zu allen seinen Erweiterungen $\frac{k \cdot m}{k \cdot n}$ und schreibt ihn als Äquivalenzklasse $[(m, n)]$. Die Äquivalenzrelation auf $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ zwischen zwei solchen "Brüchen" ist dann

$$(m, n) \approx (p, q) \text{ genau dann, wenn } m \cdot q = n \cdot p \text{ gilt,}$$

d.h. $\frac{m}{n} = \frac{p}{q}$ im herkömmlichen Sinn. Die Operationen sind dann wie in der Schule

$$[(m, n)] + [(u, v)] := [(m \cdot v + u \cdot n, n \cdot v)]$$
$$[(m, n)] \cdot [(u, v)] := [(m \cdot u, n \cdot v)]$$

[1] http://de.wikipedia.org/wiki/Sch%C3%B6nhage-Strassen-Algorithmus
[2] http://de.wikipedia.org/wiki/Hash-Funktion
[3] http://de.wikipedia.org/wiki/Divisions-Rest-Methode
[4] http://de.wikipedia.org/wiki/Rationale_Zahl

für alle $m, n, u, v \in \mathbb{Z}$ definiert, was den üblichen Operationen

$$\begin{aligned} \frac{m}{n} + \frac{u}{v} &= \frac{m \cdot v + u \cdot n}{n \cdot v} \\ \frac{m}{n} \cdot \frac{u}{v} &= \frac{m \cdot u}{n \cdot v} \end{aligned}$$

aus der Schule entspricht. Wieder haben wir die Operationen und die neuen Zahlen so definiert, daß die altgewohnten Rechenregeln als notwendiges Ergebnis herauskommen.

Aufgabe: Warum sind Addition und Multiplikation wohldefiniert?

Die Äquivalenzklasse $[(0, n)] = [(0, 1)]$ mit $n \neq 0$ fungiert als Null, die Klasse $[(n, n)] = [(1, 1)]$ mit $n \neq 0$ als Eins. Alle Äquivalenzklassen, die von der Null verschieden sind, haben die Form $[(m, n)]$ mit $m, n \neq 0$. Sie haben eine multiplikative Inverse, nämlich $[(n, m)]$ mit $[(m, n)] \cdot [(n, m)] = [(m \cdot n, m \cdot n)] = [(1, 1)]$. Die positiven rationalen Zahlen werden durch die Äquivalenzklassen der Form $[(m, n)]$ mit $m, n \in \mathbb{N} \setminus \{0\}$ dargestellt.

Auch mit diesen Zahlen kann man ganz wie in der Schule rechnen, wobei man wieder die Standardnotation einführt. Die Vorlesung "Diskrete Mathematik" beweist, daß die rationalen Zahlen einen kommutativen **Körper**[1] bilden. Sie bilden eine abelsche **Gruppe** mit neutralem Element $0 = [(0, 1)]$ unter der Addition, und $\mathbb{Q} \setminus \{0\}$ ist eine abelsche Gruppe unter der Multiplikation, wobei $[(1, 1)]$ das neutrale Element ist.

Wir stellen die Rechenregeln für Brüche noch einmal zusammen:

$$\begin{aligned} \frac{m}{n} + \frac{p}{q} &= \frac{m \cdot q + p \cdot n}{n \cdot q} \\ \frac{m}{n} \cdot \frac{p}{q} &= \frac{m \cdot p}{n \cdot q} \\ \frac{m \cdot n}{q \cdot n} &= \frac{m}{q} \end{aligned}$$

für alle $m, p \in \mathbb{Z}$, $n, q \in \mathbb{Z} \setminus \{0\}$, weil die Erfahrung lehrt, daß Studienanfänger in Mathematik oder Informatik oft immer noch Schwierigkeiten mit der Bruchrechnung haben.

[1] `http://de.wikipedia.org/wiki/K%C3%B6rper_%28Algebra%29`

Die allgemeinen Rechenregeln in kommutativen Körpern[1], wie sie für rationale, reelle und komplexe Zahlen gelten, stellen wir hier übersichtlich zusammen, verweisen aber auf die später folgenden Regeln für die Ordnungsrelationen.

1. Gegeben sei eine Menge $\mathbb{K}$ mit zwei Abbildungen $+$ und $\cdot$ von $\mathbb{K} \times \mathbb{K}$ in $\mathbb{K}$, geschrieben in Infixform. Dabei sollte man sich als $\mathbb{K}$ die Menge der rationalen Zahlen oder eine Obermenge davon vorstellen, mit der üblichen Addition und Multiplikation.

2. Die beiden Abbildungen sind kommutativ, assoziativ und distributiv:

$$\begin{array}{rcll} x+y & = & y+x & \text{Kommutativität von } + \\ (x+y)+z & = & x+(y+z) & \text{Assoziativität von } + \\ x \cdot y & = & y \cdot x & \text{Kommutativität von } \cdot \\ (x \cdot y) \cdot z & = & x \cdot (y \cdot z) & \text{Assoziativität von } \cdot \\ (x+y) \cdot z & = & (x \cdot y) + (y \cdot z) & \text{Distributivität} \end{array}$$

   für alle $x, y, z \in \mathbb{K}$.

3. $\mathbb{K}$ hat mindestens zwei spezielle Elemente, die 0 und 1 genannt werden und verschieden sind. Sie haben die Eigenschaften

$$\begin{aligned} x + 0 &= x \\ x \cdot 1 &= x \end{aligned}$$

   für alle $x \in \mathbb{K}$.

4. Zu jedem $x \in \mathbb{K}$ gibt es genau ein Element $-x \in \mathbb{K}$ mit $x + (-x) = 0$.

5. Zu jedem $x \in \mathbb{K} \setminus \{0\}$ gibt es genau ein Element $x^{-1} \in \mathbb{K} \setminus \{0\}$ mit $x \cdot x^{-1} = 1$.

Man sagt dann, $\mathbb{K}$ sei ein kommutativer Körper mit den Operationen $+$ und $-$.

Wir kennen bisher nur das Beispiel $\mathbb{K} = \mathbb{Q}$, aber findige Leser werden leicht nachrechnen können, daß man auf der Menge $\{0, 1\}$ eine Addition und eine Multiplikation so definieren kann, daß man den kleinsten aller denkbaren Körper herausbekommt.

[1] `http://de.wikipedia.org/wiki/K%C3%B6rper_%28Algebra%29`

Es gibt auch nichtkommutative Körper, die dann **Schiefkörper** genannt werden, aber nach einem ziemlich tiefliegenden Satz von **Wedderburn**[1][2] haben solche Körper immer unendlich viele Elemente.

Neben den Grundregeln kann man noch definieren

$$x - y := x + (-y)$$

$$x : z := \frac{x}{z} := x \cdot z^{-1}$$

für alle $x, y \in \mathbb{K}$ und alle $z \in \mathbb{K} \setminus \{0\}$. Die Operationen $-$ und $:$ sind im allgemeinen weder kommutativ noch assoziativ. Es gilt aber

$$\begin{aligned} x - y &= -(y - x) \\ u : v &= (v : u)^{-1} \\ -(-x) &= x \\ (v^{-1})^{-1} &= v \\ (-x) \cdot y &= -(x \cdot y) \\ (-x) : v &= -(x : v) \\ x : (-v) &= -(x : v) \\ (-x) \cdot (-y) &= x \cdot y \\ (-x) : (-v) &= x : v \end{aligned}$$

für alle $x, y \in \mathbb{K},\ u, v \in \mathbb{K} \setminus \{0\}$. Für die Bruchschreibweise bzw. die Division $:$ gelten die Regeln der Bruchrechnung wie oben.

Ferner braucht man die Regeln der naiven Potenzrechnung:

1. Für alle $x \neq 0$ definiert man $x^0 := 1$.
2. Für alle $x \neq 0$ und alle $n \in \mathbb{N}$ definiert man $x^{n+1} := x \cdot x^n$.
3. Für alle $x \neq 0$ und alle $n \in \mathbb{N}$ definiert man $x^{-n} := (x^{-1})^n$.

Dann gilt für alle $x \in \mathbb{K} \setminus \{0\}$ und alle $m, n \in \mathbb{Z}$

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

Exotische Körper müssen nicht $\mathbb{N}$ als Teilmenge enthalten. Dann muß man noch rekursiv definieren

$$\begin{aligned} (n+1) \cdot x &:= x + n \cdot x \\ (-n) \cdot x &:= -(n \cdot x) \end{aligned}$$

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Wedderburn.html`
[2] `http://de.wikipedia.org/wiki/Satz_von_Wedderburn`

für alle $x \in \mathbb{K}$ und $n \in \mathbb{N}$, aber für die rationalen, die reellen und die komplexen Zahlen ist das nicht nötig. In jedem Falle gilt aber

$$\begin{aligned}(m+n)\cdot x &:= m\cdot x + n\cdot x\\ (m\cdot n)\cdot x &:= m\cdot(n\cdot x) = n\cdot(m\cdot x)\end{aligned}$$

für alle $x \in \mathbb{K}$ und $m, n \in \mathbb{Z}$, in Analogie zur Potenzrechnung.

Man kann den obigen Sachverhalt auch durch zwei Abbildungen

$$\begin{array}{lcll} A & : & \mathbb{Z}\times\mathbb{K}\to\mathbb{K}, & A(n,x) := n\cdot x,\\ P & : & \mathbb{Z}\times\mathbb{K}\setminus\{0\}\to\mathbb{K}\setminus\{0\}, & A(n,x) := x^n \end{array}$$

beschreiben, die auf $\mathbb{K}$ bzw. $\mathbb{K}\setminus\{0\}$ "operieren".

Aufgabe: Was für Eigenschaften habe diese Abbildungen bezüglich der Operationen auf $\mathbb{Z}$, $\mathbb{K}$ und $\mathbb{K}\setminus\{0\}$?

In der Informatik sind endliche Körper[1] von besonderer Bedeutung, weil sie in der Codierungstheorie vorkommen. Aber das gehört zum Standardrepertoire der Diskreten Mathematik. Insbesondere führt die **Restklassenarithmetik** modulo n (siehe (3.6)) zu einem endlichen Körper mit n Elementen, wenn n eine Primzahl ist. Wer nicht glaubt, daß man in der Informatik endliche Körper und Vektorräume (das folgt im nächsten Kapitel) braucht, sehe sich mal die linearen Codes an[2].

3.4 Ordnungsrelationen auf Zahlen

3.4.1 Anordnungsaxiome

Für natürliche, ganze, rationale und auch für die später definierten reellen Zahlen definiert man zuerst, was "positiv[3]" bedeutet. Bei den natürlichen Zahlen ist dies klar, denn "positiv" sind alle Zahlen außer der Null. Bei den ganzen Zahlen kann man dann definieren

$$[(m,0)] > 0 \text{ genau dann, wenn } m > 0,\ m \in \mathbb{N},$$

und bei den rationalen Zahlen setzt man

$$[(m,n)] > 0 \text{ genau dann, wenn } [(m,n)] \approx [(u,v)]$$

[1] http://de.wikipedia.org/wiki/Endlicher_K%C3%B6rper
[2] http://de.wikipedia.org/wiki/Linearer_Code
[3] http://de.wikipedia.org/wiki/Positive_und_negative_Zahlen

mit $u, v \in \mathbb{Z},\ u, v > 0$. Natürlich ist dann auch klar, wie man "nichtnegativ", "negativ" und "nichtpositiv" definieren muß.

Man weiß also, was $z > 0$ und $z \geq 0$ für diese Zahlen bedeutet. Danach definiert man $x > y$ bzw. $x \geq y$ durch $x = y + z$ mit einem geeigneten $z > 0$ bzw. $z \geq 0$. Machen wir daraus eine kleine Übung für Quantoren:

$$\begin{array}{lll} \forall\, x\ \forall\, y\ (x > y & \leftrightarrow & \exists z > 0\ :\ x = y + z) \\ \forall\, x\ \forall\, y\ (x \geq y & \leftrightarrow & \exists z \geq 0\ :\ x = y + z) \end{array}$$

Analog definiert man $x < y$ bzw. $x \leq y$ durch $y > x$ bzw. $y \geq x$ für alle x, y. Mit Absicht haben wir dabei offengelassen, aus welcher Zahlenmenge x, y, z sind. Man sieht, daß diese Relationen alle transitiv und antisymmetrisch sind, und unter $\leq$ oder $\geq$ sind die Zahlenmengen total geordnet.

Man kann das Ganze auch axiomatisieren:

Definition 3.7 *Es sei $\mathbb{K}$ ein kommutativer Körper. Er heißt* **geordnet**[1] *wenn es einen* **Positivbereich** *$P \subseteq \mathbb{K}$ gibt mit*

1. *Die Mengen $-P := \{-x\ :\ x \in P\}, P$ und $\{0\}$ sind disjunkt.*
2. *Ihre Vereinigung ist $\mathbb{K}$.*
3. *Aus $x, y \in P$ folgt $x + y \in P$ und $x \cdot y \in P$.*

Für reelle und rationale Zahlen ist $P = \{x\ :\ x > 0\}$, und man definiert wie oben

$$x > y \ \leftrightarrow\ x - y \in P$$

für alle $x, y \in \mathbb{K}$.

3.4.2 Rechnen mit Ungleichungen

In Abschnitt 1.3.2 auf Seite 31 haben wir gesehen, daß man auf die beiden Seiten einer Gleichung eine beliebige Abbildung anwenden kann und wieder eine Gleichung erhält. Wir machen nun dasselbe für Ungleichungen.

Definition 3.8 *Es seien $R \subseteq M \times M$ und $S \subseteq N \times N$ zweistellige Relationen auf Mengen M bzw. N. Eine Abbildung $f\ :\ M \to N$ heißt* **monoton***, wenn für alle $x, y \in M$ mit xRy auch $f(x)Sf(y)$ gilt. Mit anderen Worten: in Relation R stehende Paare (x, y) werden auf in Relation S stehende Paare $(f(x), f(y))$ abgebildet.*

[1] `http://de.wikipedia.org/wiki/Geordneter_K%C3%B6rper`

Daß diese Definition dem üblichen Monotoniebegriff entspricht, sieht man am Beispiel der Relation $\leq$ auf $\mathbb{R}$. Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt **monoton**, wenn für alle $x, y \in \mathbb{R}$ aus $x \leq y$ auch $f(x) \leq f(y)$ folgt. Mit anderen Worten: wenn x größer wird, so wächst auch $f(x)$, und auf eine Ungleichung

$$x \leq y$$

kann man eine monotone Funktion anwenden, um eine neue Ungleichung

$$f(x) \leq f(y)$$

zu bekommen. Erlaubte Umformungen von Ungleichungen sind also durch monotone Abbildungen realisierbar.

Wenn die Anwendung einer Funktion $f : \mathbb{R} \to \mathbb{R}$ auf eine Ungleichung $x \leq y$ stets zu $f(y) \leq f(x)$ führt, also die Ungleichung umkehrt, ist f **antimonoton**.

Einfache Beispiele für monotone Funktionen auf den rationalen oder reellen Zahlen sind

$$\begin{array}{lcll} f(x) & = & x + c & \text{mit einer beliebigen Zahl } c \\ f(x) & = & x \cdot c & \text{mit einer beliebigen } \textbf{positiven} \text{ Zahl } c \\ f(x) & = & x, x^3, x^5, \ldots & \end{array}$$

und antimonoton ist

$$f(x) = x \cdot c \quad \text{mit einer beliebigen } \textbf{negativen} \text{ Zahl } c.$$

Man kann sich die Monotonieeigenschaften leicht durch eine Zeichnung klarmachen. Aus diesen Beispielen ergeben sich einfache Rechenregeln für Ungleichungen:

- Man kann zu beiden Seiten einer Ungleichung eine beliebige Zahl addieren oder subtrahieren.
- Man kann beide Seiten einer Ungleichung mit einer positiven Zahl multiplizieren.
- Multipliziert man eine Ungleichumg mit einer negativen Zahl, so kehrt sich die Ungleichung um.
- Man kann beide Seiten einer Ungleichung zu derselben ungeraden Potenz erheben.

Was gilt für das Quadrieren? Die Funktion $f(x) = x^2$ ist nur für $x \geq 0$ monoton, und deshalb kann man die beiden Seiten einer Ungleichung nicht quadrieren, wenn man nicht weiß, daß beide Seiten nichtnegativ sind. Aus $2 < 5$ folgt $2^2 < 5^2$, aber aus $-5 < 2$ folgt eben nicht $(-5)^2 = 25 < 2^2 = 4$. Ebenso ist Vorsicht geboten, wenn man von x zu $1/x = x^{-1}$ übergehen will. Diese Funktion ist zwar antimonoton auf jeweils den positiven und den negativen Zahlen, aber sie springt bei Null von großen negativen zu großen positiven Werten. Also:

- Das Quadrieren einer Ungleichung ist erlaubt, wenn beide Seiten nichtnegativ sind.
- Wenn beide Seiten einer Ungleichung negativ sind, führt das Quadrieren zur Umkehrung der Ungleichungsrelation.
- Die Bildung des Kehrwerts auf beiden Seiten einer Ungleichung kehrt die Ungleichung um, wenn beide Seiten dasselbe Vorzeichen haben. Sind die Vorzeichen verschieden, so bleibt die Ungleichung erhalten.

Das Rechnen mit Ungleichungen muß **unbedingt** geübt werden!

3.4.3 Absolutbetrag

Definition 3.9 *Zu einer reellen oder rationalen Zahl x definiert man*

$$|x| := \left\{ \begin{array}{cc} x & x \geq 0 \\ -x & x < 0 \end{array} \right\}$$

als den **Absolutbetrag**[1] *oder* **Betrag** *von x.*

Theorem 3.10 *Für den Absolutbetrag gelten die Regeln*

$$\begin{array}{rcl} |x| & \geq & 0 \\ |-x| & = & |x| \\ |x \cdot y| & = & |x| \cdot |y| \\ |x+y| & \leq & |x| + |y| \\ |x-y| & \geq & ||x| - |y|| = ||y| - |x|| \end{array}$$

für alle $x, y \in \mathbb{R}$.

Die ersten drei Regeln sind nicht weiter erklärungsbedürftig. Die vierte wird (auch in verallgemeinerter Form) die **Dreiecksungleichung** genannt, aber das wird erst später klar.

[1] `http://de.wikipedia.org/wiki/Betragsfunktion`

Zum Beweis der Dreiecksungleichung machen wir eine Fallunterscheidung.

Sind x und y nicht von verschiedenem Vorzeichen, so gilt $|x+y| = |x| + |y|$. Denn wenn beide nichtnegativ sind, hat man $x + y = |x + y|$ und $x + y = |x| + |y|$. Sind beide negativ, so folgt $|x + y| = -(x + y)$ und $-(x + y) = -x + (-y) = |x| + |y|$.

Sind x und y von verschiedenem Vorzeichen, so nehmen wir ohne Beschränkung der Allgemeinheit an, daß $|x| \geq |y|$ gilt und bekommen

$$|x + y| = |x| - |y| \leq |x| - |y| + 2|y| = |x| + |y|.$$

Ganz ähnlich beweist man die letzte Ungleichung des Satzes.

Sind x und y nicht von gleichem Vorzeichen, so gilt $|x-y| = |x|+|y| \geq |x| \geq |x| - |y|$ und aus Symmetriegründen auch $|x-y| = |x|+|y| \geq |y| \geq |y| - |x|$. Es folgt $|x - y| \geq ||y| - |x|| = ||x| - |y||$.

Sind x und y von gleichem Vorzeichen, so nehmen wir ohne Beschränkung der Allgemeinheit an, daß $|x| \geq |y|$ gilt und bekommen

$$|x - y| = |x| - |y| = ||x| - |y|| = ||y| - |x||.$$

Man mache sich klar, daß die Dreiecksungleichung zur Gleichung wird ("scharf" ist), wenn x und y gleiches Vorzeichen haben, während die Ungleichung $|x - y| \geq ||x| - |y|| = ||y| - |x||$ scharf ist, wenn die Vorzeichen verschieden sind.

Anfänger fragen immer wieder, wie man denn die zu den obigen Ungleichungen komplementären Ungleichungen

$$\begin{array}{rcl} |x + y| & \geq & ? \\ |x - y| & \leq & ? \end{array}$$

bekommt. Das ist aber nichts Neues, weil man y durch $-y$ ersetzen kann. Ingesamt braucht man sich nur

$$||x| - |y|| = ||y| - |x|| \quad \leq \quad |x + y| \quad \leq \quad |x| + |y|$$

zu merken.

3.5 Zahldarstellungen

Die natürlichen Zahlen werden durch **Ziffernsysteme** oder **Stellenwertsysteme**[1] dargestellt. Eine **Ziffer**[2] ist ein Zeichen, das eine Zahl darstellt. Man spezifiziert eine **Basis**–Zahl b, die im **Dezimalsystem**[3] gleich 10 ist, im **Dualsystem**[4] gleich 2 und im **Hexadezimalsystem**[5] gleich 16. Dann braucht man b Zeichen für **Ziffern** zwischen 0 und $b-1$ (im Hexadezimalsystem: 0 bis 9, dann A, B, C, D, E, F). Dann schreibt man natürliche Zahlen n in der b–**adischen** mathematischen Form

$$n = b_0 \cdot \underbrace{b^0}_{=1} + b_1 \cdot b^1 + b_2 \cdot b^2 + \ldots + b_k \cdot b^k \tag{3.11}$$

und als Zeichenkette

$$b_k \ldots b_2 b_1 b_0$$

in umgekehrter Reihenfolge, weil man normalerweise die Einerstelle rechts notiert. Wir identifizieren hier die Ziffern mit den Zahlen, die sie repräsentieren, aber das ist wohl verzeihlich. Hinzu kommt die Einschränkung, daß die höchste Ziffer nicht gleich Null ist, wenn die Zahl n nicht selbst Null ist, d.h.

$$b_k > 0 \text{ falls } n > 0. \tag{3.12}$$

Weil die Notation $\ldots$ verpönt ist, verwendet man für indizierte Summen das Zeichen $\sum$ und bekommt

$$n = \sum_{j=0}^{k} b_j \cdot b^j.$$

Man liest das als "Summe von j gleich Null bis k über $b_j \cdot b^j$".

Theorem 3.13 *Jede natürliche Zahl n hat eine eindeutige Darstellung der Form* (3.11) *mit* (3.12) *zu jeder Basis $b > 1$.*

Der Beweis kann per Induktion geführt werden und gleichzeitig konstruktiv sein. Wir überlassen ihn der Diskreten Mathematik. Die rekursive Grundidee folgt aus der einfachen Beobachtung, daß in (3.11) die Einerziffer b_0 der Divisionsrest von n bei Division durch b ist, denn $n - b_0$ ist ein Vielfaches von b. Man bildet dann $m := (n - b_0)/b$ und bekommt

$$m = b_1 \cdot b^0 + b_2 \cdot b^1 + \ldots + b_k \cdot b^{k-1}$$

[1] `http://de.wikipedia.org/wiki/Stellenwertsystem`
[2] `http://de.wikipedia.org/wiki/Ziffer`
[3] `http://de.wikipedia.org/wiki/Dezimalsystem`
[4] `http://de.wikipedia.org/wiki/Dualsystem`
[5] `http://de.wikipedia.org/wiki/Hexadezimalsystem`

aus (3.11). Jetzt ist also b_1 der Divisionsrest von m bei Division durch b, und dieses Verfahren kann man fortsetzen.

Wir probieren das für die Berechnung der Binärdarstellung von 23. Der Divisionsrest von 23 durch 2 ist 1, weil 23 ungerade ist. Wir haben also $b_0 = 1$, und unser erstes m ist $m = (23-1)/2 = 22/2 = 11$. Das ist wieder ungerade, also folgt $b_1 = 1$. Weiter so:

$$\begin{array}{lllllll} m & = & (11-1)/2 & = & 5 & b_2 = 1 & \text{weil 5 ungerade} \\ m & = & (5-1)/2 & = & 2 & b_3 = 0 & \text{weil 2 gerade} \\ m & = & (2-0)/2 & = & 1 & b_4 = 1 & \text{weil 1 ungerade} \\ m & = & (1-1)/2 & = & 0 & b_5 = 0 & \text{Ende des Verfahrens} \end{array}$$

Die Darstellung der dezimal als 23 geschriebenen Zahl als binäre Zeichenkette $b_4b_3b_2b_1b_0$ ist also 10111. Wie man das Verfahren als sauberes Programm aufschreibt, soll die Informatikvorlesung behandeln.

3.5.1 Binäre Arithmetik

Die in heutigen Rechnern üblichen Zahldarstellungen benutzen für ganze Zahlen stets das Binärsystem. Man kann auch in allgemeinen Stellenwertsystemen rechnen[1], aber das wollen wir nicht allgemein beschreiben. Hat man 16, 32 oder 64 Bits zur Verfügung, kann man Zahlen zwischen 0 und $2^{16}-1 = 64K-1$, $2^{32}-1 = 4G-1$ und $2^{64}-1$ darstellen. Kommt ein Vorzeichen hinzu, verliert man bei fester Darstellungslänge ein Bit. Das wird im nächsten Abschnitt behandelt.

Die Addition kann durch sukzessive Addition der Binärstellen mit Übertrag geschehen, ganz wie in der Schule. Beispiel:

```
  1011010   x
 11001111   y
 1 1111     Übertrag
---------
100101001
```

Man macht also immer aus drei Input–Bits (x–Bit, y–Bit, Übertrag) ein neues Bit und einen Übertrag. Das Ergebnis kann nur (dezimal) gleich 0, 1, 2 oder 3 sein und ist gleich der Anzahl der gegebenen Bits. Binär sind die Ergebnisse als $00, 01, 10$ und 11 zu schreiben. Der Übertrag ist also genau dann binär gleich 1, wenn die Summe der Eingabebits 2 oder 3 ist. Und die

[1] `http://de.wikipedia.org/wiki/Arithmetik_in_Stellenwertsystemen`

neue Binärstelle ist 0, wenn die Summe der Eingabebits gerade (0 oder 2) ist, und 1, wenn sie ungerade (1 oder 3) ist. Als Boolesche Funktionen von x, y und dem Übertrag u geschrieben bekommt man also

- für die neue Binärstelle:

x	y	u	Stelle
1	1	1	1
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	1
0	0	0	0

- für den Übertrag:

x	y	u	Übertrag
1	1	1	1
1	1	0	1
1	0	1	1
1	0	0	0
0	1	1	1
0	1	0	0
0	0	1	0
0	0	0	0

Diese Booleschen Funktionen haben die disjunktiven Normalformen

- für die neue Binärstelle:

$$\begin{aligned} \text{Stelle} = \quad & x \wedge y \wedge u \\ \vee \quad & x \wedge \neg y \wedge \neg u \\ \vee \quad & \neg x \wedge y \wedge \neg u \\ \vee \quad & \neg x \wedge \neg y \wedge u \end{aligned}$$

- für den Übertrag:

$$\begin{aligned} \text{Übertrag} = \quad & x \wedge y \wedge u \\ \vee \quad & x \wedge y \wedge \neg u \\ \vee \quad & x \wedge \neg y \wedge u \\ \vee \quad & x \wedge \neg y \wedge \neg u \\ \vee \quad & \neg x \wedge y \vee u \end{aligned}$$

Aufgabe: Wer sieht das Rezept, das aus den Wertetabellen leicht eine disjunktive Normalform macht?

Die beiden oben angegebenen Booleschen Funktionen bilden zusammen einen **Volladdierer**, den man bei schrittweiser Anwendung von rechts nach links stellenweise das binäre Addieren bewerkstelligen lassen kann. Man braucht so viel Addierschritte wie man Stellen hat, weil man sequentiell von Stelle zu Stelle vorgeht. Wenn man parallel arbeitet und für jede Stelle ein solches Rechenwerk hat, geht es natürlich schneller, und wenn es keine Überträge gäbe, ginge es mit einem Schritt. Das Problem beim Bau effektiver Addierwerke sind also die Überträge. Es gibt trickreiche Strategien, den Übertrag vorherzusagen (**carry lookahead**), aber die Informatikvorlesungen sollen ja auch noch spannend bleiben. Wir wollen hier nichts verraten, auch nicht, wie man mit der **Methode des Kalifen** oder der **Conditional–Sum–Addition**[1] addiert.

Und wer immer noch nicht glaubt, daß man für das Verstehen wichtiger Informatik–Algorithmen wie der Multiplikation großer ganzer Zahlen hochkarätige High–Tech–Mathematik braucht, sollte sich mal den Schönhage–Strassen–Algorithmus[2] ansehen.

3.5.2 Zweierkomplement

Man könnte die ganzen Zahlen darstellen, indem man ein Vorzeichenbit zu einer b–adischen Zifferndarstellung hinzufügt. Das ist aber rechentechnisch aufwändiger als die in Computern übliche Methode. Man versucht, die Addition und Multiplikation von Zahlen in Binärdarstellung so zu organisieren, daß man Rechenwerke bauen kann, die unabhängig von der Vorzeichenwahl arbeiten. Die Grundidee ergibt sich zwangsläufig daraus, daß man die Operation $x + (-x) = 0$ problemlos ausführen können muß, d.h. die ganz normale bitweise Addition der Darstellungen von x und $-x$ muß Null ergeben. Und das sollte beispielsweise mit den oben beschriebenen Volladdierern Stelle für Stelle möglich sein, ohne daß die Volladdierer wissen müssen, was negative Zahlen sind.

Man muss also dafür sorgen, daß die Addition der Darstellungen von x und $-x$ in jedem Bit Null ergibt. Einfacher wäre es, in jedem Bit eine Eins zu produzieren, und zwar dann, wenn an die Bits von x einfach umdreht, um die von $-x$ zu bekommen. Dann liefert die bitweise Addition von x und $-x$ immer 1 ohne Übertrag. Dazu ein Beispiel mit 11 Binärstellen:

[1] `http://de.wikipedia.org/wiki/Conditional_Sum_Addition`
[2] `http://de.wikipedia.org/wiki/Sch%C3%B6nhage-Strassen-Algorithmus`

```
 00001011001  x
 11110100110  Bitinversion
------------
 11111111111
```

Wenn man zu dieser Bitdarstellung aus lauter Einsen noch eine 1 addiert, gibt es lauter Überträge, die nach links laufen und es entstehen rechts Nullen. Beispiel:

```
  00001011001  x
  11110100110  Bitinversion
 ------------
   1111111111
  00000000001  1 dazu
  11111111110  neuer Übertrag
 ------------
 100000000000  Null mit Übertrag
```

Das ist dann der Trick der **Zweierkomplementdarstellung**[1]. Man stellt $-x$ so dar, daß man die Bits von x umdreht, dann 1 addiert und den Übertrag nach vorn einfach ignoriert. Also:

```
  00001011001  x

  11110100110  Bitinversion
  00000000001  1 dazu
  11110100110  neuer Übertrag
 ------------
  11110100111  Darstellung von -x im Zweierkomplement
```

Jetzt testen wir das:

```
 00001011001  x
 11110100111  Darstellung von -x im Zweierkomplement
 11111111110  neuer Übertrag
------------
100000000000  Null mit Übertrag
```

[1] http://de.wikipedia.org/wiki/Zweierkomplement

Ein Nachteil ist, daß man die Arithmetik so implementieren muß, daß sie Überträge ignoriert, denn diese treten beim Addieren von Zahlen verschiedenen Vorzeichens oft auf. Ein Vorteil ist, daß man die Darstellungen negativer Zahlen an der führenden Eins erkennen kann, sofern man die positiven Zahlen und Null so schreibt, daß das vorderste Bit immer Null ist. Die größte positive Zahl bei n Binärstellen ist also die als $01\ldots1$ darzustellende Zahl $2^{n-1}-1$. Ihr Zweierkomplement ist $10\ldots01$, aber die Zahl $10\ldots00$ ist noch um 1 kleiner und legal. Sie stellt also -2^{n-1} dar, und diese Zahl hat keine positive Entsprechung. Deshalb erstreckt sich der mit 16 Bits inklusive Zweierkomplementdarstellung darstellbare Bereich von $-2^{15} = -32768$ bis $+2^{15}-1 = 32767$. Das sind zusammen $2^{16} = 64K = 65536$ Zahlen, man hat also nichts verschenkt.

Daß damit die Rechnerei sauber funktioniert, kann man verstehen, wenn man sich vorstellt, daß bei n Bits die Darstellung von $-x$ mit der $(n+1)$–bittigen von $2^{n+1}-x$ übereinstimmt, wenn man das führende Bit als Übertrag ignoriert. Die bitweise Addition einer positiven Zahl y zu einer negativen Zahl $-x$ wird also wie auf den positiven Zahlen $y+(2^{n+1}-x)$ ausgeführt, und das ist, wenn man den Übertrag ignoriert, eine Darstellung von $y-x$, entweder "normal" oder im Zweierkomplement, je nach Vorzeichen von $y-x$. Hat man zwei negative Zahlen $-x$ und $-y$ zu addieren, so erfolgt das wie $(2^{n+1}-x)+(2^{n+1}-y) = 2^{n+2}-(x+y)$, und bei Ignorierung der Überträge ist das eine Zweierkomplementdarstellung von $-(x+y)$.

Hier ist ein simples Beispiel zur Berechnung der Bitdarstellung:

```
#include <stdio.h>
#include <stdlib.h>
void printbinint(int ival)
{   /* druckt 32 Bit Binaerdarstellung von ival  */
    /* BRUTAL programmiert, nicht nachahmen :-)  */
    int i;
    int bit[32];
    for (i=0; i<32; i++)
    {
if ((ival%2)==0)
{
    bit[i]=0;  /* falls ival gerade, Einserbit = 0 */
}
else
{
    bit[i]=1; /* falls ival ungerade, Einserbit = 1 */
```

```
}
ival=(ival-bit[i])/2;    /* Reduktion von ival */
    }
    for (i=0; i<32; i++)
printf("%1d",bit[31-i]); /* rueckwaerts ausgeben */
}
int main(void)
{
    int i;
    i=37;
    printf("Binaere Darstellung von %d bei 32 bit:\n",i);
    printbinint(i);
    printf("\n");
    i=-37;
    printf("Binaere Darstellung von %d bei 32 bit:\n",i);
    printbinint(i);
    printf("\n");
}
```

mit der Ausgabe

```
Binaere Darstellung von 37 bei 32 bit:
00000000000000000000000000100101
Binaere Darstellung von -37 bei 32 bit:
11111111111111111111111111011011
```

Aufgabe: Wieso berechnet dieses Programm für negative i die 32–Bit–Zweierkomplementdarstellung? Daß es für positive i die richtige Bitfolge ausgibt, ist schnell zu sehen, aber was ist für negative i los?

3.5.3 Überlauf bei ganzen Zahlen

In realen Rechnern ist die Darstellungslänge fest, und man braucht diese feste Länge für die Zweierkomplementbildung. Beim Rechnen mit negativen Zahlen tritt dann **immer** ein Überlauf ein, der nicht zu einer Fehlermeldung führt. Leider führt dann aber auch das Rechnen mit sehr großen Zahlen nicht zu einer Fehlermeldung, sobald das führende Bit Eins wird und eine negative Zahl darstellt. Es kann also der unangenehme Fall eintreten, daß die Summe zweier positiver Zahlen negativ wird, ohne daß ein Fehler bemerkt wird. Man unterscheidet deshalb genau zwischen **Übertrag** und **Überlauf**:

- Der Übertrag (engl. **carry**) entsteht beim normalen Addieren in Stellenwertsystemen. Er wird weitergereicht und ist kein Fehler.

- Ein Überlauf (engl. **overflow**) ist ein fehlerhaftes Überschreiten der Grenzen eines Stellenwertsystems.

Nehmen wir das Beispiel einer vierstelligen Binärarithmetik. Sie kann nur Zahlen zwischen $-8 = -2^3$ und $7 = 2^3 - 1$ darstellen. Rechnet man $5 + 6$ aus, so bekommt man -5. Warum?

```
 0110   6
 0101   5
---------
 1011  11 oder -5
```

Die Darstellung von -5 im Zweierkomplement einer vierstelligen Binärarithmetik ist nämlich genau die von $2^4 - 5 = 16 - 5 = 11$.

In einer 32–bittigen Arithmetik wie in einem PC bekommt man also Überlauf–Probleme, sobald man Zahlen der Größe $2^{32}/2 = 2^{31} = 2G = 2048M$ addiert. Das ist schnell der Fall, wenn man z.B. die im Umlauf befindlichen Börsenwerte in Dollar oder Euro ansieht. **Vorsicht!**

Eine schnelle Abhilfe bekommt man, wenn man stattdessen im Datentyp `double` rechnet. Aber das sehen wir uns erst im übernächsten Abschnitt an.

Hier kommt ein Beispiel. Der primitive *C*–Code

```
#include <stdio.h>
#include <stdlib.h>
int main(void)
{
    int i,j;
    j=65536;
    printf("j         = %d\n",j);
    printf("j*j       = %d\n",j*j);
    i=32767;
    printf("i         = %d\n",i);
    printf("i*j       = %d\n",i*j);
    printf("i*j+65535 = %d\n",i*j+65535);
    printf("i*j+65536 = %d\n",i*j+65536);
}
```

hat auf einem Standard–PC mit 32–Bit–Arithmetik die Ausgabe

```
j          = 65536
j*j        = 0
i          = 32767
i*j        = 2147418112
i*j+65535 = 2147483647
i*j+65536 = -2147483648
```

ohne zu einer Fehlermeldung zu führen!

Frage: Warum ist das nicht anders zu erwarten?

Aus den obigen Überlegungen folgt nämlich, daß eine 32–Bit–Arithmetik, wenn sie im Zweierkomplement rechnet, ihre Grenzen schon bei

$$-2^{31} = -2147483648 \leq n \leq 2^{31} - 1 = 2147483647 = 2G - 1$$

hat. Wird dagegen eine reine Adreßrechnung im Typ **unsigned integer** vorgenommen, entfällt das Zweierkomplement und man kann insgesamt 4G= 2^{32} Bytes adressieren. Um den vollen Adreßraum nutzen zu können, muß die CPU zwischen Adreßrechnung umd Integer–Arithmetik unterscheiden.

3.5.4 Festkommazahlen

Wenn man den Bereich der ganzen Zahlen in Richtung auf Brüche und reelle Zahlen verlassen will, erweitert man die Darstellung (3.11) für positive Zahlen z durch Zulassung negativer Exponenten. Man bekommt

$$z = \sum_{j=-m}^{k} b_j b^j$$

und als Zeichenkette

$$b_k \dots b_2 b_1 b_0 . b_{-1} b_{-2} \dots b_{-m}$$

und spricht von **Festkommazahlen**[1], wenn k und m fest gewählt sind. Solche Zahlen könnte man auch als

$$z = b^{-m} \sum_{j=-m}^{k} b_j b^{j+m} = b^{-m} \sum_{i=0}^{k+m} b_{i-m} b^i$$

mit der Indexsubstitution $i = j+m$ schreiben. Das ist bis auf den Faktor b^{-m} wieder eine ganzzahlige Darstellung, die nichts wesentlich Neues bringt. Eine

[1] `http://de.wikipedia.org/wiki/Festkommazahl`

Festkommadarstellung zur Basis b mit m Nachkommastellen ist also gerade so gemacht, daß die in ihr darstellbaren Zahlen z nach Multiplikation mit b^m ganzzahlig werden. Oder: eine ganzzahlige Darstellung zur Basis b wird um m Stellen nach rechts verschoben.

Aber wir sollten uns ansehen, was passiert, wenn man beliebig gegebene Zahlen z in Festkommadarstellung bringen will. Man sucht also zu einer beliebigen Zahl z eine nahe gelegene Zahl $\tilde{z}$, die man in Festkommadarstellung exakt formulieren kann. Zunächst beschränkt man sich auf positive Zahlen; die negativen behandelt man, indem man $\tilde{z} := -\widetilde{(-z)}$ für negative z definiert. Unsere Festkommadarstellung habe m Nachkommastellen wie oben. Wir berechnen dann die positive Zahl $z \cdot b^m$ und wählen dazu eine nahegelegene ganze Zahl n. Das kann man immer so machen, daß $|z \cdot b^m - n| \leq 1$ oder sogar $|z \cdot b^m - n| \leq 1/2$ gilt. Im ersten Fall reicht es, die größte natürliche Zahl $n \leq z \cdot b^m$ zu nehmen, im zweiten Fall wählt man die zu $z \cdot b^m$ nächstgelegene Zahl. Den ersten Fall nennt man **truncation** oder **Abschneiderundung**, weil von $z \cdot b^m$ alle Nachkommastellen einfach abgeschnitten werden. Die zweite Strategie erfordert eine Auf– oder Abrundung, aber sie ist in der Informatik nicht üblich und soll hier ignoriert werden.

Wir finden also eine natürliche Zahl n mit

$$0 \leq z \cdot b^m - n \leq |z \cdot b^m - n| \leq 1$$

und können diese Ungleichung mit b^{-m} multiplizieren, um $|z - n \cdot b^{-m}| \leq b^{-m}$ zu erhalten. Die Zahl $n \cdot b^{-m} =: \tilde{z}$ hat eine exakte Festkommadarstellung, und wir bekommen

Theorem 3.14 *In einer Festkommaarithmetik zur Basis b mit m Nachkommastellen kann man zu jeder reellen oder rationalen Zahl z eine in der Arithmetik exakt darstellbare Zahl $\tilde{z}$ finden mit*

$$|z - \tilde{z}| \leq b^{-m}.$$

Man kann $\tilde{z}$ die **Festkommarundung** von z nennen und die **Rundungsabbildung** $rd(z) := \tilde{z}$ definieren mit

$$|z - rd(z)| \leq b^{-m}.$$

Definition 3.15 *Ist eine Zahl y eine Näherung einer Zahl x, so ist $|x - y|$ der* **absolute Fehler**. *Im Falle $x \neq 0$ ist $|x - y|/|x|$ der* **relative Fehler**.

Jetzt bekommt das Theorem 3.14 die Form

> In einer Festkommaarithmetik mit m Nachkommastellen zur Basis b ist jede reelle oder rationale Zahl mit einem absoluten Fehler von maximal b^{-m} darstellbar.

3.5.5 Gleitkommazahlen

Man kann also bei großer Stellenzahl jede reelle oder rationale Zahl beliebig genau durch Festkommazahlen darstellen. Allerdings sind dabei nur die Nachkommastellen relevant, und die führenden Stellen vor dem Komma müssen noch mitgerechnet werden. In der Praxis will man Rechenwerke mit fester Stellenzahl bauen, und dann ist es besser, nur noch an die Nachkommastellen zu denken und alle Zahlen so zu skalieren, daß sie keine Stellen vor dem Komma mehr haben. Mit einer festen positiven Zahl m, einer ganzen Zahl E und einem Vorzeichen kann man dann die **Gleitkommazahlen**[1] ("floating–point numbers")

$$\pm\, b^E \cdot \sum_{j=1}^{m} b_{-j} b^{-j} \tag{3.16}$$

definieren. Im Dezimalsystem wäre das einer Zifferndarstellung

$$z := \pm\, 10^E \cdot 0.b_{-1}b_{-2}\ldots b_{-m}$$

mit Ziffern $b_{-j} \in \{0, 1, \ldots, 9\}$ gleichbedeutend. Durch Verschiebung des Dezimalpunkts kann man sicherstellen, daß immer $b_{-1} > 0$ gilt. In (3.16) ist E der **Exponent**, b die **Basis**, m die **Stellenzahl** und $0.b_{-1}b_{-2}\ldots b_{-m}$ die **Mantisse**.

In Programmiersprachen gibt man Gleitkommazahlen natürlich im Dezimalsystem ein, aber sie werden intern in ein noch zu besprechendes binär codiertes Format transformiert, wobei Genauigkeitsverluste auftreten können. Die Standardschreibweise ist z.B.

$$\texttt{17.5678} = 10^2 \cdot 0.175678 = \texttt{0.175678e2}$$

d.h. man schreibt eine dezimale Festkommazahl hin und kann dann noch den Dezimalpunkt verschieben durch Hinzufügen eines Dezimalexponenten in der Schreibweise `e`n mit einem ganzzahligen n.

Wir stellen nun die Frage, wie genau man eine beliebige reelle oder komplexe Zahl Z durch eine Gleitkommazahl $fl(Z)$ der Form (3.16) darstellen kann.

[1] `http://de.wikipedia.org/wiki/Gleitkommazahl`

Sei Z eine beliebige von Null verschiedene und ohne Einschränkung auch positive Zahl. Wir suchen zuerst den Exponenten E, für den

$$\begin{array}{rcccl} b^{E-1} & \leq & Z & < & b^E \text{ d.h.} \\ b^{-1} & \leq & Zb^{-E} & < & 1 \end{array}$$

gilt. Das kann man durch Verschieben des Punktes oder Kommas einer Festkommadarstellung immer erreichen. Nun stellen wir Zb^{-E} näherungsweise durch eine k–stellige Mantisse dar, indem wir wie bei Festkommazahlen die "höheren" Stellen weglassen. Das bedeutet, daß wir eine m–stellige Gleitkommazahl $z := fl(Z)$ zur Basis b finden können mit

$$\begin{array}{rcll} 0 & \leq & Zb^{-E} - fl(Z)b^{-E} & \leq b^{-m} \\ 0 & \leq & Z - fl(Z) & \leq b^{E-k} = b^{E-1}b^{1-m} \\ & & & \leq Zb^{1-m}. \end{array}$$

Für negative Zahlen z definiert man $fl(z) := -fl(-z)$ und gibt der Zahl Null eine Sonderbehandlung durch $fl(0) := 0$.

Theorem 3.17 *In einer m–stelligen Gleitkommaarithmetik zur Basis b gibt es zu jeder reellen Zahl Z eine Gleitkommazahl $fl(Z)$ mit*

$$|Z - fl(Z)| \leq |Z|b^{1-m}.$$

Die von Null verschiedenen Zahlen sind darstellbar mit einem **relativen Fehler** *von höchstens b^{1-m}.*

Es geht also bei Gleitkommazahlen um den **relativen**, bei Festkommazahlen um den **absoluten** Fehler. Man nennt den Wert b^{1-m} auch die **Maschinengenauigkeit**[1]. Sie ist die kleinste positive Zahl x, für die man in der jeweiligen Rechnerarithmetik noch $1 + x$ von x unterscheiden kann.

Frage: Warum ist b^{1-m} genau diese Zahl?

Man sehe sich das Ganze für ein Beispiel noch einmal an. Wir nehmen eine zweistellige Gleitkomma–Dezimalarithmetik und wollen die Zahl $Z = 10.99$ darstellen. Zuerst bilden wir $Z \cdot 10^{-2} = 0.1099$ und müssen diese Mantisse im Stile der Festkommaarithmetik abschneiden bei 0.10, weil wir nur 2 Stellen haben und immer nur abrunden durch Abschneiden. Es folgt also $fl(Z) = 10^2 \cdot 0.10 = 10$ und wir haben

$$|Z - fl(Z)| = |10.99 - 10| = 0.99 = Z\frac{0.99}{Z} = Z \cdot 0.0900819$$

[1] `http://de.wikipedia.org/wiki/Maschinengenauigkeit`

während der obige Satz zu $|Z - fl(Z)| \leq Z \cdot 10^{1-2} = Z \cdot 0.1$ führt und somit ziemlich realistisch ist.

Wie kann man die Stellen einer Mantisse zur Basis b bestimmen, wenn man eine Zahl x mit $0 \leq x < 1$ vorgegeben hat? Es muß die Beziehung

$$x = \sum_{j=1}^{m} b_{-j} b^{-j} = b_{-1} b^{-1} + b_{-2} b^{-2} \ldots$$

gelten, und es folgt

$$b \cdot x - b_{-1} = b_{-2} b^{-1} + b_{-3} b^{-2} \ldots = \sum_{j=1}^{m-1} b_{-j-1} b^{-j}$$

für den Rest der Zahldarstellung. Man muß also nach der Multiplikation von x mit b die Ziffer b_{-1} ablesen, diese abziehen und dann weitermachen. Man mache sich klar, daß $0 \leq b \cdot x < b$ gilt, also muß b_{-1} der ganzzahlige Anteil von $b \cdot x$ sein, und der Rest hat wieder die Eigenschaft $0 \leq b \cdot x - b_{-1} < 1$. Ein Beispiel für $x = 0.1$ im Binärsystem:

$$\begin{array}{lcll}
2 \cdot 0.1 & = & 0.2 & \text{ganzzahliger Anteil } b_{-1} = 0 \quad \text{Rest } 0.2 - 0 = 0.2 \\
2 \cdot 0.2 & = & 0.4 & \text{ganzzahliger Anteil } b_{-2} = 0 \quad \text{Rest } 0.4 - 0 = 0.4 \\
2 \cdot 0.4 & = & 0.8 & \text{ganzzahliger Anteil } b_{-3} = 0 \quad \text{Rest } 0.8 - 0 = 0.8 \\
2 \cdot 0.8 & = & 1.6 & \text{ganzzahliger Anteil } b_{-4} = 1 \quad \text{Rest } 1.6 - 1 = 0.6 \\
2 \cdot 0.6 & = & 1.2 & \text{ganzzahliger Anteil } b_{-5} = 1 \quad \text{Rest } 1.2 - 1 = 0.2
\end{array}$$

und ab jetzt geht es periodisch weiter:

$$x = 0.00011001100110011 \ldots \text{ usw.} \tag{3.18}$$

3.5.6 Realisierung von Gleitkommazahlen

Die obige Darstellung hat ignoriert, daß der Exponent E einer Gleitkommazahl (3.16) theoretisch beliebig groß werden und dann in einer Maschine nicht mehr dargestellt werden kann. Dieser **Exponenten–Überlauf** bzw. **Exponenten–Unterlauf** wird im allgemeinen als Fehler behandelt und dem Benutzer gemeldet.

Die Gleitkommatypen entsprechen in der Regel dem *IEEE Standard for Binary Floating-Point Arithmetic 754*[1]. Der Coprozessor von PCs arbeitet intern anders, hält aber nach außen diese Spezifikation ein. Nehmen wir an, es gebe k Bits für den Exponenten, ein Vorzeichenbit und m Bits für die Mantisse. In Java gilt bei

[1] `http://de.wikipedia.org/wiki/IEEE_754`

- `float`: $k = 8, m = 23, 1$ Vorzeichenbit, 32 bits insgesamt
- `double`: $k = 11, m = 52, 1$ Vorzeichenbit, 64 bits insgesamt.

Mit dem Satz 3.17 hat man also die relative Genauigkeit

- `float`: $2^{1-23} = 2^{-22} = 2.38 \cdot 10^{-7}$
- `double`: $2^{1-52} = 2^{-51} = 4.44 \cdot 10^{-16}$.

Man kann also etwa 6 bzw. 15 sichere Dezimalstellen erwarten, wenn man beliebige Zahlen durch Näherungen in Gleitkommaform ersetzt.

Sehen wir uns die Struktur von Gleitkommazahlen für den Spezialfall von `double` genauer an. Die Zahlen des einfachen Java–Typs `double` erfordern 64 Bits, die wir von 0 bis 63 von links nach rechts durchnumerieren. Das vorderste (nullte) Bit ist das Vorzeichenbit s, dann folgen 11 Exponentenbits, und schließlich die 52 Mantissenbits:

```
s EEEEEEEEEEE bbbbbbbbbb...bbbbbbbbbbb
0 1        11 12                    63
```

Das Zeichenbit ist Null für positive Zahlen und Eins für negative. Der Exponent wird in sogenannter **excess–1023–Notation** gespeichert. Das bedeutet, daß der wahre Exponent E sich als $E = e - 1023_{dezimal}$ aus der 11-bit-Binärzahl e ergibt. Der wahre Exponent Null hat also die Binärdarstellung $01111111111 = 1023_{dezimal}$. Es wird kein Zweierkomplement für den Exponenten verwendet! Die normale Codierung für e benutzt $0_{dezimal} = 00000000000$ und $2047_{dezimal} = 11111111111$ für sehr spezielle Zwecke, die wir weiter unten beschreiben. Deshalb liegt der normale Bereich für den wahren Exponenten zwischen $1 - 1023 = -1022$ und $2046 - 1023 = 1023$. Die Mantisse wird so verstanden, daß man die 52 Mantissenbits zu einer Binärzahl M zusammenfaßt, eine 1 voransetzt und das Ganze mit 2^{-52} multipliziert. Im Binärsystem mit Binärpunkt bekommt man dann die Zahl mit der Binärdarstellung $1.M$. Insgesamt hat dann die dargestellte Zahl den Wert

$$(-1)^s \cdot 2^{e-1023} \cdot (1.M)$$

Wer aufgepaßt hat, wird bemerken, daß man damit nicht die Null codieren kann. Man reserviert $e = 0$ zusammen mit $M = 0$ für die Null, und dann verbleibt das Zeichenbit. Das führt zu der Kuriosität, daß $+0$ und -0 verschieden codiert sind.

Die Zahlen mit $e = 0$ und $M > 0$ spielen auch eine Sonderrolle. Sie werden als unnormalisiert bezeichnet und als

$$(-1)^s * 2^{-1022} * (0.M)$$

interpretiert. Sie schließen sich an den Standard-Zahlbereich in der Gegend der Null an.

Aber noch fehlt die Interpretation von Binärdarstellungen mit $e = 2047$. Ist $M > 0$ für so eine Zahl, so ist `NaN` (*not a number*) gemeint. Dies ist eine fiktive nicht existierende Gleitkommazahl, die man als Resultat illegaler Operationen erzeugt (Division Null durch Null, Logarithmus negativer Zahlen usw.). Operationen mit `NaN` ergeben definitionsgemäß wieder `NaN`, so daß sich ein Zwischenergebnis mit `NaN` auf alle weiteren Rechnungen auswirkt.

Die Spezialfälle mit $e = 2047, M = 0$ und $s = 0$ bzw. 1 werden als `+Infinity` bzw. `-Infinity` interpretiert. Auch diese Werte sind in der Implementierung mathematischer Funktionen zu berücksichtigen. Sie entstehen z.B. bei Division positiver oder negativer Zahlen durch Null, oder bei Überlauf des Exponentenbereichs.

Hier ist ein Beispiel:

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
int main(void)
{
    double xnan, xinf;
    xinf=log(0.0);
    xnan=log(-1.0);
    printf("xinf        = %25.15e\n",xinf);
    printf("-xinf       = %25.15e\n",-xinf);
    printf("xnan        = %25.15e\n",xnan);
    printf("xnan+xinf   = %25.15e\n",xnan+xinf);
    printf("xinf-xinf   = %25.15e\n",xinf-xinf);
    printf("xnan+xnan   = %25.15e\n",xnan+xnan);
    printf("xnan+1.0    = %25.15e\n",xnan+1.0);
    printf("exp(+1.0e12)= %25.15e\n",exp(+1.0e12));
    printf("exp(-1.0e12)= %25.15e\n",exp(-1.0e12));
    printf("1/(-xinf)   = %25.15e\n",1/(-xinf));
}
```

mit der Ausgabe

```
xinf        =                       -inf
-xinf       =                        inf
xnan        =                        nan
xnan+xinf   =                        nan
xinf-xinf   =                        nan
xnan+xnan   =                        nan
xnan+1.0    =                        nan
exp(+1.0e12)=                        inf
exp(-1.0e12)=     0.000000000000000e+00
1/(-xinf)   =     0.000000000000000e+00
```

Wir haben hier ein kleines Programm zum Ansehen der Binärdarstellung von double–Zahlen:

```
#include <stdio.h>
#include <stdlib.h>
void printbinint(int ival)
{   /* druckt 32 Bit Binaerdarstellung von ival  */
    /* BRUTAL programmiert, nicht nachahmen :-)  */
    int i;
    int bit[32];
    for (i=0; i<32; i++)
    {
if ((ival%2)==0)
{
    bit[i]=0;  /* falls ival gerade, Einserbit = 0 */
}
else
{
    bit[i]=1; /* falls ival ungerade, Einserbit = 1 */
}
ival=(ival-bit[i])/2;    /* Reduktion von ival */
    }
    for (i=0; i<32; i++)
printf("%1d",bit[31-i]); /* rueckwaerts ausgeben */
}
int main(void)
{   /* KRIMINELLES Programm, nicht nachahmen :-)  */
    double y;
    double* yadr;  /* eine double - Adresse */
```

```
    int* iadr;     /* eine int - Adresse */
    y=0.125;
    yadr=&y;   /* das holt die Adresse von y */
    iadr=(int*) yadr;  /* und deutet sie als Startadresse von 32 Bit ints um */
    /* bei 64 bits für double werden wir 2 ints dort rausholen */
    printf("y         = %25.15e\n",y);
    printf("Binaere Darstellung bei 64 bit:\n");
    printbinint(iadr[1]);
    printbinint(iadr[0]);
    printf("\n");
}
```

mit der Ausgabe

```
y         =     1.250000000000000e-01
Binaere Darstellung bei 64 bit:
0011111111000000000000000000000000000000000000000000000000000000
```

Das müssen wir uns ansehen. Daß die Rest–Mantisse Null ist, kann niemand verwundern, weil wir $x = 1/8 = 2^{-3}$ gesetzt haben. Es geht also nur um das (triviale) Zeichenbit und den Exponenten. Der wahre Exponent ist $E = -3 = e - 1023$, also $e = 1020_{dez} = 011111111100_{bin}$.

Dasselbe nochmal für $y = 0.1$. Die Ausgabe ist

```
y         =     1.000000000000000e-01
Binaere Darstellung bei 64 bit:
0011111110111001100110011001100110011001100110011001100110011010
```

und in etwas besser lesbarer Schreibweise:

```
0 01111111011 1001100110011001100110011001100110011001100110011010
```

In direkter binärer Schreibweise gilt $01111111011_{bin} = 1019_{dez} = e = E + 1023$, also $E = -4_{dez}$. Vor die Mantisse müssen wir noch die obligate 1 setzen, d.h. sie stellt die binäre Festkommazahl

```
1.1001100110011001100110011001100110011001100110011010
```

dar. Aus (3.18) wissen wir schon, daß

$$0.1_{dez} = 0.0001100110011001_{bin} \ldots \text{ usw.}$$

gilt, und das ist

$$0.1_{dez} = 2^{-4} \cdot 1.1001100110011_{bin} \ldots \text{ usw.}$$

Das Ergebnis ist also erwartungsgemäß, aber es liegt ein kleiner Rundungsfehler vor, den wir durch Vergleich von

```
1.1001100110011001100110011001100110011001100110011010
1.1001100110011001100110011001100110011001100110011001100110011001100...
```

sehen können. Er ist sicher kleiner als eine Einheit der nächsten, nicht dargestellten Binärstelle, und somit im Rahmen des Üblichen. Wer das nicht sofort sieht, sollte die erste Zahl x und die zweite y nennen, und dann nachrechnen, daß

$$y < x < y + 2^{-53}$$

gilt.

Wenn man das obige Programm geeignet modifiziert, sieht man auch

```
y         =                        -inf
Binaere Darstellung bei 64 bit:
1111111111100000000000000000000000000000000000000000000000000000
y         =                         nan
Binaere Darstellung bei 64 bit:
0111111111110000000000000000000000000000000000000000000000000000
```

was man aber nach IEEE754 nicht anders erwartet.

3.5.7 Rechnen mit Gleitkommazahlen

Bei Gleitkommarechnungen gelten die üblichen Rechenregeln **nicht**, weil

1. mit begrenzter Genauigkeit gearbeitet wird,
2. Zwischenergebnisse immer gerundet werden,
3. Ausnahmen wie `+Infinity` und `NaN` eintreten können.

Aber immerhin kann man erwarten, daß heutige Gleitkomma–Arithmetiken so rechnen, daß ihr Ergebnis der Addition, Subtraktion, Multiplikation oder Division zweier Gleitkommazahlen so ausfällt, als hätte man es exakt ausgerechnet und es am Ende wieder auf eine Gleitkommazahl gerundet. Wir sehen uns das einmal genauer an.

Sind x und y beliebige reelle Zahlen, so können wir ohnehin nicht mit x und y, sondern bestenfalls mit $fl(x)$ und $fl(y)$ rechnen. Wollen wir eine Operation $\circ$ ausführen, so wäre es immerhin schön, wenn das Ergebnis $fl(fl(x) \circ fl(y))$ herauskäme. Der Satz 3.17 liefert dann

$$|fl(fl(x) \circ fl(y)) - fl(x) \circ fl(y)| \leq |fl(x) \circ fl(y)| b^{1-m} \tag{3.19}$$

d.h. auch das Rechnen mit Gleitkommazahlen liefert schlimmstenfalls relative Fehler der Größenordnung b^{1-m} pro Operation. So eine Bedingung halten heutige Rechnerarithmetiken ein.

Dabei wurde aber der Fehler bezüglich der exakten Operation $fl(x) \circ fl(y)$ auf den **gerundeten** Zahlen, nicht der Fehler bezüglich des **wahren** Ergebnisses $x \circ y$ abgeschätzt. Der relative Fehler bezüglich $x \circ y$ kann aber **dramatisch größer** sein. Das wollen wir uns an einem Beispiel klarmachen.

Die wahren Zahlen seien in Dezimalnotation

```
x = 0.5678964398765
y = 0.5678962101234
```

und wir benutzen eine achtstellige Dezimalarithmetik. Also gilt $fl(x) = 0.56789643$ bzw. $fl(y) = 0.56789621$. Die Operation $\circ$ sei die Subtraktion. Wir bekommen mit unserer Dezimalarithmetik $fl(x) - fl(y) = fl(fl(x) - fl(y)) = 0.00000022 = 0.22 \cdot 10^{-6}$ weil die Arithmetik bei diesen Zahlen keinen Fehler macht. Insbesondere ist (3.19) erfüllt. So weit, so gut. Die Rechnerarithmetik trifft keine Schuld, es ist ja scheinbar auch gar nichts passiert.

Aber jetzt vergleichen wir das Ergebnis mit dem wahren Resultat $x - y = 0.2297531 \cdot 10^{-6}$. Der relative Fehler ist

$$\frac{|fl(fl(x) - fl(y)) - (x - y)|}{|x - y|} = \frac{|0.22 \cdot 10^{-6} - 0.2297531 \cdot 10^{-6}|}{0.2297531 \cdot 10^{-6}} = 0.0424504$$

also über 4%, obwohl wir mit 8 Dezimalstellen rechnen!

Man sehe sich den Effekt noch einmal genauer an. Wir subtrahieren die fast gleichgroßen Zahlen x und y bzw. $fl(x)$ und $fl(y)$. Durch die Subtraktion löschen sich die führenden Mantissenstellen gegenseitig aus. Das Ergebnis ist 0.0000002297531 bzw. 0.00000022, und man kann sehen, daß das Gleitkommaergebnis nur noch zwei brauchbare Stellen hat, während das wahre Ergebnis noch 7 hat. Das Gleitkommaergebnis verhält sich so, als hätte man nur mit zwei statt mit 8 Dezimalstellen gerechnet, und der bei nur zweistelliger Rechnung zu erwartende schlimmste relative Fehler ist $10^{1-2} = 0.1 = 10\%$. Das entspricht genau unserer Beobachtung.

Man kann zeigen, daß die Subtraktion etwa gleichgroßer Zahlen die einzige "bösartige" Operation dieser Art ist. Das liegt daran, daß es bei allen anderen Operationen nicht eintreten kann, daß die Mantisse des Resultats führende Nullen bekommt, bevor das Ergebnis renormalisiert wurde.

Beim Rechnen mit Gleitkommazahlen ist die Subtraktion fast gleichgroßer Zahlen zu vermeiden!

Es gilt die Faustregel, daß der Verlust von j Stellen durch **Auslöschung**[1] einem Verlust von j Stellen in der Genauigkeit der Rechnerarithmetik entspricht. In unserem Falle hatten wir 6 der 8 Stellen der Rechnerarithmetik verloren.

Zum Beispiel ist es ein Kunstfehler, die Funktion

$$f(x) := \frac{1}{x} - \frac{1}{x+1}$$

für große x so zu berechnen, wie sie definiert ist. Die Form

$$f(x) = \frac{1}{x(x+1)}$$

kann ohne Auslöschung berechnet werden. Aber der Auslöschungseffekt tritt nicht in dramatischer Form ein, wenn $x \approx 0$ oder $x \approx -1$ gilt.

Hier ist ein kleines Beispielprogramm:

```
#include <stdio.h>
#include <stdlib.h>
int main(void)
{
    double y, z, diff, relf, tru;
    y  =1.123456789012345;
    z  =1.123456789000000;
    tru=0.000000000012345;
    diff=y-z;
    printf("y         = %25.15e\n",y);
    printf("z         = %25.15e\n",z);
    printf("diff      = %25.15e\n",diff);
    printf("tru       = %25.15e\n",tru);
    relf=(diff-tru)/tru;
    printf("relf      = %25.15e\n",relf);
}
```

mit der Ausgabe

[1] http://de.wikipedia.org/wiki/Ausl%C3%B6schung_%28numerische_Mathematik%29

```
y         =     1.123456789012345e+00
z         =     1.123456789000000e+00
diff      =     1.234501390001697e-11
tru       =     1.234500000000000e-11
relf      =     1.125963302164969e-06
```

3.6 Reelle Zahlen

Wir gehen zurück auf den Begriff des angeordneten Körpers aus Definition 3.7 und halten fest, daß die rationalen Zahlen ein Standardbeispiel sind. Die reellen Zahlen sind, wie man aus der Schule "weiß", auf der Zahlengeraden angeordnet und füllen sie "lückenlos", aber das kann man nicht ganz so einfach in eine saubere Form bringen. Man kann außerdem leicht beweisen (mündlich in der Vorlesung), daß die wichtige Zahl $\sqrt{2}$ (sie ist die Länge der Diagonale des Einheitsquadrats) nicht rational ist, und deshalb wird es dringend, die reellen Zahlen einzuführen. Man könnte die unendlichen Dezimalbrüche heranziehen, aber dann müßte man ihre "Konvergenz" untersuchen. Eine andere Variante wird durch folgende Begriffe vorbereitet:

Definition 3.20 *Es sei K eine Menge mit einer Ordnungsrelation $\leq$.*

1. *Ein Element $y \in K$ heißt* **obere Schranke** *einer Teilmenge $M \subseteq K$, wenn für alle $x \in M$ die Relation $x \leq y$ gilt.*

2. *Hat eine Teilmenge M von K eine obere Schranke, so heißt M* **nach oben beschränkt**. *Ganz analog definiert man die* **untere Schranke** *und die Beschränktheit nach unten.*

3. *Eine Teilmenge M von K heißt M* **beschränkt**, *wenn sie nach oben und nach unten beschränkt ist.*

4. *Eine obere Schranke y einer Teilmenge M von K heißt* **Maximum** *von M, wenn y in M liegt. Analog definiert man das* **Minimum**.

5. *Eine obere Schranke y einer Teilmenge M von K heißt* **Supremum** *von M, wenn y die kleinstmögliche obere Schranke in K ist. Analog definiert man das* **Infimum**.

Man mache sich klar, daß die obere Schranke einer Menge M nicht zu M selbst gehören muß. Ein Supremum muß nicht immer existieren. Man sehe sich die Beispiele aus [4], S. 224/225 an.

An der Menge $\{x \in \mathbb{Q} \ : \ x^2 < 2\}$, die nach oben beschränkt ist aber kein Maximum und in $\mathbb{Q}$ auch kein Supremum hat, wird nun klar, wie man

die reellen Zahlen definieren kann. Man will unter $\sqrt{2}$ genau diese oder eine äquivalente Menge verstehen. Das führt zu folgender Konstruktion:

1. Man betrachtet alle Teilmengen von $\mathbb{Q}$, die eine obere Schranke haben.
2. Zwei solche Teilmengen erklärt man als äquivalent, wenn sie dieselben Mengen von oberen Schranken haben. Das ist natürlich eine Äquivalenzrelation.
3. Die entstehenden Äquivalenzklassen nennt man **reelle Zahlen**, und die Menge dieser Zahlen bezeichnet man mit $\mathbb{R}$.

Die hier verfolgte Konstruktionsmethode ist nicht die einzige[1], aber die anderen sind keineswegs einfacher zu verstehen.

Die rationalen Zahlen sind dann als Äquivalenzklassen der einelementigen Mengen $\{x\}$ für $x \in \mathbb{Q}$ in den reellen Zahlen enthalten. Man hat natürlich jetzt die Definition der Rechenoperationen und der Anordnungsrelationen neu durchzuführen. Man bekommt die positiven reellen Zahlen, indem man sich in der obigen Definition auf Teilmengen aus positiven rationalen Zahlen beschränkt. Und auf den positiven reellen Zahlen kann man die Operationen als

$$\begin{aligned} [M] + [N] &:= [\{x + y \,:\, x \in M,\ y \in N\}] \\ [M] \cdot [N] &:= [\{x \cdot y \,:\, x \in M,\ y \in N\}] \end{aligned}$$

für Äquivalenzklassen $[M]$ und $[N]$ zweier nach oben beschränkter Teilmengen M und N von rationalen Zahlen definieren. Das erweitert man sinngemäß auf alle reellen Zahlen, und man bekommt wieder alle Gesetze eines kommutativen Körpers mit einer durch einen Positivitätsbereich definierten Anordnung. Die rationalen Zahlen sind darin enthalten. Und man kann dann zeigen, daß gilt

$$[\{x \in \mathbb{Q} \,:\, x^2 < 2\}] \cdot [\{x \in \mathbb{Q} \,:\, x^2 < 2\}] = [\{2\}].$$

Die Definition der reellen Zahlen ist so gemacht, daß man die "Lückenlosigkeit" dieser "Zahlen" genau formulieren und beweisen kann:

Theorem 3.21 *Jede nach oben bzw. unten beschränkte Menge reeller Zahlen hat ein Supremum bzw. Infimum. (***Vollständigkeit** *der reellen Zahlen).*

[1] `http://de.wikipedia.org/wiki/Reelle_Zahlen`

Man kann dann unter anderem beweisen, daß man beliebige m–te Wurzeln positiver Zahlen z ziehen kann:

$$\sqrt[m]{z} := [\{x \in \mathbb{Q} \ : \ x^m < z\}]$$

und was π sein soll, ist etwa durch

$$\pi^{-1} := [\{r \in \mathbb{Q} \ : \ \text{Kreis mit Radius } r \text{ hat Fläche } < 1\}]$$

anzudeuten, aber an dieser Stelle noch nicht exakt nachvollziehbar, weil die Begriffe "Kreis" und "Fläche" noch klärungsbedürftig sind.

Man kann beweisen, daß die reellen Zahlen $\mathbb{R}$ in einem gewissen Sinne den "einzigen" vollständigen angeordneten Körper bilden.

Definition 3.22 *Als Teilmengen der reellen Zahlen definiert man* **Intervalle** *wie folgt:*

$$\begin{aligned} [a,b] &:= \{x \in \mathbb{R} \ : \ a \leq x \leq b\} \\ (a,b] &:= \{x \in \mathbb{R} \ : \ a < x \leq b\} \\ [a,b) &:= \{x \in \mathbb{R} \ : \ a \leq x < b\} \\ (a,b) &:= \{x \in \mathbb{R} \ : \ a < x < b\} \end{aligned} \tag{3.23}$$

für alle $a, b \in \mathbb{R}$. Solche Intervalle nennen wir **beschränkt**. *Sinngemäß kann man auch $\mathbb{R} = (-\infty, \infty)$ oder $(-\infty, b) \subset (-\infty, b]$ sowie $(a, \infty) \subseteq [a, \infty)$ als unbeschränkte Intervalle definieren, wobei ∞ das Symbol für "Unendlich" ist.*

An dieser Stelle wollen wir Ungleichungen mit reellen Zahlen üben.

Das geometrische Mittel $\sqrt{x \cdot y}$ zweier positiver Zahlen ist der Flächeninhalt des "mittleren" Quadrats, das dieselbe Fläche hat wie das Rechteck mit Seitenlängen x und y. Das arithmetische Mittel von x und y ist der **Mittelwert** $\frac{x+y}{2}$, der "mitten zwischen" x und y auf der Zahlengeraden liegt.

Theorem 3.24 *(***Arithmetisch–geometrisches Mittel**[1]*)*
Es seien x und y positive reelle Zahlen. Dann gilt

$$\min(x,y) \leq \sqrt{x \cdot y} \leq \frac{x+y}{2} \leq \max(x,y),$$

d.h. das **geometrische Mittel** $\sqrt{x \cdot y}$ *ist nicht größer als das* **arithmetische Mittel** $\frac{x+y}{2}$.

[1] `http://de.wikipedia.org/wiki/Arithmetisch-geometrisches_Mittel`

Wir beweisen erst einmal für **alle** reellen Zahlen

$$\min(x, y) \leq \frac{x+y}{2} \leq \max(x, y).$$

Die Aussage ist symmetrisch gegen Vertauschung von x und y. Also können wir $x \leq y$ annehmen und es ist dann zu zeigen

$$x \leq \frac{x+y}{2} \leq y.$$

Weil wir $x \leq y$ haben, folgt auch $x/2 \leq y/2$. Addieren wir $x/2$ zu dieser Ungleichung, so folgt $x \leq \frac{x+y}{2}$. Ganz analog beweist man die rechte Ungleichung.

Der Beweis für $\sqrt{x \cdot y} \leq \frac{x+y}{2}$ vereinfacht sich, wenn wir $x = a^2$ und $y = b^2$ setzen und

$$2a \cdot b \leq a^2 + b^2$$

beweisen. Das ist einfach:

$$2a \cdot b = a^2 + b^2 - (a-b)^2 \leq a^2 + b^2,$$

weil $(a-b)^2 \geq 0$ gilt.

Jetzt brauchen wir nur noch

$$\min(x, y) \leq \sqrt{x \cdot y}$$

zu beweisen, und wieder können wir aus Symmetriegründen $x \leq y$ annehmen. Wir zielen dann auf

$$x \leq \sqrt{x \cdot y} \leq y.$$

Aus $x \leq y$ und folgt wegen der Monotonie der Wurzelfunktion auch $\sqrt{x} \leq \sqrt{y}$, und durch Multiplikation mit $\sqrt{x}$ bzw. $\sqrt{y}$ ergibt sich die Behauptung. □

4 Lineare Algebra

[1] Wir haben jetzt die Zahlen hinter uns (nur noch die komplexen Zahlen stehen aus) und wollen mit den Zahlen etwas anfangen. Wir bilden erst einmal Paare oder Tripel von Zahlen. Das führt uns später auf Geometrie[2], denn es dürfte aus der Schule bekannt sein, daß das cartesische Produkt $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ eine "Ebene" bildet, in der man z.B. Punkte, Geraden und Kreise definieren kann. Die Ebene hat zwei Dimensionen, und es liegt nahe, zu vermuten, daß das n–fache cartesische Produkt $\mathbb{R}^n$ ein Gebilde ist, das n Dimensionen hat. Allerdings ist zu klären, was "Dimension" bedeuten soll.

Aber das ist noch nicht alles. Der $\mathbb{R}^n$ kann zum Begriff des Vektorraums verallgemeinert werden, und damit kann man z.B. auch Räume behandeln, in denen jeder "Punkt" eine Funktion ist, wie $\sin(x)$ oder x^{17}. Große Teile des wissenschaftlichen Rechnens auf Hochleistungscomputern spielen sich in Vektorräumen von Funktionen ab, und deshalb müssen wir etwas weiter ausholen und dürfen nicht im $\mathbb{R}^2$ steckenbleiben.

In diesem Kapitel wird erst einmal alles weggelassen, was eine Abstandsmessung erfordert. Wir holen das später nach. Auch ohne Abstandsmessung kann man definieren, was Punkte, Geraden und Ebenen sind, und was unter "Dimension" zu verstehen ist. In diesem Sinne treiben wir jetzt schon Geometrie, aber im engeren Sinne des Wortes erfordert Geometrie einen Abstandsbegriff (*Geometrie* = Wissenschaft vom Messen der Erde).

4.1 Vektorräume

4.1.1 Grundbegriffe

Wir steuern also zuerst auf den Begriff des "Raumes"[3] mit einer "Dimension"[4] zu. In der Mathematik gibt es, wie die angegebenen Links zeigen, viele verschiedene hochinteressante Begriffe von "Raum" und "Dimension", aber wir machen uns hier das Leben etwas leichter und beschränken uns auf Vektorräume und deren Dimension. Zuerst betrachten wir die cartesischen Produkte $\mathbb{R}^n$. Im Falle $n = 2$ hat man die aus der Schule bekannte Veranschaulichung des zweidimensionalen Raumes durch zwei Koordinatenachsen. Die beiden Achsen werden beschrieben durch die Punktmengen $\{(x_1, 0) \ : \ x_1 \in \mathbb{R}\}$ und $\{(0, x_2) \ : \ x_2 \in \mathbb{R}\}$. Ein beliebiges Paar

[1] http://de.wikipedia.org/wiki/Lineare_Algebra
[2] http://de.wikipedia.org/wiki/Geometrie
[3] http://de.wikipedia.org/wiki/Raum_%28Mathematik%29
[4] http://de.wikipedia.org/wiki/Dimension_%28Mathematik%29

$(x_1, x_2) \in \mathbb{R}^2$ kann auf die Paare $(x_1, 0)$ und $(0, x_2)$ auf den Achsen projiziert werden. Man erinnere sich hier an die Projektionen in relationalen Datenbanken.

Ein Paar $(x_1, x_2) \in \mathbb{R}^2$ kann man als **Punkt** eines Raumes ansehen, aber man kann die gerichtete Strecke vom Nullpunkt $(0, 0)$ hin zu (x_1, x_2) auch als **Vektor**[1] auffassen. In der Geometrie[2] wird zwischen Punkträumen und Vektorräumen sorgfältig unterschieden. Hier können wir den Begriff "gerichtete Strecke" noch nicht definieren, und wir benutzen den allgemein üblichen Zugang über den abstrakten Vektorbegriff, ohne diesen geometrisch zu interpretieren.

Im allgemeinen kann man die n–Tupel $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$ als Punkte oder Vektoren eines "Raumes" ansehen. Durch Multiplikation mit einer weiteren reellen Zahl α (einem **Skalar**[3] im Gegensatz zu x, wir nehmen dafür griechische Buchstaben $\alpha, \beta, \gamma, \ldots$) kann man den neuen Vektor

$$\alpha \cdot x := (\alpha \cdot x_1, \ldots, \alpha \cdot x_n) \tag{4.1}$$

bilden. Ferner definiert man die komponentenweise Addition

$$x + y := (x_1, \ldots, x_n) + (y_1, \ldots, y_n) := (x_1 + y_1, \ldots, x_n + y_n) \tag{4.2}$$

für alle $x := (x_1, \ldots, x_n),\ y := (y_1, \ldots, y_n) \in \mathbb{R}^n$.

Man mache sich im $\mathbb{R}^2$ klar, was die Vektoraddition und die Skalarmultiplikation geometrisch bedeuten (mündlich in der Vorlesung).

Das Ganze klappt auch für eine beliebigen Skalarenkörper $\mathbb{K}$, der $\mathbb{R}$ ersetzt. Insbesondere kann man an $\mathbb{K} = \mathbb{C}$ oder $\mathbb{K} = \mathbb{Q}$ denken, aber in der Theorie der linearen Codes braucht man Vektorräume von Polynomen über etwas exotischeren endlichen Körpern.

Definition 4.3 *Eine nichtleere Menge V heißt* **Vektorraum**[4] *über einem Körper $\mathbb{K}$, wenn gilt:*

1. *Es gibt eine* **Addition** $V \times V \to V \ : \ (u, v) \mapsto u + v$, *unter der V eine abelsche Gruppe (siehe Definition 3.5) ist, d.h. die Addition*

[1] http://de.wikipedia.org/wiki/Vektor
[2] http://de.wikipedia.org/wiki/Geometrie
[3] http://de.wikipedia.org/wiki/Skalar_%28Mathematik%29
[4] http://de.wikipedia.org/wiki/Vektorraum

ist assoziativ und kommutativ, es gibt ein neutrales Element 0 *und zu jedem Element* $v \in V$ *ein eindeutiges Inverses* $-v \in V$*. Im Detail:*

$$\begin{array}{rcll} (u+v)+w & = & u+(v+w) & \text{für alle } u,v,w \in V \\ u+v & = & v+u & \text{für alle } u,v \in V \end{array}$$

und es gibt einen speziellen Vektor $0 \in V$ *mit*

$$u + 0 \;=\; u \quad \text{für alle } u \in V$$

und zu jedem $v \in V$ *gibt es ein (eindeutig bestimmtes) Element* $-v \in V$ *mit*

$$v + (-v) = 0.$$

2. *Es gibt eine* **Skalarmultiplikation** *als bilineare Abbildung* $\mathbb{K} \times V \to V \;:\; (\alpha, v) \mapsto \alpha \cdot v$*, d.h.*

$$\begin{array}{rcl} (\alpha+\beta)\cdot v & = & \alpha \cdot v + \beta \cdot v \\ \alpha \cdot (u+v) & = & \alpha \cdot u + \alpha \cdot v \end{array}$$

für alle $\alpha, \beta \in K$ *und* $u, v \in V$*.*

3. *Es gilt ferner*

$$\begin{array}{rcl} (\alpha \cdot \beta)\cdot v & = & \alpha \cdot (\beta \cdot v) \\ 1 \cdot v & = & v \end{array}$$

für alle $\alpha, \beta \in K$ *und* $v \in V$*.*

Ist $U \subseteq V$ *eine Teilmenge eines Vektorraums* V *über einem Körper* $\mathbb{K}$*, und gelten alle Vektorraumaxiome auch für* U*, so heißt* U *ein* **Untervektorraum** *oder* **Unterraum** *von* V*.*

Oben ist die 1 natürlich die Eins im Körper $\mathbb{K}$. Vektorräume haben in der Regel keine Multiplikation, und deshalb auch keine "Eins", die neutrales Element einer Multiplikation wäre. Anders ist das mit der Null. Es gibt eine im Körper und eine im Vektorraum V, und es gilt z.B.

$$0 \cdot v = 0 \text{ für alle } v \in V.$$

Hier ist links die Null in $\mathbb{K}$ gemeint, und rechts steht die in V. Es hat sich eingebürgert, die beiden Nullen nicht mit verschiedenen Bezeichnungen zu versehen, weil es in der Regel keine Probleme gibt.

Zum Beweis der obigen Gleichung schließt man wie folgt:

$$0 \cdot v = (1 + (-1)) \cdot v = 1 \cdot v + (-1) \cdot v = v - v = 0.$$

Wir sollten uns die wichtigsten Vektorräume etwas genauer ansehen. Als Standardbeispiele kommen $\mathbb{K}^n$ bzw. $\mathbb{R}^n$ oder $\mathbb{Q}^n$ mit komponentenweiser Addition (4.2) und der Skalarmultiplikation (4.1) in Frage. Es wird sich herausstellen, daß n in solchen Fällen die "Dimension" des Raumes angibt. In der Bezeichnungsweise werden wir die Vektoren der Räume der Form $\mathbb{K}^n$ als $x = (x_1, \ldots, x_n)$ schreiben, während wir für allgemeine Vektorräume lieber $u, v, \ldots$ verwenden. Man mache sich klar, daß für solche Vektoren keine indizierten "Koordinaten" oder "Komponenten" wie v_j existieren.

Jeder Vektorraum hat zwei triviale Untervektorräume: sich selbst und den **Nullraum**, der nur aus der Null besteht. Und wenn man irgendeinen festen Vektor $v \in V \setminus \{0\}$ hernimmt, ist

$$\text{span}\,(\{v\}) := \{\alpha \cdot v \ : \ \alpha \in K\}$$

ein Untervektorraum, der nicht der Nullraum ist.

Frage: Warum? Wie sieht so ein Raum "geometrisch" aus?

Aufgabe: Wie sehen die Unterräume der Vektorräume $\mathbb{R}$, $\mathbb{R}^2$ und $\mathbb{R}^3$ aus?.

Es gibt noch viel mehr Möglichkeiten, Vektorräume zu erzeugen:

Theorem 4.4 *Es sei M eine Menge und $\mathbb{K}$ ein kommutativer Körper. Dann ist die Menge $\mathbb{K}^M$ der Abbildungen von M in $\mathbb{K}$ ein Vektorraum über $\mathbb{K}$ mit den Verknüpfungen*

$$\begin{array}{rcl} f + g & := & x \mapsto f(x) + g(x) \text{ für alle } x \in M \\ \alpha \cdot f & := & x \mapsto \alpha \cdot f(x) \text{ für alle } x \in M \end{array}$$

für alle $\alpha \in \mathbb{K}$ und alle $f, g \ : \ M \to \mathbb{K}$.

Mit $M := \{1, 2, \ldots, n\}$ stimmen $\mathbb{K}^M$ und $\mathbb{K}^n$ überein. Man mache sich klar, daß eine Funktion $f \ : \ \{1, 2, \ldots, n\} \to \mathbb{K}$ als n–Tupel von Werten $(f(1), \ldots, f(n)) \in \mathbb{K}^n$ geschrieben werden kann. Für unendliche Mengen wie $M = \mathbb{N}$ bekommt man noch interessantere, nämlich unendlichdimensionale Räume heraus, zum Beispiel den Raum $\mathbb{R}^{\mathbb{N}}$ aller reellen **Zahlenfolgen** oder den Raum $\mathbb{R}^{\mathbb{R}}$ aller reellwertigen Funktionen auf $\mathbb{R}$.

Beispiel: Polynome[1] als linearer Unterraum von $\mathbb{R}^{\mathbb{R}}$.
Man definiere sich in $\mathbb{R}^{\mathbb{R}}$ die abstrakten "Vektoren"

$$u_k \ : \ x \mapsto x^k \text{ für alle } x \in \mathbb{R} \quad \text{für alle } k \geq 0.$$

[1] `http://de.wikipedia.org/wiki/Polynom`

Jeder "Vektor" ist also eine Funktion. Man kann die Skalarmultiplikation und die Vektoraddition durchführen, und

$$3 \cdot u_0 - 5 \cdot u_1 \ : \ x \mapsto 3x - 5x^3 \text{ für alle } x \in I\!R$$

ist z.B. wieder eine Funktion. Ebenso für $2 \cdot \sin(x) - 7 \cdot e^x$. Die Funktionen u_k bezeichnet man auch als **Monome** und schreibt sie etwas lax als

$$x^k \ : \ x \mapsto x^k, \ k \in I\!N.$$

In der obigen Zeile steht links das Symbol x^k für eine komplette Funktion aus $I\!R^{I\!R}$, während rechts eine reelle Zahl x^k steht. Ein allgemeines **Polynom** vom **Grad** n mit reellen Koeffizienten ist dann eine Abbildung

$$p \ : \ x \mapsto \sum_{k=0}^{n} \alpha_k x^k \tag{4.5}$$

mit $\alpha_0, \ldots, \alpha_n \in I\!R$ und $\alpha_n \neq 0$.

4.2 Komplexe Zahlen

Auf $I\!R^2$ oder $I\!R^3$ oder anderen Obermengen von $I\!R$ könnte man versuchen, eine Addition und eine Multiplikation so einzuführen, daß man wieder einen Körper erhält. Das funktioniert nur mit Abstrichen an den Eigenschaften angeordeneter und vollständiger kommutativer Körper, denn die reellen Zahlen bilden in gewissem Sinne den einzigen Körper mit diesen Eigenschaften.

Auf $I\!R^2$ hat man die Vektorraumaddition, die alle Gesetze einer abelschen Gruppe erfüllt und obendrein die geometrische Vektoraddition realisiert.

Man kann aber auch eine Multiplikation so einführen, daß man zwar die Anordnung verliert, aber die Lösbarkeit der Gleichung $x^2 = -1$ bekommt. Auf $\mathbb{Q}$ oder $I\!R$ hat nämlich die Anordnung der Zahlen zur Folge, daß man die Gleichung $x^2 = -1$ nicht lösen kann, denn für jedes $x \neq 0$ gilt $x^2 > 0 > -1$. Die Menge $\{x \in \mathbb{Q} \ : \ x^2 < -1\}$ ist leer, und deshalb bringt der Konstruktionstrick der reellen Zahlen nichts.

Wenn man aber auf die Anordnung verzichtet und Paare reeller Zahlen bildet, kann man neben der Vektoraddition auf $I\!R^2$ die Multiplikation

$$(x, y) \cdot (u, v) \ := \ (x \cdot u - y \cdot v, x \cdot v + u \cdot y) \tag{4.6}$$

für alle Paare $(x, y), (u, v) \in I\!R \times I\!R$ bilden und nachrechnen, daß man einen kommutativen Körper $\mathbb{C}$ bekommt, der $I\!R$ als $\{(x, 0) \ : \ x \in I\!R\}$ enthält.

Man nennt ihn den **Körper der komplexen Zahlen**[1]. Das Paar $i := (0,1)$ hat dann die schöne Eigenschaft

$$i^2 = (0,1) \cdot (0,1) = (-1,0)$$

und jedes beliebige Paar $(x,y) \in \mathbb{C}$ hat die Darstellung

$$\begin{aligned}(x,y) &= x \cdot (1,0) + y \cdot (0,1) \\ &=: x + iy.\end{aligned}$$

Die erste Zeile ist im Sinne der Vektorraumeigenschaften von $\mathbb{R}^2$ klar, und die zweite ist eine klammerfreie Kurzschreibweise, die sich sehr bewährt hat, die aber nichts als eine Abkürzung für die Vektorschreibweise ist. Man behandelt i einfach wie eine Variable mit der Eigenschaft $i^2 = -1$ und rechnet formal wie mit Polynomen, also z.B. $2i \cdot (4 - 3i) = 8i - 6i^2 = 8i + 6 = 6 + 8i$. Die Gleichung (4.6) hat dann die Form

$$\begin{aligned}(x,y) \cdot (u,v) &= (x + iy) \cdot (u + iv) \\ &= x \cdot u + i^2 \cdot y \cdot v + i(x \cdot v + u \cdot y) \\ &= x \cdot u - y \cdot v + i(x \cdot v + u \cdot y) \\ &= (x \cdot u - y \cdot v, x \cdot v + u \cdot y)\end{aligned}$$

für alle $x, y, u, v \in \mathbb{R}$ oder alle $x + iy, u + iv \in \mathbb{C}$.

Mit dieser Multiplikation und der komponentenweisen Addition

$$\begin{aligned}(x,y) + (u,v) &= (x + iy) + (u + iv) \\ &= x + u + i(y + v) \\ &= (x + u, y + v)\end{aligned}$$

werden die komplexen Zahlen zu einem kommutativen **Körper**, der die reellen Zahlen enthält.

Man darf nicht glauben, die komplexen Zahlen seien eine verrückte Idee der Mathematiker, die in der Praxis unbrauchbar sei. Informatik–Anfänger auf technischen Hochschulen müssen durch die Grundlagen der Elektrotechnik, und dort sind komplexe Zahlen nicht wegzudenken, z.B. wenn man den Wechselstrom behandelt [2]. Und das setzt sich fort, wenn man **Digitale Signalverarbeitung**[3] betreibt, aber das können wir erst ganz am Ende dieser Vorlesung etwas besser erklären.

[1] `http://de.wikipedia.org/wiki/Komplexe_Zahl`
[2] `http://de.wikipedia.org/wiki/Komplexe_Wechselstromrechnung`
[3] `http://de.wikipedia.org/wiki/Digitale_Signalverarbeitung`

Im Vorgriff auf die trigonometrischen Funktionen, und weil das beim Wechselstrom, in der Nachrichtentechnik und bei der digitalen Signalverarbeitung unumgänglich ist, erklären wir hier noch kurz die **Polarkoordinaten** des $\mathbb{R}^2$, die sich vorzüglich eignen, mit komplexen Zahlen zu rechnen. Im $\mathbb{R}^2$ kann man jedes Paar (x, y) als $r \cdot (\cos\varphi, \sin\varphi)$ schreiben, und zwar mit dem Nullpunktsabstand ("**Radius**") $r = \sqrt{x^2 + y^2}$ und dem **Winkel** φ, der zwischen der x–Achse und dem durch den Nullpunkt und (x, y) definierten Vektor besteht. (*Zeichnung in der Vorlesung*).

Die auf **Euler**[1] zurückgehende Formel

$$e^{i\cdot\varphi} = \cos(\varphi) + i\sin(\varphi) \text{ für alle } \varphi \in \mathbb{R}$$

ist hier noch nicht direkt verständlich, aber macht das Rechnen mit Wechselspannungen und allgemeinen Signalen technisch sehr einfach. Wir kommen darauf zurück, wenn wir die Exponentialfunktion und die Winkelfunktionen behandeln.

Die Abbildung zwischen (x, y) und (r, φ) mit $x = r\cos\varphi$ und $y = r\sin\varphi$ ist nicht ganz unproblematisch. Außerhalb des Nullpunkts kann die Abbildung nur dann injektiv werden, wenn man φ auf eine Periode einschränkt, z.B. auf $[0, 2\pi)$ oder $[-\pi, \pi)$. Und im Nullpunkt wird die Abbildung ohnehin nicht eindeutig, weil der Winkel irrelevant ist. Die Abbildung ist in der Richtung

$$(r, \varphi) \mapsto (r\cos\varphi, r\sin\varphi) \tag{4.7}$$

rechentechnisch einfach, die Umkehrung ist schwieriger. Man kann φ aus (x, y) durch die Gleichung

$$\tan\varphi = \frac{y}{x} \text{ oder } \arctan\left(\frac{y}{x}\right) = \varphi \tag{4.8}$$

berechnen, was aber in der Praxis wegen des eventuell verschwindenden Nenners und der Mehrdeutigkeit der Tangensfunktion problematisch ist. Deshalb halten die meisten Programmiersprachen für diese Umrechnung eine spezielle Lösung parat. In MATLAB etwa gibt es das Funktionspaar

$$\begin{aligned} [x, y] &= pol2cart(\varphi, r) \\ [\varphi, r] &= cart2pol(x, y). \end{aligned}$$

In anderen Sprachen gibt es in Gleitkommaarithmetik die Funktion $\varphi = atan2(y, x)$ für den "harten" Teil (4.8) der Umrechnung. Diese Funktion arbeitet sauber auf dem ganzen $\mathbb{R}^2$.

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Euler.html`

Bei komplexen Zahlen $z := x+iy$ bekommt man das Quadrat $r^2 = x^2+y^2$ des Radius durch Multiplikation von $z := x+iy$ mit der **konjugiert komplexen** Zahl $\overline{z} := x - iy$ als

$$(x + iy) \cdot (x - iy) = x^2 + y^2 + i \cdot 0.$$

Man schreibt dann auch $|z| := \sqrt{x^2 + y^2} = \sqrt{z \cdot \overline{z}}$ oder $|x|^2 = z \cdot \overline{z}$ und nennt $|z|$ den **Absolutbetrag** von z. Der Winkel φ in der Polarkoordinatendarstellung von (x, y) wird auch das **Argument** $\arg(z)$ der komplexen Zahl $z = x + iy$ genannt. Für die **Konjugationsabbildung** $z \mapsto \overline{z}$ gelten die Regeln

$$\begin{array}{rcl} \overline{z_1 \pm z_2} & = & \overline{z_1} \pm \overline{z_2} \\ \overline{z_1 \cdot z_2} & = & \overline{z_1} \cdot \overline{z_2} \\ z_1 = \overline{z_1} & \rightarrow & z_1 \in I\!R \\ \overline{i} & = & -i \end{array}$$

für alle $z_1, z_2 \in \mathbb{C}$. Und jetzt kann man sofort nachrechnen, daß die multiplikative Inverse zu $z := x + iy \neq 0$ sich als

$$z^{-1} = 1/z = \overline{z}/|z|^2 = \frac{x - iy}{x^2 + y^2}$$

berechnen läßt, denn es folgt

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

Für den Absolutbetrag gelten die Regeln

$$\begin{array}{rcl} |z_1|^2 & = & z_1\overline{z_1} \\ |z_1 \cdot z_2| & = & |z_1| \cdot |z_2| \\ |\overline{z_1}| & = & |z_1| \\ |z_1 + z_2| & \leq & |z_1| + |z_2| \end{array}$$

für alle $z_1, z_2 \in \mathbb{C}$. Wir werden die **Dreiecksungleichung** $|z_1 + z_2| \leq |z_1| + |z_2|$ später allgemeiner beweisen. Die anderen Identitäten sind einfach zu zeigen.

> **Es ist ein krasser Fehler, komplexe Zahlen in Ungleichungen zu verwenden. Weder $<$ noch $\leq$ sind definiert. Es gibt nur den reellen Absolutbetrag $|z|$ zur Angabe der "Größe" der Zahl.**

Mit der Polarkoordinatendarstellung kann man die Multiplikation komplexer Zahlen

$$\begin{array}{rcccccl} (x, y) & = & x + iy & = & (r\cos\varphi, r\sin\varphi) & = & r\cos\varphi + ir\sin\varphi \\ (u, v) & = & u + iv & = & (s\cos\psi, s\sin\psi) & = & s\cos\psi + is\sin\psi \end{array}$$

geometrisch interpretieren, wenn man in

$$\begin{aligned}(x,y)\cdot(u,v) &= (r\cos\varphi + ir\sin\varphi)\cdot(s\cos\psi + is\sin\psi)\\ &= rs(\cos\varphi\cos\psi - \sin\varphi\sin\psi) + irs(\sin\varphi\cos\psi + \cos\varphi\sin\psi)\\ &= rs\cos(\varphi+\psi) + irs\sin(\varphi+\psi)\end{aligned}$$

die Additionstheoreme anwendet. Die Multiplikation zweier komplexer Zahlen multipliziert die Absolutbeträge und addiert die Winkel.

Mit den komplexen Zahlen kann man nicht nur die Gleichung $z^2 + 1 = 0$ lösen, sondern **jede** polynomiale Gleichung

$$\sum_{k=0}^{n} \alpha_k z^k = 0$$

vom Grad $n \geq 1$ mit reellen Koeffizienten $\alpha_0, \ldots, \alpha_n \in \mathbb{R}$ und $\alpha_n \neq 0$. Dieses wichtige Ergebnis nennt man den **Fundamentalsatz der Algebra**. Es wird im normalen Curriculum in der Vorlesung "Funktionentheorie" bewiesen, die Eigenschaften von Funktionen $f : \mathbb{C} \to \mathbb{C}$ untersucht.

Ein wichtiger Spezialfall von algebraischen Gleichungen ist die **Kreisteilungsgleichung**, bei der man zu festem $n \in \mathbb{N}$ alle $z \in \mathbb{C}$ sucht mit

$$z^n = 1.$$

Ihre n Lösungen heißen die n–ten **Einheitswurzeln** und sie teilen den **Einheitskreis**, d.h. den Kreis um 0 in $\mathbb{C}$ mit Radius 1, in genau n gleiche Teile.

Aufgabe: Für $n = 4$ bekommt man die Einheitswurzeln ± 1, $\pm i$. Was kommt für $n = 5$ oder für $n = 17$ in Polarkoordinaten heraus? Diese speziellen Einheitswurzeln kann man mit Zirkel und Lineal konstruieren, aber das überlassen wir lieber **Gauss**[1].

Im folgenden kommen auch die komplexen Zahlen als mögliche Grundkörper bei Vektorräumen in Frage. Die Skalare sind dann also komplexe Zahlen, man hat dabei aber auf die Anordnung dieser Skalare zu verzichten.

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Gauss.html`

4.2.1 Linear–, Affin– und Konvexkombinationen

Nach diesem Exkurs wollen wir uns erst einmal einige besonders einfache Unterräume von Vektorräumen ansehen. In diesem Abschnitt sind bis auf weiteres alle Vektorräume über $\mathbb{R}$ genommen. Wir werden etwas später auch allgemeine Grundkörper $\mathbb{K}$ zulassen.

Der "kleinste" Unterraum jedes Vektorraums ist $\{0\}$, er besteht nur aus dem neutralen Element der additiven Gruppe, das bei Vektorräumen auch **Nullpunkt** oder **Ursprung** genannt wird. Nimmt man einen vom Nullpunkt verschiedenen Vektor v aus einem Vektorraum V über $\mathbb{R}$ her, so ist

$$\mathbb{R} \cdot v := \{\alpha \cdot v \ : \ \alpha \in \mathbb{R}\}$$

ein Unterraum von V, der geometrisch gesehen eine Gerade bildet, die v und den Nullpunkt enthält. Ferner ist die Strecke zwischen dem Nullpunkt und v die Menge

$$[v, 0] := \{\alpha \cdot v + (1 - \alpha) \cdot 0 \ : \ \alpha \in [0, 1]\}$$

und allgemeiner ist

$$[u, v] := \{\alpha \cdot u + (1 - \alpha) \cdot v \ : \in [0, 1]$$

die Verbindungsstrecke zwischen zwei als Punkte aufgefaßten Elementen u und v eines Vektorraums über $\mathbb{R}$. Verlängert man die Verbindungsstrecke über die Endpunkte hinaus, so bekommt man mit

$$\begin{array}{rl} & \{\alpha \cdot u + (1 - \alpha) \cdot v \ : \ \alpha \in \mathbb{R}\} \\ = & \{v + \alpha \cdot (u - v) \ : \ \alpha \in \mathbb{R}\} \\ =: & v + \mathbb{R} \cdot (u - v) \\ = & \{u + \alpha \cdot (v - u) \ : \ \alpha \in \mathbb{R}\} \\ =: & u + \mathbb{R} \cdot (v - u) \end{array}$$

die komplette Gerade durch die Punkte u und v.

Diese geometrischen Sachverhalte werden in der Vorlesung an der Tafel am Beispiel des $\mathbb{R}^2$ und des $\mathbb{R}^3$ illustriert.

Wir verallgemeinern diese einfachen geometrischen Beobachtungen:

Definition 4.9 *Es seien $v_1, \ldots, v_m$ Elemente eines Vektorraums V über einem Skalarenkörper $\mathbb{K}$, und es sei $\alpha := (\alpha_1, \ldots, \alpha_m)$ ein m–Tupel aus $\mathbb{K}^m$ für ein $m \in \mathbb{N} \setminus \{0\}$.*

1. *Das Element*
$$\sum_{j=1}^{m} \alpha_j \cdot v_j$$
von V heißt dann **Linearkombination** *von $v_1, \ldots, v_m$ mit dem* **Koeffizienten***vektor $\alpha := (\alpha_1, \ldots, \alpha_m) \in \mathbb{K}^m$.*

2. *Gilt zusätzlich die Bedingung*
$$\sum_{j=1}^{m} \alpha_j = 1,$$
so spricht man von einer **Affinkombination**.

3. *Gelten zusätzlich noch die Bedingungen $\mathbb{K} = \mathbb{R}$ und*
$$\alpha_j \in [0,1] \text{ für alle } j,\ 1 \leq j \leq m,$$
so spricht man von einer **Konvexkombination**.

Im Beispiel oben ist die Strecke $[u, v]$ zwischen u und v genau die Menge aller Konvexkombinationen aus u und v. Die Menge aller Affinkombinationen erzeugt nicht nur die Strecke, sondern die komplette Gerade durch u und v.

Beispiel: Polynome als linearer Unterraum von $\mathbb{R}^{\mathbb{R}}$.
Die Polynome über $\mathbb{R}$ sind die Elemente des Unterraums von $\mathbb{R}^{\mathbb{R}}$, der aus allen Linearkombinationen der **Monome**
$$x^k \ : \ x \mapsto x^k \text{ für alle } x \in \mathbb{R} \quad \text{für alle } k \geq 0$$
gebildet wird. Sie bilden also einen Vektorraum, den wir mit $\mathbb{P}$ bezeichnen wollen.

Definition 4.10 *Es sei V ein Vektorraum über $\mathbb{K}$ und es sei U eine Teilmenge von V.*

1. *Wenn U zu zu beliebigen endlichen Teilmengen $\{u_1, \ldots, u_m\}$ von U auch alle* **Linear***kombinationen von $u_1, \ldots, u_m$ mit beliebigen Koeffizienten aus $\mathbb{K}$ enthält, ist U ein* **linearer Unterraum** *oder Untervektorraum von V. Das stimmt mit der bisherigen Definition von Untervektorräumen überein.*

2. *Wenn U zu zu beliebigen endlichen Teilmengen $\{u_1, \ldots, u_m\}$ von U auch alle* **Affin***kombinationen von $u_1, \ldots, u_m$ mit beliebigen Koeffizienten aus $\mathbb{K}$ enthält, ist U ein* **affiner Unterraum** *von V.*

3. *Wenn U im Falle $\mathbb{K} = \mathbb{R}$ zu beliebigen endlichen Teilmengen $\{u_1, \ldots, u_m\}$ von U auch alle* **Konvex***kombinationen von $u_1, \ldots, u_m$ mit beliebigen Koeffizienten aus $\mathbb{R}$ enthält, ist U eine* **konvexe Teilmenge** *von V.*

Man kann leicht zeigen, daß es für die obige Definition auch gereicht hätte, sich auf $m = 2$ zu beschränken. Ferner sind lineare Unterräume immer auch affine Unterräume, und affine Unterräume sind immer auch konvexe Mengen.

Die geometrische Interpretation von konvexen Mengen besagt:

> Eine Teilmenge M eines Vektorraums V über $\mathbb{R}$ ist genau dann konvex, wenn sie zu je zwei beliebigen Punkten $u, v \in M$ auch die Verbindungsstrecke $[u, v]$ enthält.

Man mache sich das an Beispielen klar (Gerade, Strecke, Dreieck, Kreisscheibe, Ellipse, Kugel).

Für eine spätere Anwendung brauchen wir noch

Theorem 4.11 *Eine Teilmenge von $\mathbb{R}$ ist genau dann konvex, senn sie ein Intervall ist.*

Beweis Die Definition von beschränkten Intervallen findet sich in (3.23) auf Seite 96, und etwas später werden die unbeschränkten Intervalle definiert. Klar ist, daß jedes Intervall konvex ist. Zu einer konvexen Teilmenge K von $\mathbb{R}$ kann man das Intervall

$$I := (\inf_{x \in K} x, \sup_{x \in K} x)$$

bilden, und es folgt $I \subseteq K$ wegen der Konvexität. Falls die "Endpunkte" von I endlich sind und zu K gehören, kann man an den entsprechenden Stellen die runden Klammern in der obigen Definition durch eckige ersetzen, und damit folgt insgesamt die Behauptung. □

Theorem 4.12 *Es sei V ein Vektorraum über einem Skalarenkörper $\mathbb{K}$. Dann gilt:*

1. *Der mengentheoretische Durchschnitt von linearen Unterräumen ist der Nullraum $\{0\}$ oder ein linearer Unterraum.*

2. *Der mengentheoretische Durchschnitt von affinen Unterräumen ist leer oder ein affiner Unterraum.*

3. *Der mengentheoretische Durchschnitt von konvexen Teilmengen ist leer oder eine konvexe Teilmenge. Hierzu muß $\mathbb{K} = \mathbb{R}$ vorausgesetzt werden, damit Konvexität einen Sinn macht.*

Definition 4.13 *Es sei V ein Vektorraum über $\mathbb{R}$ und es sei U eine nichtleere Teilmenge von V. Die lineare bzw. affine bzw. konvexe* **Hülle** *von U ist der Durchschnitt aller linearen bzw. affinen Unterräume bzw. konvexen Teilmengen von V, die U enthalten. Im linearen und affinen Fall kann man allgemeine Grundkörper $\mathbb{K}$ nehmen.*

Jetzt sei eine endliche Teilmenge $U := \{u_1, \ldots, u_m\}$ von V **fest** *vorgegeben. Dann gilt:*

1. *Die Menge aller* **Linear***kombinationen*[1] *von $u_1, \ldots, u_m$ bildet einen* **linearen** *Unterraum von V, den man als*

$$span\ \{u_1, \ldots, u_m\}$$

und als den von $u_1, \ldots, u_m$ **aufgespannten** *Untervektorraum bezeichnet. Er ist die lineare Hülle*[2] *von $\{u_1, \ldots, u_m\}$.*

2. *Die Menge aller* **Affin***kombinationen von $u_1, \ldots, u_m$ bildet einen* **affinen** *Unterrraum von V. Er ist die* **affine** *Hülle von $\{u_1, \ldots, u_m\}$.*

3. *Die Menge aller* **Konvex***kombinationen von $u_1, \ldots, u_m$ bildet den von $u_1, \ldots, u_m$* **aufgespannten** *konvexen* **Simplex** *in V. Er ist die* **konvexe** *Hülle*[3] *von $\{u_1, \ldots, u_m\}$.*

Aufgabe: Man nehme im $\mathbb{R}^2$ die drei Punkte $u_1 = (0,0), u_2 = (1,0)$ und $u_3 = (0,1)$. Dann bestimme man die lineare, affine und konvexe Hülle von $\{u_1, \ldots, u_m\}$ für $m = 1, 2, 3$.

Aufgabe: Man nehme im $\mathbb{R}^2$ die drei Punkte $u_1 = (0,0), u_2 = (1,1)$ und $u_3 = (2,2)$. Dann bestimme man die lineare, affine und konvexe Hülle von $\{u_1, \ldots, u_m\}$ für $m = 1, 2, 3$.

Aufgabe: Man zeichne sich im $\mathbb{R}^2$ ein paar Punkte ein und bilde deren konvexe Hülle. Warum kann man das Ergebnis die "Gummibandkonstruktion" nennen? Welche Punkte sind überflüssig?

[1] `http://de.wikipedia.org/wiki/Linearkombination`
[2] `http://de.wikipedia.org/wiki/Lineare_H%C3%BClle`
[3] `http://de.wikipedia.org/wiki/Konvexe_H%C3%BClle`

Definition 4.14 *Sind U und V Vektorräume über demselben Skalarenkörper $\mathbb{K}$, so ist $U \times V$ wieder ein Vektorraum über $\mathbb{K}$. Addition und Skalarmultiplikation sind auf naheliegende Weise definiert.*

Es mag Informatiker wundern, warum dies alles in solcher Allgemeinheit entwickelt wird. Der Hintergrund ist (unter anderem), daß moderne Graphikkarten auf **Polygonen**[1] arbeiten. Das sind von m Punkten $u_1, \ldots, u_m$ des $\mathbb{R}^3$ aufgespannte Simplices (Mehrzahl von Simplex), die in je einer Ebene liegen, und Ebenen sind zweidimensionale affine Unterräume des $\mathbb{R}^3$. Der einfachste Fall liegt für $m = 3$ vor, und man bekommt Dreiecke im dreidimensionalen Raum, wenn die drei gegebenen Punkte nicht auf einer Geraden liegen. Dreiecke und Strecken sind die konvexen Hüllen ihrer Ecken bzw. Endpunkte.

Im **Computer–Aided Design**[2] und in der Computergraphik[3] baut man ganze virtuelle Welten aus mathematischen Objekten auf, die man mit Affintransformationen im Raum verschieben und skalieren kann. Mit projektiven Abbildungen bildet man sie auf Bildebenen ab, um sie graphisch darzustellen (**rendering**). Affine bzw. projektive Transformationen gehören in die affine bzw. projektive Geometrie. Aber so weit sind wir noch lange nicht, es bleibt noch viel zu tun.

Aufgabe: Warum ist eine Linearkombination von Vektoren $u_1, \ldots, u_n$ immer auch eine Affinkombination von $0, u_1, \ldots, u_n$?

4.2.2 Darstellungen geometrischer Objekte

Wenn man im Raum gewisse mathematische Objekte (Geraden, Kreise, Ebenen, Kugeln, Kegel, Kurven, und mathematisierte Tische, Stühle, Wände, Gegenstände...) darstellen will, hat man mehrere Möglichkeiten. Man faßt in allen Fällen aber den Raum geometrisch als Raum von Punkten auf und definiert die Objekte als Punktmengen. Die in der Mathematik üblich gewordene Sichtweise von Räumen als Vektorräume mit einer Sonderrolle des Nullpunkts ist für die geometrische Anschauung eher schädlich, denn die Nullpunktslage ist ziemlich irrelevant bei praktischen Fragestellungen. Man arbeitet eher in einem Punktraum mit affiner Geometrie.

Aber man kann die Punktmengen verschieden darstellen:

1. **implizit** als Punkte, die irgendwelchen Bedingungen genügen,

[1] `http://de.wikipedia.org/wiki/Polygon`
[2] `http://de.wikipedia.org/wiki/Computer_Aided_Design`
[3] `http://de.wikipedia.org/wiki/Computergrafik`

2. **explizit** als Bildpunkte von Abbildungen.

Das kann man z.B. am Beispiel einer Geraden im $\mathbb{R}^2$ einfach veranschaulichen:

1. bei impliziter Schreibweise:

$$\{(x_1, x_2) \in \mathbb{R}^2 \ : \ \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \alpha_3 \cdot 1 = 0\}$$

   wobei α_1 und α_1 nicht beide Null sind,

2. bei expliziter Schreibweise als Bild der Abbiildung $f \ : \ \mathbb{R} \to \mathbb{R}^2$ mit $f(t) := u + t(v - u)$ wobei u und v im $\mathbb{R}^2$ zwei verschiedene Punkte sind:

$$f(\mathbb{R}) = \{u + t(v - u) \in \mathbb{R}^2 \ : \ t \in \mathbb{R}\}.$$

Ein weiteres typisches Beispiel ist der Einheitskreis im $\mathbb{R}^2$ als

$$\begin{array}{rcl} E & := & \{(x_1, x_2) \ : \ x_1^2 + x_2^2 = 1\} \\ & = & f([0, 2\pi)) \text{ mit } f(t) := (\cos(t), \sin(t)) \end{array}$$

in impliziter und expliziter Schreibweise. Die implizite Form definiert das Objekt als "**geometrischer Ort**" mit einer Bedingung, die normalerweise Gleichungs– oder Ungleichungsform hat, während die explizite Form eine konkrete Rechenvorschrift angibt, mit der man Punkte des Objekts angeben kann. Die implizite Form erlaubt einen schnellen Test, ob ein beliebiger Punkt zum Objekt gehört, kann aber nicht ohne weiteres alle Punkte des Objekts konstruktiv produzieren.

Beide Formen haben also ihre Vor–und Nachteile. Beim **Ray–Tracing**[1] hat man Schnitte von Sehstrahlen mit Objekten zu berechnen, und dann sind implizite Darstellungen besser. Beim **Computer–Aided Design**[2] hat man konkrete Objekte zu produzieren, und deshalb verwendet man explizite Darstellungen. Beim Modellieren dreidimensionaler massiver Körper gibt es beide Varianten:

1. beim **boundary–representation–modelling** stellt man einen Körper durch explizite Darstellung seiner Begrenzungsflächen dar, während man

[1] `http://de.wikipedia.org/wiki/Raytracing`
[2] `http://de.wikipedia.org/wiki/Computer_Aided_Design`

2. bei impliziten Darstellungen von Körpern (durch Bedingungen) die Berechnung von Schnitten und Vereinigungen leicht durch Boolesche Funktionen auf den Bedingungen realisieren kann. Beispiel: eine Bohrung durch einen Körper ist durch mengentheoretische Differenz zwischen dem Körper k und dem zylindrischen Bohrkern b darstellbar. Man nimmt die Punkte, die in k und nicht in b liegen und modelliert das mit einer passenden Booleschen Funktion.

Die Transformation von impliziten zu expliziten Darstellungen und umgekehrt ist ziemlich schwierig, wie man schon am Beispiel des Kreises sehen kann.

Für das Folgende sollte festgehalten werden, daß geometrische Objekte stets Punktmengen in Vektorräumen sind, wobei der Vektoraspekt der Punkte als Vektoren zwischen Punkt und Nullpunkt im allgemeinen irrelevant ist. Computer können nur mit einigem Aufwand rein geometrisch arbeiten (d.h. mit den Begriffen "Punkt liegt auf Gerade, Gerade schneidet Kugel" usw.). Man verwendet Funktionen oder (Un–)Gleichungsbedingungen und benutzt Vektorräume vom Typ $\mathbb{R}^n$, wobei man jede Komponente durch Gleitkommazahlen darstellt, d.h. man arbeitet in $(\texttt{double})^n$.

4.3 Lineare, affine und konvexe Abbildungen

4.3.1 Linear– und andere Kombinationen

In der Computergraphik und beim Computer–Aided Design muß man geometrische Objekte verschieben, verkleinern, drehen und auf Bildebenen abbilden können. Das leisten Abbildungen zwischen Vektorräumen:

Definition 4.15 *Es seien U und V Vektorräume über einem gemeinsamen Skalarenkörper $\mathbb{K}$ und es sei $T : U \to V$ eine Abbildung von U nach V.*

1. *T ist eine* **lineare Abbildung**[1]*, wenn für beliebige* **Linear***kombinationen gilt*

$$T\underbrace{\left(\sum_{j=1}^{n}\alpha_j u_j\right)}_{\in U} = \sum_{j=1}^{n}\alpha_j T(u_j) \in V.$$

2. *T ist eine* **affine Abbildung**[2]*, wenn die obige Gleichung für beliebige* **Affin***kombinationen gilt.*

[1] `http://de.wikipedia.org/wiki/Lineare_Abbildung`
[2] `http://de.wikipedia.org/wiki/Affine_Abbildung`

3. *Ist V angeordnet unter $\leq$, und gilt $\mathbb{K} = \mathbb{R}$, so ist T eine* **konvexe** *Abbildung, wenn für beliebige* **Konvex***kombinationen gilt*

$$T\underbrace{\left(\sum_{j=1}^{n}\alpha_j u_j\right)}_{\in U} \leq \sum_{j=1}^{n}\alpha_j T\left(u_j\right) \in V.$$

4. *Eine lineare Abbildung mit Werten in $\mathbb{K}$ heißt lineares* **Funktional**.

Lineare Abbildungen sind immer affin, und affine Abbildungen mit Werten in einem geordneten Vektorraum V sind immer konvex. Erwartungsgemäß gilt

Theorem 4.16 *Unter einer linearen bzw. affinen Abbildung ist das Bild eines linearen bzw. affinen Unterraums wieder ein linearer oder affiner Unterraum.*

Sehen wir uns kurz die besonders einfachen "konstanten" Abbildungen $T(u) := v$ für alle $u \in U$ mit festem $v \in V$ an. Wenn so eine Abbildung linear sein soll, muß $T(0) = 0$ gelten, d.h. die Nullabbildung ist die einzige konstante lineare Abbildung. Aber man sieht sofort, daß jede konstante Abbildung affin ist, weil man

$$T\underbrace{\left(\sum_{j=1}^{n}\alpha_j u_j\right)}_{\in U} = v = 1\cdot v = \underbrace{\left(\sum_{j=1}^{n}\alpha_j\right)}_{=1}\cdot v = \sum_{j=1}^{n}(\alpha_j\cdot v) = \sum_{j=1}^{n}\alpha_j T(u_j)$$

für jede Affinkombination hat.

Die Identität Id_U als Abbildung $U \to U$ ist immer linear und affin. Wenn man eine Punktmenge $P \subseteq U$ im Raum um einen festen Vektor $v \in U$ verschieben will, wendet man die Abbildung $u \mapsto u + v$ an. Das ist als Summe der Identität mit einer konstanten Abbildung interpretierbar. Und wenn man, vom Nullpunkt her gesehen, eine Punktmenge $P \subseteq U$ auf das Doppelte "aufblasen" will, wird man die Abbildung $u \mapsto 2\cdot u$ verwenden, und das kann man als eine neue Abbildung $2\cdot Id_U$ sehen. Man sieht also, daß man Abbildungen addieren und mit Skalaren multiplizieren kann. Allgemeiner:

Theorem 4.17 *Sind U und V Vektorräume über einem gemeinsamen Skalarenkörper $\mathbb{K}$, so bilden die linearen und die affinen Abbildungen von U*

in V jeweils einen Vektorraum $Lin(U,V)$ bzw. $Aff(U,V)$ über $\mathbb{K}$ mit den Operationen

$$\begin{aligned}(S+T)(u) &:= S(u)+T(u) \text{ für alle } u \in U\\ (\alpha \cdot T)(u) &:= \alpha \cdot T(u) \text{ für alle } u \in U\end{aligned}$$

für beliebige Abbildungen S, T und Skalare $\alpha \in \mathbb{K}$.

Das ist sehr einfach zu beweisen, und deshalb lassen wir den Beweis weg.

Lineare und affine Abbildungen unterscheiden sich nur um die konstanten Abbildungen. Das kann man zur Definition affiner Abbildungen[1] machen, aber die Definition 4.15 ist besser, weil sie über Invarianz affiner Beziehungen erfolgt und nicht über eine Schreibweise oder eine Formel.

Theorem 4.18 *1. Ist S eine lineare Abbildung von U nach V und ist $v \in V$ beliebig, so ist die Abbildung $T : u \mapsto S(u)+v$ affin.*

2. Ist T eine affine Abbildung von U nach V, so ist die Abbildung $S : u \mapsto T(u)-T(0)$ linear.

Der erste Teil ist klar, weil S und alle konstanten Abbildungen affin sind, damit auch die Summe.

Zum Beweis des zweiten Teils nehmen wir eine beliebige Linearkombination

$$u := \sum_{j=1}^{n} \alpha_j u_j \in U$$

von Vektoren $u_1, \ldots, u_n \in U$ her und schließen auf

$$\begin{aligned}S(u) &= T(u)-T(0)\\ &= T\left(\sum_{j=1}^{n}\alpha_j u_j\right)-T(0)\\ &= T\left(\sum_{j=1}^{n}\alpha_j u_j+\left(1-\sum_{j=1}^{n}\alpha_j\right)\cdot 0\right)-T(0)\\ &= \sum_{j=1}^{n}\alpha_j T(u_j)+\left(1-\sum_{j=1}^{n}\alpha_j\right)T(0)-T(0)\\ &= \sum_{j=1}^{n}\alpha_j\left(T(u_j)-T(0)\right)\\ &= \sum_{j=1}^{n}\alpha_j S(u_j).\end{aligned}$$

[1] `http://de.wikipedia.org/wiki/Affine_Abbildung`

□

Später werden wir uns noch für spezielle Abbildungen, z.B. Drehungen und Spiegelungen interessieren, aber das erfordert zusätzliche Annahmen.

Aufgabe: Es sei $\mathbb{P}$ der Untervektorraum von $\mathbb{R}^{\mathbb{R}}$, der aus allen Polynomen besteht. Dann ist die Differentiation

$$p \mapsto p'$$

eine lineare Abbildung. Warum?

Aufgabe: Warum ist

$$p \mapsto \int_0^1 p(t)dt \in \mathbb{R}$$

auf $\mathbb{P}$ ein lineares Funktional?

Diese beiden Beispiele aus der Schulmathematik zeigen, daß es auch außerhalb geometrischer Konstruktionen sinnvolle lineare Abbildungen gibt.

Definition 4.19 *Sind U und V lineare Unterräume eines Vektorraums W, so ist*

$$U + V := \{u + v \ : \ u \in U, \ v \in V\} \subseteq W$$

ein linearer Unterraum von W. Die Vektorraumsumme $U + V$ ist eine **direkte Summe***, wenn $U \cap V = \{0\}$ gilt.*

Diese Definition enthielt illegalerweise eine Behauptung, aber diese ist leicht zu beweisen, indem man den Vektorraum $U \times V$ hernimmt (siehe Definition 4.14 auf Seite 111) und darauf die lineare Abbildung $(u, v) \mapsto u + v$ definiert. Das Bild ist nach Theorem 4.16 ein linearer Unterraum, und zwar gleich $U + V$. Wir fügen noch zwei einleuchtende und einfach beweisbare Fakten an:

Theorem 4.20 *Sind $S \ : U \to V$ und $T \ : \ V \to W$ lineare bzw. affine Abbildungen zwischen Vektorräumen über demselben Grundkörper $\mathbb{K}$, so ist auch $T \circ S \ : \ U \to W$ eine lineare bzw. affine Abbildung.*

Theorem 4.21 *Ist $S \ : U \to V$ eine bijektive lineare bzw. affine Abbildung zwischen Vektorräumen über demselben Grundkörper $\mathbb{K}$, so ist auch S^{-1} eine lineare bzw. affine Abbildung.*

Man nehme eine beliebige Linearkombination von Vektoren v_j aus dem Bildraum und schreibe diese Vektoren als Bilder $v_j = S(u_j)$ mit Vektoren u_j aus dem Urbildraum. Es folgt

$$\begin{aligned} S^{-1}\left(\sum_{j=1}^{n} \alpha_j \underbrace{v_j}_{=S(u_j)}\right) &= S^{-1}\left(\sum_{j=1}^{n} \alpha_j S(u_j)\right) \\ &= S^{-1}S\left(\sum_{j=1}^{n} \alpha_j u_j\right) \\ &= \sum_{j=1}^{n} \alpha_j u_j \\ &= \sum_{j=1}^{n} \alpha_j S^{-1}(v_j). \end{aligned}$$

□

Definition 4.22 *Man bezeichnet eine lineare Abbildung zwischen Vektorräumen auch als* **Vektorraumhomomorphismus**. *Ist die Abbildung bijektiv, so ist sie ein* **Vektorraumisomorphismus**. *Zwei Vektorräume heißen* **isomorph**, *wenn zwischen ihnen ein Vektorraumisomorphismus definiert ist.*

Natürlich ist Isomorphie eine Äquivalenzrelation auf der Menge der Vektorräume über gleichem Grundkörper $\mathbb{K}$. Wir schreiben die **Isomorphie** zwischen Vektorräumen U und V als binäre Relation $U \simeq V$.

In der Mathematik verwendet man den Begriff **Homomorphismus** oder **Morphismus** allgemein für Abbildungen mit strukturerhaltenden Eigenschaften. Hier geht es um die Vektorraumstruktur, und deshalb spricht man von Vektorraumhomomorphismen. Die "Morphismen"–Sprechweise hat sich in der theorieorientierten Mathematik durchgesetzt, in den technisch–naturwissenschaftlichen Anwendungen bisher nicht. Informatiker wissen vielleicht, was "*morphing*[1]" in der Bildverarbeitung ist, und wundern sich deshalb nicht.

4.3.2 Dualraum

Der wichtigste Fall von Vektorräumen linearer Abbildungen ergibt sich in

[1] `http://de.wikipedia.org/wiki/Morphing`

Definition 4.23 *Ist U ein Vektorraum über einem Grundkörper $\mathbb{K}$, so wird $Lin(U, \mathbb{K})$ als (algebraischer)* **Dualraum**[1] *von U bezeichnet. Er besteht aus den linearen Funktionalen auf U mit Werten in $\mathbb{K}$. Er wird in der Mathematik oft als U' oder U^* bezeichnet, und ist vom später zu definierenden topologischen Dualraum zu unterscheiden.*

Im Raum $\mathbb{R}^{\mathbb{R}}$ ist für jedes feste $x \in \mathbb{R}$ die Abbildung

$$\delta_x \; : \; f \mapsto f(x) \text{ für alle } f \in \mathbb{R}^{\mathbb{R}}$$

eine lineare Abbildung mit Werten in $\mathbb{R}$. Dieses **Auswertungsfunktional** wird von Physikern auch als **Deltafunktion** bezeichnet, ist aber keine Funktion, sondern ein Funktional. Natürlich kann man δ_x auch auf $M^{\mathbb{K}}$ bei allgemeinem Grundkörper $\mathbb{K}$ und für alle x aus einer allgemeinen Menge M definieren.

Auf dem Polynomraum $\mathbb{P}$ gibt es Funktionale wie

$$p \mapsto p'(x) \text{ für alle } p \in \mathbb{P} \tag{4.24}$$

für jedes feste x, wobei p' die Ableitung von p ist. Oder man bildet

$$p \mapsto \int_0^1 p(t)dt \text{ für alle } p \in \mathbb{P}$$

als lineares Funktional. Differentiation in einem Punkt und Integration über ein festes Intervall sind lineare Funktionale auf dem Polynomraum, die von eminenter praktischer Bedeutung sind.

Aber das ist noch nicht alles. Zu jeder Gleichung der Form

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n \;\; = \;\; c \tag{4.25}$$

für n reelle Unbekannte $x_1, \ldots, x_n$ kann man die lineare Abbildung

$$(x_1, \ldots, x_n) \mapsto a_1x_1 + a_2x_2 + \ldots + a_nx_n \tag{4.26}$$

definieren, und sie ist ein lineares Funktional auf dem $\mathbb{R}^n$. Die Gleichungen entsprechen also linearen Funktionalen.

Das kann auch in sehr viel allgemeinerem Rahmen auftreten, nämlich dann, wenn man ein Element u aus einem abstrakten Vektorraum U sucht, so daß

[1] `http://de.wikipedia.org/wiki/Dualraum`

n Bedingungen

$$\begin{aligned} \lambda_1(u) &= c_1 \\ \lambda_2(u) &= c_2 \\ \vdots &\quad \vdots \\ \lambda_n(u) &= c_n \end{aligned}$$

mit festen linearen Funktionalen aus $Lin(U, \mathbb{K})$ und Skalaren $c_1, \ldots, c_n \in \mathbb{K}$ erfüllt sind. Viele Probleme des Wissenschaftlichen Rechnens sind von dieser Form, sogar mit unendlich vielen Funktionalen.

Definition 4.27 *Ist $S : U \to V$ eine lineare Abbildung zwischen Vektorräumen über einem gemeinsamen Grundkörper $\mathbb{K}$, so ist*

$$S^d : V^* \to U^*,\ (S^d(v^*))(u) := v^*(S(u)) \text{ für alle } u \in U$$

die **duale Abbildung** *zu S.*

Das klingt etwas gewaltsam, ist aber im obigen Beispiel schon zur Anwendung gekommen. Die Differentiation

$$D : \mathbb{P} \to \mathbb{P},\ p \mapsto p' \text{ für alle } p \in \mathbb{P}$$

ist eine lineare Abbildung, und das Funktional aus (4.24) ist dann

$$((D^d)(\delta_x))(p) = \delta_x(D(p)) = p'(x).$$

Die duale Abbildung $S : U \to V$ tritt auch auf, wenn man ohne Verluste eine in U mit einem festen Funktional $\lambda \in U^*$ geltende Gleichung $\lambda(u) = c$ so transformieren will, daß man eine äquivalente Gleichung für $v := S(u)$ bekommt. Man muß ein Funktional $\mu \in V^*$ finden mit $\lambda = S^d(\mu)$ und bekommt mit

$$\begin{aligned} \lambda(u) &= S^d(\mu)(u) \\ &= \mu(S(u)) \\ &= \mu(v) \end{aligned}$$

die Äquivalenz der Gleichungen $\lambda(u) = c$ und $\mu(v) = c$ für $v = S(u)$. Kurz: Transformiert man die Unbekannten mit S, so muß man die Gleichungen mit S^d transformieren.

Die duale Abbildung zu S kann leicht mit der Adjungierten verwechselt werden, die oft als S' oder S^* geschrieben wird. Wir brauchen diese Bezeichnungen anderweitig und schreiben deshalb S^d.

Als kleine Aufwärmaufgabe formulieren wir

Theorem 4.28 *Ist $S : U \to V$ eine bijektive lineare Abbildung zwischen zwei Vektorräumen über demselben Skalarenkörper $\mathbb{K}$, so gilt*

$$(S^{-1})^d = (S^d)^{-1}.$$

Beweis: Man mache sich zuerst klar, daß man

$$(S^{-1})^d \circ S^d = Id_{V^*}$$

beweisen sollte. (Frage: Warum?)
Nimmt man dann ein beliebiges Funktional $\lambda \in V^*$ und ein beliebiges $v \in V$, so folgt

$$\begin{aligned}(((S^{-1})^d \circ S^d)(\lambda))(v) &= ((S^{-1})^d(S^d(\lambda)))(v)\\ &= S^d(\lambda)(S^{-1}(v))\\ &= \lambda(S(S^{-1}(v)))\\ &= \lambda((S \circ S^{-1})(v))\\ &= \lambda(v)\end{aligned}$$

und das ist die Behauptung. □

4.4 Matrizen

In diesem Abschnitt geht es darum, lineare Abbildungen computergerecht zurechtzuschneiden. In der Praxis betreibt man nämlich fast die gesamte lineare Algebra durch Manipulationen an sogenannten **Matrizen**. Es gibt dafür sogar eine spezielle Sprache MATLAB©[1], die wir uns noch ansehen werden. Aber erst einmal müssen wir auf die Matrixdarstellung linearer Abbildungen hinarbeiten.

4.4.1 Erzeugendensysteme

Im Vektorraum $\mathbb{K}^n$ haben alle n–Tupel die Form $x = (\xi_1, \ldots, \xi_n)$ mit Skalaren $\xi_j \in \mathbb{K}$, aber so eine schöne und einfache Darstellung gibt es in allgemeinen Vektorräumen nicht. Aus Gründen, die erst später klar werden, schreibt man die Vektoren des $\mathbb{K}^n$ als **Spaltenvektoren**

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}$$

[1] `http://de.wikipedia.org/wiki/MATLAB`

und definiert die **Einheitsvektoren** e_j für $1 \leq j \leq n$ so, daß sie an der j–ten Stelle eine Eins und sonst Nullen haben:

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \ldots, e_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Dann kann man jedes $x \in I\!K^n$ als Linearkombination

$$x = \sum_{j=1}^{n} \xi_j e_j = \xi_1 \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \xi_2 \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \ldots + \xi_n \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

schreiben. Man sieht, daß jeder Vektor eine Linearkombination der Einheitsvektoren ist, und die skalaren Koeffizienten sind eindeutig bestimmt.

Definition 4.29 *Es sei V ein Vektorraum über $I\!K$.*

1. *Eine Teilmenge M von V heißt* **Erzeugendensystem** *von V, wenn jeder Vektor $x \in V$ eine Linearkombination aus endlich vielen Elementen von M ist.*

2. *Ein mengentheoretisch minimales Erzeugendensystem heißt* **Basis**.

Man kann sich im $I\!K^n$ klarmachen, daß es sehr viele Erzeugendensysteme gibt, z.B. die Einheitsvektoren, aber alle Basen haben nur n Elemente, wie sich noch zeigen wird. Etwas schneller werden wir sehen, daß die Einheitsvektoren eine Basis bilden, und wenn man dann weitere Vektoren hinzufügt, bekommt man Erzeugendensysteme.

Aufgabe: Man mache sich klar, daß im $I\!R^2$ auch die Vektoren $(1, 1)$ und $(1, -1)$ ein Erzeugendensystem bilden. Dazu verfertige man eine Zeichnung und sehe sich das neue Koordinatensystem an, in dem diese Vektoren die Rolle der Einheitsvektoren spielen.

Eine alternative Definition von Erzeugendensystemen ergibt sich aus

Theorem 4.30 *Eine Teilmenge M eines Vektorraums V ist genau dann Erzeugendensystem von V, wenn V die lineare Hülle von M ist.*

Beweis: Sei M ein Erzeugendensystem von V. Wir wollen zeigen, daß V die lineare Hülle von M ist. Sei U ein Untervektorraum von V, der M enthält. Wir sind fertig, wenn V in U liegt bzw. $V = U$ gilt. Ein beliebiges Element $v \in V$ ist aber darstellbar duch eine Linearkombination von Vektoren aus M, weil M ein Erzeugendensystem von V ist, und liegt in U, weil U ein Untervektorraum ist, der M enthält. Also ist V gleich U.

Sei nun V die lineare Hülle von M. Wir wollen zeigen, daß M ein Erzeugendensystem von V ist. Es sei U der Unterraum von V, der aus allen endlichen Linearkombinationen von Vektoren aus M besteht. Dieser Unterraum liegt in V und enthält M, also liegt die lineare Hülle von M in diesem Unterraum, d.h. es folgt $V \subseteq U \subseteq V$ und $U = V$. □

Beispiel: Nach Definition des Polynom–Vektorraums $\mathbb{P}$ bilden die abzählbar vielen **Monome**

$$\mathbf{x}^k \ : \ x \mapsto x^k, \ k \in \mathbb{N}$$

ein Erzeugendensystem des Polynomraums.

4.4.2 Matrixdarstellung linearer Abbildungen

Wir untersuchen nun, wie sich lineare oder affine Abbildungen $S \ : \ U \to V$ zwischen Vektorräumen U und V mit einem gemeinsamen Skalarenkörper $\mathbb{K}$ schreiben lassen, wenn U und V Erzeugendensysteme $\{u_1, \ldots, u_n\} \subset U$ bzw. $\{v_1, \ldots, v_m\} \subset V$ haben. Weil die affinen Abbildungen sich nur um konstante Vektoren von den linearen Abbildungen unterscheiden, beschränken wir uns auf lineare Abbildungen.

Unter den obigen Voraussetzungen bildet die lineare Abbildung $S \ : \ U \to V$ die Vektoren $u_k \in U$ in Vektoren $S(u_k) \in V$ ab, die sich im Erzeugendensystem $\{v_1, \ldots, v_m\} \subset V$ darstellen lassen müssen. Es gibt dann zu jedem $k, \ 1 \leq k \leq n$ jeweils m Koeffizienten $\alpha_{1k}, \ldots, \alpha_{mk}$ in $\mathbb{K}$ mit

$$S(u_k) = \sum_{j=1}^{m} \alpha_{jk} v_j, \ 1 \leq k \leq n.$$

Die Abbildung wirkt dann auf beliebige Vektoren $u \in U$ so:

$$\begin{aligned} \text{Wenn } u &= \sum_{k=1}^{n} \xi_k u_k, \text{ so gilt} \\ S(u) &= \sum_{k=1}^{n} \xi_k S(u_k) \\ &= \sum_{k=1}^{n} \xi_k \sum_{j=1}^{m} \alpha_{jk} v_j \\ &= \sum_{j=1}^{m} \left(\sum_{k=1}^{n} \alpha_{jk} \xi_k \right) v_j. \end{aligned} \tag{4.31}$$

Man kann nun von den Erzeugendensystemen und von S absehen und ganz allgemein die Abbildung $M_S : \mathbb{K}^n \to \mathbb{K}^m$ mit

$$(\xi_1, \ldots, \xi_n) \mapsto \left(\sum_{k=1}^{n} \alpha_{1k} \xi_k, \ldots, \sum_{k=1}^{n} \alpha_{mk} \xi_k \right) \in \mathbb{K}^m \tag{4.32}$$

betrachten. Sie bildet den $\mathbb{K}^n$ in den $\mathbb{K}^m$ ab, und sie beschreibt, wie die Wirkung der allgemeinen linearen Abbildung $S : U \to V$ sich in den Erzeugendensystemen von U und V ausdrücken läßt. Man mache sich klar, das S und M_S **verschiedene** Abbildungen sind, obwohl sie natürlich ganz eng zusammenhängen. Man kann M_S eine Darstellung von S durch die Erzeugendensysteme von U und V nennen, und diese Darstellung hängt von S und den beiden Erzeugendensystemen ab. Klar ist auch, daß man in Computern mit einer allgemeinen linearen Abbildung S nicht gut arbeiten kann, wohl aber mit der obigen Darstellung, wie wir gleich sehen werden.

Man faßt die Koeffizienten α_{jk} in ein rechteckiges Schema (**Matrix**[1], Mehrzahl: Matrizen) zusammen und "multipliziert" diese Matrix mit dem als Spaltenvektor geschriebenen Vektor der Koeffizienten ξ_k. Das sieht dann so aus:

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} \end{pmatrix} \cdot \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} \alpha_{1k} \xi_k \\ \sum_{k=1}^{n} \alpha_{2k} \xi_k \\ \vdots \\ \sum_{k=1}^{n} \alpha_{mk} \xi_k \end{pmatrix} \tag{4.33}$$

Die horizonalen Einträge von je n Elementen der Matrix heißen **Zeilen**, die vertikalen aus m Elementen heißen **Spalten**. Das Ganze wird dann eine

[1] `http://de.wikipedia.org/wiki/Matrix_(Mathematik)`

$m \times n$–Matrix genannt, und man benutzt die Kurzschreibweise

$$\begin{aligned} M_S &= (\alpha_{jk})_{\substack{1 \le j \le m \\ 1 \le k \le n}} \in \mathbb{K}^{m \times n} \\ x &= (\xi_k)_{1 \le k \le n} \in \mathbb{K}^n \end{aligned}$$

$$M_S \cdot x := \left(\sum_{k=1}^{n} \alpha_{jk} \xi_k \right)_{1 \le j \le m} \in \mathbb{K}^m.$$

Üblicherweise wird der **Zeilenindex** zuerst genannt, dann der **Spaltenindex** (hier j bzw. k). Wir identifizieren hier die Notation für die Abbildung M_S mit der Notation für die Matrix. Das ist erlaubt, wenn man lineare Abbildungen wie M_S von $\mathbb{K}^n$ nach $\mathbb{K}^m$ hat. Dann sind die Erzeugendensysteme in Bild– und Urbildraum durch die Einheitsvektoren fest gegeben, und die Rechenvorschrift ist immer von der Form (4.32).

An dieser Stelle wird klar, daß man verbindlich festlegen muß, ob Elemente des $\mathbb{K}^n$ als Spalten– oder Zeilenvektoren geschrieben werden. Obwohl wir bei allgemeinen n–fachen cartesischen Produkten von **Mengen** die Notation der n–Tupel eingeführt haben, was den Zeilenvektoren entspricht, wollen wir ab jetzt die Elemente der **Vektorräume** $\mathbb{R}^n$ oder $\mathbb{C}^n$ oder $\mathbb{K}^n$ immer als Spaltenvektoren verstehen. Wir identifizieren also $\mathbb{K}^{n \times 1}$ mit $\mathbb{K}^n$.

Fassen wir zusammen:

Theorem 4.34 *Zu jeder linearen Abbildung $S : U \to V$ zwischen Vektorräumen über demselben Grundkörper $\mathbb{K}$ gibt es eine Darstellung durch eine Matrix, wenn U und V endliche Erzeugendensysteme haben. Wählt man Erzeugendensysteme*

$$\begin{aligned} u_1, \ldots, u_n \quad &\textit{für} \quad U \\ v_1, \ldots, v_m \quad &\textit{für} \quad V, \end{aligned}$$

so kann die Wirkung von S durch (4.31) *beschrieben und berechnet werden, wobei die Matrix–Vektor–Multiplikation* (4.33) *anzuwenden ist.*

Zwischen Räumen der Form $\mathbb{K}^n$ kann man das schärfer fassen:

Theorem 4.35 *Zu jeder linearen Abbildung $S : \mathbb{K}^n \to \mathbb{K}^m$ gibt es genau eine $m \times n$–Matrix M_S in $\mathbb{K}^{m \times n}$, die S im obigen Sinne darstellt, wenn man die Einheitsvektoren als Erzeugendensysteme wählt. Die Aktion von S auf Vektoren aus $\mathbb{K}^m$ wird durch die Matrix–Vektor–Multiplikation*

$$S(x) := M_S \cdot x$$

beschrieben, wobei im Urbild– und Bildraum die Darstellung von Vektoren durch die Einheitsvektoren unterstellt wird.

Bei linearen Abbildungen zwischen beliebigen Vektorräumen ist die Matrixdarstellung nicht eindeutig, denn sie hängt entscheidend von der Wahl der Erzeugendensysteme ab.

Oft werden Matrizen und lineare Abbildungen verwechselt. Erstens sind lineare Abbildungen sehr viel allgemeiner definiert. Zweitens liefern Matrizen nur spezielle Darstellungsformen für die Wirkung gewisser linearer Abbildungen, wenn man endliche Erzeugendensysteme wählen kann, und deshalb kann ein und dieselbe lineare Abbildung viele verschiedene Matrixdarstellungen haben. Umgekehrt: hat man eine $m \times n$ Matrix über $\mathbb{K}$, so hat man auch eine lineare Abbildung $\mathbb{K}^m \to \mathbb{K}^n$, wenn man die Einheitsvektoren als Erzeugende nimmt.

4.4.3 Operationen auf Matrizen

Wir wissen aus Theorem 4.35, daß $m \times n$–Matrizen alle linearen Abbildungen $\mathbb{K}^m \to \mathbb{K}^n$ liefern, wenn man sich darauf festlegt, die Einheitsvektoren als Erzeugende zu nehmen. Insbesondere ist klar, daß eine Matrix $A \in \mathbb{K}^{m\times n}$ über das Matrix–Vektor–Produkt durch

$$x \mapsto A \cdot x \text{ für alle } x \in \mathbb{R}^n \tag{4.36}$$

eine lineare Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^m$ definiert. Die letztere hat A als Darstellung im Erzeugendensystem der Einheitsvektoren.

Aus Theorem 4.17 folgt aber auch, daß dann die $m \times n$–Matrizen mit Koeffizienten in $\mathbb{K}$ einen Vektorraum über $\mathbb{K}$ bilden. Die Addition zweier Matrizen $A = (\alpha_{jk})$ und $B = (\beta_{jk})$ erfolgt komponentenweise:

$$A + B := (\alpha_{jk} + \beta_{jk}),$$

und die Skalarmultiplikation ebenfalls:

$$\beta \cdot A := (\beta \cdot \alpha_{jk}).$$

Das ist nichts Neues, denn diese Operationen entsprechen gerade denen im Vektorraum $Lin(\mathbb{K}^m, \mathbb{K}^n)$ über $\mathbb{K}$.

Die neuartige **Matrix–Vektor–Multiplikation** in (4.33) ist aber etwas Anderes. Man macht man sie sich am besten so klar, daß man sich in (4.33)

nur jeweils eine Zeile ansieht. In der j–ten Zeile hat man

$$\begin{aligned}(\alpha_{j1},\alpha_{j2},\ldots,\alpha_{jn})\cdot\begin{pmatrix}\xi_1\\ \xi_2\\ \vdots\\ \xi_n\end{pmatrix} &= \sum_{k=1}^{n}\alpha_{jk}\xi_k \\ &= \alpha_{j1}\cdot\xi_1+\alpha_{j2}\cdot\xi_2+\ldots+\alpha_{jn}\cdot\xi_n\end{aligned} \tag{4.37}$$

auszuführen. Man erkennt, daß man diese "Multiplikation" für beliebige Vektoren des $\mathbb{K}^n$ definieren kann, aber sie führt nicht in den Vektorraum $\mathbb{K}^n$, sondern in den Skalarenkörper $\mathbb{K}$. Man nennt sie das (reelle) **Skalarprodukt** und definiert es als

$$\sum_{j=1}^{n}u_jv_j$$

für alle Vektoren $u,v\in\mathbb{R}^n$ mit Komponenten $u_j,v_j\in\mathbb{R},\ 1\le j\le n$. Wir sehen uns das Skalarprodukt später genauer an, insbesondere weil es in der euklidischen Geometrie des $\mathbb{R}^n$ eine zentrale Rolle spielt.

Wir betrachten nun eine weitere Abbildung $T : V\to W$ mit Werten in einem Vektorraum W mir Erzeugendensystem $\{w_1,\ldots,w_p\}$. Auch diese Abbildung hat eine Matrizendarstellung M_T mit einer $p\times m$–Matrix $M_T=(\beta_{\ell j})$, wobei

$$T(v_j)=\sum_{\ell=1}^{p}\beta_{\ell j}w_\ell,\ 1\le j\le m$$

gilt. Wir wollen nun die Matrizendarstellung $M_{T\circ S}$ von $T\circ S$ ausrechnen. Deshalb berechnen wir

$$\begin{aligned}(T\circ S)(u_k) &= T(S(u_k)) \\ &= T\left(\sum_{j=1}^{m}\alpha_{jk}v_j\right) \\ &= \sum_{j=1}^{m}\alpha_{jk}T(v_j) \\ &= \sum_{j=1}^{m}\alpha_{jk}\left(\sum_{\ell=1}^{p}\beta_{\ell j}w_\ell\right) \\ &= \sum_{\ell=1}^{p}\underbrace{\left(\sum_{j=1}^{m}\beta_{\ell j}\alpha_{jk}\right)}_{=:\gamma_{\ell k}}w_\ell\end{aligned}$$

und bekommen eine neue $p \times m$–Matrix $M_{T\circ S} = (\gamma_{\ell k})$ als das **Matrizenprodukt**

$$M_{T\circ S} =: M_T \cdot M_S,\ (\gamma_{\ell k}) = (\beta_{\ell j}) \cdot (\alpha_{jk})$$

mit den Rechenregeln

$$\gamma_{\ell k} \ = \ \sum_{j=1}^{m} \beta_{\ell j}\alpha_{jk},\ 1 \le \ell \le p,\ 1 \le k \le n. \tag{4.38}$$

Theorem 4.39 *Werden zwei lineare Abbildungen $S : U \to V$, $T : V \to W$ bei geeignet gewählten Erzeugendensystemen in U, V, W durch Matrizen M_S und M_T dargestellt, so wird $T \circ S$ durch das Matrizenprodukt $M_{T\circ S} =: M_T \cdot M_S$ dargestellt.* □

Die Formel (4.38) ist von zentraler Bedeutung, weil das Matrizenprodukt unerwartet oft in mathematischen Rechnungen auftritt.

Wir können jetzt den Hintergrund der linearen Abbildungen vergessen und uns ganz auf die durch (4.38) beschriebene Multiplikation einer $p\times m$–Matrix $(\beta_{\ell j})$ mit einer $m\times n$–Matrix (α_{jk}) zu einer $p\times n$–Matrix $(\gamma_{\ell k})$ konzentrieren. Die Spaltenzahl m des"linken" Faktors muß immer gleich der Zeilenzahl des "rechten" Faktors sein, und bei der Ausführung der Multiplikation wird eine Summe über m Produkte berechnet. In Kurzform:

$$\begin{array}{rcrcl} (p \times m) & \cdot & (m \times n) & \mapsto & (p \times n) \\ (\beta_{\ell j}) & \cdot & (\alpha_{jk}) & \mapsto & (\gamma_{\ell k}) \\ & & & = & \left(\displaystyle\sum_{j=1}^{m} \beta_{\ell j}\alpha_{jk}\right) \end{array}$$

Man berechnet das Element $\gamma_{\ell k}$ der Ergebnismatrix so, daß man ein Skalarprodukt der ℓ–ten Zeile des "linken" Faktors mit der k–ten Spalte des "rechten" Faktors berechnet. Man sieht hier auch wieder die Bedeutung des Skalarprodukts, und man erkennt das Skalarprodukt (4.37) selbst als Matrixmultiplikation einer $1 \times n$–Matrix mit einer $n \times 1$–Matrix. Das Ergebnis ist eine 1×1–Matrix, und diese Matrizen identifizieren wir mit den Skalaren, aus denen sie bestehen.

Durch die **Transposition** macht man aus einem Zeilenvektor einen Spaltenvektor und umgekehrt. Allgemeiner ist die **Transponierte** A^T einer $m \times n$–Matrix A mit Elementen a_{jk} die $n \times m$–Matrix mit Elementen a_{kj} für $1 \le j \le m,\ 1 \le k \le n$. Die Transposition ist eine einstellige Operation, die

wir mit dem hoch– und nachgestellten ^{T} bezeichnen (Postfixnotation). Man vertauscht Zeilen– und Spaltenindex. Die **Diagonale** der Matrix, nämlich die Schrägreihe der Elemente a_{jj} mit gleichen Indizes, bleibt dabei erhalten. Wir fassen das Wichtigste über Matrizen hier zusammen, obwohl einige der Begriffe erst später verständlich werden:

Definition 4.40 *1. Eine $m \times n$–***Matrix** *A über einem Skalarenkörper $\mathbb{K}$ ist ein rechteckiges Schema von Skalaren $\alpha_{jk} \in \mathbb{K}$, $1 \leq j \leq m$, $1 \leq k \leq n$ der Form*

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix}$$

mit dem **Zeilenindex** *j mit Werten zwischen 1 und m sowie dem* **Spaltenindex** *k mit Werten zwischen 1 und n.*

2. Als Kurzschreibweise für diesen Sachverhalt dient $A = (\alpha_{jk}) \in \mathbb{K}^{m \times n}$.

3. Vektoren aus dem $\mathbb{K}^n$ schreiben wir als Matrizen aus $\mathbb{K}^{n \times 1}$, d.h. als Spaltenvektoren.

4. Die **Transponierte** *von A ist $A^T = (\alpha_{kj}) \in \mathbb{K}^{n \times m}$.*

5. A heißt **symmetrisch***, wenn $A = A^T$ gilt.*
Das erzwingt $m = n$, d.h. die Matrix muß quadratisch sein.

6. Für Matrizen $A = (a_{jk}) \in \mathbb{C}^{m \times n}$ ist
$\overline{A} := (\overline{a_{jk}}) \in \mathbb{C}^{m \times n}$ und $A^* := (\overline{a_{kj}}) \in \mathbb{C}^{n \times m} = \overline{A^T} = (\overline{A})^T$.

7. A heißt **hermitesch***, wenn $A = A^* := \overline{A^T}$ gilt.*
Auch das erzwingt $m = n$, d.h. die Matrix muß quadratisch sein.

8. Die **Einheitsmatrix** *$I_n = (\delta_{jk}) \in \mathbb{K}^{n \times n}$ ist die $n \times n$–Matrix, deren Elemente durch das* **Kroneckersymbol**

$$\delta_{jk} := \left\{ \begin{array}{ll} 1 & \text{falls } j = k \\ 0 & \text{falls } j \neq k \end{array} \right\} \in \mathbb{K},\ 1 \leq j \leq n$$

gegeben sind. Im Sinne von Theorem 4.35 stellt diese Matrix die Identitätsabbildung auf $\mathbb{K}^n$ dar.

9. *Das* **Matrixprodukt** *einer $m \times n$–***Matrix** $A = (a_{jk})$ *mit einer $n \times p$–***Matrix** $B = (b_{k\ell})$ *ist die $m \times p$–***Matrix** $C = (c_{j\ell}) = A \cdot B$ *mit*

$$c_{j\ell} = \sum_{k=1}^{n} a_{jk} b_{k\ell},\ 1 \leq j \leq m,\ 1 \leq \ell \leq p.$$

10. *Eine Matrix $A = (\alpha_{jk}) \in \mathbb{K}^{m \times n}$ stellt über das Matrix-Vektor Produkt aine lineare Abbildung $\mathbb{K}^n \to \mathbb{K}^n$ mit*

$$x \mapsto A \cdot x \text{ für alle } x \in \mathbb{K}^n$$

dar.

11. *Eine Matrix $A \in \mathbb{K}^{n \times n}$ heißt* **invertierbar** *oder* **nichtsingulär**, *wenn es eine Matrix $A^{-1} \in \mathbb{K}^{n \times n}$ gibt mit $A \cdot A^{-1} = I_n$. Die Matrix A^{-1} heißt dann* **Inverse** *von A.*

12. *$A \in \mathbb{R}^{n \times n}$ heißt* **orthogonal**, *wenn $A \cdot A^T = A^T \cdot A = I_n$ gilt.*

13. *$A \in \mathbb{C}^{n \times n}$ heißt* **unitär**, *wenn $A \cdot A^* = A^* \cdot A = I_n$ gilt.*

Theorem 4.41 *Ist $S : \mathbb{K}^n \to \mathbb{K}^n$ eine bijektive lineare Abbildung, die durch die $n \times n$–Matrix M_S dargestellt wird, so ist die Matrix M_S invertierbar und stellt die Abbildung S^{-1} dar. In beiden Fällen ist die Matrixdarstellung bezüglich der Einheitsvektoren gemeint.*

Zum Beweis bemerken wir zuerst, daß $M_{S^{-1}} \in \mathbb{K}^{n \times n}$ als eine die Abbildung S^{-1} darstellende Matrix existieren muß, denn S^{-1} ist eine lineare (und auch bijektive) Abbildung $\mathbb{K}^n \to \mathbb{K}^n$. Zu zeigen ist $M_{S^{-1}} \cdot M_S = I_n$. Aber weil wir wissen, daß $Id_{\mathbb{K}^n} = S^{-1} \circ S$ gilt, und weil man sehr einfach einsieht, daß die Identitätsabbildung $Id_{\mathbb{K}^n}$ durch die Einheitsmatrix I_n dargestellt wird, folgt aus Theorem 4.39 und Theorem 4.35, daß

$$M_{Id_{\mathbb{K}^n}} = I_n = M_{S^{-1}} \cdot M_S$$

gilt. Mit demselben Argument sieht man auch, daß $I_n = M_S \cdot M_{S^{-1}}$ gilt. □

Theorem 4.42 1. *Das Matrizenprodukt ist assoziativ, aber nicht kommutativ.*

2. *Die Matrizen aus $\mathbb{K}^{m \times n}$ bilden einen Vektorraum über $\mathbb{K}$ unter komponentenweiser Addition und Skalarmultiplikation.*

3. *Die invertierbaren Matrizen aus $\mathbb{K}^{n\times n}$ bilden die* **allgemeine lineare Gruppe** *$GL(n,\mathbb{K})$ über $\mathbb{K}$ unter der Matrixmultiplikation. Dabei gilt $A\cdot A^{-1}=A^{-1}\cdot A=I_n$, obwohl die Multiplikation im allgemeinen nicht kommutativ ist. Ferner hat man die in Gruppen geltenden allgemeinen Regeln, aber die Matrixmultiplikation ist nicht kommutativ. Die Matrizen in $GL(n,\mathbb{K})$ stellen genau die bijektiven linearen Abbildungen $\mathbb{K}^n\to\mathbb{K}^n$ dar, wenn man die Einheitsvektoren als Erzeugende wählt.*

4. *Es gelten die Rechenregeln*

$$\begin{aligned}
(A^T)^T &= A\\
\overline{(\overline{A})} &= A\\
(A^*)^* &= A\\
(A\cdot B)^T &= B^T\cdot A^T\\
(A\cdot B)^* &= B^*\cdot A^*\\
(A\cdot B)\cdot C &= A\cdot(B\cdot C)\\
A\cdot(B+C) &= A\cdot B+A\cdot C\\
(A+B)\cdot C &= A\cdot C+B\cdot C\\
\alpha\cdot(B\cdot C) &= (\alpha\cdot B)\cdot C)=B\cdot(\alpha\cdot C)
\end{aligned}$$

   soweit die Matrizenprodukte überhaupt definiert sind.

5. *Das Skalarprodukt (u,v) von Vektoren $u,v\in\mathbb{K}^n$ läßt sich als Matrizenprodukt*

$$u^T\cdot v=v^T\cdot u$$

   schreiben.

6. *Ist I_n die $n\times n$–Einheitsmatrix, so gelten für alle $m\times n$–Matrizen $A\in\mathbb{K}^{m\times n}$ und alle $n\times p$–Matrizen $B\in\mathbb{K}^{n\times p}$ die Gleichungen*

$$A=A\cdot I_n=I_m\cdot A\ \text{ und }\ B=I_n\cdot B=B\cdot I_p.$$

7. *Sind $A,B\in\mathbb{K}^{n\times n}$ invertierbar, so gilt*

$$\begin{aligned}
(A^T)^{-1} &= (A^{-1})^T\\
(\overline{A})^{-1} &= \overline{A^{-1}}\\
(A^*)^{-1} &= (A^{-1})^*\\
(A\cdot B)^{-1} &= B^{-1}\cdot A^{-1}.
\end{aligned}$$

8. *Ist $A\in GL(n,K)$ symmetrisch, so auch A^{-1}. Die symmetrischen Matrizen aus $GL(n,K)$ bilden eine Untergruppe, die* **spezielle lineare Gruppe** *$SL(n,K)$.*

9. *Ist $A \in GL(n, \mathbb{R})$ orthogonal, so auch $A^{-1} = A^T$. Die orthogonalen Matrizen aus $GL(n, \mathbb{R})$ bilden eine Untergruppe von $GL(n, \mathbb{R})$, die* **orthogonale Gruppe** *$O(n)$.*

10. *Ist $A \in GL(n, \mathbb{C})$ unitär, so auch $A^{-1} = A^*$. Die unitären Matrizen aus $GL(n, \mathbb{C})$ bilden eine Untergruppe von $GL(n, \mathbb{C})$, die* **unitäre Gruppe** *$U(n)$.*

11. *Zu jeder Matrix $A \in \mathbb{K}^{m \times n}$ ist die Matrix $A^*A \in \mathbb{K}^{n \times n}$ hermitesch.*

Machen wir einen exemplarischen Beweis vor, und zwar für die Aussage:

Ist $A \in GL(n, K)$ symmetrisch, so auch A^{-1},

wobei wir der Einfachheit halber alle anderen obigen Aussagen voraussetzen. Dann ist es nicht schwer, denn man hat

$$A^{-1} = (A^T)^{-1} = (A^{-1})^T$$

falls A invertierbar und symmetrisch ist.

Neben den Rechenregeln für Matrizen sollten man auch noch auf Fallen hinweisen:

1. Es wird immer wieder der Fehler gemacht, die Faktoren in einem Matrizenprodukt zu vertauschen, d.h. so zu tun, als würde das Kommutativgesetz gelten. Ein typisches Beispiel ist, die in $\mathbb{R}$ gültige binomische Formel
$$(a + b)^2 = a^2 + 2ab + b^2$$
naiv auf Matrizen anzuwenden. Korrekt ist aber nur
$$(A + B) * (A + B) = A * A + A * B + B * A + B * B.$$

2. Aus $A * B = 0$ folgt keineswegs $A = 0$ oder $B = 0$, sofern man Matrizen mit mindestens zwei Zeilen oder Spalten betrachtet. Dazu eine kleine Aufgabe: Man gebe eine Matrix $A \in \mathbb{R}^{2\times 2}$ an, für die $A^2 = A * A = 0$ gilt.

3. Aber man kann von $A * B = 0$ auf $B = 0$ schließen, wenn A invertierbar ist.

4. Analog kann man aus einer Matrizengleichung der Form
$$A * B = A * C$$
nur dann problemlos $B = C$ folgern, wenn A invertierbar ist. Es reicht nicht, $A \neq 0$ vorauszusetzen.

Wir wollen noch klären, wie wir mit dem Dualraum $(K^n)^*$ umgehen wollen. Aber erst einmal überlegen wir uns

Lemma 4.43 *Hat man für einen Vektor $x \in I\!K^n$ die Gleichungen*

$$x^T y = 0 \text{ für alle } y \in I\!K^n,$$

so folgt $x = 0$.

Das ist leicht zu beweisen, wenn man $y = e_j,\ 1 \leq j \leq n$ setzt. □.

Wir werden das gleich in der folgenden Form anwenden:

Lemma 4.44 *Hat man für zwei Vektoren $x,\ z \in I\!K^n$ die Gleichungen*

$$x^T y = z^T y \text{ für alle } y \in I\!K^n,$$

so folgt $x = z$. □

Frage: Warum folgt das aus dem obigen Lemma?

Aber jetzt ist es Zeit, auf den Dualraum zuzusteuern. Jeder Vektor $x \in I\!K^n$ definiert ein lineares Funktional $S_n(x)$ aus $(K^n)^*$ durch

$$S_n(x)\ :\ y \mapsto x^T y = (S_n(x))(y) \in I\!K \text{ für alle } y \in I\!K^n,$$

die durch das Skalarprodukt gegeben ist. Damit ist

$$S_n\ :\ I\!K^n \to (I\!K^n)^*$$

eine lineare Abbildung. Man kann mit obigem Lemma sehr leicht sehen, daß diese Abbildung injektiv ist.

Sie ist aber auch bijektiv, weil wir eine Inverse angeben können. Zu jedem linearen Funktional $\lambda \in (I\!K^n)^*$ bilden wir den Vektor

$$R_n(\lambda) := (\lambda(e_1), \ldots, \lambda(e_n))^T$$

und bekommen eine lineare Abbildung von $(K^n)^*$ in K^n. Sie erfüllt die Gleichung

$$\begin{aligned} R_n(\lambda)^T \cdot y &= (\lambda(e_1), \ldots, \lambda(e_n))y \\ &= \lambda(y) \text{ für alle } y \in I\!K^n \end{aligned}$$

und deshalb auch

$$\begin{aligned} (S_n(R_n(\lambda)))(y) &= (R_n(\lambda))^T y \\ &= \lambda(y) \text{ für alle } y \in I\!K^n,\ \lambda \in (K^n)^*. \end{aligned}$$

Das bedeutet $S_n \circ R_n = Id_{(K^n)^*}$ und S_n muß auch surjektiv, d.h. insgesamt bijektiv sein, und die Inverse ist R_n. Wir fassen zusammen:

Theorem 4.45 *Die Vektorräume $\mathbb{K}^n$ und $(\mathbb{K}^n)^*$ sind isomorph bezüglich der oben angegebenen Abbildungen*

$$S_n \ : \ \mathbb{K}^n \to (\mathbb{K}^n)^*, \ R_n \ : \ (\mathbb{K}^n)^* \to \mathbb{K}^n. \square$$

Deshalb kann man den Dualraum $(K^n)^*$ so uminterpretieren, daß man Zeilenvektoren aus n Komponenten aus K als ein Funktional λ auffaßt und einfach

$$\lambda(x) := \lambda \cdot x \text{ für alle } x \in \mathbb{K}^n$$

definiert. Mit anderen Worten: schreibt man die Vektoren aus K^n als Spaltenvektoren und die Funktionale aus $(K^n)^*$ als Zeilenvektoren von je n Komponenten, so ist die Wirkung eines linearen Funktionals gerade durch das Matrixprodukt *Zeilenvektor* $\cdot$ *Spaltenvektor*, d.h. durch das Skalarprodukt der beiden Vektoren gegeben. Man sieht also, daß das Duale mit der Transposition zusammenhängt, und das gilt auch für die linearen Abbildungen:

Theorem 4.46 *Stellt eine Matrix $A \in \mathbb{K}^{m\times n}$ eine lineare Abbildung $\mathbb{K}^n \to \mathbb{K}^m$ dar, so wird die duale Abbildung durch die Transponierte von A dargestellt, und zwar als Abbildung zwischen Zeilenvektoren:*

$$A^d(y^T) = (A^T y)^T \text{ für alle } y^T \in (\mathbb{K}^m)^*.$$

Beweis: Es sei eine Matrix $A \in \mathbb{K}^{m\times n}$ als Abbildung $\mathbb{K}^n \to \mathbb{K}^m$ gegeben. Die duale Abbildung $A^d \ : \ (K^m)^* \to (K^n)^*$ erfüllt $(A^d(\lambda))(x) = \lambda(A(x))$ für alle $\lambda \in (K^m)^*$ und alle $x \in \mathbb{K}^n$. Benutzen wir die Darstellung beliebiger Funktionale λ durch Zeilenvektoren y^T für $y \in \mathbb{K}^m$, so folgt

$$\begin{aligned} (A^d(y^T))(x) &= y^T(A(x)) \\ &= y^T \cdot A \cdot x \\ &= (A^T y)^T \cdot x \text{ für alle } x \in K^n, \ y \in \mathbb{K}^m, \end{aligned}$$

und aus Lemma 4.43 folgt die Behauptung. $\square$

Mit Satz 4.46 und Satz 4.28 kann man dann auch sinsehen daß

$$(A^T)^{-1} = (A^{-1})^T \text{ für alle } A \in \mathbb{K}^{n\times n}$$

gilt.

Das Rechnen mit Matrizen muß unbedingt geübt werden. Mit den oben schon definierten Einheitsvektoren e_j kann man aus einer Matrix $A = (\alpha_{jk}) \in$

$\mathbb{K}^{m\times n}$ die k–te Spalte als Vektor $A \cdot e_k \in \mathbb{K}^m$ und die j–te Zeile als n–Tupel $e_j^T \cdot A$ herausziehen. Das Element α_{jk} ist nichts anderes als $e_j^T A e_k$. Die Einheitsvektoren haben die Skalarprodukte

$$e_j^T \cdot e_k = \delta_{jk} = e_k^T \cdot e_j.$$

Einen beliebigen Vektor $x \in \mathbb{K}^n$ kann man schreiben als

$$x = \sum_{j=1}^{n} e_j^T \cdot x \cdot e_j = \sum_{j=1}^{n} x^T \cdot e_j \cdot e_j$$

und eine Matrix $A = (\alpha_{jk}) \in \mathbb{K}^{m\times n}$ als

$$A = \sum_{j=1}^{m}\sum_{k=1}^{n} e_j^T \cdot A \cdot e_k \cdot e_j \cdot e_k^T.$$

4.5 Basis und Dimension

4.5.1 Basen und lineare Unabhängigkeit

Sehen wir uns an, was passiert, wenn in einem Erzeugendensystem $\{v_1, \ldots, v_N\}$ ein Vektor v_j überflüssig ist. Man kann dann v_j durch die übrigen Vektoren ausdrücken, d.h. es gilt

$$\begin{aligned} v_j &= \sum_{\substack{k=1 \\ k\neq j}}^{N} \alpha_k v_k \text{ oder} \\ 0 &= \sum_{k=1}^{N} \alpha_k v_k \text{ mit } \alpha_j = -1 \end{aligned}$$

Wenn man das j nicht vorher weiß, kann man sagen, daß die Menge $\{v_1, \ldots, v_N\}$ sicher dann keine Basis ist, wenn es Koeffizienten $\alpha_1, \ldots, \alpha_N$ gibt, die nicht alle Null sind, so daß

$$0 = \sum_{k=1}^{N} \alpha_k v_k \tag{4.47}$$

gilt, denn man kann dann nach den v_j mit $\alpha_j \neq 0$ auflösen und damit v_j durch die anderen Vektoren ausdrücken.

Definition 4.48 *1. Eine endliche Teilmenge $\{v_1, \ldots, v_N\}$ von Vektoren eines Vektorraums V über einem Skalarenkörper $\mathbb{K}$ heißt* **linear unabhängig**, *wenn aus einer Gleichung der Form (4.47) immer folgt, daß alle Koeffizienten α_j gleich Null sind.*

2. Man nennt eine Linearkombination der Null aus Vektoren mit Koeffizienten gleich Null auch eine **triviale** *Linearkombination. Lineare Unabhängigkeit von $\{v_1, \ldots, v_N\}$ bedeutet also, daß die einzige mögliche Linearkombination aus $\{v_1, \ldots, v_N\}$, die Null ergibt, trivial sein muß.*

3. Eine unendliche Teilmenge X von Vektoren eines Vektorraums V über $\mathbb{R}$ heißt **linear unabhängig**[1], *wenn jede endliche Teilmenge linear unabhängig ist.*

Theorem 4.49 *Basen sind linear unabhängig.*

Das ist nach der obigen Argumentation klar, denn bei linearer Abhängigkeit ist mindestens ein Basisvektor überflüssig. □

Theorem 4.50 *Es seien die Vektoren $\{v_1, \ldots, v_N\}$ eines Vektorraums V linear unabhängig. Dann gilt: In der Darstellung*

$$x = \sum_{j=1}^{n} \alpha_j v_j$$

eines beliebigen Vektors aus der linearen Hülle der $\{v_1, \ldots, v_N\}$ sind die Koeffizienten α_j eindeutig bestimmt.

Zum Beweis nehmen wir an, es gäbe eine weitere Darstellung

$$x = \sum_{j=1}^{n} \beta_j v_j.$$

Dann ist

$$0 = \sum_{j=1}^{n} (\beta_j - \alpha_j) v_j$$

eine Darstellung der Null, und alle Koeffizienten müssen verschwinden wegen der linearen Unabhängigkeit. Also gilt $\alpha_j = \beta_j,\ 1 \leq j \leq n$. □

Theorem 4.51 *Die Monome sind linear unabhängig als Elemente des Vektorraums $\mathbb{R}^{\mathbb{R}}$.*

[1] `http://de.wikipedia.org/wiki/Lineare_Unabh%C3%A4ngigkeit`

Allgemeiner gilt, daß ein Polynom vom Grade $n \geq 0$ höchstens an n verschiedenen Punkten (**Nullstellen**) verschwinden kann, aber das wollen wir hier nicht beweisen, sondern nur benutzen. Es wird vermutlich in der Diskreten Mathematik bewiesen. Wenn dann aber ein Polynom überall verschwindet, so kann es keinen Grad $n \geq 0$ haben, und das Polynom kann nur Nullen als Koeffizienten haben. □

Theorem 4.52 *Der Vektorraum $\mathbb{K}^n$ hat die Menge der n Einheitsvektoren $\{e_1, \ldots, e_n\}$ als Basis.*

Das ist klar,weil die Einheitsvektoren ein Erzeugendensystem sind und linear unabhängig sind.

Theorem 4.53 *Jeder Vektorraum mit einem endlichen Erzeugendensystem hat auch eine Basis.*

Das ist klar, weil das Minimum der Anzahl der Elemente aller denkbaren Erzeugendensysteme existiert und dann eine Basis liefert. □

Das Resultat gilt sinngemäß auch für beliebige Vektorräume, aber es ist dann schwieriger zu beweisen.

Es gibt dazu auch eine konstruktive Variante:

Theorem 4.54 *Jedes endliche Erzeugendensystem enthält eine Basis.*

Ist ein Erzeugendensystem keine Basis, so ist es linear abhängig, und wir wissen schon, daß man dann einen Vektor weglassen kann, um ein kleineres Erzeugendensystem zu bekommen. Dann fahren wir induktiv fort, bis wir bei einer Basis ankommen. □

Die lineare Unabhängigkeit hat aber auch noch eine geometrische Nebenbedeutung, die nicht unterschlagen werden sollte. Gehen wir dazu mit Schulkenntnissen in den $\mathbb{R}^2$. Wie kann man dort die Menge aller Geraden durch den Nullpunkt beschreiben? In der Schule lernen manche, daß Nullpunktsgeraden immer die Form

$$\{(x, y) \ : \ y = mx, \ x, y \in \mathbb{R}\} \text{ für alle } \ m \in \mathbb{R}$$

haben, aber dann ist die y–Achse $\{0\} \times \mathbb{R}$ nicht dabei. Man muß die definierende Gleichung symmetrisch zu x und y machen, etwa indem man $0 = mx + ny$ schreibt. Es dürfen aber nicht m und n beide Null sein, sonst ist die Menge der ganze $\mathbb{R}^2$. Man bekommt deshalb **alle** Nullpunktsgeraden durch die "homogene" Schreibweise

$$G_{m,n} := \{(x, y) \ : \ 0 = mx + ny, \ x, y \in \mathbb{R}\} \text{ für alle } \ m, n \in \mathbb{R}, \ m^2 + n^2 > 0.$$

Wann fallen zwei Geraden zusammen, die durch die Paare (m_1, n_1) und (m_2, n_2) des $\mathbb{R}^2 \setminus \{(0,0)\}$ gegeben sind? Offenbar genau dann, wenn für alle $(x, y) \in \mathbb{R}^2$ die Gleichungen $0 = m_1 x + n_1 y$ und $0 = m_2 x + n_2 y$ logisch äquivalent sind. Nehme wir an, die Geraden fielen zusammen. Auf der ersten Geraden liegt der Punkt $(n_1, -m_1)$, und dieser muß dann auf der zweiten liegen, d.h. es folgt

$$m_1 n_2 = m_2 n_1.$$

Dann folgt aber

$$\begin{aligned} m_2(m_1, n_1) - m_1(m_2, n_2) &= (0,0) \\ n_2(m_1, n_1) - n_1(m_2, n_2) &= (0,0) \end{aligned}$$

und weil nicht beide Linearkombinationen trivial sein können, sind die Vektoren (m_1, n_1) und (m_2, n_2) linear abhängig. Wenn die Geraden zusammenfallen, sind also die Vektoren der definierenden Gleichungen linear abhängig.

Gilt auch die Umkehrung? Es seien nun die Vektoren (m_1, n_1) und (m_2, n_2) linear abhängig, d.h. es gibt Koeffizienten α und β, nicht beide Null, so daß

$$\begin{aligned} \alpha(m_1, n_1) + \beta(m_2, n_2) &= (0,0) \\ \alpha(m_1, n_1) &= -\beta(m_2, n_2) \end{aligned}$$

gilt. Wir können ohne Einschränkung annehmen, daß $\alpha = 1$ gilt, und es folgt

$$\begin{aligned} m_1 &= -\beta m_2 \\ n_1 &= -\beta n_2 \\ m_1 n_2 &= -\beta m_2 n_2 \\ &= m_2 n_1 \end{aligned}$$

und nach der obigen Argumentation schließt man darauf, daß die Geraden übereinstimmen.

Das Ganze läßt sich auch etwas allgemeiner sehen. Nehmen wir an, für die Unbekannten $x_1, \ldots, x_n$ bestünden die zwei Gleichungen

$$\begin{aligned} a_1 x_1 + a_2 x_2 + \ldots + a_n x_n &= c \\ b_1 x_1 + b_2 x_2 + \ldots + b_n x_n &= d. \end{aligned}$$

Wir interessieren uns dafür, wann eine der beiden Gleichungen "überflüssig" ist, d.h. sich aus der anderen ergibt. Das ist dann der Fall, wenn man eine der Gleichungen durch Multiplikation mit einem Faktor aus der anderen bekommen kann, d.h. wenn eine der Gleichungen

$$\begin{aligned} (a_1, a_2, \ldots, a_n, c) &= \beta(b_1, b_2, \ldots, b_n, d) \\ (b_1, b_2, \ldots, b_n, d) &= \beta(a_1, a_2, \ldots, a_n, c) \end{aligned}$$

gilt, und das ist der Fall, wenn die Vektoren $(a_1, a_2, \ldots, a_n, c)$ und $(b_1, b_2, \ldots, b_n, d)$ linear abhängig sind. Umgekehrt folgt aus der linearen Abhängigkeit dieser Vektoren auch die Überflüssigkeit einer der beiden Gleichungen.

4.5.2 Dimension

Wir beginnen mit

Theorem 4.55 *Hat ein Vektorraum V über einem Skalarenkörper $\mathbb{K}$ eine endliche Basis $X := \{v_1, \ldots, v_n\}$, so ist er isomorph zu $\mathbb{K}^n$.*

Zum Beweis definiere man die Abbildung $S : \mathbb{K}^n \to V$ mit

$$(\xi_1, \ldots, \xi_n)^T \mapsto \sum_{j=1}^{n} \xi_j v_j \tag{4.56}$$

und rechne nach, daß sie linear, injektiv und surjektiv ist. □

Wir werden im folgenden zu einem Vektorraum V über $\mathbb{K}$ mit Basis $\{v_1, \ldots, v_n\}$ die Abbildung (4.56) den **Standard–Isomorphismus** zwischen $\mathbb{K}^n$ und V nennen.

Definition 4.57 *Die Anzahl der Elemente einer Basis eines Vektorraums V heißt* **Dimension** *des Vektorraums.*

Theorem 4.58 *Die Dimension eines Vektorraums ist eindeutig bestimmt. Sie ist entweder unendlich oder gleich der Anzahl der Elemente einer endlichen Basis. Alle Vektorräume der Dimension n über einem Grundkörper $\mathbb{K}$ sind zueinander und zu $\mathbb{K}^n$ isomorph.*

Das folgt natürlich aus dem bisher schon Gesagten. Die Dimension eines Vektorraums muß Unendlich sein, wenn es kein endliches Erzeugendensystem gibt. Andernfalls verkleinere man ein solches zu einer Basis, und man kann eine Basis kleinster Länge $n < \infty$ finden. Dieses n ist eindeutig bestimmt und gibt die Dimension des Vektorraums an. Mit der Abbildung 4.56 bekommt man dann die Isomorphie zu $\mathbb{K}^n$. □

Wir fassen zusammen, was man über surjektive, injektive und bijektive lineare Abbildungen im Zusammenhang mit Erzeugendensystemen und Basen sagen kann:

Theorem 4.59

1. *Die Umkehrabbildung eines Vektorraumisomorphismus ist ein Vektorraumisomorphismus.*

2. *Das Bild einer Menge linear unabhängiger Vektoren unter einer injektiven linearen Abbildung zwischen Vektorräumen ist linear unabhängig.*

3. *Das Bild eines Erzeugendensystems unter einer surjektiven linearen Abbildung zwischen Vektorräumen ist ein Erzeugendensystem.*

4. *Ein Vektorraumisomorphismus bildet Basen in Basen ab.*

5. *Die Dimensionen isomorpher Vektorräume sind gleich.*

Die Beweise sind durchweg einfach, wenn man sie in der obigen Reihenfolge ausführt. Teil 1 ist in Theorem 4.21 schon enthalten. Zu Teil 2 nehmen wir linear unabhängige Vektoren $u_1, \ldots, u_n$ aus U und bilden mit einer injektiven linearen Abbildung $T : U \to V$ ab auf $T(u_1), \ldots, T(u_n)$. Wären diese Vektoren linear abhängig, so gäbe es eine nichttriviale Linearkombination

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \alpha_j T(u_j) \\ &= T\left(\sum_{j=1}^{n} \alpha_j u_j \right) \end{aligned}$$

und wegen der Injektivität von T folgt, daß

$$0 = \sum_{j=1}^{n} \alpha_j u_j$$

eine nichttriviale Linearkombination der u_j ist, was nicht möglich ist.

Zu Teil 3 nimmt man bei gleichen Bezeichnungen wie oben ein Erzeugendensystem M von U her und bildet $N := T(M) \subset V$. Um zu zeigen, daß N ein Erzeugendensystem für V ist, nehmen wir einen beliebigen Vektor $v \in V$ und stellen ihn wegen der Surjektivität von T als Bild $v = T(u)$ eines Vektors in U dar. Dieser ist im Erzeugendensystem M als Linearkombination darstellbar, und wenn wir diese Linearkombination mit T in V abbilden, bekommen wir eine Darstellung von $v = T(u)$ durch $N = T(M)$.

Teil 4 folgt sofort aus den Teilen 1, 2 und 3, und Teil 5 ist Theorem 4.58. □

4.5.3 Isomorphiesatz

Wir nehmen jetzt eine lineare Abbildung $T : U \to V$ zwischen Vektorräumen U und V über einem gemeinsamen Grundkörper $\mathbb{K}$ her (die Anordnung brauchen wir auch in diesem Abschnitt nicht). Der Satz 1.36 kann hier angewendet und entscheidend verschärft werden. Er besagt, daß $T(U)$ und die Menge der Äquivalenzklassen von U unter der Äquivalenzrelation $R \subseteq U \times U$ mit

$$uRv \leftrightarrow T(u) = T(v) \text{ für alle } u, v \in U$$

bijektiv aufeinander abgebildet werden können. Wir werden sehen, daß diese Bijektion eine lineare Abbildung zwischen Vektorräumen ist.

Aber dazu stellen wir erst einmal fest, daß wegen der Linearität von T gilt

$$uRv \leftrightarrow T(u) = T(v) \leftrightarrow T(u - v) = 0 \text{ für alle } u, v \in U.$$

Definition 4.60 *Ist $T : U \to V$ eine lineare Abbildung zwischen Vektorräumen U und V über einem gemeinsamen Grundkörper $\mathbb{K}$, so sind*

$$\begin{aligned} \ker T &:= \{u \in U : T(u) = 0\} \\ \text{range } T &:= T(U) \subseteq V \end{aligned}$$

der **Kern** *und das* **Bild** *von T.*

Theorem 4.61 *Kern und Bild von linearen Abbildungen $T : U \to V$ sind lineare Unterräume von U bzw. V. Eine lineare Abbildung ist genau dann injektiv, wenn ihr Kern nur der Nullraum ist.*

Aufgabe: Man führe diese einfachen Beweise übungshalber durch.

Jetzt müssen wir uns mit den Äquivalenzklassen bezüglich der obigen Relation näher befassen. Wir können das etwas allgemeiner tun, indem wir den Kern $\ker T$ von T durch einen allgemeinen Unterraum U_0 von U ersetzen und die Relation

$$uRv \leftrightarrow u - v \in U_0 \text{ für alle } u, v \in \mathbb{R}$$

betrachten. Das ist eine Äquivalenzrelation, und wir wollen die Menge

$$U/U_0 := \{[u] : u \in U\}$$

der Äquivalenzklassen zu einem Vektorraum über $\mathbb{K}$ machen, dem **Faktorraum** oder **Quotientenraum** von U nach U_0. Die Vektorraumaddition definieren wir als

$$[u] + [v] := [u + v] \text{ für alle } u, v \in U$$

und müssen Wohldefiniertheit zeigen, d.h. aus $[u] = [u_1]$ und $[v] = [v_1]$ muß $[u + v] = [u_1 + v_1]$ folgen für alle $u, u_1, v, v_1 \in U$. Das ist einfach, wenn wir benutzen, daß $[u] = [u + w]$ gilt für alle $w \in U_0, u \in U$. Man bekommt

$$[u + v] \;=\; [u + v + \underbrace{(u_1 - u) + (v_1 - v)}_{\in U_0}] = [u_1 + v_1].$$

Die Skalarmultiplikation definiert man als

$$\alpha \cdot [u] := [\alpha \cdot u] \text{ für alle } \alpha \in \mathbb{K},\ u \in U$$

und wenn $[u] = [v]$ gilt, folgt $v - u \in U_0$ und

$$\alpha \cdot [u] = [\alpha \cdot u] = [\alpha \cdot u + \underbrace{\alpha \cdot (v - u)}_{\in U_0}] = [\alpha \cdot v] = \alpha \cdot [v],$$

d.h. auch die Skalarmultiplikation ist wohldefiniert. Mit einfachen, aber lästigen Schlüssen folgt

Theorem 4.62 *Ist U_0 ein linearer Unterraum eines Vektorraums U, so ist der* **Quotientenraum** *U/U_0 wieder ein Vektorraum unter den obigen Operationen. Man nennt die Dimension von U/U_0 die* **Codimension** *von U_0 bezüglich U.*

Jetzt nehmen wir wieder $U_0 = \ker T$ und bekommen

Theorem 4.63 *Ist $T : U \to V$ eine lineare Abbildung zwischen Vektorräumen U und V über einem gemeinsamen Grundkörper $\mathbb{K}$, so gibt es zwischen dem Quotientenraum $U/\ker T$ und dem Bildraum* range $(T) = T(U)$ *eine bijektive lineare Abbildung*

$$\tilde{T} : [u] \mapsto T(u) \textit{ für alle } u \in U$$

und es gilt die Isomorphie

$$T(U) \simeq U/\ker T.$$

Daß diese Abbildung wohldefiniert und bijektiv ist, wissen wir schon. Die Linearität folgt aus

$$\begin{aligned}\tilde{T}(\alpha[u]+\beta[v]) &= \tilde{T}([\alpha u+\beta v])\\ &= T(\alpha u+\beta v)\\ &= \alpha T(u)+\beta T(v)\\ &= \alpha\tilde{T}([u])+\beta\tilde{T}([v])\end{aligned}$$

für alle $u, v \in U,\ \alpha, \beta \in K$. □

Hat man nur einen Unterraum U_0 aber keine Abbildung T, so kann man sich durch $T : U \to U/U_0$ mit $u \mapsto [u] \in U/U_0$ eine lineare Abbildung verschaffen mit $T(U) = U/U_0$ und $\ker T = U_0$. Die beiden Fälle von Theorem 4.62 und Theorem 4.63 sind also nicht wesentlich verschieden.

Wir wissen also jetzt, daß $U/\ker T$ zu $T(U)$ isomorph ist, und wir wissen dann auch nach Theorem 4.59, daß die Dimensionen dieser Räume gleich sind. Aber was ist im allgemeinen die Dimension von $U/\ker T$, wenn man die Dimensionen von U und von $\ker T$ kennt? Und was ist die Dimension von U/U_0, wenn man die Dimensionen von U und von einem Unterraum U_0 kennt? Wir formulieren und beweisen eine abgeschwächte Form:

Theorem 4.64 *Ist die Dimension von U/U_0 endlich, so gilt*

$$\begin{array}{rcccl} U & \simeq & U_0 & \times & (U/U_0)\\ \dim U & = & \dim U_0 & + & \dim(U/U_0).\end{array}$$

Zum Beweis nehmen wir eine Basis von U/U_0 in der Form $\{[v_1], \ldots, [v_n]\}$ und definieren eine lineare Abbildung

$$S\ :\ U_0 \times (U/U_0) \to U,\ \left(u_0, \sum_{k=1}^{n} \alpha_k[v_k]\right) \mapsto u_0 + \sum_{k=1}^{n} \alpha_k v_k$$

Diese Abbildung ist injektiv, weil aus

$$u_0 + \sum_{k=1}^{n} \alpha_k v_k = 0$$

folgt, daß

$$\begin{aligned} 0 &= \left[u_0 + \sum_{k=1}^{n} \alpha_k v_k\right]\\ &= \left[\sum_{k=1}^{n} \alpha_k v_k\right]\\ &= \sum_{k=1}^{n} \alpha_k [v_k]\end{aligned}$$

gilt und deshalb alle α_k verschwinden müssen. Dann folgt aber auch $u_0 = 0$.

Die Abbildung ist auch surjektiv, weil man zu jedem $u \in U$ die Klasse $[u]$ darstellen kann als

$$[u] = \sum_{k=1}^{n} \alpha_k [v_k]$$

und man bekommt, daß

$$\left[u - \sum_{k=1}^{n} \alpha_k v_k \right] = 0$$

gilt, also mit irgendeinem $u_0 \in U_0$ auch

$$u - \sum_{k=1}^{n} \alpha_k v_k = u_0$$

gilt. Wir haben also die Isomorphie von U und $U_0 \times (U/U_0)$. Ist U endlichdimensional, so ist jeder der auftretenden Räume isomorph zu einem Raum der Form $\mathbb{K}^m$, und die Behauptung des Satzes folgt. Ist U nicht endlichdimensional, so kann U_0 auch nicht endlichdimensional sein, und wir sind ebenfalls fertig. □

Der Sinn des Isomorphiesatzes ist (unter anderem), daß die Dimension des Bildes einer linearen Abbildung immer um die Dimension des Kernes kleiner ist als die Dimension des vollen Urbildraums. Genauer gilt

Theorem 4.65 *Ist $T : U \to V$ eine lineare Abbildung zwischen zwei Vektorräumen über demselben Skalarenkörper $\mathbb{K}$ und ist U endlichdimensional, so gilt*

$$\dim T(U) = \dim U - \dim \ker T.$$

Außer bei injektiven Abbildungen gehen also immer Dimensionen "verloren", und zwar genau so viele, wie der Kern hat.

Eine weitere wichtige Konsequenz des Isomorphiesatzes ist

Theorem 4.66 *Ist $T : U \to V$ eine lineare Abbildung zwischen Vektorräumen der Dimension n, so sind Injektivität, Surjektivität und Bijektivität von T äquivalent.*

Beweis: Man sehe sich die aus dem Isomorphiesatz folgende Dimensionsgleichung

$$n - \dim \ker T = \dim T(U) \leq n$$

mit Verstand an. □

Am Schluß noch eine kleine Entspannungsübung:

Theorem 4.67 *Sind U und V zwei endlichdimensionale Unterräume eines Vektorraums W, so gilt*

$$\dim(U + V) + \dim(U \cap V) = \dim U + \dim V.$$

Dies ist eine lehrreiche Anwendung des Isomorphiesatzes. Weil

$$\dim(U \times V) = \dim U + \dim V$$

gilt (Frage: warum?), wird man die Abbildung

$$U \times V \to U + V, \;\; (u, v) \mapsto u + v \text{ für alle } u \in U, v \in V$$

heranziehen. Der Rest sollte jetzt klar sein. Oder?

4.5.4 Rang von Matrizen

Wir wollen nun noch unser Wissen über lineare Unabhängigkeit und Dimension auf Matrizen anwenden. Dazu gehen wir wieder auf die Matrizendarstellung linearer Abbildungen aus Abschnitt 4.5.1 zurück. Dort hatten wir noch nicht den Begriff der Basis. Wenn wir Satz 4.34 durch Benutzung von Basen verschärfen, bekommen wir

Theorem 4.68 *Eine lineare Abbildung $T : U \to V$ zwischen endlichdimensionalen Vektorräumen U und V über einem gemeinsamen Grundkörper $\mathbb{K}$ hat eine eindeutig bestimmte Matrixdarstellung im Sinne des Abschnitts 4.4, wenn man in U und V Basen $\{u_1, \ldots, u_n\} \subset U$ bzw. $\{v_1, \ldots, v_m\} \subset V$ wählt.*

Das ist klar, weil man in Abschnitt 4.4 die Bilder $T(u_k)$ der Basisvektoren u_k **eindeutig** in der Basis $\{v_1, \ldots, v_m\} \subset V$ darstellen kann.

Aber man kann auch die lineare Unabhängigkeit der Zeilen– und Spaltenvektoren von Matrizen untersuchen:

Definition 4.69 *Sei $A \in \mathbb{K}^{m \times n}$ eine Matrix. Sie hat m Zeilen als Vektoren des $\mathbb{K}^n$ und n Spalten als Vektoren des $\mathbb{K}^m$.*

1. *Die Dimension der linearen Hülle der Spaltenvektoren nennt man den* **Spaltenrang** *von A. Er ist höchstens gleich m.*

2. *Die Dimension der linearen Hülle der Zeilenvektoren nennt man den* **Zeilenrang** *von A. Er ist höchstens gleich n.*

Theorem 4.70 *Sei $A \in \mathbb{K}^{m\times n}$ eine Matrix, und wir identifizieren sie mit der linearen Abbildung $\mathbb{K}^n \to \mathbb{K}^m$ mit $x \mapsto A \cdot x$, $x \in \mathbb{K}^n$.*

1. *Der Spaltenrang von A ist genau dann gleich m, wenn A surjektiv ist.*

2. *Der Spaltenrang von A ist genau dann gleich n, wenn A injektiv ist.*

3. *A ist genau dann bijektiv, wenn n und m gleich sind und mit dem Spaltenrang von A übereinstimmen.*

4. *Zeilen– und Spaltenrang sind höchstens gleich* $\min(m, n)$.

Beweis: Der Spaltenrang ist die Dimension von $A(\mathbb{K}^n) \subseteq \mathbb{K}^m$. Daraus folgt die erste Behauptung.

Zum Beweis der zweiten stellen wir die Null als Linearkombination der Spalten von A dar als

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \alpha_j (Ae_j) \\ &= A\left(\sum_{j=1}^{n} \alpha_j e_j\right). \end{aligned}$$

Sind die Spalten linear abhängig, so gibt es eine nichttriviale solche Linearkombination, und A ist nicht injektiv. Ist A nicht injektiv, so kann man Koeffizienten finden, die nicht alle verschwinden, so daß die rechte Seite Null ist. Dann sind aber auch die Spalten linear abhängig.

Die dritte Behauptung kombiniert die ersten beiden.

Aus $A(\mathbb{K}^n) \subseteq \mathbb{K}^m$ folgt, daß der Spaltenrang höchstens gleich $\min(m, n)$ ist. Weil der Zeilenrang von A der Spaltenrang von A^T ist, folgt dasselbe für den Zeilenrang. □

Ist A bijektiv, so hat A eine Inverse A^{-1}, und deren Transponierte $(A^{-1})^T = (A^T)^{-1}$ ist die Inverse der Transponierten. Also ist auch die Transponierte bijektiv, und es folgt, daß auch der Zeilenrang von A gleich m und n ist, weil der Zeilenrang von A der Spaltenrang von A^T ist. Im bijektiven Falle sind also Zeilen– und Spaltenrang gleich. Es gibt eigentlich keinen guten Grund, warum das immer so sein müßte, aber wunderbarerweise gilt

Theorem 4.71 *Zeilenrang und Spaltenrang von Matrizen sind gleich, und man spricht deshalb vom* **Rang** *einer Matrix.*

Beweis: Wir üben das Rechnen mit Matrizen und wollen damit die "ziemlich trickreiche Indexfieselei" aus [4], S. 138 vermeiden. Hat man eine Matrizenmultiplikation $C = A \cdot B$, so besagt dies, daß sich die Spalten von C linear durch die Spalten von A kombinieren lassen, wobei die Matrix B die Koeffizienten enthält. Dies liegt an der Gleichung

$$Ce_k = A \cdot B \cdot e_k = \underbrace{\sum_j Ae_j e_j^T}_{=A} Be_k = \sum_j \underbrace{e_j^T Be_k}_{\in K} Ae_j.$$

Wir nehmen jetzt an, die Matrix $A \in I\!K^{m\times n}$ habe einen Spaltenrang $r \leq n$. Dann lassen sich alle n Spalten von A aus einer Teilmenge von nur r Spalten linear kombinieren. Es gibt also eine Matrixgleichung $A = \tilde{A} \cdot B$ mit einer Matrix $\tilde{A} \in I\!K^{m\times r}$, die aus den r linear unabhängigen Spalten von A besteht, und einer Koeffizientenmatrix $B \in I\!K^{r\times n}$. Dann gilt aber auch $A^T = B^T \cdot \tilde{A}^T$, und dies besagt, daß sich die Spalten von A^T aus den r Spalten von B^T linear kombinieren lassen. Also ist der Spaltenrang von A^T und damit auch der Zeilenrang von A höchstens gleich r. Wir haben damit bewiesen, daß

$$\text{Zeilenrang } \leq \text{ Spaltenrang}$$

gilt, und aus Symmetriegründen folgt die umgekehrte Relation auch. □

Theorem 4.72 *Für ein Matrizenprodukt $C = A \cdot B$ gilt immer*

$$\textit{Rang}\ (C) \leq \min(\ \textit{Rang}\ (A),\ \textit{Rang}\ (B)).$$

Beweis: Die Spalten von C sind Linearkombinationen der Spalten von A. Also folgt $\text{Rang}\ (C) \leq \text{Rang}\ (A)$. Durch Transposition folgt auch

$$\begin{array}{rcl} \text{Rang}\ (C) & = & \text{Rang}\ (C^T) \\ & = & \text{Rang}\ (B^T \cdot A^T) \\ & \leq & \text{Rang}\ (B^T) \\ & = & \text{Rang}\ (B) \end{array}$$

und insgesamt folgt die Behauptung. □

Die Verschärfung

$$\text{Rang}\ (C) = \min(\ \text{Rang}\ (A),\ \text{Rang}\ (B)). \tag{4.73}$$

heben wir auf für später.

4.5.5 Lineare Gleichungen

An dieser Stelle sollten wir beginnen, über lineare Gleichungen und Gleichungssysteme zu reden. Im Vektorraum $\mathbb{K}^n$ ist eine **lineare Gleichung** durch die Forderung

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b \tag{4.74}$$

gegeben, wobei die Skalare $b, a_1, \ldots, a_n$ fest gegeben sind und man einen Vektor $x = (x_1, \ldots, x_n)^T \in \mathbb{K}^n$ mit (4.74) sucht. Mehrere gleichzeitig zu erfüllende lineare Gleichungen schreibt man als **lineares Gleichungssystem**

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \ldots & + & a_{mn}x_n & = & b_m \end{array} \tag{4.75}$$

oder eleganter in Matrixform

$$A \cdot x = b \tag{4.76}$$

mit $A = (a_{jk}) \in \mathbb{K}^{m\times n}$, $b \in \mathbb{K}^m$, $x \in \mathbb{K}^n$. Nun ist allerdings der Vektorraum $\mathbb{K}^n$ sehr speziell, und man sollte besser eine feste lineare Abbildung $T : U \to V$ nehmen und von Vektoren $u \in U$ verlangen, daß sie mit einem fest gegebenen $v \in V$ die Gleichung

$$T(u) = v \tag{4.77}$$

erfüllen. Dies verallgemeinert (4.74) und (4.75) auf ganz natürliche Weise.

Definition 4.78 *Eine lineare Abbildungsgleichung* (4.77) *heißt* **homogen**, *wenn die rechte Seite v die Null ist, andernfalls* **inhomogen**.

Theorem 4.79 *Ist $T : U \to V$ eine lineare Abbildung zwischen Vektorräumen über demselben Grundkörper $\mathbb{K}$, so kann man über die Lösbarkeit der Gleichung* (4.77) *folgendes sagen:*

1. *Die Menge der Lösungen einer homogenen Gleichung* (4.77) *ist nie leer und immer gleich dem linearen Unterraum* $\ker T$ *von U.*
2. *Die Menge der Lösungen einer inhomogenen Gleichung* (4.77) *ist leer oder ein affiner Unterraum von U.*
3. *Für ein festes $v \in V$ ist die Gleichung lösbar, wenn v im Bildraum $T(U)$ liegt.*

4. *Für alle $v \in V$ ist die Gleichung genau dann lösbar, wenn T surjektiv ist.*

5. *Zwei Lösungen einer festen Abbildungsgleichung unterscheiden sich um eine Lösung des homogenen Systems.*

6. *Man bekommt alle Lösungen einer festen inhomogenen Abbildungsgleichung, indem man zu einer speziellen Lösung der inhomogenen Gleichung beliebige Lösungen der homogenen Gleichung addiert.*

7. *Eine Lösung u einer einzelnen Gleichung der Form* (4.77) *ist genau dann eindeutig, wenn T injektiv ist.*

8. *Ist T injektiv, so sind alle Gleichungen der Form* (4.77) *bei beliebigem v eindeutig lösbar, sofern sie überhaupt lösbar sind.*

9. *Die Gleichung ist genau dann für alle $v \in V$ eindeutig lösbar, wenn T bijektiv ist.*

Beweis: Die Aussagen 1 bis 6 sind klar. Bei den beiden nächsten achte man auf die Formulierung: ist auch nur eine einzige Gleichung nicht eindeutig lösbar, so ist T nicht injektiv und es sind **alle** Gleichungen nicht eindeutig lösbar. Gelten nämlich für ein festes $v \in V$ die Gleichungen $T(u_1) = v = T(u_2)$ für zwei verschiedene $u_1, u_2 \in U$, so folgt $T(u_1 - u_2) = 0$ und T ist nicht injektiv. Dann kann aber auch jede andere Gleichung $T(u) = w$ nicht eindeutig lösbar sein, weil auch $T(u+u_1-u_2) = w$ gilt. Die "Uneindeutigkeit" besteht immer aus dem kompletten Kern von T, egal ob man eine oder alle Gleichungen betrachtet. Damit sind dann aber auch die letzten Aussagen unmittelbar einsichtig. □

Wichtig ist der Spezialfall eines linearen Funktionals λ auf einem Vektorraum V. Ist λ das Nullfunktional (d.h. $\lambda(v) = 0$ für alle $v \in V$), so ist der Bildraum $\{0\}$, der Kern ist V, und der Quotientensatz wird trivial: $V/V \simeq \{0\}$. Ist das Funktional nicht das Nullfunktional, so folgt $\mathbb{K} \simeq V/\ker\lambda$, d.h. der Kern des Funktionals hat Codimension 1. Hat V die Dimension n, so hat $V/\ker\lambda$ die Dimension $n-1$, und man bekommt eine **Hyperebene** durch den Nullpunkt.

Betrachten wir eine Gleichung (4.74), so ist die lineare Abbildung

$$T \;:\; x \mapsto a^T x$$

ein Funktional, und wenn $a \neq 0$ gilt, ist der lineare Raum $\ker T$ eine Hyperebene durch den Nullpunkt (im $\mathbb{R}^3$ eine Ebene durch den Nullpunkt, im

$\mathbb{R}^2$ eine Gerade durch den Nullpunkt). Der Raum der inhomogenen Lösungen ist ein affiner Unterraum, der aus einer festen inhomogenen Lösung und Addition von beliebigen homogenen Lösungen besteht. Er ist eine allgemeine Hyperebene (Ebene, Gerade im $\mathbb{R}^3$ bzw. $\mathbb{R}^2$).

Ein lineares Gleichungssystem (4.75) im $\mathbb{K}^n$ beschreibt die Schnittmenge von m Hyperebenen im $\mathbb{K}^n$, Es wird vermittelt durch eine Matrix $A \in \mathbb{K}^{m\times n}$, und wir sehen uns jetzt an, was Satz 4.79 dann besagt:

Theorem 4.80 *Ein inhomogenes lineares Gleichungssystem* (4.76) *mit einer $m \times n$–Matrix A ist genau dann für alle rechten Seiten lösbar, wenn*

$$m = Rang(A) \leq n$$

gilt. Eindeutigkeit der Lösungen des homogenen oder inhomogenen Systems hat man genau dann, wenn

$$n = Rang(A) \leq m$$

gilt. Allgemeine und eindeutige Lösbarkeit hat man genau dann, wenn der Rang von A gleich n und m ist.

4.6 Lineare Algebra in der Praxis

4.6.1 Speichertechnik

Das Rechnen mit reellen Zahlen wird normalerweise in Programmiersprachen wie C oder Java durch den Datentyp `double` und seine Standardoperationen ausgeführt. Aber wie arbeitet man mit Vektoren und Matrizen?

Seit den Anfängen des elektronischen Rechnens verwendet man dazu intern den indizierten Speicherzugriff und nutzt die wortweise linear adressierbare Struktur des Speichers des von–Neumann–Rechners aus. Vektoren werden im effizientesten Idealfall also im Speicher durch lückenlos aneinandergereihte `double`–Zahlen dargestellt. Man nennt diese indizierten Datentypen **Arrays**, während der Begriff *Vektor* in objektorientierten Sprachen wie Java für eine abstrahierte Klasse steht, die es erlaubt, mit Indexzugriff und dynamischer Speicherverwaltung auf geordnete Listen von Objekten zuzugreifen. Im numerischen Rechnen sind *arrays* immer vorzuziehen, weil Vektor–Klassen eine zusätzliche Dereferenzierung erfordern. Wir gehen im folgenden immer davon aus, daß Vektoren als *arrays* gespeichert sind.

Weil bei heutigen Rechnern komplizierte hierarchische Speicherzugriffs– und Verarbeitungsmethoden (**Paging**, **Cache**, **Pipelining**) fest implementiert

sind, sollten alle Zugriffe auf Vektoren oder *arrays* **datenlokal** ablaufen, d.h. immer auf im Speicher unmittelbar benachbarte Zahlen zugreifen.

Das läßt sich bei Vektoren relativ einfach machen, bei Matrizen aber nicht, denn der von–Neumann–Rechner hat keinen zweidimensionalen Speicher. Man muß Matrizen intern als Vektoren speichern, und das kann man entweder zeilen– oder spaltenweise tun. Eine Matrix $A = (a_{jk}) \in I\!R^{m \times n}$ kann man vektoriell entweder zeilenweise als

$$(a_{11}, a_{12}, \ldots, a_{1n}, a_{21}, a_{22}, \ldots, a_{2n}, \ldots, a_{m1}, a_{m2}, \ldots, a_{mn})$$

oder spaltenweise als

$$(a_{11}, a_{21}, \ldots, a_{m1}, a_{12}, a_{22}, \ldots, a_{m2}, \ldots, a_{1n}, a_{2n}, \ldots, a_{mn})$$

speichern. Aber schon bei der Matrixmultiplikation $C = A \cdot B$ sieht man das hier versteckte Problem: man muß die Zeilen von A mit den Spalten von B skalar multiplizieren, und das geht nur dann datenlokal und ohne Tricks, wenn man A zeilenweise und B spaltenweise speichert. Weil die einzelnen Programmierumgebungen aber die Speichertechnik für Matrizen fest definieren (in FORTRAN und MATLAB wird spaltenweise gespeichert, in C und JAVA zeilenweise), muß man zu mathematisch–informatischen Tricks greifen, die hier kurz erwähnt werden sollen. Dabei gehen wir davon aus, daß Vektoradditionen und Skalarprodukte $x^T y = y^T x = (x, y)_2$ sich problemlos berechnen lassen.

4.6.2 Matrix–Vektor–Multiplikation

Zu berechnen sei der Vektor $z = Ax \in I\!R^m$ als Produkt einer $m \times n$–Matrix A mit einem Vektor $x \in I\!R^n$. Bei zeilenweiser Speicherung von A gibt es keine Probleme, weil man die Komponenten von Ax gemäß $e_k^T Ax = (e_k^T A)x,\ 1 \leq k \leq m$ als Folge von Skalarprodukten von x mit den Zeilen $e_k^T A$ von A ausrechnen kann. Bei spaltenweiser Speicherung von A verwendet man die **column–sweep–Methode**

$$z = Ax = \sum_{j=1}^{n} (Ae_j e_j^T)x = \sum_{j=1}^{n} e_j^T x \cdot Ae_j = \sum_{j=1}^{n} x_j \cdot Ae_j$$

d.h. man summiert die Spalten von A auf, nachdem man sie jeweils mit den Faktoren $x_1, x_2, \ldots, x_n$ multipliziert hat. Vom theoretischen Aufwand her sind die beiden Formen gleich, auf konkreten Rechnern kann das Laufzeitverhalten aber sehr unterschiedlich sein, insbesondere dann, wenn der Speicherbedarf der Matrix den Umfang des Cache oder des physikalischen Hauptspeichers übersteigt.

4.6.3 Matrizenmultiplikation

Will man eine $\ell \times m$–Matrix $A = (a_{ik})$ mit einer $m \times n$–Matrix $B = (b_{kj})$ multiplizieren, so erfordert die naive Vorgehensweise eine zeilenweise Speicherung von A und eine spaltenweise Speicherung von B. Man kann das Matrizenprodukt aber bei zeilenweiser Speicherung umschreiben in

$$e_i^T C = e_i^T AB = e_i^T A \sum_{k=1}^{m} e_k e_k^T B = \sum_{k=1}^{m} \underbrace{e_i^T A e_k}_{=a_{ik}} \cdot e_k^T B, \tag{4.81}$$

weil man B als Summe seiner Zeilen

$$B = \sum_{k=1}^{m} e_k e_k^T B$$

darstellen kann. In (4.81) hat man dann eine Summation von skalierten Zeilen von B, um die Zeilen von C auszurechnen. Ganz analog geht das bei spaltenweiser Organisation:

$$Ce_j = ABe_j = \underbrace{\left(\sum_{k=1}^{m} Ae_k e_k^T \right)}_{=A} Be_j = \sum_{k=1}^{m} \underbrace{e_k^T B e_j}_{=b_{kj}} \cdot Ae_k,$$

d.h. die Spalten von C sind gewichtete Summen der Spalten von A.

Bei einigermaßen trickreicher Programmierung läßt sich bei der Matrizenmultiplikation einiges an Geschwindigkeit herausholen. Dazu gibt es ein Programm und eine zugehörige Ausgabe.

4.6.4 Dünn besetzte Matrizen und Vektoren

In praktischen Anwendungen treten nicht selten gigantische Matrizen auf, die allerdings sehr viele Nullen enthalten. Man nennt solche Matrizen **dünn besetzt** oder engl. **sparse**. Man speichert dann die einzelnen Spalten oder Zeilen als dünn besetzte Vektoren, je nach zeilen– oder spaltenweiser Speichertechnik der Matrizen. Und von einem dünn besetzten Vektor $V \in \texttt{double}^N$ speichert man in einem `double`–*array* $v \in \texttt{double}^n$ nur die $n << N$ von Null verschiedenen Komponenten. Deren Indizes hat man dann anderswo zu speichern. Man könnte einfach die Indizes in ein weiteres `int`–*array* $I \in \texttt{int}^n$ der Länge n setzen und dann mit $v_j = V_{I(j)},\ 1 \le j \le n$ die von Null verschiedenen Komponenten von V durchlaufen. Der Zugriff auf eine einzelne Komponente V_k ist dann zwar nicht so einfach, tritt aber viel seltener auf als

das Durchlaufen des ganzen Vektors. Um bei den Indizes Speicherplatz zu sparen, speichert man statt der Indizes in der Regel nur die **offsets** oder Indexsprünge $J(j) := I(j+1) - I(j)$ bis zum nächsten nicht verschwindenden Element. Wenn man den Index des ersten nichtverschwindenden Elements hat, kann man sich damit leicht durch den Vektor "durchhangeln" und hat stets Datenlokalität.

4.6.5 Programmpakete

Es sollte nach diesen Bemerkungen klar sein, daß hocheffiziente Verfahren zum Rechnen mit großen Vektoren und Matrizen sehr sorgfältig konzipiert und implementiert sein müssen. Anfänger sollten die Finger davon lassen und sich auf bewährte Programmpakete stützen. Unter `http://www.netlib.org` findet man solche Pakete. Grundlage ist

> BLAS (Basic Linear Algebra Subprograms)
> `http://www.netlib.org/blas/faq.html`

in FORTRAN mit einem C–Interface. Das Projekt

> ATLAS (Automatically Tuned Linear Algebra Software)
> `http://sourceforge.net/projects/math-atlas/`

liefert optimierte Versionen für spezielle Architekturen. Programmpakete, die über die lineare Algebra hinausgehen, sind ohne Anspruch auf Vollständigkeit

- GSL (GNU Scientific Library)
  Eine numerische Freeware–Bibliothek in C und C++ unter der *GNU General Public License*.
  `http://www.gnu.org/software/gsl/`

- IMSL
  Umfassende Fortran-Unterprogrammbibliothek mit vielen numerischen Verfahren in den Bereichen Algebra und Analysis

- IMSL-C/MATH
  Umfangreiche C-Funktionsbibliothek mit Verfahren in den Bereichen Algebra und Analysis

- NAG
  Umfassende Unterprogrammbibliothek für Fortran77, Fortran90 und C mit vielen numerischen Verfahren in den Bereichen Algebra und Analysis

- Numerical Recipes
  Sammlung von Routinen aus Algebra und Analysis als C- und Fortran-Unterprogrammbibliothek.

Diese und andere kann man in Göttingen über
`http://www.gwdg.de/service/software/software-rz/sw_numerisch.html`
abrufen.

4.6.6 MATLAB

Für Projekte, die nicht an die Grenze der Leistungsfähigkeit von Computersystemen gehen, braucht man keine eigene Programmierung auf Ebene der Elemente von Vektoren und Matrizen. Man kann sich auf Programmsysteme wie MAPLE, Mathematica, MATLAB oder MuPAD stützen, die eine eigene Kommandosprache haben, in der man mit Matrizen und Vektoren rechnen kann. Die anderen Systeme sind stärker auf symbolisches Rechnen als auf lineare Algebra ausgerichtet. Deshalb wird hier wird eine kurze Anleitung zur Benutzung von MATLAB gegeben, wobei die praktische Handhabung auf den Göttinger Rechnern im Vordergrund steht.

> **Es wird dringend empfohlen, den folgenden Text direkt am Rechner durchzuarbeiten und die MATLAB–Kommandos sofort auszuprobieren!**

Mit der UNIX–Kommandozeile

```
matlab &
```

auf einem der lokalen Rechner ruft man MATLAB auf. Nach einem schnell verschwindenden Begrüßungsfenster sieht man ein Arbeitsfenster, das u.a. ein *Command Window* enthält, in dem man durch Direkteingabe Kommandos ausführen kann. Man kann aber über die üblichen Menüeinträge (*File/Open*) auch vorgefertigte Kommandosequenzen im *Command Window* ausführen, die man als *m–files* bezeichnet und mit jedem beliebigen ASCII–Texteditor bearbeiten kann.

Hier ist ein simples Beispiel, das im folgenden kommentiert werden soll. Man kann die Befehle einzeln (ohne den mit % beginnenden Kommentarteil) in das jeweilige Kommandofenster eingeben, um zu sehen, was passiert.

```
clear all;        % bereinigt die komplette Vorgeschichte
A=[1 0.2 ; -0.3 4]% eine Matrix mit 2 Zeilen und Spalten
                  % ein Semikolon faengt eine neue Zeile an
```

```
x=[5; -0.6]       % ein Vektor als Spaltenvektor, 2 Komponenten
z=A*x             % Matrix mal Vektor
B=A*A             % Matrix mal Matrix
rank(A)           % Dimension von Zeilen/Spaltenraum
y=1:7             % ein Folgenstueck als Zeile
z=(1:7)'          % dito als Spalte, transponiert
C=0.1*[1:7;2:8;3:9]% eine 3x7-Matrix, skalar multipliziert
C'*C              % liefert eine 7x7-Matrix
C*C'              % liefert eine 3x3-Matrix
D=ones(2,3)       % Matrix mit Einsen, 2x3
E=eye(5)          % Einheitsmatrix, 5x5
F=exp(-z)         % Operationen bilden Matrizen auf Matrizen ab
G=exp(-C)         % und werden komponentenweise ausgerechnet
G(:,3)            % dritte Spalte
G(2,:)            % zweite Zeile
u=A\x             % loest Gleichung A*u=x
x-A*u             % Test
C+C'              % Fehlermeldung
```

Wie alle anderen Systeme dieser Art arbeitet auch MATLAB als dynamischer Interpreter, d.h. das "Wissen" von MATLAB und die Nutzung des internen Speichers hängt von der Vorgeschichte ab. Die Zuweisung

```
A=B
```

weist dem Bezeichner `A` die Bedeutung zu, die vorher dem Bezeichner `B` zukam. Die vorherige Bedeutung des Bezeichners `A` ist verloren. Eine weitere Zuweisung

```
A=C
```

überschreibt dies durch die Bedeutung des Bezeichners `C`. Das im Beispiel zuerst auftretende Kommando

```
clear all;
```

bereinigt die komplette Vorgeschichte und löscht alle Bedeutungen von Bezeichnern. Das ist zu Beginn eines neuen und unabhängigen *m–files* sinnvoll, damit auch der bisher reservierte Speicher (*workspace* in MATLAB) freigegeben wird. Man kann sich übrigens in einem über das Hauptfenster aufrufbaren Teilfenster stets den aktuellen *workspace* und seine Nutzung ansehen.

Die Kommandostruktur von MATLAB ist zeilenorientiert (ein Zeile = ein Kommando), wobei man mit ... auf eine Verlängerungszeile gehen kann,

wenn nötig. Wenn man ein Kommando mit einem Semikolon abschließt, wird die Ausgabe unterdrückt. Das ist bei großen Matrizen und Vektoren lebenswichtig.

In MATLAB sind alle "normalen" Objekte Matrizen von `double`–Zahlen, wobei Vektoren als Matrizen mit einer Spalte, n–Tupel als Matrizen mit einer Zeile und Skalare als 1×1–Matrizen aufgefaßt werden. Im Normalfall arbeiten alle Operationen auf **kompletten Matrizen**. Sonderfälle muß man speziell behandeln. Man tut gut daran, dieses Grundkonzept nicht künstlich zu verwässern, indem man statt mit Vektoren und Matrizen zu arbeiten, auf deren Komponenten zurückgeht. Schleifenprogrammierung ist möglich, sollte aber wie die Pest vermieden werden, wenn sie sich nicht auf komplette Matrizen bezieht.

Kleine Matrizen kann man in MATLAB direkt eingeben, indem man z.B. die Matrix

$$A = \begin{pmatrix} 1 & 0.2 \\ -0.3 & 4 \end{pmatrix}$$

als

```
A=[1 0.2 ; -0.3 4]
```

spezifiziert und sofort ausgibt (kein Semikolon als Abschluß). Die Eingabe geschieht innerhalb der Klammern [ ] zeilenweise mit Leerzeichen als Trennzeichen, wobei das Semikolon eine neue Zeile beginnt. Dann ist auch klar, was

```
x=[5; -0.6]      % ein Vektor als Spaltenvektor, 2 Komponenten
```

bewirkt. Ganz gemäß der MATLAB–Philosophie kann man, wenn die Größen stimmen, in dieser Klammernotation auch Matrizen einsetzen, um z,B. `[A -A]` oder `[A ; 7 -2.3]` zu bilden.

Die drei Kommandos

```
z=A*x            % Matrix mal Vektor
B=A*A            % Matrix mal Matrix
rank(A)          % Dimension von Zeilen/Spaltenraum
```

zeigen, wie einfach man nun in MATLAB mit Matrizen und Vektoren umgeht. Dabei kann man die üblichen Operationen $+, -, *$ zwischen Matrizen verwenden, muß aber aufpassen, ob die Operationen bei den vorliegenden Matrixgrößen auch ausführbar sind, sonst erfolgt eine Fehlermeldung. Es gibt

allerdings eine sehr praktische Sonderregelung, wenn einer der Operanden skalar ist. Dann wird die Operation als komponentenweise Skalaroperation ausgeführt. So kann man Konstanten zu Matrizen addieren oder Matrizen mit festen Faktoren komponentenweise multiplizieren. Der Zugriff auf Matrixelemente erfolgt über Indizes in runden Klammern, z.B. mit `A(1,2)` auf A_{12}, sollte aber nur im absoluten Notfall benutzt werden.

Zum Erzeugen von Standardmatrizen gibt es die Befehle

```
zeros(m,n)          % Matrix mit Nullen, m x n
ones(m,n)           % Matrix mit Einsen, m x n
eye(n)              % Einheitsmatrix, n x n
```

und man kann natürlich auch größere Matrizen aus Dateien einlesen, indem man das Kommando `load` verwendet (man gebe im Kommandofenster `help load` ein, um die genaue Syntax zu sehen).

Einer der wichtigsten Operatoren in MATLAB ist der Doppelpunkt oder *colon*–Oberator. Steht er zwischen Skalaren, so erzeugt er Zeilenvektoren von Werten:

```
3:7                 % liefert [3 4 5 6 7]
4:2:9               % liefert [4 6 8]
0:0.15:1            % liefert [0 0.15 0.3 0.45 0.6 0.75 0.9]
```

Das ist extrem hilfreich zur Erzeugung von Wertetabellen, denn Funktionen wie `sin` arbeiten immer komponentenweise auf kompletten Vektoren oder Matrizen:

```
sin(0:0.15:1)       % liefert die Sinuswerte auf
                    % [0 0.15 0.3 0.45 0.6 0.75 0.9]
```

Und wer gerne etwas Graphisches sehen möchte, sollte

```
x=0:0.01:2*pi;
plot(x,sin(x),x,cos(x))
```

versuchen, aber das Semikolon nicht vergessen.

Man kann die *colon*–Notation auch sehr gut auf Indexbereiche anwenden. Zum Beispiel kann man die obere linke 2×3–Teilmatrix aus einer Matrix `A` herausholen und nach `B` speichern mit

```
B=A(1:2,1:3)
```

Aber der Doppelpunkt kann auch als Platzhalter mit der Bedeutung "für alle" stehen. Ist etwa `A` eine $m \times n$–Matrix in MATLAB, so ist `A(:,3)` die dritte Spalte und `A(2,:)` die zweite Zeile von `A`.

Die Transposition einer Matrix wird durch ein nachgestelltes Apostroph bewirkt, z.B. in

```
C=0.1*[1:7;2:8;3:9]  % eine 3x7-Matrix, skalar multipliziert
C'*C                 % liefert eine 7x7-Matrix
C*C'                 % liefert eine 3x3-Matrix
```

Manchmal sind auch noch die komponentenweisen Skalaroperationen von MATLAB nützlich. Wenn `A`= (a_{jk}) und `B`= (b_{jk}) zwei Matrizen gleicher Größe sind, so besteht `A.*B` aus der Matrix $(a_{jk} \cdot b_{jk})$. Analog ist `A./B` definiert.

Aber die Stärke von MATLAB liegt in der einfachen Verfügbarkeit höherer Operationen, die das Lösen von Gleichungssystemen, das Bestimmen des Rangs oder des Kerns von Matrizen erlauben. Der Rang von `A` wird mit `rank(A)` abgerufen, während eine Orthonormalbasis des Kerns mit `null(A)` und des Bildes mit `orth(A)` produziert wird. Hier fließt die rechengenauigkeitsbedingte Unsicherheit des Rangentscheids ein (vgl. Abschnitt 6.4 auf Seite 190). Die Lösung x eines linearen Gleichungssystems $A \cdot x = b$ bekommt man einfach mit

```
x=A\b                % löst Ax=b
```

unter unsichtbarer Verwendung des Gaußschen Eliminationsverfahrens mit Pivotisierung, sofern die Voraussetzungen für die Lösbarkeit gegeben sind. Aber man kann auch eine QR–Zerlegung von `A` nach Householder mit

```
[Q,R]=qr(A)    % A=Q*R, Q orthogonal, R obere Dreiecksmatrix
```

bekommen. Varianten dieses Kommandos sieht man nach Eingabe von `help qr` im Kommandofenster.

Das soll hier erst einmal genügen. Es wird dringend empfohlen, mit MATLAB herumzuspielen.

5 Räume mit metrischer Struktur

Bisher haben wir nur die Vektorraumstruktur benutzt, d.h. die Addition von Vektoren und die Skalarmultiplikation Jetzt führen wir Abstandsbegriffe ein, die man einerseits für Grenzprozesse und andererseits für weitergehende geometrische Sachverhalte braucht.

5.1 Metriken und Normen

Es wird jetzt Zeit, in Vektorräumen oder allgemeinen Mengen "Geometrie" zu treiben, und das heißt wörtlich "Erdvermessung". Man sollte dazu mindestens den Abstand von Punkten "messen" können. Dazu ist nicht unbedingt eine Vektorraumstruktur nötig, es würde ein Abstandsbegriff reichen.

Definition 5.1 *Eine* **Metrik** *auf einer Menge M ist eine Abbildung*

$$d \ : \ M \times M \to I\!R$$

mit den Eigenschaften

$$\begin{array}{lcl} d(x,x) & = & 0 \\ d(x,y) & \geq & 0 \\ d(x,y) & = & 0 \textit{ impliziert } x = y \\ d(x,y) & = & d(y,x) \\ d(x,z) & \leq & d(x,y) + d(y,z) \quad \textit{(Dreiecksungleichung)} \end{array}$$

für alle $x, y, z \in M$. Dann heißt M mit d ein **metrischer Raum**[1][2].

Insbesondere für das "Messen" auf diskreten Strukturen ist der Begriff der Metrik hilfreich, denn dort hat man keine Vektorraumeigenschaften. Zum Beispiel braucht man in der Codierungstheorie auf den Binärwörtern in B^n die **Hamming–Distanz**[3]

$$d((b_1, \ldots, b_n), (c_1, \ldots, c_n)) := \sum_{j=1}^{n} |b_j - c_j|$$

[1] `http://de.wikipedia.org/wiki/Metrischer_Raum`

[2] Südlich von Göttingen gibt es einen Hügel ("die Gleichen") mit zwei Kuppen. David Hilbert soll seine Studenten immer gefragt haben, warum "die Gleichen" so heißen. Sie sind aber weder gleich hoch noch sehen sie gleich aus usw. und Hilberts Antwort ist: "Weil sie gleichen Abstand voneinander haben!" Siehe dazu die Eigenschaft $d(x,y) = d(y,x)$ der Metrik.

[3] `http://de.wikipedia.org/wiki/Hamming-Abstand`

welche die Anzahl der verschiedenen Bits von $(b_1, \ldots, b_n)$ und $(c_1, \ldots, c_n)$ angibt.

Aufgabe: Warum ist das eine Metrik?

Auf der Erdkugel–Oberfläche ist der kürzeste Abstand zweier Punkte gleich der von einem Piloten geflogenen Luftlinie auf einem Großkreis. Die Kugeloberfläche ist kein Vektorraum, aber dennoch bildet sie mit dieser **sphärischen Metrik** einen metrischen Raum. Im engeren Sinne ist die "Geo"–Metrie also gar keine Vektorraumgeometrie, sondern eine Geometrie im metrischen Raum. Man kann Dreiecke definieren, aber die Winkelsumme ist nicht gleich 180 Grad.

Wir haben schon Kenntnisse über Vektorräume, und deshalb wollen wir die Vektorraumstruktur benutzen. Nach der allgemeinen Definition von Vektorräumen kann man aber den Vektoren eines Vektorraums nicht ohne Zusatzvoraussetzungen eine "Länge" zuweisen, und zwei "Punkte" u und v haben nicht notwendig einen Abstand. Natürlich könnte man den Abstand über die Länge als

$$\text{Abstand } (u, v) = \text{ Länge } (u - v) = \text{ Länge } (v - u)$$

einführen, aber die Länge ist eben nicht definiert.

Für Vektoren x des $\mathbb{R}^n$ geht das aber, indem man als Länge

$$\|x\|_2 := \|(x_1, \ldots, x_n)^T\| := \sqrt{\sum_{j=1}^{n} x_j^2} \quad \text{für alle } x \in \mathbb{R}^n$$

setzt. Man mache sich diese Formel im $\mathbb{R}^2$ und $\mathbb{R}^3$ klar. Im Komplexen sollte man analog

$$\|x\|_2 := \|(x_1, \ldots, x_n)^T\| := \sqrt{\sum_{j=1}^{n} x_j \overline{x_j}} = \sqrt{\sum_{j=1}^{n} |x_j|^2} \quad \text{für alle } x \in \mathbb{C}^n$$

nehmen. Diese Längenbegriffe erfüllen die folgende

Definition 5.2 *Es sei V ein Vektorraum über $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$.*

1. *Eine* **Norm** *auf V ist eine Abbildung*

$$\|.\| \ : \ V \to \mathbb{R}, \ \ v \mapsto \|v\| \in \mathbb{R} \text{ für alle } v \in V$$

mit den Eigenschaften

$$\begin{array}{rcl} \|v\| & \geq & 0 \\ \|v\| & = & 0 \text{ } \textit{impliziert } v = 0 \\ \|\alpha v\| & = & |\alpha|\|v\| \\ \|u+v\| & \leq & \|u\| + \|v\| \quad \textit{(Dreiecksungleichung)} \end{array}$$

für alle $\alpha \in \mathbb{K}$ und $u, v \in V$.

2. *Ein Vektorraum heißt* **normiert**[1]*, enn auf ihm eine Norm definiert ist.*

Man sieht, daß man hier im Skalarenkörper $\mathbb{K}$ den Absolutbetrag braucht, und deshalb verwenden wir für den Rest des Kapitels immer $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$.

Bis auf die Dreiecksungleichung[2], die wir erst später beweisen wollen, sollte klar sein, daß die obigen Fälle im $\mathbb{R}^n$ oder $\mathbb{C}^n$ Normen definieren. Es gibt aber noch andere Möglichkeiten, z.B.

$$\begin{array}{rcl} \|x\|_\infty & := & \max\{|x_j| \ : \ 1 \leq j \leq n\} \\ \|x\|_p & := & \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p} \end{array} \tag{5.3}$$

für alle $x \in \mathbb{K}^n$ und alle $p \in [1, \infty)$. Die Dreiecksungleichung ist im ersten Fall relativ einfach, im zweiten Fall deutlich schwieriger zu beweisen. Den zweiten Fall lassen wir erst einmal weg, und im ersten Fall benutzen wir die skalare Dreiecksungleichung:

$$\begin{array}{rcl} \|x+y\|_\infty & = & \max\limits_{1\leq j\leq n} |x_j + y_j| \\ & \leq & \max\limits_{1\leq j\leq n} (|x_j| + |y_j|) \\ & \leq & \max\limits_{1\leq j\leq n} |x_j| + \max\limits_{1\leq j\leq n} |y_j| \\ & = & \|x\|_\infty + \|y\|_\infty. \end{array}$$

Im $\mathbb{K}^n$ ist die **Einheitskugel** gegeben durch

$$\{x \in \mathbb{K}^n \ : \ \|x\|_2 \leq 1\},$$

und die **Einheitssphäre** ist

$$\{x \in \mathbb{K}^n \ : \ \|x\|_2 = 1\}.$$

[1] `http://de.wikipedia.org/wiki/Normierter_Raum`
[2] `http://de.wikipedia.org/wiki/Dreiecksungleichung`

Ersetzt man die Norm $\|.\|_2$ durch eine allgemeinere Norm $\|.\|_p$, $1 \leq p \leq \infty$, so bekommt man "p–Einheitskugeln". In Abbildung 1 sieht man das im Falle des $\mathbb{R}^2$. Von innen nach außen sieht man die "Kugeln" (in diesem Falle besser "Kreise") für $p = 1,\ 1.5,\ 2,\ 5,\ 2000$. Dazu gibt es ein MATLAB-Programm.

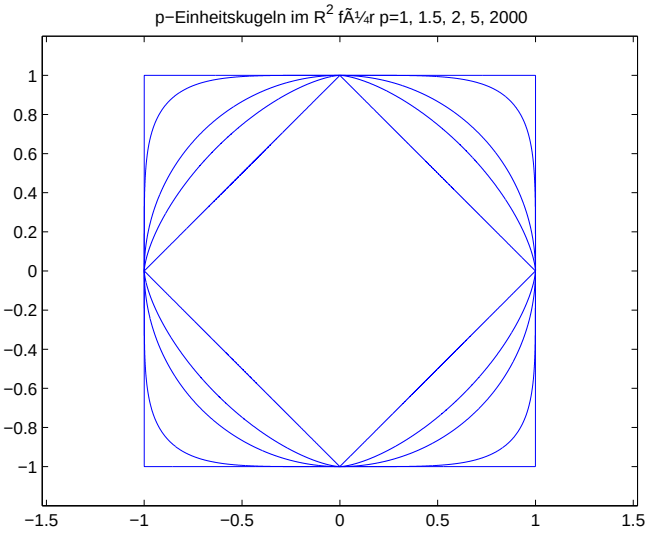


Abbildung 1: p–Einheitskugeln

Für viele Anwendungen ist folgende einfache Abschätzung des Skalarproduktes wichtig:

Theorem 5.4 *Für beliebige Vektoren $x, y \in \mathbb{K}^n$ gilt*

$$|x^T y| = \left| \sum_{j=1}^{n} x_j y_j \right| \leq \sum_{j=1}^{n} |x_j||y_j| \leq \max_{1 \leq k \leq n} |x_k| \sum_{j=1}^{n} |y_j| = \|x\|_\infty \cdot \|y\|_1.$$

Wir unterdrücken hier den Beweis der allgemeineren **Hölder–Minkowski–Ungleichung**

$$|x^T y| \leq \|x\|_p \cdot \|y\|_q$$

für alle $x, y \in \mathbb{K}^n$ und für alle $p, q \in [1, \infty]$ mit $\frac{1}{p} + \frac{1}{q} = 1$.

Man sollte sich klarmachen, daß Normen Abstandsbegriffe sind, die in der Praxis sehr wichtige und unterschiedliche Bedeutung haben. Ist etwa $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ ein Vektor, dessen Komponenten Kostenanteile an einem

Produkt sind, so gibt $\|x\|_1$ die Gesamtkosten des Produkts und $\|x\|_\infty$ die Kosten des teuersten Anteils an. Ist $x \in \mathbb{R}^2$ ein Vektor der Ebene, so ist $\|x\|_2$ der normale euklidische Luftlinienabstand zum Nullpunkt, während man $\|x\|_1$ als "Taxifahrerabstand" zum Nullpunkt ansehen kann (Skizze in der Vorlesung).

Noch interessanter wird es, wenn man in **unendlichdimensionalen** Vektorräumen Normen einführt. Das kann man zum Beispiel auf dem Raum

$$\mathbb{R}_0^{\mathbb{N}} := \{f \ : \ \mathbb{N} \to \mathbb{R}, \ f(n) \neq 0 \text{ nur für endlich viele } n \in \mathbb{N}\}$$

machen, indem man ganz analog

$$\begin{aligned} \|f\|_\infty &:= \max\{|f(j)| \ : \ j \in \mathbb{N}\} \\ \|f\|_p &:= \left(\sum_{j=1}^{n} |f(j)|^p\right)^{1/p} \end{aligned}$$

für alle $f \in \mathbb{R}_0^{\mathbb{N}}$ definiert. Im Vorgriff auf Späteres definiert man im Raum $C[a,b]$ der stetigen reellwertigen Funktionen auf einem Intervall $[a,b] \subset \mathbb{R}$ die Normen

$$\begin{aligned} \|f\|_\infty &:= \max\{|f(t)| \ : \ a \leq t \leq b\} \\ \|f\|_p &:= \left(\int_a^b |f(t)|^p dt\right)^{1/p}. \end{aligned}$$

Dies soll erst einmal als Beispielsammlung reichen. Wichtig ist dabei nur, daß die Normeigenschaften gelten, und daß man keinesfalls bei der Normdefinition auf simple Räume wie $\mathbb{R}^n$ oder $\mathbb{C}^n$ eingeschränkt ist.

Man bekommt aus normierten Räumen immer auch metrische Räume:

Theorem 5.5 *Ist $\|.\|$ eine Norm auf einem Vektorraum V, so ist V mit*

$$d(u,v) := \|u - v\| \text{ für alle } u, v \in V$$

ein metrischer Raum.

5.2 Normäquivalenz

Es ist hier und auch im nächsten Kapitel wichtig zu wissen, wie sehr sich die auf einem festen Vektorraum möglichen Normen unterscheiden können.

Definition 5.6 *Es sei V ein Vektorraum über einem Grundkörper $\mathbb{K}$. Zwei Normen $\|.\|_A$ und $\|.\|_B$ auf V heißen* **äquivalent**, *wenn es positive reelle Konstanten c, C gibt mit*

$$c \cdot \|v\|_A \leq \|v\|_B \leq C \cdot \|v\|_A \text{ für alle } v \in V.$$

Es ist einfach zu beweisen (wie?), daß man dadurch eine Äquivalenzrelation auf der Menge der Normen auf V hat. Obwohl der folgende Satz auch in allgemeinen endlichdimensionalen Vektorräumen gilt (er ist in unendlichdimensionalen nicht richtig), beweisen wir ihn in diesem Text nur für einen Spezialfall, und in diesem Abschnitt nur zur Hälfte:

Theorem 5.7 *Auf den Vektorräumen $\mathbb{R}^n$ sind alle Normen äquivalent.*

Beweis: Wir haben bisher nur $\|.\|_\infty$ als Norm auf $\mathbb{R}^n$ nachgewiesen. Jetzt sei $\|.\|$ eine beliebige andere Norm und $x \in \mathbb{R}^n$ vorgegeben. Mit den Normeigenschaften und Theorem 5.4 folgt

$$\begin{aligned} \|x\| &= \left\| \sum_{j=1}^{n} x_j e_j \right\| \\ &\leq \sum_{j=1}^{n} |x_j| \, \|e_j\| \\ &\leq \|x\|_\infty \sum_{j=1}^{n} \|e_j\| \end{aligned}$$

und man setzt $C := \sum_{j=1}^{n} \|e_j\|$ um $\|x\| \leq C \cdot \|x\|_\infty$ zu erhalten. Die umgekehrte Ungleichung 8.21 sparen wir uns für das nächste Kapitel auf. Wir haben also, daß die Norm $\|.\|_\infty$ zu allen anderen Normen äquivalent ist. Dann sind wegen der Transitivität alle Normen auf $\mathbb{R}^n$ äquivalent. □

An dieser Stelle sollte man die einfachsten Normäquivalenzkonstanten im $\mathbb{R}^n$ oder $\mathbb{C}^n$ angeben:

$$\begin{array}{rcrll} \|x\|_p & \leq & \sqrt[p]{n} \cdot & \|x\|_\infty & \text{für alle } p \geq 1 \\ \|x\|_\infty & \leq & 1 \cdot & \|x\|_p & \text{für alle } p \geq 1, \text{ und daraus auch} \\ \|x\|_p & \leq & \sqrt[p]{n} \cdot & \|x\|_q & \text{für alle } p, q \geq 1. \end{array}$$

Aufgabe: Man beweise das. Im dritten Fall gibt es bessere Abschätzungen, aber die sind schwieriger und werden weggelassen.

Theorem 5.8 *Auf endlichdimensionalen Vektorräumen über $\mathbb{R}$ sind alle Normen äquivalent.*

Beweisskizze: Ist V ein n–dimensionaler Vektorraum über $\mathbb{R}$ mit einer beliebigen Norm $\|.\|_V$, so kann man sich eine Basis $\{v_1, \ldots, v_n\}$ verschaffen und den Standard–Isomorphismus

$$T \; : \; \mathbb{R}^n \to V, \; (\alpha_1, \ldots, \alpha_n)^T \mapsto \sum_{j=1}^{n} \alpha_j v_j$$

benutzen, um eine entsprechende Norm

$$\|x\| := \|T(x)\|_V \text{ für alle } x \in \mathbb{R}^n$$

auf $\mathbb{R}^n$ zu definieren. Macht man dies sinngemäß für zwei Normen auf V, so folgt die Äquivalenz dieser Normen aus der Äquivalenz der entsprechenden Normen auf $\mathbb{R}^n$. □

Die obigen Resultate gelten sinngemäß auch für die Räume $\mathbb{C}^n$ und endlichdimensionale Räume über $\mathbb{C}$. Die Beweise werden hier aber nicht in voller Breite ausgeführt. Es sollte reichen, daß man Vektoren des $\mathbb{C}^n$ durch Zerlegung in Real– und Imaginärteil der Koeffizienten als Vektoren des $\mathbb{R}^n \times \mathbb{R}^n$ auffassen kann:

$$\begin{aligned}\sum_{j=1}^{n}(x_j + iy_j)e_j &= \sum_{j=1}^{n} x_j e_j + i \cdot \sum_{j=1}^{n} y_j e_j \\ &\simeq \left(\sum_{j=1}^{n} x_j e_j, \sum_{j=1}^{n} y_j e_j\right).\end{aligned}$$

Dabei stimmen die $\|.\|_2$–Normen überein:

$$\begin{aligned}\left\|\sum_{j=1}^{n}(x_j + iy_j)e_j\right\|_{2,\mathbb{C}^n} &= \sum_{j=1}^{n}(|x_j|^2 + |y_j|^2) \\ &= \left\|\left(\sum_{j=1}^{n} x_j e_j, \sum_{j=1}^{n} y_j e_j\right)\right\|_{2,\mathbb{R}^{2n}}\end{aligned}$$

Zusammen mit dem Standardisomorphismus kann man sich deshalb beim Studium endlichdimensionaler Vektorräume über $\mathbb{R}$ oder $\mathbb{C}$ in vielen Fällen auf den $\mathbb{R}^n$ zurückziehen. Wir werden im folgenden diese Bijektion (etwas lax) ebenfalls als **Standardisomorphismus** zwischen dem $\mathbb{C}^n$ und dem $\mathbb{R}^{2n}$ bezeichnen.

In unendlichdimensionalen Räumen gilt die Normäquivalenz im allgemeinen nicht. Und in der Praxis erlaubt die Normäquivalenz nicht, sich auf eine spezielle Norm zu beschränken, denn die verschiedenen Normen haben ja sehr wichtige praktische Bedeutungen, wie wir oben schon angedeutet haben.

5.3 Innere Produkte

Aber wir gehen zurück zu den Räumen $\mathbb{R}^n$ bzw. $\mathbb{C}^n$ unter der Norm $\|.\|_2$, weil wir in diesem Fall mehr als nur eine Norm bekommen. Wir schreiben

eine Variante des Skalarprodukts zweier Vektoren $x = (x_1, \ldots, x_n)^T$ und $y = (y_1, \ldots, y_n)^T$ als

$$(x, y)_2 := \sum_{j=1}^{n} x_j \overline{y_j} = x^T \overline{y}$$

hin und stellen fest, daß $\|x\|_2^2 = (x, x)_2$ gilt. Im reellen Fall sind die Querstriche natürlich überflüssig. Das ist ein Spezialfall der folgenden Situation:

Definition 5.9 *Eine skalarwertige binäre Abbildung*

$$(.,.) \; : \; V \times V \to I\!K$$

auf einem Vektorraum V über $I\!K = I\!R$ oder $I\!K = \mathbb{C}$ heißt **inneres Produkt** *oder* **Skalarprodukt**[1]*, wenn für alle $x, y \in V$ gilt*

$$\begin{array}{rcl} (x,x) & \geq & 0 \\ (x,x) & = & 0 \quad \textit{impliziert } x = 0 \\ (x,y) & = & \overline{(y,x)} \\ (.,y) & & \textit{ist linear, d.h.} \\ (\alpha u + \beta v, y) & = & \alpha(u,y) + \beta(v,y) \textit{ für alle } \alpha, \beta \in I\!K, \; u, v \in V. \end{array}$$

Aus der Definition folgt sofort auch die **Antilinearität**

$$\begin{array}{rcl} (x, \alpha u + \beta v) & = & \overline{(\alpha u + \beta v, x)} \\ & = & \overline{\alpha(u,x) + \beta(v,x)} \\ & = & \overline{\alpha}\overline{(u,x)} + \overline{\beta}\overline{(v,x)} \\ & = & \overline{\alpha}(x,u) + \overline{\beta}(x,v) \text{ für alle } \alpha, \beta \in I\!K, \; u, v \in V. \end{array}$$

Im reellen Fall bezeichnet man ein inneres Produkt auch als **Bilinearform**, im komplexen Fall als **Sesquilinearform**.

Theorem 5.10 *Hat ein Vektorraum V über $I\!K = I\!R$ oder $I\!K = \mathbb{C}$ ein inneres Produkt $(.,.)$, so ist durch*

$$\|v\| := \sqrt{(v,v)} \textit{ für alle } v \in V$$

eine Norm auf V definiert und es gelten die **Cauchy–Schwarz'sche Ungleichung**[2] [3] [4]

$$|(u,v)| \leq \|u\|\|v\|$$

[1] `http://de.wikipedia.org/wiki/Skalarprodukt`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Cauchy.html`
[3] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Schwarz.html`
[4] `http://de.wikipedia.org/wiki/Cauchy-Schwarzsche_Ungleichung`

und die **Parallelogrammgleichung**[1]

$$\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

für alle $u, v \in V$. Zwei Vektoren $u, v \in V$ heißen **orthogonal**[2]*, wenn $(u, v) = 0$ gilt. Für orthogonale Vektoren u, v gilt der Satz des* **Pythagoras**[3][4]*:*

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2.$$

Beweis: Bis auf die Dreiecksungleichung sind die Normeigenschaften klar. Mit Schulkenntnissen kann man die Cauchy–Schwarz'sche Ungleichung beweisen, indem man für feste Vektoren $u, v \in V$ eine Kurvendiskussion der Funktion

$$\begin{aligned} f(t) &:= \|u + t\cdot v\|^2 \geq 0 \text{ für alle } t \in K \\ &= \|u\|^2 + |t|^2\|v\|^2 + \underbrace{\overline{t}(u,v) + t(v,u)}_{reell} \end{aligned}$$

ausführt. Im komplexen Fall setzt man noch

$$t = r\frac{(u,v)}{|(u,v)|}, \quad |t| = r$$

mt reellem r an, und man kann ohne Einschränkung annehmen, daß $(u, v) \neq 0$ gilt. Es folgt jetzt

$$g(r) := \|u\|^2 + r^2\|v\|^2 + 2r|(u,v)| \geq 0 \text{ für alle } r \in I\!R.$$

Eine quadratische Funktion der Form $ar^2 + 2br + c$ mit $a > 0$ ist genau dann nichtnegativ für alle r, wenn $b^2 \leq ac$ gilt, denn man hat

$$\begin{aligned} ar^2 + 2br + c &= \left(r\sqrt{a} + \frac{b}{\sqrt{a}}\right)^2 - \frac{b^2}{a} + c \\ &\geq -\frac{b^2}{a} + c \end{aligned}$$

und das Minimum wird angenommen. Es ist genau dann nichtnegativ, wenn $b^2 \leq ac$ gilt, und wir bekommen $|(u, v)|^2 \leq \|u\|^2\|v\|^2$.

Die Parallelogrammgleichung ergibt sich durch Ausmultiplizieren:

$$\begin{aligned} \|u+v\|^2 + \|u-v\|^2 &= \|u\|_2 + (u,v) + (v,u) + \|v\|^2 \\ &+ \|u\|_2 - (u,v) - (v,u) + \|v\|^2 \\ &= 2\|u\|^2 + 2\|v\|^2. \end{aligned}$$

[1] `http://de.wikipedia.org/wiki/Parallelogrammgleichung`
[2] `http://de.wikipedia.org/wiki/Orthogonalit%C3%A4t`
[3] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Pythagoras.html`
[4] `http://de.wikipedia.org/wiki/Satz_des_Pythagoras`

Die Dreiecksungleichung folgt aus der Cauchy–Schwarz'schen Ungleichung mit

$$\begin{aligned} \|u+v\|^2 &= \|u\|^2 + \|v\|^2 + (u,v) + (v,u) \\ &\leq \|u\|^2 + \|v\|^2 + 2|(u,v)| \\ &\leq \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| \\ &= (\|u\| + \|v\|)^2 \end{aligned}$$

und die erste Zeile dieses Arguments liefert den Satz des Pythagoras. □

Definition 5.11 *Ein Vektorraum mit innerem Produkt heißt auch* **Prä–Hilbert–Raum**[1] [2]*. Ist der Grundkörper über $\mathbb{K}$ reell, so heißt der Raum* **euklidisch** [3][4]*. Im euklidischen Raum ist der* **Winkel**[5] *$\angle(u,v)$ zwischen zwei Vektoren $u, v \in V \setminus \{0\}$ definiert durch*

$$\cos(\angle(u,v)) = \frac{(u,v)}{\|u\|\|v\|}. \tag{5.12}$$

Das ist motiviert durch den Cosinussatz[6] der ebenen Trigonometrie, der oft auch in der Form

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2\|u\|\|v\|\cos(\angle(u,v))$$

als Verallgemeinerung des Satzes von Pythagoras[7] formuliert wird. Durch Vergleich mit

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2(u,v)$$

ergibt sich dann die Gleichung (5.12). Den Cosinussatz kann man mit elementaren Mitteln beweisen (siehe unten), wenn man schon weiß, was ein Winkel ist. Umgekehrt kann man ihn, wie oben, zur Definition von Winkeln in allgemeinen Vektorräumen über $\mathbb{R}$ mit innerem Produkt verwenden.

Hier kommt ein kleiner Einschub zum Cosinussatz. Man sehe sich Figur 2 an.

---

[1] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Hilbert.html
[2] http://de.wikipedia.org/wiki/Pr%C3%A4hilbertraum
[3] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Euclid.html
[4] http://de.wikipedia.org/wiki/Euklidischer_Raum
[5] http://de.wikipedia.org/wiki/Winkel
[6] http://de.wikipedia.org/wiki/Kosinussatz
[7] http://de.wikipedia.org/wiki/Satz_des_Pythagoras

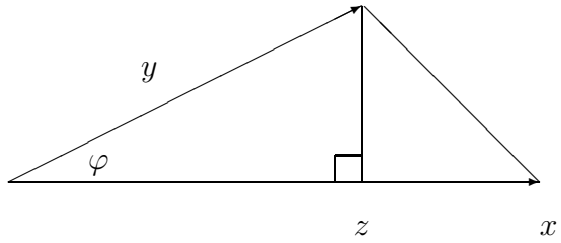


Abbildung 2: Cosinussatz

Der Vektor $y - z$ ist orthogonal zu x, und z ist ein Vielfaches von x. Dann folgt aus dem Ansatz $z = \alpha x$ und

$$\begin{aligned} (y - z, x) &= 0 \\ (y, x) &= (z, x) \\ &= \alpha(x, x) \\ &= \alpha \|x\|_2^2 \end{aligned}$$

notwendig die Gleichung

$$\begin{aligned} z &= \frac{(x, y)}{\|x\|_2^2} x \\ \|z\|_2 &= \frac{|(x, y)|}{\|x\|_2} \end{aligned}$$

und man kann den Satz des Pythagoras anwenden:

$$\begin{aligned} \|y\|_2^2 &= \|z\|_2^2 + \|y\|_2^2 \sin^2 \varphi \\ \|y\|_2^2 \cos^2 \varphi &= \|z\|_2^2 \\ &= \frac{(x, y)_2^2}{\|x\|_2^2} \\ \|y\|_2 \cos \varphi &= \frac{(x, y)_2}{\|x\|_2} \end{aligned}$$

wobei die letzte Gleichung das Vorzeichen bei spitzen und stumpfen Winkeln richtig setzt.

Wir sehen uns noch an, was im euklidischen Raum unter dem üblichen inneren Produkt passiert, wenn man eine Matrix $A = (a_{jk}) \in I\!K^{m \times n}$ hat

und für zwei Vektoren $u \in \mathbb{K}^n$ und $v \in \mathbb{K}^m$ die inneren Produkte

$$\begin{aligned}(Au, v)_2 &= u^T A^T \overline{v} \\ &= u^T \overline{A^* v} \\ &= \overline{(u, A^* v)_2} \\ (v, Au)_2 &= \overline{(Au, v)_2} \\ &= \overline{(u, A^* v)_2} \\ &= (A^* v, u)_2\end{aligned}$$

ausrechnet. Man sieht, daß in beiden Fällen die Matrix auf das andere Argument des inneren Produktes verschoben werden kann, wenn man sie transponiert und zum konjugiert Komplexen übergeht. Diese Rechentechnik ist von großer praktischer Bedeutung und muß unbedingt beherrscht werden.

Ohne hier die Wohldefiniertheit zeigen zu können, geben wir an, wie sich dieser Trick verallgemeinern läßt:

Definition 5.13 *Es sei T eine lineare Abbildung zwischen zwei Prä–Hilbert–Räumen U und V über $\mathbb{K}$ mit inneren Produkten $(.,.)_U$ und $(.,.)_V$. Dann ist die* **Adjungierte**[1][2] *zu T eine Abbildung*

$$T^* \ : \ V \to U$$

mit der Eigenschaft

$$(T(u), v)_V \ = \ (u, T^*(v))_U \text{ für alle } u \in U, \ v \in V.$$

Das ist etwas anderes als die duale Abbildung von V^* in U^*, obwohl in vielen Fällen ein enger Zusammenhang besteht. Man kann diesen Zusammenhang erahnen, wenn man das innere Produkt benutzt, um zu jedem $u \in U$ ein Funktional $\lambda_u \in U^*$ mit

$$\lambda_u(w) := (w, u)_U \text{ für alle } w \in U$$

zu definieren. Es ist aber keineswegs klar, ob man damit alle denkbaren Funktionale aus U^* bekommt. Wenn das aber doch so ist (und man erfährt in der Disziplin "Funktionalanalysis" Bedingungen dafür), so kann man zu jedem $v \in V$ das Funktional $u \mapsto (T(u), v)_V$ in U^* bilden und schreibt es als ein λ_z mit $\lambda_z(u) = (u, z)_U = (T(u), v)_V$ und definiert $T^*(v) := z$. Das klappt.

Im $\mathbb{R}^n$ oder $\mathbb{C}^n$ kann man eine Vielzahl von inneren Produkten definieren, indem man sie über spezielle quadratische Matrizen einführt.

[1] `http://de.wikipedia.org/wiki/Adjungierter_Operator`
[2] `http://de.wikipedia.org/wiki/Adjungierte_Matrix`

Definition 5.14

1. *Jede $n \times n$–Matrix A über $\mathbb{K}$ definiert eine* **quadratische Form**[1]

$$q_A(x) := x^T A\overline{x} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_j\overline{x_k} \text{ für alle } x \in \mathbb{K}^n$$

   mit Werten in $\mathbb{K}$.

2. *Ist A reell und symmetrisch, oder ist A komplex und hermitesch, so ist die quadratische Form q_A reellwertig.*

3. *Die Matrix A heißt in diesen beiden Fällen* **positiv semidefinit**, *wenn $q_A(x) \geq 0$ für alle $x \in \mathbb{K}^n$ gilt.*

4. *Sie heißt* **positiv definit**[2], *wenn zusätzlich $x = 0$ aus $q_A(x) = 0$ folgt.*

5. *In diesem Falle ist*

$$(x, y)_A := x^T A\overline{y} \ : \ K^n \times \mathbb{K}^n \to \mathbb{R} \tag{5.15}$$

   ein inneres Produkt.

6. *Zu jeder Matrix $A \in \mathbb{K}^{m\times n}$ ist die Matrix $A^*A \in \mathbb{K}^{n\times n}$ hermitesch und positiv semidefinit. Sie ist positiv definit, wenn A den Rang n hat.*

In der Definition sind mehrere Behauptungen versteckt, die zu beweisen sind. Zuerst die Reellwertigkeit im Falle $A = A^*$:

$$\begin{aligned}
\overline{q_A(x)} &= \overline{x^T A\overline{x}} \\
&= \overline{x}^T\overline{A}x \\
&= x^T\overline{A}^T\overline{x} \\
&= x^T A^*\overline{x} \\
&= x^T A\overline{x} \\
&= q_A(x) \text{ für alle } x \in \mathbb{K}^n.
\end{aligned}$$

Die nach Definition 5.9 zu fordernden Eigenschaften eines inneren Produktes rechnet man leicht nach, wenn A positiv definit ist.

Nehmen wir eine Matrix $A \in \mathbb{K}^{m\times n}$ her und berechnen die quadratische Form

$$q_{A^*A}(\overline{x}) = \overline{x}^T A^* Ax = \|\overline{Ax}\|_2^2 \geq 0.$$

[1] `http://de.wikipedia.org/wiki/Quadratische_Form`
[2] `http://de.wikipedia.org/wiki/Definitheit`

Das liefert die letzte Aussage der "Definition". □

Positiv definite Matrizen und die von ihnen erzeugten inneren Produkte der Form (5.15) treten in der Optimierung und in der Statistik häufig auf. Wir werden auf dieses Thema zurückkommen.

5.4 Orthogonalität und Orthonormalbasen

Definition 5.16 *Es sei V ein Prä–Hilbert–Raum über $\mathbb{R}$ oder $\mathbb{C}$.*

1. *Zwei Vektoren $u, v \in V$ heißen* **orthogonal**[1]*, wenn $(u, v) = 0$ gilt.*
2. *Zwei Unterräume U und W heißen orthogonal, wenn $(u, w) = 0$ für alle $u \in U$ und alle $w \in W$ gilt.*
3. *Ist U ein Unterraum von V, so ist*

$$U^{\perp} := \{v \in V \ : \ (v, u) = 0 \text{ für alle } u \in U\}$$

   der **Orthogonalraum** *zu U (lies: "U senkrecht"). Der Orthogonalraum wird auch als* **orthogonales Komplement**[2] *bezeichnet.*
4. *Eine Basis aus paarweise orthogonalen Vektoren heißt* **Orthogonalbasis**.
5. *Eine Orthogonalbasis, deren Vektoren alle die Länge 1 haben, heißt* **Orthonormalbasis**[3].

Natürlich sind die Einheitsvektoren des $\mathbb{K}^n$ eine Orthonormalbasis des $\mathbb{K}^n$, und der Raum $\mathbb{R}^n$ ist euklidisch unter dem Skalarprodukt als innerem Produkt. Weitere Orthonormalbasen bekommt man durch die Zeilen– und Spaltenvektoren orthogonaler Matrizen:

Theorem 5.17 *Es sei A eine orthogonale oder unitäre $n \times n$–Matrix. Dann gilt:*

1. *$\|A \cdot x\| = \|x\|$ für alle $x \in \mathbb{K}^n$, d.h. A läßt Längen invariant.*
2. *$(A \cdot x, A \cdot y) = (x, y)$ für alle $x, y \in \mathbb{K}^n$ d.h. A läßt innere Produkte invariant, und im reellen Fall läßt A auch alle Winkel invariant.*
3. *A bildet Orthogonalbasen in Orthogonalbasen ab.*

[1] http://de.wikipedia.org/wiki/Orthogonalit%C3%A4t
[2] http://de.wikipedia.org/wiki/Komplement_(lineare_Algebra)
[3] http://de.wikipedia.org/wiki/Orthonormalbasis

4. *Dasselbe gilt für Orthonormalbasen.*

5. *Die Zeilen und Spalten von orthogonalen oder unitären Matrizen sind Orthonormalsysteme.*

6. *Ein Basiswechsel im $\mathbb{R}^n$ von der Standardbasis $\{e_1, \ldots, e_n\}$ in eine andere Orthonormalbasis $\{u_1, \ldots, u_n\}$ ist durch eine Orthogonalmatrix U mit den Spalten $u_1, \ldots, u_n$ gegeben.*

Beweis: Es reicht, die zweite Eigenschaft zu zeigen, denn daraus folgen alle anderen (wieso?). Es gilt aber unter den obigen Bezeichnungen

$$\begin{aligned}(A \cdot x, A \cdot y) &= x^T \underbrace{A^T A^*}_{I_n} y^* \\ &= x^T y^* \\ &= (x, y).\end{aligned}$$

□

In Orthonormalbasen haben Vektoren besonders schöne Koeffizientendarstellungen. Ist etwa $v_1, \ldots, v_n$ eine Orthonormalbasis eines n–dimensionalen Vektorraums V, so hat jedes $v \in V$ die Darstellung

$$v = \sum_{j=1}^{n} (v, v_j) v_j$$

wie man leicht nachrechnet, und es folgt sofort auch

$$\|v\|^2 = \sum_{j=1}^{n} |(v, v_j)|^2.$$

Die Koeffizienten (v, v_j) sind ganz eng an die Länge $\|v\|$ des Vektors gebunden.

Wegen dieser wunderbaren Eigenschaften sind Orthonormalbasen für euklidische Räume extrem wichtig, und wir werden darauf zurückkommen. Vorerst aber noch ein nicht ganz so naheliegendes, für die digitale Signalverarbeitung zentrales Beispiel. Man betrachte den Vektorraum der reellwertigen Funktionen f auf $\mathbb{R}$ mit Periode 2π, d.h. mit

$$f(t + 2\pi) = f(t) \text{ für alle } t \in \mathbb{R}.$$

Darin liegen alle trigonometrischen Funktionen $\sin(jt)$ und $\cos(jt)$ für $j = 0, 1, 2, \ldots$ usw. Mit dem reellwertigen inneren Produkt

$$(u, v) := \frac{1}{\pi} \int_{-\pi}^{\pi} u(t) v(t) dt$$

sind die genannten Funktionen zusammen mit der konstanten Funktion $1/\sqrt{2}$ orthonormal. Dies kann man mit guten Schulkenntnissen beweisen, aber wir werden das später etwas eleganter machen. Die infiniten Linearkombinationen

$$f(t) := a_0 + \sum_{j=1}^{\infty} (a_j \cos(jt) + b_j \sin(jt)) \tag{5.18}$$

nennt man **Fourierreihen**, und man benutzt sie zur Darstellung periodischer Signale. Davon später mehr.

Die orthogonalen $n \times n$–Matrizen beschreiben nach Satz 5.17 die Übergänge zwischen dem Standard–Orthonormalsystem der Einheitsvektoren des $\mathbb{R}^n$ und beliebigen anderen Orthonormalsystemen. In diesem Einschub sehen wir uns an, was dies bei Anwendung auf Matrizen bedeutet.

Eine beliebige lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^m$ ist, wie wir schon wissen, durch eine Matrix $A \in \mathbb{R}^{m \times n}$ als $x \mapsto A \cdot x$ darstellbar. Ein Basiswechsel im $\mathbb{R}^n$ von der Standardbasis $\{e_1, \ldots, e_n\}$ in eine andere Orthonormalbasis $\{u_1, \ldots, u_n\}$ ist durch eine Orthogonalmatrix U mit den Spalten $u_1, \ldots, u_n$ gegeben, denn dann gilt ja $Ue_j = u_j,\ 1 \le j \le n$. Analog ist ein Basiswechsel im $\mathbb{R}^m$ auf eine neue Orthonormalbasis $\{v_1, \ldots, v_m\}$ durch eine $m \times m$–Orthogonalmatrix V beschreibbar.

Die Frage ist nun, ob man durch eine geeignete Basenwahl im Bildraum und im Urbildraum erreichen kann, daß man in den beiden neuen Basen die Abbildung A auf eine besonders einfache Form bringen kann. In den neuen Basen wird die Abbildung durch $V \cdot A \cdot U^{-1} = V \cdot A \cdot U^T$ beschrieben, weil

$$\begin{array}{ccc} \text{span } \{e_j\} & \stackrel{A}{\to} & \text{span } \{e_k\} \\ U \downarrow & & \downarrow V \\ \text{span } \{u_j\} & \stackrel{V \cdot A \cdot U^{-1}}{\to} & \text{span } \{v_k\} \end{array}$$

gilt. Die Orthogonalmatrizen lassen euklidische Längen, Winkel und Abstände unverändert. Deshalb muß man von der transformierten Matrix erwarten, daß sie Längenveränderungen vornimmt. Im Idealfall ist sie so simpel strukturiert, dass sie die Einheitsvektoren e_ℓ entweder in Null oder in Streckungen $\sigma_\ell e_\ell$ mit $\sigma_\ell > 0$ überführt, d.h. sie hat die Form $S = (s_{jk}) \in \mathbb{R}^{m \times n}$ mit $s_{jk} = \sigma_j \delta_{jk} = \sigma_k \delta_{jk},\ 1 \le j \le m,\ 1 \le k \le n$ mit $\sigma_j \ge 0,\ 1 \le j \le \min(m, n)$.

Und es zeigt sich in der Tat, daß man zu jeder $m \times n$–Matrix A geeignete Orthogonalmatrizen $U \in O(n)$ und $V \in O(m)$ finden kann, so daß $V \cdot A \cdot U^{-1} = S$ mit einer so einfachen $m \times n$–Streckungsmatrix S gilt. Umgekehrt heißt das

aber auch, daß man jede Matrix A als $A = U^T \cdot S \cdot V$ schreiben kann, und das nennt man eine **Singulärwertzerlegung**[1]. Für das praktische Rechnen ist die Singulärwertzerlegung sehr wichtig, weil sie zu gegebener Matrix A eine sehr gute Wahl neuer Orthonormalbasen durchführt, in denen die Wirkung von A besonders einfach beschreibbar ist. Denn es ist einfach, Rang, Bild und Kern von S zu bestimmen (wie?), und damit bekommt man sofort auch Rang, Bild und Kern von A mit passenden Basen. Aber das Verfahren zur Berechnung einer Singulärwertzerlegung ist an dieser Stelle noch zu schwierig. Wir greifen das Thema in Abschnitt 9.2 auf Seite 245 wieder auf, sobald wir den Konvergenzbegriff für Folgen im $\mathbb{R}^n$ zur Verfügung haben. Jacobi–Matrizen bzw. Givens–Rotationen werden sich dabei als sehr nützlich erweisen. Wir beschränken uns dabei aber auf reelle und symmetrische Matrizen $A = A^T$, und dann wird man wegen

$$A = U^T \cdot S \cdot V = A^T = (U^T \cdot S \cdot V)^T = V^T S^T U$$

anstreben, die Zerlegung mit $U = V$ zu machen, zumal S automatisch symmetrisch ist. Man bekommt dann also

$$A = U^T S U = U^{-1} S U$$

mit einer Diagonalmatrix

$$S = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \end{pmatrix}$$

deren Diagonalelemente **Eigenwerte** heißen. Doch darüber mehr in in Abschnitt 9 auf Seite 243.

Orthonormalbasen sind also eine feine Sache, aber man muß beweisen, daß es sie in beliebigen Prä–Hilbert–Räumen auch gibt. Dazu

Theorem 5.19 *Ist V ein endlichdimensionaler Prä–Hilbert–Raum, so hat V eine Orthonormalbasis.*

Beweis: Wir beginnen mit einer beliebigen Basis $\{v_1, \dots . v_n\}$ und konvertieren sie mit dem **Orthogonalisierungsverfahren** von Erhard **Schmidt**[2] in eine Orthogonalbasis $\{u_1, \dots . u_n\}$. Das machen wir induktiv und fangen

[1] `http://de.wikipedia.org/wiki/Singul%C3%A4rwertzerlegung`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Schmidt.html`

damit an, daß mit $u_1 := v_1$ die Vektormenge $\{u_1\}$ aus paarweise orthogonalen Vektoren besteht. Nun nehmen wir an, wir hätten schon aus $\{v_1, \ldots . v_k\}$ eine Menge $\{u_1, \ldots . u_k\}$ aus paarweise orthogonalen Vektoren erzeugt. Dann machen wir einen Ansatz

$$u_{k+1} := v_{k+1} + \sum_{j=1}^{k} \alpha_j u_j$$

und bestimmen die Koeffizienten so, daß $(u_{k+1}, u_m) = 0$ für $1 \leq m \leq k$ gilt. Das funktioniert, wenn man wegen

$$\begin{aligned} (u_{k+1}, u_m) &= (v_{k+1}, u_m) + \sum_{j=1}^{k} \alpha_j (u_j, u_m) \\ &= (v_{k+1}, u_m) + \alpha_m (u_m, u_m) \end{aligned}$$

die Koeffizienten als

$$\alpha_m = -\frac{(v_{k+1}, u_m)}{(u_m, u_m)}, \quad 1 \leq m \leq k$$

wählt. Nun haben wir per Induktion eine Orthogonalbasis, und durch Renormierung $u_j \mapsto u_j / \|u_j\|$ bekommen wir eine Orthonormalbasis. Für spätere Zwecke halten wir noch fest, daß unsere Konstruktion garantiert, daß stets gilt

$$(v_j, u_k) = 0, \ 1 \leq j < k \leq n. \tag{5.20}$$

□

Theorem 5.21 *In euklidischen Vektorräumen V gelten für jeden endlichdimensionalen linearen Unterraum U die Beziehungen*

$$\begin{aligned} U + U^{\perp} &= V \\ U \cap U^{\perp} &= \{0\} \end{aligned}$$

d.h. die Summe $V = U + U^{\perp}$ ist direkt.

Beweis: Für ein $u \in U \cap U^{\perp}$ gilt $(u, u) = \|u\|^2 = 0$, also $u = 0$. Wir nehmen uns eine Orthonormalbasis $\{u_1, \ldots . u_n\}$ von U her und definieren zu jedem Vektor $v \in V$ die Vektoren

$$\begin{aligned} u_v &:= \sum_{j=1}^{n} (v, u_j) u_j \in U \\ u_v^{\perp} &:= v - u_v. \end{aligned}$$

Es folgt

$$\begin{aligned}(u_v^\perp, u_k) &= (v - u_v, u_k)\\ &= (v - \sum_{j=1}^{n}(v,u_j)u_j, u_k)\\ &= (v,u_k) - \sum_{j=1}^{n}(v,u_j)(u_j,u_k)\\ &= (v,u_k) - (v,u_k)\\ &= 0 \text{ für alle } 1 \le k \le n\end{aligned}$$

und deshalb liegt $u_v^\perp$ in $U^\perp$ und $v \in U + U^\perp$. □

Die im Beweis benutzte Abbildung $v \mapsto u_v$ ist von enormer Bedeutung:

Definition 5.22 *Ist U ein n–dimensionaler Unterraum eines Prä–Hilbert–Raums V, und hat U eine Orthonormalbasis $\{u_1, \ldots, u_n\}$, so ist die Abbildung*

$$\begin{aligned}P_U &: V \to U\\ v \mapsto P_U(v) &:= \sum_{j=1}^{n}(v,u_j)u_j \in U\end{aligned}$$

der orthogonale **Projektor** *von V auf U.*

Theorem 5.23 *Unter den obigen Bezeichnungen hat ein orthogonaler Projektor P_U die Eigenschaften*

$$\begin{aligned}P_U &\quad \textit{ist linear}\\ P_U(u) &= u \textit{ für alle } u \in U\\ P_U(u^\perp) &= 0 \textit{ für alle } u^\perp \in U^\perp\\ P_U \circ P_U &= P_U \textit{ (Idempotenz)}\\ Id_V - P_U &: V \to U^\perp\\ (Id_V - P_U) \circ (Id_V - P_U) &= Id_V - P_U \textit{ (Idempotenz)}\\ \|v - P_U(v)\| &= \min_{u\in U}\|v-u\| \textit{ für alle } v \in V.\end{aligned}$$

Beweis: Die ersten vier Eigenschaften sind elementar nachzurechnen, und die fünfte folgt aus Theorem 5.21. Die sechste folgt sofort aus der vierten. Wir müssen nur noch die Minimaleigenschaft beweisen, und dazu nehmen wir einen beliebigen Vektor $v \in V$ her und schreiben einen beliebigen Vektor $u \in U$ als $u = u - P_U(v) + P_U(v)$. Dann rechnen wir den Abstand aus und benutzen den Satz des Pythagoras:

$$\begin{aligned}\|v-u\|^2 &= \|v - (u - P_U(v) + P_U(v))\|^2\\ &= \|\underbrace{v - P_U(v)}_{\in U^\perp} - \underbrace{(u - P_U(v))}_{\in U}\|^2\\ &= \|v - P_U(v)\|^2 + \|u - P_U(v)\|^2\\ &\ge \|v - P_U(v)\|^2\end{aligned}$$

mit Gleichheit genau dann, wenn $u = P_U(v)$ gilt. □

Der orthogonale Projektor P_U bildet also einen beliebigen Vektor $v \in V$ auf den eindeutig bestimmten Vektor $P_U(v) \in U$ ab, der zu v unter allen anderen Vektoren aus U den kürzesten Abstand hat. Die Verbindungsgerade von v zu $P_U(v) \in U$ steht auf U senkrecht (Skizze in der Vorlesung).

Die Minimaleigenschaft der orthogonalen Projektoren ist in vielen Anwendungen von zentraler Bedeutung, zum Beispiel in der auf **Gauss**[1] zurückgehenden **Ausgleichsrechnung** mach der **Methode der kleinsten Quadrate**. Dabei hat man einen großen Vektor $b \in \mathbb{R}^N$ von Meßergebnissen, die sich, wenn keine Fehler vorliegen würden, als Wert $A \cdot x$ einer linearen Abbildung schreiben lassen müßten, die durch eine $N \times n$–Matrix A mit $n << N$ (n sehr klein gegen N) gegeben ist. Man will den Vektor $x \in \mathbb{R}^n$ berechnen, aber wegen der Fehler kann das lineare Gleichungssystem $A \cdot x = b$ nicht lösbar sein, und es hat ohnehin im Normalfall viel mehr Gleichungen als Unbekannte.

Man zieht sich aus der Affäre, indem man ein x sucht, das die Länge des Fehlervektors $b - A \cdot x$ minimal macht. Wenn man sich $\|b - A \cdot x\|_2^2$ ansieht, minimiert man dabei die Summe der Quadrate der Komponenten des Fehlers, was den Namen der Methode erklärt. Mit unseren bisherigen Kenntnissen ist schon klar, wie man das Problem jetzt angehen sollte: man definiert den Unterraum $U := A(\mathbb{R}^n) \subset \mathbb{R}^N$ und projiziert b auf U, um das Bild $P_U(b)$ zu bekommen, das unter allen Elementen von U den kürzesten Abstand zu b hat. Die Lösung hat die Form

$$P_U(b) \;=\; \sum_{j=1}^{n} (b, u_j) u_j$$

mit einer Orthonormalbasis $\{u_1, \ldots, u_n\}$ von $U = A(\mathbb{R}^n)$. Dazu hat man aus den Spalten von A eine Orthonormalbasis mit gleicher linearer Hülle zu erzeugen, aber wir wollen erst später ausführen, wie man das praktisch macht. Im Prinzip kann man das Orthogonalisierungsverfahren von Erhard Schmidt nehmen, aber es ist rechentechnisch heikel.

Aber wir können einiges direkt ausrechnen. Zuerst ein

Lemma 5.24 *Hat eine $N \times n$–Matrix A über $\mathbb{R}$ mit $N \geq n$ den maximalen Rang n, so ist $A^T A$ symmetrisch, nicht singulär und positiv definit.*

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Gauss.html`

Daß A^TA symmetrisch ist, folgt aus

$$(A^TA)^T = A^T(A^T)^T = A^TA.$$

Sowohl die positive Definitheit als auch die Invertierbarkeit von A^TA folgt, wenn wir zeigen können, daß aus $Ax = 0$ immer $x = 0$ folgt, d.h. wenn A injektiv ist (Frage: warum?). Wegen der linearen Unabhängigkeit der n Spalten von A muß A aber injektiv sein. □

Wegen des Lemmas kann man unter dessen Voraussetzungen immer das **Gaußsche Normalgleichungssystem**

$$A^TAx = A^Tb$$

lösen, als Ersatz für das eventuell überbestimmte lineare Gleichungssystem $Ax = b$.

Theorem 5.25 *Die Lösung des Gaußschen Normalgleichungssystems liefert auch das Minimum von*

$$\|Ax - b\|_2^2$$

unter allen $x \in \mathbb{R}^n$.

Weil die obige Minimierung eine Summe von Quadraten möglichst klein macht, heißt diese Technik auch **Methode der kleinsten Quadrate**[1].

Sei $x \in \mathbb{R}^n$ eine Lösung des Gaußschen Normalgleichungssystems, und sei $y \in \mathbb{R}^n$ ein beliebiger Vektor. Dann untersuchen wir

$$\begin{aligned}
\|A(x+y) - b\|_2^2 &= (A(x+y)-b)^T(A(x+y)-b) \\
&= (x+y)^TA^TA(x+y) - 2b^TA(x+y) + b^Tb \\
&= x^TA^TAx + 2x^TA^TAy + y^TA^TAy \\
&\quad -2(A^Tb)^Tx - 2(\underbrace{A^Tb}_{A^TAx})^Ty + b^Tb \\
&= x^TA^TAx + y^TA^TAy - 2(A^Tb)^Tx + b^Tb \\
&\geq x^TA^TAx - 2(A^Tb)^Tx + b^Tb \\
&= \|Ax - b\|_2^2
\end{aligned}$$

weil wegen der positiven Definitheit immer $y^TA^TAy \geq 0$ gilt. □.

Noch ein kleiner "*hack*" für Interessierte: Wir haben bei der Definition des orthogonalen Projektors P_U vorausgesetzt, dass U endlichdimensional ist. Man kann dann aber auch $P_{U^\perp} := Id_V - P_U$ definieren und bekommt ganz analoge Eigenschaften für $P_{U^\perp}$. Vertauscht man U und $U^\perp$ in dieser Argumentation, so stellt man fest, dass sich Orthogonalprojektoren P_U auch dann definieren lassen, wenn $U^\perp$ statt U endlichdimensional ist. Sind weder U noch $U^\perp$ endlichdimensional, braucht man Zusatzvoraussetzungen.

[1] `http://de.wikipedia.org/wiki/Methode_der_kleinsten_Quadrate`

5.5 Geraden, Hyperebenen, Spiegelungen, Drehungen

Wir sehen uns jetzt noch einmal im $\mathbb{R}^n$ die Geraden[1] und Hyperebenen[2] an. **Gerade** sind Punktmengen der Form

$$G(x,r) := \{x + \alpha \cdot r \ : \ \alpha \in \mathbb{R}\} =: x + \mathbb{R} \cdot r$$

für Vektoren $x \in \mathbb{R}^n$ und $r \in \mathbb{R}^n \backslash \{0\}$. Der Vektor r heißt **Richtungsvektor** und gibt die **Richtung**[3] der Geraden an. Er kann normiert werden zu $\|r\|_2 = 1$. Gilt dies, und sind zwei Punkte $u_1 := x + \alpha_1 \cdot r, u_2 := x + \alpha_2 \cdot r$ auf der Geraden $G(x,r)$ gegeben, so ist deren Abstand durch $|\alpha_1 - \alpha_2|$ aus den Koeffizienten des Richtungsvektors ablesbar, weil

$$\begin{aligned} \|u_1 - u_2\|_2 &= \|x + \alpha_1 \cdot r - (x + \alpha_2 \cdot r)\|_2 \\ &= \|(\alpha_1 - \alpha_2) \cdot r\|_2 \\ &= |\alpha_1 - \alpha_2| \|r\|_2 \\ &= |\alpha_1 - \alpha_2| \end{aligned}$$

gilt.

Geraden sind affine Unterräume. Zwei Geraden heißen **parallel**, wenn ihre Richtungsvektoren linear abhängig sind. Eine Gerade $G(x,r)$ geht genau dann durch den Nullpunkt, wenn x und r linear abhängig sind, und dann gilt $G(x,r) = G(0,r)$ und man hat einen eindimensionalen linearen Unterraum.

Frage: Warum gelten diese Aussagen?

Aufgabe: Wie berechnet man für einen beliebigen Punkt $v \in \mathbb{R}^n$ dessen Abstand zu einer gegebenen Geraden $G(x,r)$?

Hyperebenen sind beschreibbar durch nichttriviale inhomogene lineare Gleichungen der Form (4.74). Dazu braucht man einen Vektor $a \in \mathbb{R}^n \setminus \{0\}$ und einen Skalar β, und man betrachtet den affinen Unterraum

$$H(a,\beta) = \{x \in \mathbb{R}^n \ : \ a^T x = \beta\}.$$

Der zugehörige lineare Unterraum mit Codimension 1 ist die Nullpunktshyperebene $H(a,0) = \ker\,(x \mapsto a^T x)$, und sie ist der Orthogonalraum der Nullpunktsgeraden

$$G(0,a) = \mathbb{R} \cdot a := \{\alpha \cdot a \ : \ \alpha \in \mathbb{R}\} = H^{\perp}(a,0).$$

[1] `http://de.wikipedia.org/wiki/Gerade`
[2] `http://de.wikipedia.org/wiki/Hyperebene`
[3] `http://de.wikipedia.org/wiki/Richtungsvektor`

Diese Gerade (und damit auch der feste Vektor a) steht auf allen Hyperebenen $H(a,\beta)$ senkrecht, was bei einem affinen Unterraum besagt, daß sie auf allen Differenzen von Elementen des affinen Unterraums senkrecht steht.

In der Darstellung von $H(a,\beta)$ kann man a und β gemeinsam mit einem festen, von Null verschiedenen Faktor multiplizieren und erhält wieder dieselbe Hyperebene. Deshalb normiert man gern a so, daß $a^Ta = 1$ gilt und nennt a dann den **Normaleneinheitsvektor** zu den Hyperebenen $H(a,\beta)$. Die Richtung bzw. das Vorzeichen von a teilt dann den Raum $\mathbb{R}^n$ in drei Teile:

$$\begin{array}{rl} \text{“oberer” Halbraum} & \{x \in \mathbb{R}^n \ : \ a^Tx > \beta\} \\ \text{Hyperebene} & \{x \in \mathbb{R}^n \ : \ a^Tx = \beta\} \\ \text{“unterer” Halbraum} & \{x \in \mathbb{R}^n \ : \ a^Tx < \beta\}. \end{array}$$

Hat man die Darstellung $H(a,\beta)$ einer Hyperebene durch $a^Ta = 1$ normiert, so hat der Nullpunkt von dieser Ebene den Abstand $|\beta|$. Allgemeiner hat dann ein beliebiger Punkt $y \in \mathbb{R}^n$ den Abstand $|\beta - a^Ty|$ von der Hyperebene. Der vorzeichenbehaftete Wert $\beta - a^Ty$ gibt an, ob sich y im unteren oder oberen **Halbraum**[1] oder sogar auf der Hyperebene befindet.

Das kann man konkret ausrechnen, indem man die Gerade $G(y,a)$, die ja durch y geht und auf der Hyperebene senkrecht steht, mit der Hyperebene schneidet. Der Schnittpunkt $y + \alpha \cdot a$ erfüllt dann

$$\begin{array}{rcl} a^T(y + \alpha \cdot a) & = & \beta \\ \alpha & = & \dfrac{\beta - a^Ty}{a^Ta} \\ & = & \beta - a^Ty \end{array}$$

und dies ist bis auf das Vorzeichen der Abstand von y zu $y + \alpha \cdot a$ auf $G(y,a)$, wie wir schon wissen. Wie im Beweis der Minimaleigenschaft der Projektion im Satz 5.23 zeigt man noch, daß man durch diese Konstruktion den kürzesten Abstand von y zur Hyperebene $H(a,\beta)$ realisiert hat.

Mit dem Begriff der Hyperebene kann man nun definieren, was eine **Spiegelung** an einer solchen Ebene sein soll. Ein Vektor $y \in \mathbb{R}^n$ wird an einer Hyperebene $H(a,\beta)$ gespiegelt, wenn der Bildpunkt $S(y)$ nach der obigen Konstruktion genau der Punkt $y + 2\alpha \cdot a$ ist (Skizze in der Vorlesung). Es folgt, daß die Spiegelung die affine Abbildung

$$\begin{array}{rcl} S(y) & = & y + 2\alpha \cdot a \\ & = & y + 2(\beta - a^Ty) \cdot a \text{ für alle } y \in \mathbb{R}^n \end{array}$$

[1] `http://de.wikipedia.org/wiki/Halbraum`

ist. Wenn es sich um eine Nullpunktshyperebene handelt, gilt $\beta = 0$ und die Spiegelung wird eine lineare Abbildung

$$\begin{aligned} S(y) &= y - 2a^T y \cdot a \\ &= y - 2 \cdot a \cdot a^T y \\ &= (Id_{I\!R^n} - 2a \cdot a^T)y \text{ für alle } y \in I\!R^n. \end{aligned}$$

Theorem 5.26 *Ist $a \in I\!R^n \setminus \{0\}$ ein Vektor, so wird die Spiegelung an der Nullpunktshyperebene $H(a,0)$ durch die lineare Abbildung mit der* **Householder**[1]*–Matrix*[2] *$Id_{I\!R^n} - 2a \cdot a^T$ beschrieben. Diese Matrizen sind symmetrisch und orthogonal.*

Aufgabe: Man beweise diese Behauptungen.

Obwohl Spiegelungen durch Householder–Matrizen beschrieben werden, sollte man in der Praxis nie eine Matrizenmultiplikation ausführen, um eine Spiegelung zu bewirken. Man geht besser zurück auf die oben schon benutzte Form

$$S(y) = y - 2a^T y \cdot a,$$

indem man erst $a^T y$ ausrechnet, dann $\gamma := 2a^T y$ und schließlich $S(y) = y - \gamma \cdot a$ als reine Vektoroperation. das ist erheblich effizienter.

Nun wollen wir Drehungen im $I\!R^2$ untersuchen. Dazu bedarf es eines Winkels ψ um den wir drehen wollen, und ein beliebiger Punkt $(x,y)^T = (r\cos\varphi, r\sin\varphi)^T$ in Polarkoordinaten soll in

$$\begin{aligned} \begin{pmatrix} r\cos(\varphi+\psi) \\ r\sin(\varphi+\psi) \end{pmatrix} &= \begin{pmatrix} r\cos\varphi\cos\psi - r\sin\varphi\sin\psi \\ r\cos\varphi\sin\psi + r\sin\varphi\cos\psi \end{pmatrix} \\ &= \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & +\cos\psi \end{pmatrix} \cdot \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} \end{aligned}$$

übergehen. Die **Drehmatrix**[3]

$$D_\psi := \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & +\cos\psi \end{pmatrix}$$

ist orthogonal, aber bei beliebigem Winkel nicht symmetrisch. Sie stellt die Drehung als lineare Abbildung im Orthonormalsystem der Einheitsvektoren dar. Man beweist mit Hilfe der bekannten Rechenregeln leicht

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Householder.html`
[2] `http://de.wikipedia.org/wiki/Householdertransformation`
[3] `http://de.wikipedia.org/wiki/Drehmatrix`

Theorem 5.27 *Die reellen 2×2–Drehmatrizen bilden eine Gruppe unter der Matrizenmultiplikation. Es gilt*

$$\begin{array}{rcl} D_0 & = & Id_{\mathbb{R}^2} \\ D_{\phi+\psi} & = & D_\phi \circ D_\psi \text{ für alle } \phi, \psi \in \mathbb{R} \\ D_\phi^{-1} & = & D_{-\phi} \text{ für alle } \phi \in \mathbb{R} \end{array}$$

Drehungen im R^n sind praktisch nicht einfach handzuhaben, wenn man sie nicht auf den zweidimensionalen Fall reduziert. Man betrachtet einfach nur zwei fest gewählte Indizes $j < k \in \{1, \ldots, n\}$ und dreht um einen Winkel ψ in der durch die Einheitsvektoren e_j und e_k aufgespannten zweidimensionalen Ebene. Die Drehmatrix hat dann die Form

$$\begin{array}{cc} & \begin{array}{ccccccccc} & & j & & & & k & & \\ & & \downarrow & & & & \downarrow & & \end{array} \\ \begin{array}{c} \\ \\ \\ j \rightarrow \\ \\ \\ k \rightarrow \\ \\ \\ \end{array} & \left(\begin{array}{ccccccccc} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\psi & & & -\sin\psi & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & \sin\psi & & & \cos\psi & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{array} \right) \end{array}$$

wobei überall sonst Nullen stehen. Das kann man etwas kompakter schreiben als

$$Id_{\mathbb{R}^n} + (\cos\psi - 1)(e_j e_j^T + e_k e_k^T) + \sin\psi(e_k e_j^T - e_j e_k^T).$$

Man nennt diese Matrizen **Jacobi**[1]–Matrizen. Sie sind orthogonal. In der Praxis benutzt man aber auch hier keine Matrixmultiplikation, sondern berechnet zu einem gegebenen Vektor $v = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$ die Komponenten des transformierten Vektors $w = (w_1, \ldots, w_n)^T \in \mathbb{R}^n$ als

$$\begin{array}{rcl} w_j & = & v_j \cos\psi - v_k \sin\psi \\ w_k & = & v_j \cos\psi + v_k \sin\psi \\ w_i & = & v_i \text{ für alle } i \neq j,\ i \neq k,\ 1 \leq i \leq n. \end{array}$$

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Jacobi.html`

6 Lösung linearer Gleichungssysteme

Wir greifen jetzt zurück auf Abschnitt 4.5.5 auf Seite 147 und wollen ein **lineares Gleichungssystem** der Form (4.75) oder in Matrixschreibweise

$$A \cdot x = b$$

mit einer Matrix $A = (a_{jk}) \in I\!K^{n\times n}$ und einer rechten Seite $b \in I\!R^n$ nach dem Vektor $x \in I\!R^n$ der Unbekannten auflösen. Um Probleme zu vermeiden, setzen wir erst einmal voraus, der Rang von A sei maximal, d.h. gleich n, und dann wissen wir nach Satz 4.80, daß es eine eindeutige Lösung x geben muß. Wir wollen die Lösung ausrechnen. Es ist dabei erlaubt, die Gleichungen beliebig zu vertauschen, und das entspricht einer Vertauschung der Zeilen der Matrix A und der rechten Seite b. Auch die Spalten von A kann man beliebig vertauschen, denn das ist nur eine Umnumerierung der Unbekannten.

6.1 Orthogonalisierungsverfahren

Als erste Lösungsidee benutzen wir, was wir schon über Orthonormalbasen wissen, und berechnen eine neue Orthonormalbasis $\{u_1, \ldots, u_n\}$ aus den Spalten $v_j := Ae_j$ von A. Das klappt mit der Orthogonalisierungsmethode[1] von Erhard Schmidt aus dem Beweis des Satzes 5.19 oder einem anderen, noch zu schildernden Verfahren. Wenn wir das neue Orthonormalsystem benutzen, um uns eine orthogonale oder unitäre Matrix Q mit den Spalten $u_j = Qe_j$ zu beschaffen, so muß dann die Gleichung $A = Q \cdot R$ mit einer $n \times n$–Matrix $R = Q^* \cdot A$ gelten, weil die Spalten von A aus denen von Q linear kombinierbar sind. Aus (5.20) folgt dann aber

$$\begin{aligned}
(Ae_j, u_k) &= 0,\ 1 \le j < k \le n \\
&= (Ae_j, Qe_k),\ 1 \le j < k \le n \\
&= (Q^*Ae_j, e_k),\ 1 \le j < k \le n \\
&= (Re_j, e_k),\ 1 \le j < k \le n. \\
&= e_j^T R^T e_k,\ 1 \le j < k \le n.
\end{aligned}$$

Das bedeutet, daß die j–te Spalte Re_j von R nur Nullen in den Komponenten $j+1$ bis n hat, oder daß alle Elemente unterhalb der Diagonale gleich Null sind.

Definition 6.1 *1. Eine Matrix $R = (r_{kj}) \in I\!K^{m\times n}$ hat* **obere Dreiecksform**, *wenn gilt*

$$r_{kj} = e_k^T R e_j = e_j^T R^T e_k = 0 \text{ für alle } j < k,\ 1 \le k \le m,\ 1 \le j \le n. \tag{6.2}$$

[1] `http://de.wikipedia.org/wiki/Gram-Schmidtsches_Orthogonalisierungsverfahren`

2. *Eine QR–***Zerlegung**[1] *einer $n \times n$–Matrix A hat die Form $A = Q \cdot R$ mit einer unitären Matrix $U \in U(n)$ und einer oberen Dreiecksmatrix R.*

Theorem 6.3 *Jede $n \times n$–Matrix A hat eine QR–Zerlegung.*

Den Beweis haben wir schon für den Fall von Matrizen mit maximalem Spaltenrang geführt. Die allgemeine Situation erfordert etwas mehr Sorgfalt, soll hier aber nicht behandelt werden. □.

Jetzt können wir das lineare Gleichungssystem umschreiben in

$$Rx = U^*Ax = U^*b =: c$$

und müssen noch das System $Rx = c$ lösen. Weil U maximalen Rang n hat, ist nach (4.73) der Rang von A derselbe wie der von R. Das System hat die schöne Form

$$\begin{array}{ccccccccccc}
r_{11}x_1 & + & r_{12}x_2 & + & \dots & + & r_{1,n-1}x_{n-1} & + & r_{1n}x_n & = & c_1 \\
 & & r_{22}x_2 & + & \dots & + & r_{2,n-1}x_{n-1} & + & r_{2n}x_n & = & c_2 \\
 & & & & \ddots & & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & & & & & & r_{n-1,n-1}x_{n-1} & + & r_{n-1,n}x_n & = & c_{n-1} \\
 & & & & & & & & r_{nn}x_n & = & c_n
\end{array}$$

und läßt sich durch die Rekursion

$$\begin{aligned}
x_n &= \frac{c_n}{r_{nn}} \\
x_{n-1} &= \frac{1}{r_{n-1,n-1}}\left(c_{n-1} - r_{n-1,n}x_n\right) \\
x_k &= \frac{1}{r_{kk}}\left(c_k - \sum_{j=k+1}^{n} r_{k,j}x_j\right),\ k = n-2, n-3, \dots, 1
\end{aligned}$$

"rückwärts" auflösen, wenn alle Diagonalelemente r_{kk} nicht Null sind. Das ist aber erfüllt, wenn A und damit auch R invertierbar sind, denn es gilt

Theorem 6.4 *Eine obere $n \times n$–Dreiecksmatrix R ist genau dann nichtsingulär und invertierbar, wenn alle Diagonalelemente $e_j^T Re_k$ nicht Null sind.*

Beweis: Es sei das Diagonalelement r_{kk} gleich Null. Dann bildet R die lineare Hülle von $\{e_1, \dots, e_k\}$ auf die lineare Hülle von $\{e_1, \dots, e_{k-1}\}$ ab, wie man aus (6.2) ablesen kann, denn aus

$$Re_j = \sum_{i=1}^{j} e_i^T Re_j \cdot e_i,\ 1 \le j \le n$$

[1] `http://de.wikipedia.org/wiki/QR-Zerlegung`

folgt

$$Re_k = \underbrace{e_k^T Re_k}_{=0} \cdot e_k + \sum_{i=1}^{k-1} e_i^T Re_j \cdot e_i.$$

Also kann R nicht Maximalrang haben. Sind aber alle Diagonalelemente von Null verschieden, und hat man eine verschwindende Linearkombination der Spalten von R, so folgt

$$\begin{aligned}
0 &= \sum_{j=1}^{n} \alpha_j Re_j \\
&= \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{j} e_i^T Re_j \cdot e_i \\
&= \sum_{i=1}^{n} \left(\sum_{j=i}^{n} \alpha_j e_i^T Re_j \right) e_i \\
0 &= \sum_{j=i}^{n} \alpha_j e_i^T Re_j,\ 1 \le i \le n \\
&= \alpha_i \underbrace{e_i^T Re_i}_{\neq 0} + \sum_{j=i+1}^{n} \alpha_j e_i^T Re_j,\ 1 \le i \le n
\end{aligned}$$

und nacheinander $\alpha_n = 0,\ \alpha_{n-1} = 0, \ldots, \alpha_1 = 0$. Also hat R genau n linear unabhängige Spalten und ist nichtsingulär. □

6.2 Householder–Verfahren

Informatiker werden das praktische Lösen linearer Gleichungssysteme rekursiv versuchen, und das führt in der Tat zu sehr effizienten Methoden, die man dann, wenn man sie gefunden hat, aber nicht rekursiv programmiert. Nehmen wir uns zuerst im reellen Fall den Orthogonalisierungsprozeß der Spalten von A vor. Wir versuchen, durch eine Householder–Spiegelung die erste Spalte Ae_1 von A in ein Vielfaches αe_1 von e_1 zu transformieren. Das ist ein Übergang zu einer neuen Orthonormalbasis, in der die erste Spalte von A Basisvektor ist. Weil Längen invariant unter Orthogonaltransformationen sind, muß $|\alpha| = \|Ae_1\|_2$ gelten, d.h. α ist bis auf ein Vorzeichen bekannt. Wir suchen also einen nichtverschwindenden Vektor a mit $\|a\| = 1$ und

$$\begin{aligned}
(Id_{\mathbb{R}^n} - 2a \cdot a^*)Ae_1 &= \alpha e_1 \\
&= Ae_1 - 2a \cdot a^* Ae_1 \\
&= Ae_1 - 2a^* Ae_1 \cdot a \\
a &= (Ae_1 - \alpha e_1)\frac{1}{2a^* Ae_1}
\end{aligned}$$

Bis auf einen Skalar gilt also $a = Ae_1 - \alpha e_1$ und wegen der Normierung folgt

$$a = \frac{Ae_1 - \alpha e_1}{\|Ae_1 - \alpha e_1\|_2,}$$

was sich umgekehrt auch als hinreichend für $(Id_{\mathbb{R}^n} - 2a \cdot a^*)Ae_1 = \alpha e_1$ erweist. Das Vorzeichen von α ist immer noch frei, und man kann es so wählen, daß in der ersten Komponente $a_{11} - \alpha$ von $Ae_1 - \alpha e_1$ keine Auslöschung eintritt. Man setzt also $\alpha = -\text{sgn}\,(a_{11})\|Ae_1\|_2$.

Jetzt hat man erreicht, daß

$$(Id_{\mathbb{R}^n} - 2a \cdot a^*)\, A = \begin{pmatrix} \alpha & u^T \\ 0 & \tilde{A} \end{pmatrix}$$

mit einer $(n-1) \times (n-1)$–Matrix $\tilde{A}$ gilt. Dabei bestimmt man die rechte Seite und damit $\tilde{A}$ nicht durch Matrixmultiplikation, sondern durch die Berechnung der Transformation

$$A - (2a) \cdot (a^* A) =: \begin{pmatrix} \alpha & u^T \\ 0 & \tilde{A} \end{pmatrix}$$

unter Berücksichtigung der Klammerung. Wir werden später noch eine andere Methode sehen, zu einer Gleichung der Form

$$T \cdot A = \begin{pmatrix} \alpha & u^T \\ 0 & \tilde{A} \end{pmatrix} \tag{6.5}$$

zu kommen, und wir haben es bisher für die orthogonale Householder–Transformation[1] $T := Id_{\mathbb{R}^n} - 2a \cdot a^*$ geschafft. Aus der obigen Gleichung folgt aber für unser Gleichungssystem $A \cdot x = b$ eine rekursive Form mit

$$T \cdot A \cdot x = T \cdot b = \begin{pmatrix} \alpha & u^T \\ 0 & \tilde{A} \end{pmatrix} \cdot x,$$

wenn man x und Tb geeignet aufspaltet. In der Tat ergibt sich mit der Aufspaltung

$$T \cdot b =: \begin{pmatrix} \gamma \\ \tilde{b} \end{pmatrix} \text{ und } x =: \begin{pmatrix} x_1 \\ \tilde{x} \end{pmatrix}$$

die Beziehung

$$T \cdot b = \begin{pmatrix} \gamma \\ \tilde{b} \end{pmatrix} = \begin{pmatrix} \alpha & u^T \\ 0 & \tilde{A} \end{pmatrix} \cdot x = \begin{pmatrix} \alpha & u^T \\ 0 & \tilde{A} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \tilde{x} \end{pmatrix} = \begin{pmatrix} \alpha x_1 + u^T \tilde{x} \\ \tilde{A}\tilde{x} \end{pmatrix}. \tag{6.6}$$

[1] `http://de.wikipedia.org/wiki/Householdertransformation`

Dann löst man erst rekursiv $\tilde{A}\tilde{x} = \tilde{b}$, setzt das Ergebnis ein in die erste Gleichung ein und löst diese durch

$$x_1 = (\gamma - u^T\tilde{x})\frac{1}{\alpha}.$$

Man nennt diese Methode das **Householder–Verfahren**. Es leistet nichts wesentlich anderes als die Orthogonalisierung nach Erhard Schmidt, ist aber in einem Sinn, den wir hier nicht weiter erklären können, stabiler.

6.3 Eliminationsverfahren nach Gauß

Es gibt aber auch noch andere Methoden, die Zerlegung (6.5) zu erreichen. Man kann die Transformationsmatrix T, die man sich in der Theorie an dieser Stelle denkt, aber nie praktisch mit A multipliziert, mit einem Vektor $a \in I\!K^{n-1}$ auch als

$$T = \begin{pmatrix} 1 & 0 \\ a & I_{n-1} \end{pmatrix}$$

statt als Householdermatrix ansetzen. Man zerlegt A als

$$A = \begin{pmatrix} \beta & v^T \\ w & C \end{pmatrix} \tag{6.7}$$

mit bekannten Größen $\beta \in I\!K,\ v, w \in I\!K^{n-1},\ \tilde{A} \in I\!K^{n-1\times n-1}$ und setzt in (6.5) ein, um zu sehen, ob und wie man $\alpha,\ u$ und $\tilde{A}$ berechnen kann. Es folgt

$$\begin{aligned} \begin{pmatrix} 1 & 0 \\ a & I_{n-1} \end{pmatrix}\begin{pmatrix} \beta & v^T \\ w & C \end{pmatrix} &= \begin{pmatrix} \alpha & u^T \\ 0 & \tilde{A} \end{pmatrix} \\ &= \begin{pmatrix} \beta & v^T \\ \beta a + w & av^T + C \end{pmatrix} \end{aligned}$$

Das kann man erfüllen, indem man

$$\begin{aligned} \alpha &= \beta \\ u &= v \\ a &= -w/\beta \\ \tilde{A} &= av^T + C \end{aligned}$$

setzt, wobei man allerdings darauf achten muß, daß β nicht verschwindet. Etwas übersichtlicher geschrieben, hat man die Matrixgleichung

$$\begin{pmatrix} 1 & 0 \\ -w/\beta & I_{n-1} \end{pmatrix}\begin{pmatrix} \beta & v^T \\ w & C \end{pmatrix} = \begin{pmatrix} \beta & v^T \\ 0 & C - wv^T/\beta \end{pmatrix}, \tag{6.8}$$

die auf eine andere, und technisch sogar einfachere Weise die Zerlegung (6.5) erreicht. Man macht danach wie in (6.6) weiter. Durch Vertauschen von Zeilen und/oder Spalten des linearen Gleichungssystems kann man bei einer nichtsingulären Matrix A immer erreichen, daß das linke obere Element nicht Null ist.

Diese rekursive Methode geht auf **Gauss**[1] zurück und heißt **Eliminationsverfahren**[2] Bei Rechnung mit Papier und Bleistift vertauscht man erst die Gleichungen und Unbekannten, bis im System (4.75)

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{array}$$

das Element a_{11} nicht Null ist. Nun zieht man von der zweiten Zeile das a_{21}/a_{11}–fache der ersten Zeile ab, um dort als erstes Element eine Null zu bekommen. Das macht man Zeile für Zeile, mit dem Ergebnis

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ \underbrace{\left(a_{21} - a_{11}\frac{a_{21}}{a_{11}}\right)}_{=0} x_1 & + & \underbrace{\left(a_{22} - a_{12}\frac{a_{21}}{a_{11}}\right)}_{=:\tilde{a}_{22}} x_2 & + & \dots & + & \underbrace{\left(a_{2n} - a_{1n}\frac{a_{21}}{a_{11}}\right)}_{=:\tilde{a}_{2n}} x_n & = & \underbrace{b_2 - b_1\frac{a_{21}}{a_{11}}}_{=:\tilde{b}_2} \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ \underbrace{\left(a_{m1} - a_{11}\frac{a_{m1}}{a_{11}}\right)}_{=0} x_1 & + & \underbrace{\left(a_{m2} - a_{12}\frac{a_{m1}}{a_{11}}\right)}_{=:\tilde{a}_{m2}} x_2 & + & \dots & + & \underbrace{\left(a_{mn} - a_{1n}\frac{a_{m1}}{a_{11}}\right)}_{=:\tilde{a}_{mn}} x_n & = & \underbrace{b_m - b_1\frac{a_{m1}}{a_{11}}}_{=:\tilde{b}_m} \end{array}$$

wobei man von der j–ten Zeile das a_{j1}/a_{11}–fache der ersten Zeile abzieht. Es resultiert ein kleineres System, dessen allgemeines Element die Form

$$\tilde{a}_{jk} = a_{jk} - a_{1k}\frac{a_{j1}}{a_{11}},\ 2 \le j,k \le n$$

hat. Wenn man genauer hinsieht, ist dies die elementweise ausgeschriebene Matrixgleichung $\tilde{A} = C - wv^T/\beta$, denn es gelten wegen (6.7) die Beziehungen

$$\begin{array}{rcl} \beta & = & a_{11} \\ w & = & (a_{21}, a_{31}, \dots, a_{m1})^T \\ v & = & (a_{12}, a_{13}, \dots, a_{1n})^T \\ C & = & (a_{jk}),\ 2 \le j,k \le n. \end{array}$$

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Gauss.html`
[2] `http://de.wikipedia.org/wiki/Gau%C3%9Fsches_Eliminationsverfahren`

Auch die Transformation der rechten Seite finden wir hier wieder, denn es gilt

$$T \cdot b = \begin{pmatrix} 1 & 0 \\ -w/\beta & I_{n-1} \end{pmatrix} \cdot b = (b_1, \tilde{b}_2, \ldots, \tilde{b}_m)^T$$

wenn man die obigen Bezeichnungen einsetzt und vergleicht.

Man macht nun mit dem verkleinerten System weiter, wobei man eventuell erst einmal wieder die Zeilen oder Spalten vertauschen muß, bis das Element in der oberen linken Ecke nicht Null ist. Am Ende hat man ein System mit einer oberen Dreiecksmatrix und kann "von unten her" die Unbekannten ausrechnen. Diese Rechentechnik ist natürlich von ihrer Logik her rekursiv, wird aber durch geeignete Schleifenprogrammierung und Überspeicherung der Matrixeinträge in nicht–rekursiver Form abgewickelt. Keinesfalls wird eine Matrixmultiplikation wie in (6.8) ausgeführt, obwohl das Ergebnis dasselbe ist.

Schreibt man die Gleichung 6.8 rekursiv als

$$\underbrace{\begin{pmatrix} 1 & 0 \\ * & I_{n-1} \end{pmatrix}}_{=:L_1} \cdot A = \underbrace{\begin{pmatrix} * & * \\ 0 & * \end{pmatrix}}_{=:R_1}$$

$$L_2 \cdot L_1 \cdot A = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & * & I_{n-2} \end{pmatrix}}_{=:L_2} \cdot \underbrace{\begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \end{pmatrix}}_{=:R_1 = L_1 \cdot A} = \underbrace{\begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{pmatrix}}_{=:R_2}$$

usw., so bekommt man

$$\begin{aligned} \underbrace{L_{n-1} \cdots L_2 \cdot L_1}_{=:L^{-1}} \cdot A &= R_{n-1} =: R \\ A &= L \cdot R. \end{aligned}$$

Dabei treten spezielle Matrizenprodukte auf:

Definition 6.9 *Eine $n \times n$–Matrix $L = (\ell_{jk})$ heißt* **normierte untere Dreiecksmatrix**, *wenn gilt*

$$\begin{aligned} \ell_{jj} &= 1 \text{ für alle } j,\ 1 \le j \le n \\ \ell_{jk} &= 0 \text{ für alle } j,k,\ 1 \le j < k \le n. \end{aligned}$$

Theorem 6.10 *Die normierten unteren $n \times n$–Dreiecksmatrizen bilden eine Gruppe unter der Matrizenmultiplikation.*

Beweis: Es seien $L = (\ell_{ij})$ und $M = (m_{jk})$ zwei solche Matrizen. Dann folgt

$$\begin{aligned} e_i^T LMe_k &= \sum_{j=1}^{n} \ell_{ij} m_{jk} \\ &= \sum_{\substack{j = 1 \\ i \geq j \\ j \geq k}}^{n} \ell_{ij} m_{jk} \\ &= \begin{cases} 0 & i < k \\ 1 & i = k. \end{cases} \end{aligned}$$

Um zu beweisen, daß auch L^{-1} wieder eine normierte untere Dreiecksmatrix ist, sehen wir uns die linearen Gleichungssysteme

$$L \cdot x^{(j)} = e_j,\ 1 \leq j \leq n$$

$$\begin{pmatrix} 1 & & & & \\ \vdots & \ddots & & & \\ * & \dots & 1 & & \\ \vdots & & \vdots & \ddots & \\ * & \dots & * & \dots & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1^{(j)} \\ \vdots \\ x_j^{(j)} \\ \vdots \\ x_n^{(j)} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j$$

an. Man kann sie von oben her eindeutig auflösen, und im obigen Fall sind die ersten $j-1$ Komponenten der Lösung gleich Null, die j–te gleich Eins. Bildet man die Matrix B mit Spalten $x^{(1)}, x^{(2)}, \dots, x^{(n)}$, so bekommt man $B = L^{-1}$, weil aus $L \cdot x^{(j)} = L \cdot B \cdot e_j = e_j,\ 1 \leq j \leq n$ die Gleichung $L \cdot B = I_n$ folgt. □

Das Eliminationsverfahren liefert also, wenn man Zeilen– und Spaltenvertauschungen ignoriert, eine LR–**Zerlegung**[1] $A = L \cdot R$ mit einer normierten unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix R.

6.4 Pivotisierung und Rangentscheid

Man sieht, daß bei der Transformation immer durch das Diagonalelement a_{11} und später $\tilde{a}_{22}$ usw. dividiert werden muß. Dabei muß Auslöschung im Ergebnis verhindert werden, und deshalb sollte man sicherstellen, daß die voneinander abgezogenen Zahlen nicht zu groß werden. Das ist natürlich nicht generell erreichbar, aber es ist zumindestens günstig, durch geeignetes Vertauschen von Zeilen oder Spalten (**Pivotisierung**) dafür zu sorgen, daß

[1] `http://de.wikipedia.org/wiki/LR-Zerlegung`

diese **Pivotelemente**[1] betragsmäßig groß sind. Man sortiert immer das betragsmäßig größte Element nach vorn und oben. Genaueres lernt man in der Vorlesung "Numerische Mathematik" (vgl. auch [5]).

Auch beim Householder–Verfahren wird pivotisiert, aber nur durch Spaltenvertauschung. Man setzt in jedem Rekursionsschritt den Spaltenvektor mit der größten euklidischen Länge nach vorn. Dies ist insbesondere dann nötig, wenn man eine Matrix $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ hat und eine Orthogonalbasis des Spaltenraums bestimmen will, wobei man den Rang von A gar nicht kennt. Das ist der Normalfall bei Ausgleichsproblemen nach der Methode der kleinsten Quadrate. Man wendet das Householder–Verfahren schrittweise auf die Spalten von A an, wobei man immer die Teilspalten mit größter Länge nach vorn sortiert und abbricht, wenn man nur noch Spalten hat, die man nicht klar von Nullspalten unterscheiden kann. Für diesen **Rangentscheid** hat man ein von der Rechengenauigkeit abhängiges Kriterium zu wählen, das aber immer fragwürdig bleibt, sofern man nicht exakt rechnet.

6.5 Inversion

Hat man mehrere lineare Gleichungssystem mit derselben $n \times n$–Koeffizientenmatrix A zu lösen, so lohnt sich die separate QR– oder LR–Zerlegung, weil man dann Systeme der Form $Ax = b$ für verschiedene b z.B. bei einer LR–Zerlegung nacheinander durch die beiden dreiecksförmigen Systeme

$$\begin{aligned} Ly &= b \\ Rx &= y \end{aligned}$$

lösen kann, denn es folgt $LRx = Ly = b = Ax$.

Die **Inverse**[2] A^{-1} einer Matrix $A \in \mathbb{K}^{n \times n}$ muß man in der Praxis nur selten berechnen. Aber sie ist ein Spezialfall der obigen Situation, weil man die n rechten Seiten $e_j,\ 1 \leq j \leq n$ nehmen kann und damit die Spalten $A^{-1}e_j$ von A^{-1} bekommt:

$$A \cdot (A^{-1}e_j) = I_n e_j = e_j,\ 1 \leq j \leq n.$$

Das **Gauß–Jordan–Verfahren**[3] zur Inversion können wir hier aus Zeitgründen nicht behandeln.

[1] `http://de.wikipedia.org/wiki/Pivotelement`
[2] `http://de.wikipedia.org/wiki/Inverse`
[3] `http://de.wikipedia.org/wiki/Gau%C3%9F-Jordan-Algorithmus`

6.6 Determinanten

Der Begriff der Determinante[1] ist für die Theorie wichtig, verleitet die Studierenden aber dazu, Determinanten auf Computern ausrechnen zu wollen, was in der Regel unpraktisch ist. Der Vollständigkeit halber (und weil wir sie bei der mehrdimensionalen Integration leider nicht vermeiden können) müssen sie hier behandelt werden.

Die geometrische Idee der Determinante ist einfach zu veranschaulichen. Gegeben seien zwei Vektoren z_1, z_2 des $\mathbb{R}^2$. Sie sind genau dann linear abhängig, wenn der Flächeninhalt des von ihnen und dem Nullpunkt aufgespannten Parallelogramms (d.h. der konvexen Hülle von $0, z_1, z_2, z_1 + z_2$) Null ist. Man kann also diesen Flächeninhalt als Kriterium für den "Grad der linearen Abhängigkeit" der Vektoren nehmen. Genauso funktioniert dies für n Vektoren $z_1, \ldots, z_n$ des $\mathbb{R}^n$: man sollte das "Volumen" der Bildmenge der Abbildung

$$T \; : \; [0,1]^n \to \mathbb{R}^n, \; (t_1, \ldots, t_n)^T \mapsto \sum_{j=1}^{n} t_j z_j \tag{6.11}$$

untersuchen. Es ist für zwei Vektoren $(x_1, y_1), (x_2, y_2)$ nach einiger Rechnung zu sehen (Zeichnung siehe Abbildung 3, mit Dank an Anna Eggers), daß dieses Volumen gleich $x_1 y_2 - x_2 y_1$ ist. Die Fläche des Parallelogramms ergibt sich nämlich durch Abziehen der sechs Restflächen von der Gesamtfläche, d.h.

$$\begin{aligned} & (x_1 + x_2) * (y_1 + y_2) - 2 * F(A) - 2 * F(B) - 2 * F(C) \\ = \; & (x_1 + x_2) * (y_1 + y_2) - 2 * \frac{y_2 * x_2}{2} - 2 * \frac{y_1 * x_1}{2} - 2 * y_1 * x_2 \\ = \; & y_2 * x_1 - y_1 * x_2. \end{aligned}$$

Schreibt man die Vektoren als Spalten einer Matrix, so bekommt man

$$\det \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix} := x_1 y_2 - x_2 y_1 \tag{6.12}$$

als einfachste Form der Determinante. Die Vertauschung der Vektoren kehrt das Vorzeichen um, so daß man genaugenommen keinen Flächeninhalt hat, sondern eine vorzeichenbehaftete skalare Größe, deren Betrag ein Flächeninhalt ist. Im Komplexen kann man diese Zahl auch definieren, aber sie verliert die Bedeutung einer Fläche. Sie ist aber immerhin noch genau dann Null, wenn die Matrix singulär ist.

[1] `http://de.wikipedia.org/wiki/Determinante_(Mathematik)`

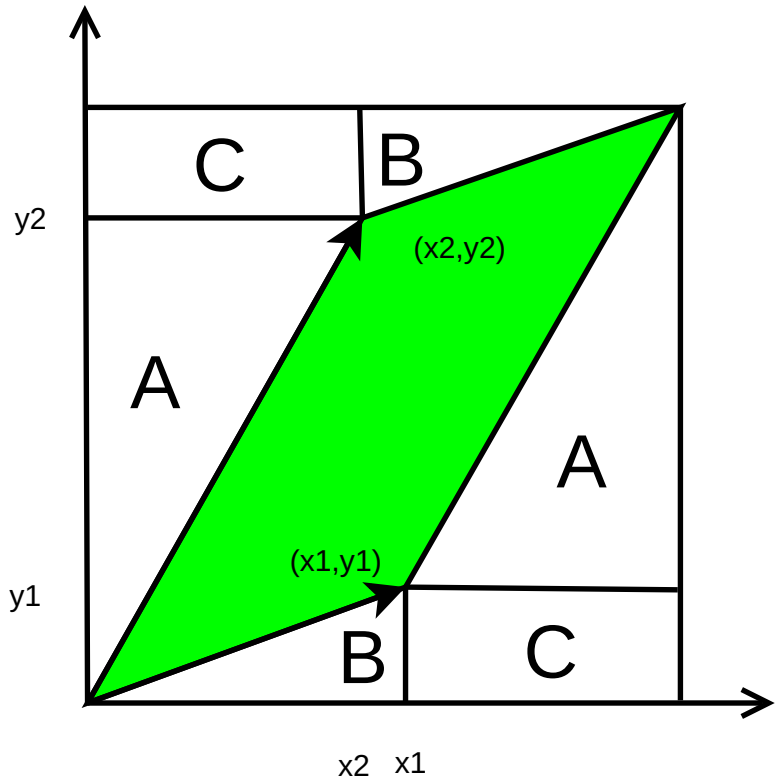


Abbildung 3: Flächeninhalt des Parallelogramms

Allgemeiner wird man also zu einer quadratischen Matrix aus n Spaltenvektoren $z_1, \ldots, z_n$ des $\mathbb{K}^n$ eine Zahl det $(z_1, \ldots, z_n)$ definieren wollen, die **Determinante** genannt wird und folgende Eigenschaften hat:

det $(e_1, \ldots, e_n) = 1$

det ist linear in jedem Argument

det wechselt das Vorzeichen beim Vertauschen zweier Argumente

Im reellen Fall ist |det | das Volumen der Bildmenge aus (6.11).

Jetzt ist es aber Zeit, die Determinante sauber und allgemein zu definieren. Im Falle $n = 1$ haben wir einen 1×1–Vektor $x \in K$ und setzen det $(x) := x$. Dann treffen die obigen Eigenschaften zu, denn u.a. ist $|x|$ die Länge der Strecke zwischen 0 und x. Im zweidimensionalen Fall nehmen wir die

Definition (6.12) und sind zufrieden. Diese Definition zeigt aber auch, wie man allgemeiner verfahren kann, wenn man sie als

$$\det \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix} := x_1y_2 - x_2y_1 = x_1\det(y_2) - y_1\det(x_2)$$

interpretiert. Bei einer allgemeinen Matrix $A = (a_{jk}) \in I\!K^{n\times n}$ definiert man zuerst die $(n-1)\times(n-1)$ Untermatrizen A_{jk} dadurch, dass man die j–te Zeile und die k–te Spalte von A streicht. Dann setzt man rekursiv

$$\begin{aligned} \det(A) &:= \det(Ae_1, \ldots, Ae_n) \\ &:= a_{11}\det(A_{11}) - a_{12}\det(A_{12}) \pm \ldots + (-1)^{n-1}a_{1n}\det(A_{1n}) \\ &= \sum_{k=1}^{n} a_{1k}(-1)^{k-1}\det(A_{1k}). \end{aligned}$$

Obwohl Informatiker für Rekursionen schwärmen, ist diese Definition praktisch unbrauchbar. Denn wenn der Aufwand für eine $n \times n$–Determinante $A(n)$ ist, folgt $A(n) = n * A(n-1) + 2n - 1$ mit fakultativem Wachstum.

Es ist nicht einfach einzusehen, dass unsere Definition zu den oben geforderten Eigenschaften führt. Eine Beweismöglichkeit benutzt eine alternative Form der Determinante, die kombinatorische Begriffe erfordert, und die man oft als Definition der Determinante findet.

Definition 6.13 *Eine* **Permutation**[1] *der Zahlenmenge $Z_n := \{1, 2, \ldots, n\}$ ist eine Bijektion auf dieser Menge. Die Menge der Permutationen von Z_n werde mit $\mathcal{S}_n$ bezeichnet. Zu jeder Permutation $\pi \in \mathcal{S}_n$ ist der* **Fehlstand** *$\epsilon(\pi)$ definiert als die minimale Zahl von Elementvertauschungen, die $\pi(Z_n)$ in Z_n überführt. Das* **Vorzeichen** *der Permutation π ist $\sigma(\pi) := (-1)^{\epsilon(\pi)}$.*

In der Vorlesung "Diskrete Mathematik" sollte folgendes bewiesen worden sein:

Theorem 6.14

- *Die Menge $\mathcal{S}_n$ hat genau $n!$ Elemente.*
- *Sie ist als Automorphismengruppe von Z_n eine Gruppe unter der Verkettung $\circ$ und heißt* **symmetrische Gruppe**.
- *Es gilt $\sigma(\pi) = (-1)^{\epsilon(\pi)} = (-1)^m$, wenn man $\pi(Z_n)$ in Z_n mit m Vertauschungen überführen kann.*
- *σ ist ein Homomorphismus auf die multiplikative Gruppe $\{1, -1\}$.*

[1] `http://de.wikipedia.org/wiki/Permutation`

- *Die Permutationen π mit $\sigma(\pi) = 1$ bzw. mit geradem $\epsilon(\pi)$ bilden eine Untergruppe, die* **alternierende Gruppe $\mathcal{A}_n$**.

Theorem 6.15 *Die Determinante* $\det(A)$ *einer Matrix* $A = (a_{jk}) \in I\!K^{n\times n}$ *läßt sich schreiben als*

$$\det(A) = \sum_{\pi\in\mathcal{S}_n} \sigma(\pi) a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)}. \tag{6.16}$$

Beweis: Es gilt

$$\begin{aligned} & \sum_{\pi\in\mathcal{S}_n} \sigma(\pi) a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)} \\ = & \sum_{k=1}^{n} a_{1k} \sum_{\substack{\pi\in\mathcal{S}_n \\ \pi(1)=k}} \sigma(\pi) a_{2\pi(2)} \cdots a_{n\pi(n)} \end{aligned}$$

Eine Permutation $\pi \in \mathcal{S}_n$ mit $\pi(1) = k$ kann man schreiben als Tupel $(k, \pi(2), \ldots, \pi(n))$. Ist N_k die aufsteigend sortierte Menge $\{1, \ldots, n\} \setminus \{k\}$, so entspricht $(\pi(2), \ldots, \pi(n))$ einer Umsortierung der Zahlen aus N_k. Alle diese Umsortierungen kann man mit je einer Permutation $\psi \in \mathcal{S}_{n-1}$ beschreiben. Man braucht $\epsilon(\psi)$ Vertauschungen, um $(\pi(2), \ldots, \pi(n))$ in die Reihenfolge der Elemente von N_k zu bringen. Mit weiteren $k-1$ Vertauschungen bringt man dann noch das in $(k, \pi(2), \ldots, \pi(n))$ vorn stehende Element k an seine richtige Position. Deshalb gilt $\sigma(\pi) = (-1)^{k-1}\sigma(\psi)$. Mit dieser Überlegung kann man die obige Gleichung weiter umformen zu

$$\begin{aligned} & \sum_{k=1}^{n} a_{1k} \sum_{\substack{\pi\in\mathcal{S}_n \\ \pi(1)=k}} \sigma(\pi) a_{2\pi(2)} \cdots a_{n\pi(n)} \\ = & \sum_{k=1}^{n} a_{1k} (-1)^{k-1} \sum_{\psi\in\mathcal{S}_{n-1}} \sigma(\psi) a_{2,\psi(2)} \cdots a_{n,\psi(n)} \end{aligned}$$

wobei die Permutationen ψ in der Summe so gemeint sind, dass sie den Zahlen $2, \ldots, n$ die Zahlen aus N_k eindeutig zuordnen. Deshalb ist die zweite Summe genau $\det(A_{1k})$. □

Bis auf die Volumeneigenschaft folgen nun leicht die geforderten Eigenschaften der Determinante. Es folgt aber auch bei Matrizenschreibweise

$$\det(A) = \det(A^T),$$

weil

$$\begin{aligned}\det(A^T) &= \sum_{\pi\in\mathcal{S}_n} \underbrace{\sigma(\pi)}_{=\sigma(\pi^{-1})} \underbrace{a_{\pi(1)1}a_{\pi(2)2}\cdots a_{\pi(n)n}}_{\text{vertauschen}} \\ &= \sum_{\pi\in\mathcal{S}_n} \sigma(\pi^{-1})a_{1\pi^{-1}(1)}a_{2\pi^{-1}(2)}\cdots a_{n\pi^{-1}(n)} \\ &= \det(A).\end{aligned}$$

Nun gehen wir auf die effektivere Berechnung von Determinanten zu. Dazu nehmen wir wieder die Schreibweise $\det(z_1,\ldots,z_n)$ für n Vektoren $z_1,\ldots,z_n$ des $\mathbb{R}^n$, die wir als Zeilen oder Spalten in eine $n\times n$–Matrix schreiben können. Wegen des Zeichenwechsels bei Vertauschung folgt für den Spezialfall $z_1 = z_2$ die Gleichung

$$\begin{aligned}\det(z_1,z_2,\ldots,z_n) &= -\det(z_2,z_1,\ldots,z_n) \\ &= -\det(z_1,z_2,\ldots,z_n) \\ &= 0,\end{aligned}$$

was natürlich auch wegen der hier nicht bewiesenen Volumeneigenschaft gelten sollte. Also verschwindet die Determinante $\det(z_1,\ldots,z_n)$ sobald zwei der Vektoren gleich sind.

Wegen der Linearität der Determinante gilt für alle $k \neq 1$ die Gleichung

$$\begin{aligned}\det(z_1+\alpha\cdot z_k,z_2,\ldots,z_n) &= \det(z_1,z_2,\ldots,z_n)+\alpha\cdot\det(z_k,z_2,\ldots,z_n) \\ &= \det(z_1,z_2,\ldots,z_n)+0\end{aligned}$$

d.h. die Determinante ändert sich nicht, wenn man zu einem der Vektoren $z_1,z_2,\ldots,z_n$ ein Vielfaches eines anderen addiert. Das wiederum bedeutet, dass bei der Gauß–Elimination ohne Pivotisierung die Determinante erhalten bleibt. Permutiert man Zeilen oder Spalten mit einer Permutation π, so ändert sich das Vorzeichen der Determinante um $(-1)^{\sigma(\pi)}$.

Theorem 6.17

- *Die Determinante einer oberen oder unteren Dreiecksmatrix ist das Produkt der Diagonalelemente.*
- *Hat eine $n\times n$–Matrix A eine LR–Zerlegung $A = L\cdot R$ mit einer normierten unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix R, so gilt* $\det(A) = \det(R)$ *und ist gleich dem Produkt der Diagonalelemente von R.*
- *Eine LR–Zerlegung ist mit dem Gaußschen Eliminationsverfahren berechenbar, die damit auch die Determinante liefert.*

- *Bei Pivotisierung ändert sich das Vorzeichen der Determinante gemäß den Zeilenvertauschungen: vertauscht man zwei Zeilen mit Abstand k, so ändert sich das Vorzeichen um den Faktor $(-1)^k$.*
- *Eine $n \times n$–Matrix ist genau dann singulär, wenn ihre Determinante Null ist.*

Beweis: Ist A eine obere Dreiecksmatrix, so sind in Gleichung 6.16 alle Terme $a_{1\pi(1)}a_{2\pi(2)}\cdots a_{n\pi(n)}$ gleich Null, wenn nicht $\pi(k) \geq k$ für alle k gilt. Dann folgt aber nacheinander $\pi(n) = n,\ \pi(n-1) = n-1, \ldots, \pi(1) = 1$ und es bleibt als Determinante das Produkt der Diagonalelemente übrig. Der Rest ist einfach, weil bei den Zeilentransformationen der Gauß–Elimination die Determinante unverändert bleibt, sofern nicht pivotisiert wird. Und bei Zeilen– oder Spaltenvertauschungen gilt die genannte Vorzeichenänderung. □

Theorem 6.18 *Für Produkte von Matrizen $A, B \in \mathbb{K}^{n\times n}$ gilt*

$$\det(A \cdot B) = \det(A) \cdot \det(B).$$

Man kann einen Beweis über die LR–Zerlegung des vorigen Satzes führen, aber das wollen wir hier bis auf eine knappe Andeutung unterlassen. Wenn wir Pivotisierung ingorieren und aus einer LR–Zerlegung $A = L_A R_A$ und einer analogen, aber mit Spaltentransformationen ausgeführten Zerlegung $B = R_B L_B$ die LR–Zerlegung $ABL_B^{-1} = L_A R_A R_B$ hinschreiben, ergibt sich $\det(ABL_B^{-1}) = \det(R_A)\det(R_B)$. Und weil sich nach der vor dem vorigen Satz angegebenen Überlegung eine Determinante nicht ändert, wenn man mit normierten Dreiecksmatrizen multipliziert, folgt die Behauptung.

Interessant ist schließlich der Fall reeller Orthogonalmatrizen, die wegen $A^{-1} = A^T$ und

$$1 = \det(I) = \det(A) \cdot \det(A^{-1}) = \det(A) \cdot \det(A^T) = \det(A)^2$$

die Determinante 1 oder -1 haben. Man kann ferner zeigen, daß Drehungen die Determinante 1 und Spiegelungen die Determinante -1 haben. Unitäre komplexe Matrizen haben eine im allgemeinen komplexe Determinante mit Betrag 1.

Die allgemeine Behandlung der Volumeneigenschaft verschieben wir auf den Abschnitt 9.3. Speziell für $n \times n$–Diagonalmatrizen D mit Zahlen $\lambda_1, \ldots, \lambda_n$ in der Diagonale ist das Volumen der Bildmenge aus (6.11) genau $|\lambda_1 \cdots \lambda_n|$ und dies stimmt mit dem Betrag der Determinante überein.

6.7 Vektorprodukt

Im $\mathbb{R}^3$ kann man zu 3 Vektoren a, b, c die manchmal auch als **Spatprodukt**[1] bezeichnete Determinante $[a, b, c] := det(a, b, c)$ bilden, die bis auf das Vorzeichen das Volumen des von den drei Vektoren aufgespannten dreidimensionalen Parallelogramms angibt. Das Spatprodukt verschwindet also genau dann, wenn die drei Vektoren linear abhängig sind. Das Spatprodukt ist linear in jeder Komponente und ändert sein Vorzeichen bei Vertauschung von Argumenten. Weil es linear in a und skalar ist, kann man einen als $b \times c$ bezeichneten Vektor angeben mit $[a, b, c] = a^T(b \times c)$. Man nennt $b \times c$ das **Vektorprodukt** oder **Kreuzprodukt**[2] von b und c, und aus der Determinantenform des Spatprodukts folgt

$$b \times c := (b_2c_3 - b_3c_2, b_3c_1 - b_1c_3, b_1c_2 - b_2c_1)^T.$$

Aus $[a, b, c] = a^T(b \times c)$ ergeben sich einige Eigenschaften des Vektorprodukts:

$$\begin{aligned} b \times c &= -(c \times b) \\ b^T(b \times c) &= 0 \\ c^T(b \times c) &= 0 \\ a^T(b \times c) &= -b^T(a \times c) \\ &= c^T(a \times b). \end{aligned}$$

Der Vektor $b \times c$ steht also auf b und c senkrecht, was ihn bei linear unabhängigen b, c bis auf einen Faktor festlegt. Normiert man ihn dann zur Länge 1, so gibt $[\frac{b \times c}{\|b \times c\|_2}, b, c] = \|b \times c\|_2$ bis auf ein Vorzeichen das Volumen des von den drei Vektoren aufgespannten dreidimensionalen Parallelogramms an, und weil $b \times c$ auf b und c senkrecht steht und normiert ist, ist das Volumen des dreidimensionalen Parallelogramms numerisch gleich dem Flächeninhalt des von b und c aufgespannten zweidimensionalen Parallelogramms. Also gibt $\|b \times c\|_2$ die Fläche des von b und c aufgespannten Parallelogramms an.

[1] `http://de.wikipedia.org/wiki/Spatprodukt`
[2] `http://de.wikipedia.org/wiki/Kreuzprodukt`

7 Geometrie

Die reellen Zahlen werden üblicherweise auf einer **Zahlengeraden** veranschaulicht, auf der jede reelle Zahl x als ein **Punkt** "liegt", und zwei Punkte $x, y \in \mathbb{R}$ den **Abstand** $|x - y|$ haben. In der Geometrie[1] spielen die Begriffe "*Punkt, Gerade, liegt auf, Abstand*" eine zentrale Rolle, und das wollen wir in diesem Kapitel ein wenig abstrahieren (David **Hilbert**[2] hat gesagt: "Man muß jederzeit an Stelle von 'Punkte, Gerade, Ebenen' 'Tische, Bänke, Bierseidel' sagen können".[3] Diese Abstraktion treiben wir aber nicht allzu weit, sondern beschränken uns auf den "normalen" zwei– und dreidimensionalen Raum. Dabei entwickeln wir Grundbegriffe und Grundoperationen aus der Geometrie, soweit sie für die **Computergraphik**, das **Computer–Aided Design** und gewisse Anwendungen im Wissenschaftlichen Rechnen nötig sind.

7.1 Geometrische Objekte

Im dreidimensionalen Raum sind die wichtigsten Objekte Punkte, Geraden und (Hyper–) Ebenen. Sie können wie allgemeinere Objekte, z.B. Kurven und Körper, notfalls als Punktmengen mathematisch beschrieben werden. Bei einer rein geometrischen Sichtweise gibt es aber zunächst keine Koordinatensysteme und keinen Abstandsbegriff. Dann ist die Beschreibung von Punktmengen nicht möglich und wird durch eine abstrakte Axiomatik ersetzt, die wir hier nicht in voller Ausführlichkeit schildern können. Dadurch geht leider auch die saubere Unterscheidung zwischen (z.B.) Euklidischer, affiner und projektiver Geometrie verloren. Statt eines puristisch geometrischen Ansatzes arbeiten wir hier ganz pragmatisch im $\mathbb{R}^3$ mit seinem Standard–Koordinatensystem und seinem Standard–Abstandsbegriff, der euklidischen Norm. Es wird sich später herausstellen, daß das nicht reicht, aber das ist jetzt noch nicht abzusehen. Aber als Kompromiß wollen wir versuchen, eine geometrische Sichtweise soweit möglich durchzuhalten.

Jedes Element $V = (x, y, z)$ des reellen Vektorraums $\mathbb{R}^3$ hat zwei Interpretationen. Es definiert einerseits einen Vektor als gerichtete Verbindungsstrecke zwischen dem Nullpunkt und einem **Punkt** V des $\mathbb{R}^3$, andererseits aber

[1] `http://de.wikipedia.org/wiki/Geometrie`

[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Hilbert.html`

[3] Diesen Spruch soll Hilbert, so O. Blumenthal, 1891 auf der Heimfahrt von Halle nach Königsberg nach dem Anhören eines Vortrages von Hermann Wiener geäußert haben. Quelle: Schreiber, Peter (1987). Euklid. Biographien hervorragender Naturwissenschaftler, Techniker und Mediziner. Bd. 87.. Leipzig: Teubner, S. 140

auch eben diesen Punkt. In der Geometrie sollte man aber zwischen Punkten und Vektoren klar unterscheiden. In der **Euklidischen**, **affinen** und der **projektiven** Geometrie gibt es keinen ausgezeichneten Punkt, und deshalb erst recht keinen "Nullpunkt". Punkte sind dann eben Punkte, und Vektoren kann man, wenn unbedingt nötig, als geordnete Paare von zwei Punkten definieren.

Die Auszeichnung eines Nullpunkts ist eine ziemlich willkürliche Sache und sollte erst im Zusammenhang mit der Einführung und Benutzung eines Koordinatensystems vorgenommen werden. Auch im praktischen Leben ist die Festlegung eines "Nullpunkts" immer nebensächlich, z.B. die Festlegung des Nullmeridians auf die Länge von Greenwich. Sie ändert nichts an der Geometrie der Gebäude Göttingens. Wenn man in der Computergraphik eine "Szene" zusammenbaut, ist die genaue Lage eines "Nullpunkts" ebenfalls nebensächlich. Wenn wir ab hier von Punkten reden, meinen wir keine Vektoren und ignorieren Koordinatensysteme und ihre Nullpunkte. Wenn die Sache aber konkret wird und man Punkte auf einen Rechner bringen will, führt man Koordinaten ein und stellt Punkte in einem Koordinatensystem dar.

7.2 Euklidische und affine Geometrie

Die **Euklidische** Geometrie ist die "übliche", die zwar keinen Nullpunkt kennt, aber sehr wohl Winkel definieren und von "senkrecht" reden kann. Sie entspricht, grob gesagt, allem, was man mit Zirkel und Lineal anstellen kann. Wird ein (Null–) Punkt ausgezeichnet, so spricht man genaugenommen von "polareuklidischer Geometrie". Hier werden wir die Euklidische Geometrie stillschweigend voraussetzen und nicht weiter behandeln.

Die Grundobjekte praktischer Geometrie sind Punkte, Geraden und (Hyper)–Ebenen. Man fragt nun danach, unter welchen Transformationen Punkte, Geraden und (Hyper)–Ebenen wieder in Punkte, Geraden und Ebenen übergehen. Aus der linearen Algebra kennen wir lineare und affine Transformationen, aber die linearen zeichnen den Nullpunkt aus und lassen ihn unverändert, was für unsere geometrischen Pläne unbrauchbar ist. Wir wollen natürlich auch Verschiebungen erlauben, und daher sind Affintransformationen richtig. Sie lassen sich im Vektorraum als Summe einer Verschiebung und einer linearen Transformation schreiben. Unter affinen Transformationen gehen Punkte, Geraden und Ebenen wieder in Punkte, Geraden und Ebenen über. Ferner führen Affinkombinationen von Punkten aus Geraden oder Ebenen nicht aus diesen hinaus. Deshalb werden wir den üblichen $I\!R^3$

geometrisch als affinen Punktraum sehen und unsere geometrischen Objekte mit Affintransformationen behandeln.

Leider reicht das aber für Computergraphik nicht aus. Wenn im $\mathbb{R}^3$ etwa ein gerader, bis zum Horizont reichender Schienenstrang durch zwei parallele, sich nirgends schneidende Geraden dargestellt wird, so brauchen wir eine Transformation, die ein photorealistisches Abbild auf einen Bildschim bringt. Dort ergeben sich dann aber zwei Geradenstücke, die sich in der Bildmitte schneiden, wenn der Horizont in der Bildmitte liegt. Deshalb kann der Abbildungsmechanismus nicht affin sein. Er würde Geraden in Geraden und Punkte in Punkte abbilden, aber ein affines Bild zweier sich nirgends schneidender Geraden kann sich nicht schneiden oder schneidet sich überall.

Rechnen wir das kurz vor. Die beiden sich nicht schneidenden Geraden seien als

$$\begin{aligned} G_1 &:= \{x \in \mathbb{R}^3 \ : \ x = x_1 + \alpha \cdot r, , \ \alpha \in \mathbb{R}\}, \\ G_2 &:= \{x \in \mathbb{R}^3 \ : \ x = x_2 + \alpha \cdot r, , \ \alpha \in \mathbb{R}\} \end{aligned}$$

mit linear unabhängigen Vektoren r und $x_1 - x_2$ dargestellt. Sie werden mit $x \mapsto Ax + b$ mit $b \in \mathbb{R}^2$ und einer Matrix $A \in \mathbb{R}^{3\times 2}$ affin transformiert. Schneiden sich zwei Bilder, so folgt mit $\alpha_1, \alpha_2 \in \mathbb{R}$ die Gleichung

$$\begin{aligned} A(x_1 + \alpha_1 r) + b &= A(x_2 + \alpha_2 r) + b \\ Ax_1 + b &= Ax_2 + \alpha_2 \cdot Ar + b - \alpha_1 \cdot Ar \\ A(x_1 + \beta r) + b &= Ax_1 + b + \beta Ar \\ &= Ax_2 + \alpha_2 \cdot Ar + b - \alpha_1 Ar + \beta Ar \\ &= A(x_2 + (\alpha_2 - \alpha_1 + \beta)r) + b, \end{aligned}$$

für alle $\beta \in \mathbb{R}$, d.h. die Bildgeraden fallen komplett zusammen.

Wir müssen also die bequeme affine und euklidische Geometrie verlassen. Erhalten bleiben soll, daß Geraden in Geraden und Ebenen in Ebenen übergehen sollten, und daß die Aussage "Der Punkt P liegt auf der Geraden G" nach Transformation mit T in "Der Punkt $T(P)$ liegt auf der Geraden $T(G)$" übergehen sollte. Wir machen das ganz allgemein, indem wir eine Geometrie aufbauen, in der man von Punkten, Geraden, "liegt auf" und "schneidet" reden kann (**Inzidenzgeometrie**). Weil sich schneidende Geraden in sich schneidende Geraden übergehen sollten, wird der Ausweg sein, die beiden parallelen Schienen sich in einem unendlich fernen Punkt schneiden zu lassen, der dann ganz konkret in einen Schnittpunkt auf dem Bildschirm transformiert wird.

7.3 Ebene projektive Geometrie

In jeder vernünftigen Geometrie ist es so, daß man von einer Menge $\mathcal{P}$ von Punkten ausgeht und dann zu je zwei verschiedenen Punkten $P, Q \in \mathcal{P}$ eine **Gerade** $G := \overline{PQ}$ definiert. Dadurch bekommt man eine Menge $\mathcal{G}$ von Geraden, und man fordert eine **Inzidenzrelation** auf $\mathcal{P} \times \mathcal{G}$, die zu einem Paar $(P, G) \in \mathcal{P} \times \mathcal{G}$ angibt, ob der Punkt P auf der Geraden G "liegt" oder mit ihr "inzidiert". Man schreibt $P \in G$, falls P auf G liegt, bzw. wenn die Inzidenzrelation erfüllt ist. Dadurch bekommt man durch die Hintertür eine Interpretation von Geraden als Punktmengen, aber eigentlich sollte man ein anderes Symbol als $\in$ verwenden, denn unsere einzigen Mengen sind $\mathcal{P}$ und $\mathcal{G}$. Natürlich sollte immer $P \in \overline{PQ}$ und $Q \in \overline{PQ}$ gelten, aber es ist nicht klar, ob noch weitere Punkte auf so einer "Geraden" "liegen" und wieviele es sind.

Im nächsten Schritt legt man fest, ob und wann sich zwei verschiedene Geraden $G, H \in \mathcal{G}$ "schneiden", d.h. ob es einen "Schnittpunkt" $P \in \mathcal{P}$ mit $P \in G$ und $P \in H$ gibt.

Definition 7.1 *Eine* **projektive Ebene** *besteht aus nichtleeren Mengen $\mathcal{P}$ und $\mathcal{G}$ sowie einer Relation $\in$ auf $\mathcal{P} \times \mathcal{G}$ mit den Eigenschaften*

1. *Für alle $P \neq Q \in \mathcal{P}$ gibt es genau ein $G \in \mathcal{G}$ mit $P \in G$, $Q \in G$.*
2. *Für alle $G \neq H \in \mathcal{G}$ gibt es genau ein $P \in \mathcal{P}$ mit $H \ni P$, $G \ni P$.*

Sie heißt **endlich**, *wenn $\mathcal{P}$ und $\mathcal{G}$ endlich sind.*

Wenn man nur eine Ebene damit geometrisch modellieren will, ist hier Schluß, denn es gibt nur Punkte und Geraden. Um zu demonstrieren, daß man sich an eine formale Axiomatik halten muß, folgt ein Beispiel. Wir definieren eine endliche projektive Ebene durch $\mathcal{P} := \{P, Q, R\}$ und $\mathcal{G} = \{F, G, H\}$ mit der Inzidenzrelation aus Tabelle 2. Eine Veranschaulichung bietet Figur 4. In der Vorlesung wird auch noch eine andere endliche projektive Ebene angegeben, bei der es 7 Punkte und 7 Geraden gibt, so daß je drei Punkte genau eine Gerade definieren und je drei Geraden sich in genau einem Punkt schneiden.

Es zeigt sich, daß man die Begriffe "Gerade" und "Punkt" vertauschen kann, wenn man gleichzeitig "liegt auf" mit "schneidet in" vertauscht oder $\in$ in $\ni$ umdreht. Denn zu je zwei verschiedenen Punkten gibt es genau eine Gerade, auf der diese Punkte liegen, und zu je zwei verschiedenen Geraden gibt es genau einen Punkt, in dem sich die Geraden schneiden.

	F	G	H
P	T	T	F
Q	T	F	T
R	F	T	T

Tabelle 2: Inzidenzrelation einer dreipunktigen projektiven Ebene.

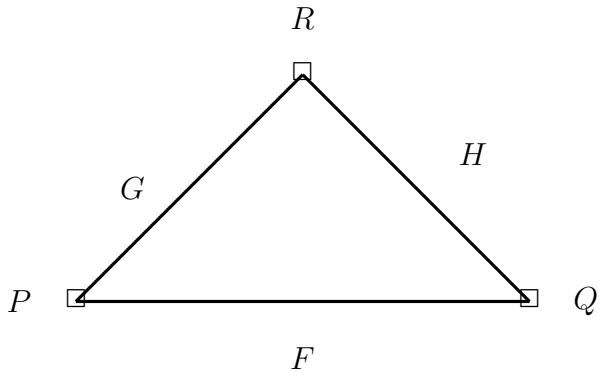


Abbildung 4: Dreipunktige projektive Ebene

In der “normalen” Ebene kann man Punkte und Geraden “normal” veranschaulichen, aber man sagt, zwei verschieden Geraden seien **parallel**, wenn sie sich **nicht** schneiden. Und weil es bei normaler Anschauung parallele Geraden gibt, die sich nicht schneiden, ist die “normale” Ebene nicht projektiv, denn in einer projektiven Ebene schneiden sich alle voneinander veschiedenen Geraden wegen der obigen Definition.

Zwei parallele Geraden in der “normalen” Ebene haben aber immer eine eindeutige gemeinsame vorzeichenlose “Richtung”. Die “Richtung” einer Geraden kann man dann als “unendlich fernen Punkt” ansehen, der auf der Geraden ”liegt”. Man erweitert die Menge der “normalen” oder “endlichen” Punkte durch diese “unendlich fernen Punkte” und bekommt eine neue, größere formale Punktmenge. Je ein endlicher und ein unendlich ferner Punkt definieren dann genau eine “normale” Gerade, nämlich diejenige, die erstens durch den gegebenen endlichen Punkt geht und zweitens die durch den unendlich fernen Punkt gegebene Richtung hat. Jetzt schneiden sich alle Geraden in genau einem Punkt. Wunderbar.

Aber man braucht dann noch eine Gerade zu je zwei verschiedenen unendlich

fernen Punkten. Eine "normale" Gerade mit zwei verschiedenen Richtungen gibt es nicht. Deshalb definiert man eine neue "unendlich ferne Gerade", auf der genau die unendlich fernen Punkte liegen, sonst keine. Man kann sie sich veranschaulichen als einen unendlich fernen Kreis um die Ebene, wobei die Verbindung eines Kreispunkts (d.h. eines unendl;ich fernen Punkts) mit einem gedachten Kreismittelpunkt die "Richtung" der zum unendlich fernen Punkt gehörenden Geraden angibt (Zeichnung in der Vorlesung). Damit bekommt man eine Erweiterung der "normalen" Ebene zu einer projektiven Ebene **über $\mathbb{R}$**, die allen Anforderungen genügt.

Aber wie soll man so etwas in der Informatik realisieren? Ist das ein rein theoretisches Gedankenspiel?

Nein, denn moderne Methoden der Computergraphik und des Computer–Aided Design benutzen sogar die projektive Geometrie des **Raumes**, die wir weiter unten behandeln, und zwar beim Abbilden dreidimensionaler Szenen auf den Bildschirm. Damit das nicht zu schwierig wird, wollen wir erst einmal die Realisierung der projektiven **Ebene** über $\mathbb{R}$ behandeln.

Dazu überlegen wir uns, wie man Geraden realisieren sollte. Die Schulform $y = a \cdot x + b,\ a, b \in \mathbb{R}$ ist unbrauchbar, weil man damit keine Geraden realisieren kann, die parallel zur y–Achse sind. Die Form $x = a \cdot y + b,\ a, b \in \mathbb{R}$ ist unbrauchbar, weil man damit keine Geraden realisieren kann, die parallel zur x–Achse sind. Besser ist $a \cdot x + b \cdot y + c = 0$ mit $a, b, c \in \mathbb{R},\ a^2 + b^2 \neq 0$. Die Punktmenge

$$\{(x, y) \in \mathbb{R}^2 \ : \ a \cdot x + b \cdot y + c = 0\} \tag{7.2}$$

ist dann immer eine "normale" Gerade mit einer durch $(a, b) \neq 0$ bestimmten Richtung. Eine solche Gerade kann man aber auch durch

$$\{(x, y) \in \mathbb{R}^2 \ : \ a \cdot d \cdot x + b \cdot d \cdot y + c \cdot d = 0\}$$

mit beliebigem $d \neq 0$ beschreiben. Die Geraden lassen sich also darstellen durch Äquivalenklassen $[a, b, c]$ von Tripeln $(a, b, c) \in \mathbb{R}^3 \setminus \{0\}$,wenn wir zwei Tripel $(a, b, c), (a', b', c') \in \mathbb{R}^3 \setminus \{0\}$ als äquivalent bezeichnen, wenn die durch Multiplikation mit einer von Null verschiedenen reellen Zahl ineinander transformierbar sind:

$$[a, b, c] = [a', b', c'], \quad \text{wenn } d \cdot (a, b, c) = (a', b', c') \text{ mit } d \neq 0. \tag{7.3}$$

Die "normalen" Geraden brauchen zusätzlich noch $(a, b) \neq (0, 0)$, und wir werden sehen, daß die unendlich ferne Gerade der Äquivalenzklasse $[0, 0, 1]$ entspricht.

Die Darstellung von Punkten sollte aber ziemlich genau der von Geraden entsprechen, damit der Dualismus zwischen Punkten und Geraden funktioniert. Man sollte also ebenfalls die Punkte durch Äquivalenzklassen $[u, v, w]$ mit $(u, v, w) \neq 0 \in \mathbb{R}^3$ darstellen. Die "normalen" Punkte $(a, b) \in \mathbb{R}^2$ sind Äquivalenzklassen der Form $[a, b, 1]$, während unendlich ferne Punkte Äquivalenzklassen der Form $[a, b, 0]$ sind, wobei (bis auf einen gemeinsamen Faktor) das Paar $(a, b) \neq 0$ der Richtung der "normalen" Geraden zur Äquivalenzklasse $[a, b, c]$ mit beliebigem c entspricht. Ein endlicher oder unendlich ferner Punkt $[u, v, w]$ liegt auf der "normalen" oder unendlich fernen Geraden $[a, b, c]$ genau dann, wenn die homogene Gleichung $au + bv + cw = 0$ gilt. Man spricht deshalb von **homogenen Koordinaten**.

Theorem 7.4 *Die projektive Ebene über $\mathbb{R}$ kann in homogenen Koordinaten folgendermaßen realisiert werden:*

1. *Grundmenge ist*

$$\mathbb{P}^2 := \{[a, b, c] \ : \ (a, b, c) \in \mathbb{R}^3 \setminus \{0\}\}$$

   mit der Äquivalenzrelation (7.3).

2. *Projektive Punkte und Geraden sind eindeutig bestimmt durch Elemente von $\mathbb{P}^2$ bzw. durch eindimensionale Unterräume von $\mathbb{R}^3$.*

3. *Ein projektiver Punkt $P = [u, v, w]$ liegt genau dann auf einer projektiven Geraden $[a, b, c]$, wenn $au + bv + cw = 0$ gilt.*

4. *Zwei projektive Punkte oder Geraden $X = [a, b, c]$, $Y = [a', b', c']$ sind genau dann verschieden, wenn die Vektoren (a, b, c) (a', b', c') des $\mathbb{R}^3$ linear unabhängig sind. In diesem Falle ist der Vektor $(a, b, c) \times (a', b', c')$ des $\mathbb{R}^3$ nicht Null und steht auf (a, b, c) (a', b', c') senkrecht. Seine Äquivalenzklasse kann man mit $X \times Y := [a, b, c] \times [a', b', c']$ bezeichnen und sowohl als projektiven Punkt als auch als projektive Gerade interpretieren.*

   (a) *Sind X und Y projektive Punkte, so ist $X \times Y$ die projektive Gerade, auf der X und Y liegen.*

   (b) *Sind X und Y projektive Geraden, so ist $X \times Y$ der projektive Punkt, in dem sich X und Y schneiden.*

   (c) *Ist X ein projektiver Punkt und Y eine projektive Gerade, so definiert $X \times Y$ sowohl einen Punkt P als euch eine Gerade G, und zwar so, daß sich einerseits Y und G in P schneiden und andererseits G die X und Q verbindende projektive Gerade ist.*

5. *Ein Punkt* (a, b) *der üblichen Ebene* $\mathbb{R}^2$ *entspricht einem Punkt* $[a, b, 1]$ *der projektiven Ebene.*

6. *Eine "normale" Gerade* (7.2) *mit* $(a, b) \neq 0$ *ist eine Gerade* $[a, b, c]$ *der projektiven Ebene.*

7. *Die unendlich ferne Gerade ist* $[0, 0, 1]$.

8. *Unendlich ferne Punkte sind* $[a, b, 0]$ *mit* $(a, b) \neq 0$.

Diesen Satz wollen wir nicht im Detail beweisen, aber wir ziehen ein paar exemplarische Folgerungen, die das Ganze etwas beleuchten.

> Der klassische Schnittpunkt zweier nicht–paralleler klassischer Geraden entspricht dem endlichen projektiven Punkt, der sich als Schnittpunkt der projektiven Geraden ergibt.

Beweis: Schreibt man die Geraden mit (a, b, c) und (a', b', c') im Sinne von (7.2), so sind sie genau dann nicht–parallel, wenn $ab' \neq a'b$ gilt. Das ist wiederum äquivalent dazu, daß die dritte Komponente c'' von $(a, b, c) \times (a', b', c') = (a'', b'', c'')$ nicht verschwindet. Dann ist $[a'', b'', c''] = [a''/c'', b''/c'', 1]$ ein endlicher projektiver Punkt, und $(a''/c'', b''/c'')$ liegt im klassischen Sinn auf beiden klassischen Geraden, denn $(a''/c'', b''/c'', 1)$ steht in klassischem Sinn senkrecht auf (a, b, c) und (a', b', c').

> Auf der unendlich fernen Geraden $[0, 0, 1]$ liegen nur unendlich ferne Punkte.

Beweis: Liegt $[a, b, c]$ auf $[0, 0, 1]$, so folgt $a \cdot 0 + b \cdot 0 + c \cdot 1 = c = 0$, d.h. der Punkt ist ein unendlich ferner.

> Die unendlich ferne und eine andere projektive Gerade schneiden sich immer in einem unendlich fernen Punkt.

Beweis: Schneidet man $[0, 0, 1]$ und $[a, b, c]$, so folgt $[0, 0, 1] \times [a, b, c] = [-b, a, 0]$.

> Ein endlicher und ein unendlich ferner Punkt definieren genau eine "normale" Gerade.

Beweis: Weil kein endlicher Punkt auf der unendlich fernen Geraden liegen kann, ist das klar. Man kann es aber auch ausrechnen: Die Punkte seien $[a, b, 1]$ und $[a', b', 0]$. Es folgt $[a, b, 1] \times [a', b', 0] = [-b', a', ab' - ba']$. Wäre dies die unendlich ferne Gerade, so müßte $a' = b' = 0$ gelten, was unerlaubt ist, denn es muß $(a', b', 0) \neq 0 \in \mathbb{R}^3$ gelten.

In homogenen Koordinaten kann man Kegelschnitte sehr viel schöner darstellen als "herkömmlich", weil (z.B.) Ellipsen, Parabeln und Hyperbeln keinen, genau einen oder genau zwei Schnittpunkte mit der unendlich fernen Geraden haben. Aber leider ist für so etwas weder Zeit noch Raum in dieser Vorlesung.

7.4 Projektive Geometrie des Raumes

Will man einen "dreidimensionalen" Raum modellieren (Vorsicht: "Dimension" ist hier bei uns nur als Begriff der linearen Algebra definiert), so braucht man auch noch (Hyper-) Ebenen. Diese bilden eine Menge $\mathcal{E}$, und man fordert, daß es zu je einer Geraden $g \in \mathcal{G}$ und zu einem Punkt P, der nicht mit G inzidiert, es eine Ebene E geben soll, die mit P und G inzidiert. Zu je zwei verschiedenen Ebenen soll es obendrein genau eine Gerade geben, in der sich die Ebenen schneiden.

Wir gehen jetzt ziemlich brutal vor und modellieren das Ganze mit homogenen Koordinaten in Analogie zum Fall der projektiven Ebene.

Theorem 7.5 *Der projektive Raum über $\mathbb{R}$ kann in homogenen Koordinaten folgendermaßen realisiert werden:*

1. *Grundmenge ist*

   $$\mathbb{P}^3 := \{[a,b,c,d] \ : \ (a,b,c,d) \in \mathbb{R}^4 \setminus \{0\}\}$$

   mit der Äquivalenzrelation

   $[a,b,c,d] = [a',b',c',d']$ *genau dann, wenn* $(a,b,c,d) = z\cdot(a',b',c',d'),\ z \neq 0$.

   Wir schreiben auch $[p] \in \mathbb{P}^3$ mit $p \in \mathbb{R}^4 \setminus \{0\}$. Zwei Elemente $[p]$ und $[q]$ von $\mathbb{P}^3$ sind genau dann verschieden, wenn die Vektoren p und q des $\mathbb{R}^4$ linear unabhängig sind.

2. *Projektive Punkte und Ebenen sind eindeutig bestimmt durch Elemente von $\mathbb{P}^3$.*

3. *Projektive Punkte $P = [p]$ sind eindeutig bestimmt durch die eindimensionalen Teilräume span (p) von $\mathbb{R}^4$.*

4. *Projektive Ebenen $E = [e]$ sind eindeutig bestimmt durch die dreidimensionalen Teilräume $(\text{span}\ (e))^\perp$ von $\mathbb{R}^4$.*

5. *Ein projektiver Punkt $P = [p]$ liegt genau dann auf einer projektiven Ebene $E = [e]$, wenn $p^T e = 0$ gilt, d.h. p auf e senkrecht steht.*

6. *Projektive Geraden sind bestimmt durch zweidimensionale Teilräume U von $\mathbb{R}^4$. Man kann sie auf zwei Weisen beschreiben:*

   (a) *in* **Punktdarstellung** *als Spann von zwei verschiedenen Punkten $P = [p]$ und $Q = [q]$. Dann enthalten sie alle Punkte $R = [r]$ mit $r \in span(p, q)$ und sind in allen Ebenen $E = [e]$ mit $e^T p = 0 = e^T q$ enthalten. Die Vektoren p und q spannen U auf.*

   (b) *in* **Ebenendarstellung** *als Schnitt von zwei verschiedenen Ebenen $E = [e]$ und $F = [f]$. Dann enthalten sie alle Punkte $P = [p]$ mit $p^T e = 0 = p^T f$ und sind in allen Ebenen $G = [g]$ mit $g \in span(e, f)$ enthalten. Die Vektoren e und f spannen $U^\perp$ auf.*

7. *Zwei projektive Punkte oder Ebenen $X = [x]$, $Y = [y]$ sind genau dann verschieden, wenn die Vektoren x, y des $\mathbb{R}^4$ linear unabhängig sind. Zu dem von beiden Vektoren aufgespannten Teilraum U des $\mathbb{R}^4$ gibt es einen eindeutig bestimmten zweidimensionalen Orthogonalraum V. Beide Teilräume können als projektive Gerade G_U bzw. G_V aufgefaßt werden.*
   Zwei aufspannende Vektoren aus U beschreiben G_U in Punktdarstellung, zwei aufspannende Vektoren aus V beschreiben G_U in Ebenendarstellung.
   Zwei aufspannende Vektoren aus V beschreiben G_V in Punktdarstellung, zwei aufspannende Vektoren aus U beschreiben G_V in Ebenendarstellung.
   Sind X und Y projektive Punkte, so ist G_U die X und Y verbindende projektive Gerade.
   Sind X und Y projektive Ebenen, so ist G_V die projektive Schnittgerade von X und Y.

8. *Zwei verschiedene projektive Ebenen schneiden sich in genau einer projektiven Geraden.*

9. *Zwei verschiedene projektive Punkte liegen auf genau einer projektiven Geraden.*

10. *Eine projektive Ebene und eine darin nicht enthaltene projektive Gerade schneiden sich in genau einem projektiven Punkt.*

11. *Ein projektiver Punkt und eine projektive Gerade, auf der der projektive Punkt nicht liegt, definieren genau eine projektive Ebene, auf der beide liegen.*

12. *Ein Punkt* (a, b, c) *des üblichen Raumes* $\mathbb{R}^3$ *entspricht einem Punkt* $[a, b, c, 1]$ *des projektiven Raums.*

13. *Eine "normale" (Hyper-) Ebene*

$$\{(x, y, z) \in \mathbb{R}^3 \ : \ a \cdot x + b \cdot y + c \cdot z + d = 0\}$$

mit $(a, b, c) \neq 0$ *ist eine projektive Ebene* $[a, b, c, d]$.

14. *Die unendlich ferne projektive Ebene ist* $[0, 0, 0, 1]$.

15. *Unendlich ferne Punkte sind* $[a, b, c, 0]$ *mit* $(a, b, c) \neq 0$.

16. *Unendlich ferne Geraden sind zweidimensionale Teilräume von* $\mathbb{R}^3 \times \{0\}$.

Auch mit diesen Aussagen sollte man herumspielen, aber das wollen wir jetzt den Übungen überlassen und uns stattdessen fragen, welche Transformationen die Punkte, Geraden und Ebenen des projektiven Raumes wieder in Punkte, Geraden und Ebenen des projektiven Raumes transformieren und dabei Inzidenzen erhalten. Wir ignorieren dabei große Teile der projektiven Geometrie und beschränken uns auf das technisch Nötige.

Zunächst stellen wir fest, daß eine "normale" Affinkombination

$$\sum_{j=1}^{n} \alpha_j x_j \text{ mit } x_j \in \mathbb{R}^3 \text{ und } \alpha_j \in \mathbb{R} \text{ mit } 1 = \sum_{j=1}^{n} \alpha_j$$

wenn man sie stattdessen auf endliche projektive Punkte der Form $[x_j, 1]$ anwendet, zu einem endlichen Punkt

$$\sum_{j=1}^{n} \alpha_j (x_j, 1) = (\sum_{j=1}^{n} \alpha_j x_j, 1)$$

führt, der auch "normal" herausgekommen wäre. Umgekehrt kann man Affinkombinationen auf die homogenen Koordinaten projektiver Punkte anwenden und sieht, das sie die endlichen Punkte genau so transformieren wie das in der affinen Punktgeometrie stattfinden würde. Man kann also Affinkombinationen uneingeschränkt auf Punkte von Geraden oder Ebenen in homogenen Koordinaten ausführen, ohne die Geraden und Ebenen zu verlassen. Schließlich kann man noch sehen, daß solche Kombinationen die unendlich ferne Ebene fest lassen.

Es sei nun $x \mapsto Ax + b$ eine affine Transformation auf dem $\mathbb{R}^3$. Sie sollte auf endlichen projektiven Punkten $[x, 1]$ als $[Ax + b, 1]$ wirken. Das kann man als vierdimensionale lineare Transformation

$$\begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} Ax + b \\ 1 \end{pmatrix}$$

schreiben. Deshalb verliert man nichts, wenn man sich auf lineare Transformationen auf dem $\mathbb{R}^4$ beschränkt, die auf homogene Koordinaten wirken.

7.5 Projektionen in der Computergraphik

Wir wollen eine dreidimensionale Szene, die wir kunstvoll im $\mathbb{R}^3$ aufgebaut haben, auf einem Bildschirm anzeigen. Die Szene selbst kann aus Punkten, Kurven, Flächen und Körpern bestehen, die in dreidimensionalen "**Weltkoordinaten**" dargestellt sind. Zunächst ignorieren wir das Problem, ob unser Bildausschnitt die gesamte Szene oder nur einen Ausschnitt zeigen soll. Aber wir wollen die Szene genau so darstellen, wie sie dem Auge des Betrachters erscheint, insbesondere dann, wenn sich der Betrachter relativ zur Szene bewegt. Dabei ist die Tiefeninformation relativ zum Betrachter wichtig, denn der Vordergrund muß den Hintergrund überdecken. Diese Information gehört nicht zur Szene, sondern sie hängt von der Szene und dem jeweiligen Betrachter ab. Sie muß bei bewegten Betrachtern immer neu berechnet werden, z.B. wenn der Betrachter durch ein Säulenlabyrinth läuft.

In den Weltkoordinaten befindet sich also auch ein **Augpunkt** A, von dem aus der Betrachter die Szene sieht. Ferner ist die Betrachtungsrichtung V (**view vector**) wichtig, die als Einheitsvektor im Augpunkt die Blickrichtung des Betrachters angibt. Es wrd ferner angenomen, daß der Betrachter die Szene durch einen Sichtrahmen (**viewport**) sieht, der einem fiktiven Bildschirm entspricht, der zwischen Augpunkt und Szene liegt und die konkrete Bildinformation trägt. Was zum Bild des Betrachters beiträgt, liegt auf Sehstrahlen des Betrachters vom Augpunkt durch den Sichtrahmen in die Szene. Das ergibt einen Kegel sichtbarer Punkte, und wenn man die Tiefe der Szene als begrenzt annimmt, ergibt sich ein **Sichtvolumen** (**view volume**).

Die wichtigste Abbildungsoperation der Computergraphik ist die Projektion von Szenenpunkten auf den Sichtrahmen. Das geschieht durch Verbinden des Augpunktes A mit einem Szenenpunkt S durch eine Gerade, din im Bildpunkt B den Sichtrahmen schneidet. Der Sichtrahmen befindet sich im Abstand d vom Augpunkt auf dem Sichtvektor V und steht auf dem Sichtvektor senkrecht. Die Bildebene des Sichtrahmens wird durch zwei orthogonale

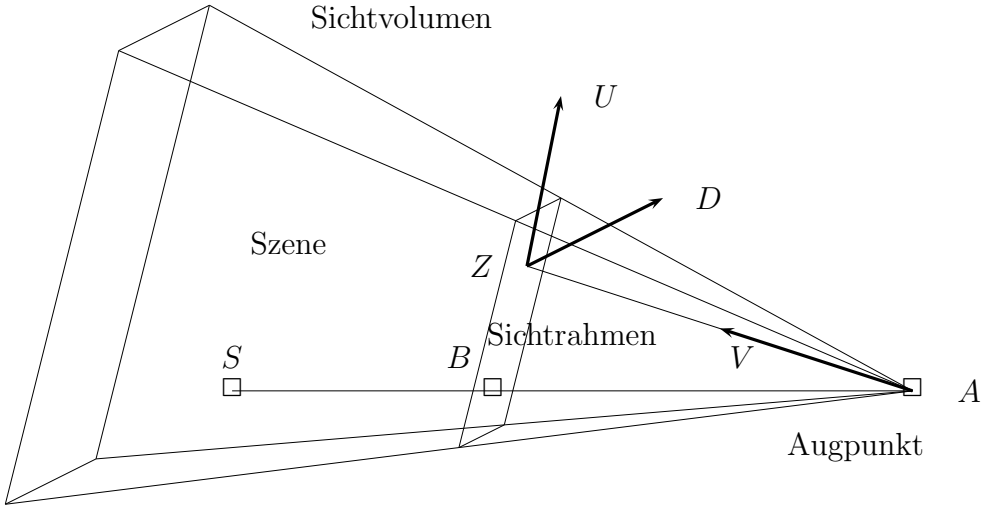


Abbildung 5: Sichttransformation

Vektoren U und D aufgespannt. Man nennt U den **view–up vector**, denn er gibt an, was im Bild "oben" ist. Weil wir hier alles in affiner Geometrie darstellen wollen, benutzen wir keinen Nullpunkt in Weltkoordinaten. Aber der Punkt Z kann als neuer Nullpunkt der Bildkoordinaten angesehen werden. Ein beliebiger Bildpunkt B hat dann als Vektor des $\mathbb{R}^3$ die Form $Z+\alpha U+\beta D$ mit den Bildkoordinaten $\alpha,\ \beta$. Die drei Vektoren V, U, D bilden ein Orthonormalsystem.

Jetzt wollen wir die Sichttransformation ausrechnen. Der Punkt B im Bild liegt auf der Verbindungsstrecke zwischen Augpunkt A und Szenenpunkt S. Er hat also die Form $B = \lambda A + (1-\lambda)S = S + \lambda(A-S)$. Damit haben wir in

$$S + \lambda(A - S) \quad = \quad Z + \alpha U + \beta D$$

drei Gleichungen mit drei Unbekannten. Wir nutzen aus, was wir wissen, nämlich die Orthonormalitäten und die Gleichung $Z = A + dV$. Weil wir später den Augpunkt A nach Unendlich schicken wollen, rechnen wir alles

auf Z um. Es folgt

$$\begin{aligned} S + \lambda(Z - dV - S) &= Z + \alpha U + \beta D \\ (1-\lambda)(S - Z) &= d\lambda V + \alpha U + \beta D \\ (1-\lambda)V^T(S - Z) &= d\lambda \\ \lambda &= \frac{V^T(S-Z)}{V^T(S-Z)+d} \\ 1-\lambda &= \frac{d}{V^T(S-Z)+d} \\ (1-\lambda)U^T(S-Z) &= \alpha \\ (1-\lambda)D^T(S-Z) &= \beta. \end{aligned}$$

Weil der Augpunkt nie im Sichtrahmen liegen sollte, kann man $d \neq 0$ annehmen und $\sigma := 1/d$ einführen. Dann ergibt sich

$$\begin{aligned} \lambda &= \frac{\sigma V^T(S-Z)}{\sigma V^T(S-Z)+1} \\ 1-\lambda &= \frac{1}{\sigma V^T(S-Z)+1} \\ \alpha &= \frac{U^T(S-Z)}{\sigma V^T(S-Z)+1} \\ \beta &= \frac{D^T(S-Z)}{\sigma V^T(S-Z)+1}. \end{aligned}$$

Alle drei Transformationen sind als Funktion von S weder linear noch affin, sondern rational! Genaugenommen sind sie Quotienten von zwei affinen Abbildungen. Aber wenn wir in der Bildebene homogene Koordinaten einführen, folgt

$$\begin{aligned} [\alpha, \beta, 1] &= \left[\frac{U^T(S-Z)}{\sigma V^T(S-Z)+1}, \frac{D^T(S-Z)}{\sigma V^T(S-Z)+1}, 1\right] \\ &= \left[U^T(S-Z), D^T(S-Z), \sigma V^T(S-Z)+1\right] \\ &= \left[(U^T, D^T, \sigma V^T)S - (U^T, D^T, \sigma V^T)Z + (0,0,1)\right] \\ &= \left[\begin{pmatrix} U^T \\ D^T \\ \sigma V^T \end{pmatrix}(S-Z) + (0,0,1)\right] \end{aligned}$$

und das ist in homogenen Koordinaten eine rein affine Transformation.

Die Tiefeninformation steckt in λ, wobei $\lambda = 0$ gilt, wenn $S - Z$ senkrecht ist zu V, d.h. wenn S im Bildrahmen liegt. Hinter dem Bildrahmen liegende Punkte S haben eine positive Tiefenkoordinate. Deshalb kann man die Szene so umrechnen, daß sie im orthonormalen Koordinatensystem von V, U und

D mit Ursprung in Z neu dargestellt wird, und zwar mit den üblichen Koordinaten (α, β, λ) oder den homogenen Koordinaten $[\alpha, \beta, \lambda, 1]$. Es folgt

$$\begin{aligned}
[\alpha, \beta, \lambda, 1] &= \left[\frac{U^T(S-Z)}{\sigma V^T(S-Z)+1}, \frac{D^T(S-Z)}{\sigma V^T(S-Z)+1}, \frac{\sigma V^T(S-Z)}{\sigma V^T(S-Z)+1}, 1\right] \\
&= \left[U^T(S-Z), D^T(S-Z), V^T(S-Z), \sigma V^T(S-Z)+1\right] \\
&= \left[\begin{pmatrix} U^T \\ D^T \\ V^T \\ \sigma V^T \end{pmatrix} (S-Z) + (0,0,0,1)\right]
\end{aligned} \tag{7.6}$$

und auch dies ist eine affine Transformation, diesmal im $\mathbb{R}^4$.

In diesem neuen Koordinatensystem hat der Augpunkt wegen $A = Z - dV$ immer die Form $A = [0, 0, -d, 1]$. Wird er nach Unendlich verschoben, so strebt d gegen Unendlich und σ gegen 0. In homogenen Koordinaten wird dann A über $A = [0, 0, -d, 1] = [0, 0, -1, 1/d] \to [0, 0, -1, 0]$ zum unendlich fernen Punkt $[0, 0, -1, 0]$, der von Z aus in Richtung $-V$ im Unendlichen liegt. Die Sehstrahlen sind dann alle parallel zu V, und man spricht von einer **Parallelprojektion**. Bei endlichem Augpunkt hat man eine **Zentralprojektion**. Die Darstellung in homogenen Koordinaten zeigt, daß man beide Fälle sauber parallel behandeln kann. Wenn man alle geometrischen Transformationen als affine Transformationen in vierdimensionalen homogenen Koordinaten schreibt, hat man keinerlei Fallunterscheidungen zu machen.

7.6 Tiefenpufferverfahren

Bei der Anzeige von Szenen gehört zu jeder Bitposition des Bildes je ein Paar (α, β) von diskreten Bildkoordinaten. Zu jedem solchen Koordinatenpaar gehört ein Strahl von Augpunkt A durch den Bildpunkt $B = Z + \alpha U + \beta D$. Auf diesem Strahl ist nur diejenige Bildinformation anzuzeigen, die vom am weitesten "vorn" gelegenen Szenenpunkt kommt, d.h. die mit kleinstem λ. Dabei ist es durchaus erlaubt, das der Augpunkt im Unendlichen liegt, denn Parallelprojektion ist nicht nur zulässig, sondern oft auch wünschenswert.

Die Standard-Anzeigetechnik, wie sie von heutigen Grafiksystemen realisiert wird, speichert zu jedem Koordinatenpaar nicht nur die Bildinformation, sondern in einem zusätzlichen **Tiefenpuffer** (z–buffer) zu allen bisherigen Szenenpunkten, die auf dem Sehstrahl durch (α, β) liegen, auch die bislang kleinste Tiefenkoordinate λ. Soll ein neuer Szenenpunkt angezeigt werden, so wird zuerst das Bildkoordinatenpaar (α, β) zusammen mit der Tiefenkoordinate λ ausgerechnet, und zwar nach der Formel (7.6) in homogenen

Koordinaten. Dann wird im Tiefenpuffer nachgesehen, ob der dort gespeicherte Wert größer als λ ist, Wenn nein, ist der neue Punkt "hinten", und man kann einen neuen Szenenpunkt ausrechnen und den soeben berechneten vergessen. Andernfalls ist der Punkt "vorn", und seine reale Farbinformation muß berechnet werden und in den Bildspeicher eingetragen werden. Im Tiefenpuffer wird der neue Wert von λ abgelegt und ein neuer Szenenpunkt kommt dran.

Das Verfahren hat den Vorteil, daß man den Durchlauf durch die Szenenpunkte nicht strukturieren muß. Es wird in allen modernen Hochleistungs–Graphiksystemen zusammen mit der vierdimensionalen Transformation (7.6) in homogenen Koordinaten realisiert. Deshalb gehört es zusammen mit dem Verständnis projektiver Geometrie zum Grundwissen der Informatik–Studierenden.

8 Folgen

Wir haben bisher Mengenlehre, Logik, ein wenig Zahlentheorie und dann lineare Algebra und Geometrie betrieben. Dabei gab es schon mehrfach Anlaß, von "Unendlich" zu reden, zum Beispiel als es um $\mathbb{N}$, um abzählbar unendliche Mengen oder um unendlichdimensionale Vektorräume wie den Polynomraum ging. In diesem Kapitel wird nun der Begriff des Grenzwerts oder des Limes eingeführt. Damit beginnt die Disziplin "Analysis[1]", und der Umgang mit "Unendlich" wird vom Sonderfall zum Normalfall. Die Differential– und Integralrechnung sowie die Differentialgeometrie basieren direkt auf dem Grenzwertbegriff, und weitergehende anwendungsbezogene Disziplinen wie Wahrscheinlichkeitstheorie und Differentialgleichungen sind ohne das Differenzieren und Integrieren undenkbar. Auch in der informatik brauchen viele Anwendungen Ergebnisse der Differential– und Integralrechnung, der Wahrscheinlichkeitstheorie oder Differentialgeometrie, und weil diese Disziplinen sämtlich den Grenzwertbegriff voraussetzen, ist dieses Kapitel, obwohl es nur indirekt Bezug zur Informatik hat, sehr wichtig für die Ausbildung der Informatikstudierenden.

8.1 Reelle Zahlenfolgen

Wir halten uns hier zuerst an Folgen reeller Zahlen. Die allgemeinen Folgen in metrischen Räumen holen wir aber nach.

8.1.1 Konvergenz von Folgen

Definition 8.1 *Eine reelle* **Zahlenfolge**[2] *(kurz "***Folge***" genannt) ist eine Abbildung von $\mathbb{N}$ in $\mathbb{R}$ bzw. ein Element von $\mathbb{R}^{\mathbb{N}}$. In Anlehnung an die n–Tupel verwendet man für reelle Zahlenfolgen oft die Schreibweise $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ statt $a : \mathbb{N} \to \mathbb{R}$, $n \overset{a}{\mapsto} a(n)$. Die Zahlen a_n heißen* **Glieder** *der Folge $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$. Eine* **Teilfolge** *von $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ ist eine Folge $(a_n)_n \in \mathbb{R}^N$ mit einer unendlichen Teilmenge N von $\mathbb{N}$.*

[1] `http://de.wikipedia.org/wiki/Analysis`
[2] `http://de.wikipedia.org/wiki/Folge_(Mathematik)`

Hier sind ein paar Beispiele:

$$\begin{array}{lcll} a_n & := & \dfrac{1}{n+1} & 1, 1/2, 1/3, 1/4, \ldots \\ a_n & := & 2^n & 1, 2, 4, 8, 16, \ldots \\ a_n & := & 2^{-n} & 1, 1/2/1/4, 1/8, 1/16, \ldots \\ a_n & := & \left(1 + \dfrac{1}{n+1}\right)^{n+1} & \dfrac{2^1}{1^1}, \dfrac{3^2}{2^2}, \dfrac{4^3}{3^3}, \dfrac{5^4}{4^4}, \ldots \\ a_n & := & (-1)^n & 1, -1, 1, -1, 1, -1, \ldots \end{array}$$

Man "sieht", daß die ersten beiden Folgen gegen Null und die dritte gegen Unendlich "streben", während es nicht ganz so klar ist, daß die vierte Folge gegen $e \approx 2.71828$ "strebt". Die letzte wiederum kann sich nicht zwischen 1 und -1 entscheiden.

Bei Folgen interessiert man sich nicht besonders für die ersten Terme; man will wissen, wie sich die Terme a_n für große n verhalten. Deshalb werden wir Folgen statt bei $n = 0$ auch oft bei $n = 1$ oder einem anderen Index beginnen lassen. In solchen Fällen schreiben wir z.B. $(a_n)_{n\geq 1}$.

Etwas interessanter als die obigen Beispiele sind rekursiv definierte und praktisch wichtigere Folgen wie

$$\begin{array}{rcl} a_0 & := & z \text{ für eine feste Zahl } z > 0 \\ a_{n+1} & := & \dfrac{a_n}{2} + \dfrac{z}{2a_n},\ n \geq 0. \end{array} \tag{8.2}$$

Diese Folge ist eine sehr effiziente (und in Computern inklusive einiger Zusatztricks auch implementierte) Methode, die Wurzel aus z näherungsweise zu berechnen, denn die Folge "strebt ziemlich rapide" gegen $\sqrt{z}$. Für $z = 2$ bekommt man

$$\begin{array}{lcllcl} a_0 & = & 2.0 & a_0^2 & = & 4.0 \\ a_1 & = & 1.5 & a_1^2 & = & 2.25 \\ a_2 & = & 1.416666666666666 & a_2^2 & = & 2.006944444444444 \\ a_3 & = & 1.414215686274509 & a_3^2 & = & 2.000006007304882 \\ a_4 & = & 1.414213562374689 & a_4^2 & = & 2.000000000004510 \\ a_5 & = & 1.414213562373095 & a_5^2 & = & 1.999999999999999 \end{array}$$

Aber warum in aller Welt "strebt" die rekursiv definierte Folge

$$\begin{aligned} a_0 &:= 1 \\ a_n &:= a_{n-1} + (-1)^n \frac{1}{2*n+1} \\ a_0 &= 1.0 \\ a_1 &= 0.6666666666666667 \\ a_2 &= 0.8666666666666667 \\ a_3 &= 0.7238095238095239 \\ a_4 &= 0.8349206349206351 \\ a_5 &= 0.7440115440115441 \\ a_6 &= 0.8209346209346211 \\ \vdots & \quad \vdots \\ a_{1499} &= 0.7852314967492998 \\ a_{1500} &= 0.7855647190085467 \\ \vdots & \quad \vdots \end{aligned}$$

so schrecklich langsam gegen $\pi/4 \approx 0.7853981633974483$, daß sie zur Berechnung von π leider unbrauchbar ist? Immerhin ist es überraschend, daß sie ausgerechnet gegen $\pi/4$ "streben" soll, aber es ist noch ein weiter Weg, bis wir das beweisen können. Das Ergebnis geht auf **Gregory**[1] und **Leibniz**[2] zurück, siehe auch den sehr lesenswerten Text über π in
`http://www-gap.dcs.st-and.ac.uk/~history/HistTopics/Pi_through_the_ages.html`.

Es ist jetzt Zeit, den Begriff "die Folge ... strebt gegen..." sauber zu fassen. Wenn wir den Grenzwert α nennen, so sollten die Folgenglieder der Zahl α beliebig nahe kommen, d.h. der Abstand $|a_n - \alpha|$ sollte beliebig klein werden. Was heißt "beliebig"? Das faßt man so, daß man zu jeder positiven (kleinen) Zahl ϵ eine ϵ–**Umgebung** von α als

$$U_\epsilon(\alpha) := \{\beta \in I\!R \ : \ |\alpha - \beta| < \epsilon\} = (\alpha - \epsilon, \alpha + \epsilon)$$

definiert und verlangt, daß in jeder vorgegebenen kleinen ϵ–Umgebung von α immer ein komplettes Endstück $\{a_n \ : \ n \geq N\}$ der Folge liegen muß. Wenn man also ein beliebig kleines ϵ vorgegeben bekommt, muß man ein $N \in I\!N$ angeben können, so daß das Endstück $\{a_n \ : \ n \geq N\}$ in der ϵ–Umgebung von α liegt. Also:

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Gregory.html`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Leibniz.html`

Definition 8.3 *Eine reelle* **Zahlenfolge** $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ *ist* **konvergent** *gegen einen* **Grenzwert** *oder* **Limes**[1] [2] $\alpha \in \mathbb{R}$*, wenn es zu jedem reellen* $\epsilon > 0$ *ein* $N \in \mathbb{N}$ *gibt, so daß für alle* $n \geq N$ *die Abschätzung*

$$|a_n - \alpha| < \epsilon$$

gilt. Man schreibt

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

Folgen, die gegen Null konvergieren, nennt man **Nullfolgen**[3]*.*
Folgen, die nicht konvergieren, nennt man **divergent***.*
Der Grenzwert einer konvergenten Teilfolge einer Folge heißt **Häufungspunkt**[4] *der Folge.*

Das klingt sehr abstrakt und ist es auch. Die Leser sollten unbedingt noch einmal den vorausgehenden Text durcharbeiten, wenn das Probleme macht. In der Praxis hat man eine Art Tauschgeschäft:

> Wenn Du mir ein $\epsilon > 0$ Deiner Wahl vorgibst, dann muß ich, wenn ich beweisen will, daß die Folge $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ gegen α konvergiert, Dir ein $N \in \mathbb{N}$ zurückgeben können, so daß $|a_n - \alpha| < \epsilon$ für alle $n \geq N$ gilt.

Mathematisch gesehen hat man eigentlich eine Funktion $N : (0, \infty) \to \mathbb{N},\ \epsilon \mapsto N(\epsilon)$ anzugeben, so daß für jedes $\epsilon > 0$ stets $|a_n - \alpha| < \epsilon$ für alle $n \geq N(\epsilon)$ gilt.

Wir werden das an den obigen Beispielen üben. Aber vorerst halten wir noch fest:

> Bei der Konvergenzuntersuchung kann man eine beliebige, aber endliche Anzahl von Folgengliedern einfach ignorieren, ohne die Konvergenzeigenschaften zu verändern.

Unsere Beispiele zum Umgang mit der Grenzwertdefinition fangen wir an mit

Theorem 8.4 *Ist eine Folge konvergent, so ist ihr Limes eindeutig bestimmt.*

[1] `http://de.wikipedia.org/wiki/Grenzwert_(Folge)`

[2] Das lateinische Wort *Limes* hat den Plural *Limites* und nicht *Limiten*. Im Englischen benutzt man *limit* und *limits*.

[3] `http://de.wikipedia.org/wiki/Nullfolge`

[4] `http://de.wikipedia.org/wiki/H%C3%A4ufungspunkt`

Zum Beweis nehmen wir an, eine konvergente Folge $(a_n)_n$ habe zwei verschiedene Limites $\alpha \neq \beta$ in $\mathbb{R}$. Dann muß jede beliebig kleine Umgebung beider Limites je ein komplettes Endstück der Folge enthalten, was natürlich nicht sein kann. Wenn wir zu Übungszwecken dieses Argument mathematisch sauber ausführen wollen, definieren wir

$$\delta := |\alpha - \beta| > 0$$

und wählen $\epsilon := \delta/2$ damit die ϵ–Umgebungen von α und β sich nicht überschneiden. Wegen der Konvergenz der Folge, und weil α und β Limites sein sollen, gibt es dann ein $N \in \mathbb{N}$ und ein $M \in \mathbb{N}$ mit

$$\begin{array}{rcl} |a_m - \alpha| & < & \epsilon \text{ für alle } m \geq M \\ |a_n - \beta| & < & \epsilon \text{ für alle } n \geq N. \end{array}$$

Wir nehmen $K := \max(M, N)$ und ein $k \geq K$ und bekommen

$$\begin{array}{rcl} \delta = |\alpha - \beta| & = & |\alpha - a_k + a_k - \beta| \\ & \leq & |\alpha - a_k| + |a_k - \beta| \\ & < & 2\epsilon = \delta \end{array}$$

was nicht sein kann. □

Durch Hinschreiben der Definition des Limes bekommt man sofort heraus, daß konstante Folgen $(a_n)_n = (\alpha)_n$ konvergent sind mit Limes α.

Aber jetzt sehen wir uns die wichtigsten Nullfolgen an.

Behauptung: Die Folge mit Gliedern $a_n := \frac{1}{n+1}$ hat den Grenzwert Null.
Zu zeigen ist: Zu jedem $\epsilon > 0$ gibt es ein $N \in \mathbb{N}$, so daß $\frac{1}{n+1} < \epsilon$ für alle $n \geq N$ gilt.
Inoffizielle Zwischenrechnung: Es muß $n+1 > 1/\epsilon$ für alle $n \geq N$ gelten, und das kriegen wir hoffentlich hin, wenn wir

$$n + 1 \geq N + 1 > 1/\epsilon$$

wählen. Wir müssen jetzt aber zurück zu einem regulären Beweis.
Lösung: Wir wählen zu gegebenem ϵ ein N mit $N + 1 > 1/\epsilon$.
Für alle $n \geq N$ gilt dann $n+1 \geq N+1 > 1/\epsilon$ und $1/(n+1) \leq 1/(N+1) < \epsilon$.
□

Man beachte, daß die inoffizielle Zwischenrechnung nicht zum Beweis gehört, sondern nur die richtige Idee liefert. Wenn man das passende N zum ϵ angegeben hat, kann man erst den Beweis beginnen.

Jetzt zur speziellen **geometrischen Folge** $a_n := q^n$, wobei wir oben $q = 2$ und $q = 1/2$ hatten. Behauptung: Für jedes feste $q \in \mathbb{R}$ mit $|q| < 1$ hat die Folge mit Gliedern $a_n := q^n$ den Grenzwert Null.
Zu zeigen ist: Zu jedem $\epsilon > 0$ gibt es ein $N \in \mathbb{N}$, so daß $|q^n| < \epsilon$ für alle $n \geq N$ gilt.
Inoffizielle Zwischenrechnung: Wir formen um:

$$\begin{aligned} |q^n| &< \epsilon \\ |q|^n &< \epsilon \\ n \cdot \log|q| &< \log\epsilon \\ n &> \frac{\log\epsilon}{\log|q|} \end{aligned}$$

wobei wir benutzen, daß aus $|q| < 1$ stets $\log|q| < 0$ folgt und der Logarithmus monoton ist.
Lösung: Wir wählen zu gegebenem ϵ ein N mit $N > \frac{\log\epsilon}{\log|q|}$.
Für alle $n \geq N$ gilt dann $n \geq N > \frac{\log\epsilon}{\log|q|}$ und bei Umkehrung der Schlußkette der inoffiziellen Zwischenrechnung folgt $|q^n| < \epsilon$. Falls es stört, daß wir hier den Logarithmus benutzt haben: es folgt später ein anderer Beweis. □

Es ist leicht zu beweisen, daß die Folgen mit Gliedern 2^n oder $(-1)^n$ nicht konvergieren, d.h. keinen Grenzwert haben. Immerhin hat die Folge $(-1)^n$ zwei Häufungspunkte.

Aufgabe: Man beweise das.

Wir brauchen für spätere Zwecke noch

Definition 8.5 *Eine reelle Zahlenfolge $(a_n)_n$ strebt gegen $+\infty$, wenn es zu jedem $K > 0$ ein $N \in \mathbb{N}$ gibt, so daß*

$$a_n \geq K \text{ für alle } n \geq N.$$

Man schreibt $\lim_{n\to\infty} a_n = \infty$. Analog definiert man $\lim_{n\to\infty} a_n = -\infty$, falls $\lim_{n\to\infty}(-a_n) = \infty$ gilt.

Wir werden solche Folgen nicht als konvergent bezeichnen, aber dennoch die Notation $\lim_{n\to\infty} a_n = \pm\infty$ benutzen.

8.1.2 Konvergenzsätze für Folgen

Für die Untersuchung von Folgen auf Konvergenz gibt es einen guten Werkzeugkasten. Wir beginnen mit Aussagen über Folgen, deren Konvergenz wir schon wissen.

Theorem 8.6 *1. Konvergente Folgen sind (als Zahlenmengen) beschränkt. Umgekehrt, und für die Praxis wichtiger: ist eine Folge nicht beschränkt, so kann sie nicht konvergent sein. Der Limes einer konvergenten Zahlenfolge liegt innerhalb beliebiger Schranken der Folge, und insbesondere zwischen Infimum und Supremum der Folge.*

2. Die konvergenten reellen Zahlenfolgen bilden einen unendlichdimensionalen Untervektorraum von $\mathbb{R}^{\mathbb{N}}$. Die Abbildung $(a_n)_n \mapsto \lim a_n$ ist eine lineare Abbildung auf diesem Unterraum. Mit anderen Worten: Sind zwei reelle Zahlenfolgen $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ und $(b_n)_n \in \mathbb{R}^{\mathbb{N}}$ konvergent und sind $\alpha, \beta \in \mathbb{R}$ beliebig, so ist die Folge $(\alpha a_n + \beta b_n)_n$ konvergent und es gilt

$$\lim_{n\to\infty} (\alpha a_n + \beta b_n) = \alpha \lim_{n\to\infty} a_n + \beta \lim_{n\to\infty} b_n.$$

Insbesondere sind Linearkombinationen von Nullfolgen wieder Nullfolgen, denn die Nullfolgen sind ein Unterraum der konvergenten Folgen, sie bilden den Kern der obigen linearen Abbildung.

3. Sind $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ und $(b_n)_n \in \mathbb{R}^{\mathbb{N}}$ konvergente reelle Zahlenfolgen, so ist auch $(a_n \cdot b_n)_n$ eine konvergente Zahlenfolge und es gilt

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

Insbesondere sind Produkte und Potenzen von Nullfolgen wieder Nullfolgen.

4. Sind $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ und $(b_n)_n \in \mathbb{R}^{\mathbb{N}}_{\neq 0}$ konvergente reelle Zahlenfolgen, und ist der Grenzwert von $(b_n)_n$ nicht Null, so ist auch $(a_n/b_n)_n$ eine konvergente Zahlenfolge und es gilt

$$\lim_{n\to\infty} (a_n/b_n) = \Big(\lim_{n\to\infty} a_n\Big) \,/\, \Big(\lim_{n\to\infty} b_n\Big).$$

5. Ist $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ eine konvergente Zahlenfolge, so auch $(|a_n|)_n \in \mathbb{R}^{\mathbb{N}}$ mit $\lim_{n\to\infty} |a_n| = |\lim_{n\to\infty} a_n|$.

Die Beweise sind durchweg elementar, und gute Übungsaufgaben. Weil der Beweis von Teil 3 einige typische Eigenarten hat, die auch anderswo nützlich sind, führen wir ihn in aller Breite vor.

Voraussetzung: $(a_n)_n \in \mathbb{R}^{\mathbb{N}}$ und $(b_n)_n \in \mathbb{R}^{\mathbb{N}}$ sind konvergente reelle Zahlenfolgen, die Grenzwerte sind $\alpha = \lim a_n$ und $\beta = \lim b_n$.

Zu zeigen: Die Folge $(a_n \cdot b_n)_n$ ist konvergent und hat den Limes $\alpha \cdot \beta$.
Zu zeigen: Zu jedem $\epsilon > 0$ gibt es ein $N \in \mathbb{N}$, so daß für alle $n \geq N$ die Aussage $|a_n \cdot b_n - \alpha \cdot \beta| < \epsilon$ folgt.
Voraussetzung A: Zu jedem $\epsilon_A > 0$ gibt es ein $N_A \in \mathbb{N}$, so daß für alle $n \geq N_A$ die Aussage $|a_n - \alpha| < \epsilon_A$ folgt. Ferner gibt es wegen der Beschränktheit konvergenter Folgen ein $K_A > 0$ so daß $|a_n| \leq K_A$ und $|\alpha| \leq K_A$ gilt.
Voraussetzung B: Zu jedem $\epsilon_B > 0$ gibt es ein $N_B \in \mathbb{N}$, so daß für alle $n \geq N_B$ die Aussage $|b_n - \beta| < \epsilon_B$ folgt. Ferner gibt es ein $K_B > 0$ so daß $|b_n| \leq K_B$ und $|\beta| \leq K_B$ gilt.

Inoffizielle Zwischenrechnung: Man muß irgendwie von $|a_n - \alpha| < \epsilon$ und $|b_n - \beta| < \epsilon$ auf $|a_n \cdot b_n - \alpha \cdot \beta| < \epsilon$ kommen. Weil das n in $a_n \cdot b_n$ doppelt vorkommt, in den entsprechenden Voraussetzungen aber nur einfach, sollte man sich eine Brücke über $a_n \cdot \beta$ bauen. Dieser Wert ist ein "Mittelding" zwischen $a_n \cdot b_n$ und $\alpha \cdot \beta$. Der Beweis geht ganz ähnlich mit der entsprechenden Brücke über $\alpha \cdot b_n$. Es folgt:

$$\begin{aligned} |a_n \cdot b_n - \alpha \cdot \beta| &= |a_n \cdot b_n - a_n \cdot \beta + a_n \cdot \beta - \alpha \cdot \beta| \\ &\leq |a_n \cdot b_n - a_n \cdot \beta| + |a_n \cdot \beta - \alpha \cdot \beta| \\ &= |a_n| \cdot |b_n - \beta| + |\beta| \cdot |a_n - \alpha| \\ &\leq K_A \cdot |b_n - \beta| + K_B \cdot |a_n - \alpha| \\ &< K_A \epsilon_B + K_B \epsilon_A \end{aligned}$$

wenn wir zu gewissen ϵ_A und ϵ_B die Indizes n mit $n \geq N_A$ und $n \geq N_B$ nehmen. Man muß das Ganze so hinkriegen, daß $K_A \epsilon_B + K_B \epsilon_A < \epsilon$ wird, wenn ϵ vorgegeben ist. Also verteilt man je $\epsilon/2$ auf diese beiden Summanden. Dann ergibt sich die

Lösung: Zu gegebenem $\epsilon > 0$ wählt man ϵ_A und ϵ_B so klein, daß $K_B \epsilon_A < \epsilon/2$ und $K_A \epsilon_B < \epsilon/2$ gilt (man nimmt z.B. $\epsilon_A = \epsilon/(2(K_A + 1))$ und $\epsilon_B = \epsilon/(2(K_B + 1))$). Dazu bekommt man je ein N_A und ein N_B aus den Voraussetzungen A und B mit $|a_n - \alpha| < \epsilon_A$ und $|b_n - \beta| < \epsilon_B$ für alle $n \geq N := \max(N_A, N_B)$. Wir wählen zu unserem ϵ genau dieses N. Dann folgt für alle $n \geq N = \max(N_A, N_B)$ die Aussage

$$\begin{aligned} |a_n \cdot b_n - \alpha \cdot \beta| &< K_A \epsilon_B + K_B \epsilon_A \\ &< \epsilon/2 + \epsilon/2 = \epsilon \end{aligned}$$

indem wir wie in der inoffiziellen Zwischenrechnung verfahren und unsere gute Wahl von ϵ_A und ϵ_B einsetzen. □

Typische Anwendungen dieses Satzes sehen etwa so aus:

Aufgabe: Ist die Folge mit den Gliedern

$$a_n := \frac{4n^2 - 3n + 2}{1 + n(n+1)}$$

konvergent und wenn ja, was ist der Limes?

Die Grundidee ist, den Ausdruck so umzuformen, daß man möglichst viele konvergente Folgen ablesen kann, um deren Limites einzusetzen. Wir ignorieren das einzelne Folgenglied mit $n = 0$ und dividieren den Bruch durch n^2:

$$a_n := \frac{4 - 3\frac{1}{n} + 2\frac{1}{n^2}}{\frac{1}{n^2} + 1 + \frac{1}{n}}$$

Wir wissen aus dem obigen Satz, daß alle Folgen der Form $b_n = \frac{1}{n+k}$ mit festem k gegen Null konvergieren, denn sie stimmen bis auf endlich viele Glieder mit $\frac{1}{n+1}$ überein. Deshalb konvergiert $\frac{1}{n}$ gegen Null und nach unserem Satz auch jede Potenz davon, d.h. insbesondere auch $\frac{1}{n^2}$. Die Folge im Zähler hat also nach unserem Satz den Limes $4 - 3 \cdot 0 + 2 \cdot 0 = 4$, während die Folge im Nenner gegen $0 \cdot 1 + 1 + 0 = 1$ konvergiert. Dann liefert der Satz die Konvergenz der gesamten Folge gegen 4.

Zu Übungszwecken notieren wir noch ein paar einfache Fakten ohne Beweis:

- Die beschränkten Folgen bilden einen Untervektorraum des allgemeinen Folgenraums $\mathbb{R}^{\mathbb{N}}$.
- Auf beschränkten Folgen ist

  $$\|(a_n)_n\|_\infty := \sup_{n \in \mathbb{N}} |a_n|$$

  eine Norm.
- Die konvergenten Folgen sind ein Untervektorraum der beschränkten Folgen.
- Die konstanten Folgen sind ein zu $\mathbb{R}$ isomorpher Untervektorraum des Vektorraums der konvergenten Folgen.

Eine andere Gruppe von Sätzen wirkt auf Folgen, deren Konvergenz man noch nicht weiß und deren Limes man noch nicht kennt. Hier ist wichtig, daß die reellen Zahlen **vollständig** angeordnet sind, d.h. jede nach oben bzw. unten beschränkte Menge reeller Zahlen hat nach Satz 3.21 ein Supremum bzw. Infimum. Man sieht sich Folgen an, die angeordnet sind:

Definition 8.7 *Eine Folge $(a_n)_n$ heißt (schwach)* **monoton**[1] **wachsend**, *wenn*

$$a_n \leq a_{n+1} \text{ für alle } n \in \mathbb{N}$$

gilt. Hat man $<$ statt $\leq$, so ist die Folge **streng monoton** *wachsend oder* **stark monoton** *wachsend. Analog, aber mit umgedrehter Ordnungsrelation, definiert man eine Folge als monoton* **fallend**.

Theorem 8.8 *Jede schwach monoton wachsende und nach oben beschränkte Folge ist konvergent und ihr Supremum ist ihr Limes. Das gilt analog auch für fallende Folgen und das Infimum.*

Beweis: Ist $(a_n)_n$ eine monoton wachsende und nach oben beschränkte Folge, so hat die Punktmenge $\{a_n : n \in \mathbb{N}\}$ ein Supremum $\alpha \in \mathbb{R}$, weil die reellen Zahlen vollständig sind. Es gilt also

$$a_n \leq a_{n+1} \leq \alpha \text{ für alle } n \in \mathbb{N}$$

und α ist die kleinste reelle Zahl mit dieser Eigenschaft. Es sei nun ein $\epsilon > 0$ vorgegeben, und wir behaupten, daß es dazu ein $N \in \mathbb{N}$ gibt mit $\alpha - a_N < \epsilon$. Wäre das nicht so, müßte für alle $n \in \mathbb{N}$ die Aussage $\alpha - a_n \geq \epsilon$ gelten, aber dann wäre $\alpha - \epsilon$ eine kleinere obere Schranke als α, was nicht geht. Wir finden also immer ein N mit $\alpha - a_N < \epsilon$ und bekommen für alle $n \geq N$ wegen $a_N \leq a_n \leq \alpha$ erst recht

$$|\alpha - a_n| = \alpha - a_n \leq \alpha - a_N < \epsilon.$$

□

Damit hat man einen einfachen Beweis für die Konvergenz aller Folgen $a_n := q^n$ gegen Null, sofern $0 < q < 1$ gilt, denn aus

$$a_n = q^n = q^{n+1}/q = a_{n+1}/q > a_{n+1} > 0$$

ergibt sich, daß die Folge monoton fällt und durch Null nach unten beschränkt ist. Sie konvergiert also gegen ein Infimum $\alpha \geq 0$. Die Folge $b_n := (qa_n)_n$ konvergiert dann gegen $q\alpha$ und ist bis auf das erste Glied identisch mit der Folge $(a_n)_n$, hat also denselben Limes. Es folgt $\alpha = q\alpha$, und wegen $q \neq 1$ kann α nicht positiv sein. □.

Wenn man nicht weiß, ob eine Folge konvergiert, kann man also nach Monotonie und Beschränktheit fragen, um Konvergenz nachzuweisen. Geht das nicht,

[1] `http://de.wikipedia.org/wiki/Monotonie_(Mathematik)`

so reicht es manchmal zu wissen, daß die Folgenglieder ziemlich dicht beieinander liegen. Das ist der Inhalt des Konvergenzkriteriums von **Cauchy**[1]:

Theorem 8.9 *Eine reelle Zahlenfolge $(a_n)_n$ ist genau dann konvergent, wenn es zu jedem $\epsilon > 0$ ein $N \in \mathbb{N}$ gibt, so daß für alle $m, n \geq N$ die Abschätzung $|a_n - a_m| < \epsilon$ gilt.*

Hat eine Folge diese Eigenschaft, so spricht man auch von einer **Cauchy–Folge**[2] und der obige Satz bekommt die Form

- Jede Cauchy–Folge ist konvergent, und jede konvergente Folge ist eine Cauchy–Folge.

Es ist sehr einfach zu zeigen (Aufgabe: wie?), daß jede konvergente Folge eine Cauchy–Folge ist. Die Umkehrung ist erheblich schwieriger, aber wir werden es versuchen. Es gelte also das Cauchy–Kriterium, und wir wollen zeigen, daß die Folge konvergiert. Weil wir (noch) keinen Limes haben, müssen wir Satz 8.8 benutzen, und dazu brauchen wir monotone und beschränkte Folgen, die wir uns erst noch bauen müssen. Mit dieser Grundidee ist es nicht mehr ganz so schwierig.

Aus dem Kriterium folgt sofort, daß die Folge $(a_n)_n$ beschränkt sein muß, denn wenn man das N_1 zu $\epsilon = 1$ wählt, folgt, daß alle a_n mit $n \geq N_1$ die Eigenschaft $|a_N - a_n| < 1$ haben, was $|a_n| < 1 + |a_N|$ bedeutet. Diese Schranke gilt für das komplette Endstück, und die ersten N_1 Glieder machen keinen wesentlichen Unterschied, weil sie auch beschränkt sind.

Jetzt verschaffen wir uns schwach monotone und beschränkte Folgen, indem wir Infimum und Supremum der Endstücke bilden:

$$b_n := \inf\{a_m \ : \ m \geq n\} \leq c_n := \sup\{a_m \ : \ m \geq n\}$$

und bekommen nach Satz 8.8 Limites β und γ mit

$$b_n \leq b_{n+1} \leq \lim_{k\to\infty} b_k =: \beta \leq \gamma := \lim_{k\to\infty} c_k \leq c_{n+1} \leq c_n \text{ für alle } n \in \mathbb{N}.$$

Dann geben wir ein beliebiges $\epsilon > 0$ vor und erhalten aus dem Kriterium ein $N \in \mathbb{N}$ mit $|a_n - a_m| < \epsilon$ für alle $n, m \geq N$. Also gilt $|a_N - a_m| < \epsilon$ für alle $m \geq N$ und es folgt

$$a_N - \epsilon \leq a_m \leq a_N + \epsilon \text{ für alle } m \geq N$$

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Cauchy.html`
[2] `http://de.wikipedia.org/wiki/Cauchy-Folge`

und das bedeutet

$$a_N - \epsilon \leq b_N \leq c_N \leq a_N + \epsilon$$

und $0 \leq \gamma - \beta \leq c_N - b_N \leq 2\epsilon$. Weil ϵ beliebig war, folgt $\beta = \gamma$, und wir wollen jetzt beweisen, daß dies der gesuchte Limes der Folge $(a_n)_n$ ist.

Zu jedem $\epsilon > 0$ bekommen wir wegen der Konvergenz der Folge $(b_n)_n$ ein N_B mit $|b_n - \beta| < \epsilon$ für alle $n \geq N_B$ sowie analog ein N_C mit $|c_n - \gamma| = |c_n - \beta| < \epsilon$ für alle $n \geq N_C$. Wählen wir $N := \max(N_B, N_C)$, so folgt für alle $n \geq N$ die Aussage

$$\beta - \epsilon < b_n \leq a_m \leq c_n < \beta + \epsilon$$

für alle $m \geq n \geq N$, und deshalb gilt $|a_m - \beta| < \epsilon$ für alle $m \geq N$. □

Ein weiterer nützlicher Trick zum Beweis der Konvergenz von reellen Zahlenfolgen ist die Einschließung:

Theorem 8.10 *Sind $(a_n)_n$, $(b_n)_n$, $(c_n)_n$ reelle Zahlenfolgen mit*

$$a_n \leq b_n \leq c_n \text{ für alle } n \in \mathbb{N},$$

und sind $(a_n)_n$, $(c_n)_n$ konvergent mit demselben Grenzwert $\alpha \in \mathbb{R}$, so konvergiert auch $(b_n)_n$ gegen diesen Grenzwert.

Beweis: Zu gegebenem $\epsilon > 0$ gibt es $N_A,\ N_C \in \mathbb{N}$, so daß für alle $n \geq \max(N_A, N_C)$ stets $|a_n - \alpha| < \epsilon$ und $|c_n - \alpha| < \epsilon$ gilt. Dann folgt wegen

$$-|a_n - \alpha| \leq a_n - \alpha \leq b_n - \alpha \leq c_n - \alpha \leq |c_n - \alpha|$$

die Abschätzung

$$|b_n - \alpha| \leq \max(|a_n - \alpha|, |c_n - \alpha|) < \epsilon.$$

Wir wählen zu gegebenem $\epsilon > 0$ also $N := \max(N_A, N_C)$ und bekommen die Konvergenz von $(b_n)_n$. □

Der wichtigste Anwendungsfall sieht so aus: ist $(c_n)_n$ eine Nullfolge, und gilt

$$0 \leq b_n \leq c_n \text{ für alle } n \in \mathbb{N},$$

so ist auch $(b_n)_n$ eine Nullfolge. Man sagt dann, die Folge $(c_n)_n$ sei eine **Majorante** von $(b_n)_n$.

Wenn wir nur Beschränktheit einer Folge haben, können wir nicht auf Konvergenz schließen, wie man am Beispiel der Folge $1, -1, 1, -1, \ldots$ sieht. Aber jetzt kommt eine unscheinbar klingende, aber sehr wichtige Aussage über beschränkte reelle Zahlenfolgen:

Theorem 8.11 *(Satz von* **Bolzano**[1]**–Weierstrass**[2]*)*
Jede beschränkte reelle Zahlenfolge hat eine konvergente Teilfolge und damit auch einen Häufungspunkt.

Beweis: Es sei eine beschränkte Folge $(d_n)_n \in \mathbb{R}^{\mathbb{N}}$ gegeben, und weil sie beschränkt ist, können wir annehmen, sie liege im Intervall $[-K, K]$ mit einem positiven $K \in \mathbb{R}$. Wir streben einen konstruktiven rekursiv–induktiven Beweis an, der das vorige Ergebnis benutzt und den Standardtrick der **Intervallschachtelung** anwendet. Dazu bezeichnen wir unsere Folge neu mit $(d_n^{(0)})_n \in [-K \cdot 2^0, K \cdot 2^0] =: [a_0, c_0]$ und wählen ein beliebiges Folgenelement aus dieser Folge, das wir b_0 nennen. Es folgt der Induktionsanfang

$$\begin{array}{rcccl} a_0 & \leq & b_0 & \leq & c_0 \\ 0 & < & c_0 - a_0 & \leq & 2K2^0 \\ b_0 & \in & [a_0, c_0] & \cap & \{d_j \ : \ j \in \mathbb{N}\} \\ d_n^{(0)} & \in & [a_0, c_0] & \cap & \{d_j \ : \ j \in \mathbb{N}\} \text{ für alle } n \in \mathbb{N}. \end{array}$$

Nehmen wir an, wir hätten für ein $k \geq 0$ schon

$$\begin{array}{rcccl} a_k & \leq & b_k & \leq & c_k \\ 0 & < & c_k - a_k & \leq & 2K2^{-k} \\ b_k & \in & [a_k, c_k] & \cap & \{d_j \ : \ j \in \mathbb{N}\} \\ d_n^{(k)} & \in & [a_k, c_k] & \cap & \{d_j \ : \ j \in \mathbb{N}\} \text{ für alle } n \in \mathbb{N}. \end{array}$$

Jetzt teilen wir das Intervall $[a_k, c_k]$ in zwei Teile und greifen einen Teil heraus, der unendlich viele der $d_n^{(k)}$ enthält. Diese Teilfolge nennen wir $(d_n^{(k+1)})_n$, und das halbierte Intervall wird $[a_{k+1}, c_{k+1}]$. Eines der Folgenelemente von $d_n^{(k+1)}$ nehmen wir heraus und nennen es b_{k+1}. Damit ist klar, daß wir den Induktionsschritt vollzogen haben, aber es folgt auch die Monotonie

$$a_k \leq a_{k+1} < c_{k+1} \leq c_k \text{ für alle } k \in \mathbb{N}$$

weil wir immer die linke oder die rechte Hälfte von $[a_k, c_k]$ als $[a_{k+1}, c_{k+1}]$ nehmen.

Die beiden Folgen $(a_k)_k$ und $(c_k)_k$ sind monoton und beschränkt. Sie haben deshalb Limites α bzw γ, die mit dem Supremum bzw. Infimum zusammenfallen. Diese Limites müssen gleich sein, weil $0 < c_k - a_k \leq 2K2^{-k}$ und

$$a_k \leq \alpha \leq \gamma \leq c_k \text{ für alle } k \in \mathbb{N}$$

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Bolzano.html`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Weierstrass.html`

gilt. Nach Satz 8.10 folgt dann die Konvergenz der Folge $(b_k)_k$ gegen denselben Limes. Diese Folge ist aber eine Teilfolge der ursprünglich gegebenen Folge. □

Manche Puristen (genauer: **Intuitionisten**[1], z.B. **Brouwer**[2] oder die **Konstruktivisten**[3]) würden so einen Beweis nicht gelten lassen, weil hier unendlich oft aus einer jeweils unendlichen Menge eine unendliche Teilmenge ausgewählt wurde. Aber wir sollten uns so einen schönen Beweis nicht durch irgendein Genörgele kaputtmachen lassen. Oder?

Als konkretes Beispiel untersuchen wir die Folge

$$a_0 := z > 0,\ a_{n+1} := \frac{a_n}{2} + \frac{z}{2a_n},\ n \geq 0$$

aus (8.2). Da wir vermuten, daß der Limes $\sqrt{z}$ ist, ziehen wir ihn von beiden Seiten ab und bekommen nach einiger Rechnung

$$a_{n+1} - \sqrt{z} = \frac{(a_n - \sqrt{z})^2}{2a_n}. \tag{8.12}$$

Weil die Folge nur positive Elemente haben kann (Induktion) gilt $a_{n+1} - \sqrt{z} \geq 0$ und deshalb

$$a_{n+2} - a_{n+1} = \frac{z}{2a_{n+1}} - \frac{a_{n+1}}{2} = \frac{z - a_{n+1}^2}{2a_{n+1}} \leq 0.$$

Also ist die Folge spätestens vom zweiten Glied an monoton fallend und nach unten wegen (8.12) durch $\sqrt{z} > 0$ beschränkt. Sie hat also einen Limes $\alpha \geq \sqrt{z} > 0$. Die drei Folgen mit den Gliedern $a_{n+1},\ a_n/2$ und $z/(2a_n)$ sind nach Satz 8.6 konvergent mit den Limites $\alpha, \alpha/2$ und $z/(2\alpha)$. Es gilt dann

$$\alpha = \frac{\alpha}{2} + \frac{z}{2\alpha}$$

und daraus folgt nach kurzer Rechnung $\alpha^2 = z$.

Die Gleichung (8.12) zeigt, warum die Folge sehr schnell konvergiert. Für $n \geq 1$ und $z \geq 1$ gilt nämlich

$$0 \leq a_{n+1} - \sqrt{z} \leq \frac{(a_n - \sqrt{z})^2}{2}$$

[1] `http://de.wikipedia.org/wiki/Intuitionismus`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Brouwer.html`
[3] `http://de.wikipedia.org/wiki/Konstruktive_Mathematik`

wegen $a_n \geq \sqrt{z} \geq 1$. Wenn nun der Fehler $a_n - \sqrt{z}$ für ein n schon ziemlich klein ist, z.B. 10^{-5}, so ist der Fehler $a_{n+1} - \sqrt{z}$ schon kleiner als 10^{-10}. Die Anzahl der korrekten Dezimalstellen verdoppelt sich mit jedem Schritt! Das haben wir oben schon beaobachtet.

Diese Methode zur Wurzelberechnung ist ein Spezialfall des **Newton**[1]–Verfahrens zur Lösung allgemeiner Gleichungen und Gleichungssysteme.

Manchmal hat man Ausdrücke, die auf zwei verschiedene Weisen als Folgen aufgefaßt werden können. Dann ist es nicht gleichgültig, welchen der beiden möglichen Limites man zuerst bildet. Beispiel:

$$f(n,m) := \left(1 - \frac{1}{n}\right)^m \text{ für alle } m, n \geq 1.$$

Das liefert

$$\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^m = 1^m = 1, \lim_{m\to\infty} \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^m = 1$$

und

$$\lim_{m\to\infty} \left(1 - \frac{1}{n}\right)^m = 0, \lim_{n\to\infty} \lim_{m\to\infty} \left(1 - \frac{1}{n}\right)^m = 0.$$

Merksatz:

> **Beim Vertauschen zweier Grenzprozesse ist größte Vorsicht geboten!**

8.2 Landau–Symbole

In der Informatik hat man oft Verfahren A und B zu vergleichen, die in Abhängigkeit von einem Parameter n, etwa der Länge eines binären Eingabewortes oder der Anzahl der zu sortierenden Objekte, den Aufwand $A(n)$ oder $B(n)$ haben. Welches ist schneller? Das kann man für feste n untersuchen, aber auch für "sehr große" n im "asymptotischen Grenzfall". Es geht hier weniger um konvergente Folgen, sondern darum, welche der beiden Folgen "schneller" gegen Unendlich strebt. Und dabei sind feste Faktoren nicht besonders relevant; man sieht Verfahren, die den Aufwand $n^3/3$ oder $2n^3/3$ haben, als vergleichbar schnell an. Zusätzliche Terme wie $17n^2$ fallen für große n nicht ins Gewicht und werden ignoriert.[2] Genauer:

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Newton.html`

[2] Nebenbei: das Gaußsche Eliminationsverfahren braucht im wesentlichen $n^3/3$ Multiplikationen zum Lösen eines $n \times n$–Gleichungssystems, während das QR–Verfahren nach Householder $2n^3/3$ Multiplikationen braucht. Es gibt trickreiche Verfahren, die mit etwa $n^{\log_2 7} \approx n^{2.807}$ Multiplikationen auskommen, und das Rennen nach dem kleinsten Exponenten ist offen.

Definition 8.13 *(**Landau**[1]–Symbole[2] Es seien $(a_n)_n$, $(b_n)_n \in \mathbb{R}^{\mathbb{N}}$ zwei reelle Zahlenfolgen.*

1. *Man schreibt*
$$a_n = \mathcal{O}(b_n)$$
und sagt "a_n ist Groß–Oh von b_n", falls es Konstanten $C \in \mathbb{R}_{>0}$ und $N \in \mathbb{N}$ gibt mit
$$|a_n| \leq C \cdot |b_n| \text{ für alle } n \geq N.$$
Sind alle b_n von Null verschieden, so ist dies gleichbedeutend damit, daß die Folge $(|a_n/b_n|)_n$ durch C nach oben beschränkt ist.

2. *Die Aussage $a_n = \Omega(b_n)$ besagt, daß es Konstanten $C \in \mathbb{R}_{>0}$ und $N \in \mathbb{N}$ gibt mit*
$$|a_n| \geq C \cdot |b_n| \text{ für alle } n \geq N.$$
Man sagt "a_n ist Groß–Omega von b_n". Sind alle b_n von Null verschieden, so ist dies gleichbedeutend damit, daß die Folge $(|a_n/b_n|)_n$ durch C nach unten durch eine positive Konstante beschränkt ist.

3. *Mit $a_n = o(b_n)$ ("a_n ist Klein–Oh von b_n") ist gemeint, daß es zu jedem $\epsilon > 0$ ein $N \in \mathbb{N}$ gibt mit*
$$|a_n| \leq \epsilon \cdot |b_n| \text{ für alle } n \geq N.$$
Sind alle b_n von Null verschieden, so ist dies gleichbedeutend damit, daß die Folge $(|a_n/b_n|)_n$ gegen Null konvergiert..

4. *Die Aussage $a_n = \Theta(b_n)$, d.h. "a_n ist Theta von b_n" bedeutet*
$$a_n = \mathcal{O}(b_n) \textbf{ und } a_n = \Omega(b_n),$$
d.h. es gibt Konstanten C_1, $C_2 \in \mathbb{R}_{>0}$ und $N \in \mathbb{N}$ mit
$$C_1 \cdot |b_n| \leq |a_n| \leq C_2 \cdot |b_n| \text{ für alle } n \geq N.$$

Theorem 8.14 *1. Als Relationen auf Folgen gesehen sind alle diese Begriffe transitiv. Z.B. folgt aus $a_n = \mathcal{O}(b_n)$ und $b_n = \mathcal{O}(c_n)$ stets auch $a_n = \mathcal{O}(c_n)$.*

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Landau.html`
[2] % `http://de.wikipedia.org/wiki/Landau-Symbole`

2. *Die Relation Θ ist eine Äquivalenzrelation.*
3. *An den obigen Relationen zwischen Folgen ändert sich nichts, wenn man*
   - *die Folgen mit positiven Konstanten multipliziert oder*
   - *zu Beträgen übergeht oder*
   - *nur Endstücke betrachtet.*
4. *Gilt $a_n = \mathcal{O}(b_n)$ und $a'_n = \mathcal{O}(b'_n)$, so gilt*

$$\begin{aligned} |a_n + a'_n| &= \mathcal{O}(\max(|b_n|, |b'_n|)) \\ |a_n \cdot a'_n| &= \mathcal{O}(|b_n| \cdot |b'_n|)). \end{aligned}$$

5. *Die obige Ausage bleibt richtig, wenn man $\mathcal{O}$ durch o ersetzt.*
6. *Aus $a_n = o(b_n)$ folgt $a_n = \mathcal{O}(b_n)$.*
7. *Eine Folge ist genau dann beschränkt, wenn sie $\mathcal{O}(1)$ ist.*
8. *Eine Folge ist genau dann eine Nullfolge, wenn sie $o(1)$ ist.*

Aufgabe: Man beweise das.

In der Informatik–Praxis sind die meisten der in der Landau–Notation auftretenden Folgen positiv und streben gegen Unendlich. In diesem Falle sieht man sich das Konvergenzverhalten von a_n/b_n an und bekommt

$a_n = \mathcal{O}(b_n)$	genau dann, wenn a_n/b_n nach oben beschränkt ist
$a_n = \Omega(b_n)$	genau dann, wenn a_n/b_n nach unten positiv beschränkt ist
$a_n = o(b_n)$	genau dann, wenn a_n/b_n gegen Null konvergiert
$a_n = \Theta(b_n)$	genau dann, wenn a_n/b_n nach oben und unten positiv beschränkt ist.

Nützlich sind für Folgen $(a_n)_n$, $(b_n)_n \in \mathbb{R}^{\mathbb{N}}_{>0}$ auch die Aussagen

$$\begin{aligned} \frac{a_n}{b_n} &= \mathcal{O}(1), \text{ falls } a_n = \mathcal{O}(b_n) \\ \frac{a_n}{b_n} &= o(1), \text{ falls } a_n = o(b_n). \end{aligned}$$

Für die Anwendungen in der Informatik braucht man dringend

Theorem 8.15 *Für alle $k \in \mathbb{N}$ und alle reellen $\alpha > 0$ gilt*

$$\begin{aligned} \log^k n &= \mathcal{o}(n^\alpha) \\ n^k &= \mathcal{o}((1+\alpha)^n) \\ \alpha^n &= \mathcal{o}(n!),\ n! := 1 \cdot 2 \cdot \ldots \cdot n \text{ (sprich: } n \text{ Fakultät).} \end{aligned}$$

Das bedeutet

logarithmisches Wachstum	ist langsamer als	polynomiales Wachstum
polynomiales Wachstum	ist langsamer als	exponentielles Wachstum
exponentielles Wachstum	ist langsamer als	fakultatives Wachstum.

Der Beweis erfordert etliches Wissen über Exponentialfunktion und Logarithmus, wenn er kurz sein soll. Wir könnten ihn später nachholen, aber es geht teilweise auch mit Umwegen "zu Fuß", die man beim ersten Lesen getrost überspringen kann.

Zuerst beweisen wir $n^k = \mathcal{o}((1+\alpha)^n)$ für alle $\alpha > 0$ und $k \in \mathbb{N}$. Man mache sich klar, daß große k und kleine α problematisch sind. Zu beliebigem $\alpha > 0$, beliebigem $k > 0$ und beliebigem $\epsilon > 0$ sehen wir uns mit dem binomischen Satz die Größe

$$\begin{aligned} \epsilon(1+\alpha)^{n+k+1} &= \epsilon \sum_{j=0}^{n+k+1} \binom{n+k+1}{j} \alpha^j \\ &\geq \epsilon \binom{n+k+1}{k+1} \alpha^{k+1} \\ &= \epsilon \frac{(n+k+1)!}{(k+1)!n!} \alpha^{k+1} \\ &\geq n^k \epsilon \frac{n}{(k+1)!} \alpha^{k+1} \end{aligned}$$

an. Wenn man n so groß macht, daß

$$\epsilon \frac{n}{(k+1)!} \alpha^{k+1} \geq 1$$

gilt, folgt $n^k \leq \epsilon(1+\alpha)^{n+k+1} = \epsilon(1+\alpha)^{k+1}(1+\alpha)^n$, d.h. $n^k = \mathcal{o}((1+\alpha)^n)$, weil der feste Faktor $(1+\alpha)^{k+1}$ an der $\mathcal{o}$–Relation nichts ändert.

Jetzt gehen wir an den Beweis der dritten Relation und lassen uns ein beliebiges $\epsilon > 0$ geben. Ist $n \geq 2$ gerade, so folgt

$$\begin{aligned} \epsilon n! &= \epsilon n(n-1) \cdots (n/2)((n/2-1) \cdots 1 \\ &\geq \epsilon (n/2-1)^{n/2+1} \\ &\geq \epsilon (n/2-1) \left(\sqrt{n/2-1}\right)^n \\ &\geq \alpha^n \end{aligned}$$

wenn man n so groß wählt, daß

$$\epsilon(n/2-1) \geq 1,\ \sqrt{n/2-1} \geq \alpha$$

gelten. Ganz ähnlich kann man für ungerade n argumentieren.

Bei der ersten Wachstumsrelation geht der Logarithmus schon in die Formulierung ein, deshalb braucht man diesen auch im Beweis. Wir versuchen, mit elementaren Eigenschaften des Logarithmus und der Exponentialfunktion auszukommen und setzen $n = e^{x_n}$ mit reellem $x_n = \log n$ und wählen für große n natürliche Zahlen m_n mit $m_n \leq x_n < 1 + m_n$. Zu fest gegebenem $k \in I\!N$ und $\alpha > 0$ lassen wir uns ein $\epsilon > 0$ geben und schätzen ab:

$$\begin{aligned} \epsilon n^\alpha &= \epsilon e^{\alpha x_n} \\ &\geq \epsilon e^{\alpha m_n} \\ &= \epsilon e^{-\alpha} \left(e^{\frac{\alpha}{k}}\right)^{k(m_n+1)} \\ &\geq k(m_n+1) \\ &\geq kx_n \\ &= \log^k n \end{aligned}$$

für genügend große n, wobei wir die schon bewiesene Relation

$$j = o(\left(e^{\frac{\alpha}{k}}\right)^j) \text{ für große } j = k(m_n+1)$$

ausgeschlachtet haben. □

Man sehe sich zu den verschiedenen Wachstumsgeschwindigkeiten die Tabellen in [1], S. 278 und [4], S. 241 an.

Wichtiger ist, die bisher bereitgestellten Werkzeuge richtig anzuwenden. Deshalb folgen jetzt ein paar Beispiele. Viele weitere sind in der Literatur anzutreffen.

Problem: Gegeben sei die Folge mit Gliedern $a_n := 4n^2 + 15n + 3\log^5(n)$. Gesucht ist eine möglichst einfache Folge $(b_n)_n$ mit $a_n = \mathcal{O}(b_n)$ oder sogar $a_n = \Theta(b_n)$.

Lösung: Es ist klar, daß die Folge gegen Unendlich strebt, und daß $4n^2 = \mathcal{O}(n^2),\ 15n = \mathcal{O}(n),\ 3\log^5(n) = \mathcal{O}(\log^5(n)) = o(n)$ gilt. Der am stärksten wachsende Teil ist also $\mathcal{O}(n^2)$, und man wird $b_n = n^2$ nehmen. Nach dieser eher informellen Vorüberlegung berechnet man

$$\frac{a_n}{b_n} = 4 + \frac{15}{n} + 3\frac{\log^5(n)}{n^2}$$

und dies konvergiert gegen 4, weil die beiden zusätzlichen Folgen Nullfolgen bzw. $o(1)$ sind. Also folgt $a_n = \Theta(n^2)$.

Problem: Zwei Sortierprogramme brauchen zum Sortieren von n Elementen jeweils $2n \log n$ bzw. $n(n+1)/2$ Vergleichsoperationen. Welches Verfahren ist bei großen n schneller?

Lösung: Der Aufwand des zweiten ist $\Theta(n^2)$, weil $n(n+1)/2 = n^2(1+1/n)/2$ gilt. Der Aufwand des ersten ist geringer, weil man mit Theorem 8.15 auf

$$\frac{2n \log n}{n^2} = \frac{2 \log n}{n} = o(1)$$

schließen kann.

8.3 Folgen in metrischen Räumen

Bisher haben wir von Anordnung, Vollständigkeit und Monotonie Gebrauch gemacht, und deshalb haben wir uns auf reelle Zahlenfolgen beschränkt. Aber schon allein für Folgen komplexer Zahlen, und erst recht für Folgen von Vektoren, Matrizen oder Funktionen brauchen wir eine erheblich allgemeinere Theorie der Folgen.

Definition 8.16 *Es sei M eine beliebige Menge. Eine* **Folge** *von Elementen von M ist eine Abbildung von $\mathbb{N}$ in M bzw. ein Element von $M^{\mathbb{N}}$. In Anlehnung an die n–Tupel verwendet man für Folgen oft die Schreibweise $(a_n)_n \in M^{\mathbb{N}}$.*

Um Grenzwerte und Konvergenz behandeln zu können, brauchen wir ϵ–Umgebungen, und dazu wiederum brauchen wir einen Abstandsbegriff oder etwas Allgemeineres, etwa eine **Topologie**[1]. Eine additive abelsche Gruppenstruktur wie in Vektorräumen ist nicht unbedingt nötig. Ist M ein metrischer Raum (siehe Definition 5.1) mit einer Metrik $d : M \times M \to \mathbb{R}_{\geq 0}$, so wird man eine ϵ–Umgebung eines Elementes $x \in M$ als

$$\{y \in M : d(x, y) < \epsilon\}$$

definieren, und dann ist klar, was Konvergenz bedeuten soll:

Definition 8.17 *Eine* **Folge** $(a_n)_n \in M^{\mathbb{N}}$ *in einem metrischen Raum M mit Metrik d ist* **konvergent** *gegen einen* **Grenzwert** *oder* **Limes** $x \in M$,

[1] `http://de.wikipedia.org/wiki/Topologie_(Mathematik)`

wenn es zu jedem reellen $\epsilon > 0$ ein $N \in \mathbb{N}$ gibt, so daß für alle $n \geq N$ die Abschätzung

$$d(a_n, x) < \epsilon$$

gilt, d.h. wenn die reelle Zahlenfolge $(d(a_n, x))_n \in \mathbb{R}^{\mathbb{N}}$ der Distanzen zwischen x und den Folgengliedern a_n eine Nullfolge ist. Man schreibt auch hier

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

Weil wir weder eine Addition noch eine Skalarmultiplikation zur Verfügung haben, gibt es wenig Möglichkeiten, indirekt die Konvergenz einer Folge zu erschließen. Konstante Folgen $(a_n)_n = (x)_n$ sind natürlich immer konvergent gegen x.

Theorem 8.18 *Ist in einem metrischen Raum M mit Metrik d eine Folge $(a_n)_n \in M^{\mathbb{N}}$ konvergent gegen $x \in M$ und ist $(b_n)_n \in M^{\mathbb{N}}$ eine weitere Folge in M, so daß die reelle Zahlenfolge $(d(a_n, b_n))_n$ eine Nullfolge ist, so ist auch $(b_n)_n$ gegen x konvergent.*

Der Beweis ist nicht schwierig, weil man mit der Dreiecksungleichung

$$0 \leq d(b_n, x) \leq d(b_n, a_n) + d(a_n, x)$$

hat, und die beiden rechtsstehenden Zahlenfolgen sind Nullfolgen. □

Sehen wir uns erst einmal den Fall $M = \mathbb{R}^k$ mit der aus der Maximumsnorm folgenden Distanz

$$d(x, y) = \|x - y\|_\infty = \max_{1\leq m\leq k} |x_m - y_m|$$

an. Was folgt dann aus der Konvergenz einer Folge $(x^n)_n$ von Vektoren $x^n \in \mathbb{R}^k$ gegen einen Vektor $y \in \mathbb{R}^k$? Das ist einfach: die reelle Zahlenfolge $(a_n)_n$ mit Gliedern

$$a_n := d(x^n, y) = \|x^n - y\|_\infty \geq |x^n_m - y_m| \text{ für alle } m,\ 1 \leq m \leq k$$

muß eine Nullfolge sein. Also sind die Folgen $(|x^n_m - y_m|)_n$ für alle $m,\ 1 \leq m \leq k$ Nullfolgen, d.h. man hat Konvergenz der reellen Zahlenfolge $(x^n_m)_n$ gegen y_m in der m–ten Komponente, und das gilt für alle Komponenten. Diese Schlußweise läßt sich leicht umkehren (Frage: wie?):

Theorem 8.19 *Im Raum $\mathbb{R}^k$ mit dem Maximumsnorm–Abstand ist die Konvergenz von Folgen von Vektoren äquivalent zur Konvergenz der Zahlenfolgen in jeder Komponente.*

Im $\mathbb{R}^k$ kann man also Folgen von Vektoren $(x^n)_n$ untersuchen, indem man die k reellen Zahlenfolgen $(x_k^n)_n$ der k Komponenten untersucht.

Gilt Satz 8.19 auch für andere Abstandsbegriffe? Weil die Konvergenz der Zahlenfolgen in allen Komponenten von der Wahl der Metrik gar nicht abhängt, kann man das vermuten. Die eine Richtung des folgenden Satzes ist einfach zu beweisen. die andere wird Probleme machen.

Theorem 8.20 *Im Raum $\mathbb{R}^k$ mit einer beliebigen Norm ist die Konvergenz von Folgen von Vektoren äquivalent zur Konvergenz der Zahlenfolgen in jeder Komponente.*

Beweis: Wir beweisen erst nur die Aussage

> Im Raum $\mathbb{R}^k$ sei eine Folge $(x^n)_n$ von Vektoren $x^n \in \mathbb{R}^k$ gegeben, deren Komponenten sämtlich konvergieren, d.h. es gibt einen Vektor $y \in \mathbb{R}^k$ mit $\lim_{n\to\infty} x_j^n = y_j,\ 1 \le j \le k$. Dann folgt in jeder durch eine Norm erzeugten Metrik die Konvergenz gegen y.

Wir benutzen den schon bewiesenen Teil von Theorem 5.7 um

$$\|x\| \le C \cdot \|x\|_\infty$$

für alle $x \in R^k$ mit einer Konstanten C zu bekommen. Hat man dann eine Folge $(x^n)_n \in (\mathbb{R}^k)^{\mathbb{N}}$ von Vektoren, die komponentenweise gegen einen Vektor $y \in \mathbb{R}^k$ konvergieren, so ist nach Satz 8.19 die Zahlenfolge $(\|y - x^n\|_\infty)_n$ eine Nullfolge, und nach der obigen Abschätzung muß auch $(\|y - x^n\|)_n$ eine Nullfolge sein. □

Die Umkehrung der Aussage von Satz 8.20 ist eng mit der noch unbewiesenen Abschätzung

$$c \cdot \|x\|_\infty \le \|x\| \text{ für alle } x \in \mathbb{R}^k \tag{8.21}$$

aus Theorem 5.7 verbunden, wobei die rechts stehende Norm beliebig und c positiv ist. Wenn man diese Abschätzung beweisen kann, folgen zu den Sätzen 5.7 und 8.20 die noch unbewiesenen Umkehrungen. Es zeigt sich, daß der Beweis auf den Satz 8.11 zurückgeht, den wir im Kapitel über Vektorräume noch nicht zur Verfügung hatten.

Wir definieren

$$c := \inf\{\|z\| \ :\ z \in \mathbb{R}^k,\ \|z\|_\infty = 1\} \ge 0$$

und wollen beweisen, daß c positiv ist. Wenn wir das geschafft haben, nehmen wir ein beliebiges $x \ne 0$, setzen $z := x/\|x\|_\infty$ und bekommen wegen $\|z\|_\infty = 1$

die Behauptung (8.21) in der Form $c \leq \|z\| = \|x\|/\|x\|_\infty$, denn für $x = 0$ ist nichts zu beweisen.

Wir nehmen jetzt also an, die Zahl c sei gleich Null und wählen eine Folge $(z^n)_n$ von Vektoren des $\mathbb{R}^k$ mit $\|z^n\|_\infty = 1$ und $\lim_{n\to\infty} \|z^n\| = c = 0$. Die Komponenten z^n_m der Vektoren z^n liegen alle in $[-1, 1]$ und es ist zu jedem n immer mindestens eine der Komponenten z^n_m, $1 \leq m \leq k$ gleich 1 oder -1. In mindestens einer Komponente r, $1 \leq r \leq k$ tritt dieser Fall unendlich oft auf, und wir gehen zu einer Teilfolge über, die wir wieder $(z^n)_n$ nennen, und die $|z^n_r| = 1$ für alle $n \in \mathbb{N}$ erfüllt. Da es auf Faktoren ± 1 bei der Infimumsbildung nicht ankommt, können wir sogar $z^n_r = 1$ für alle $n \in \mathbb{N}$ annehmen. Von dieser Teilfolge bilden wir nacheinander Teilfolgen, die in der ersten, zweiten und schließlich k–ten Komponente konvergieren, denn die Zahlenfolgen z^n_m sind ja beschränkt auf $[-1, 1]$. Die resultierende Folge nennen wir wieder $(z^n)_n$, und sie ist jetzt in jeder Komponente konvergent, erfüllt $z^n_r = 1$ und $\|z^n\|_\infty = 1$ für alle n und liefert immer noch $\lim_{n\to\infty} \|z^n\| = c = 0$, weil sie Teilfolge der ursprünglichen Folge ist. Der Limes ist ein Vektor $y \in [-1, 1]^k$ mit $y_r = 1$. Die Abschätzung

$$\begin{aligned} \|y\| &= \|y - z^n + z^n\| \\ &\leq \|y - z^n\| + \|z^n\| \\ &\leq C \cdot \|y - z^n\|_\infty + \|z^n\| \end{aligned}$$

hat auf der rechten Seite Nullfolgen, liefert also $\|y\| = 0$. Dann folgt $y = 0$ im Widerspruch zu $y_r = 1$. □

Theorem 8.22 *(mehrdimensionaler Satz von* **Bolzano**[1]**–Weierstrass**[2]*)*
In endlichdimensionalen Vektorräumen über $\mathbb{R}$ oder $\mathbb{C}$ hat jede in irgendeiner Norm beschränkte Folge eine konvergente Teilfolge.

Beweis: Wir führen den Beweis für $\mathbb{R}^k$ aus und verweisen im allgemeinen Fall auf den Standardisomorphismus. Ist eine Folge im $\mathbb{R}^k$ in einer beliebigen Norm beschränkt, so ist sie wegen (8.21) in der Norm $\|.\|_\infty$ beschränkt. Dann kann man wie im Beweis des vorigen Satzes schrittweise eine Teilfolge auswählen, die in allen Komponenten konvergiert. Nach Satz 8.20 konvergiert die Folge dann auch in der Metrik zur gegebenen Norm. □

Die Verfolgung der Frage, wann sich aus Beschränktheit einer Folge bereits die Konvergenz einer Teilfolge ergibt, führt in die Disziplin **Topologie**[3],

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Bolzano.html`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Weierstrass.html`
[3] `http://de.wikipedia.org/wiki/Topologie_(Mathematik)`

wenn man abstrahiert, und in die **Funktionalanalysis**, wenn man konkrete unendlichdimensionale normierte Vektorräume studiert.

In Richtung auf die Funktionalanalysis machen wir aber noch einen kleinen Schritt. Wenn wir in Vektorräumen arbeiten wollen, haben wir keine Anordnung mehr (siehe $\mathbb{C}$), und man kann aus beliebigen Folgen keine monotonen Teilfolgen auswählen. Die Vollständigkeit, die uns Grenzwerte indirekt über Suprema und Infima zusichert, kann nicht mehr so wie bei den reellen Zahlen (vgl. Satz 3.21) ausgedrückt werden. Was tun?

Der Ausweg besteht darin, das Konvergenzkriterium 8.9 von **Cauchy**[1] nicht als Satz anzustreben, sondern zur Definition der Vollständigkeit zu machen.

Definition 8.23 *Es sei M ein metrischer Raum mit Metrik d.*

1. *Eine Folge $(x_n)_n \in M^{\mathbb{N}}$ heißt* **Cauchyfolge***, wenn es zu jedem $\epsilon > 0$ ein $N \in \mathbb{N}$ gibt mit*

$$d(x_n, x_m) < \epsilon \text{ für alle } n, m \geq N.$$

2. *Ein metrischer Raum heißt* **vollständig***, wenn in ihm jede Cauchyfolge konvergiert.*

Nach Satz 3.21 ist $\mathbb{R}$ auch in diesem Sinne vollständig, aber wir wollen natürlich etwas über die Vollständigkeit von allgemeineren Vektorräumen wissen:

Theorem 8.24 *Jeder normierte endlichdimensionale Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$ ist vollständig, d.h. jede Cauchyfolge konvergiert.*

Wir führen den **Beweis** nur für $\mathbb{R}^k$ vor. Es sei also $(x^n)_n \in (\mathbb{R}^k)_{\mathbb{N}}$ eine Cauchyfolge von Vektoren, d.h. zu jedem $\epsilon > 0$ gibt es ein $N \in \mathbb{N}$ mit

$$\|x^n - x^m\| < \epsilon \text{ für alle } n, m \geq N.$$

Wegen der Normäquivalenz, insbesondere (8.21), hat man dann auch eine Cauchyfolge in der Norm $\|.\|_\infty$, und es folgt für alle Komponenten

$$|x_j^n - x_j^m| \leq \|x^n - x^m\|_\infty < \epsilon \text{ für alle } n, m \geq N.$$

Also liegt in jeder Komponente eine reelle Cauchyfolge vor, die nach Satz 8.9 konvergiert. Mit Satz 8.20 folgt dann die Konvergenz der Vektoren in der

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Cauchy.html`

Norm. Im Falle $\mathbb{R}^k$ ist damit der Beweis des Satzes erbracht. Die Erweiterung auf endlichdimensionale Vektorräume über $\mathbb{R}$ oder $\mathbb{C}$ erfolgt mit dem Standard–Isomorphismus. □

Im Spezialfall des Körpers $\mathbb{C}$ verfahren wir wie im $\mathbb{R}^2$ unter der euklidischen Norm bzw. der dadurch definierten Metrik. Die euklidische Norm von $(x, y) \in \mathbb{R}^2$ stimmt mit dem Absolutbetrag von $x + iy \in \mathbb{C}$ überein, und Konvergenz einer Folge wird deshalb wie im Reellen definiert, wobei aber der Absolutbetrag eine andere Bedeutung hat. Sofern keine Ordnung oder Monotonie benutzt wird, gelten alle Aussagen ganz analog, z.B. der Satz von Bolzano–Weierstraß. Jede Cauchyfolge ist konvergent, und konvergente Folgen sind immer Cauchyfolgen. Konvergenz einer komplexen Folge ist gleichbedeutend mit gleichzeitiger Konvergenz der durch Real– und Imaginärteil definierten reellen Folgen, und das nutzt man aus, wenn man komplexe Folgen auf Konvergenz untersucht.

8.4 Abgeschlossene und offene Mengen

Für das nächste Kapitel benötigen wir noch Begriffe, die in allgemeinerer Form der Disziplin **Topologie**[1] (Wissenschaft von der "Lage", d.h. eine Art Geometrie ohne Metrik) zuzuordnen sind:

Definition 8.25 *Es sei M ein metrischer Raum mit Metrik d, und es sei N eine Teilmenge von M. Der (topologische oder metrische)* **Abschluß** $\overline{M}$ *(lies: M–quer) von N ist dann die Menge aller Grenzwerte von Folgen, die in N liegen und in M konvergent sind. Eine Menge N heißt* **abgeschlossen***, wenn $N = \overline{N}$ gilt, d.h. wenn sie zu jeder in M konvergenten und in N liegenden Folge $(x_n)_n$ auch den Grenzwert $x = \lim_{n\to\infty} x_n$ enthält.*

Man mache sich das am Beispiel reeller Intervalle klar. Ein Intervall der Form $[a, b]$ mit $a < b$ ist abgeschlossen, ein Intervall $(a, b]$ oder (a, b) nicht, weil man gegen a konvergente Folgen finden kann, deren Grenzwert a nicht zum Intervall gehört. Hat man ein Rechenverfahren, das ein Ergebnis als Limes einer Folge berechnet (z.B. beim Wurzelziehen), so muß das Verfahren immer in einer abgeschlossenen Menge ablaufen, wenn gesichert sein soll, daß der Grenzwert wieder zur Menge gehört. Obendrein gilt

Theorem 8.26 *Eine nichtleere abgeschlossene Teilmenge eines vollständigen metrischen Raumes ist wieder ein vollständiger metrischer Raum.*

[1] `http://de.wikipedia.org/wiki/Topologie_(Mathematik)`

Das wollen wir hier nicht beweisen, sondern als Übung lassen.

Reelle Intervalle, die zwar beschränkt, aber nicht abgeschlossen sind, haben nicht notwendig ein minimales oder maximales Element. Aber:

Theorem 8.27 *Nichtleere beschränkte und abgeschlossene Mengen in endlichdimensionalen normierten Vektorräumen über $\mathbb{R}$ oder $\mathbb{C}$ haben ein in der Norm minimales bzw. maximales Element.*

Beweis: Es sei $M \subset \mathbb{R}^k$ beschränkt, d.h. es gelte $\|x\| \leq K$ für alle $x \in M$ mit einer positiven Konstanten K. Dann ist die Menge $\{\|x\| \ : \ x \in M\} \subseteq [0, K] \subset \mathbb{R}$ beschränkt und besitzt ein Infimum s^- und ein Supremum s^+. Man kann also eine Folge $(x_n)_n$ in M finden, so daß z.B. $\lim_{n\to\infty} \|x_n\| = s^+$ gilt. Nach dem Satz von Bolzano–Weierstraß gibt es dazu eine konvergente Teilfolge, die wir wieder $(x_n)_n$ nennen. Deren Limes x muß wegen der Abgeschlossenheit von M wieder in M liegen, und es folgt

$$\|x\| = \lim_{n\to\infty} \|x_n\| = s^+,$$

d.h. x ist ein in der Norm maximales Element von M. Analog verfährt man beim minimalen Element. □

Man mache sich klar, daß aus Abgeschlossenheit nicht Beschränktheit folgt, denn jeder metrische Raum, auch $\mathbb{R}$ und $\mathbb{C}^k$, ist selbst abgeschlossen. Und am Beispiel $(0, 1)$ sieht man, daß Beschränktheit nicht zu Abgeschlossenheit führt.

Theorem 8.28 *Die Vereinigung von endlich vielen und der Durchschnitt von beliebig vielen abgeschlossenen Mengen ist abgeschlossen.*

Der Beweis ist nicht schwierig, soll aber dennoch übergangen werden. □

Wir wollen jetzt noch Umgebungen und offene Mengen behandeln, aber die Beweise nicht ausführen.

Definition 8.29 *Es sei x ein Element eines metrischen Raums M mit Distanzfunktion d.*
Eine (offene) ϵ–Umgebung von x ist dann die Menge

$$U_\epsilon(x) := \{y \in M \ : \ d(x, y) < \epsilon\}.$$

Eine **Umgebung** *von x ist eine Menge, die mindestens eine ϵ–Umgebung von x enthält.*
Eine Teilmenge N von M ist eine **offene Menge**, *wenn es zu jedem Punkt $x \in N$ auch eine Umgebung $U(x)$ gibt, die ganz in N liegt.*

Man mache sich klar, daß eine wie oben definierte ϵ–Umgebung immer offen ist. Eine **abgeschlossene** ϵ–Umgebung ist

$$\overline{U}_\epsilon(x) := \{y \in M \ : \ d(x,y) \leq \epsilon\}.$$

Das logische Gegenteil von "offen" ist nicht "abgeschlossen", und dasselbe gilt auch in der anderen Richtung. Ein Intervall der Form $I := [a,b)$ ist weder offen noch abgeschlossen, denn b gehört nicht dazu und a hat keine Umgebung in I. Der gesamte metrische Raum M ist immer eine offene und abgeschlossene Teilmenge von sich selbst.

Theorem 8.30 *Die Vereinigung beliebig vieler und der Durchschnitt endlich vieler offener Teilmengen desselben metrischen Raums ist offen.*
Der Abschluß einer Teilmenge N eines metrischen Raums ist der Durchschnitt aller abgeschlossenen Teilmengen, die N enthalten.

Definition 8.31 *Der* **offene Kern** $\overset{\circ}{N}$ *einer Teilmenge N eines metrischen Raums ist die Vereinigung aller offenen Teilmengen von N.*
Der **Rand** ∂N *einer Teilmenge N eines metrischen Raums ist die mengentheoretische Differenz $\partial N := \overline{N} - \overset{\circ}{N}$ zwischen Abschluß und offenem Kern.*

8.5 Schreibweisen für allgemeine Grenzprozesse

Es tritt oft die Situation ein, daß eine Aussage, die für eine gegen ein x konvergente Folge $(x_n)_n$ gilt, gar nicht von der Auswahl der Folge, sondern nur von x abhängt. Man verwendet für solche allgemeinen Grenzprozesse auch oft die etwas nachlässige Notation

$$\lim_{y \to x} A(y,x) = B(x)$$

mit der Bedeutung

Für alle gegen x konvergenten Folgen $(x_n)_n$ gilt

$$B(x) = \lim_{n \to \infty} A(x_n, x)$$

wenn die Ausdrücke A und B Sinn machen. Typische Beispiele sind

$$\lim_{x \to 0} x^2 = 0 \text{ oder } \lim_{z \to 1} \frac{z^n - 1}{z - 1} = n.$$

Wir werden diese Notation auch verwenden, wenn Folgen gegen $\pm\infty$ streben im Sinne von Definition 8.5 auf Seite 220, z.B. wenn wir schreiben

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

In analoger Weise verallgemeinert man die **Landau–Symbole** $\mathcal{O}$ und o für allgemeinere Grenzprozesse. Mit der Schreibweise

$$A(y) = \mathcal{O}(B(y)) \text{ für } y \to x$$

meint man, daß es eine Konstante K und eine Umgebung U von x gibt, so daß für alle $y \in U$ die Ungleichung $|A(y)| \leq K|B(y)|$ gilt. Anders ausgedrückt: es gibt ein $K > 0$ und ein $\epsilon > 0$, so dass für alle y mit $|x - y| < \epsilon$ auch $|A(y)| \leq K|B(y)|$ folgt. Jede gegen x konvergente Folge $(x_n)_n$ hat dann ein Endstück in der festen ϵ–Umgebung von x und es gilt $|A(x_n)| \leq K|B(x_n)|$ für alle diese Endstücke. Die wegzulassenden Anfangsstücke der Folgen braucht man nicht zu spezifizieren, wenn man die ϵ–Umgebung spezifiziert hat.

Bei o verfährt man analog, denn

$$A(y) = o(B(y)) \text{ für } y \to x$$

soll bedeuten, daß es zu jedem $\epsilon > 0$ eine Umgebung U von x gibt, so daß für alle $y \in U$ die Ungleichung $|A(y)| \leq \epsilon|B(y)|$ gilt. Anders ausgedrückt: für alle $\epsilon > 0$ gibt es ein $\delta > 0$, so dass für alle y mit $|x - y| < \delta$ auch $|A(y)| \leq \epsilon|B(y)|$ folgt. Wir werden dies in Satz 12.2 auf Seite 304 benutzen.

9 Eigenwerte

Weil wir jetzt Folgen im $\mathbb{R}^n$ zur Verfügung haben, können wir unsere Argumentation aus Abschnitt 5.4 auf Seite 174 wieder aufgreifen. Es wird sich herausstellen, daß wir nicht nur Folgen von Vektoren, sondern auch Folgen von Matrizen behandeln müssen. Ein "folgenloses" Vorgehen ist nicht möglich, denn das gestellte Problem läßt sich nicht mit endlich vielen Rechenschritten lösen.

9.1 Grundlagen

Definition 9.1 *Ist $A \in K^{n\times n}$ eine quadratische Matrix, und gilt $A \cdot x = \lambda x$ mit einem Skalar $\lambda \in K$ und einem vom Nullvektor verschiedenen Vektor $x \in K^n$, so heißt λ* **Eigenwert**[1] *von A und x ist der zu λ und A gehörige* **Eigenvektor**.

Diese Definition besagt, dass die Wirkung von A als lineare Abbildung auf die durch x definierte Nullpunktsgerade eine reine **Streckung** um den Faktor λ ist. Auf diesem Unterraum hat A also eine besonders einfache Form. Wir tragen einige elementare Eigenschaften von Eigenwerten und Eigenvektoren zusammen:

Theorem 9.2

1. *Eigenwerte von reellen symmetrischen bzw. von komplexen hermiteschen Matrizen sind immer reell.*
2. *Eigenvektoren zu verschiedenen Eigenwerten von reellen symmetrischen bzw. von komplexen hermiteschen Matrizen sind immer orthogonal.*
3. *Eigenwerte von positiv semidefiniten Matrizen sind reell und nicht negativ.*
4. *Eigenwerte von positiv definiten Matrizen sind reell und positiv.*
5. *Eigenwerte λ erfüllen die Gleichung $\det(A - \lambda \cdot I) = 0$. Die Funktion $p(\lambda) := \det(A - \lambda \cdot I)$ heißt* **charakteristisches Polynom**.

[1] `http://de.wikipedia.org/wiki/Eigenwert`

Beweis: Es gelte $A = A^* \in K^{n\times n}$ und $Ax = \lambda x,\ Ay = \mu y$ mit evtl. komplexen Eigenwerten $\lambda \neq \mu$ und evtl. komplexen Eigenvektoren $x, y \in \mathbb{C}^n \setminus \{0\}$. Dann folgt

$$\lambda\|x\|_2^2 = \lambda x^T\overline{x} = (Ax)^T\overline{x} = x^TA^T\overline{x} = x^T\overline{A}\overline{x} = x^T\overline{\lambda}\overline{x} = \overline{\lambda}\|x\|_2^2$$

und λ muss reell sein, ebenso wie dann auch μ. Ferner gilt

$$\lambda(x,y) = \lambda x^T\overline{y} = x^TA^T\overline{y} = x^T\overline{A}\overline{y} = x^T\overline{\mu y} = \mu x^T\overline{y} = \mu(x,y)$$

und aus $\lambda \neq \mu$ folgt $(x,y) = 0$. Ist A positiv semidefinit, so können wir die quadratische Form q_A aus (5.15) von Seite 170 auf einem Eigenvektor auswerten und bekommen aus

$$0 \leq q_A(\overline{x}) = \overline{x}^TAx = \overline{x}^T\lambda x = \lambda\|x\|_2^2$$

die Nichtnegativität des passenden Eigenwerts. Genauso folgt die Positivität der Eigenwerte positiv definiter Matrizen. Die letzte Behauptung folgt daraus, daß ein Eigenvektor $x \neq 0$ zum Eigenwert λ eine nichttriviale Lösung des homogenen linearen Gleichungssystems $(A - \lambda\cdot I)x = 0$ ist, und deshalb muss die Determinante der Koeffizientenmatrix verschwinden. Aus Theorem 6.15 folgt dann, daß die Funktion $\det(A - \lambda\cdot I)$ ein Polynom ist. □

Die Frage ist nun, ob man zu jeder linearen Abbildung eines n–dimensionalen Raumes in sich eine geeignete Basenwahl treffen kann, so dass sich die Abbildung bei Matrixdarstellung in dieser Basis als reine Streckung schreiben läßt. Wir wollen also, wenn die Abbildung $A : K^n \to K^n$ durch eine $n\times n$–Matrix A gegeben ist, eine Basis $\{v_1,\ldots,v_n\}$ von K^n finden, so daß mit geeigneten Skalaren $\lambda_j \in K$ die Gleichungen $A\cdot v_j = \lambda_j v_j,\ 1 \leq j \leq n$ gelten. Bauen wir die gesuchten Basisvektoren v_j als Spalten in eine nichtsinguläre $n\times n$–Matrix V ein, so würde

$$\begin{aligned} Av_j &= \lambda_j v_j \\ AVe_j &= \lambda_j Ve_j \\ V^{-1}AVe_j &= \lambda_j V^{-1}Ve_j \\ &= \lambda_j e_j,\ \ 1 \leq j \leq n \end{aligned}$$

gelten, wenn unser Vorhaben gelänge. Das würde aber bedeuten, dass die Matrix $V^{-1}AV$ eine reine Diagonalmatrix

$$D = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} = V^{-1}AV$$

wäre, was wiederum bedeutet, daß sich A als $A = VDV^{-1}$ schreiben läßt.

Leider ist es falsch, daß **jede** $n \times n$–Matrix A über K in diesem Sinne **diagonalisierbar** ist. Wir sehen uns einen Spezialfall an. Ist die obige Basis sogar orthonormal, so ist V eine Orthogonalmatrix, und weil D symmetrisch ist, folgt dann aus

$$A^T = (VDV^{-1})^T = (VDV^T)^T = VD^TV^T = VDV^T = A$$

die Symmetrie von A. Dieser Sonderfall ist leichter zu behandeln:

Theorem 9.3 *Jede symmetrische reelle $n \times n$–Matrix ist durch eine geeignete Orthogonalmatrix diagonalisierbar. Es gibt eine Orthonormalbasis des $\mathbb{R}^n$ aus Eigenvektoren von A.*

Der konstruktive Beweis dieses Satzes folgt im nächsten Abschnitt. Hier soll nur noch erwähnt werden, daß die Diagonalisierung symmetrischer Matrizen in Physik und Technik oft auftritt, und zwar im Zusammenhang mit Schwingungsproblemen von mechanischen oder elektrischen Systemen. Die gesuchten Eigenwerte entsprechen den Wellenlängen der "Grundschwingungen" des Systems, und die Eigenvektoren beschreiben die "Moden" der Grundschwingungen. Ein typischer Fall ist die Berechnung der Obertöne einer schwingenden Saite zusammen mit den dazu passenden stehenden Wellen.

Die Transformation allgemeiner, nicht notwendig symmetrischer quadratischer Matrizen auf **Jordan–Normalform** ist schwieriger und wird hier nicht behandelt, auch weil sie mehr für die Theorie als für die Praxis von Bedeutung ist. Stattdessen bringen wir in Abschnitt 9.3 auf Seite 250 die praktisch wesentlich wichtigere **Singulärwertzerlegung**.

9.2 Das Jacobi-Verfahren für symmetrische Matrizen

C.G.J. Jacobi[1] hat 1845/46 ein Verfahren zur Behandlung des Eigenwertproblems symmetrischer $n \times n$-Matrizen angegeben, das für nicht zu große n auch heute noch brauchbar ist. Für große oder unsymmetrische Matrizen nimmt man andere Verfahren, die in der Veranstaltung "Numerische Mathematik II" oder "Numerische lineare Algebra" behandelt werden. Das Verfahren berechnet **alle** Eigenwerte (und wenn nötig, auch die Eigenvektoren) und beruht auf

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Jacobi.html`

Theorem 9.4 *Ist $A = A^T = (a_{ij})$ eine reelle symmetrische $n \times n$-Matrix, so ist die Größe*

$$\sum_{k,j=1}^{n} a_{jk}^2$$

invariant gegen orthogonale Transformationen.

Der **Beweis** ist eine einfache Folgerung aus Satz 5.17, denn die euklidischen Normen aller Spaltenvektoren sind invariant, deshalb auch die Quadratsummen der Spalten und der gesamten Matrix. □

Setzt man

$$N(A) := \sum_{i \neq k} a_{ik}^2,$$

so folgt

$$\sum_{k,j=1}^{n} a_{jk}^2 = N(A) + \sum_{j=1}^{n} a_{jj}^2. \tag{9.5}$$

Da die linke Seite dieser Gleichung gegenüber orthogonalen Transformationen invariant ist, wird man versuchen, durch geeignete orthogonale Transformationen die Größe $N(A)$ zu verkleinern und damit durch Vergrößern von $\sum_{j=1}^{n} a_{jj}^2$ die Matrix A in eine Diagonalmatrix zu überführen. Dazu kann man ein Element $a_{ij} \neq 0$ mit $i \neq j$ auswählen und in der durch e_i und e_j aufgespannten Ebene eine Transformation ausführen, die a_{ij} in Null überführt. Setzt man die Transformation im $\mathbb{R}^2$ als Drehung um einen Winkel α an, so liefert die Ähnlichkeitstransformation (JACOBI-Transformation oder GIVENS-Rotation genannt)

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \cdot \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix} \cdot \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

eine Diagonalmatrix, wenn das Nebendiagonalelement

$$\begin{aligned} & a_{ij}(\cos^2\alpha - \sin^2\alpha) + (a_{jj} - a_{ii}) \cos\ \alpha\ \sin\ \alpha \\ = \ & a_{ij} \cos\ 2\alpha + (a_{jj} - a_{ii})\ \tfrac{1}{2}\ \sin\ 2\alpha \end{aligned}$$

verschwindet. Man könnte also den Winkel α aus

$$\cot 2\alpha = \frac{a_{ii} - a_{jj}}{2a_{ij}}$$

bestimmen, aber es ist möglich, die Winkelfunktionen zu vermeiden, wenn man zunächst die Größe $\tau := \cos\ 2\alpha = \cos^2\alpha - \sin^2\alpha$ einführt, dann $\cos\ \alpha = \sqrt{(1+\tau)/2}$ und $\sin\alpha = \sigma \cdot \sqrt{(1-\tau)/2}$ definiert sowie das Vorzeichen σ des Sinus so wählt, daß

$$a_{ij} \cdot \tau + (a_{jj} - a_{ii}) \cdot \frac{\sigma}{2}\ \sqrt{1-\tau^2} = 0$$

gilt. Das wiederum ist erzielbar, wenn $\sigma = \operatorname{sgn} a_{ij}$ und

$$\tau = (a_{ii} - a_{jj})/(4a_{ij}^2 + (a_{ii} - a_{jj})^2)^{1/2}$$

gesetzt wird. Damit wäre das Problem für 2×2-Matrizen gelöst.

Im allgemeinen Fall verwendet man Transformationsmatrizen

$$T_{ij}(\alpha) := E + (c-1)(e_j e_j^T + e_i e_i^T) + s(e_j e_i^T - e_i e_j^T) \tag{9.6}$$

mit

$$c := (\frac{1+\tau}{2})^{1/2},\ \ s := \sigma \cdot (\frac{1-\tau}{2})^{1/2} \quad \text{und}$$

$$\sigma := \operatorname{sgn}\,(a_{ij}),$$

$$\tau := \frac{a_{ii} - a_{jj}}{\sqrt{(a_{ii} - a_{jj})^2 + 4a_{ij}^2}},$$

wenn $a_{ij} \neq 0$ gilt. Es folgt

Lemma 9.7 *Wählt man zwei Indizes i, j mit $a_{ij} \neq 0$, so verschwindet $b_{ij} = b_{ji}$ für die Matrix*

$$B := T_{ij}(\alpha) \cdot A \cdot T_{ij}^T(\alpha), \tag{9.8}$$

und es gilt

$$N(B) = N(A) - 2a_{ij}^2. \tag{9.9}$$

Beweis: Aus der Invarianz der Gleichung (9.5) gegenüber orthogonalen Transformationen folgt

$$N(A) - N(B) = \sum_{k=1}^{n} (b_{kk}^2 - a_{kk}^2) = b_{jj}^2 + b_{ii}^2 - a_{jj}^2 - a_{ii}^2, \tag{9.10}$$

da B aus A durch Umformung der Zeilen und Spalten mit den Indizes i und j entsteht. Die rechte Seite von (9.10) kann man aber bereits bei 2×2-Matrizen betrachten:

$$\begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \cdot \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix} \cdot \begin{pmatrix} c & s \\ -s & c \end{pmatrix}. \tag{9.11}$$

In diesen Teilmatrizen werden nämlich die Größen $b_{ij},\ b_{ji},\ b_{jj},\ b_{ii}$ ebenso berechnet wie in der Gleichung (9.8). Die Invarianz von (9.5) für die Teilmatrizen liefert

$$b_{ii}^2 + b_{jj}^2 + 2b_{ij}^2 = a_{ii}^2 + a_{jj}^2 + 2a_{ij}^2,$$

d.h. mit (9.10) gilt

$$N(A) - N(B) = 2(a_{ij}^2 - b_{ij}^2), \tag{9.12}$$

was wegen $b_{ij} = 0$ zu (9.9) führt. □

Um Mißverständnissen vorzubeugen: die Matrix B aus (9.8) wird in der Praxis nicht berechnet. Die Rechenorganisation wird weiter unten beschrieben.

Durch eine (orthogonale) Transformation mit $T_{ij}(\alpha)$ kann man also jeweils eines der Nichtdiagonalelemente in Null überführen und die Summe der Quadrate der Nebendiagonalelemente verkleinern. Durch sukzessive Anwendung von orthogonalen Transformationen $T_{ij}(\alpha)$ für verschiedene i, j kann man damit erreichen, daß A gegen eine Diagonalmatrix strebt, auch wenn nachfolgende Transformationen die schon erzielten Nullen wieder verändern. Je nach Auswahl des nächsten zu annullierenden Elementes a_{ij} erhält man verschiedene Varianten des Verfahrens. Das **klassische** JACOBI**-Verfahren** wählt in jedem Schritt das betragsmäßig größte Nichtdiagonalelement aus und bekommt dann

$$N(B) \leq N(A)\ \left(1 - \frac{1}{n(n-1)}\right),$$

d.h. die Folge der Werte $N(A)$ strebt mindestens so schnell gegen Null wie die geometrische Folge $(q^k)_k$ mit $q = 1 - \frac{1}{n(n-1)}$. Weil die Größe $N(A)$ mit jedem Schritt des Verfahrens kleiner wird, bekommt man für alle Nebendiagonalelemente der Matrix eine Nullfolge. Weil die Gesamt–Quadratsumme invariant bleibt, konzentriert sie sich immer mehr auf die Diagonale. Es liegt im Laufe des Verfahrens eine Folge von Matrizen vor, deren Nichtdiagonalelemente jeweils gegen Null konvergieren. Die Diagonalelemente sind beschränkt, aber nicht notwendig konvergent. Man bekommt lediglich nach dem Satz 8.22 von Bolzano–Weierstraß eine konvergente Teilfolge.

Theorem 9.13 *Das Jacobi–Verfahren zur Berechnung der Eigenwerte und eines vollen orthonormalen Systems von Eigenvektoren einer symmetrischen reellen $n \times n$–Matrix A erzeugt eine Folge von reellen symmetrischen $n \times n$–Matrizen $A^{(k)}$, wobei $A^{(k+1)} = T_{ij}(\alpha)A^{(k)}T_{ij}^T(\alpha)$ mit von k anhängigen Indizes $1 \leq i < j \leq n$ und einem von k abhängigen Drehwinkel α gilt. Alle Häufungspunkte der Folge sind Diagonalmatrizen mit den Eigenverten von A in der Diagonale.*

Es sollte klargestellt werden, daß man keinesfalls Matrizenmultiplikationen ausführt. Aus (9.6) entnimmt man

$$\begin{aligned}
T_{ij}(\alpha)A &= A + (c-1)(e_j e_j^T A + e_i e_i^T A) + s(e_j e_i^T A - e_i e_j^T A) \\
e_k^T T_{ij}(\alpha)A &= e_k^T A + (c-1)(\underbrace{e_k^T e_j}_{=\delta_{kj}} e_j^T A + \underbrace{e_k^T e_i}_{=\delta_{ki}} e_i^T A) \\
&\quad + s(\underbrace{e_k^T e_j}_{=\delta_{kj}} e_i^T A - \underbrace{e_k^T e_i}_{=\delta_{ki}} e_k^T e_i e_j^T A) \\
e_k^T T_{ij}(\alpha)A &= e_k^T A \text{ falls } j \neq k \neq i \\
e_i^T T_{ij}(\alpha)A &= c e_i^T A - s e_i^T A \\
e_j^T T_{ij}(\alpha)A &= c e_j^T A + s e_j^T A
\end{aligned}$$

d.h. nur die i–te und j–te Zeile von $T_{ij}(\alpha)A$ sind zu berechnen, und zwar als Linearkombinationen der entsprechenden beiden Zeilen von A. Ganz analog, aber als Spaltenoperation auf $T_{ij}(\alpha)A$, führt man die Transformation $T_{ij}(\alpha)A \mapsto T_{ij}(\alpha)AT_{ij}^T(\alpha)$ aus.

Hat man nach etlichen Iterationen eine Matrix mit (relativ) kleinen Nichtdiagonalelementen gefunden, so sind die Diagonalglieder Näherungen für die Eigenwerte der Matrix. Das folgt aus einem allgemeinen Satz von **Gerschgorin**:

Theorem 9.14 *Ist $A = (a_{jk})_{1\leq j,k\leq n}$ eine $n \times n$-Matrix mit komplexen Koeffizienten a_{jk}, so erfüllt jeder Eigenwert λ von A wenigstens eine der Ungleichungen*

$$|\lambda - a_{jj}| \leq \sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}| \qquad \text{für } j = 1, \ldots, n.$$

Beweis: Zum Eigenwert λ von A gibt es einen Eigenvektor $x \in \mathbb{C}^n \setminus \{0\}$ mit $Ax = \lambda x$. Eine betragsmäßig größte Komponente x_j von x kann dabei zu 1 normiert werden. Aus $Ax - \lambda x = 0$ folgt

$$(a_{jj} - \lambda)x_j + \sum_{\substack{k=1 \\ k\neq j}}^{n} a_{jk}x_k = 0,$$

d.h. es gilt

$$|a_{jj} - \lambda| = |(a_{jj} - \lambda)x_j| = |\sum_{\substack{k=1 \\ k\neq j}}^{n} a_{jk}x_k| \leq \sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}|\,|x_k| \leq \sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}|.$$

□

Wir müssen noch die Berechnung der Eigenvektoren behandeln:

Theorem 9.15 *Führt man alle Transformationen $T_{ij}(\alpha)$ des Jacobi–Verfahrens nacheinander an der Einheitsmatrix aus, so erhält man eine Folge von orthogonalen Matrizen, deren Häufungspunkte Orthogonalmatrizen sind, die je einen orthonormalen Satz von Eigenvektoren enthalten.*

Beweis: Nach m Schritten gilt

$$B_m = Q_m \cdot A \cdot Q_m^T$$

mit einer geeigneten Orthogonalmatrix Q_m, die sich gerade als Anwendung der $T_{ij}(\alpha)$ auf I_n schreiben läßt. Da die Elemente von Orthogonalmatrizen immer in $[-1, 1]$ liegen müssen, bilden die Elemente der Matrizen Q_m in jeder Komponente eine beschränkte Folge, und man kann mindestens eine Teilfolge dieser Matrizenfolge auswählen, so daß alle Komponenten der Matrizen konvergieren. Dann ist die Grenzwert–Matrix Q wieder orthogonal, weil die Gleichungen $Q_m Q_m^T = I_n = Q_m^T Q_m$ aus Summen von Produkten bestehen und Satz 8.6 dann auch $QQ^T = I_n = Q^T Q$ liefert. Ebenso gelten die Gleichungen $D = QAQ^T$ und $AQ^T = Q^T D$ im Grenzfall. Deshalb enthält Q^T einen vollständigen Satz orthonormaler Eigenvektoren. □

9.3 Singulärwertzerlegung

Wir nehmen jetzt eine reelle und weder symmetrische noch quadratische $m \times n$–Matrix A her und wollen für A eine entsprechende "Diagonalisierung" versuchen. Das gelingt nicht, aber man kann etwas viel Praktischeres als die in der Theorie hergeleitete **Jordan–Normalform** bekommen, nämlich eine **Singulärwertzerlegung**. Obwohl man diese Zerlegung der Praxis anders berechnet, untersuchen wir erst einmal die Diagonalisierung $V^T A^T A V = VD$ der symmetrischen $n \times n$–Matrix $A^T A$ mit Eigenwerten $\lambda_1, \ldots, \lambda_n$, die eine Diagonalmatrix D bilden, und einer $n \times n$ Orthogonalmatrix V, deren Spalten $v_k = Ve_k$ die Eigenvektoren von A^A sind, d.h. es gilt

$$A^T A v_k = \lambda_k v_k,\ 1 \leq k \leq n.$$

Für einen Moment werden wir $m = n$ und $\mathrm{Rang}(A) = n$ voraussetzen, weil man dann besser sehen kann, was passiert. Wir nehmen die Bildvektoren $w_k = Av_k$ und prüfen sie auf Orthonormalität:

$$w_j^T w_k = v_j^T A^T A v_k = v_j^T \lambda_j v_k = \lambda_j v_j^T v_k = \lambda_j \delta_{jk},\ 1 \leq j, k \leq n.$$

Sie sind orthogonal, aber nicht orthonormal. Im Falle $\text{Rang}(A) = n$ kann keiner der Vektoren w_k verschwinden, und wir bekommen wegen der oben folgenden Gleichung $\|w_k\|_2^2 = \lambda_k$ die Positivität der λ_k. Setzen wir $\mu_k := \sqrt{\lambda_k}$, so sind die Vektoren $u_k := w_k/\mu_k$ orthonormal, und wir können sie als Spalten in eine $n \times n$–Orthogonalmatrix U zusammenfassen. Für diese gilt dann

$$Ue_j = u_j = w_j/\mu_j,\ 1 \leq j \leq n.$$

Jetzt sehen wir uns die Matrix U^TAV an, indem wir die Komponenten ausrechnen:

$$\begin{aligned} e_j^T U^T AV e_k &= w_j^T AV e_k/\mu_j \\ &= v_j^T A^T AV e_k/\mu_j \\ &= v_j^T \lambda_k v_k/\mu_j \\ &= \frac{\lambda_k}{\mu_j} v_j^T v_k \\ &= \frac{\lambda_k}{\mu_j}\delta_{jk},\ 1 \leq j,k \leq n \\ &= \mu_j \delta_{jk},\ 1 \leq j,k \leq n \end{aligned}$$

Packen wir die μ_j in eine Diagonalmatrix $\sqrt{D}$ mit $\sqrt{D}\sqrt{D} = D$, so folgt $U^TAV = \sqrt{D}$ und $A = U\sqrt{D}V^T$. In **zwei** geeignet gewählten Orthogonalbasen läst sich A also als reine Streckung schreiben.

Wie man dieses Argument variiert, wenn A nicht vollen Rang hat oder sogar nicht quadratisch ist, bleibt dem Leser überlassen. Es funktioniert, wenn man erst nur die w_j mit $\lambda_j > 0$ benutzt und dann die so gebildeten u_j zu einer Orthonormalbasis des $\mathbb{R}^m$ ergänzt.

Theorem 9.16 *Jede reelle Matrix $A \in \mathbb{R}^{m\times n}$ kann man in der Form $A = U\sqrt{D}V^T$ schreiben, wobei $u \in O(m)$ und $V \in O(n)$ Orthogonalmatrizen sind. Die Matrix $\sqrt{D} \in \mathbb{R}^{m\times n}$ ist außerhalb der Hauptdiagonalen gleich Null und enthält auf der Diagonalen die Wurzeln $\sqrt{\lambda_k}$ der Eigenwerte $\lambda_1, \ldots, \lambda_n \geq 0$ der symmetrischen und positiv semidefiniten Matrix A^TA. Diese heißen* **Singulärwerte** *von A.*

Wir haben im vorigen Kapitel schon angegeben, daß Orthogonalmatrizen die Determinante ± 1 haben, und daß die Determinante eines Matrizenprodukts gleich dem Produkt der Matrizen ist. Deshalb folgt aus dem vorigen Satz, daß die Determinante einer quadratischen reellen Matrix bis auf das Vorzeichen gleich dem Produkt der Singulärwerte ist. Daraus ergibt sich aber auch die allgemeine Volumeneigenschaft der Determinante. Denn aus $A = U\sqrt{D}V^T$ und der Invarianz von Längen unter orthogonalen Transformationen folgt, daß das Volumen der Menge aus 6.11 gleich dem Produkt der Beträge der Singulärwerte ist, denn das Transformationsverhalten von A wird in den durch U und V definierten "richtigen" orthonormalen Koordinatensystemen durch $\sqrt{D}$ bestimmt.

Theorem 9.17 *Der Absolutbetrag der Determinante einer reellen $n \times n$–Matrix A gibt an, um welchen Faktor sich Volumina von Mengen bei Transformation mit A vergrößern bzw. verkleinern.*

10 Reihen

10.1 Konvergenz von Reihen

Reihen sind summierte Folgen, also Summen über unendlich viele Zahlen. Genauer:

Definition 10.1 *Es sei $(a_n)_n$ eine reelle Zahlenfolge. Dann versteht man unter der* **Reihe**[1] $\sum_{k=0}^{\infty} a_k$ *die Folge* $(s_n)_n = \left(\sum_{k=0}^{n} a_k\right)_n$ *der* **Partialsummen**. *Wenn die Folge der Partialsummen gegen eine reelle Zahl α konvergiert, sagt man, die Reihe* **konvergiere** *gegen α oder habe den* **Summenwert** *oder Wert α und schreibt*

$$\sum_{k=0}^{\infty} a_k = \alpha.$$

Reihen, die nicht konvergieren, nennt man **divergent**.

Wie bei Folgen kann man natürlich den Anfangsindex anders wählen. Komplexe Reihen behandelt man durch Zerlegung in Real– und Imaginärteil wie zwei getrennte reelle Reihen. man mache sich klar, daß eine Reihe aus **zwei** Folgen besteht: aus der Folge der Partialsummen und der Folge der Reihenglieder. Diese beiden Folgen sind unbedingt auseinanderzuhalten.

Hier sind einige Beispiele:

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{1}{n} &\quad \text{divergent, “harmonische” Reihe} \\
\sum_{n=1}^{\infty} \frac{1}{n^2} &= \frac{\pi^2}{6} \\
\sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} &= \frac{\pi}{4} \\
\sum_{n=0}^{\infty} q^n &= \frac{1}{1-q},\ -1 < q < 1,\ \text{“geometrische” Reihe} \\
\sum_{n=0}^{\infty} \frac{1}{n!} &= e \approx 2.71828
\end{aligned}$$

Es kommt in der Informatik gelegentlich vor, die Partialsummen der divergenten **harmonischen Reihe** $\sum_{n=1}^{\infty} \frac{1}{n}$ nach oben und unten abschätzen zu

[1] `http://de.wikipedia.org/wiki/Reihe_(Mathematik)`

müssen, und deshalb sehen wir uns eine Partialsumme bis zu einem Maximalindex der Form $K = 2^{k+1} - 1$ an. Es folgt, wenn wir die Summation in Gruppen ausführen, die Divergenz aus der Abschätzung

$$\begin{aligned}
\sum_{n=1}^{2^{k+1}-1} \frac{1}{n} &= \sum_{j=0}^{k} \sum_{2^j \le n < 2^{j+1}} \frac{1}{n} \\
&\ge \sum_{j=0}^{k} \sum_{2^j \le n < 2^{j+1}} \frac{1}{2^{j+1}-1} \\
&= \sum_{j=0}^{k} \frac{2^j}{2^{j+1}-1} \\
&\ge \sum_{j=0}^{k} \frac{2^j}{2^{j+1}} \\
&= \sum_{j=0}^{k} \frac{1}{2} \\
&= \frac{k+1}{2}.
\end{aligned}$$

Man kann aber auch nach oben abschätzen und bekommt

$$\begin{aligned}
\sum_{n=1}^{2^{k+1}-1} \frac{1}{n} &= \sum_{j=0}^{k} \sum_{2^j \le n < 2^{j+1}} \frac{1}{n} \\
&\le \sum_{j=0}^{k} \sum_{2^j \le n < 2^{j+1}} \frac{1}{2^j} \\
&= \sum_{j=0}^{k} \frac{2^j}{2^j} \\
&= \sum_{j=0}^{k} 1 \\
&= k+1.
\end{aligned}$$

Insgesamt folgt also

$$\frac{k+1}{2} \le \sum_{n=1}^{2^{k+1}-1} \frac{1}{n} \le k+1.$$

Setzt man $N := 2^{k+1} - 1$, so folgt $k + 1 = \log_2(N + 1)$ und

$$\frac{\log_2(N+1)}{2} \le \sum_{n=1}^{N} \frac{1}{n} \le \log_2(N+1).$$

Diese Einschließung ist zwar nicht für alle N bewiesen, beschreibt aber die Asymptotik der Partialsummen ziemlich genau als

$$\sum_{n=1}^{N} \frac{1}{n} = \Theta(\log_2(N+1)).$$

In der Informatik entstehen gewisse logarithmische Ausdrücke auf genau diese Weise.

Auch die **geometrische Reihe** $\sum_{n=0}^{\infty} q^n$ kann man durch Partialsummen behandeln, denn es folgt durch Induktion leicht

$$\sum_{n=0}^{N} q^n = \frac{1-q^{N+1}}{1-q},\ q \neq 1.$$

Die rechtsstehende Folge konvergiert für $N \to \infty$ gegen $1/(1-q)$, wenn $|q| < 1$ gilt, weil die geometrische Folge $(q^N)_N$ nach der Argumentation auf Seite 224 dann eine Nullfolge ist. Für alle anderen q divergiert sie, und für $q > 1$ haben die Partialsummen das geometrische Wachstumsverhalten

$$\sum_{n=0}^{N} q^n = \frac{q^{N+1}-1}{q-1} = \Theta(q^N).$$

Alle anderen Beispiele behandeln wir nicht direkt, sondern durch Anwendung der Ergebnisse des nächsten Abschnitts.

Aber es sollte noch darauf hingewiesen werden, daß die Informatik nicht ohne Reihen auskommt. Wir werden im letzten Kapitel die Zerlegung von periodischen analogen Signalen in ihre Grundfrequenzen zu behandeln haben. Darauf haben wir schon im Kapitel 5 vergewiesen, und in (5.18) auf Seite 173 steht schon eine Reihe, allerdings eine über Funktionen, und sie zerlegt ein Signal in seine Grundfrequenzen.

10.2 Konvergenzsätze für Reihen

Auch für die Konvergenzuntersuchung von Reihen gibt es einen Werkzeugkasten. Wie bei Folgen beginnen wir mit Ergebnissen, die schon Konvergenz vorausetzen.

Theorem 10.2 *1. Ist eine Reihe konvergent, so bilden die Reihenglieder eine Nullfolge. Für die Praxis ist die Umkehrung wichtiger: wenn die Reihenglieder keine Nullfolge bilden, kann die Reihe nicht konvergieren.*

2. *Im Raum $\mathbb{R}^{\mathbb{N}}$ bilden die konvergenten Reihen, als Folgen der Reihenglieder gesehen (nicht als Folge der Partialsummen!) einen linearen Unterraum des Unterraums der Nullfolgen.*

3. *Die Abbildung, die einer konvergenten Reihe ihren reellen Grenzwert zuordnet, ist auf diesem Unterraum linear. Also sind Linearkombinationen von konvergenten Reihen wieder konvergent, und der Grenzwert einer Linearkombination ergibt sich durch die Linearkombination der Grenzwerte der gegebenen konvergenten Reihen.*

Das ist einfach zu beweisen, deshalb wird kein Platz verschwendet. □

Bei reellen Folgen waren die monotonen und beschränkten nach Satz 8.8 automatisch konvergent. Dem entsprechen die reellen Reihen mit nichtnegativen Gliedern:

Theorem 10.3 *Eine reelle Reihe mit nichtnegativen Gliedern ist genau dann konvergent, wenn ihre Partialsummen nach oben beschränkt sind.*

Eine reelle **alternierende Reihe***, d.h. eine mit abwechselnden Vorzeichen der Glieder, ist konvergent, wenn die Absolutbeträge der Glieder eines Endstücks eine monotone Nullfolge sind.*

Zum Beweis des zweiten Teils schreiben wir das Endstück einer alternierenden Reihe ohne Einschränkung der Allgemeinheit als Reihe $\sum_{k=0}^{\infty}(-1)^k a_k$ mit $a_k \geq a_{k+1} \geq 0$. Für die Partialsummen $s_n = \sum_{k=0}^{n}(-1)^k a_k$ gilt dann $s_{2n} - s_{2n-2} = a_{2n} - a_{2n-1} \leq 0$ und $s_{2n+1} - s_{2n-1} = -a_{2n+1} + a_{2n} \geq 0$. Die Folgen $(s_{2n})_n$ und $(s_{2n+1})_n$ sind also monoton fallend bzw. steigend. Sie sind wegen $s_{2n} - s_{2n-1} = a_{2n} \geq 0$ auch beschränkt, weil $s_{2n} \geq s_{2n-1}$ gilt. Also sind beide Folgen konvergent, und sie haben wegen $s_{2n} - s_{2n-1} = a_{2n} \to 0$ denselben Limes. □

Der obige Satz beweist die Konvergenz der Leibniz'schen Reihe $\displaystyle\sum_{n=0}^{\infty}(-1)^n \frac{1}{2n+1}$, aber er liefert leider nicht ihren Summenwert $\pi/4$. Dazu braucht man wesentlich mehr Maschinerie. Wie langsam die leibnizreihe konvergiert, sieht man in Abbildung 6. Problematisch sind Reihen, deren Glieder chaotische Vorzeichen haben. Das kann hier nicht genau untersucht werden. Es gibt unangenehme Beispiele, die zwar konvergieren, die nach **unendlichen** Umsortierungsprozessen aber verschiedene Summenwerte haben, weil die Folgen der Partialsummen sich durch die Umsortierung wesentlich ändern. Aber es gibt wenigstens teilweise Abhilfe, nämlich dann, wenn die Reihe der Absolutbeträge konvergiert.

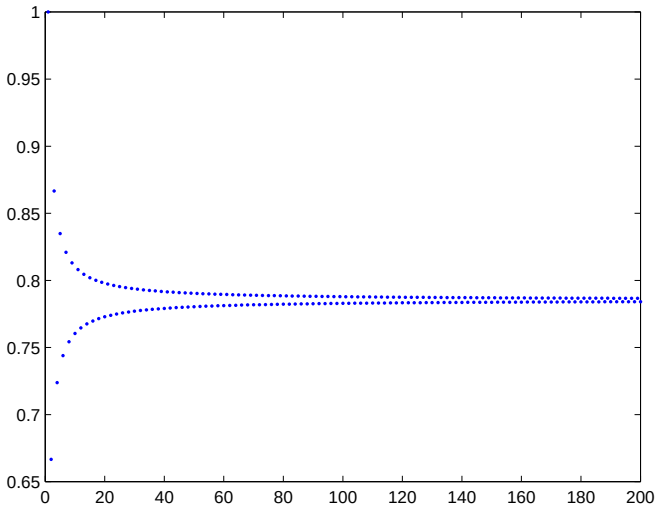


Abbildung 6: Partialsummen der Leibnizreihe

Definition 10.4 *Eine Reihe mit Gliedern a_n heißt* **absolut konvergent**, *wenn die Reihe mit den Gliedern $|a_n|$ konvergiert.*

Theorem 10.5 *Ist eine Reihe absolut konvergent, so ist sie konvergent. Wenn man sie beliebig umsortiert, bekommt man immer den gleichen Summenwert.*

Beweis des ersten Teils im reellen Fall: Wir definieren die Partialsummen

$$s_k := \sum_{n=0}^{k} a_n,\ s_k^+ := \sum_{\substack{n=0 \\ a_n \geq 0}}^{k} a_n,\ s_k^- := \sum_{\substack{n=0 \\ a_n < 0}}^{k} a_n,\ s_k = s_k^+ + s_k^-.$$

Nach Satz 10.3 und der Voraussetzung der absoluten Konvergenz sind die Folgen $(s_k^+)_k$ und $(s_k^-)_k$ beide monoton und beschränkt, also konvergent. Dann ist nach Satz 8.6 auch die Summe dieser Folgen konvergent, und nach Definition 10.1 ist die gegebene Reihe konvergent. Im komplexen Fall argumentiert man wie üblich mit Real– und Imaginärteil.

Der zweite Teil der Behauptung ist schwieriger und wird hier nicht bewiesen, weil er für die Praxis unwichtig ist. □

Jetzt kommen noch zwei sehr nützliche Hilfsmittel zum Beweis der Konvergenz von Reihen:

Theorem 10.6 *Es sei $\sum_{n=0}^{\infty} a_n$ eine Reihe.*

1. *Gilt die Abschätzung*

$$|a_n| \leq b_n \text{ für alle } n \in \mathbb{N},\ n \geq n_0$$

*für eine konvergente Reihe $\sum_{n=0}^{\infty} b_n$, so ist $\sum_{n=0}^{\infty} a_n$ absolut konvergent (***Majorantenkriterium***).*

2. *Gibt es eine reelle Zahl $q \in (0,1)$ mit*

$$|a_{n+1}| \leq q|a_n| \text{ für alle } n \in \mathbb{N},\ n \geq n_0,$$

*so ist so ist $\sum_{n=0}^{\infty} a_n$ absolut konvergent (***Quotientenkriterium***).*

Beweis Der erste Teil folgt direkt aus

$$\sum_{n \geq n_0}^{N} |a_n| \leq \sum_{n \geq n_0}^{N} b_n < \infty$$

und die Folge dieser Partialsummen ist beschränkt. Im zweiten Teil benutzt man den ersten, indem man als Majorante die geometrische Reihe verwendet. Induktiv folgt ja

$$|a_{n_0+k}| \leq q|a_{n_0+k-1}| \leq q^2|a_{n_0+k-2}| \leq \ldots \leq q^k|a_{n_0}| =: b_{n_0+k}, \quad \text{für alle } k \geq 0$$

und dann auch

$$\sum_{n \geq n_0}^{\infty} |a_n| \leq |a_{n_0}| \sum_{k=0}^{\infty} q^k \leq \frac{|a_{n_0}|}{1-q}$$

erst für Partialsummen und dann für die ganzen Reihen. □

Mit dem Majorantenkriterium kann man leicht die Reihe $\sum_{n=1}^{\infty} n^{-2}$ als konvergent nachweisen, weil man für das Endstück

$$\frac{1}{n^2} \leq \frac{1}{n(n-1)} = \frac{1}{n-1} - \frac{1}{n}, \quad \text{für alle } n \geq 2$$

hat und die Majorante umsortieren kann zu der Folge

$$s_N := \sum_{n=2}^{N} \left(\frac{1}{n-1} - \frac{1}{n} \right) = 1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \ldots + \frac{1}{N-1} - \frac{1}{N} = 1 - \frac{1}{N}$$

mit Limes 1. Aber wieder bleibt der Summenwert der eigentlichen Reihe im Dunkeln.

Die **Exponentialreihe** $\sum_{n=1}^{\infty} 1/n!$ ist ziemlich leicht mit dem Majorantenkriterium oder dem Quotientenkriterium als konvergent nachzuweisen. Aber wir wollen etwas mehr...

10.3 Potenzreihen

Eine wichtige Art von Reihen definiert Funktionen eines reellen oder komplexen Arguments x durch eine **Potenzreihe**

$$x \mapsto f(x) := \sum_{n=0}^{\infty} a_n x^n.$$

Das Supremum aller reellen Zahl R, so daß die Reihe für alle $x \in \mathbb{R}$ oder $\mathbb{C}$ mit $|x| < R$ konvergiert, heißt **Konvergenzradius**.

Wenn man die absolute Konvergenz von Potenzreihen beweisen will, bekommt man die schwach mit N monoton steigenden Partialsummen

$$s_N := \sum_{n=0}^{N} |a_n||x|^n.$$

Man kann das als eine gewichtete geometrische Reihe sehen, und wenn man weiß, das die Koeffizienten nicht allzu stark ansteigen, z.B. sich wie

$$|a_n| \leq K^n \tag{10.7}$$

mit einem $K > 0$ verhalten, so folgt

$$\begin{aligned} s_N &\leq \sum_{n=0}^{N} K^n |x|^n \\ &\leq \frac{1}{1 - K|x|} \end{aligned}$$

sofern man

$$|x| < \frac{1}{K}$$

hat. Also ist der Konvergenzradius in so einem Fall mindestens $1/K$, denn die Partialsummen sind monoton und nach oben beschränkt. Dieses Argument funktioniert auch dann, wenn man (10.7) nur für ein Endstück hat. Erwartungsgemäß wird der Konvergenzradius kleiner, wenn die Koeffizienten stärker gegen Unendlich gehen. Aus dieser Vorüberlegung wird auch klar, daß Potenzreihen einen riesigen Baukasten für Funktionen liefern, denn man kann ja allerhand Folgen $(a_n)_n$ finden, die der Einschränkung (10.7) für ein Endstück genügen. In der Tat sind viele spezielle Funktionen[1] in Potenzreihen "entwickelbar".

[1] `http://de.wikipedia.org/wiki/Spezielle_Funktionen`

Hier sind ein paar Beispiele, bei denen wir zunächst nur auf die Reihen, nicht auf die links stehenden klassischen Funktionen schauen sollten:

$$\begin{aligned} \exp(x) &:= \sum_{n=0}^{\infty} \frac{x^n}{n!} \\ cos(x) &:= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} \\ sin(x) &:= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \\ log(1+x) &:= \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} \end{aligned} \tag{10.8}$$

Wir haben hier statt der üblichen Bezeichnungen cos, sin und log die Schreibweisen cos, sin und log benutzt, weil wir die Funktionen durch die Reihen definiert haben und es keineswegs klar ist, ob wir es wirklich mit dem Sinus oder dem Logarithmus zu tun haben. Der Logarithmus ist normalerweise definiert als Umkehrfunktion der Exponentialfunktion[1], und diese wiederum hat mehrere Definitionen. Wir können hier von der obigen Reihendefinition der Exponentialfunktion ausgehen und leiten die typischen weiteren Eigenschaften dann her. Deshalb haben wir oben exp statt exp geschrieben, aber ob die obige Funktion log mit der Umkehrfunktion log der Exponentialfunktion übereinstimmt, ist nicht bewiesen.

Die geometrische Definition[2] der trigonometrischen Funktionen als Verhältnisse der Längen von Katheten zur Hypotenuse in rechtwinkligen Dreiecken sollte aus der Schule bekannt sein. Ferner ist wegen der Definition von π als Verhältnis von Kreisumfang zu Kreisdurchmesser klar, daß man Winkel im Bogenmaß durch reelle Zahlen φ beschreiben kann, Also wird die klassische Definition von Sinus un Cosinus schon auf der Schule so erweitert, daß sin und cos als Funktionen eines Winkels φ im Bogenmaß umgeschrieben werden können. Dadurch wird aus einer geometrischen Definition als Längenverhältnis eine Definition als reelle Funktion. Ob die obige Definition dieser entspricht, ist hier noch nicht klar, weil wir erstens die Länge eines Kreisbogens noch nicht berechnen können und zweitens noch nachrechnen müssen, daß dann diese Reihen genau die richtigen sind. Und das ist keineswegs klar, denn nach der üblichen Definition sind Sinus und Cosinus periodische Funktionen, aber die obigen Reihen sehen alles andere als periodisch aus. Oder ist es etwa

[1] `http://de.wikipedia.org/wiki/Exponentialfunktion`

[2] `http://de.wikipedia.org/wiki/Sinus`

klar, daß

$$0 = \sum_{n=0}^{\infty} (-1)^n \frac{\pi^{2n+1}}{(2n+1)!} = sin(\pi)$$

gilt? Offenbar scheint π sich als Lösung einer Gleichung mit unendlich vielen polynomialen Termen schreiben zu lassen, aber ist auch bewiesen (nach **Lindemann**[1]) daß es nicht mit endlich vielen polynomialen Termen geht.

Die ersten drei Fälle sind leicht mit dem Quotientenkriterium als konvergent für jedes x nachzuweisen. Für die Funktion exp haben wir

$$\frac{x^{n+1}}{(n+1)!} \leq q\frac{|x^n|}{n!}$$

zu zeigen. Wir machen erst eine informelle Zwischenrechnung und bekommen für alle $x \neq 0$ die Abschätzung

$$x\frac{x^n}{|x^n|} \leq (n+1)q.$$

Wenn wir die Konvergenz für alle $|x| < R$ und mit $q = 1/2$ beweisen wollen, sehen wir daran, daß man ein N mit $(N+1) \geq 2R$ nehmen sollte. Dann folgt

$$\begin{aligned} \frac{x^{n+1}}{(n+1)!} &= x\frac{x^n}{(n+1)!} \\ &\leq \frac{R}{n+1}\frac{|x^n|}{n!} \\ &\leq \frac{1}{2}\frac{|x^n|}{n!} \end{aligned}$$

für alle $n \geq N$ und wir haben die absolute Konvergenz der Reihe bewiesen, weil das Quotientenkriterium für $q = 1/2$ erfüllt ist. Dieses Argument funktioniert für jedes R, aber die Konvergenzanalyse erfordert ein mit wachsendem R ebenfalls wachsendes N, d.h. die Konvergenz "setzt später ein". Für große $|x|$ ist dieser Effekt in Abbildung 7 klar zu sehen. Betrachtet man den absoluten Fehler (siehe Abbildung 8), so wird klar, daß die Exponentialreihe für große negative Argumente unbrauchbar ist. Weil die Zwischenergebnisse in die Größenordnung von 10^{20} gehen, das Ergebnis e^{-50} aber nahe bei Null ist, tritt eine gewaltige **Auslöschung** ein, und das Endergebnis liegt bei etwa 10^5 statt bei 0. Weil man auf Standardrechnern etwa 15 korrekte Dezimalstellen hat, müssen etwa 20-15=5 falsche Stellen vorliegen.

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Lindemann.html`

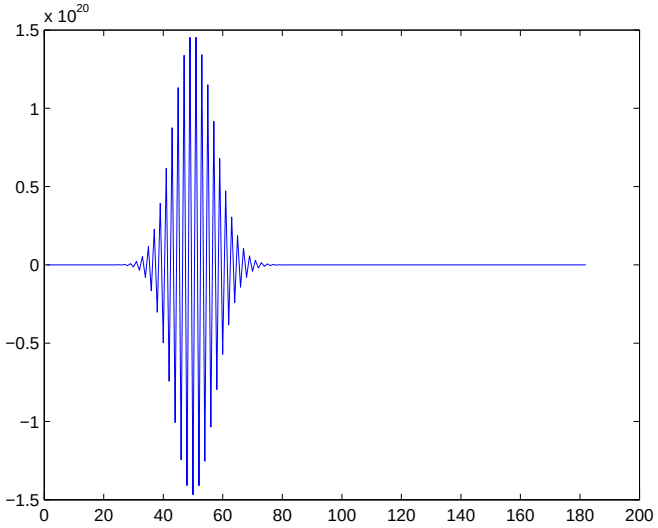


Abbildung 7: Partialsummen der Exponentialreihe für $x = -50$.

Aufgabe: Man schreibe ein kleines Programm, das $\exp(-10)$ und $\exp(1)$ über die Potenzreihe näherungsweise ausrechnet. Was ist zu beobachten?

Die Konvergenzanalyse der Reihen von *sin* und *cos* ist damit im Prinzip auch schon erledigt, denn diese Reihen sind Teilreihen der Exponentialfunktion, wenn man zu Beträgen übergeht. Die obige Reihe für $log(1+x)$ erweist sich bei einer entsprechenden Argumentation als konvergent für alle $|x| < 1$, und mehr kann man nicht erwarten, weil der Wert $x = -1$ nicht erlaubt sein kann, wenn es sich wirklich um den Logarithmus handelt.

Zumindestens mit den Reihen auf den rechten Seiten von (10.8) kann man jetzt arbeiten, aber es nicht klar, wieso die Reihen die links stehenden Funktionen darstellen. Wir machen einen kleinen Schritt in diese Richtung mit

Theorem 10.9 *Die Reihen aus* (10.8) *haben die Eigenschaften*

$$\exp(x+y) = \exp(x) \cdot \exp(y) \text{ } \textit{für alle } x, y \in \mathbb{C} \tag{10.10}$$

und

$$\exp(iz) = cos(z) + i \cdot sin(z) \text{ } \textit{für alle } z \in \mathbb{C}.$$

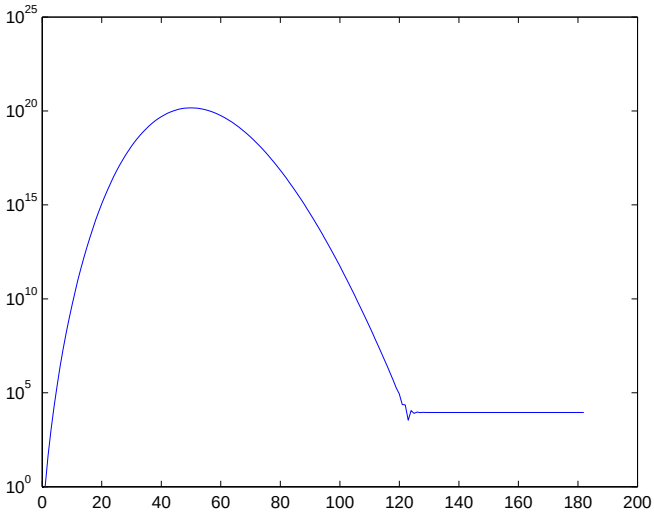


Abbildung 8: Absoluter Fehler der Partialsummen der Exponentialreihe für $x = -50$.

Der **Beweis** des ersten Teils benutzt die in der Vorlesung "Diskrete Mathematik" hoffentlich bewiesene **binomische Formel**

$$(x+y)^n = \sum_{j=0}^{n} \binom{n}{j} x^j y^{n-j}$$

und summiert die Terme

$$\begin{aligned} \frac{(x+y)^n}{n!} &= \sum_{j=0}^{n} \frac{\binom{n}{j}}{n!} x^j y^{n-j} \\ &= \sum_{j=0}^{n} \frac{x^j}{j!} \frac{y^{n-j}}{(n-j)!} \\ &= \sum_{\substack{j \geq 0 \\ k \geq 0 \\ j+k=n}} \frac{x^j}{j!} \frac{y^k}{k!} \end{aligned}$$

auf geeignete Weise, nämlich

$$\begin{aligned}\sum_{n=0}^{2N}\frac{(x+y)^n}{n!} &= \sum_{n=0}^{2N}\sum_{\substack{j\geq 0\\ k\geq 0\\ j+k=n}}\frac{x^j}{j!}\frac{y^k}{k!}\\ &= \left(\sum_{j=0}^{N}\frac{x^j}{j!}\right)\cdot\left(\sum_{k=0}^{N}\frac{y^k}{k!}\right)\\ &\quad+\sum_{n>N+1}^{2N}\sum_{\substack{j>N\\ k>N\\ j+k=n}}\frac{x^j}{j!}\frac{y^k}{k!}.\end{aligned}$$

Wir müssen hier vorsichtig über die Partialsummen argumentieren, weil wir keinen Satz über Produkte von Reihen zu Verfügung haben. Weil wir aber den entsprechenden Satz für Folgen benutzen können, ist nur noch zu zeigen, daß die zuletzt stehende Doppelsumme in Abhängigkeit von N eine Nullfolge ist. Dazu benutzen wir $|x|<R,\ |y|<R$ und bekommen

$$\begin{aligned}\left|\sum_{n>N+1}^{2N}\sum_{\substack{j>N\\ k>N\\ j+k=n}}\frac{x^j}{j!}\frac{y^k}{k!}\right| &\leq \sum_{n>N+1}^{2N}\sum_{\substack{j>N\\ k>N\\ j+k=n}}\frac{|x|^j}{j!}\frac{|y|^k}{k!}\\ &\leq \sum_{n=N+1}^{2N}\sum_{\substack{j>N\\ k>N\\ j+k=n}}R^n\frac{1}{(N!)^2}\\ &\leq \frac{R^{2N}}{(N!)^2}\sum_{n=N+1}^{2N}\sum_{\substack{j>N\\ k>N\\ j+k=n}}1\\ &\leq N^2\frac{R^{2N}}{(N!)^2}.\end{aligned}$$

Mit Satz 8.15 folgt die Behauptung des ersten Teils.

Für den zweiten sehen wir uns die Partialsummen an und erhalten

$$\begin{aligned}
\exp(iz) &\approx \sum_{n=0}^{N} \frac{(iz)^n}{n!} \\
&= \sum_{\substack{n=0 \\ n \text{ gerade}}}^{N} \frac{i^n z^n}{n!} + \sum_{\substack{n=0 \\ n \text{ ungerade}}}^{N} \frac{i^n z^n}{n!} \\
&= \sum_{\substack{m=0 \\ 2m \le N}}^{N} \frac{i^{2m} z^{2m}}{(2m)!} + \sum_{\substack{m=0 \\ 2m+1 \le N}}^{N} \frac{i^{2m+1} z^{2m+1}}{(2m+1)!} \\
&= \sum_{\substack{m=0 \\ 2m \le N}}^{N} \frac{(-1)^m z^{2m}}{(2m)!} + i \cdot \sum_{\substack{m=0 \\ 2m+1 \le N}}^{N} \frac{(-1)^m z^{2m+1}}{(2m+1)!} \\
&\approx cos(z) + i \cdot sin(z)
\end{aligned}$$

mit Gleichheit im Limes.

□

Mit dem Ergebnis dieses Satzes kann man allerhand anstellen, aber aus Platzgründen wird nicht alles bewiesen. Klar ist

$$\begin{aligned}
\exp(0) &= 1 \\
\exp(x) &> 1, \qquad x \in I\!R,\ x > 0 \\
\exp(2x) &= \exp(x)^2 \ge 0 \quad x \in I\!R
\end{aligned}$$

und man kann aus der Funktionalgleichung (10.10) der Exponentialfunktion ablesen, daß sie überall Null sein müßte, wenn sie an einer einzigen Stelle Null wäre. Also ist die Exponentialfunktion bei reellem Argument überall positiv. Sie ist auch streng monoton, denn wenn man $x < y$ hat, folgt

$$\exp(y) = \exp(x) \cdot \underbrace{\exp(y-x)}_{>1} > \exp(x).$$

Also ist die Exponentialfunktion auf ihrem Bildbereich umkehrbar. Der Bildbereich ist ganz $(0, \infty)$, aber das beweisen wir später. Die Umkehrfunktion heißt (natürlicher) **Logarithmus**[1] und wird als log oder manchmal auch ln geschrieben. Ob log mit der oben definierten Funktion *log* übereinstimmt,

[1] `http://de.wikipedia.org/wiki/Logarithmus`

ist noch offen. Aber aus (10.10) folgt dann auch die Funktionalgleichung des Logarithmus:

$$\log(x \cdot y) = \log(x) + \log(y) \text{ für alle } x, y > 0 \tag{10.11}$$

denn es folgt

$$\begin{aligned} \exp(\log(x) + \log(y)) &= \exp(\log(x)) \cdot \exp(\log(y)) \\ &= x \cdot y, \end{aligned}$$

und nach Anwendung des Logarithmus ergibt sich die Behauptung. Die Funktionalgleichung (10.11) des Logarithmus bildet die Grundlage der **Logarithmentafel** und der **Rechenschieber**. Man kann zwei positive Zahlen multiplizieren, indem man ihre Logarithmen addiert und auf das Ergebnis die Exponentialfunktion anwendet.

10.4 Darstellungen reeller Zahlen durch Reihen

Wir schieben hier noch einen Nachtrag zur Darstellung reeller Zahlen ein. Das ist für das mathematische Verständnis wichtig, aber in der Praxis irrelevant.

Im Abschnitt 3.6 auf Seite 94 haben wir reelle Zahlen als Äquivalenzklassen beschränkter Mengen von rationalen Zahlen eingeführt, aber aus dem Abschnitt 3.5.5 auf Seite 84 über **Gleitkommazahlen** sollte auch klar sein, daß man reelle Zahlen durch infinite b–adische Darstellungen beschreiben kann, die wir hier im Kapitel über Reihen genau wie im Abschnitt 3.6, aber anders als z.B. [4] indizieren:

$$\begin{aligned} x &= \pm \sum_{j=-m}^{\infty} b_{-j} b^{-j} \\ \text{in Ziffern} &= \pm b_m b_{m-1} \cdots b_0 \,.\, b_{-1} b_{-2} \cdots b_{-n} b_{-n-1} \cdots \end{aligned} \tag{10.12}$$

mit **Ziffern** $b_j \in \{0, 1, \ldots, b-1\}$ und einer natürlichen Zahl als **Basis** $b > 1$.

Theorem 10.13 *Die infinite b–adische Darstellung* (10.12) *einer reellen Zahl x ist eine konvergente Reihe.*

Beweis: Wir sehen uns die positiv genommenen Reihenreste für große n an:

$$s_n := \sum_{j=n}^{\infty} b_{-j} b^{-j}$$

und erhalten die Behauptung aus

$$\begin{aligned} s_n &= \sum_{j=n}^{\infty} b_{-j} b^{-j} \\ &\leq \sum_{j=n}^{\infty} (b-1) b^{-j} \\ &= (b-1) \sum_{j=n}^{\infty} b^{-j} \\ &= (b-1) b^{-n} \sum_{j=0}^{\infty} \left(\frac{1}{b}\right)^j \\ &= (b-1) b^{-n} \frac{1}{1-\frac{1}{b}} \\ &= b^{1-n}, \end{aligned}$$

weil b^{1-n} gegen Null konvergiert für $n \to \infty$. □

Man sieht an der obigen Argumentation aber auch, daß die Reihendarstellung nicht eindeutig ist, denn z.B. gilt im Dezimalsystem

$$1.0000 \cdots = 0.9999 \cdots$$

Unabhängig von dieser Uneindeutigkeit kann man rekursiv zu jeder reellen Zahl x eine b–adische Darstellung (10.12) angeben, die x darstellt. Sehen wir uns das für positive x an. Wir können durch Abtrennen des ganzzahligen Anteils ohne Einschränkung annehmen, daß $0 \leq x < 1$ gilt und wir beginnen die Rekursion mit $x_1 := x \in [0, 1)$. Dann gilt $bx_1 \in [0, b)$ und die erste Nachkomma–Dezimalziffer b_{-1} von $x_1 = x$ sei genommen als der ganzzahlige Anteil in $\{0, 1, \ldots, b-1\}$ von $b \cdot x_1$. Wir berechnen $x_2 := bx_1 - b_{-1}$ und bekommen $x_2 \in [0, 1)$ nach unserer Wahl von b_{-1}.

Jetzt sollte klar sein, wie es weitergeht: zu $x_n \in [0, 1)$ bestimmt man b_{-n} als ganzzahligen Teil von bx_n und geht zu $x_{n+1} := bx_n - b_{-n} \in [0, 1)$ über. Man zeigt dann leicht per Induktion, dass die Gleichung

$$x_{n+1} = b^n x - \sum_{j=1}^{n} b_{-j} b^{n-j} \text{ für alle } n \geq 0$$

gilt, die man auch als Zerlegung von $b^n x$ in einen b–adisch dargestellten ganzzahligen Teil plus Rest $x_{n+1} \in [0, 1)$ deuten kann. Jetzt stellt man die Gleichung um zu

$$x = b^{-n} x_{n+1} + \sum_{j=1}^{n} b_{-j} b^{-j}$$

und sieht sofort die Konvergenz der Partialsummen von (10.12) gegen x, weil $b^{-n}x_{n+1} \in [0, b^{-n})$ für $n \to \infty$ gegen Null strebt.

Theorem 10.14 *Das obige Verfahren konstruiert zu jeder reellen Zahl x genau eine b–adische Entwicklung, die x als Reihe darstellt.* □

11 Standardfunktionen und Stetigkeit

In diesem Kapitel betrachten wir Abbildungen $f : \mathbb{R} \to \mathbb{R}$, die man auch **reelle** oder **reellwertige Funktionen**[1] nennt. Der Definitionsbereich wird oft auf ein beschränktes oder unbeschränktes **Intervall** I eingeschränkt (vgl. Definition 3.22 auf Seite 96). Reelle Funktionen mit wilden Teilmengen von $\mathbb{R}$ als Definitionsbereich werden wir hier nicht behandeln. Unsere Definitionsbereiche sind immer Intervalle und deshalb immer konvex.

11.1 Stetige Funktionen

11.1.1 Funktionen und Graphen

Definition 11.1 *Sei $f : \mathbb{R} \to \mathbb{R}$ eine Funktion. Man bezeichnet f als*

gerade	*wenn*	$f(x) = f(-x)$	*für alle* $x \in \mathbb{R}$
ungerade	*wenn*	$f(x) = -f(-x)$	*für alle* $x \in \mathbb{R}$
(schwach) monoton wachsend	*wenn*	$f(x) \leq f(y)$	*für alle* $x \leq y \in \mathbb{R}$
streng monoton wachsend	*wenn*	$f(x) < f(y)$	*für alle* $x < y \in \mathbb{R}$
periodisch mit Periode T	*wenn*	$f(x) = f(x+T)$	*für alle* $x \in \mathbb{R}$.

Man sehe sich die Beispiele in der Wikipedia oder hier an:

$f(x)$	$=$	x^2	ist gerade
$f(x)$	$=$	x^3	ist ungerade
$f(x)$	$=$	$\max(x, 1)$	ist schwach monoton wachsend
$f(x)$	$=$	$\exp(x)$	ist streng monoton wachsend
$f(x)$	$=$	$\cos(x)$	hat die Periode 2π.

Funktionen $f : \mathbb{R} \supseteq I \to \mathbb{R}$ werden oft durch **Funktionsgraphen**[2] veranschaulicht. Dazu markiert man im $\mathbb{R}^2$ alle Punkte der Form $(x, f(x))$ mit beliebigen $x \in \mathbb{R}$. Die Menge

$$\{(x, f(x)) \in \mathbb{R}^2 \ : \ x \in I\}$$

ist der **Graph** von f. Dieser Graphenbegriff ist von dem in der Graphentheorie[3] verschieden.

Aufgabe: Für jede der obigen Eigenschaften gebe man ein weiteres Beispiel durch eine Funktion mit ihrem Graphen an und deute die Eigenschaften geometrisch.

In MATLAB kann man einfache Funktionsgraphen durch Befehlsfolgen wie

[1] `http://de.wikipedia.org/wiki/Funktion_(Mathematik)`
[2] `http://de.wikipedia.org/wiki/Funktionsgraph`
[3] `http://de.wikipedia.org/wiki/Graph_(Graphentheorie)`

```
x=-1:0.01;1;
plot(x,2*x.^3-x)
```

zeichnen lassen. Hier haben wir durch den ersten Befehl einen Zeilenvektor aus äquidistanten Punkte im Intervall $[-1,1]$ mit Abstand 0.01 definiert, aber die resultierenden 201 Zahlen wegen des Semikolons nicht ausgegeben. Der nachfolgende Befehl zeichnet dann die Funktion $f(x) = 2x^3 - x$ auf diesem Intervall. Durch Setzen eines Punktes bei `x.^3` erzwingt man die komponentenweise Anwendung der dritten Potenz auf den Zeilenvektor `x`.

Aufgabe: Man benutze MATLAB, um je eine auf das Intervall $[-1,1]$ eingeschränkte Funktion mit den obigen Eigenschaften zu zeichnen.

Theorem 11.2 *1. Ist ein Intervall $I \subseteq \mathbb{R}$ fest gegeben, so bildet die Menge $I^{\mathbb{R}}$ der reellen Funktionen mit Definitionsbereich I einen reellen Vektorraum unter den Operationen*

$$\begin{aligned}(f+g)(x) &:= f(x)+g(x) \text{ für alle } x \in I,\ f,g \in I^{\mathbb{R}}\\ (\alpha f)(x) &:= \alpha \cdot f(x) \text{ für alle } x \in I,\ f \in I^{\mathbb{R}}.\end{aligned}$$

2. Im Falle $I = \mathbb{R}$ bilden die geraden bzw. ungeraden Funktionen sowie die Funktionen mit fester Periode einen Untervektorraum von $\mathbb{R}^{\mathbb{R}}$.

3. Zu zwei reellen Funktionen f und g mit gemeinsamem Definitionsbereich I kann man das **punktweise Funktionenprodukt** *$f \cdot g$ durch $(f \cdot g)(x) := f(x) \cdot g(x)$ für alle $x \in I$ definieren. Man bekommt eine lineare Abbildung $I^{\mathbb{R}} \times I^{\mathbb{R}} \to I^{\mathbb{R}}$ mit $(f,g) \mapsto f \cdot g$.*

4. Gilt zusätzlich $g(x) \neq 0$ für alle $x \in I$, so ist f/g analog definiert.

Aufgabe: Man beweise Teile dieses Satzes.

Frage: Was ist im Raum $I^{\mathbb{R}}$ das Analogon zu den Einheitsvektoren des $\mathbb{R}^n$?

Frage: Warum ist der Raum $I^{\mathbb{R}}$ unendlichdimensional?

11.1.2 Stetigkeit reeller Funktionen

Angehende Informatiker wollen reelle Funktionen konkret ausrechnen. Aber wie soll das gehen, wenn man ein Argument $x \in \mathbb{R}$, für das man $f(x)$ ausrechnen will, nicht exakt im Rechner darstellen kann? Ist z.B. $\tilde{x}$ eine Gleitkommazahl, die nahe bei x liegt, so sollte das Ergebnis $f(\tilde{x})$, wenn man es exakt ausrechnen könnte, nahe bei $f(x)$ sein. Will man $f(x)$ **beliebig** genau durch Werte der Form $f(\tilde{x})$ ausrechnen, so wird man verlangen müssen,

daß auch $\tilde{x}$ "entsprechend nahe" bei x liegt. Und wenn wir $\tilde{x}$ immer näher an x heranschieben, sollte $f(\tilde{x})$ immer näher an $f(x)$ herankommen. Macht man das mit Gliedern x_n einer gegen x konvergenten Folge, so sollte $f(x_n)$ gegen $f(x)$ konvergieren. Also:

Definition 11.3 *Eine reelle Funktion f ist (folgen–)* **stetig** *in einem Punkte x ihres Definitionsintervalls I, wenn sie jede in I liegende und gegen x konvergente Folge $(x_n)_n$ auf eine gegen $f(x)$ konvergente Folge abbildet, d.h. wenn aus $x = \lim_{n\to\infty} x_n$ immer $f(x) = \lim_{n\to\infty} f(x_n)$ folgt. Ist eine Funktion in allen Punkten ihres Definitionsbereichs I stetig, so wird sie* **auf I stetig** *genannt.*

Man mache sich klar, daß diese Definition impliziert, daß der Wert $f(x)$ als Limes aller Folgen $(f(x_n))_n$ immer derselbe ist, gleichgültig welche gegen x konvergente Folge $(x_n)_n$ man nimmt. Im Sinne des Abschnitts 8.5 auf Seite 241 bekommt die Stetigkeit von f in x auch die Form

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

Es ist jetzt kein großes Problem, mit Hilfe des Satzes 8.6 auf Seite 221 folgendes zu beweisen:

Theorem 11.4 *1. Die in einem festen Punkte oder in einem festen Intervall stetigen Funktionen mit gemeinsamem Definitionsbereich bilden einen reellen Vektorraum.*

2. *Sind f und g zwei Funktionen, die in einem gemeinsamen Punkt x ihrer Definitionsbereiche I_f und I_g stetig sind, so ist das punktweise genommene Produkt $f \cdot g$ in x stetig als Funktion auf $I_f \cap I_g$.*

3. *Ist unter den obigen Voraussetzungen zusätzlich $g(x)$ nicht Null, so ist der punktweise genommene Quotient f/g in x stetig als Funktion auf $I_f \cap I_g \cap \{y \,:\, g(y) \neq 0\}$.*

4. *Sind f und g stetige Funktionen auf I, so ist auch $f \cdot g$ stetig auf I.*

5. *Sind f und g stetige Funktionen auf I und gilt $g(x) \neq 0$ für alle $x \in I$, so ist auch f/g stetig auf I.*

Beweis: Es seien f und g zwei Funktionen mit Definitionsbereich I, die in einem Punkte $x \in I$ stetig seien. Ferner seien α und β reelle Zahlen, und wir betrachten eine Folge $(x_n)_n$ in I mit $\lim_{n\to\infty} x_n = x$. Wegen der Stetigkeit von f und g bekommen wir $f(x) = \lim_{n\to\infty} f(x_n)$ und $g(x) = \lim_{n\to\infty} g(x_n)$. Auf

diese beiden Folgen wenden wir Satz 8.6 auf Seite 221 an und folgern, dass die Linearkombinations–Folge $(\alpha f + \beta g)(x_n)$ gegen $(\alpha f + \beta g)(x)$ konvergiert.

Diese Beweisidee funktioniert auch für alle anderen Aussagen: man bildet einfach Produkte und Quotienten der Folgen $(f(x_n))_n$ und $(g(x_n))_n$. □

Der obige Satz besagt, dass man durch Bilden von Linearkombinationen, Produkten und Quotienten von stetigen Funktionen wieder stetige Funktionen bekommt. Aber der Baukasten für stetige Funktionen ist noch größer:

Theorem 11.5 *Es seien f und g reellwertige Funktionen mit Definitionsbereichen I_f und $I_g \supseteq f(I_f)$, so dass die Hintereinanderanwendung $g \circ f$ Sinn macht. Ferner sei f stetig in $x \in I_f$ und g stetig in $f(x) \in f(I_f) \subseteq I_g$. Dann ist die Hintereinanderanwendung $g \circ f$ in x stetig.*

Beweis: Eine beliebige gegen x konvergente Folge $(x_n)_n$ wird mit f wegen der Stetigkeit von f auf eine gegen $f(x)$ konvergente Folge $(f(x_n))_n$ abgebildet. Weil g in $f(x)$ als stetig vorausgesetzt wurde, kann man diesen Schluss wiederholen und bekommt, dass die Folge $(g(f(x_n)))_n$ gegen $g(f(x))$ konvergiert. □

Wir wenden jetzt unseren Baukasten an, um Standardfunktionen als stetig nachzuweisen. Sehen wir uns erst einmal die **Monome** $x \mapsto x^n$ für $n \geq 0$ an. Im Falle $n = 0$ haben wir die konstante Funktion $x \mapsto 1$, und diese ist überall stetig, weil die Bilder aller Folgen konstant, also konvergent sind. Die **Identität** $x \mapsto x = x^1$ ist auch auf $\mathbb{R}$ stetig, weil sie konvergente Folgen auf sich selber abbildet. Weil aber nach Satz 11.4 alle Produkte stetiger Funktionen wieder stetig sind, müssen alle Monome stetig auf $\mathbb{R}$ sein. Linearkombinationen stetiger Funktionen sind nach Satz 11.4 stetig, also sind auch alle Polynome stetig.

Quotienten von zwei Polynomen heißen **rationale Funktionen**. Diese sind nach Satz 11.4 überall dort stetig, wo der Nenner nicht Null wird. Punkte, in denen Funktionen oder Polynome den Wert Null annehmen, heißen **Nullstellen**. Die Nullstellen des Nennerpolynoms einer rationalen Funktion heißen **Pole** der rationalen Funktion. In diesem Sinne hat die rationale Funktion $\frac{x-3}{x^2-1}$ die Nullstelle 3 und die Pole $+1$ und -1.

Wir wissen aus dem vorigen Kapitel, daß viele Funktionen um den Nullpunkt herum als Potenzreihen zu schreiben sind. Dann sind sie auch stetig:

Theorem 11.6 *Potenzreihen sind innerhalb ihres Konvergenzradius stetig.*

Zum Beweis sehen wir uns eine für $|x| < R$ absolut konvergente Potenzreihe

$$x \mapsto f(x) := \sum_{n=0}^{\infty} a_n x^n.$$

an und verwenden für Partialsummen den dritten binomischen Lehrsatz, der weiter unten als (11.14) steht. Wir nehmen $|x|,\ |y| \leq r < R$ an und bekommen

$$\begin{aligned}
& \left| \sum_{n=0}^{N} a_n (x^n - y^n) \right| \\
\leq & \sum_{n=0}^{N} |a_n| |x^n - y^n| \\
\leq & \sum_{n=0}^{N} |a_n| |(x-y) \sum_{j=1}^{n} x^{n-j} y^{j-1}| \\
\leq & \sum_{n=0}^{N} |a_n| |x-y| \sum_{j=1}^{n} |x|^{n-j} |y|^{j-1} \\
\leq & |x-y| \sum_{n=0}^{N} |a_n| \sum_{j=1}^{n} r^{n-j} r^{j-1} \\
\leq & |x-y| \sum_{n=0}^{N} |a_n| n r^{n-1}.
\end{aligned}$$

Hier braucht man eine kleine Zusatzüberlegung. Die Multiplikation der Folge $(r^{n-1})_n$ mit n verschlechtert zwar die Lage, aber die Folge bleibt "geometrisch" für ein Endstück:

Lemma 11.7 *Für jedes s mit $r < s$ gibt es ein $N \in \mathbb{N}$, so daß*

$$n r^n \leq s^n$$

für alle $n \geq N$ gilt.

Beweis: Weil $\log(n) = o(n)$ für $n \to \infty$ gilt, gibt es ein N, so daß

$$\frac{\log(n)}{n} \leq \log(s/r)$$

für alle $n \geq N$ gilt. Aber dann folgt

$$\begin{aligned}
\log(n) &\leq n \log(s/r) \\
n &\leq (s/r)^n \\
n r^n &\leq s^n. \square
\end{aligned}$$

Dieses Lemma wird auch hilfreich sein, wenn wir Differenzierbarkeit von Potenzreihen beweisen wollen.

Um den Beweis des Satzes 11.6 abzuschließen, nehmen wir ein s mit $r < s < R$ und können dann auf

$$\sum_{n=0}^{\infty} |a_n| n r^n \leq C + \sum_{n=N}^{\infty} |a_n| s^n < \infty$$

schließen. Weil Multiplikation mit r oder $1/r$ nichts Wesentliches verändert, bekommen wir also eine Konstante K, so daß

$$|f(x) - f(y)| \leq |x - y| \cdot K$$

für alle $|x|,\ |y| < r$ gilt. Wenn wir statt y eine gegen x konvergente Folge $(x_n)_n$ einsetzen, folgt

$$|f(x) - f(x_n)| \leq |x - x_n| \cdot K \to 0$$

für $n \to \infty$, und das liefert Stetigkeit von f in x. □

11.1.3 Zwischenwertsatz

Am liebsten würden wir jetzt auch die Stetigkeit der Umkehrfunktion einer streng monotonen Funktion zeigen. Aber wir wissen noch nicht, ob die Bildmenge einer stetigen Funktion wieder ein Intervall ist. Dazu brauchen wir einige Vorbereitungen.

Theorem 11.8 *(***Nullstellensatz***)*
Es sei f eine stetige Funktion auf einem Intervall I und es gebe zwei Zahlen $a, b \in I$ mit

$$f(a) < 0 < f(b).$$

Dann gibt es ein c zwischen a und b mit $f(c) = 0$. Man nennt c eine **Nullstelle** *von f.* □

Beweis: Wir machen das mit einer **Intervallschachtelung**[1], wie schon beim Beweis des Satzes 8.11 von Bolzano–Weierstraß. Zu Beginn setzen wir $a_0 := a,\ b_0 := b$ und haben $f(a_0) \leq 0 < f(b_0)$. Wir konstruieren Folgen $(a_k)_k$ und $(b_k)_k$ mit $f(a_k) \leq 0 < f(b_k)$ und $|b_k - a_k| \leq 2^{-k}|b_0 - a_0|$. Das ist für $k = 0$ schon klar, und den Übergang von k nach $k + 1$ machen wir so, daß

[1] `http://de.wikipedia.org/wiki/Intervallschachtelung`

wir im Punkte $c_k := (a_k + b_k)/2$ mitten zwischen a_k und b_k testen, ob der Funktionswert dort positiv oder negativ ist. Wir setzen

$$\begin{array}{lll} a_{k+1} = c_k,\ b_{k+1} = b_k & \text{falls} & f(c_k) \leq 0 \\ a_{k+1} = a_k,\ b_{k+1} = c_k & \text{falls} & f(c_k) > 0. \end{array}$$

Das liefert den Induktionsschritt und damit dann zwei schwach monotone und beschränkte, deshalb konvergente Folgen mit gleichem Limes c. Aus der Stetigkeit folgt

$$\lim_{k\to\infty} f(a_k) = f(c) = \lim_{k\to\infty} f(b_k)$$

und da für alle k die Ungleichungen $f(a_k) \leq 0 < f(b_k)$ gelten, kann der Limes $f(c)$ nur Null sein. □.

Der obige Satz ist natürlich auch in der Praxis anwendbar, wenn man Nullstellen von Funktionen ausrechnen will. Es gibt effizientere Verfahren, aber die brauchen mehr Voraussetzungen als Stetigkeit. Was dem Wert Null recht ist, ist allen anderen Werten billig:

Theorem 11.9 *(***Zwischenwertsatz***)* [1]
Es sei f eine stetige Funktion auf einem Intervall I. Dann hat jede reelle Zahl z, die echt zwischen zwei Funktionswerten $f(a)$ und $f(b)$ mit $a, b \in I$ liegt, ein Urbild echt zwischen a und b, d.h. es gibt ein c zwischen a und b mit $z = f(c)$.

Beweis: Man wendet den Nullstellensatz auf die stetige Funktion $g(x) := f(x) - z$ an. □

Korollar 11.10 *Der Bildbereich einer stetigen reellen Funktion auf einem Intervall ist konvex.* □

Beweis: Das folgt sofort aus dem Zwischenwertsatz. Hat man nämlich zwei Werte $f(x)$ und $f(y)$ aus dem Bildbereich, so wird jeder Zwischenwert angenommen, d.h. im Falle $f(x) \leq f(y)$ liegt auch das Intervall $[f(x), f(y)]$ im Bildbereich. □

Theorem 11.11 *Eine stetige reelle Funktion auf einem abgeschlossenen und beschränkten Intervall nimmt dort Minimum und Maximum an. Insbesondere werden* **abgeschlossene** *Intervalle der Form $[a, b]$ wieder auf abgeschlossene Intervalle abgebildet.*

[1] `http://de.wikipedia.org/wiki/Zwischenwertsatz`

Beweis: Es sei f stetig auf einem abgeschlossenen und beschränkten Intervall $[a, b] \subset I\!R$. Wir benutzen den Beweisgang des Satzes 8.27 und setzen

$$s^- := \inf\{f(x) \ : \ x \in [a, b]\} \leq \sup\{f(x) \ : \ x \in [a, b]\} =: s^+.$$

Es gibt eine Folge $(x_n)_n$ in $[a, b]$ so dass die Folge $(f(x_n))_n$ gegen s^+ strebt oder beliebig groß wird für $n \to \infty$, je nachdem ob s^+ endlich ist oder nicht. Wir können nach dem Satz von Bolzano–Weierstraß eine gegen eine reelle Zahl x konvergente Teilfolge aussuchen und sie wieder $(x_n)_n$ nennen. Wegen der Abgeschlossenheit des Intervalls folgt dann $x \in [a, b]$ und wir bekommen $f(x) = \lim_{n\to\infty} f(x_n)$ wegen der Stetigkeit von f. Also nimmt f sein Supremum oder Maximum auf $[a, b]$ an. □

Der obige Satz ist zentral für alle praktischen **Optimierungsprobleme**, bei denen man eine Kosten– oder Nutzenfunktion auf einer ziemlich allgemeinen Menge von "zulässigen" Punkten maximiert oder minimiert.

11.2 Umkehrfunktionen

Wir wissen schon aus Definition 1.32 auf Seite 35, was eine Umkehrabbildung ist, und dass jede bijektive Abbildung eine Umkehrabbildung hat.

Theorem 11.12 *Ist eine reelle Funktion auf ihrem Definitionsbereich streng monoton, so hat sie auf ihrem Bildbereich eine Umkehrfunktion*[1]*, die ebenfalls monoton ist.*

Achtung: Es ist hier nicht gesagt, daß der Bildbereich wieder ein Intervall ist (dazu braucht man Stetigkeit), und deshalb kann es sein, daß die Umkehrfunktion auf einer ziemlich wilden Menge reeller Zahlen definiert ist.

Beweis: Es sei $f \ : \ I \to I\!R$ streng monoton. Dann ist f auch injektiv, denn aus $f(x) = f(y)$ kann weder $x < y$ noch $y < x$ folgen, sondern es muss $x = y$ gelten. Dann ist die Funktion, wenn man ihren Wertebereich auf die reale Bildmenge einschränkt, bijektiv und umkehrbar.

Wir müssen noch die Monotonie der Umkehrfunktion zeigen. Gilt fur zwei Werte $u = f(x)$ und $v = f(y)$ die Relation $u < v$, so ist $f^{-1}(u) < f^{-1}(v)$ zu zeigen. Das ist aber dasselbe wie $x < y$, und diese Aussage folgt wegen der Monotonie von f aus $u = f(x) < f(y) = v$, weil $x = y$ und $x > y$ nicht möglich sind. □

[1] `http://de.wikipedia.org/wiki/Umkehrfunktion`

Theorem 11.13 *Die Umkehrfunktion einer streng monotonen stetigen Funktion ist stetig.*

Beweis: Es sei $(y_n)_n$ mit $y_n = f(x_n)$ eine gegen $y = f(x)$ konvergente Folge. Zu zeigen ist $\lim_{n\to\infty} x_n = x$. Wir können wegen der Monotonie von f annehmen, daß $f(x_n)$ schwach monoton steigend gegen $f(x)$ konvergiert. Dann ist die Folge $(x_n)_n$ schwach monoton und nach oben durch x beschränkt. Sie hat einen Grenzwert $z \leq x$ mit $f(z) = \lim_{n\to\infty} f(x_n) = f(x)$ wegen der Stetigkeit von f. Aus der strengen Monotonie ergibt sich dann $z = x$, d.h. die Folge $(x_n)_n$ konvergiert gegen x. □

Wir wollen jetzt die **Monome** $x \mapsto x^n$ für $n \geq 0$ etwas genauer untersuchen. Wir wissen schon, daß die Monome stetig sind, und deshalb bilden sie abgeschlossene Intervalle in abgeschlossene Intervalle ab. Der Definitionsbereich der Monome ist also immer $\mathbb{R}$, und der Bildbereich von $x \mapsto x^n$ ist $\mathbb{R}$, wenn n ungerade ist, sonst $[0, \infty)$.

Jetzt sehen wir uns die Monotonieeigenschaften an. Im Falle $n = 0$ haben wir eine konstante Abbildung, und die ist schwach monoton steigend und fallend. Die Identität $x \mapsto x$ ist trivialerweise streng monoton, weil jede Ungleichung $x < y$ auf sich selbst abgebildet wird. Die quadratische Funktion $x \mapsto x^2$ ist, wie wir aus der Schule wissen, für positive x streng monoton steigend und für negative x streng monoton fallend. Wir werden also den Definitionsbereich der Monome auf $I := \mathbb{R}_{\geq 0} := [0, \infty)$ einschränken und dort strenge Monotonie aller Monome $x \mapsto x^n$ für alle $n \geq 1$ behaupten. Diese folgt aber sofort aus der per Induktion leicht beweisbaren Variante

$$x^n - y^n = (x - y)\sum_{j=1}^{n} x^{n-j}y^{j-1} \text{ für alle } x, y \in \mathbb{C},\ n \geq 0 \qquad (11.14)$$

der dritten binomischen Formel, denn im Falle $n \geq 1$ und $x > y \geq 0$ ist die rechte Seite immer positiv.

Aus dem vorigen Satz folgt dann die Injektivität der Monome $x \mapsto x^n$ mit $n \geq 1$ auf den nichtnegativen reellen Zahlen. Dann existiert die inverse Abbildung, hier auch **Umkehrfunktion** genannt, auf der Bildmenge von $x \mapsto x^n$. Diese bildet dann eine Zahl z der Form $z = x^n$ auf x ab, d.h. die Umkehrabbildung "zieht die n–te Wurzel". Bei unserer Einschränkung des Definitionsbereichs auf $[0, \infty)$ wissen wir aber, daß die Bildmenge aus allen nichtnegativen reellen Zahlen besteht. Also gilt

Theorem 11.15 *Zu jeder nichtnegativen reellen Zahl z und jedem $n \in \mathbb{N},\ n \geq 1$ gibt es genau eine reelle nichtnegative Zahl $\sqrt[n]{x}$ mit $(\sqrt[n]{x})^n = x$.* □

Frage: Welche Monotonieeigenschaften haben die Monome auf den negativen reellen Zahlen?

Frage: Welche Monome sind gerade Funktionen und welche sind ungerade Funktionen?

11.3 Standardfunktionen

Wir wissen jetzt, daß $\sqrt[2]{\sqrt{2}}$ definiert ist, aber $\sqrt[\sqrt{2}]{2}$ ist noch undefiniert. Man kann wie in der Schule für positive x und ganze Zahlen $n \neq 0,\ m > 0$ die Potenz $x^{\frac{n}{m}}$ als $x^{\frac{n}{m}} := (\sqrt[m]{x})^n$ definieren, aber es ist nicht klar, was x^z für beliebige reelle $x, z > 0$ sein soll. Man lernt in der Schule die Definition $x^z := \exp(z \log x)$. und deshalb gehen wir jetzt auf die Exponentialfunktion und ihre Umkehrung, den Logarithmus ein.

Die in (10.8) auf Seite 260 durch die Potenzreihe

$$\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!} \text{ für alle } x \in \mathbb{C}$$

definierte Funktion exp hat nach Satz 10.9 auf Seite 262 die Eigenschaft

$$\exp(x+y) = \exp(x) \cdot \exp(y) \text{ für alle } x, y \in \mathbb{C}. \tag{11.16}$$

Wir haben schon im vorigen Kapitel die strenge Monotonie der Exponentialfunktion auf ganz $\mathbb{R}$ nachgewiesen. Wenn wir die Stetigkeit zeigen wollen, können wir natürlich auf Satz 11.6 zurückgehen, aber hier ist nochmal eine kleine Wiederholung der Beweisidee. Wir nehmen $|x|, |y| \leq R$ an und bekommen mit (11.14) die Abschätzung

$$\begin{aligned}
|\exp(x) - \exp(y)| &\leq |x-y| \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{j=1}^{n} |x|^{n-j} |y|^{j-1} \\
&\leq |x-y| \sum_{n=1}^{\infty} \frac{1}{n!} n R^{n-1} \\
&= |x-y| \sum_{n=1}^{\infty} \frac{1}{(n-1)!} R^{n-1} \\
&= |x-y| \exp(R),
\end{aligned}$$

woraus sich die Stetigkeit sofort ablesen läßt, wenn man statt y Elemente einer gegen x konvergenten Folge einsetzt. Während unsere Monotonieüberlegung auf $\mathbb{C}$ keinen Sinn macht, weil dort gar keine Ordnung existiert, macht der obige Stetigkeitsbeweis auch auf $\mathbb{C}$ Sinn.

Wir definieren $e := \exp(1) \approx 2.171828$ und erhalten $\exp(n) = e^n$ für alle $n \geq 0$ aus (11.16) per Induktion. Wegen $e > 2$ strebt $(e^n)_n$ gegen Unendlich für $n \to \infty$ und gegen Null für $n \to -\infty$, ohne die Null jemals zu erreichen. Deshalb finden wir den Bildbereich $\exp(\mathbb{R}) = \mathbb{R}_{>0}$ und bekommen die monotone Umkehrfunktion $\log(x)$ auf $\mathbb{R}_{>0}$ mit Bild $\mathbb{R}$. Es gilt also

$$\exp(\log(x)) = x \text{ für alle } x > 0 \text{ und } \log(\exp(y)) = y \text{ für alle } y \in \mathbb{R}.$$

Wendet man diese Umkehrfunktion auf die Funktionalgleichung (11.16) an, so folgt (10.11), wie wir schon gesehen haben. Das Potenzieren einer positiven Zahl x ist dann wegen der Funktionalgleichung (11.16) durch

$$x^n = (\exp(\log(x))^n = \exp(n \cdot \log(x))$$

möglich, und die m–te Wurzel $\sqrt[m]{x}$ ist

$$\sqrt[m]{x} = \exp(\frac{1}{m}\log(x))$$

wegen

$$\left(\exp(\frac{1}{m}\log(x))\right)^m = \exp(\log(x)) = x.$$

Zusammen folgt

$$x^{\frac{n}{m}} = \exp\left(\frac{n}{m}\log(x)\right) \text{ für alle } n, m \in \mathbb{Z},\ m > 0,\ x \in \mathbb{R}_{>0}.$$

Jetzt halten wir ein positives $a \in \mathbb{R}$ fest und untersuchen die Funktion $f_a(x) := \exp(x \cdot \log(a))$ auf $\mathbb{R}$. Es gilt

$$\begin{array}{rcll} f_a(1) & = & \exp(1 \cdot \log(a)) & = & a \\ f_a(n) & = & \exp(n \cdot \log(a)) & = & a^n \\ f_a\left(\frac{n}{m}\right) & = & \exp\left(\frac{n}{m} \cdot \log(a)\right) & = & a^{\frac{n}{m}} \end{array}$$

und deshalb ist die Definition

$$a^x := \exp(x \cdot \log(a)) \text{ für alle } a, x \in \mathbb{R},\ a > 0$$

eine sinnvolle und stetige Erweiterung der üblichen Potenzfunktion. Die Funktionalgleichung (11.16) liefert dann alle Regeln der Potenzrechnung.

Eine Gleichung $y = a^x$ wird üblicherweise "zur Basis" a logarithmiert durch die zunächst unbegründete Forderung

$$x =: \log_a(y) \Leftrightarrow y = a^x \text{ für alle } a, y \in \mathbb{R},\ a, y > 0.$$

Aus $y = a^x$ folgt aber

$$\begin{aligned} y &= \exp(x \cdot \log(a)) \\ \log(y) &= x \cdot \log(a) \\ x &= \frac{\log(y)}{\log(a)} =: \log_a(y) \end{aligned}$$

und man bekommt die allgemeine Logarithmusdefinition

$$\log_a(y) := \frac{\log(y)}{\log(a)} \text{ für alle } a, y > 0.$$

Wir gehen jetzt zu den trigonometrischen Funktionen sin und cos über, können aber ihre geometrische Bedeutung noch nicht erklären. Wir nehmen die Reihendefinitionen für verwandte Funktionen *sin* und *cos* aus (10.8) auf Seite 260 durch die Potenzreihen

$$\begin{aligned} sin(x) &:= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \text{ für alle } x \in \mathbb{C} \\ cos(x) &:= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \text{ für alle } x \in \mathbb{C} \end{aligned}$$

und verwenden neben (11.16) die Funktionalgleichung

$$\exp(iz) = cos(z) + i \cdot sin(z) \text{ für alle } z \in \mathbb{C}. \tag{11.17}$$

die wir schon in Theorem 10.9 auf Seite 262 hergeleitet haben. Um die Funktionen *sin* und *cos* reell zu erhalten, wählt man $z := x \in \mathbb{R}$ und bekommt

$$\exp(ix) = cos(x) + i \cdot sin(x) \text{ für alle } x \in \mathbb{R}$$

als Zerlegung der komplexen Funktion $\exp(ix)$ in Real–und Imaginärteil. Wir wollen auf die Additionstheoreme hinaus und untersuchen

$$\begin{aligned} \exp(i(x+y)) &= cos(x+y) + i \cdot sin(x+y) \\ &= \exp(ix)\exp(iy) \\ &= (cos(x) + i \cdot sin(x)) \cdot (cos(y) + i \cdot sin(y)) \\ &= (cos(x)cos(y) - sin(x)sin(y)) \\ &\quad + i \cdot (sin(x)cos(y) + cos(x)sin(y)) \end{aligned}$$

woraus durch Koeffizientenvergleich die bekannten Gleichungen

$$\begin{aligned} cos(x+y) &= cos(x)cos(y) - sin(x)sin(y) \\ sin(x+y) &= sin(x)cos(y) + cos(x)sin(y) \end{aligned}$$

für alle $x,\ y \in \mathbb{R}$ folgen.

Wir sehen aus den Reihen direkt, dass *cos* gerade und *sin* ungerade ist. Damit konjugieren wir $\exp(ix)$ und bekommen

$$\begin{aligned}\overline{\exp(ix)} &= \overline{cos(x) + i \cdot sin(x)} \\ &= cos(x) - i \cdot sin(x) \\ &= cos(-x) + i \cdot sin(-x) \\ &= \exp(-ix),\end{aligned}$$

was zu

$$\begin{aligned}1 &= \exp(0) \\ &= \exp(ix)\exp(-ix) \\ &= \exp(ix)\overline{\exp(ix)} \\ &= |\exp(ix)|^2 \\ &= cos^2(x) + sin^2(x)\end{aligned}$$

führt. Die letzte Gleichung hätte man auch schon aus den Additionstheoremen folgern können, aber wir ziehen hier lieber die Konsequenz, daß die komplexen Zahlen $\exp(ix)$ alle den Betrag 1 haben und somit auf dem Rand des Einheitskreises der komplexen Zahlenebene liegen!

Die komplexe Zahl $\exp(ix)$ hat also im cartesischen Koordinatensystem der komplexen Zahlenebene die Koordinaten $(cos(x), sin(x))$. In Polarkoordinaten habe $\exp(ix)$ einen geometrischen Winkel $\alpha(x)$. Mit den auf herkömmliche Weise definierten trigonometrischen Funktionen sin und cos gilt also $cos(x) = \cos(\alpha(x))$ und $sin(x) = \sin(\alpha(x))$ wegen der Winkeldefinition aus (5.12) auf Seite 167. Wenn man die geometrische Definition von $\sin(\alpha)$ und $\cos(\alpha)$ als Verhältnisse von Gegenkathete bzw. Ankathete zur Hypothenuse bezüglich eines Winkels α in einem rechtwinkligen Dreieck nimmt, folgt dasselbe, aber nicht $\alpha(x) = x$. Erst wenn wir zeigen, daß die Länge des Einheitskreisbogens von $(1, 0)$ nach $(cos(x), sin(x))$ gleich x ist, stimmt die geometrische Definition der trigonometrischen Funktionen mit der Definition durch Reihen überein.

11.4 Stetigkeit von Abbildungen

Es sollte aus dem Abschnitt 8.3 auf Seite 234 klar sein, daß man Stetigkeit auch für wesentlich allgemeinere Abbildungen definieren kann, denn man braucht nur einen Limesbegriff im Urbild– und Bildraum. Man sehe sich dazu auch den hinteren Teil der entsprechenden Wikipedia-Seite [1] an.

[1] `http://de.wikipedia.org/wiki/Stetigkeit`

Definition 11.18 *Es seien M bzw. N metrische Räume. Eine Abbildung $f : M \to N$ ist (folgen–)* **stetig** *in einem Punkte $x \in M$, wenn sie jede in M liegende und gegen x konvergente Folge $(x_n)_n$ auf eine gegen $f(x)$ konvergente Folge abbildet, d.h. wenn aus $x = \lim_{n\to\infty} x_n$ immer $f(x) = \lim_{n\to\infty} f(x_n)$ folgt. Ist f in allen Punkten ihres Definitionsbereichs M stetig, so wird sie* **auf** M **stetig** *genannt.*

Wir kennen viele interessante metrische Räume, nämlich die normierten Vektorräume. Alle Abbildungen zwischen Vektorräumen, die man rechnerisch sauber auswerten will, sind also auf Stetigkeit zu untersuchen. Ein simples Beispiel für eine stetige Abbildung ist $x \mapsto \|x\|$ auf einem normierten Vektorraum, und ebenso ist der Abstand $x \mapsto d(x, y)$ zu einem festen Punkt y eines metrischen Raumes mit Distanzfunktion d immer eine stetige Abbildung.
Frage: Warum?

Der praktisch wichtigste Fall besteht aber aus **linearen** Abbildungen $T : U \to V$ zwischen Vektorräumen U und V. Ist so eine Abbildung stetig, wenn wir Urbild– und Bildraum normieren? Zunächst einmal folgt

Lemma 11.19 *Ist eine lineare Abbildung zwischen normierten Vektorräumen stetig im Nullpunkt, so ist sie überall stetig. Das heißt auch: Wenn eine solche lineare Abbildung Nullfolgen in Nullfolgen abbildet, ist sie überall stetig.*

Beweis: Wir zeigen Stetigkeit in einem beliebigen $u \in U$ für eine lineare Abbildung $T : U \to V$, die in 0 stetig ist. Dazu sei $(u_n)_n$ eine gegen u konvergente Folge in U. Dann ist $(u_n - u)_n$ eine Nullfolge, und wegen Stetigkeit von T in Null und $T(0) = 0$ folgt, daß $(T(u_n - u))_n$ eine Nullfolge sein muß. Es gilt aber $T(u_n - u) = T(u_n) - T(u)$, und deshalb konvergiert $(T(u_n))_n$ gegen $T(u)$, was die Stetigkeit von T in u beweist. □

Sehen wir uns einen gutartigen und einen bösartigen Fall an. Der gutartige besteht aus einer linearen Funktion $f : \mathbb{R} \to \mathbb{R}$, die dann ja immer die Form $x \mapsto a \cdot x$ mit einer reellen Zahl a (einer 1×1–Matrix) haben muss. Man bekommt

$$|f(x)| \leq |a||x| \text{ für alle } x \in \mathbb{R} \tag{11.20}$$

und jede Nullfolge $(x_n)_n$ in $\mathbb{R}$ wird auf eine Nullfolge der Form $(a \cdot x_n)_n$ abgebildet. Diese linearen Abbildungen sind also stetig, aber das wußten wir schon.

Jetzt ein bösartiger Fall. Wir nehmen die Abbildung aus (4.24), die jedes Polynom p auf seine Ableitung p' abbildet. Diese Abbildung ist linear, unabhängig davon, wie der Polynomraum P^1_∞ normiert wird. Nehmen wir die

Norm

$$\|p\|_\infty := \max_{x\in[-1,1]} |p(x)| \text{ für alle } p \in P^1_\infty$$

und die Folge $(p_n)_n$ mit den skalierten Monomen $p_n(x) := x^n/n$. Das ist eine Nullfolge wegen

$$\|p_n\|_\infty = \max_{x\in[-1,1]} |p_n(x)| = \max_{x\in[-1,1]} |x^n|/n \le \frac{1}{n}$$

und wenn die Ableitungsabbildung bei Verwendung derselben Norm im Bildraum stetig wäre, müßte auch

$$\|p_n'\|_\infty = \max_{x\in[-1,1]} |p_n'(x)| = \max_{x\in[-1,1]} |x^{n-1}| = 1$$

eine Nullfolge sein, was nicht stimmt.

Das Differenzieren ist (unter Umständen)
eine unstetige lineare Abbildung!

Das hat sehr unangenehme Konsequenzen für das Wissenschaftliche Rechnen, denn man kann das Berechnen von Ableitungen nicht vermeiden, wenn man wichtige Probleme in Wissenschaft und Technik lösen will.

Lineare Abbildungen zwischen normierten Räumen sind also leider nicht immer stetig. Aber wenn wir unendlichdimensionale Räume ausschließen, kommen wir wieder zurück zu der gutartigen Situation (11.20), die man folgendermaßen verallgemeinern kann:

Definition 11.21 *Eine lineare Abbildung T zwischen normierten Vektorräumen U und V heißt* **beschränkt**[1]*, wenn es eine Konstante $K \in \mathbb{R}$ gibt mit*

$$\|T(u)\|_V \le K \cdot \|u\|_U \text{ für alle } u \in U. \tag{11.22}$$

Theorem 11.23

1. *Beschränkte lineare Abbildungen sind stetig.*
2. *Stetige lineare Abbildungen auf endlichdimensionalen Vektorräumen sind beschränkt.*
3. *Alle linearen Abbildungen auf* **endlichdimensionalen** *Vektorräumen sind beschränkt und damit stetig.*

[1] `http://de.wikipedia.org/wiki/Linearer_Operator%23Beschr%C3%A4nkte_lineare_Operatoren`

4. *Die beschränkten linearen Abbildungen zwischen zwei festen normierten Vektorräumen bilden einen normierten Vektorraum $BL(U,V)$, und die* **natürliche Norm** *oder* **Operatornorm**[1] *einer beschränkten linearen Abbildung $T : U \to V$ ist*

$$\|T\|_{U,V} := \sup_{u \in U \setminus \{0\}} \frac{\|T(u)\|_V}{\|u\|_U}. \tag{11.24}$$

5. *Damit gilt*

$$\|T(u)\|_V \leq \|T\|_{U,V} \|u\|_U \text{ für alle } u \in U. \tag{11.25}$$

6. *Definiert man zu einem weiteren normierten Vektorraum W mit Norm $\|.\|_W$ die zugeordnete Norm $\|.\|_{V,W}$ auf dem Raum $BL(V,W)$ der beschränkten linearen Abbildungen von V in W, so hat die zu $\|.\|_U$ und $\|.\|_W$ zugeordnete Norm auf dem Raum $BL(U,W)$ der beschränkten linearen Abbildungen von U in W die Eigenschaft*

$$\|S \circ T\|_{U,W} \leq \|T\|_{U,V} \cdot \|S\|_{V,W} \tag{11.26}$$

   für alle $S \in BL(V,W)$, $T \in BL(U,V)$.

7. *Die zugeordnete Norm $\|T\|_{U,V}$ ist das Minimum aller Konstanten K, die in* (11.22) *auftreten können.*

Beweisskizze: Die Gleichung (11.22) zeigt sofort, daß Nullfolgen auf Nullfolgen abgebildet werden, und das beweist den ersten Teil.

Ist eine lineare Abbildung $T : U \to V$ nicht beschränkt, so gibt es zu jeder Konstanten $K \in I\!N$ ein $u_K \in U$ mit $\|T(u_K)\|_V \geq K \cdot \|u_K\|_U$. Wenn wir die u_K renormieren zu $\|u_K\|_U = 1$, können wir nach dem Satz 8.22 von Bolzano–Weierstraß auf Seite 237 eine konvergente Teilfolge auswählen, die gegen ein $u \in U$ mit $\|u\|_U = 1$ konvergiert. Dann folgt aber $T(u) = \lim T(u_K)$ aus der Stetigkeit von T, und das widerspricht der Aussage $\|T(u_K)\|_V \geq K \cdot \|u_K\|_U$, weil die rechte Seite gegen Unendlich strebt und die linke beschränkt bleibt.

Zum Beweis des dritten Teils wählen wir eine Basis $u_1, \ldots, u_n$ im Urbildraum U und bilden dann einen beliebigen Vektor $u \in U$ ab gemäß

$$u \;=\; \sum_{j=1}^{n} \alpha_j u_j \;\mapsto\; T(u) \;=\; \sum_{j=1}^{n} \alpha_j T(u_j).$$

[1] `http://de.wikipedia.org/wiki/Operatornorm%23Operatornormen`

Die Normen $\|u\|_U$ und

$$\left\| \sum_{j=1}^{n} \alpha_j u_j \right\|_* := \sum_{j=1}^{n} |\alpha_j|$$

sind nach Satz 5.8 auf Seite 163 äquivalent. Es gibt also eine Konstante C_1 mit $\|u\|_* \leq C_1 \cdot \|u\|_U$ für alle $u \in U$. Dann folgt

$$\begin{aligned} \|T(u)\|_V &= \left\| \sum_{j=1}^{n} \alpha_j T(u_j) \right\|_V \\ &\leq \sum_{j=1}^{n} |\alpha_j| \, \|T(u_j)\|_V \\ &\leq \max_{1 \leq j \leq n} \|T(u_j)\|_V \sum_{j=1}^{n} |\alpha_j| \\ &= \|u\|_* \max_{1 \leq j \leq n} \|T(u_j)\|_V \\ &= C_1 \cdot C_2 \|u\|_U \text{ für alle } u \in U, \end{aligned}$$

wenn man

$$C_2 := \max_{1 \leq j \leq n} \|T(u_j)\|_V$$

definiert. Das zeigt die Beschränktheit von T.

Zum Beweis des vierten Teils schließen wir aus (11.22), daß das Supremum in (11.24) existiert, weil es kleiner oder gleich K sein muß. Die reelle Zahl $\|T\|_{U,V}$ ist also immer wohldefiniert, und einige hier nicht ausgeführte Überlegungen zeigen, daß man eine Norm hat und die beschränkten linearen Abbildungen damit einen normierten Vektorraum bilden. Die fünfte Aussage folgt dann sofort aus (11.24).

Zum Beweis der sechsten nimmt man das Supremum von

$$\|(S \circ T)(u)\|_W \leq \|S\|_{V,W} \|T(u)\|_V \leq \|S\|_{V,W} \|T\|_{U,V} \|u\|_U.$$

Gilt (11.22) mit einer Konstanten K, so folgt sofort auch $\|T(u)\|_V \leq K$, wenn man in (11.24) einsetzt. Und weil wir schon wissen, daß $\|T(u)\|_V$ die Rolle von K in (11.22) spielen kann, ist auch die letzte Aussage bewiesen. □.

Das Supremum in (11.24) ist natürlich nicht kleiner als jeder einzelne Term. Man hat also eine untere Abschätzung

$$\frac{\|T(u)\|_V}{\|u\|_U} \leq \|T\|_{U,V}$$

für jedes beliebige $u \neq 0$ aus U. Das werden wir später brauchen können.

Um den Umgang mit Suprema zu üben, beweisen wir noch die Dreiecksungleichung

$$\|S+T\|_{U,V} \leq \|S\|_{U,V} + \|T\|_{U,V} \tag{11.27}$$

für beliebige beschränkte lineare Abbildungen $S,\ T \in BL(U,V)$. Wir beginnen mit der normalen Dreiecksungleichung in V für ein beliebiges $u \in U \setminus \{0\}$ und wenden (11.25) zweimal an:

$$\begin{array}{rcl} \|(S+T)(u)\|_V & \leq & \|S(u)\|_V + \|T(u)\|_V \\ & \leq & \|S\|_{U,V}\|u\|_U + \|T\|_{U,V}\|u\|_U \\ & = & (\|S\|_{U,V} + \|T\|_{U,V})\,\|u\|_U. \end{array}$$

Wenn wir das durch $\|u\|_U$ dividieren, folgt

$$\frac{\|(S+T)(u)\|_V}{\|u\|_U} \leq \|S\|_{U,V} + \|T\|_{U,V}.$$

Also existiert das Supremum der linken Seite, und es folgt die Behauptung (11.27).

Die Definition von Operatornormen wenden wir jetzt auf Matrizen an. Weil eine Matrix $A = (a_{jk}) \in \mathbb{R}^{m\times n}$ über $x \mapsto A \cdot x$ eine lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^m$ darstellt, und weil nach dem vorigen Satz alle diese Abbildungen stetig und beschränkt sind, kann man zwei Normen $\|.\|_p$ auf dem $\mathbb{R}^n$ und $\|.\|_q$ auf dem $\mathbb{R}^m$ wählen und definiert

$$\|A\|_{p,q} := \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|A\cdot x\|_q}{\|x\|_p}$$

als natürliche oder zugeordnete Norm. Dabei wähle man $p, q \in [1,\infty]$ mit den Normdefinitionen aus (5.3) auf Seite 160. Diese Matrixnormen[1] hängen zwar von p und q, also von den festgelegten Vektornormen ab, aber sie sind im Sinne von Definition 5.6 auf Seite 162 äquivalent.

Aufgabe: Man beweise diese Aussage.

Die Wikipedia[2] nennt eine Matrixnorm der Form $\|.\|_{p,p}$ durch die Vektornorm $\|.\|_p$ **induziert**.

[1] `http://de.wikipedia.org/wiki/Matrixnorm%23Matrixnormen`

[2] `http://de.wikipedia.org/wiki/Matrixnorm%23Matrixnormen`

Wir üben jetzt das Rechnen mit Normen und bestimmen die Matrixnorm $\|A\|_{\infty,\infty}$, die sich ergibt, wenn man im Urbild– und Bildraum die Maximumsnorm verwendet. Nach dem vorigen Satz sollten wir die Abschätzung

$$\|A \cdot x\|_\infty \leq K \cdot \|x\|_\infty \text{ für alle } x \in I\!R^n$$

mit der kleinstmöglichen Konstanten K beweisen. Man bekommt, wenn man möglichst haarscharf in Richtung auf $\|x\|_\infty$ abschätzt, die Beziehung

$$\begin{aligned}
\|A \cdot x\|_\infty &= \max_{1\leq j\leq m} \left| \sum_{k=1}^{n} a_{jk} x_k \right| \\
&\leq \max_{1\leq j\leq m} \sum_{k=1}^{n} |a_{jk}||x_k| \\
&\leq \max_{1\leq k\leq n} |x_k| \max_{1\leq j\leq m} \sum_{k=1}^{n} |a_{jk}| \\
&= \|x\|_\infty \cdot \max_{1\leq j\leq m} \sum_{k=1}^{n} |a_{jk}| \text{ für alle } x \in I\!R^n.
\end{aligned}$$

Also folgt

$$\|A\|_{\infty,\infty} = \sup_{x\in I\!R^n\setminus\{0\}} \frac{\|A \cdot x\|_\infty}{\|x\|_\infty} \leq \max_{1\leq j\leq m} \sum_{k=1}^{n} |a_{jk}|.$$

Wir vermuten, daß es nicht besser geht, und dazu reicht es, ein x anzugeben, für das Gleichheit eintritt. Es gibt ein $i,\ 1 \leq i \leq m$ mit

$$\max_{1\leq j\leq m} \sum_{k=1}^{n} |a_{jk}| = \sum_{k=1}^{n} |a_{ik}|,$$

und wir nehmen dann $x_k := \text{sgn}\,(a_{ik}),\ 1 \leq k \leq n$. Dann gilt $a_{ik}x_k = |a_{ik}| \geq 0$ und es folgt die umgekehrte Ungleichung

$$\begin{aligned}
\|A \cdot x\|_\infty &= \max_{1\leq j\leq m} \left| \sum_{k=1}^{n} a_{jk} x_k \right| \\
&\geq \left| \sum_{k=1}^{n} a_{ik} x_k \right| \\
&= \sum_{k=1}^{n} |a_{ik} x_k| \\
&= 1 \cdot \sum_{k=1}^{n} |a_{ik}| \\
&= \|x\|_\infty \cdot \max_{1\leq j\leq m} \sum_{k=1}^{n} |a_{jk}|
\end{aligned}$$

für dieses spezielle x.

Aus naheliegenden Gründen nennt man

$$\|A\|_{\infty,\infty} := \max_{1\le j\le m} \sum_{k=1}^{n} |a_{jk}|$$

die **Zeilensummennorm** von A.

Aufgabe: Man beweise, daß die **Spaltensummennorm** von A die natürliche Norm der Form

$$\|A\|_{1,1} := \max_{1\le k\le n} \sum_{j=1}^{m} |a_{jk}|$$

ist.

Die wichtige Norm $\|A\|_{2,2}$ ist nicht ganz so einfach auszurechnen. Man bildet besser das Quadrat von (11.22) und versucht dann, die Größe $\|A \cdot x\|_2^2 = x^T A^T A x$ so gut wie möglich in Richtung auf $\|x\|^2$ abzuschätzen. Wählen wir als x einen Eigenvektor $\neq 0$ zu einem Eigenwert λ von $A^T A$, so folgt $A^T A x = \lambda x$ und $\|A \cdot x\|_2 = x^T A^T A x = \lambda x^T x = \lambda \|x\|_2^2$. Also gilt $\lambda \ge 0$ für jeden solchen Eigenwert, und wir bekommen

$$\|A \cdot x\|_2 \le \sqrt{\lambda}\|x\|_2$$

für den zugehörigen Eigenvektor x. Im schlimmsten Fall müssen wir also mindestens mit der Konstanten

$$\max\{\sqrt{\lambda} \ : \ \lambda \text{ ist Eigenwert von } A^T A\}$$

rechnen. Diese tritt in unserer Abschätzung auf, wenn wir den Eigenvektor zum größten Eigenwert von $A^T A$ einsetzen, es gibt also keine kleinere Konstante, die das Gewünschte leistet.

Das beweist aber noch nicht. daß wir **für alle** x mit dieser Konstanten auskommen. Dazu müssen wir uns an die Diagonalisierbarkeit symmetrischer Matrizen erinnern. Die Matrix $A^T A$ ist symmetrisch und erfüllt Satz (9.3) auf Seite 245. Es gibt also eine $n \times n$–Orthogonalmatrix V und eine Diagonalmatrix D mit Diagonalelementen $\lambda_1, \ldots, \lambda_n$ so dass $A^T A = V D V^T$ gilt. Dann folgt $\|x\|_2 = \|V^T x\|_2 =: \|y\|_2$ für $y := V^T x$ aus der Orthogonalität von V und $V^T = V^{-1}$ (siehe Satz 4.42 auf Seite 129 und 5.17 auf Seite 171) und

man erhält

$$\begin{aligned}
\|A \cdot x\|_2^2 &= x^T A^T A x = x^T V D V^T x = y^T D y \\
&= \sum_{j=1}^{n} \lambda_j y_j^2 \\
&\leq \left(\max_{1 \leq k \leq n} \lambda_k \right) \sum_{j=1}^{n} y_j^2 \\
&= \|y\|_2^2 \max_{1 \leq k \leq n} \lambda_k \\
&= \|x\|_2^2 \max_{1 \leq k \leq n} \lambda_k \\
\|A \cdot x\|_2 &\leq \|x\|_2 \sqrt{\max_{1 \leq k \leq n} \lambda_k}.
\end{aligned}$$

Also ist die **Spektralnorm**

$$\|A\|_{2,2} := \max\{\sqrt{\lambda} \ : \ \lambda \text{ ist Eigenwert von } A^T A\}$$

die natürliche Matrixnorm zur euklidischen Vektornorm, aber sie ist leider alles andere als leicht handzuhaben, denn sie erfordert eine Eigenwertberechnung oder eine Singulärwertzerlegung (siehe Satz 9.16 auf Seite 251).

Man kann eine etwas größere und sehr viel einfacher berechenbare Norm, nämlich die **Frobeniusnorm**

$$\|A\|_F := \sqrt{\sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk}^2}$$

nehmen, um immerhin noch die Abschätzung

$$\|A \cdot x\|_2 \leq \|A\|_F \cdot \|x\|_2 \text{ für alle } x \in I\!R^n$$

zu bekommen. Das beweist man mit der schon in Abschnitt 4.6.1 benutzten Zerlegung der Matrix A als Summe $\sum_{k=1}^{n} A e_k e_k^T$ der Spalten, und mit der

Cauchy–Schwarzschen Ungleichung aus

$$\begin{aligned}
\|A \cdot x\|_2 &= \left\| \sum_{k=1}^{n} A e_k e_k^T x \right\|_2 \\
&= \left\| \sum_{k=1}^{n} A e_k x_k \right\|_2 \\
&\leq \sum_{k=1}^{n} |x_k| \|A e_k\|_2 \\
&\leq \sqrt{\sum_{k=1}^{n} x_k^2} \sqrt{\sum_{k=1}^{n} \|A e_k\|_2^2} \\
&= \|x\|_2 \sqrt{\sum_{k=1}^{n} \sum_{j=1}^{m} a_{jk}^2} \\
&= \|x\|_2 \|A\|_F.
\end{aligned}$$

Wegen dieser Ersetzung der Spektralnorm durch die Frobeniusnorm verallgemeinert man (11.24) und (11.26) in geeigneter Weise:

Definition 11.28 *Eine Norm $\|.\|_M$ auf $BL(U,V)$ ist* **passend** *zu oder* **verträglich**[1] *mit $\|.\|_U$ und $\|.\|_V$, wenn für alle $T \in BL(U,V)$ die Abschätzung*

$$\|T(u)\|_V \leq \|A\|_M \cdot \|u\|_U \text{ für alle } u \in U$$

gilt. Die entsprechende Verallgemeinerung von (11.26) *wird* **Multiplikativität** *genannt.*

Beim praktischen Rechnen beschränkt man sich auf passende und multiplikative Normen. Wenn man zugeordnete Normen (d.h. Operatornormen) verwendet, können nach Satz 11.23 keine Probleme auftreten, und für die Frobeniusnorm gilt

Theorem 11.29 *Die Frobeniusnorm ist passend zur $\|.\|_2$–Norm und multiplikativ.*

Beweis: Wir müssen nur noch die Multiplikativität beweisen. Dazu nehmen wir zwei Matrizen $A = (a_{jk}) \in I\!R^{n \times m}$, $B = (b_{ij}) \in I\!R^{m \times \ell}$ her, bilden das Matrizenprodukt $C := (c_{ik}) = A * B \in I\!R^{n \times \ell}$ und die Quadrate der

[1] `http://de.wikipedia.org/wiki/Matrixnorm%23Matrixnormen`

Frobeniusnormen

$$\begin{aligned}
\|C\|_F^2 &= \sum_{i=1}^{n}\sum_{k=1}^{\ell} c_{i,k}^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{\ell}\left(\sum_{j=1}^{m} a_{ij}b_{jk}\right)^2 \\
&\leq \sum_{i=1}^{n}\sum_{k=1}^{\ell}\left(\sum_{j=1}^{m} a_{ij}^2\right)\left(\sum_{j=1}^{m} b_{jk}^2\right) \\
&= \left(\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2\right)\left(\sum_{k=1}^{\ell}\sum_{j=1}^{m} b_{jk}^2\right) \\
&= \|A\|_F^2 \cdot \|B\|_F^2.
\end{aligned}$$

Dabei war wieder einmal die Cauchy–Schwarzsche Ungleichung hilfreich. □

11.5 Gleichmäßige Stetigkeit und Konvergenz

Es gibt zur Stetigkeit eine unter den hier vorliegenden Umständen äquivalente Definition, die in anderen Büchern verwendet wird. Dazu gehen wir auf den Fall reeller Funktionen zurück und untersuchen, unter welchen Umständen wir eine Funktion f in einem Punkt x ihres Definitionsintervalls I stabil und sicher ausrechnen können. Wenn wir $f(x)$ bis auf einen beliebig vorgegebenen absoluten Fehler ϵ ausrechnen wollen, so müssen wir eine von ϵ, f und x abhängige Schranke δ angeben können, so daß die Rechnung mit einem fehlerhaften $\tilde{x}$ mit $|x-\tilde{x}| < \delta$ immer noch mit Sicherheit zu $|f(x)-f(\tilde{x})| < \epsilon$ führt.

Definition 11.30 *1. Eine reelle Funktion f ist* **stetig** *in einem Punkte x ihres Definitionsintervalls I, wenn es zu jedem $\epsilon > 0$ ein $\delta > 0$ gibt, so dass für alle $y \in I$ aus $|x-y| < \delta$ stets $|f(x)-f(y)| < \epsilon$ folgt. Dabei darf δ auch von x abhängen.*

2. Eine Funktion f ist stetig in ihrem ganzen Definitionsbereich I, wenn sie in jedem Punkte von I stetig ist.

Theorem 11.31 *Stetigkeit (Definition 11.30 auf Seite 291) und Folgenstetigkeit (Definition 11.3 auf Seite 271) sind bei reellen Funktionen äquivalent.*

Beweis: Es sei f in x stetig, und wir wollen Folgenstetigkeit in x zeigen. Gegeben sei also eine gegen x konvergente Folge $(x_n)_n$ und wir wollen Konvergenz von $f(x_n)$ gegen $f(x)$ zeigen. Dazu müssen wir uns ein $\epsilon > 0$ vorgeben

lassen und dann ein $N \in \mathbb{N}$ finden, so dass aus $n \geq N$ stets $|f(x)-f(x_n)| \leq \epsilon$ folgt. Wir nehmen das ϵ dankend entgegen und benutzen es erst einmal in der Stetigkeitsdefinition. Daraus bekommen wir ein $\delta > 0$, so dass aus $y \in I$ mit $|x - y| < \delta$ stets $|f(x) - f(y)| < \epsilon$ folgt. Dieses δ stecken wir in die Konvergenzdefinition der Folge $(x_n)_n$ hinein und bekommen ein $N \in \mathbb{N}$ so dass aus $n \geq N$ immer $|x_n - x| \leq \delta$ folgt. Diese x_n setzen wir als y ein und erhalten wie gewünscht $|f(x) - f(x_n)| \leq \epsilon$.

Jetzt setzen wir Folgenstetigkeit in x voraus und wollen Stetigkeit zeigen. Wir schließen indirekt und nehmen an, die Funktion f sei in x nicht stetig, und beweisen dann, dass sie auch nicht folgenstetig ist. Jetzt müssen wir die Stetigkeitsdefinition negieren, und alle Quantoren "umdrehen". Es gibt also ein $\epsilon > 0$ so dass für alle $\delta > 0$ ein $y \in I$ existiert mit $|x - y| < \delta$ und $|f(x)-f(y)| \geq \epsilon$. Setzt man hier der Einfachheit halber $\delta := 1/n$, so bekommt man für jedes n ein $y_n \in I$ mit $|x - y_n| < 1/n$ und $|f(x) - f(y_n)| \geq \epsilon$. Die Folge $(y_n)_n$ konvergiert also gegen x, aber die Ungleichung $|f(x)-f(y_n)| \geq \epsilon$ zeigt, dass $f(y_n)$ nicht gegen $f(x)$ konvergiert. Also ist f nicht folgenstetig. □

An dieser Stelle machen wir einen kleinen Exkurs und greifen die Definition 8.29 offener Mengen auf Seite 240 auf. Der folgende Satz dient in der **Topologie**[1] als Definition der Stetigkeit von Abbildungen, ist hier aber "nur" ein Satz.

Theorem 11.32 *Ist f eine stetige Abbildung zwischen metrischen oder normierten Räumen, so sind die Urbilder offener Mengen immer offen.*

Zum Beweis nehmen wir der Einfachheit halber den Fall $f : U \to V$ mit normierten Räumen an und nehmen eine beliebige offene Menge V_0 im Bildraum V her. Die Urbildmenge ist dann

$$U_0 := \{u \in U \ : \ f(u) \in V_0\}$$

und wir müssen zeigen, daß auch U_0 offen ist. Dazu nehmen wir uns ein beliebiges $u \in U_0$ und das zugehörige $v_0 := f(u_0) \in V_0$ vor. Weil V_0 offen ist, gibt es eine ϵ–Umgebung von v_0, die ganz in V_0 liegt. Mit diesem ϵ wenden wir die Stetigkeitsdefinition an und bekommen ein $\delta > 0$ mit der Eigenschaft, daß aus $\|u - u_0\|_U < \delta$ immer $\|f(u) - f(u_0)\|_V < \epsilon$ folgt. Also bildet f eine komplette δ–Umgebung von u_0 in die ϵ–Umgebung von $f(u_0)$ ab, die in V_0 liegt, und somit liegt die komplette δ–Umgebung von u_0 in der Menge U_0.

[1] `http://de.wikipedia.org/wiki/Topologie_(Mathematik)`

Diese ist also offen, denn wir haben zu einem beliebigen ihrer Elemente eine offene Umgebung gefunden, die ganz in der Menge liegt. □

Die nächste Frage betrifft das Problem, ob man bei einer überall stetigen Funktion zu einem gegebenen ϵ die Wahl des δ unabhängig vom Stetigkeitspunkt x treffen kann. Die Frage klingt sehr theoretisch, ist es aber nicht, weil man sehr oft gezwungen ist, so einen Schluß auszuführen.

Definition 11.33 *Eine reelle Funktion f heißt* **gleichmäßig stetig**[1] *auf ihrem Definitionsintervall I, wenn es zu jedem $\epsilon > 0$ ein $\delta > 0$ gibt, so daß für alle $x, y \in I$ mit $|x - y| < \delta$ auch $|f(x) - f(y)| < \epsilon$ folgt.*

Theorem 11.34 *(Satz von* **Heine**[2][3] *oder von* **Cantor**[4]*)*
Stetige Funktionen auf abgeschlossenen und beschränkten Intervallen I sind dort gleichmäßig stetig.

Beweis: Nehmen wir das Gegenteil an. Dann gibt es ein $\epsilon > 0$, so daß zu jedem $\delta > 0$ zwei Punkte $x_\delta, y_\delta \in I$ existieren mit $|x_\delta - y_\delta| < \delta$, aber $|f(x_\delta) - f(y_\delta)| \geq \epsilon$. Wir nehmen $\delta := 1/n$ für $n \in \mathbb{N}_{>0}$ und bekommen, mit leichter Änderung der Schreibweise, zu jedem $n \in \mathbb{N}_{>0}$ zwei Punkte $x_n, y_n \in I$ mit $|x_n - y_n| < 1/n$, aber $|f(x_n) - f(y_n)| \geq \epsilon$. Wegen der Abgeschlossenheit und Beschränktheit des Intervalls I gibt es eine gegen ein $x \in I$ konvergente Teilfolge der Folge $(x_n)_n$. Die Teilfolge der Folge $(y_n)_n$ mit denselben Indizes konvergiert dann wegen $|x_n - y_n| < 1/n$ auch gegen x, und die Stetigkeit von f liefert für diese Teilfolgen

$$\lim_n f(x_n) = f(x) = \lim_n f(y_n),$$

was im Widerspruch zu $|f(x_n) - f(y_n)| \geq \epsilon$ für alle $n \in \mathbb{N}_{>0}$ steht. □

Aufgabe: Die Funktion $x \to x^2$ ist zu jedem $a > 0$ auf $I := [-a, a]$ gleichmäßig stetig. Was ist zu gegebenem ϵ das denkbar größte δ, das man bei der gleichmäßigen Stetigkeit benutzen kann?

Die Definition 11.33 der gleichmäßigen Stetigkeit gilt vollkommen analog auch für Funktionen auf Teilmengen des $\mathbb{R}^k$ oder eines metrischen Raumes, man ersetzt nur die Beträge durch Normen oder Abstände in der Metrik. Wenn man dann aber den Satz 11.34 erweitern will, braucht man den Satz von

[1] http://de.wikipedia.org/wiki/Gleichm%C3%A4%C3%9Fige_Stetigkeit
[2] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Heine.html
[3] http://de.wikipedia.org/wiki/Satz_von_Heine
[4] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Cantor.html

Bolzano–Weierstraß, der zu beschränkten Folgen immer konvergente Teilfolgen liefert. Und deren Limes sollte wieder im Definitionsbereich der Funktion liegen, so dass man insgesamt vom Definitionsbereich der stetigen Funktion verlangen muss, dass er eine abgeschlossene und beschränkte Teilmenge eines $\mathbb{R}^k$ ist. Mit fast wörtlich gleichem Beweis, den wir hier nicht wiederholen, gilt dann auch

Theorem 11.35 *Stetige Funktionen auf abgeschlossenen und beschränkten Teilmengen eines $\mathbb{R}^k$ sind auf ihrem Definitionsbereich gleichmäßig stetig.*□

In Vorbereitung der Integralrechnung fragen wir uns jetzt, wie man stetige Funktionen durch einfachere Funktionen ersetzen kann, ohne einen allzugroßen absoluten Fehler zu begehen. In der Mathematik nennt man dann die Ersatzfunktion eine **Approximation** der gegebenen Funktion. Wir betrachten erst die Ersetzung einer Funktion durch eine stückweise konstante Funktion (Zeichnung in der Vorlesung).

Theorem 11.36 *Es sei f eine auf einem abgeschlossenen und beschränkten Intervall $[a,b]$ stetige Funktion. Wir betrachten Intervallzerlegungen Δ der Form*

$$\Delta \; : \quad a = x_0 < x_1 < \ldots < x_{n+1} = b$$

mit beliebigem $n = n(\Delta)$ und der **Fülldichte** $h := h(\Delta) := \max\limits_{0 \le j \le n} (x_{j+1} - x_j)$.
Ferner gehöre zu Δ die Auswahl eines beliebigen $y_j \in [x_j, x_{j+1}], 0 \le j \le n$. Dann wird die Funktion f ersetzt durch die stückweise konstante Funktion

$$f_\Delta(x) \;\; := \;\; f(y_j) \text{ für alle } x \in [x_j, x_{j+1}],\; 0 \le j \le n,$$

wobei die eventuelle Mehrfachdefinition in den Punkten x_j durch beliebige Auswahl einer der beiden Alternativen behoben werden kann. Es gilt:
Zu jedem $\epsilon > 0$ gibt es ein $h_0 > 0$, so daß für alle Zerlegungen Δ mit $h(\Delta) \le h_0$ die Abschätzung

$$|f_\Delta(x) - f(x)| < \epsilon \text{ für alle } x \in [a,b]$$

folgt.

Beweis: Die Funktion f ist gleichmäßig stetig auf $[a,b]$. Es gibt also zu jedem $\epsilon > 0$ ein $\delta > 0$, so daß für alle $x, y \in [a,b]$ mit $|x - y| < \delta$ stets $|f(x) - f(y)| < \epsilon$ folgt. Wir wählen $h_0 := \delta$ und eine beliebige Zerlegung Δ mit $h(\Delta) \le h_0 < \delta$. Dann folgt

$$|f_h(x) - f(x)| = |f(y_j) - f(x)| < \epsilon \text{ für alle } x \in [x_j, x_{j+1}]$$

weil $|y_j - x| \leq x_{j+1} - x_j \leq h < \delta$ gilt. Also gilt auch $|f_h(x) - f(x)| < \epsilon$ für alle $x \in [a, b]$. □

Es ist bemerkenswert, daß der Satz für alle hinreichen feinen Zerlegungen gilt, egal wie sie im Detail aussehen. Die Ersetzung einer Funktion durch eine stückweise konstante Funktion sieht unpraktisch aus, ist es aber nicht, weil schnelle und extrem rundungsgenaue Verfahren zur Auswertung komplizierter Funktionen gar nichts berechnen, sondern in einer Tabelle vorausberechneter Werte nachsehen.

Wir behandeln jetzt noch die wichtigste Anwendung des obigen Satzes in der Mathematik. Unter den Voraussetzungen von Satz 11.36 auf Seite 294 legen wir jetzt die Funktionswerte **zweier** stückweise konstanter Approximationen an die gegebene stetige Funktion f fest. Dabei wollen wir die gegebene Funktion f von oben und unten sauber abschätzen und "einschließen", und wir werden das in Abschnitt 13.1.1 auf Seite 357 bei der Definition des bestimmten (Riemann-) Integrals[1] dringend brauchen.

Theorem 11.37 *Es gelten die Voraussetzungen von Satz 11.36. Bei den Funktionen*

$$f_\Delta^{oben}(t) := \max_{x_i \leq x \leq x_{i+1}} f(t) \text{ für alle } t \in [x_i, x_{i+1}],\ 0 \leq i \leq n$$
$$f_\Delta^{unten}(t) := \min_{x_i \leq x \leq x_{i+1}} f(t) \text{ für alle } t \in [x_i, x_{i+1}],\ 0 \leq i \leq n$$

sei die eventuelle Mehrfachdefinition in den Punkten x_j so behoben, daß bei f_Δ^{oben} das Maximum und bei f_Δ^{unten} das Minimum der beiden Alternativen genommen werde. Es gilt:
Zu jedem $\epsilon > 0$ gibt es ein $h_0 > 0$, so daß für alle Zerlegungen Δ mit $h(\Delta) \leq h_0$ die Abschätzungen

$$\begin{aligned} |f_\Delta^{oben}(x) - f_\Delta^{unten}(x)| &< \epsilon \\ f_\Delta^{unten}(x) \leq f(x) &\leq f_\Delta^{oben}(x) \end{aligned}$$

für alle $x \in [a, b]$ gelten.

Beweis: Man wendet den vorigen Satz für $\epsilon/2$ an. Das ist möglich, weil die Funktion f auf den abgeschlossenen Teilintervallen $[x_i, x_{i+1}]$ nach Satz 11.11 auf Seite 275 jeweils ihr Minimum und Maximum annimmt. □

In der Literatur spricht man bei der obigen Konstruktion gelegentlich auch von **Oberfunktionen** und **Unterfunktionen**.

[1] `http://de.wikipedia.org/wiki/Riemann-Integral`

Wen es stört, daß die Ersatzfunktionen nicht stetig sind, kann stückweise lineare und stetige Funktionen verwenden. Aber man bekommt leider keine Einschließung von oben bzw. unten.

Theorem 11.38 *Es sei f eine auf einem abgeschlossenen und beschränkten Intervall $[a,b]$ stetige Funktion. Wir betrachten Intervallzerlegungen Δ der Form*

$$\Delta : \quad a = x_0 < x_1 < \ldots < x_{n+1} = b$$

mit beliebigem $n = n(\Delta)$ und $h := h(\Delta) := \max\limits_{0 \le j \le n} x_{j+1} - x_j$. Dann wird die Funktion f ersetzt durch die stückweise affin–lineare Funktion

$$\begin{aligned} f_\Delta(x) &:= \frac{x - x_j}{x_{j+1} - x_j} f(x_{j+1}) + \frac{x_{j+1} - x}{x_{j+1} - x_j} f(x_j) \\ &\quad \text{für alle } x \in [x_j, x_{j+1}],\ 0 \le j \le n, \end{aligned}$$

wobei die eventuelle Mehrfachdefinition in den Punkten x_j nicht stört.
Es gilt:
Zu jedem $\epsilon > 0$ gibt es ein $h_0 > 0$, so daß für alle Zerlegungen Δ mit $h(\Delta) \le h_0$ die Abschätzung

$$|f_\Delta(x) - f(x)| < \epsilon \text{ für alle } x \in [a,b]$$

folgt.

Beweis: Die Funktion f ist gleichmäßig stetig auf $[a,b]$. Es gibt also zu jedem $\epsilon > 0$ ein $\delta > 0$, so daß für alle $x, y \in [a,b]$ mit $|x - y| < \delta$ stets $|f(x) - f(y)| < \epsilon$ folgt. Wir wählen $h_0 := \delta$ und eine beliebige Zerlegung Δ mit $h(\Delta) \le h_0 < \delta$. Dann folgt

$$\begin{aligned} |f_h(x) - f(x)| &= |\frac{x - x_j}{x_{j+1} - x_j} f(x_{j+1}) \\ &\quad + \frac{x_{j+1} - x}{x_{j+1} - x_j} f(x_j) - f(x)| \\ &= |\frac{x - x_j}{x_{j+1} - x_j} (f(x_{j+1}) - f(x)) \\ &\quad + \frac{x_{j+1} - x}{x_{j+1} - x_j} (f(x_j) - f(x))| \\ &< \frac{x - x_j}{x_{j+1} - x_j} \epsilon + \frac{x_{j+1} - x}{x_{j+1} - x_j} \epsilon \\ &= \epsilon. \end{aligned}$$

Also gilt auch $|f_h(x) - f(x)| < \epsilon$ für alle $x \in [a,b]$. □

11.6 Funktionenfolgen

Der nächste Schritt betrifft Folgen, aber nicht Zahlenfolgen, sondern Folgen von Funktionen. Beispielsweise hat man in einem iterativen Algorithmus nacheinander Funktionen $f_1, f_2, \ldots$ auf einem gemeinsamen Definitionsbereich I berechnet, und man möchte wissen, ob es eine Limesfunktion f gibt und ob die Limesfunktion stetig ist. Nehmen wir eine solche Folge $(f_n)_n$ von reellen Funktionen auf I als gegeben an.

Zuerst kann man für feste $x \in I$ die Zahlenfolgen $(f(x_n))_n$ ansehen und auf Konvergenz untersuchen.

Definition 11.39 *Eine Funktionenfolge $(f_n)_n$ auf I konvergiert* **punktweise** *gegen eine Funktion f auf I, wenn für jedes $x \in I$ die Konvergenz $\lim_{n\to\infty} f_n(x) = f(x)$ eintritt.*

Beispiel: Die Funktionen $x \to x^n$ konvergieren punktweise auf $I = [0, 1]$ gegen die unstetige Funktion

$$f(x) = \begin{cases} 0 & 0 \leq x < 1 \\ 1 & x = 1. \end{cases}$$

Das ist ein abschreckendes Beispiel, denn die Monome sind sehr anständige stetige Funktionen, beschränkt auf $[0, 1]$, die Folgen $(x^n)_n$ sind monoton und beschränkt, und man könnte doch bei all diesen schönen Voraussetzungen erwarten, daß die Grenzfunktion wieder stetig ist. Ist sie aber nicht, und man braucht stärkere Voraussetzungen, um die Stetigkeit einer Grenzfunktion nachzuweisen.

Die punktweise Konvergenz von Funktionenfolgen kann man definieren und untersuchen, ohne dass man einen Funktionenraum hat. Wenn man aber einen normierten Funktionenraum betrachtet, ist zu klären, wie sich die Konvergenz einer Funktionenfolge in der Norm zur punktweisen Konvergenz verhält.

Definition 11.40 *Gegeben sei eine Funktionenfolge $(f_n)_n$ aus einem normierten Raum F von reellwertigen Funktionen auf einem Definitionsbereich I, und die Norm auf F werde mit $\|.\|_F$ bezeichnet. Wenn es eine Funktion $f \in F$ gibt, für die*

$$\lim_{n\to\infty} \|f - f_n\|_F = 0$$

gilt, so sagt man, die Folge $(f_n)_n$ sei **normkonvergent** *gegen f.*

Das ist nichts anderes als Definition 8.17 auf Seite 234 auf dem metrischen Raum, der durch die Norm auf F gegeben ist. Was gilt nun für das Verhältnis zwischen punktweiser und Normkonvergenz?

- Konvergiert $(f_n)_n$ punktweise gegen eine reellwertige Funktion auf I, so ist noch nicht einmal klar, ob f überhaupt in F liegt, d.h. Normkonvergenz ist keineswegs gesichert.
- Umgekehrt kann man aus Normkonvergenz nicht immer auf punktweise Konvergenz schließen.

Zum erstgenannten Faktum kann man das obige Beispiel der Funktionenfolge $(x^n)_n$ auf $[0,1]$ heranziehen, wenn man auf eine beliebige Weise einen Raum F stetiger Funktionen auf $[0,1]$ definiert. Um die zweite Situation anzusehen, kann man auf dem Raum $F = C[-1,1]$ der stetigen Funktionen auf $[-1,1]$ die Norm

$$\|f\|_1 := \int_{-1}^{+1} |f(t)|dt \text{ für alle } f \in C[-1,1]$$

einführen, aber wenn man dann eine Folge von stetigen "Hütchenfunktionen"

$$f_h(x) := \begin{cases} 1 + t/h & -h \leq t \leq 0 \\ 1 - t/h & 0 \leq t \leq h \\ 0 & \text{sonst} \end{cases}$$

mit $h = 1/n$ und $n \geq 1$ betrachtet, ist $\|f_{1/n}\|_1 = 1/n$ eine Nullfolge, während $f_h(0) = 1$ für alle $h \in [0,1]$ gilt. Also hat man Normkonvergenz gegen die Nullfunktion, aber punktweise konvergieren die Funktionen gegen die unstetige Funktion

$$f(x) := \left\{ \begin{array}{ll} 0 & x \neq 0 \\ 1 & x = 0 \end{array} \right\}.$$

Aufgabe: Kann man so auch für die Monomfolge $(x^n)_n$ schließen?

Die Lage sieht etwas besser aus, wenn man weiß, daß die Auswerteabbildung

$$\delta_x \; : \; f \mapsto f(x)$$

für jedes $x \in I$ als lineare Abbildung $F \to \mathbb{R}$ beschränkt ist, d.h. wenn

$$|f(x)| \leq C(x)\|f\|_F \text{ für alle } f \in F$$

mit von x abhängigen Konstanten $C(x)$ gilt. Denn dann kan man aus Normkonvergenz immer auf punktweise Konvergenz schließen, indem man

$$|f(x) - f_n(x)| \leq C(x)\|f - f_n\|_F$$

benutzt. Also ist klar, daß in obigem Beispiel eines Funktionenraums die Auswerteabbildung nicht immer beschränkt ist, und das sieht man daraus, daß

$$1 = f_{1/n}(0) \leq C(x)\|f_{1/n}\|_F = C(x)/n$$

nicht für $n \to \infty$ gelten kann.

Aber es gibt auch einen angenehmen Fall:

Theorem 11.41 *Die auf einem abgeschlossenen und beschränkten Intervall $I := [a, b]$ stetigen (und gleichmäßig stetigen) Funktionen bilden unter der Norm*

$$\|f\|_\infty := \max_{a \leq x \leq b} |f(x)| \text{ für alle } f \in C[a, b]$$

einen vollständigen normierten Vektorraum, der mit $C[a, b]$ bezeichnet wird. Cauchyfolgen in diesem Raum konvergieren gegen eine stetige Grenzfunktion. Die Konvergenz einer Folge $(f_n)_n$ in diesem Raum gegen eine Grenzfunktion f impliziert immer eine punktweise Konvergenz, und sogar eine **gleichmäßige Konvergenz** *in folgendem Sinne:*

> *Zu jedem $\epsilon > 0$ gibt es ein $N(\epsilon) \in \mathbb{N}$, so daß für alle $n \geq N(\epsilon)$ und alle $x \in [a, b]$ stets $|f_n(x) - f(x)| < \epsilon$ folgt.*

Ferner ist die Auswerteabbildung stetig und beschränkt mit Konstante 1.

Beweis: Stetige Funktionen nehmen auf abgeschlossenen Intervallen nach Satz 11.11 auf Seite 275 ihr Maximum an. Deshalb ist $\|f\|_\infty$ wohldefiniert, und man bekommt in der Tat eine Norm, wie man leicht nachrechnet. Ebenso ist die letzte Behauptung des Satzes sehr einfach zu sehen, weil nämlich $|f(x)| \leq \|f\|_\infty$ für alle $f \in C[a, b],\ x \in [a, b]$ gilt.

Zum Beweis der Vollständigkeit müssen wir zeigen, daß jede Cauchyfolge in der Norm gegen eine stetige Grenzfunktion konvergiert. Eine Cauchyfolge $(f_n)_n$ in $C[a, b]$ hat die Eigenschaft, daß es zu jedem $\epsilon > 0$ ein $N(\epsilon) \in \mathbb{N}$ gibt mit $\|f_n - f_m\|_\infty < \epsilon$ für alle $m, n \geq N(\epsilon)$. Dann folgt aber auch $|f_n(x) - f_m(x)| < \epsilon$ für alle $x \in I$, d.h. man hat zu jedem $x \in I$ eine reelle Cauchyfolge $(f_n(x))_n$, die deshalb gegen eine reelle Zahl konvergieren muß, die wir $f(x)$ nennen. Das liefert eine reellwertige Funktion $x \to f(x)$, aber wir wissen noch nicht, ob diese stetig ist und in $C[a, b]$ liegt. Übrigens beweist man ganz analog, daß eine in $C[a, b]$ konvergente Folge immer punktweise und gleichmäßig konvergiert, aber das wollen wir den geneigten Leser(innen) überlassen.

Zum Beweis der (gleichmäßigen) Stetigkeit der Grenzfunktion f müssen wir uns ein beliebiges ϵ vorgeben lassen und dann ein $\delta > 0$ finden, so daß aus $|x-y| < \delta$ immer $|f(x)-f(y)| < \epsilon$ folgt, und zwar für alle $x, y \in [a,b]$. Nehmen wir also ein solches ϵ entgegen, und wenden wir dann die Cauchyfolgeneigenschaft auf $\epsilon/3$ an. Es gibt dann ein $N(\epsilon/3) \in \mathbb{N}$ mit $\|f_n - f_m\|_\infty < \epsilon/3$ für alle $m, n \geq N(\epsilon/3)$. Wir setzen ein beliebiges x ein und lassen das m gnadenlos gegen Unendlich gehen und bekommen

$$\begin{array}{rcl} |f_n(x) - f_m(x)| & \leq & \|f_n - f_m\|_\infty < \epsilon/3 \\ |f_n(x) - f(x)| & = & \lim_{m\to\infty} |f_n(x) - f_m(x)| \leq \epsilon/3 \end{array}$$

für alle $n \geq N(\epsilon/3)$. Es folgt

$$\begin{array}{rcl} |f(x) - f(y)| & \leq & |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \\ & \leq & \epsilon/3 + |f_n(x) - f_n(y)| + \epsilon/3 \end{array}$$

für alle $n \geq N(\epsilon/3)$. Jetzt fixieren wir $n := N(\epsilon/3)$ und benutzen die gleichmäßige Stetigkeit von $f_n = f_{N(\epsilon/3)}$, um zu gegebenem $\epsilon/3$ ein $\delta > 0$ zu finden mit $|f_n(x) - f_n(y)| < \epsilon/3$ für alle $|x-y| < \delta$. Das liefert insgesamt $|f(x) - f(y)| < \epsilon$ für alle $|x-y| < \delta$. □.

Korollar 11.42 *Zu jeder stetigen Funktion $f \in C[a,b]$ gibt es eine Folge stückweise linearer und stetiger Funktionen, die gegen f in der Norm $\|.\|_\infty$ konvergiert.*

Beweis: Das liefert Satz 11.38 auf Seite 296. □.

Theorem 11.43 *(Approximationssatz von* **Weierstrass**[1]*)*
Zu jeder stetigen Funktion $f \in C[a,b]$ gibt es eine Folge von Polynomen, die gegen f in der Norm $\|.\|_\infty$ konvergiert.

Der Beweis ist für Informatik–Studierende zu schwierig. Aber es sollte klar sein, daß der Satz sehr wichtig ist, denn in Rechnern kann man sehr effizient Polynome ausrechnen, mit denen man dann gewisse stetige Funktionen sehr genau reproduzieren kann.

Andere Funktionenräume mit anderen Normen sind problematisch, wie wir schon am Beispiel $C[-1,1]$ mit der Norm $\|.\|_1$ gesehen haben.

Aufgabe: Man zeige, daß man nicht besser wegkommt, wenn man auf $C[-1,1]$ durch

$$\|f\|_p^p := \int_{-1}^{+1} |f(x)|^p dx$$

für beliebige $p \in [1,\infty)$ eine Norm definiert.

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Weierstrass.html`

12 Differentialrechnung

Hier beginnt das Kernstück der **Analysis**: die Differentialrechnung[1]. Danach folgt ihre Umkehrung, die Integralrechnung. Es sollte aus dem Schulunterricht schon bekannt sein, wie wichtig diese Kulturtechnik ist. Schließlich ist die Geschwindigkeit eines Autos, gemessen in km/h, die Ableitung des jeweiligen Ortes des Autos nach der Zeit, und die Beschleunigung ist wiederum die Ableitung der Geschwindigkeit nach der Zeit. Ein Tachometer ist ein analoges Gerät zum Differenzieren, und in der Trägheitsnavigation geht man den umgekehrten Weg: man integriert alle Beschleunigungen zweifach, um die Ortsveränderung des bewegten Objektes zu rekonstruieren. Weitere Anwendungsfelder sind natürlich die Optimierungsaufgaben, und (etwas versteckter) sind auch alle mathematischen Techniken zur Signalverarbeitung (z.B. Datenkompression) ohne Differential– und Integralrechnung nicht zu verstehen.

12.1 Differenzierbare Funktionen

12.1.1 Differenzierbarkeit

Definition 12.1 *Es sei f eine reellwertige Funktion auf einem reellen Intervall I und es sei $x \in I$ fest. Wenn der Grenzwert*

$$\lim_{n\to\infty} \frac{f(x_n) - f(x)}{x_n - x}$$

für jede gegen x konvergente Folge $(x_n)_n$ mit $x_n \neq x$ und $x_n \in I$ für alle $n \in \mathbb{N}$ existiert und denselben Wert hat, wird dieser Wert mit $f'(x)$ oder $\frac{df}{dx}$ bezeichnet und die **Ableitung** *von f in x genannt.*
Die Funktion f heißt dann **in** *x* **differenzierbar**[2]*. Ist f in allen $x \in I$ differenzierbar, so heißt f* **in** *I* **differenzierbar**, *und die Funktion $f' : I \to \mathbb{R}$ mit $x \mapsto f'(x)$ heißt (erste)* **Ableitung** *von f.*
Die Funktion f heißt **stetig differenzierbar** *in I, wenn sie in I differenzierbar ist und die Ableitung in I stetig ist.*

Man mache sich an Hand einer Skizze die geometrische Bedeutung dieses Grenzwertes klar: die durch $(x, f(x))$ und $(x_n, f(x_n))$ definierten Geraden

[1] `http://de.wikipedia.org/wiki/Differentialrechnung`
[2] `http://de.wikipedia.org/wiki/Differenzierbarkeit`

sind **Sekanten** des Funktionsgraphen, weil sie ihn an zwei Punkten schneiden, und die Sekanten streben für $n \to \infty$ gegen die Tangente[1] an den Graphen in $(x, f(x))$, wenn die Funktion f in x differenzierbar ist.

Die obige Schreibweise der Ableitung als Funktion ist manchmal etwas mißverständlich, weil das Zeichen x in doppelter Bedeutung vorkommt. Zunächst ist x eine freie Variable, die angibt, wie f definiert ist. Dabei ist es irrelevant, ob man z.B. $f(x) = 3 \cdot x \cdot \sin^2(x)$ oder $f(t) = 3 \cdot t \cdot \sin^2(t)$ schreibt. Gemeint ist in beiden Fällen dasselbe, nämlich

$$f \ : \ I\!R \to I\!R, \ \ x \mapsto 3 \cdot x \cdot \sin^2(x) \text{ für alle } x \in I\!R.$$

Die Notation $f(x) = 3 \cdot x \cdot \sin^2(x)$ ist die Kurzform dieses Sachverhalts. Die Funktion heißt f, nicht $f(x)$, und das x ist irrelevant, wenn man nicht eine die Funktion f definierende Formel angibt, die x enthält.

Die Ableitung als Funktion ist dann wieder mit einer freien Variablen zu schreiben, etwa $f'(s) = 3 \cdot \sin^2(s) + 6 \cdot s \cdot \sin(s) \cdot \cos(s)$ als Kurzform von

$$f' \ : \ I\!R \to I\!R, \ \ s \mapsto 3 \cdot \sin^2(s) + 6 \cdot s \cdot \sin(s) \cdot \cos(s) \text{ für alle } s \in I\!R.$$

Wenn man die Ableitung f' an einer festen Stelle x auswertet, schreibt man $f'(x)$, aber damit ist keine Funktion, sondern eine reelle Zahl gemeint.

Eine präzisere Notation verwendet eine Abbildung D, die einer Funktion f ihre Ableitung $D(f)$ als Funktion zuordnet (sofern diese existiert, natürlich). Diese Abbildung ist linear (siehe unten) und bildet differenzierbare Funktionen auf Funktionen ab. Die Auswertung an einer Stelle x kann man als lineare Abbildung δ_x schreiben, die auf einer Funktion f die Wirkung $\delta_x(f) := f(x)$ hat. Die Auswertung einer Ableitung von f ist dann genaugenommen $\delta_x(D(f))$.

Wenn man eine Funktion f mit einer freien Variablen t schreibt und dann in x auswertet, wird in manchen Büchern die Notation

$$\frac{d}{dt} f(t)_{|x} := f'(x) \in I\!R$$

verwendet, die zwischen der freien Variablen und dem Auswertepunkt unterscheidet.

[1] `http://de.wikipedia.org/wiki/Tangente`

Im Sinne des Abschnitts 8.5 auf Seite 241 läßt sich die Differenzierbarkeit von f in x auch in der Form

$$\lim_{y \to x} \frac{f(y) - f(x)}{y - x} = f'(x)$$

schreiben, wobei allerdings der formale Schönheitsfehler auftritt, daß man von den beliebigen gegen x konvergente Folgen voraussetzen muss, dass sie nicht x selbst enthalten. Einen Ausdruck der Form $\frac{f(y)-f(x)}{y-x}$ mit $x \neq y$ nennt man auch **Differenzenquotient**[1]. Er gibt die Steigung der Sekante an, während $f'(x)$ die Steigung der Tangente an f in $(x, f(x))$ angibt. Dabei ist die Steigung der Tangens des Steigungswinkels (siehe den Steigungswinkel α der Tangente in Abb. 9).

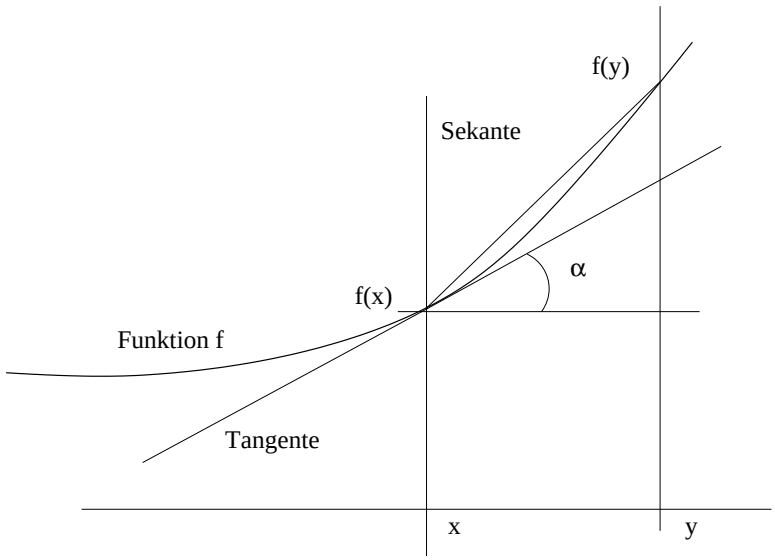


Abbildung 9: Differenzenquotient, Sekante und Tangente

Eine alternative und sehr praktische Schreibweise für Differenzierbarkeit ist

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

wobei man sich unter h beliebige Nullfolgen $(h_n)_n$ vorstellt, so daß $h_n \neq 0$ und $x + h_n \in I$ für alle $n \in I\!N$ gilt.

Man kann den lästigen, im Grenzfall verschwindenden Nenner loswerden, indem man ihn wegmultipliziert und dann eine o–Relation hinschreibt:

[1] `http://de.wikipedia.org/wiki/Differenzenquotient`

Theorem 12.2 *Eine reellwertige Funktion f ist in einem Punkte x ihres Definitionsintervalls I genau dann differenzierbar, wenn es eine mit $f'(x)$ bezeichnete reelle Zahl gibt, so daß es für jedes $\epsilon > 0$ ein $\delta > 0$ gibt, so dass aus $x, y \in I$, $|x - y| < \delta$ stets*

$$|f(y) - f(x) - f'(x)(y - x)| \leq \epsilon |y - x| \tag{12.3}$$

folgt.

Beweis: Es sei f in x differenzierbar im Sinne der Definition 12.1, und wir nehmen an, die alternative Form des obigen Satzes sei nicht erfüllt. Dann gibt es zu jeder reellen Zahl α ein ϵ, so dass für alle $\delta > 0$ Zahlen $y_\delta \in I$ mit $|x - y_\delta| < \delta$ existieren, so dass $|f(y_\delta) - f(x) - \alpha(y_\delta - x)| > \epsilon |y_\delta - x|$ gilt. Natürlich wählen wir $\alpha = f'(x)$ und $\delta = 1/n$, um eine Folge $(y_n)_n$ zu bekommen mit $|x - y_n| < 1/n$ und $|f(y_n) - f(x) - f'(x)(y_n - x)| > \epsilon |y_n - x|$. Es folgt $y_n \neq x$ und

$$\left| \frac{f(y_n) - f(x)}{y_n - x} - f'(x) \right| > \epsilon$$

im Widerspruch zur Annahme.

Die Umkehrung ist einfach. Wir wählen in der Notation von Theorem 12.2 zu gegebenem ϵ das passende δ mit

$$|f(y) - f(x) - f'(x)(y - x)| \leq \epsilon \cdot |y - x|$$

für alle x, y mit $|x - y| < \delta$. Dann nehmen wir eine beliebige gegen x konvergente Folge $(x_n)_n$ mit $x_n \neq x$ für alle n und bekommen ein N, so daß für alle $n \geq N$ immer $|x_n - x| < \delta$ gilt. Damit gehen wir in die obige Abschätzung und bekommen

$$\left| \frac{f(x_n) - f(x)}{x_n - x} - f'(x) \right| \leq \epsilon$$

für alle $n \geq N$, was die Differenzierbarkeit von f in x beweist. □

Der Satz zeigt die geometrische Bedeutung der Ableitung, denn der Graph der Funktion $g_x(y) := f(x) + f'(x)(y - x)$ ist genau die Tangente an den Graphen von f in $(x, f(x))$.

Differentiation ist eine lineare Operation auf einer Funktion. Deshalb gilt

Theorem 12.4 *Die in einem festen Punkte x ihres gemeinsamen Definitionsbereichs I differenzierbaren Funktionen bilden einen reellen Vektorraum.*

Die in ihrem gemeinsamen Definitionsbereich I differenzierbaren Funktionen bilden einen reellen Vektorraum.
Die in ihrem gemeinsamen Definitionsbereich I stetig differenzierbaren Funktionen bilden einen reellen Vektorraum.

Theorem 12.5 *Für die Differentiation von Funktionen gelten die folgenden Regeln, sofern die vorkommenden Größen Sinn machen:*

1. **Produktregel**[1]

$$(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$$

falls f und g in x differenzierbar sind.

2. **Quotientenregel**[2]

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x) \cdot g(x) - g'(x) \cdot f(x)}{g^2(x)}$$

falls f und g in x differenzierbar sind und $g(x) \neq 0$ gilt.

3. **Kettenregel**[3]

$$(g \circ f)'(x) = g'(f(x)) \cdot f'(x) \tag{12.6}$$

falls g in $f(x)$ und f in x differenzierbar ist.

4. **Ableitung der Umkehrfunktion**:

$$(f^{-1})'(f(x)) = \frac{1}{f'(x)} \tag{12.7}$$

sofern f in x differenzierbar ist und $f'(x) \neq 0$ gilt.

Die Aussagen im obigen Satz sind so zu verstehen, daß unter den angegebenen Voraussetzungen die links stehenden Funktionen differenzierbar sind und ihre Ableitung durch die rechtsstehende Formel ausrechenbar ist. Die Beweise sind durchweg einfach, stehen in allen Büchern und werden hier weggelassen. Problematisch ist höchstens die letzte Regel, weil bewiesen werden muß, daß aus der Bedingung $f'(x) \neq 0$ die Existenz der Umkehrfunktion von f in einer Umgebung von $f(x)$ folgt. □

[1] `http://de.wikipedia.org/wiki/Produktregel`
[2] `http://de.wikipedia.org/wiki/Quotientenregel`
[3] `http://de.wikipedia.org/wiki/Kettenregel`

Theorem 12.8 *Ist $f : I \to \mathbb{R}$ an einer Stelle $x \in I$ differenzierbar und gilt dort $f'(x) \neq 0$, so ist f in einer Umgebung von x streng monoton, und zwar steigend, wenn $f'(x) > 0$ gilt, sonst fallend.*

Beweis: Wir wählen uns ein $\epsilon < |f'(x)|/2$ und bekommen aus Satz 12.2 ein δ, so daß aus $|y - x| < \delta$ die Ungleichung (12.3) folgt. Für jedes solche $y \neq x$ ergibt sich

$$\left| \frac{f(y) - f(x)}{y - x} - f'(x) \right| \leq \epsilon < \frac{|f'(x)|}{2},$$

und deshalb hat der Differenzenquotient $\frac{f(y)-f(x)}{y-x}$ für diese y dasselbe Vorzeichen wie $f'(x)$. Ist $f'(x)$ positiv, so ist f um x streng monoton steigend, andernfalls fallend. Der Bereich der strengen Monotonie umfaßt alle y aus dem Definitionsbereich von f mit $|x - y| < \delta$. Dort existiert dann die Umkehrfunktion von f. □

Die Formel (12.7) für die Ableitung der Umkehrfunktion wird oft falsch angewendet. Man mache sich klar, daß links als Argument von f^{-1} ein Wert aus dem **Bild** von f stehen muß, also keinesfalls x statt $f(x)$ stehen darf. Die Formel selbst ergibt sich auch leicht aus der Kettenregel (12.6), wenn man die Gleichung $x = f^{-1}(f(x))$ differenziert:

$$\begin{aligned} 1 &= (f^{-1}(f(x)))' \\ &= (f^{-1})'(f(x)) \cdot f'(x). \end{aligned}$$

Die Ableitungen $f_n'(x) = n \cdot x^{n-1}$ der Monome $f_n : x \mapsto x^n$ bekommt man leicht durch Induktion, weil man $f_0'(x) = 0$ und $f_1'(x) = 1$ direkt ausrechnen kann und die Produktregel auch den Induktionsschluß

$$\begin{aligned} (f_{n+1})'(x) &= (f_1 \cdot f_n)'(x) \\ &= f_1'(x) \cdot f_n(x) + f_1(x) \cdot f_n'(x) \\ &= f_n(x) + x \cdot f_n'(x) \\ &= x^n + x \cdot n \cdot x^{n-1} \\ &= (n+1)x^n \end{aligned}$$

liefert. Bei der Exponentialfunktion muß man der Versuchung widerstehen, die Reihe gliedweise zu differenzieren, denn das ist eine nicht immer erlaubte Operation. Aber es folgt aus

$$\frac{\exp(x+h) - \exp(x)}{h} = \exp(x) \frac{\exp(h) - \exp(0)}{h}$$

die Differenzierbarkeit an jeder Stelle x mit $\exp'(x) = \exp(x) \cdot \exp'(0)$, wenn die Differenzierbarkeit in Null geklärt ist. Aus der Exponentialreihe folgt

dazu

$$\begin{aligned}
|\exp(h) - 1 - h| &= \left| \sum_{j=2}^{\infty} \frac{h^j}{j!} \right| \\
&\leq \sum_{j=2}^{\infty} \frac{|h|^j}{j!} \\
&= h^2 \sum_{j=2}^{\infty} \frac{|h|^{j-2}}{j!} \\
&= h^2 \sum_{j=0}^{\infty} \frac{|h|^j}{(j+2)!} \\
&\leq h^2 \sum_{j=0}^{\infty} \frac{|h|^j}{j!} \\
&= h^2 \exp(|h|)
\end{aligned}$$

und daraus bekommt man $\exp'(0) = 1$, was zu $\exp'(x) = \exp(x)$ für alle $x \in \mathbb{R}$ führt. Die Exponentialfunktion hat sich selbst als Ableitung. Als Anwendung der Formel (12.7) berechnen wir die Ableitung des Logarithmus über

$$\begin{aligned}
\log'(\exp(x)) &= \frac{1}{\exp(x)} \\
y &:= \exp(x) \\
\log'(y) &= \frac{1}{y}
\end{aligned}$$

für alle $y > 0$.

Aus der Kettenregel, angewendet auf $a^x := \exp(x \cdot \log(a))$ folgt

$$(a^x)' = \log(a) \cdot a^x.$$

Man kann das so interpretieren, daß die durch Differentiation gegebene lineare Abbildung die Funktionen a^x als "Eigenvektoren" zu Eigenwerten $\log(a)$ hat. Die Ableitung des Logarithmus erlaubt, die Gleichung $x^\alpha = \exp(\alpha \cdot \log(x))$ zu differenzieren, um die allgemeine Regel

$$\begin{aligned}
(x^\alpha)' &= \exp(\alpha \cdot \log(x)) \alpha \log'(x) \\
&= \alpha x^\alpha \frac{1}{x} \\
&= \alpha \cdot x^{\alpha - 1}
\end{aligned}$$

zu erhalten, die auch das Differenzieren von Wurzelfunktionen beschreibt. Man muss dabei darauf achten, dass oft die Differenzierbarkeit in Null fehlschlägt, und zwar bei allen x^α mit $\alpha < 1$.

Die Ableitungen von Sinus und Cosinus erhält man, indem man entweder die Funktion $\exp(ix) = cos(x) + i \cdot sin(x)$ mit einiger Frechheit nach der reellen Variablen x differenziert oder die Additionstheoreme anwendet, was sich als gleichwertig herausstellt. Mit der obigen Argumentationstechnik oder der Kettenregel sieht man, dass wegen

$$\begin{aligned}\frac{\exp(i(x+h)) - \exp(ix)}{h} &= \exp(ix)\frac{\exp(ih)) - \exp(0)}{h} \\ &= \exp(ix)\frac{1 + ih + \mathcal{O}(h^2) - 1}{h}\end{aligned}$$

die Gleichung $\exp(ix)' = i \cdot \exp(ix)$ gilt, und das bedeutet

$$\begin{aligned}(cos(x) + i \cdot sin(x))' &= i \cdot (cos(x) + i \cdot sin(x)) \\ &= -sin(x) + i \cdot cos(x)\end{aligned}$$

sodass man

$$cos'(x) = -sin(x) \text{ und } sin'(x) = cos(x)$$

bekommt.

Natürlich kann man Funktionen unter günstigen Umstanden mehrfach differenzieren. Bis zur dritten Ableitung werden wir die Schreibweise f, f', f'', f''' verwenden, und danach geht es mit $f^{(4)}, f^{(5)} \ldots$ weiter. Nicht geklammerte Exponenten an Funktionen können nach wie vor Potenzen sein.

Wir sehen uns noch einige Umkehrfunktionen an. Will man die Umkehrfunktion zur Sinusfunktion finden, so hat man erst einmal einen Bereich aufzusuchen, wo der Sinus streng monoton ist. Das ist etwa auf $[-\pi/2, \pi/2]$ der Fall, denn dort ist seine Ableitung, der Cosinus, positiv bis auf die Endpunkte. Auf diesem Intervall nimmt der Sinus jeden Wert zwischen -1 und 1 genau einmal an, und deshalb ist die Umkehrfunktion arcsin (**Arcussinus**) auf $[-1, 1]$ definiert und hat Werte in $[-\pi/2, \pi/2]$. Seine Ableitung bekommt man aus

$$\begin{aligned}\arcsin'(\sin(x)) &= \frac{1}{\sin'(x)} \\ &= \frac{1}{\cos(x)} \\ &= \frac{1}{\sqrt{1 - \sin^2(x)}} \\ \arcsin'(y) &= \frac{1}{\sqrt{1 - y^2}}.\end{aligned}$$

Wichtiger als der Arcussinus ist der **Arcustangens**, denn er wird für die Umformung von cartesischen Koordinaten (x, y) in Polarkoordinaten (r, φ)

gebraucht. Man sehe sich das nochmal auf Seite 104 an. Der **Tangens**

$$\tan(x) := \frac{\sin(x)}{\cos(x)}$$

ist wie der Sinus auf $[-\pi/2, \pi/2]$ streng monoton steigend, denn seine Ableitung ist auf $(-\pi/2, \pi/2)$ nach der Quotientenregel als

$$\begin{aligned}
\tan'(x) &= \left(\frac{\sin(x)}{\cos(x)}\right)' \\
&= \frac{\cos(x)\cos(x) - (-\sin(x))\sin(x)}{\cos^2(x)} \\
&= \frac{1}{\cos^2(x)}
\end{aligned}$$

positiv und am Rand sogar $+\infty$. Seine Umkehrfunktion, der Arcustangens, hat dann die Ableitung

$$\begin{aligned}
\arctan'(\tan(x)) &= \frac{1}{\frac{1}{\cos^2(x)}} \\
&= \frac{\cos^2(x)}{1} \\
&= \frac{\cos^2(x)}{\sin^2(x) + \cos^2(x)} \\
&= \frac{1}{\frac{\sin^2(x)}{\cos^2(x)} + 1} \\
&= \frac{1}{\tan^2(x) + 1} \\
\arctan'(y) &= \frac{1}{y^2 + 1}.
\end{aligned}$$

12.1.2 Symbolisches Differenzieren

Komplizierte Funktionen differenziert man besser symbolisch, um Fehlerquellen zu vermeiden. Dazu verwendet man Programme wie MuPAD, MAPLE, Mathematica oder MATLAB (das letztere hat eine auf MAPLE aufsetzende "Symbolic Math Toolbox"). Es folgen ein paar Beispiele, in denen wir die Funktion $f(x) = \exp(-2xy(x-z)^2)$ nach der Variablen x zweimal differenzieren. Die anderen vorkommenden Größen y und z sind bezüglich x konstant, werden also wie andere Konstanten, z.B. 2 behandelt. Man kann sich das Ergebnis gleich als gültigen C–Ausdruck zu Programmierzwecken ausgeben lassen.

In MAPLE:

```
> restart;with(CodeGeneration):
Warning, the protected name Matlab.....

> f:=exp(-2*x*y*(x-z)^2);
                              /               2\
                    f := exp\-2 x y (x - z) /
> df:=diff(f,x);
          /             2                  \    /               2\
   df := \-2 y (x - z)  - 4 x y (x - z)/ exp\-2 x y (x - z) /
> ddf:=diff(f,x,x);
                                      /              2\
   ddf := (-8 y (x - z) - 4 x y) exp\-2 x y (x - z) /

                                        2
        /             2                \     /               2\
      + \-2 y (x - z)  - 4 x y (x - z)/  exp\-2 x y (x - z) /
> simplify(ddf);
     /               2\   /                   4         3           2  2
4 exp\-2 x y (x - z) / y \-3 x + 2 z + 9 y x  - 24 y x  z + 22 y x  z

            3      4\
   - 8 y x z  + y z /
> C(simplify(ddf));
cg = 0.4e1 * exp(-0.2e1 * x * y * pow(x - z, 0.2e1))
* y * (-0.3e1 * x + 0.2e1 * z + 0.9e1 * y * pow(x, 0.4e1)
- 0.24e2 * y * pow(x, 0.3e1) * z + 0.22e2 * y * x * x * z * z
- 0.8e1 * y * x * pow(z, 0.3e1) + y * pow(z, 0.4e1));
```

In MuPAD:

```
>> f(x):=exp(-2*y*(x-z)^2)

                                          2
                             exp(- 2 y (x - z) )
>> df(x):=diff(f(x),x)

                                                 2
                    - 2 y (2 x - 2 z) exp(- 2 y (x - z) )
>> ddf(x):=diff(df(x),x)
```

```
          2             2                   2                          2
      4 y  (2 x - 2 z)  exp(- 2 y (x - z) ) - 4 y exp(- 2 y (x - z) )
>>
```

12.1.3 Eigenschaften differenzierbarer Funktionen

Zunächst setzt Differenzierbarkeit immer Stetigkeit voraus:

Theorem 12.9 *Ist eine reelle Funktion f in x differenzierbar, so ist sie in x auch stetig.*

Beweis: Wir benutzen Satz 12.2. Deshalb gibt es für jedes $\epsilon > 0$ ein $\delta > 0$, so dass aus $x, y \in I,\ |x - y| < \delta$ stets (12.3) folgt. Ist $(x_n)_n$ eine beliebige gegen x konvergente Folge, so ergibt sich

$$|f(x_n) - f(x)| \leq |f'(x)(x_n - x)| + \epsilon|x_n - x|$$

und weil die rechte Seite eine Nullfolge ist, konvergiert $(f(x_n))_n$ gegen $f(x)$. □

Leider müssen wir auch den Albtraum aller Schüler hier bringen:

Definition 12.10 *Es sei f eine auf I stetige reellwertige Funktion. Ein Punkt $x \in I$ heißt* **lokales Minimum** *bzw.* **lokales Maximum** *von f in I, wenn es eine Umgebung U von x gibt, so daß f in x sein Minimum bzw. Maximum bezüglich $U \cap I$ annimmt. Mit anderen Worten: es gibt ein $\delta > 0$, so daß*

$$f(x) = \min_{t \in I \cap [x-\delta, x+\delta]} f(t) \text{ oder } f(x) = \max_{t \in I \cap [x-\delta, x+\delta]} f(t)$$

gilt. Man nennt die Punkte, an denen lokale Minima und Maxima angenommen werden, auch lokale **Extremstellen**, *und die Werte dort sind* **Extremwerte**[1].

Achtung: x ist die Extrem**stelle** und $f(x)$ ist der Extrem**wert**!

Man mache sich klar, daß eine Extremstelle auch am Rand des Definitionsbereichs I liegen kann. Aber in diesem Fall ist der folgende Satz nicht anwendbar:

Theorem 12.11 *Es sei f eine auf $[a, b]$ stetige und in (a, b) differenzierbare reellwertige Funktion. Hat f in $x \in (a, b)$ ein lokales Minimum oder Maximum, so folgt $f'(x) = 0$.*

[1] `http://de.wikipedia.org/wiki/Extremwert`

Beweis: Wir benutzen Satz 12.8 auf Seite 306. Wenn $f'(x) \neq 0$ gelten würde, so wäre f in einer Umgebung von x streng monoton. Widerspruch. □

Wie man am Beispiel $f(x) = x^3$ in $x = 0$ sehen kann, gilt die Umkehrung **nicht**. Bevor wir die "Kurvendiskussion[1]" fortsetzen, bringen wir noch drei wichtige Verschärfungen des Zwischenwertsatzes für stetige Funktionen:

Theorem 12.12 *Es sei f eine auf $[a, b]$ stetige und in (a, b) differenzierbare reellwertige Funktion.*

- **Satz von Rolle**[2]*:*
  Gilt $f(a) = f(b)$, so gibt es ein $x \in (a, b)$ mit $f'(x) = 0$.
- **Mittelwertsatz**[3]*:*
  Es gibt ein $x \in (a, b)$ mit

$$f'(x) = \frac{f(b) - f(a)}{b - a}. \tag{12.13}$$

Beweis: Zum Beweis des Satzes von Rolle benutzen wir wieder Satz 12.8 auf Seite 306. Wenn die Aussage des Satzes von Rolle falsch wäre, d.h. $f'(x) \neq 0$ auf ganz (a, b) gelten würde, so müßte f auf (a, b) streng monoton sein. Weil f auf $[a, b]$ stetig ist, muss f also sein Minimum und Maximum in a und b annehmen, und diese beiden Funktionswerte müssen wegen der strengen Monotonie verschieden sein. Widerspruch.

Der Mittelwertsatz folgt dann durch Anwendung des Satzes von Rolle auf die Funktion

$$g(x) = f(x) - (x - a)\frac{f(b) - f(a)}{b - a}.$$

Auch g ist auf $[a, b]$ stetig und auf (a, b) differenzierbar. Weil g die Voraussetzungen des Satzes von Rolle erfüllt, denn es gilt $g(a) = g(b)$, gibt es ein $x \in (a, b)$ mit $g'(x) = 0$. Das liefert (12.13). □

Dieser Satz hat eine extrem wichtige Verallgemeinerung:

Theorem 12.14 *(Satz von* **Taylor**[4][5]*)*
Es sei f in einem Intervall $[a, b]$ mindestens n–mal stetig differenzierbar,

[1] http://de.wikipedia.org/wiki/Kurvendiskussion
[2] http://de.wikipedia.org/wiki/Satz_von_Rolle
[3] http://de.wikipedia.org/wiki/Mittelwertsatz_der_Differentialrechnung
[4] http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Taylor.html
[5] http://de.wikipedia.org/wiki/Taylorformel

und die $(n+1)$–te Ableitung existiere in (a,b). Dann gibt es zu jedem x und $x_0 \neq x$ in $[a,b]$ ein ξ echt zwischen x und x_0, so daß

$$f(x) - \sum_{j=0}^{n} f^{(j)}(x_0)\frac{(x-x_0)^j}{j!} = f^{(n+1)}(\xi)\frac{(x-x_0)^{n+1}}{(n+1)!} \tag{12.15}$$

gilt. Man nennt die Summe das **Taylorpolynom** *n–ten Grades zu f in x_0, und die rechte Seite heißt* **Restglied**.

Beweis: Seien $x \neq x_0$ fest gegeben. Wegen $x \neq x_0$ können wir eine Zahl r durch die Gleichung

$$f(x) - \sum_{j=0}^{n} f^{(j)}(x_0)\frac{(x-x_0)^j}{j!} =: r \cdot (x-x_0)^{n+1}$$

definieren. Wir nehmen vorübergehend zwei neue Variablen y und z, um die Funktion

$$g_z(y) := f(z) - \sum_{j=0}^{n} f^{(j)}(y)\frac{(z-y)^j}{j!} - r \cdot (z-y)^{n+1}$$

zu definieren. Deren Ableitung müssen wir etwas mühevoll ausrechnen:

$$\begin{aligned} g_z'(y) &= 0 - \sum_{j=0}^{n}\left(f^{(j+1)}(y)\frac{(z-y)^j}{j!} - f^{(j)}(y)\frac{(z-y)^{j-1}}{(j-1)!}\right) \\ &\quad + r(n+1)\cdot(z-y)^n \\ &= -f^{(n+1)}(y)\frac{(z-y)^n}{n!} + r(n+1)\cdot(z-y)^n \\ &= \frac{(z-y)^n}{n!}\big(r(n+1)! - f^{(n+1)}(y)\big), \end{aligned}$$

wobei wir benutzt haben, dass die Produktregel–Terme in der Summe sich gegenseitig fast alle wegheben.

Aber jetzt rechnen wir $g_x(x)$ und $g_x(x_0)$ aus:

$$\begin{aligned} g_x(x) &= f(x) - \sum_{j=0}^{n} f^{(j)}(x)\frac{(x-x)^j}{j!} - r\cdot(x-x)^{n+1} = 0 \\ g_x(x_0) &= f(x) - \sum_{j=0}^{n} f^{(j)}(x_0)\frac{(x-x_0)^j}{j!} - r\cdot(x-x_0)^{n+1} = 0. \end{aligned}$$

Die erste Gleichung gilt nach Definition von g_x, die zweite nach Definition von r. Nach dem Satz von Rolle, angewendet auf die Funktion $g_x(y)$, die in

$[a, b]$ stetig und in (a, b) differenzierbar ist, gibt es ein ξ echt zwischen x und x_0 mit $g'_x(\xi) = 0$. Das ergibt wegen $\xi \neq x$ und

$$g'_x(\xi) = \frac{(x-\xi)^n}{n!}\left(r(n+1)! - f^{(n+1)}(\xi)\right) = 0$$

die Gleichung

$$r(x-x_0)^{n+1} = f^{(n+1)}(\xi)\frac{(x-x_0)^{n+1}}{(n+1)!} = f(x) - \sum_{j=0}^{n} f^{(j)}(x_0)\frac{(x-x_0)^j}{j!}.$$

□

Man kann die Bedeutung des Taylorschen Satzes nicht hoch genug einschätzen. Er besagt, daß man eine $(n+1)$–mal differenzierbare Funktion f lokal in der Nähe eines "Entwicklungspunktes" x_0 durch ein Polynom $P_n(x)$ vom Grade n so ersetzen kann, daß der absolute Fehler in Punkten x nahe bei x_0 die genaue Form $f^{(n+1)}(\xi)\frac{(x-x_0)^{n+1}}{(n+1)!}$ hat.

Damit die Wichtigkeit glaubhaft wird, folgen sofort drei Beispiele.

Ist eine Funktion f an einer Stelle x_0 beliebig oft differenzierbar, so kann man fragen, ob die **Taylorreihe**[1]

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

für x aus einer Umgebung von x_0 oder sogar überall konvergiert. Bei der Exponentialfunktion ist das klar, weil

$$\exp(x) = \sum_{n=0}^{\infty} \frac{\exp^{(n)}(0)}{n!}x^n = \sum_{n=0}^{\infty} \frac{1}{n!}x^n$$

überall konvergiert. Aber auch die geometrische Reihe

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^j,\ |x| < 1$$

ist gleich ihrer Taylorreihe, weil man per Induktion die Ableitungen

$$\left(\frac{1}{1-x}\right)^{(n)} = \left((1-x)^{-1}\right)^{(n)} = n!(1-x)^{-(n+1)}$$

[1] `http://de.wikipedia.org/wiki/Taylorreihe`

mit dem Wert $n!$ an der Stelle $x = 0$ bekommt. Der Zusammenhang zwischen Taylorreihen und Potenzreihen wird unten noch etwas genauer zu behanden sein.

Wenn man für eine Funktion eine Methode hat, Funktionswerte auszurechnen, hat man noch lange kein Verfahren, die Ableitung exakt auszurechnen. Man kann aber durch zwei Funktionsaufrufe $f(x+h)$ und $f(x)$ den Differenzenquotienten

$$\frac{f(x+h) - f(x)}{h}$$

berechnen und ihn als Ersatz für die Ableitung $f'(x)$ ansehen. Wie groß ist der Fehler?

Der Taylorsche Satz liefert für zweimal differenzierbare Funktionen sofort

$$\left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| = \frac{1}{2} h |f''(\xi)|$$

mit einem ξ zwischen x und $x + h$. Das ist ziemlich schlecht, aber man kann nichts Besseres erwarten, weil Gleichheit gilt. Obendrein tritt schwere Auslöschung bei der Berechnung des Differenzenquotienten auf, und man kann gar keine sehr kleinen h verwenden. Das untersuchen wir später. Ist f dreimal stetig differenzierbar, so zieht man zwei Taylor–Entwicklungen um x voneinander ab, um nach einiger Rechnung

$$\left| \frac{f(x+h) - f(x-h)}{2h} - f'(x) \right| \leq \frac{1}{6} h^2 (|f'''(\xi)|$$

mi einem ξ zwischen $x - h$ und $x + h$ zu bekommen. Schon besser.

Jetzt wollen wir den Fehler grob abschätzen, der in Satz 11.38 auf Seite 296 bei der Ersetzung von Funktionen durch Geradenstücke auftrat. Wir halten zwei Punkte x und $x + h$ fest und ersetzen f dort durch

$$p(y) := f(x) + (y - x) \frac{f(x+h) - f(x)}{h} \text{ für alle } y \in [x, x+h].$$

Das verbindet die Punkte $(x, f(x))$ und $(x+h, f(x+h))$ durch eine Gerade. Wenn wir $f(x + h)$ und $f(y)$ beide um x entwickeln bis zur 2. Ableitung, bekommen wir

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2} f''(\xi_1) \\ f(y) &= f(x) + (y-x)f'(x) + \frac{(y-x)^2}{2} f''(\xi_2) \end{aligned}$$

und der Fehler wird

$$\begin{aligned} f(y)-p(y) &= f(x)+(y-x)f'(x)+\frac{(y-x)^2}{2}f''(\xi_2) \\ &\quad -\left(f(x)+(y-x)\left(f'(x)+\frac{h}{2}f''(\xi_1)\right)\right) \\ &= \frac{(y-x)^2}{2}f''(\xi_2)-(y-x)\frac{h}{2}f''(\xi_1) \\ |f(y)-p(y)| &\leq h^2\max(|f''(\xi_1)|,|f''(\xi_2)|). \end{aligned}$$

Diese Abschätzung kann man durch geschickteres Entwickeln verbessern, aber sie zeigt immerhin, daß die Ersetzung einer zweimal stetig differenzierbaren Funktion f durch einen stückweise linearen und stetigen Geradenzug f_Δ auf einer Zerlegung Δ höchstens den Fehler

$$\|f-f_\Delta\|_\infty \leq h(\Delta)^2\|f''\|_\infty$$

hat. Dieses Ergebnis ist deutlich besser als das von Satz 11.38 auf Seite 296, aber wir haben auch mehr vorausgesetzt.

Wenn man bei Untersuchungen von Funktionen auf die Situation $\frac{f(x)}{g(x)}=\frac{0}{0}$ stößt, hilft oft die folgende Anwendung des Mittelwertsatzes:

Theorem 12.16 *(Satz von* **de l'Hospital**[1]*)*
Es seien f und g Funktionen, die in $[a,b]$ differenzierbar seien, und die Ableitungen f' und g' seien noch in (a,b) stetig. Ferner sei $x\in(a,b)$ ein Punkt mit $f(x)=0,\ g(x)=0,\ g'(x)\neq 0$. Dann gilt:
Der Grenzwert $\lim_{y\to x}\frac{f(y)}{g(y)}$ existiert und ist gleich $\frac{f'(x)}{g'(x)}$.

Beweis: Wir wählen einen beliebigen Punkt $y\in(a,b)\setminus\{x\}$. Dann gilt nach dem Mittelwertsatz

$$\begin{aligned} \frac{f(y)-f(x)}{y-x} &= f'(\xi),\ \xi \text{ zwischen } y \text{ und } x \\ \frac{g(y)-g(x)}{y-x} &= g'(\eta),\ \eta \text{ zwischen } y \text{ und } x \\ \frac{f(y)}{g(y)}=\frac{f(y)-f(x)}{g(y)-g(x)} &= \frac{f'(\xi)}{g'(\eta)},\ \xi,\eta \text{ zwischen } y \text{ und } x \end{aligned}$$

und wenn wir als y beliebige Folgenglieder einer gegen x konvergenten Folge einsetzen, ergibt sich

$$\lim_{y\to x}\frac{f(y)}{g(y)}=\frac{f'(x)}{g'(x)}$$

[1] `http://de.wikipedia.org/wiki/Regel_von_L%E2%80%99Hospital`

weil die Punkte ξ und η zwischen y und x liegen und somit auch gegen x konvergieren, und weil f' und g' in x stetig sind. □

Typische Anwendungen, die man aber auch anders beweisen kann, sind

$$\lim_{x\to 1} \frac{x^n - 1}{x - 1} = n$$

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

Gilt bei speziellen Anwendungen nicht nur $\frac{f(x)}{g(x)} = \frac{0}{0}$, sondern auch noch $\frac{f'(x)}{g'(x)} = \frac{0}{0}$, so wendet man, wenn die Differenzierbarkeitsvoraussetzungen gegeben sind, den Satz von de l'Hospital einfach noch einmal an und geht zu $\frac{f''(x)}{g''(x)}$ über.

12.1.4 Differentiation von Potenzreihen

Jetzt gehen wir noch einmal etwas gründlicher auf Potenzreihen ein.

Theorem 12.17 *Es sei*

$$f(x) := \sum_{n=0}^{\infty} a_n x^n$$

eine für $|x| \leq R$ absolut konvergente Potenzreihe, d.h. es sei

$$\sum_{n=0}^{\infty} |a_n| R^n < \infty$$

konvergent. Dann ist f für alle x mit $|x| < R$ unendlich oft differenzierbar. Alle Ableitungen sind durch gliedweises Differenzieren der Potenzreihe von f selbst als Potenzreihe darstellbar, und alle diese Potenzreihen konvergieren absolut für $|x| < R$. Insbesondere stimmt f mit seiner Taylorreihe überein, d.h. es gilt $a_n = \frac{f^{(n)}(0)}{n!}$ für alle $n \geq 0$. Ferner stimmt f auch mit jeder seiner Taylorreihen in beliebigen Entwicklungspunkten x_0 mit $|x_0| < R$ überein, d.h. es gilt

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

und diese Reihe ist absolut konvergent für alle $x \in \mathbb{R}$ mit $|x - x_0| < R - |x_0|$.

Beweis: Wir bringen nur eine Skizze und untersuchen zuerst die Reihe

$$g(x) = \sum_{n=1}^{\infty} n a_n x^{n-1},$$

die f' darstellen könnte, weil sie durch gliedweises Differenzieren der Reihe von f entsteht. Die Konvergenz wird nur für $|x| < R$ behauptet, und deshalb setzen wir $|x| \le r < R$ mit einem geeigneten r voraus. Dann folgt

$$\begin{aligned}\sum_{n=1}^{\infty} n|a_n|r^{n-1} &= \sum_{n=1}^{\infty} n|a_n| \left(\frac{r}{R}\right)^{n-1} \frac{1}{R} R^n \\ &\le \sum_{n=1}^{K} n|a_n| \left(\frac{r}{R}\right)^{n-1} \frac{1}{R} R^{n-1} \\ &\quad + \sum_{n=K+1}^{\infty} |a_n| R^n < \infty\end{aligned}$$

wenn K so groß gewählt wird, daß für alle $n > K$ die Abschätzung

$$n \left(\frac{r}{R}\right)^{n-1} \frac{1}{R} \le 1$$

gilt. So ein K muss existieren, weil die Folge $\left(\frac{r}{R}\right)^{n-1}$ geometrisch–exponentiell gegen Null geht, die Folge n aber nur polynomial gegen Unendlich (siehe Satz 8.15 auf Seite 232). Also konvergiert die Reihe noch für alle $|x| \le r < R$, und weil $r < R$ beliebig war auch für alle $|x| < R$. Dasselbe gilt dann für alle höheren gliedweisen Ableitungen der Potenzreihe, und insbesondere konvergiert

$$\sum_{n=2}^{\infty} n(n-1)|a_n|r^{n-2} \text{ für alle } r < R, \tag{12.18}$$

was wir gleich brauchen werden.

Um zu beweisen, dass $g(x) = f'(x)$ gilt, sehen wir uns erst mit der Taylorformel die Entwicklung von $y \to y^n$ für $n \ge 2$ an:

$$x^n = x_0^n + (x - x_0) n x_0^{n-1} + n(n-1)(x - x_0)^2 \xi^{n-2}/2$$

mit ξ zwischen x_0 und x, und für $x := x_0 + h$ folgt

$$(x_0 + h)^n = x_0^n + nhx_0^{n-1} + n(n-1)h^2 \xi^{n-2}/2$$

mit ξ zwischen x_0 und $x_0 + h$. Das setzen wir in Partialsummen der Reihen ein:

$$\begin{aligned} & \sum_{n=0}^{N} (a_n(x_0+h)^n - a_n x_0^n - n a_n h x_0^{n-1}) \\ = \; & \frac{h^2}{2} \sum_{n=2}^{N} n(n-1) a_n \xi^{n-2}. \end{aligned}$$

Erfüllen x_0 und $x_0 + h$ die Ungleichung $|x| \leq r$, so auch ξ, und wir können (12.18) benutzen, um

$$\begin{aligned} & \sum_{n=0}^{N} \left| a_n(x_0+h)^n - a_n x_0^n - n a_n h x_0^{n-1} \right| \\ \leq \; & \frac{h^2}{2} \sum_{n=2}^{\infty} n(n-1) a_n r^{n-2} \end{aligned}$$

zu bekommen. Für festes $h > 0$ können wir den Grenzübergang $N \to \infty$ ausführen und bekommen

$$\begin{aligned} |f(x_0+h) - f(x_0) - h g(x_0)| &= \left| \sum_{n=0}^{\infty} a_n(x_0+h)^n - a_n x_0^n - n a_n h x_0^{n-1} \right| \\ &\leq \sum_{n=0}^{\infty} |a_n(x_0+h)^n - a_n x_0^n - n a_n h x_0^{n-1}| \\ &\leq \frac{h^2}{2} \sum_{n=2}^{\infty} n(n-1) a_n r^{n-2}. \end{aligned}$$

Nach Satz 12.2 auf Seite 304 folgt dann $f'(x_0) = g(x_0)$. Genauso argumentiert man für alle anderen Ableitungen.

Die Entwicklung in einem anderen Punkt x_0 hat die Form

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} a_n x^n \\ &= \sum_{n=0}^{\infty} a_n \sum_{j=0}^{n} \binom{n}{j} (x-x_0)^j x_0^{n-j} \\ &= \sum_{j=0}^{\infty} (x-x_0)^j \underbrace{\sum_{n=j}^{\infty} a_n \binom{n}{j} x_0^{n-j}}_{=:b_j}, \end{aligned}$$

wobei man

$$\begin{aligned} b_j &= \sum_{n=j}^{\infty} a_n \binom{n}{j} x_0^{n-j} \\ &= \frac{1}{j!} \sum_{n=j}^{\infty} a_n n \cdot (n-1) \cdots (n-j+1) x_0^{n-j} \\ &= \frac{1}{j!} f^{(j)}(x_0) \end{aligned}$$

hat, und dabei sind die auftretenden Reihen nach den obigen Argumenten absolut konvergent für $|x_0| < R$. Das setzt man nun in die Entwickung um x_0 ein und beweist die Konvergenz wie folgt:

$$\begin{aligned} & \sum_{j=0}^{\infty} |b_j||x - x_0|^j \\ \leq & \sum_{j=0}^{\infty} |x - x_0|^j \sum_{n=j}^{\infty} |a_n| \binom{n}{j} |x_0|^{n-j} \\ = & \sum_{n=0}^{\infty} |a_n| \sum_{j=0}^{n} \binom{n}{j} |x_0|^{n-j} |x - x_0|^j \\ = & \sum_{n=0}^{\infty} |a_n| (|x - x_0| + |x_0|)^n \\ \leq & \sum_{n=0}^{\infty} |a_n| R^n < \infty. \end{aligned}$$

□

Obwohl es auch anderweitig geht, kann man mit dem obigen Satz sofort einsehen, dass

$$\sum_{n=1}^{\infty} n x^{n-1} = \frac{1}{(1-x)^2} \text{ für alle } |x| < 1$$

gilt. Man bildet einfach die Ableitung von $1/(1-x)$, der geometrischen Reihe.

Eine weitere Anwendung betrifft die in der Signalverarbeitung wichtige sinc–Funktion $f(x) = \frac{\sin(x)}{x}$. Als Anwendung des Satzes von de l'Hospital haben wir schon gesehen, dass $f(0) = 1$ gesetzt werden kann, d.h. es liegt keine Singularität im Nullpunkt vor. Wenn wir die Reihe der Sinusfunktion verwenden, folgt

$$f(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n+1)!}$$

und diese Reihe ist überall absolut konvergent, weil für $|x| \leq R$ man gemäß

$$\begin{aligned}\sum_{n=0}^{\infty} \frac{|x|^{2n}}{(2n+1)!} &\leq \sum_{n=0}^{\infty} \frac{R^{2n}}{(2n+1)!} \\ &= \frac{1}{R} \sum_{n=0}^{\infty} \frac{R^{2n+1}}{(2n+1)!} \\ &\leq \frac{1}{R} \sum_{n=0}^{\infty} \frac{R^n}{n!} \\ &= \frac{1}{R} \exp(R)\end{aligned}$$

abschätzen kann. Also hat die Funktion beliebig viele Ableitungen, und diese Ableitungen kann man durch gliedweises Differenzieren der Potenzreihe erhalten. Nullstellen liegen vor bei $k \cdot \pi$ für alle $k \in \mathbb{Z} \setminus \{0\}$, und nach dem Satz von Rolle liegt zwischen je zwei solcher Nullstellen je eine Nullstelle der Ableitung. Dieses Argument läßt sich beliebig oft wiederholen, um unendlich viele Nullstellen jeder Ableitung zu erhalten. Die Funktion klingt gegen Unendlich ab.

12.1.5 Kurvendiskussion

Für die schon in der Schule übliche "Kurvendiskussion" brauchen wir noch ein paar Begriffe:

Definition 12.19 *Es sei f auf I eine differenzierbare Funktion.*

1. *Hat f in einem Punkt z eine Nullstelle mit dem Verhalten*

   $$f(x) = \Theta((x-z)^m) \text{ für } x \to z$$

   mit einem positiven m, so hat z die **Nullstellenordnung** *m.*

2. *Die Nullstellen von f' in I heißen* **kritische Punkte**.

3. *Ist f in einer Umgebung eines inneren Punktes x aus I sogar zweimal differenzierbar, und ist x Extremstelle von f', so wird x* **Wendepunkt**[1] *von f genannt.*

4. *Eine (nicht behebbare)* **Singularität**[2] *von f ist ein Punkt z außerhalb des Definitionsbereichs I von f, in den f nicht stetig fortsetzbar ist, für den es also Folgen $(x_n)_n$ in I gibt, die gegen z konvergieren, für die aber die Folgen $(f(x_n))_n$ nicht alle gegen einen endlichen festen gemeinsamen Grenzwert konvergieren.*

[1] `http://de.wikipedia.org/wiki/Wendepunkt`
[2] `http://de.wikipedia.org/wiki/Singularit%C3%A4t_%28Mathematik%29`

5. **Pole** *oder* **Polstellen**[1] *sind Singularitäten z, in deren Umgebung die Funktion f nicht beschränkt ist.*

6. *Gibt es ein positives m mit*

$$f(x) = \Theta((x-z)^{-m}) \text{ für } x \to z,$$

so hat z die **Polordnung** *m.*

7. *Die Funktion f ist auf I* **konvex** *bzw.* **konkav**, *wenn für alle $x, y \in I$ und alle* **Konvexkombinationen** *$z = \alpha \cdot x + (1-\alpha) \cdot y \in I,\ \alpha \in [0,1]$ die Ungleichungen*

$$\begin{array}{rcl} f(z) & \leq & \alpha \cdot f(x) + (1-\alpha) \cdot f(y) \text{ (Konvexität)} \\ f(z) & \geq & \alpha \cdot f(x) + (1-\alpha) \cdot f(y) \text{ (Konkavität)} \end{array}$$

gelten.

8. **Asymptoten**[2] *bzgl $x \to \pm\infty$ sind Funktionen g, für die*

$$|f(x_n) - g(x_n)| \to 0 \text{ für alle Folgen } (x_n)_n \text{ mit } x_n \to \infty \text{ oder } -\infty$$

gilt. Dabei sollten f und g für große Argumente, d.h. in Umgebungen von $\pm\infty$ definiert sein. Wie man an der Wikipedia sehen kann, ist der Begriff der Asymptote nicht so einfach mit voller Allgemeinheit definierbar.

Eine sogenannte **Kurvendiskussion**[3] (eine Kurve[4] ist mathematisch allerdings etwas Anderes) besteht darin, zu einer auf einer Teilmenge von $\mathbb{R}$ definierten Funktion f folgendes zu untersuchen (oder eine Teilmenge davon):

1. Definitionsbereich von f
2. Stetigkeit
3. Singularitäten, inklusive Bestimmung der Ordnung von Polen
4. Differenzierbarkeit (wie viele Ableitungen existieren wo, und sind sie noch stetig?)

[1] `http://de.wikipedia.org/wiki/Polstelle`
[2] `http://de.wikipedia.org/wiki/Asymptote`
[3] `http://de.wikipedia.org/wiki/Kurvendiskussion`
[4] `http://de.wikipedia.org/wiki/Kurve_(Mathematik)`

5. Nullstellen von f und den Ableitungen von f, inklusive Bestimmung der Nullstellenordnung
6. Monotonie
7. Extremwerte
8. Wendepunkte
9. Konkavität und Konvexität
10. Asymptoten

Man sollte eine Kurvendiskussion immer mit einer Zeichnung des Graphen beginnen.

Wir stellen nach den Begriffen auch noch einige Hilfsmittel zusammen.

Theorem 12.20 *Es sei f in allen Punkten des Definitionsbereiches zweimal stetig differenzierbar.*

1. *Lokale Extremstellen im* **Innern** *des Definitionsbereiches sind immer kritische Punkte (Satz 12.11). Die Umkehrung gilt nicht.*
2. *Kritische Punkte x im Innern des Definitionsbereiches, in denen die zweite Ableitung nicht verschwindet, sind lokale Extremstellen.*
3. *Kritische Punkte sind lokale Maxima, wenn $f''(x) < 0$ gilt, und lokale Minima, wenn $f''(x) < 0$ gilt.*
4. *In Umgebungen von Punkten x mit $f'(x) \neq 0$ ist die Funktion streng monoton (Satz 12.8).*
5. *In Umgebungen von Punkten x mit $f''(x) > 0$ bzw. $f''(x) < 0$ ist die Funktion konvex bzw. konkav.*
6. *Wendepunkte sind notwendig kritische Punkte von f'. Die Umkehrung gilt nicht.*
7. *Ein kritischer Punkt x von f' (d.h. eine Nullstelle von f'') ist Wendepunkt von f, wenn f in einer Umgebung von x dreimal stetig differenzierbar ist und $f'''(x)$ nicht verschwindet.*

Beweis: In Umgebungen von Punkten x, in denen $f''(x) \neq 0$ gilt, ist f' nach Satz 12.8 streng monoton. Im Falle $f''(x) > 0$ ist f' streng monoton wachsend, und deshalb gilt für alle nahe bei x gelegenen Punkte $x_- < x < x_+$ nach dem Mittelwertsatz die Ungleichung

$$\frac{f(x) - f(x_-)}{x - x_-} = f'(z_-) < f'(x) < \frac{f(x_+) - f(x)}{x_+ - x} = f'(z_+)$$

für gewisse Punkte $z_- \in (x_-, x),\ z_+ \in (x, x_+)$.

Gilt zusätzlich $f'(x) = 0$, so folgt daraus, dass f in x ein lokales Minimum hat, denn es gilt

$$f(x) - f(x_-) < 0 < f(x_+) - f(x).$$

Wenn wir wissen, daß zwischen zwei Punkten x_- und x_+ die zweite Ableitung noch positiv ist, können wir einen beliebigen Zwischenpunkt x als Konvexkombination der Punkte x_- und x_+ wählen:

$$x = \underbrace{\frac{x_+ - x}{x_+ - x_-}}_{=\alpha} x_- + \underbrace{\frac{x - x_-}{x_+ - x_-}}_{=1-\alpha} x_+.$$

Zum Beweis der Konvexität von f müssen wir dann

$$f(x) \leq \frac{x_+ - x}{x_+ - x_-} f(x_-) + \frac{x - x_-}{x_+ - x_-} f(x_+)$$

zeigen, und wir können dabei $x \in (x_-, x_+)$ voraussetzen. Mit diesen drei Punkten gehen wir nochmal in die erste Ungleichung und erhalten die (sogar "strikte") Konvexität aus

$$\begin{aligned}
0 &< f'(z_+) - f'(z_-) \\
0 &< \frac{f(x_-) - f(x)}{x - x_-} + \frac{f(x_+) - f(x)}{x_+ - x} \\
0 &< (x_+ - x)(f(x_-) - f(x)) + (x - x_-)(f(x_+) - f(x)) \\
(x_+ - x + x - x_-) f(x) &< (x_+ - x) f(x_-) + (x - x_-) f(x_+) \\
f(x) &< \frac{x_+ - x}{x_+ - x_-} f(x_-) + \frac{x - x_-}{x_+ - x_-} f(x_+).
\end{aligned}$$

Zur Kurvendiskussion werden diverse Beispiele in den Übungen behandelt.

Im obigen Satz bleiben einige Fälle offen, wenn mehrere aufeinanderfolgende Ableitungen an derselben Stelle verschwinden. Dazu hat man sich dann die Nullstellenordnung der entsprechenden Ableitung anzusehen und das folgende Ergebnis anzuwenden:

Theorem 12.21 *Ist f in einer Umgebung einer Nullstelle z der Ordnung $2m-1$ noch $(2m-1)$–mal stetig differenzierbar und gilt $f^{(2m-1)}(z) \neq 0$, so wechselt f in einer Umgebung von z das Vorzeichen, d.h. der Punkt z ist keine lokale Extremstelle. Im Falle einer geraden Nullstellenordnung wechselt f in einer Umgebung von z das Vorzeichen nicht, d.h. z ist lokale Extremstelle von f.*

Beweis: Hat f eine Nullstelle der genauen Ordnung $2m-1$ in z, so gilt

$$f(z) = f'(z) = \cdots = f^{(2m-2)}(z) = 0 \neq f^{(2m-1)}(z),$$

weil f lokal das Verhalten $(x-z)^{2m-1}$ hat. Nach dem Satz von Taylor gibt es zu jedem x aus einer Umgebung von z ein ξ zwischen x und z mit

$$f(x) = \frac{f^{(2m-1)}(\xi)(x-z)^{2m-1}}{(2m-1)!}.$$

Weil $f^{(2m-1)}(z)$ nicht verschwindet, ist also das Vorzeichenverhalten von $f(x)$ um z dasselbe wie das von $f^{(2m-1)}(z)(x-z)^{2m-1}$. Ganz analog argumentiert man für gerade Nullstellenordnungen. □

Den obigen Satz wendet man auf f' an, um Wendepunkte zu klären.

Die typischen logischen Fehler bei Kurvendiskussionen betreffen die Verwechslung von notwendigen und hinreichenden Bedingungen sowie das unsaubere Lösen von Gleichungen (z.B. Fehlen von Proben).

12.1.6 Differentialrechnung vektorwertiger Funktionen

Natürlich kann man auch Funktionen einer reellen Variablen betrachten, deren Bild in einem Vektorraum wie $\mathbb{R}^n$ liegt. Hier sind ein paar Beispiele:

Der Einheitskreisrand ist darstellbar als Bild von

$$f(\phi) := (\cos(\phi), \sin(\phi))^T \in \mathbb{R}^2,\ f\ :\ [0, 2\pi) \to \mathbb{R}^2.$$

Ein Strahl im $\mathbb{R}^3$, der von $x \in \mathbb{R}^3$ aus in Richtung $r \in \mathbb{R}^3 \setminus \{0\}$ geht, ist

$$f(t) := x + t \cdot r \in \mathbb{R}^3,\ ; f\ :\ [0, \infty) \to \mathbb{R}^3.$$

Die Funktion

$$f(t) := \left(\frac{1}{1+t^2}, \frac{t\sqrt{2}}{1+t^2}, \frac{t^2}{1+t^2} \right)^T$$

hat Werte auf der Oberfläche der dreidimensionalen Einheitskugel, weil die Quadratsumme der Bildkomponenten gleich 1 ist.

Jede Bildkomponente einer solchen Abbildung

$$f \; : \; I \to I\!R^n, \; t \mapsto (f_1(t), \ldots, f_n(t))^T \in I\!R^n \tag{12.22}$$

ist eine ganz "normale" reelle Funktion, und deshalb kann man, wenn die entsprechenden Ableitungen existieren, die Definition

$$f'(t) := (f_1'(t), \ldots, f_n'(t))^T \tag{12.23}$$

verwenden, d.h. die Ableitung eines Vektors ist der Vektor der Ableitungen.

Definition 12.24 *Eine Abbildung* $f \; : \; I\!R \supseteq I \to I\!R^n$ *mit* (12.22) *und* $n \geq 2$ *heißt* **Kurve**[1].
Sind alle Komponentenableitungen in $t \in I$ *definiert, so heißt* $f'(t) := (f_1'(t), \ldots, f_n'(t))^T$ *mit* (12.23) *der* **Tangentialvektor**[2] *an die Kurve* f *im Punkte* $f(t)$.
Der Skalar $t \in I$ *heißt* **Parameter** *der Kurve.*

Man berechne die Tangentialvektoren für die obigen Beispiele und mache sich den geometrischen Sachverhalt klar.

Eine schöne Sammlung von Kurven findet sich auf der mathematikgeschichtlichen website[3] der St. Andrews Universität in Schottland.

Hier beginnt das mathematische Spezialgebiet der **Differentialgeometrie**[4], in dem Kurven, Flächen und Körper im zwei– und dreidimensionalen Raum behandelt werden. Wegen der wichtigen Anwendungen im **Computer–Aided Design**[5] (**CAD**, rechnergestütztes Konstruieren) ist es auch für Informatik–Studierende wichtig. Man sehe sich geeignete websites zum Stichwort CAD an, u.a die Seiten des marktführenden Produkts CATIA und zugeordnete Demos[6]

Geschlossene geometrische Gebilde wie Kreise und Ellipsen kann man nicht komplett durch eine simple reellwertige Funktion im cartesischen Kordinatensystem beschreiben, weil man kein cartesisches Koordinatensystem so legen

[1] http://de.wikipedia.org/wiki/Kurve_(Mathematik)
[2] http://de.wikipedia.org/wiki/Tangente
[3] http://www-history.mcs.st-andrews.ac.uk/Curves/Curves.html
[4] http://de.wikipedia.org/wiki/Differentialgeometrie
[5] http://de.wikipedia.org/wiki/Computer_Aided_Design
[6] http://www.3ds.com/gallery/virtools-4-tour/

kann, daß man die Kurve eindeutig definieren kann. Man kann z.B. Polarkoordinaten nehmen, aber diese zeichnen den Nullpunkt auf besondere Weise aus und versagen für Gebilde, die auf Strahlen durch den Nullpunkt verlaufen. Deshalb kommt man nicht um den obigen Begriff herum, wenn man Geometrie auf Kurven treiben will.

Beschreibt $f(t)$ den Ort eines beweglichen Punktes in Abhängigkeit von der Zeit t, so ist $f'(t)$ der Geschwindigkeitsvektor und $f''(t)$ der Beschleunigungsvektor, sofern die Komponentenabbildungen zweimal differenzierbar sind. Die skalare Geschwindigkeit wäre dann $\|f'(t)\|_2$. Um deren Ableitung auszurechnen, könnte man die Beziehung

$$\|f'(t)\|_2 = \sqrt{\|f'(t)\|_2^2} = \sqrt{f'(t)^T f'(t)} \tag{12.25}$$

nutzen und die Kettenregel sowie die Ableitung eines Skalarproduktes heranziehen. Weil ein Skalarprodukt aber ein Spezialfall eines Matrizenproduktes ist, hilft folgendes Ergebnis:

Theorem 12.26 *Sind die Komponentenabbildungen der matrixwertigen Funktionen*

$$\begin{array}{rcll} A(t) & := & (a_{ij}(t)) \in I\!R^{\ell\times m} & \text{für alle } t \in [a,b] \\ B(t) & := & (b_{jk}(t)) \in I\!R^{m\times n} & \text{für alle } t \in [a,b] \end{array}$$

in $t \in [a,b]$ differenzierbar, so gilt die **Produktregel**

$$(A \cdot B)'(t) = A'(t) \cdot B(t) + A(t) \cdot B'(t),$$

wobei wie für Kurven die matrixwertigen Ableitungen als

$$A'(t) := (a'_{ij}(t)) \in I\!R^{\ell\times m}$$

definiert sind.

Beweis: Man differenziert die Komponenten des Matrizenprodukts $A \cdot B$ mit der Produktregel und bekommt

$$\begin{aligned} (e_i^T(A\cdot B)e_k)'(t) &= \left(\sum_{j=1}^{m} a_{ij}(t)\cdot b_{jk}(t)\right)' \\ &= \sum_{j=1}^{m} (a_{ij}(t)\cdot b_{jk}(t))' \\ &= \sum_{j=1}^{m} \left(a'_{ij}(t)\cdot b_{jk}(t) + a_{ij}(t)\cdot b'_{jk}(t)\right) \\ &= \sum_{j=1}^{m} a'_{ij}(t)\cdot b_{jk}(t) + \sum_{j=1}^{m} a_{ij}(t)\cdot b'_{jk}(t) \\ &= e_i^T A'(t)\cdot B(t)e_k + e_i^T A(t)\cdot B'(t))e_k \end{aligned}$$

für alle Komponenten der Bildmatrix. □

Damit können wir nun die Ableitung von (12.25) leicht ausrechnen:

$$(\|f'(t)\|_2)' = \frac{1}{2\|f'(t)\|_2}\left(f''(t)^T f'(t) + f'(t)^T f''(t)\right) = \frac{f'(t)^T f''(t)}{\|f'(t)\|_2}$$

sofern f in t zweimal differenzierbar ist und der Tangentialvektor $f'(t)$ nicht verschwindet.

Ein weitere Anwendung des obigen Satzes betrifft die Wirkung linearer Abbildungen auf Kurven. Ist A eine feste Matrix, die eine differenzierbare Kurve $x(t)$ linear transformiert, z.B. dreht oder spiegelt, so folgt

$$(A \cdot x(t))' = \underbrace{A'}_{=0} \cdot x(t) + A \cdot x'(t) = A \cdot x'(t),$$

d.h. der Tangentialvektor wird derselben Transformation unterworfen, zusammen mit allen eventuellen höheren Ableitungen.

Diese Dinge sehen abstrakt aus, sind aber sehr praktisch. Klären wir zur Illustration die Frage, wann ein Auto aus der Kurve fliegt. Offenbar dann, wenn die seitliche Beschleunigung zu groß wird, denn die seitlich wirkende Fliehkraft ist, wie alle dynamischen Kräfte, proportional zur Beschleunigung. Die "seitliche" Beschleunigung ist gegeben durch die Projektion des Beschleunigungsvektors $f''(t)$ auf die Richtung, die senkrecht zum Tangentialvektor $f'(t)$ ist und positives Skalarprodukt mit dem Beschleunigungsvektor hat. Die entscheidende Größe ist deshalb die Länge des Vektors

$$f''(t) - \frac{f''(t)^T f'(t)}{\|f'(t)\|_2^2} f'(t),$$

denn dieser steht auf dem Tangentialvektor senkrecht. Man sollte sich an dieser Stelle noch einmal ansehen, was in Abschnitt 5.4 über Orthogonalität und Projektoren gesagt wurde, denn oben kommt der Projektor

$$P(y) := \left(y^T \frac{f'(t)}{\|f'(t)\|_2}\right) \frac{f'(t)}{\|f'(t)\|_2}$$

auf die Richtung $\frac{f'(t)}{\|f'(t)\|_2}$ des Tangentialvektors $f'(t)$ vor. Der Zusammenhang zu Definition 5.22 auf Seite 176 besteht darin, dass man den Vektor u_1 so wählen muss, dass er die Länge 1 hat und tangential ist, d.h. man hat $u_1 = \frac{f'(t)}{\|f'(t)\|_2}$ zu setzen. Wegen Satz 5.23 auf Seite 176 steht dann $f''(t) - P(f''(t))$ auf dem Tangentialvektor senkrecht.

Wir haben oben eine Kurve als vektorwertige Abbildung f definiert, die auf einem Intervall I der reellen Zahlen definiert ist und Werte in einem $\mathbb{R}^n$ hat. Davon zu unterscheiden ist die Bildmenge $f(I) \subset \mathbb{R}^n$ als Punktmenge im $\mathbb{R}^n$. Beispielsweise kann man die Ideallinie einer Rennstrecke mit sehr unterschiedlichen Geschwindigkeiten befahren. Jede solche "Befahrung" ist eine Kurve in unserem Sinne, aber alle diese Kurven haben dieselbe Bildmenge, nämlich die Ideallinie der Rennstrecke. Die Richtung der jeweiligen Geschwindigkeitsvektoren ist im geometrischen Sinne immer tangential zur Ideallinie, aber die Länge des jeweiligen Geschwindigkeitsvektors hängt von der Fahrgeschwindigkeit ab.

Die einfachste Art, zwei Kurven mit gleicher Bildmenge zu konstruieren, besteht darin, eine Kurve $f : I \to \mathbb{R}^n$ mit einer bijektiven Abbildung $\varphi : J \mapsto I$ auf einem Intervall J durch

$$g(t) := f(\varphi(t)) \text{ für alle } t \in J$$

zu **reparametrisieren**. Man nennt dann φ eine **Umparametrisierung** von f. Die Bilder beider Kurven sind gleich. Ist die Umparametrisierung differenzierbar, so bekommen wir für die Tangentialvektoren in Punkten $s = \varphi(t)$ mit $s \in I$ und $t \in J$, die ja zu demselben Bildpunkt $f(s) = g(t)$ führen, einerseits $f'(s)$ und andererseits nach der (komponentenweisen) Kettenregel

$$g'(t) = f'(\varphi(t)) \cdot \varphi'(t) = f'(s) \cdot \varphi'(t).$$

Man sieht, dass sich die Tangentialvektoren im selben Bildpunkt $f(s) = g(t)$ um den Faktor $\varphi'(t)$ unterscheiden, aber dieselbe Richtung haben. Eine ideale Umparametrisierung wäre eine solche, die zu Tangentialvektoren der Länge Eins führen würde, d.h. es müßte überall $\varphi'(t) = 1/\|f'(\varphi(t))\|_2$ gelten, damit man $\|g'(t)\|_2 = 1$ hätte. Unter der Voraussetzung, daß f stetig differenzierbar ist und überall $f'(s) \neq 0$ gilt, klappt das, aber es erfordert entweder Kenntnisse über die Lösbarkeit von **Differentialgleichungen** oder den Begriff der **Bogenlänge** einer Kurve. In beiden Fällen müssen wir auf die Integralrechnung warten.

Beispiel: Man kann Teile des Einheitskreises als Bild von Kurven darstellen, die trigonometrische Funktionen und Wurzeln vermeiden, z.B. durch

$$s \mapsto \left(\frac{2s}{1+s^2}, \frac{1-s^2}{1+s^2} \right).$$

Die "nichtparametrische" Darstellung des Einheitskreises durch $y = \pm\sqrt{1-x^2}$ in cartesischen Koordinaten verwendet keine Kurve, kann aber je nach Vorzeichenwahl nur den oberen oder unteren Halbkreis beschreiben. Obendrein ist sie in $x = \pm 1$ nicht differenzierbar. Die Parametrisierung durch

$(\cos(\phi), \sin(\phi))$ hat den Vorteil, daß alle Tangentialvektoren die Länge 1 haben. Sie entspricht dem Durchfahren des Kreises mit konstanter Geschwindigkeit. Wie wir später sehen werden, ist der Kurvenparameter ϕ genau dann die Bogenlänge der Kurve, wenn die Tangentialvektoren alle die Länge 1 haben. Das ist verwandt mit der Frage, ob die durch Reihen definierten trigonometrischen Funktionen mit den geometrisch definierten übereinstimmen.

In der Praxis des Computer–Aided Design beschreibt man polynomiale Kurven nicht durch die Monombasis, d.h. man verwendet nicht den naheliegenden Ansatz

$$P(t)) := \sum_{j=0}^{n} a_j t^j,\ t \in I\!R,\ a_j \in I\!R^k,\ 0 \le j \le n.$$

Dabei mache man sich klar, dass die Koeffizienten a_j jetzt Vektoren sind, während die Monombasis die skalaren Koeffizienten t^j liefert. Das Ganze ist eine Linearkombination von parameterabhängigen Vektoren, also eine Kurve. Wie wir aus der Taylorformel wissen, haben dann die Koeffizienten a_j die Bedeutung $a_j = \frac{1}{j!}P^{(j)}(0)$, und das können wir auch für Kurven nachvollziehen. Entscheidend ist, daß diese Koeffizienten nur vom Verhalten von P in einer beliebig kleinen Umgebung des Nullpunkts abhängen. Das ist ein Nachteil gegenüber Koeffizienten, die man zu einer anderen Basis bildet, und die etwas über den Kurvenverlauf im Großen aussagen.

Die entscheidende Grundidee für eine praxisorientierte Wahl einer Basis der Polynome vom Grade $\le n$ ist, daß in einer Darstellung

$$P(t) := \sum_{j=0}^{n} b_j \beta_{j,n}(t),\ t \in I\!R,\ b_j \in I\!R^k,\ 0 \le j \le n \tag{12.27}$$

die Kurve in der konvexen Hülle der Koeffizienten $b_0, \ldots, b_n$ liegen sollte, d.h. der Punkt $P(t)$ muss Konvexkombination der Koffizienten $b_0, \ldots, b_n$ sein. Dazu braucht man (siehe Abschnitt 4.2.1 auf Seite 107) die Eigenschaften

$$\begin{aligned} \beta_{j,n}(t) &\in [0,1] \text{ für alle } j,\ 0 \le j \le n \\ \sum_{j=0}^{n} \beta_{j,n}(t) &= 1 \end{aligned}$$

auf einem geeigneten Definitionsintervall $I = [a, b]$. So eine Basis nennt man eine nichtnegative **Zerlegung der Eins**. Man konstruiert sie durch einen

simplen Trick:

$$\begin{aligned} 1 &= 1^n \\ &= \underbrace{\left(\frac{b-t}{b-a}+\frac{t-a}{b-a}\right)^n}_{=1} \\ &= \sum_{j=0}^{n} \underbrace{\binom{n}{j}\left(\frac{b-t}{b-a}\right)^{n-j}\left(\frac{t-a}{b-a}\right)^{j}}_{=:\beta_{j,n}(t)\geq 0} \end{aligned}$$

Theorem 12.28 *Zu festem $n \geq 0$ und festem Intervall $[a,b]$ werden die Funktionen*

$$\begin{aligned} \beta_{j,n}(t) &:= \binom{n}{j}\left(\frac{b-t}{b-a}\right)^{n-j}\left(\frac{t-a}{b-a}\right)^{j},\ 0 \leq j \leq n,\ t \in I\!R \\ \beta_{j,n}(t) &:= 0 \text{ sonst} \end{aligned}$$

Bernstein–Polynome *vom Grade n genannt. Sie haben die folgenden Eigenschaften:*

$$\begin{aligned} \beta_{j,n}(t) &\geq 0 \text{ für alle } t \in [a,b] \\ \sum_{j=0}^{n} \beta_{j,n}(t) &= 1 \text{ für alle } t \in I\!R \\ \beta_{j,n}(a) &= \delta_{j0},\ 0 \leq j \leq n \\ \beta_{j,n}(b) &= \delta_{jn},\ 0 \leq j \leq n \\ \beta_{j,n}(t) &= \frac{b-t}{b-a}\beta_{j,n-1}(t) + \frac{t-a}{b-a}\beta_{j-1,n-1}(t) \text{ für alle } n \geq 1,\ 0 \leq j \leq n \\ \beta'_{j,n}(t) &= \frac{n}{b-a}\left(\beta_{j-1,n-1}(t) - \beta_{j,n-1}(t)\right) \text{ für alle } n \geq 0,\ 0 \leq j \leq n. \end{aligned}$$

Der Beweis der letzten beiden Aussagen ist eine Übung im Rechnen mit Binomialkoeffizienten und wird hier unterdrückt. Die anderen Aussagen sind klar. □

Theorem 12.29 *Die Darstellung* (12.27) *einer polynomialen Kurve vom Grade n mit Werten im $I\!R^k$ wird als* **Bernstein–Bézier***–Darstellung oder* **Bezierkurve**[1] *bezeichnet. Die Vektoren $b_0, \ldots, b_n$ heißen* **Kontrollpunkte***, und ihre stückweise lineare Verbindung heißt* **Kontrollpolygon***. Es gilt:*

1. *Das Bild $P([a,b])$ der Kurve liegt in der konvexen Hülle der Kontrollpunkte.*

[1] `http://de.wikipedia.org/wiki/Bezierkurve`

2. *Anfangs– und Endpunkt des Kurvenstücks sind* $P(a) = b_0$ *und* $P(b) = b_n$*, also Anfangs– und Endpunkt des Kontrollpolygons.*

3. *Die Tangentialvektoren an den Enden der Kurve sind* $P'(a) = \frac{n}{b-a}(b_1 - b_0)$ *sowie* $P'(b) = \frac{n}{b-a}(b_n - b_{n-1})$*. Ihre Richtungen stimmen mit den Richtungen der Endstücke des Kontrollpolygons überein.*

4. *Die Berechnung eines Kurvenpunktes* $P(t)$ *mit* $t \in (a, b)$ *kann nach* **de Casteljau** *durch folgende rekursive affine Konstruktion geschehen:*

   - *Start: Definiere* $b_{j,n}(t) := b_j,\ 0 \le j \le n.$
   - *Gegeben* $b_{j,r}(t),\ 0 \le j \le r,\ 0 < r \le n.$
     Berechne

     $$b_{j,r-1}(t) := \frac{b-t}{b-a} b_{j,r}(t) + \frac{t-a}{b-a} b_{j+1,r}(t),\ 0 \le j \le r-1 \quad (12.30)$$

   - *Ende:* $b_{0,0}(t)$ *liefert* $P(t)$*.*

   Diese Konstruktion kann auch leicht zeichnerisch ausgeführt werden (siehe Vorlesung). Man teilt die Strecke zwischen $b_{j,r}(t)$ *und* $b_{j+1,r}(t)$ *durch den neuen Punkt* $b_{j,r-1}(t)$ *so, wie* t *das Intervall* $[a, b]$ *teilt.*

5. *Für alle* $r,\ 0 \le r \le n$ *gilt bei der obigen Konstruktion*

   $$P(t) = \sum_{j=0}^{r} b_{j,r}(t)\beta_{j,r}(t). \quad (12.31)$$

6. *Die konstruierten neuen Punkte liefern zwei neue Kontrollnetze:*

   - *Die Punkte* $b_{0,n}(t), b_{0,n-1}(t), \ldots, b_{0,0}(t)$ *sind das Kontrollnetz zu* P *über dem Teilintervall* $[a, t]$*.*
   - *Die Punkte* $b_{0,0}(t), b_{1,1}(t), \ldots, b_{n,n}(t)$ *sind das Kontrollnetz zu* P *über dem Teilintervall* $[t, b]$*.*

Beweis: Die ersten drei Aussagen folgen sofort aus dem vorigen Satz. Die dritte wird klar, wenn wir die vierte beweisen, denn (12.31) ist die Schleifeninvariante des de Casteljau–Verfahrens.

Dazu stellen wir fest, daß (12.27) und (12.30) im Falle $r = n$ übereinstimmen, und das ist der Beginn des de Casteljau–Verfahrens. Nehmen wir für eine

Rückwarts–Induktion an, dass (12.30) für ein $r, 0 < r \leq n$ gelte. Dann benutzen wir die Rekursionsformel der Bernstein–Polynome und erhalten

$$\begin{aligned}
P(t) &= \sum_{j=0}^{r} b_{j,r}(t)\beta_{j,r}(t) \\
&= \sum_{j=0}^{r} b_{j,r}(t) \left(\frac{b-t}{b-a}\beta_{j,r-1}(t) + \frac{t-a}{b-a}\beta_{j-1,r-1}(t) \right) \\
&= \sum_{j=0}^{r} \beta_{j,r-1}(t) b_{j,r}(t) \frac{b-t}{b-a} + \sum_{j=0}^{r} \beta_{j-1,r-1}(t) b_{j,r}(t) \frac{t-a}{b-a} \\
&= \sum_{j=0}^{r-1} \beta_{j,r-1}(t) b_{j,r}(t) \frac{b-t}{b-a} + \sum_{k=0}^{r-1} \beta_{k,r-1}(t) b_{k+1,r}(t) \frac{t-a}{b-a} \\
&= \sum_{j=0}^{r-1} \beta_{j,r-1}(t) \underbrace{\left(b_{j,r}(t) \frac{b-t}{b-a} + b_{j+1,r}(t) \frac{t-a}{b-a} \right)}_{=b_{j,r-1}(t)} \\
&= \sum_{j=0}^{r-1} \beta_{j,r-1}(t) b_{j,r-1}(t)
\end{aligned}$$

wobei man ausnutzen muss, dass alle "überschüssigen" Bernsteinpolynome als Null definiert sind. Damit ist die Rekursion bis herunter zu $r = 0$ bewiesen, und wegen $\beta_{0,0} = 1$ folgt das Endergebnis $P(t) = b_{0,0}(t)$ des Verfahrens von de Casteljau.

Die letzte Aussage wird **Subdivision**[1] genannt, weil sie ein Kontrollnetz in zwei neue Kontrollnetze "unterteilt". Der zugehörige Beweis ist schwieriger und wird übergangen. □

Anna Eggers hat dazu die Abbildung 10 hergestellt.

In der Praxis wendet man die Subdivision ein paarmal an und übergibt dann die entstandene Kette von Kontrollpolygonen an die Computergraphik, denn letztere erwartet Polygonzüge. Man kann beweisen, daß bei einem festen Polynom P, das man auf kleinen Intervallen $[a, b]$ durch Bernstein–Bezier–Polygonzüge ersetzt, ein absoluter Fehler der Größenordnung $\mathcal{O}((b-a)^2)$ für $b - a \to 0$ entsteht. In diesem Sinne ist die Subdivision konvergent, denn sie verkleinert die Intervalle. Jeder Subdivisionsschritt, der ein Intervall $[a, b]$ in $t = (a + b)/2$ halbiert, viertelt den absoluten Fehler, der zwischen Polynom und Kontrollpolygon besteht.

[1] `http://de.wikipedia.org/wiki/Subdivision_Surfaces`

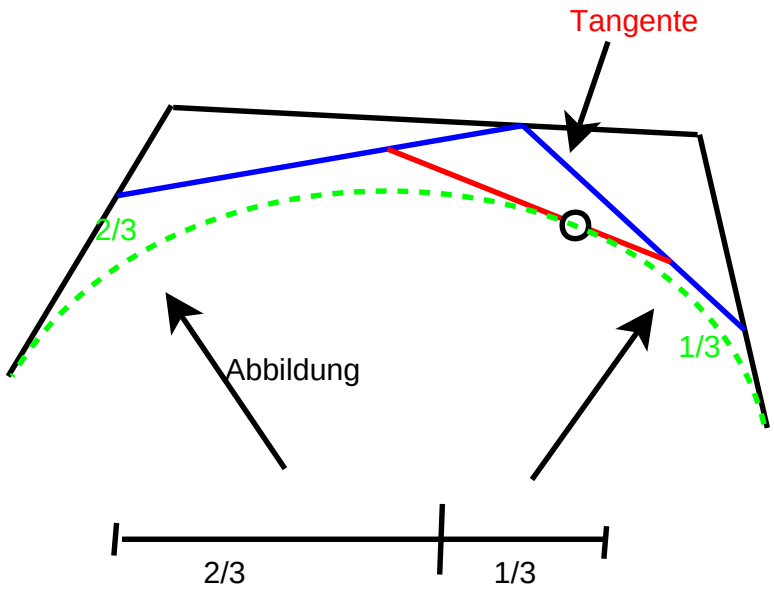


Abbildung 10: Casteljau–Verfahren und Subdivision

Weiteres sollte einer Vorlesung über Computer–Aided Design vorbehalten bleiben. Aber es macht Spaß, mit Bezierkurven zu spielen[1]

12.2 Multivariate Differentialrechnung

Funktionen mehrerer reeller Variablen können wir in der Form

$$f \; : \; I\!R^n \supseteq I \to I\!R, \; x = (x_1, \ldots, x_n)^T \mapsto f(x)$$

schreiben. Dabei sei I ein geeigneter Definitionsbereich. Wir werden hier keine exotischen Definitionsbereiche zulassen, sondern der Einfachheit halber immer annehmen, daß der Definitionsbereich ein n–faches cartesisches Produkt (siehe Def. 1.14 auf Seite 19) von Intervallen in $I\!R$ ist.. Beispiele sind $I = [-1, 1]^n$ oder $I = I\!R^n$ oder $I = [-1, 1] \times (0, \infty)$.

Natürlich kann man alle Variablen außer z.B. x_j festhalten, indem man sie einfach momentan als Konstanten ansieht, und dann nur nach x_j differenzieren. Als Funktion von x_j allein hat man eine reelle Funktion auf einem Intervall, und deshalb ist der enstprechende Ableitungsbegriff definiert. Die entstehende **partielle Ableitung**[2] wird mit $\frac{\partial f}{\partial x_j}$ oder kurz auch f_{x_j} bezeich-

[1] http://www.fh-friedberg.de/users/jingo/mathematics/bezier/bezier.html
[2] http://de.wikipedia.org/wiki/Partielle_Ableitung

net. Der **Zeilen**vektor

$$\nabla f := \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

wird **Gradient**[1] von f genannt, wenn alle partiellen Ableitungen von f existieren. Das Symbol ∇[2] wird *Nabla* genannt. Auch hier sollte man vorsichtig zwischen der Rolle von x_j als freie Variable und als Koordinate eines Auswertungspunkts unterscheiden.

Definition 12.32 *Eine Abbildung $f : \mathbb{R}^n \supseteq I \to \mathbb{R}$ heißt (partiell)* **differenzierbar** *in $x \in I$, wenn alle partiellen Ableitungen $\frac{\partial f}{\partial x_j}$, $1 \leq j \leq n$ in x existieren. Sie heißt (partiell) differenzierbar in I, wenn sie in allen Punkten $x \in I$ (partiell) differenzierbar ist.*

Hier ein Beispiel: Die Funktion

$$f(x, y) = \exp(x)\cos(y) + 3 * x^2 y \tag{12.33}$$

hat den Gradienten

$$\nabla f = \left(\exp(x)\cos(y) + 6 * xy, -\exp(x)\sin(y) + 3 * x^2 \right).$$

Wir beschränken uns hier der Einfachheit halber auf partielle Differenzierbarkeit und ignorieren die "vollständige" oder "totale" Differenzierbarkeit, die man in den mathematischen Anfängervorlesungen an dieser Stelle finden würde. Das Manko gleichen wir aus, indem wir fast immer Stetigkeit der partiellen Ableitungen fordern, z.B. auch in

Theorem 12.34 *(***Multivariate Kettenregel***) Es sei eine stetig partiell differenzierbare Funktion $f : I \supseteq R^n \to \mathbb{R}$ auf einer stetig differenzierbaren Kurve $x : [a, b] \to I \subseteq \mathbb{R}^n$ auszuwerten und entlang der Kurve zu differenzieren. Die Funktion $g(t) := F(x(t)) : [a, b] \to \mathbb{R}$ hat dann die Ableitung*

$$g'(t) = \sum_{j=1}^{n} \left(\frac{\partial f}{\partial x_j}(x(t)) \right) \cdot x_j'(t) = (\nabla f)(x(t)) \cdot x'(t)$$

deren Form genau der univariaten Kettenregel entspricht.

[1] `http://de.wikipedia.org/wiki/Gradient_(Mathematik)`
[2] `http://de.wikipedia.org/wiki/Nabla`

Den **Beweis** führen wir nur für $n = 2$ aus und schreiben, sofern die auftretenden Nenner nicht Null sind, unter Benutzung des univariaten Mittelwertsatzes

$$\begin{aligned}
& \frac{1}{h}(g(t+h) - g(t)) \\
= \; & \frac{1}{h}(f(x_1(t+h), x_2(t+h)) - f(x_1(t), x_2(t))) \\
= \; & \frac{1}{h}\left(f(x_1(t+h), x_2(t+h)) - f(x_1(t), x_2(t+h))\right) \\
& + \frac{1}{h}\left(f(x_1(t), x_2(t+h)) - f(x_1(t), x_2(t))\right) \\
= \; & \frac{f(x_1(t+h), x_2(t+h)) - f(x_1(t), x_2(t+h))}{x_1(t+h) - x_1(t)} \cdot \frac{x_1(t+h) - x_1(t)}{h} \\
& + \frac{f(x_1(t), x_2(t+h)) - f(x_1(t), x_2(t))}{x_2(t+h) - x_2(t)} \cdot \frac{x_2(t+h) - x_2(t)}{h} \\
= \; & \frac{\partial f}{\partial x_1}(\xi_1(t,h), x_2(t+h)) \cdot x_1'(\tau_1(t,h)) \\
& + \frac{\partial f}{\partial x_2}(x_1(t), \xi_2(t,h)) \cdot x_2'(\tau_2(t,h))
\end{aligned}$$

mit $\xi_1(t,h)$ zwischen $x_1(t+h)$ und $x_1(t)$ sowie $\xi_2(t,h)$ zwischen $x_2(t+h)$ und $x_2(t)$, ferner auch $\tau_1(t,h),\ \tau_2(t,h)$ zwischen $t+h$ und t. Jetzt kann man den Grenzübergang $h \to 0$ ausführen und erhält die Behauptung. $\square$.

Die Voraussetzungen dieses Satzes lassen sich abschwächen, aber das soll uns hier nicht interessieren. Es ist aber darauf hinzuweisen, daß im Gegensatz zur Differentialrechnung mit nur einer Variablen die punktweise Existenz partieller Ableitungen nicht die Stetigkeit der Funktion impliziert.

Wenn wir die Funktion (12.33) auf einem Kreis

$$t \mapsto (x(t), y(t))^T = (\cos(t), \sin(t))^T$$

auswerten wollen, bekommen wir die univariate Funktion

$$g(t) := f(x(t), y(t)) = \exp(\cos(t)) \cos(\sin(t)) + 3 * \cos^2(t) \sin(t),$$

und man kann g natürlich direkt nach t differenzieren. Nach Satz 12.34 kann

man aber auch die Kettenregel in der Form

$$\begin{aligned} g'(t) &= \nabla f(x(t), y(t)) \cdot \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} \\ &= (\exp(x(t))\cos(y(t)) + 6 * x(t)y(t), -\exp(x(t))\sin(y(t)) + 3 * x(t)^2) \cdot \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} \\ &= (\exp(\cos(t))\cos(\sin(t)) + 6 * \cos(t)\sin(t))\,(-\sin(t)) \\ &\quad + (-\exp(\cos(t))\sin(\sin(t)) + 3 * \cos(t)^2)\cos(t) \end{aligned}$$

anwenden und sollte das gleiche Ergebnis bekommen.

Wir betrachten den wichtigsten Spezialfall von Satz 12.34: die Differentiation von f entlang einer Geraden $x(t) = y + t \cdot r$ mit $y, r \in \mathbb{R}^n$. Dann gilt $g(t) := f(y + t \cdot r)$ mit

$$\begin{aligned} g'(t) &= \sum_{j=1}^{n} \left(\frac{\partial f}{\partial x_j}(y + t \cdot r) \right) \cdot x_j'(t) \\ &= \sum_{j=1}^{n} \left(\frac{\partial f}{\partial x_j}(y + t \cdot r) \right) \cdot r_j \\ &= ((\nabla f)(y + t \cdot r))\, r. \end{aligned}$$

Speziell: Die **Richtungsableitung** im Punkte y in Richtung r ist

$$g'(0) = ((\nabla f)(y)) \cdot r.$$

Man mache sich dies geometrisch klar. Dabei ist es hilfreich, sich $f(y)$ als die "Höhe" eines Gebirges über einem Punkt y vorzustellen. Wenn man über y in der Höhe $f(y)$ steht und in Richtung r einen kleinen Schritt macht, gibt die Richtungsableitung in Richtung r die Steigung des "Gebirges" in Richtung r an. Fragen wir nach der Richtung des steilsten Anstiegs oder Abstiegs, so müssen wir die Richtung r durch $\|r\|_2 = 1$ normieren und das Maximum bzw. Minimum von $((\nabla f)(y)) \cdot r$ als Funktion von r ausrechnen. Wegen der Cauchy–Schwarz–Ungleichung gilt

$$|\,((\nabla f)(y)) \cdot r| \leq \|(\nabla f)(y)\|_2$$

mit Gleichheit genau dann, wenn r und $(\nabla f)(y)$ parallel sind.

Theorem 12.35 *Der Gradient gibt eine Richtung des steilsten Anstiegs an, der negative Gradient eine Richtung des steilsten Abstiegs.* □

Wir sehen uns mal die Menge

$$\{y \in D \subseteq \mathbb{R}^n \;:\; f(y) = c\}$$

aller Punkte $y \in D$ an, für die f einen festen Wert c hat. Bei nur zwei Variablen ist das anschaulich eine **Höhenlinie**, im allgemeinen eine **Niveaumenge**[1]. Innerhalb dieser Niveaumenge nehmen wir mal die Existenz des Bildes einer glatten Kurve $y(t)$ mit Definitionsbereich $T \subset \mathbb{R}$ an. Dann ist die Funktion $f(y(t)) = c$ konstant, und ihre Ableitung ist

$$0 = (\nabla f)(y(t)) \cdot y'(t).$$

Theorem 12.36 *Der Gradient $(\nabla f)(y)$ steht senkrecht auf allen Tangentialvektoren von glatten Kurven, die durch y gehen und in der Niveaumenge von y liegen.*

Die beiden vorigen Sätze gehören zum Grundwissen aller Bergwanderer. Der steilste Abstieg oder Anstieg von einer Höhenlinie erfolgt immer in eine Richtung, die senkrecht zur Höhenlinie ist.

Im Mehrdimensionalen gibt es keinen "echten" Mittelwertsatz der Differentialrechnung und keinen Zwischenwertsatz für stetige Funktionen, weil diese auf Ordnung basieren. Als Ersatz kann man aber einen Mittelwertsatz bzw. einen Zwischenwertsatz entlang einer Geraden nehmen. Wie oben betrachtet man $g(t) := f(y+t\cdot r)$. Dann gilt nach dem eindimensionalen Mittelwertsatz

$$\frac{g(t) - g(0)}{t} = g'(\tau),\ \tau \in (0,t)$$

$$f(y + t \cdot r) - f(y) = t \cdot ((\nabla f)(y + \tau \cdot r))\, r$$

mit $\tau \in (0,t)$. Setzt man $z := y + r,\ t = 1$, so besagt dies

$$f(z) - f(y) = ((\nabla f)((1-\tau)y + \tau z))\,(z - y).$$

Theorem 12.37 *Sei $f : \mathbb{R}^n \supset D \to \mathbb{R}$ eine stetig partiell differenzierbare Funktion, und seien y und z zwei Punkte des Definitionsbereichs D von f, deren Verbindungsstrecke ganz in D liegt. Dann gibt es einen Punkt ξ auf dieser Strecke, so dass*

$$f(z) - f(y) = ((\nabla f)(\xi))\,(z - y)$$

gilt. □

[1] `http://de.wikipedia.org/wiki/Niveaumenge`

Mehrfache partielle Ableitungen erfordern eine besondere Notation und etwas Vorsicht. Differenziert man erst nach der freien Variablen x_j und dann nach x_k, so schreibt man

$$\frac{\partial}{\partial x_k}\left(\frac{\partial f}{\partial x_j}\right) =: \frac{\partial^2 f}{\partial x_k \partial x_j}.$$

Die Reihenfolge der Ableitungen ist im allgemeinen nicht vertauschbar, die man am Beispiel

$$f(x,y) := \begin{cases} xy\frac{x^2-y^2}{x^2+y^2} & (x,y) \neq (0,0) \\ \\ 0 & (x,y) = (0,0) \end{cases}$$

sieht. Es ist eine gute Übung, die Gleichungen

$$\begin{aligned} \frac{\partial f}{\partial x}(0,y) &= -y \quad \text{für alle } (x,y) \in I\!R^2 \\ \frac{\partial f}{\partial y}(x,0) &= x \quad \text{für alle } (x,y) \in I\!R^2 \end{aligned}$$

unter Anwendung des Satzes von de l'Hospital nachzurechnen. Sie ergeben

$$\frac{\partial^2 f}{\partial x \partial y}(0,0) = 1 \neq -1 = \frac{\partial^2 f}{\partial y \partial x}(0,0),$$

aber in allen anderen Punkten gilt Gleichheit. Somit kann man die zweifachen partiellen Ableitungen

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{(x^2-y^2)(x^4+10x^2y^2+y^4)}{(x^2+y^2)^3} = \frac{\partial^2 f}{\partial y \partial x}$$

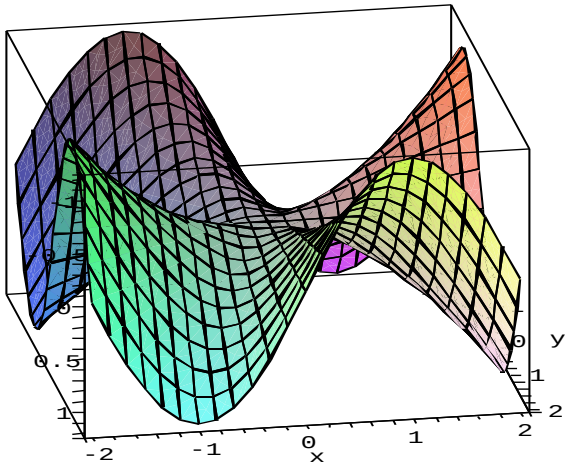
die überall bis auf $(0,0)$ existieren, stetig sind und übereinstimmen, nicht in den Punkt $(0,0)$ hinein stetig fortsetzen. Die Stetigkeit dieser **gemischten** partiellen Ableitung ist also eine unverzichtbare Voraussetzung, wenn man Vertauschbarkeit der beiden Differentiationen beweisen will.

Wir fragen MAPLE, wie diese Funktion aussieht. Die Funktion selbst sieht ganz harmlos aus, aber die zweifache gemischte partielle Ableitung hat es in sich.

```
> restart;
> f:=x*y*(x^2-y^2)/(x^2+y^2);
```

$$f := \frac{xy\,(x^2-y^2)}{x^2+y^2}$$

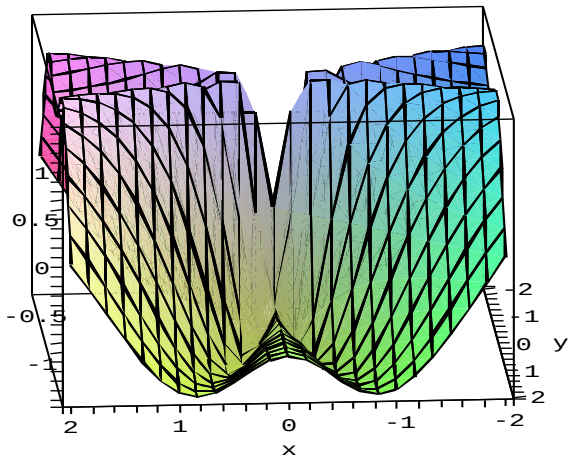
```
> plot3d(f,x=-2..2,y=-2..2,axes=boxed);
```




```
> g:=simplify(diff(f,x,y));
```

$$g := \frac{x^6 + 9\,x^4y^2 - 9\,x^2y^4 - y^6}{\left(x^2 + y^2\right)^3}$$

```
> plot3d(g,x=-2..2,y=-2..2,axes=boxed);
```



Unter Zusatzvoraussetzungen sind mehrfache partielle Ableitungen aber

durchaus in ihrer Reihenfolge vertauschbar. Im Beispiel der Funktion (12.33) bekommt man

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = \exp(x)\sin(y) + 6x.$$

Die nötigen Voraussetzungen bringt

Theorem 12.38 *Es sei f in einer Umgebung eines Punktes $(x_0, y_0) \in \mathbb{R}^2$ definiert und in der ganzen Umgebung mögen die partiellen Ableitungen*

$$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial^2 f}{\partial x \partial y}, \frac{\partial^2 f}{\partial y \partial x}$$

existieren und

$$\frac{\partial f}{\partial x}, \frac{\partial^2 f}{\partial y \partial x} \text{ oder } \frac{\partial f}{\partial y}, \frac{\partial^2 f}{\partial x \partial y}$$

mögen dort stetig sein. Dann gilt dort

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

Beweis: Wir werden Stetigkeit von

$$\frac{\partial f}{\partial x}, \frac{\partial^2 f}{\partial y \partial x}$$

voraussetzen. Mit den Abkürzungen

$$f_y := \frac{\partial f}{\partial y},\ f_x := \frac{\partial f}{\partial x},\ f_{xy} := \frac{\partial^2 f}{\partial x \partial y},\ f_{yx} := \frac{\partial^2 f}{\partial y \partial x}$$

sehen wir uns erst einmal

$$\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) = \lim_{h \to 0} \underbrace{\frac{f_y(x_0 + h, y_0) - f_y(x_0, y_0))}{h}}_{=:D(h)}$$

an. Die beiden im Limesausdruck auftretenden Größen haben die Form

$$\begin{aligned} f_y(x_0 + h, y_0) &= \lim_{k \to 0} \frac{f(x_0 + h, y_0 + k) - f(x_0 + h, y_0)}{k} \\ f_y(x_0, y_0) &= \lim_{k \to 0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k} \end{aligned}$$

und wir können deshalb schreiben

$$\begin{aligned} D(h) &= \lim_{k \to 0} \frac{(f(x_0 + h, y_0 + k) - f(x_0 + h, y_0)) - (f(x_0, y_0 + k) - f(x_0, y_0))}{hk} \\ &=: \lim_{k \to 0} \frac{Z(h, k)}{hk}. \end{aligned}$$

Jetzt definieren wir die um x_0 stetig differenzierbare Hilfsfunktion

$$\begin{aligned} g_k(x) &:= f(x, y_0 + k) - f(x, y_0) \text{ mit} \\ g_k'(x) &= f_x(x, y_0 + k) - f_x(x, y_0) \end{aligned}$$

und wenden den Mittelwertsatz zweimal an:

$$\begin{aligned} Z(h,k) &= g_k(x_0 + h) - g_k(x_0) \\ &= h \cdot g_k'(\xi(h,k)) \\ &= h \cdot (f_x(\xi(h,k), y_0 + k) - f_x(\xi(h,k), y_0)) \\ &= h \cdot k \cdot f_{yx}(\xi(h,k), \eta(h,k))) \end{aligned}$$

mit $\xi(h,k)$ zwischen x_0 und x_0+h sowie $\eta(h,k)$ zwischen y_0 und y_0+k. Dazu braucht man die nach y um y_0 stetig differenzierbare Hilfsfunktion

$$u_{h,k}(y) := f_x(\xi(h,k), y).$$

Es folgt

$$\begin{aligned} D(h) &= \lim_{k \to 0} \frac{Z(h,k)}{hk} = h \cdot f_{yx}(\xi(h,k), y_0)) \\ f_{xy}(x_0, y_0) &= \lim_{h \to 0} \frac{D(h)}{h} = f_{yx}(x_0, y_0) \end{aligned}$$

wegen Stetigkeit von f_{yx}. □

Jetzt wird untersucht, wann ein $y \in I\!R^n$ lokales Minimum einer Funktion $f : I\!R^n \to I\!R$ ist. Das soll natürlich heißen, daß

$$f(x) \geq f(y) \text{ für alle } x \text{ aus einer Umgebung von } y$$

gilt, also z.B. für alle x mit $\|x - y\| < \delta$ für ein $\delta > 0$.

Wir setzen mindestens zweifache stetige partielle Differenzierbarkeit voraus. Der eindimensionale Satz von Taylor liefert

$$f(y + t \cdot r) - f(y) = t \cdot (\nabla f)(y) r + \frac{t^2}{2} \sum_{j,k=1}^{n} r_j r_k \frac{\partial^2 f}{\partial x_k \partial x_j}(y + \tau(y,t) \cdot r)$$

mit $\tau \in (0, t)$.

Ist y ein lokales Minimum, so muss die Funktion

$$z(t) := f(y + t \cdot r) - f(y) = t \cdot (\nabla f)(y) r + \frac{t^2}{2} \sum_{j,k=1}^{n} r_j r_k \frac{\partial^2 f}{\partial x_k \partial x_j}(y + \tau(t) \cdot r)$$

in $t = 0$ ein lokales Minimum haben. Ist $t = 0$ ein innerer Punkt des Definitionsbereiches von z, muss die Ableitung von z dort gleich 0 sein, d.h.

$$(\nabla f)(y)r = 0 \text{ für alle } r$$

also $(\nabla f)(y) = 0$. Das ist die notwendige Bedingung für ein lokales Extremum in einem inneren Punkt des Definitionsbereichs. Ein Punkt y mit $(\nabla f)(y) = 0$ heißt **kritischer Punkt** von f. Er ist nicht unbedingt ein lokales Minimum oder Maximum, aber alle lokalen Minima und Maxima im Inneren sind kritische Punkte.

Definition 12.39 *Die* **Hessesche Matrix**[1] *einer zweimal partiell differenzierbaren Funktion f von n Variablen $x_1, \ldots, x_n$ ist*

$$H_f(z) := \left(\frac{\partial^2 f}{\partial x_k \partial x_j}(z)\right)_{1 \le j,k \le n}$$

und sie ist nach Satz 12.38 symmetrisch, wenn die ersten und zweiten partiellen Ableitungen von f in einer Umgebung von z stetig sind.

Es gelte in einem inneren Punkt y des Definitionsbereichs von f die notwendige Bedingung $(\nabla f)(y) = 0$ für ein lokales Minimum. Dann folgt

$$\begin{aligned} z(t) &= f(y + t \cdot r) - f(y) \\ &= \frac{t^2}{2} \sum_{j,k=1}^{n} r_j r_k \frac{\partial^2 f}{\partial x_k \partial x_j}(y + \tau(t) \cdot r) \\ &= \frac{t^2}{2} r^T H_f(y + \tau(t) \cdot r) r. \end{aligned}$$

Ist $H_f(y + \tau(t) \cdot r)$ positiv semidefinit, so folgt $r^T H_f(y + \tau(t) \cdot r)r \ge 0$ sowie $f(y + t \cdot r) \ge f(y)$ und man hat ein lokales Minimum. Die "richtigen" hinreichenden Bedingungen sind also $(\nabla f)(y) = 0$ und positive Semi-Definitheit von $H_f(z)$ in einer Umgebung von y, z.B. für alle z mit $\|z - y\| < \delta$.

Ist f zweimal stetig differenzierbar, so ist $z \mapsto H_f(z)$ eine stetige matrixwertige multivariate Funktion. Wenn die zweiten Ableitungen stetig sind, liegt die Matrix $H_f(z)$ nahe bei $H_f(y)$, sofern z nahe bei y liegt. Dann kann man aus der positiven Definitheit von $H_f(y)$ auf die von $H_f(z)$ schließen:

Theorem 12.40 *(Störungssatz für positiv definite Matrizen)*
Alle Matrizen aus einer einer genügend kleinen Umgebung einer positiv definiten symmetrischen Matrix sind positiv definit.

[1] `http://de.wikipedia.org/wiki/Hessematrix`

Deshalb reicht es, als hinreichende Bedingungen für ein lokales Minimum y zu fordern

1. $(\nabla f)(y) = 0$,
2. $H_f(y)$ ist positiv definit,
3. Stetigkeit aller Ableitungen inklusive der zweiten.

Hier noch eine Skizze des Beweises zum Störungssatz. Ist A positiv definit, so ist A nach Satz 9.3 auf Seite 245 mit einer Orthogonalmatrix V in eine Diagonalmatrix D als $A = VDV^T$ transformierbar. Die Diagonale von D enthält die Eigenwerte $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$ von A. Liegt eine symmetrische Matrix B nahe genug bei A, im Sinne von $\|A - B\|_{2,2} < \delta$, so folgt wegen der Invarianz der euklidischen Norm unter orthogonalen Matrizen

$$\|A - B\|_{2,2} = \|V^T(A - B)V\|_{2,2} = \|D - V^T BV\|_{2,2} < \delta.$$

Also weicht die symmetrische Matrix $V^T BV$ nur wenig von der Diagonalmatrix D ab, und der Satz 9.14 von Gerschgorin zeigt zusammen mit Satz 9.2 die positive Definitheit von $V^T BV$ für hinreichend kleine δ. Die Matrix B ist dann auch positiv definit, weil sie dieselben positiven Eigenwerte wie $V^T BV$ hat.

12.2.1 Vektorfelder

Wir betrachten nun vektorwertige Funktionen mehrerer Variablen, z.B. die Abbildungen $x \mapsto (\nabla f)(x)$ oder $x \mapsto H_f(x)$ oben. In diesem Abschnitt schreiben wir dann $F(x) = (F_1(x), \ldots, F_m(x))^T$ mit $x = (x_1, \ldots, x_n)^T$ als Abbildung $\mathbb{R}^n \to \mathbb{R}^m$ oder auf einer Teilmenge D von $\mathbb{R}^n$, die wie bisher ein cartesisches Produkt von reellen Intervallen sein soll.

So eine Abbildung ist im Falle $n = m = 2$ oder $n = m = 3$ als **Vektorfeld**[1] zu deuten (Skizzen und Beispiele in der Vorlesung,, z.B. Geschwindigkeitsfeld einer Strömung). Viele physikalische Felder sind Vektorfelder. Ein Spezialfall ist $F(x) = (\nabla f)(x)$ als Gradientenfeld einer skalarwertigen Funktion $f : \mathbb{R}^n \to \mathbb{R}$. So eine Funktion f bezeichnet man dann auch als **Skalarfeld**.

Definition 12.41 *Ist $F : D \supseteq \mathbb{R}^n \to \mathbb{R}^m$ mit $F(x) = (F_1(x), \ldots, F_m(x))^T$ und $x = (x_1, \ldots, x_n)^T$ eine vektorwertige Abbildung, für die alle partiellen*

[1] `http://de.wikipedia.org/wiki/Vektorfeld`

Ableitungen aller Komponenten nach allen Variablen in einem Punkte $z \in D$ existieren, so faßt man diese in der **Jacobimatrix**[1] *oder* **Funktionalmatrix**

$$F'(z) = \nabla F(z) := \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(z) & \frac{\partial F_1}{\partial x_2}(z) & \cdots & \frac{\partial F_1}{\partial x_n}(z) \\ \frac{\partial F_2}{\partial x_1}(z) & \frac{\partial F_2}{\partial x_2}(z) & \cdots & \frac{\partial F_2}{\partial x_n}(z) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1}(z) & \frac{\partial F_m}{\partial x_2}(z) & \cdots & \frac{\partial F_m}{\partial x_n}(z) \end{pmatrix}$$

zusammen und sagt, F sei in z einmal **partiell differenzierbar**.

Natürlich ist F auf D einmal partiell differenzierbar, wenn alle partiellen Ableitungen in allen Punkten von D existieren, und ist dort einmal stetig partiell differenzierbar, wenn alle partiellen Ableitungen stetige Funktionen auf D sind.

Theorem 12.42 *(***Kettenregel***)*
Es seien $F : \mathbb{R}^n \to \mathbb{R}^m$, $G : \mathbb{R}^m \to \mathbb{R}^k$ partiell differenzierbar, und wir schreiben $F(x) = F(x_1, \ldots, x_n)$ und $G(y) = G(y_1, \ldots, y_m)$ sowie $F = (F_1, \ldots, F_m)^T$ und $G := (G_1, \ldots, G_k)^T$. Dann gilt für alle $x \in \mathbb{R}^n$

$$\frac{\partial (G \circ F)_\ell}{\partial x_j}(x) = \sum_{i=1}^{m} \frac{\partial G_\ell}{\partial y_i}(F(x)) \frac{\partial F_i}{\partial x_j}(x), \ 1 \le j \le n, \ 1 \le \ell \le k$$

oder als Matrizenmultiplikation bei Auswertung in x

$$(\nabla (G \circ F))(x) = (\nabla G)(F(x)) \cdot (\nabla F)(x).$$

Die Formel gilt sinngemäß auch bei eingeschränkten Definitionsbereichen.

Beweis: Auf jede einzelne Komponente $(G \circ F)_\ell(x) = G_\ell(F(x))$ angewendet, braucht man nur Satz 12.34. □

12.2.2 Flächen

Flächen[2] im $\mathbb{R}^3$ sind Abbildungen $F(x) = (F_1(x), F_2(x), F_3(x))^T$ mit $x = (x_1, x_2)^T$ als Abb $\mathbb{R}^2 \to \mathbb{R}^3$ oder auf einer Teilmenge von $\mathbb{R}^2$. Dann ist ∇F eine 3×2–Matrix. Der modernere Begriff in der Mathematik ist **Mannigfaltigkeit**[3], aber es ist für Informatik–Studierende zu allgemein und zu kompliziert gefaßt.

[1] `http://de.wikipedia.org/wiki/Jacobimatrix`
[2] `http://de.wikipedia.org/wiki/Fl%C3%A4che_%28Topologie%29`
[3] `http://de.wikipedia.org/wiki/Mannigfaltigkeit`

In "parametrischer" Schreibweise und in cartesischen Koordinaten kann man Flächen im $\mathbb{R}^3$ als Abbildungen $F(p) = (x(p), y(p), z(p))^T$ mit $p = (u, v)^T$ als Abb $\mathbb{R}^2 \to \mathbb{R}^3$ oder auf einer Teilmenge von $\mathbb{R}^2$ schreiben. Man hat dann im $\mathbb{R}^2$ die (u, v)–Koordinaten und im $\mathbb{R}^3$ die (x, y, z)–Koordinaten. Der **Flächenparameter** ist $p = (u, v)^T$.

Eine Kurve auf der Fläche bekommt man mit einer Kurve, die erst einmal in den Parameterbereich abbildet: $t \mapsto p(t) = (u(t), v(t))^T$, dann bildet man einfach $g(t) := F(p(t)) = (x(p(t)), y(p(t)), z(p(t)))^T$ auf der Fläche. Der Tangentialvektor dazu ist nach der Kettenregel das Matrix–Vektor–Produkt

$$g'(t) = (F \circ p)'(t) = (\nabla F)(p(t)) \cdot \nabla p(t) = (\nabla F)(p(t)) \cdot p'(t)$$

Wir halten jetzt einen Punkt $w = p(t)$ und den Bildpunkt $F(w) = F(p(t))$ fest und betrachten alle Kurven durch diese Punkte, im Parameterbereich zuerst, dann im Bildbereich, d.h. auf der Fläche. Die Matrix $T := (\nabla F)(p(t)) = (\nabla F)(w)$ ist dann fest. Die Tangentialvektoren im Bildbereich sind also alle von der Form $T \cdot p'(t)$, d.h. sie sind Bild von $p'(t)$ unter der festen Matrix T. Sie sind also im $\mathbb{R}^3$ der Bildraum von T. Er heißt **Tangentialraum**[1] an die Fläche im Punkt $F(w)$. Die Jacobimatrix ist also die Matrix, die Tangentialvektoren an die Fläche produziert, wenn man sie auf Vektoren anwendet.

Schreiben wir das noch etwas konkreter hin. Die Ableitungen einer Fläche $F(u, v) = (x(u, v), y(u, v), z(u, v))^T$ bezeichnen wit etwas knapper als

$$F_u := \frac{\partial F}{\partial u} := \begin{pmatrix} \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial u} \\ \frac{\partial z}{\partial u} \end{pmatrix}$$

$$F_v := \frac{\partial F}{\partial v} := \begin{pmatrix} \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial v} \end{pmatrix}$$

und diese beiden Vektoren spannen den Tangentialraum auf. Wir wollen jetzt die Flächennormale ausrechnen. Sie sollte natürlich auf dem Tangentialraum senkrecht stehen, und das erreicht man einfach durch ein Vektorprodukt $F_u \times F_v$.

Beispiele (hier nur Skizze):

Nichtparametrische Flächen der Form $(x, y, z(x, y))^T$, z.B. die Einheits-Halbkugel $(x, y, \sqrt{1 - x^2 - y^2})^T$ auf dem Vollkreis $K = \{(x, y) : x^2 + y^2 \leq 1\}$.

[1] `http://de.wikipedia.org/wiki/Tangentialraum`

Oberfläche der Einheitskugel, parametrisch:

$(\sin\phi \cdot \cos\psi, \sin\phi \cdot \sin\psi, \cos\phi)^T$ auf $(\psi, \phi) \in [0, 2\pi) \times [0, \pi]$.

Man sehe sich diese Fläche und ihre Parametrisierung genau an! Wo liegen die Pole, wo der Äquator?

Schraubenflächen wie z.B. $(r\cos\phi, r\sin\phi, \phi)^T$

Veranschaulichung, Diskussion der Definitionsbereiche, Ausrechnen der Tangentialräume seien den Lesern überlassen. Es gibt schöne websites mit Beispielen von Flächen, z.B. die des Virtual Math Museums[1]

Bernstein-Bezier-Tensorproduktflächen:

Man nehme zwei Sätze von Bernsteinpolynomen $\beta_{i,m},\ 0 \le i \le m$ bzw. $\beta_{j,n},\ 0 \le j \le n$ und bilde alle Produkte

$$\gamma_{ij}(u,v) := \beta_{i,m}(u) \cdot \beta_{j,n}(v),\ 0 \le i \le m,\ 0 \le j \le n.$$

Das ergibt $(m+1)\cdot(n+1)$ Polynome von 2 Variablen. Jetzt nimmt man ein Kontrollnetz aus $(m+1)\cdot(n+1)$ dreidimensionalen Vektoren $b_{ij} \in I\!R^3,\ 0 \le i \le m,\ 0 \le j \le n$ hinzu und bildet die Fläche

$$\begin{aligned} F(u,v) &:= \sum_{i=0}^{m}\sum_{j=0}^{n} b_{ij}\gamma_{ij}(u,v) = \sum_{i=0}^{m}\sum_{j=0}^{n} b_{ij}\beta_{i,m}(u)\cdot\beta_{j,n}(v) \\ &= \sum_{i=0}^{m} \beta_{i,m}(u) \cdot \underbrace{\left(\sum_{j=0}^{n} b_{ij}\beta_{j,n}(v)\right)}_{=:c_i(v)\in I\!R^3} = \sum_{i=0}^{m}\beta_{i,m}(u)c_i(v) \\ &= \sum_{j=0}^{n} \beta_{j,n}(v) \cdot \underbrace{\left(\sum_{i=0}^{m} \beta_{i,m}(u)b_{ij}\right)}_{=:d_j(u)\in I\!R^3} = \sum_{j=0}^{n}\beta_{j,n}(v)d_j(u) \end{aligned}$$

als zwei verschiedene Schreibweisen von "Kurven von Kurven". Diese **Bernstein-Bezier-Tensorproduktflächen** werden in der Vorlesung etwas genauer beschrieben und gezeichnet. Man sehe sich die **isoparametrischen Kurven** $F(u_0, v)$ und $F(u, v_0)$ an und schreibe sie als Bernstein-Bezier-Kurven. Insbesondere die Randkurven des Flächenstücks. Wie sehen die Tangentialvektoren am Rand und in den Ecken aus? Wie kann man die Tangentialräume beschreiben? Bilineare und biquadratische Flächen?

[1] `http://virtualmathmuseum.org`

12.3 Implizite Funktionen

12.3.1 Implizit definierte Kurven und Flächen

Zuerst gehen wir in den $\mathbb{R}^2$ und betrachten Gleichungen der Form $g(x,y) = 0$. Beispiele sind Geradengleichungen $ax + by + c = 0$ oder Kreisgleichungen $(x - x_0)^2 + (y - y_0)^2 - r^2 = 0$. Gefragt ist, ob man daraus eine explizit definierte Kurve bestimmen kann, z.B. in nichtparametrischer Form $y(x)$ oder $x(y)$, oder in parametrischer Form $(x(t), y(t))$. Wenn man annimmt, die Funktion g sei partiell differenzierbar und die Auflösung nach x, y oder beiden sei möglich, bekommt man durch Differenzieren beispielsweise aus dem Ansatz $g(x, y(x)) = 0$ die Gleichung $g_x(x, y(x)) + g_y(x, y(x))y'(x) = 0$ und daraus die **Differentialgleichung**

$$y'(x) = -\frac{g_x(x, y(x))}{g_y(x, y(x))}$$

sofern g_y nicht verschwindet. Hier haben wir $g_x := \frac{\partial g}{\partial x}$ und analog $g_y := \frac{\partial g}{\partial y}$ gesetzt.

Unter ziemlich schwachen Voraussetzungen sind solche Differentialgleichungen lokal lösbar, aber dazu kommen wir in dieser Vorlesung leider nicht. Man kann auch ohne Differentialgleichungen aus dem obigen Ansatz auf das richtige Resultat kommen:

Theorem 12.43 *(**Satz über implizite Funktionen**)*
Ist g eine reelle Funktion zweier Variablen (x, y), die in einer Umgebung U eines Punktes (x_0, y_0) definiert und dort stetig partiell differenzierbar ist, wobei g_y nicht verschwindet und $g(x_0, y_0) = 0$ gilt, so existiert in einer Umgebung V von x_0 eine Funktion y von x mit $g(x, y(x)) = 0$ in V.

Beweis: Wir nehmen $g_y > 0$ an und betrachten $f(y) := g(x_0, y)$ in einer Umgebung von y_0. Diese Funktion ist wegen $g_y > 0$ streng monoton und stetig, und deshalb gibt es ein kleines positives ϵ so daß

$$f(y_0 - \epsilon) < 0 = f(y_0) < f(y_0 + \epsilon)$$

gilt und die Punkte (x_0, y) mit $|y - y_0| < \epsilon$ noch in U liegen. Jetzt definieren wir die um x_0 stetigen Funktionen $f_-(x) := g(x, y_0 - \epsilon)$ und $f_+(x) := g(x, y_0 + \epsilon)$ und bekommen

$$f_-(x_0) = g(x_0, y_0 - \epsilon) = f(y_0 - \epsilon) < 0 = f(y_0) < f(y_0 + \epsilon) = g(x_0, y_0 + \epsilon) = f_+(x_0).$$

Dann gibt es ein $\delta > 0$ so daß die Ungleichung

$$f_-(x) < 0 < f_+(x)$$

auch noch für $|x - x_0| < \delta$ gilt, und das definiert unsere Umgebung V von x_0. Siehe die Abbildung 11 von Anna Eggers dazu. Für ein beliebiges festes x aus diesem Intervall verläuft die Funktion $y \mapsto g(x, y)$ mindestens zwischen $g(x, y_0 - \epsilon) = f_-(x) < 0$ und $g(x, y_0 + \epsilon) = f_+(x) > 0$. Also gibt es nach dem Zwischenwertsatz oder Nullstellensatz zu unserem x ein $y(x)$ mit $g(x, y(x)) = 0$. □

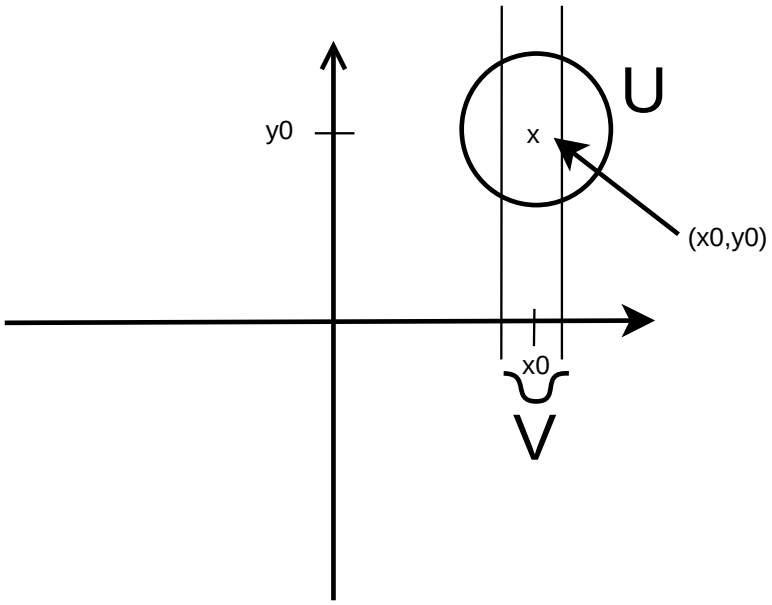


Abbildung 11: Die Umgebungen U und V im Beweis von Theorem 12.43.

Analog bekommt man ein Resultat zur Auflösbarkeit nach x, wenn g_x nicht verschwindet. Aus beiden Ergebnissen zusammen kann man auf die parametrische Auflösbarkeit $g(x(t), y(t)) = 0$ schließen, sobald $g_y^2 + g_x^2 > 0$ gilt. Allerdings sind alle diese Auflösungsmöglichkeiten nur lokal definiert und keineswegs elegant, besonders nicht im parametrischen Fall.

Beispiele: Auflösen der Kreisgleichung.

Zur Vorbereitung auf den Flächenfall geben wir noch an, wie man aus einer stetig differenzierbaren expliziten zweidimensionalen Kurvengleichung $g(x, y) = 0$ in einem festen Kurvenpunkt (x_0, y_0) mit $g(x_0, y_0) = 0$ einen Normalenvektor ausrechnet, der auf dem Tangentialvektor senkrecht steht. Der parametrische Ansatz $g(x(t), y(t)) = 0$ entlang der Kurve liefert bei Differentiation die Gleichung

$$g_x(x(t), y(t)) \cdot x'(t) + g_y(x(t), y(t)) \cdot y'(t) = 0.$$

Der Tangentialvektor ist $(x'(t), y'(t))$, und deshalb steht der Vektor

$$(g_x(x(t), y(t)), g_y(x(t), y(t))) = (\nabla g)(x(t), y(t)$$

darauf senkrecht. An der Stelle (x_0, y_0) kann man sich die ganzen Ableitungen sparen und einfach den Normalenvektor als $(g_x(x_0, y_0), g_y(x_0, y_0))$ ansetzen. Die ganze Rechnung ist nur eine Wiederholung des Arguments, daß der Gradient auf den Höhenlinien einer Funktion senkrecht steht. Denn unsere implizite Kurve ist die Höhenlinie zu g vom Niveau 0, und der Gradient von g muß darauf überall senkrecht stehen.

Der Satz über implizite Funktionen gilt sinngemäß und mit praktisch gleichem Beweis auch für reelle Funktionen von mehreren Variablen. Man hat eine Gleichung der Form $g(y, x_1, \ldots, x_n) = 0$ mit stetig partiell differenzierbarem g in einer Umgebung eines Punktes $y_0, \tilde{x}_1, \ldots, \tilde{x}_n$ mit $g(y_0, \tilde{x}_1, \ldots, \tilde{x}_n) = 0$. Ist dann in dieser Umgebung g_y nirgends Null, so kann man in einer Umgebung von $(\tilde{x}_1, \ldots, \tilde{x}_n)$ nach y auflösen und damit die Gleichung $g(y(x_1, \ldots, x_n), x_1, \ldots, x_n) = 0$ dort erfüllen. Das tritt ein, wenn man im $\mathbb{R}^3$ eine Fläche durch eine Gleichung $g(x, y, z) = 0$ in cartesischen Koordinaten beschreibt und lokal etwa nach $z(x, y)$ auflösen will, so daß $g(x, y, z(x, y)) = 0$ gilt.

Will man einen Normalenvektor oder die Tangentialvektoren aus einer impliziten Flächendarstellung $g(x, y, z) = 0$ ausrechnen, ist das für die Normale am einfachsten, denn sie ist bis auf einen Skalarfaktor durch $(\nabla g)^T = (\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}, \frac{\partial g}{\partial z})^T$ gegeben. Zwei weitere dazu orthogonale Vektoren findet man leicht. Denn man kann zu einem beliebigen Vektor $a \in \mathbb{R}^3 \setminus \{0\}$ mit $a = (a_1, a_2, a_3)$ und $(a_1, a_2) \neq (0, 0)$ erst $b := (-a_2, a_1, 0)$ und dann $c := a \times b$ berechnen.

Schwieriger ist die Auflösung impliziter Gleichungen, wenn man mehrere Gleichungen hat und nach mehr als einer Funktion gleichzeitig auflösen will. Das tritt z.B. auf, wenn man zwei implizit durch Gleichungen der Form

$$\begin{aligned} g_1(x, y, z) &= 0 \\ g_2(x, y, z) &= 0 \end{aligned}$$

gegebene Flächen F_1 und F_2 im $\mathbb{R}^3$ schneiden will. Wenn es eine Schnittkurve $(x(t), y(t), z(t))$ mit einem Parameter t gibt, muss

$$\begin{aligned} g_1(x(t), y(t), z(t)) &= 0 \\ g_2(x(t), y(t), z(t)) &= 0 \end{aligned}$$

gelten. Falls man die Kurve über x als $(x, y(x), z(x)$ parametrisieren kann, hat man

$$\begin{aligned} g_1(x, y(x), z(x)) &= 0 \\ g_2(x, y(x), z(x)) &= 0 \end{aligned}$$

zu lösen. Man löst also diese 2 Gleichungen nach 2 der Unbekannten auf und nimmt die verbleibende Unbekannte als unabhängige Variable.

Hat man zwei explizit gegebene Flächen als Bilder der Abbildungen $F_1(u, v), F_2(r, s) \in \mathbb{R}^3$, so ist das Schnittproblem noch unangenehmer. Mit $F_1(u, v) = F_2(r, s)$ hat man 3 Gleichungen mit 4 Unbekannten, und man kann sich eine Variable, z.B. r herauspicken und die drei anderen als Funktion von r zu schreiben versuchen. Das läuft auf die Lösung von $F_1(u(r), v(r)) - F_2(r, s(r)) = 0$ heraus, d.h. man löst diese 3 Gleichungen nach 3 der Unbekannten auf und nimmt die verbleibende Unbekannte als unabhängige Variable.

Der allgemeine Fall hat die Form $F(z) = 0$, wobei F Werte im $\mathbb{R}^n$ hat und z aus mehr als n Unbekannten besteht. Wenn man dieses Gleichungssystem nach n der Variablen aus $z \in \mathbb{R}^m$ auflösen will, teilt man z in n "abhängige" Variablen y und $m-n$ "unabhängige" Variablen x auf. Damit bekommt man die alternative Form $F(x, y) = 0$, wobei nun nach $y \in R^n$ in Abhängigkeit von $x \in \mathbb{R}^{m-n}$ gefragt ist, d.h. das Gleichungssystem $F(x, y(x)) = 0$ ist zu lösen. Die "richtige" Bedingung ist natürlich an den (partiellen) Gradienten von F bezüglich y gekoppelt.

Theorem 12.44 *Ist $P := (x_0, y_0) \in \mathbb{R}^m = \mathbb{R}^{m-n} \times \mathbb{R}^n$ ein Punkt, und ist F eine in einer Umgebung U von P stetig partiell differenzierbare Abbildung mit Werten im $\mathbb{R}^n$, mit $F(x_0, y_0) = 0$ und mit in U nirgends verschwindender Determinante der $n \times n$–Jacobimatrix $\nabla_y F$ von $G(y) := F(x, y)$ bei festem x, so kann man in einer Umgebung V von $x_0 \in \mathbb{R}^{m-n}$ eine Abbildung y mit Werten im $\mathbb{R}^n$ angeben mit $F(x, y(x)) = 0$ für alle $x \in V$.*

Dieser Satz kann per Induktion bewiesen werden, aber das wollen wir uns nicht antun.

Stattdessen sehen wir uns an, ob man ein Verfahren zur Berechnung von $y(x)$ aus x angeben kann. Das entspricht bei festem x der Bestimmung einer Nullstelle von $G(y) := F(x, y)$. Dabei ist eine Näherung (x_0, y_0) schon bekannt. Wir vergessen jetzt das x und bestimmen nur noch eine Lösung y eines Gleichungssystems $G(y) = 0$ mit n Gleichungen und n Unbekannten, wobei wir Nichtsingularität von ∇G voraussetzen und wissen, dass wir eine Näherung y_0 für den gesuchten Vektor y haben. Das machen wir im nächsten Abschnitt.

12.3.2 Nichtlineare Gleichungen und Gleichungssysteme

Wir können den gesamten Kontext der impliziten Funktionen wieder ignorieren und uns auf die altgewohnten Bezeichnungen zurückziehen. Im eindimensionalen Fall haben wir eine Gleichung $f(x) = 0$ mit einer reellen Unbekannten und einer reellwertigen Funktion zu lösen, wobei eine Näherung x_0 mit $f(x_0) \approx 0$ bekannt und f in einer Umgebung U von x_0 definiert sei. Ist $\tilde{x} \in U$ die gesuchte Nullstelle, so folgt nach dem Satz von Taylor

$$0 = f(\tilde{x}) = f(x_0) + f'(x_0)(\tilde{x} - x_0) + \mathcal{O}(\tilde{x} - x_0)^2$$

sofern f in U zweimal stetig differenzierbar ist. Ist $f'(x_0)$ von Null verschieden, so kann man nach $\tilde{x}$ formell auflösen und bekommt

$$\tilde{x} = x_0 - \frac{f(x_0)}{f'(x_0)}$$

bis auf einen Term mit dem Verhalten $\frac{1}{f'(x_0)}\mathcal{O}(\tilde{x} - x_0)^2$. Setzt man

$$x_1 := x_0 - \frac{f(x_0)}{f'(x_0)},$$

so folgt

$$\tilde{x} - x_1 = \frac{1}{f'(x_0)}\mathcal{O}(\tilde{x} - x_0)^2.$$

Wenn man weiß, daß in ganz U die Ableitung f' nicht Null ist, folgt aus der obigen Gleichung, daß beim Start dicht bei $\tilde{x}$ der absolute Fehler bis auf einen festen Faktor quadriert wird. Wenn der Anfangsfehler $\tilde{x} - x_0$ hinreichend klein ist, verdoppelt sich die Anzahl der korrekten Stellen bei jedem Schritt! Man rechnet also mit dem **Newton–Verfahren**

$$x_{i+1} := x_i - \frac{f(x_i)}{f'(x_i)}$$

und bekommt eine sehr schnell gegen die Nullstelle konvergierende Iteration. Geometrisch ersetzt man die Nullstellenberechnung für f in jedem Schritt durch die Nullstellenberechnung einer Tangente. Abbildung 12 zeigt die geometrische Konstruktion, während 13 und 14 Beispiele für Fehlschläge zeigen. Die Abbildungen hat wieder Anna Eggers beigesteuert.

Theorem 12.45 *Ist f eine in einer Umgebung U einer Nullstelle $\tilde{x}$ definierte zweimal stetig differenzierbare Funktion mit in U nirgends verschwindender Ableitung, so konvergiert das Newton–Verfahren gegen die Nullstelle, sofern der Startwert x_0 hinreichend dicht bei der Nullstelle liegt.*

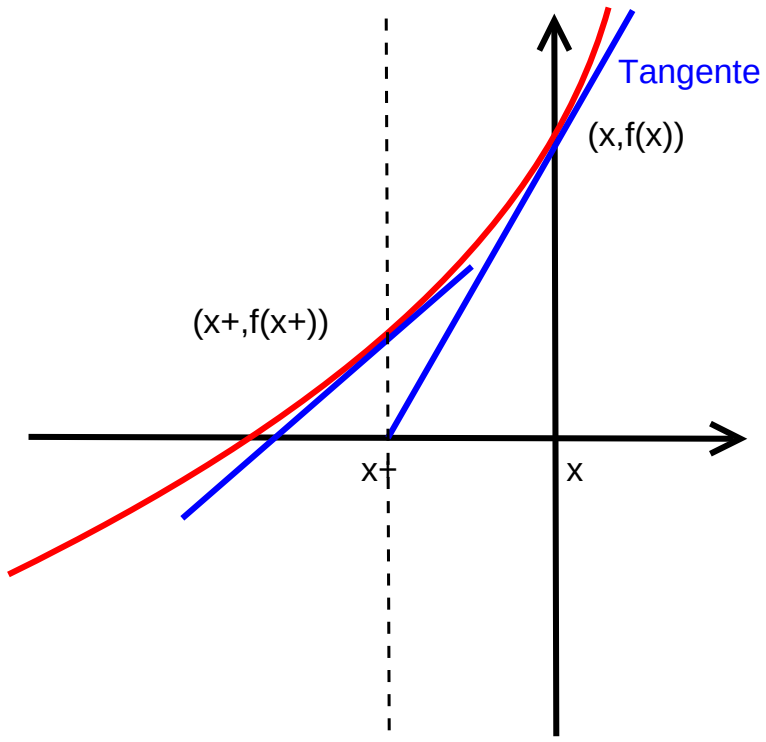


Abbildung 12: Newtonverfahren

Wir verzichten auf eine strikte Beweisführung, weil sie im wesentlichen die obige Argumentation wiederholt. Es gibt stärkere Konvergenzsätze, die nicht von der Existenz der Nullstelle ausgehen, aber diese gehören in die Numerische Mathematik.

Jetzt zeigen wir noch, dass man im Falle eines Gleichungssystems $F(x) = 0$ mit einer in einer Umgebung U einer Nullstelle $\tilde{x} \in \mathbb{R}^n$ definierten Funktion F mit Werten im $\mathbb{R}^n$ ganz analog verfahren können. Die näherungsweise gültige Beziehung

$$0 = F(\tilde{x}) \approx F(x_0) + (\nabla F)(x_0)(\tilde{x} - x_0)$$

wird unter Voraussetzung der Nichtsingularität der Jacobimatrix $(\nabla F)(x)$ auf U umgestellt und in das Newton–Verfahren

$$x_{i+1} := x_i - ((\nabla F)(x_i))^{-1} F(x_i)$$

verwandelt. In der Praxis invertiert man die Jacobimatrix nicht, sondern löst bei gegebenem x_i das lineare Gleichungssystem

$$((\nabla F)(x_i))(x_{i+1} - x_i) = F(x_i).$$

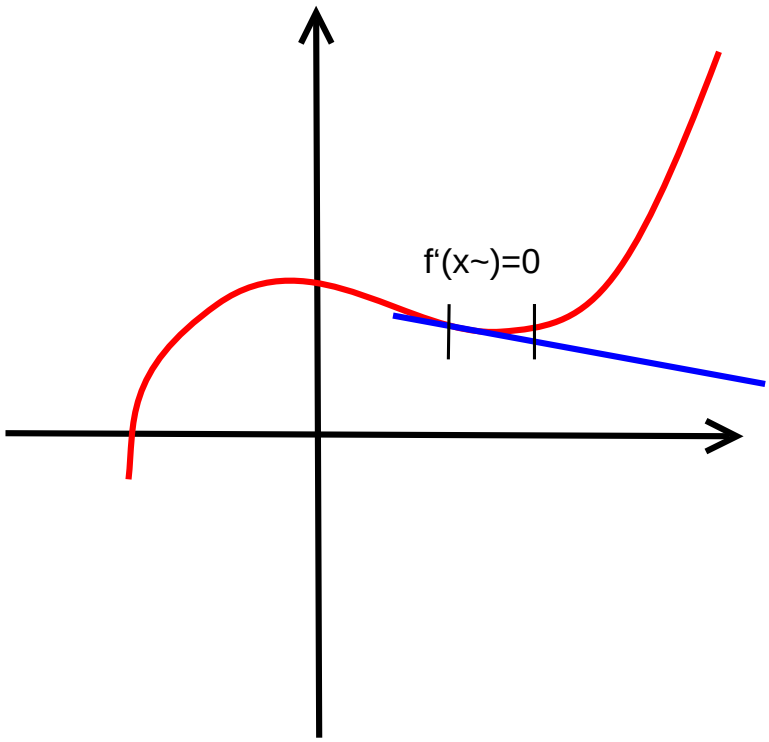


Abbildung 13: Newtonverfahren auf Abwegen

Auch dieses Verfahren konvergiert bei gutem Startwert und entsprechenden Differenzierbarkeitsvoraussetzungen sehr gut.

12.4 Vektoranalysis

Ab hier betrachten wir nur Vektorfelder $F : \mathbb{R}^3 \to \mathbb{R}^3$ und skalare Felder $f : \mathbb{R}^3 \to \mathbb{R}$. Dabei zielen wir auf wichtige Felder aus der Physik, z.B. die aus dem Elektromagnetismus. Alle in diesem Abschnitt auftretenden Abbildungen seien mindestens zweimal stetig partiell differenzierbar. Erst noch eine Wiederholung:

Der **Gradient**

$$\nabla f := \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3} \right)$$

bildet $\mathbb{R}^3 \to \mathbb{R}^3$ ab und ist also ein aus einem skalaren Feld f abgeleitetes Vektorfeld.

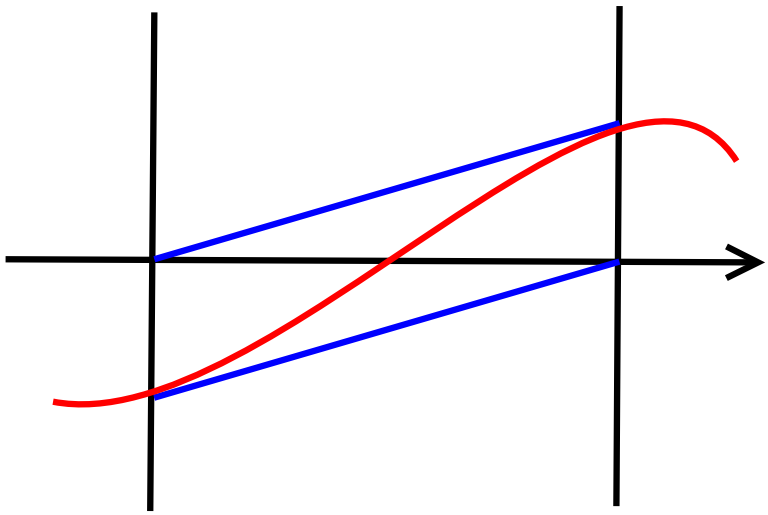


Abbildung 14: Kreisendes Newtonverfahren

Die **Divergenz**

$$div\ F := \sum_{j=1}^{3} \frac{\partial F_j}{\partial x_j}$$

bildet $I\!R^3 \to I\!R$ ab und ist also ein aus einem Vektorfeld F abgeleitetes skalares Feld. Aus Gründen, die man erst nach der multivariaten Integration gut versteht, wird ein Feld F **quellenfrei** genannt, wenn $div\ F = 0$ gilt. Das trifft für das magnetostatische Feld zu.

Für das Skalarprodukt zweier Felder F, G gilt nach Kettenregel

$$\nabla(F^T G) = (\nabla F)^T G + F^T \nabla G$$

und für ein Produkt aus einem Vektorfeld F und einem Skalarfeld f

$$div\ (f \cdot F) = f \cdot div\ F + (\nabla f) \cdot F.$$

Das sind Übungsaufgaben zur Produktregel.

Läßt sich ein Vektorfeld F als Gradient $F = \nabla f$ eines Skalarfeldes f schreiben, so heißt f ein **Potential** zu F. Für die Divergenz eines solchen Feldes gilt

$$div(\nabla f) = \sum_{j=1}^{3} \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_j} = \sum_{j=1}^{3} \frac{\partial^2 f}{\partial^2 x_j} =: \Delta f$$

und das ist der **Laplace**–Operator. **Harmonische Funktionen** sind skalare Funktionen u mit $\Delta u = 0$. Ihr Gradient ist also quellenfrei. Es ist ein

wichtiges Problem des wissenschaftlichen Rechnens, "partielle Differentialgleichungen" wie $\Delta u = f$ zu lösen. Dabei ist f vorgegeben und u gesucht.

Die **Rotation** von F ist ein Vektorfeld

$$rot\ F := \left(\frac{\partial F_3}{\partial x_2} - \frac{\partial F_2}{\partial x_3}, \frac{\partial F_1}{\partial x_3} - \frac{\partial F_3}{\partial x_1}, \frac{\partial F_2}{\partial x_1} - \frac{\partial F_1}{\partial x_2}, \right)^T .$$

Es bildet $I\!R^3 \to I\!R^3$ ab und ist also ein aus einem Vektorfeld f abgeleitetes weiteres Vektorfeld. Ein Vektorfeld F heißt **wirbelfrei**, wenn $rot\ F = 0$ gilt. Das gilt für die Felder der Elektrostatik. Glatte Gradientenfelder von Potentialen sind immer wirbelfrei:

$$rot\ \nabla f = 0$$

wegen

$$rot\ \nabla f := \left(\frac{\partial}{\partial x_2}\frac{\partial f}{\partial x_3} - \frac{\partial}{\partial x_3}\frac{\partial f}{\partial x_2}, \frac{\partial}{\partial x_3}\frac{\partial f}{\partial x_1} - \frac{\partial}{\partial x_1}\frac{\partial f}{\partial x_3}, \frac{\partial}{\partial x_1}\frac{\partial f}{\partial x_2} - \frac{\partial}{\partial x_2}\frac{\partial f}{\partial x_1}, \right)^T = 0$$

"Umgekehrt" sind glatte Wirbelfelder immer quellenfrei, denn

$$div\ rot\ F = \frac{\partial^2 F_3}{\partial x_1 x_2} - \frac{\partial^2 F_2}{\partial x_1 x_3} + \frac{\partial^2 F_1}{\partial x_2 x_3} - \frac{\partial^2 F_3}{\partial x_2 x_1} + \frac{\partial^2 F_2}{\partial x_3 x_1} - \frac{\partial^2 F_1}{\partial x_3 x_2} = 0.$$

Diese Differentiationsabbildungen sind auf ein cartesisches Koordinatensystem bezogen. Es ist eine gute Übung, sie auf Polar-, Kugel- oder Zylinderkoordinaten umzurechnen.

Aufgabe: Man sehe sich die Felder

$$\begin{aligned} F(x,y,z) &= (x,y,z)^T \\ F(x,y,z) &= (-y,x,0)^T \end{aligned}$$

durch Zeichnung an und rechne nach, ob sie quellen- oder wirbelfrei sind. Was könnte hier mit "Quelle" oder "Wirbel" geometrisch gemeint sein? Bestimmen Sie zu einem der Felder ein Potential. Warum funktioniert das für das andere Feld garantiert nicht?
Tip: Der Vektor $(-y,x)^T$ steht auf $(x,y)^T$ senkrecht.

13 Integralrechnung

Die Integralrechnung[1] kann man als Umkehrung der Differentialrechnung ansehen. Wie bei der Differentialrechnung beginnen wir mit reellen Funktionen einer reellen Variablen und gehen dann zu Integralen multivariater Funktionen über.

13.1 Univariate Integrale

13.1.1 Bestimmte Integrale

Wir gehen zurück auf Satz 11.37 auf Seite 295. Dort war eine stetige Funktion f auf einem abgeschlossenen und beschränkten Intervall $[a, b]$ vorgegeben. Dann kann man zu jedem $\epsilon > 0$ ein $h_0 > 0$ angeben, so daß für alle Zerlegungen

$$\Delta : \quad a = x_0 < x_1 < \ldots < x_{n+1} = b$$

mit Maximalschrittweite

$$h(\Delta) := \max |x_{j+1} - x_j| < h_0$$

die Funktion f durch je eine auf den Teilintervallen $[x_j, x_{j+1}]$ der Zerlegung stückweise konstante "Unter"- und "Oberfunktion" f^{unten} bzw. f^{oben} angenähert werden kann, so daß die Abschätzungen

$$\begin{array}{rcccl} 0 & \leq & f_\Delta^{oben}(x) - f_\Delta^{unten}(x) & \leq & \epsilon \\ f_\Delta^{unten}(x) & \leq & f(x) & \leq & f_\Delta^{oben}(x) \end{array}$$

für alle $x \in [a, b]$ gelten. Nun bildet man die "Unter"- und "Obersumme"[2][3]

$$\begin{aligned} s_\Delta^{unten} &:= \sum_{j=0}^{n} (x_{j+1} - x_j) \cdot f_\Delta^{unten}\left(\frac{x_j + x_{j+1}}{2}\right) \\ &= \sum_{j=0}^{n} (x_{j+1} - x_j) \cdot \min_{x \in [x_j, x_{j+1}]} f(x) \\ s_\Delta^{oben} &:= \sum_{j=0}^{n} (x_{j+1} - x_j) \cdot f_\Delta^{oben}\left(\frac{x_j + x_{j+1}}{2}\right) \\ &= \sum_{j=0}^{n} (x_{j+1} - x_j) \cdot \max_{x \in [x_j, x_{j+1}]} f(x) \end{aligned}$$

[1] http://de.wikipedia.org/wiki/Integralrechnung

[2] http://de.wikipedia.org/wiki/Riemann-Integral#Ober-_und_Untersummen

[3] http://www.geogebra.at/de/upload/files/dynamische_arbeitsblaetter/lwolf/oberuntersumme/obe

die alle rechteckigen Teilflächen (Zeichnung!) vorzeichenbehaftet aufsummieren und bekommt

$$\begin{aligned} 0 &\leq s_\Delta^{oben} - s_\Delta^{unten} &&\leq (b-a)\epsilon \\ s_\Delta^{unten} &\leq \sum_{j=0}^{n}(x_{j+1}-x_j)\cdot f(\xi_j) &&\leq s_\Delta^{oben} \end{aligned}$$

wobei die Auswertungspunkte ξ_j in $[x_j, x_{j+1}]$ beliebig gewählt sein können. Wenn man die Zerlegungen immer feiner wählt, strebt $s_\Delta^{oben} - s_\Delta^{unten}$ gegen Null.

Das reicht aber noch nicht, um einzusehen, daß s_Δ^{oben} und s_Δ^{unten} gegen einen gemeinsamen Grenzwert streben. Dazu kann man Monotonie ausnutzen, und zwar in folgender Weise, die wir hier aber nur skizzieren. Eine Zerlegung Δ_1 heißt **Verfeinerung** einer Zerlegung Δ_0, wenn alle Teilpunkte von Δ_0 auch Teilpunkte von Δ_1 sind. Dann kann man beweisen (und an einer Zeichnung sehen), daß

$$\begin{aligned} 0 &\leq s_{\Delta_1}^{oben} - s_{\Delta_1}^{unten} &&\leq s_{\Delta_0}^{oben} - s_{\Delta_0}^{unten} \\ s_{\Delta_0}^{unten} &\leq s_{\Delta_1}^{unten} &&\leq \sum_{j=0}^{n}(x_{j+1}-x_j)\cdot f(\xi_j) \leq s_{\Delta_1}^{oben} \leq s_{\Delta_0}^{oben} \end{aligned}$$

gilt. Man sehe sich dazu an, wie eine Zerlegung durch Hinzufügen eines weiteren Punktes verfeinert wird (Skizze in der Vorlesung). Durch immer feiner werdende Zerlegungen folgt dann

Theorem 13.1 *Für jede stetige Funktion f auf einem abgeschlossenen und beschränkten Intervall $[a,b]$ existiert der gemeinsame Limes der Ober- und Untersummen, wenn die Maximalschrittweiten der betreffenden Zerlegungen gegen Null streben. Der Limes heißt* **bestimmtes Integral** *imr Riemannschen Sinne*[1] *von f auf $[a,b]$ und wird mit*

$$\int_a^b f(x)dx$$

bezeichnet. Dabei ist die Bezeichnung der **Integrationsvariablen** *x beliebig. Das bestimmte Integral ist gleichzeitig der Limes der* **Riemannschen Summen**

$$\sum_{j=0}^{n}(x_{j+1}-x_j)\cdot f(\xi_j)$$

[1] `http://de.wikipedia.org/wiki/Riemann-Integral`

wobei die Punkte ξ_j beliebig aus $[x_j, x_{j+1}]$ gewählt werden können und die Maximalschrittweite der zugrundeliegenden Zerlegung gegen Null strebt.
In gleicher Weise ist

$$\int_{\alpha}^{\beta} f(x)dx$$

für alle $[\alpha, \beta] \subseteq [a, b]$ definiert. Man nennt dann f den **Integranden** *und α und β die* **Integrationsgrenzen**.

Definition 13.2 *Unter den Voraussetzungen des vorigen Satzes legt man noch fest. daß*

$$\int_{\beta}^{\alpha} f(t)dt = -\int_{\alpha}^{\beta} f(t)dt, \ \int_{\alpha}^{\alpha} f(t)dt = 0$$

für alle $\alpha < \beta$ aus $[a, b]$ gilt.

Diese Integraldefinition geht auf **Riemann**[1] zurück. Sie benutzt Monotonie und Stetigkeit, und sie erweist sich für weitergehende Anwendungen als unzureichend. Stattdessen verwendet man das **Lebesgue**[2]–Integral[3], das allerdings eine saubere **Masstheorie**[4] erfordert, die wir hier unterdrücken müssen.

13.1.2 Eigenschaften des Integrals

Theorem 13.3 *Falls die auftretenden Integranden zwischen den Integrationsgrenzen stetig sind, gilt*

$$\begin{aligned}
\int_a^c f(x)dx &= \int_a^b f(x)dx + \int_b^c f(x)dx \\
\int_a^b (f+g)(x)dx &= \int_a^b f(x)dx + \int_a^b g(x)dx \\
\int_a^b (\alpha \cdot f(x))dx &= \alpha \cdot \int_a^b f(x)dx \\
\int_a^b f(x)dx &\leq \int_a^b g(x)dx \text{ falls } f(x) \leq g(x) \text{ für alle } x \in [a, b]
\end{aligned}$$

Die Integration ist also eine bezüglich des Integranden lineare und monotone Funktion. Die Beweise zu den obigen Aussagen ergeben sich leicht, indem man die entsprechenden Riemannschen Summen bildet.

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Riemann.html`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Lebesgue.html`
[3] `http://de.wikipedia.org/wiki/Lebesgue-Integral`
[4] `http://de.wikipedia.org/wiki/Ma%C3%9Ftheorie`

Theorem 13.4 *(***Mittelwertsatz** *der Integralrechnung*[1]*)*
Ist f auf $[a,b]$ stetig, so existiert ein $y \in [a,b]$ mit

$$\frac{1}{b-a}\int_a^b f(x)dx = f(y).$$

Beweis: Es gilt

$$\begin{array}{rcccl}
\min\limits_{a\le x\le b} f(x) & \le & f(x) & \le & \max\limits_{a\le x\le b} f(x) \\
(b-a)\min\limits_{a\le x\le b} f(x) & \le & \int_a^b f(x)dx & \le & (b-a)\max\limits_{a\le x\le b} f(x) \\
\min\limits_{a\le x\le b} f(x) & \le & \frac{1}{b-a}\int_a^b f(x)dx & \le & \max\limits_{a\le x\le b} f(x)
\end{array}$$

und deshalb gibt es nach dem Zwischenwertsatz für stetige Funktionen das gewünschte y. □

13.1.3 Stammfunktionen

Definition 13.5 *Eine differenzierbare Funktion g auf einem Intervall I heißt* **Stammfunktion**[2] *zu einer Funktion f auf I, wenn $g' = f$ gilt.*

Theorem 13.6 *Sind g_1 und g_2 Stammfunktionen zu derselben Funktion, so unterscheiden sich g_1 und g_2 nur um eine Konstante.*

Beweis: Es ist wegen $g_1' - g_2' = (g_1 - g_2)' = 0$ zu zeigen, daß eine differenzierbare Funktion g mit $g' = 0$ eine Konstante ist. Das folgt aus dem Mittelwertsatz, denn zu beliebigen $x < y$ aus I folgt

$$\frac{g(y)-g(x)}{y-x} = g'(\xi) = 0$$

mit beliebigem ξ zwischen x und y. □

Theorem 13.7 *(***Hauptsatz der Differential- und Integralrechnung**[3]*)*
Es sei $f : I \to I\!R$ eine stetige Funktion und $a \in I$ beliebig. Dann ist die Funktion

$$g_a(x) := \int_a^x f(t)dt$$

[1] `http://de.wikipedia.org/wiki/Mittelwertsatz_der_Integralrechnung`
[2] `http://de.wikipedia.org/wiki/Stammfunktion`
[3] `http://de.wikipedia.org/wiki/Fundamentalsatz_der_Analysis`

die eindeutig bestimmte Stammfunktion von f, die in a verschwindet. Deswegen nennt man eine Stammfunktion auch **unbestimmtes Integral** *oder* **Integral mit variabler oberer Integrationsgrenze**. *Ist g eine beliebige Stammfunktion zu f, so gilt*

$$\int_x^y f(t)dt = g(y) - g(x) =: g|_x^y$$

für alle $x, y \in I$.

Beweis: Wir rechnen einen Differenzenquotienten von g_a aus und bekommen nach dem Mittelwertsatz der Integralrechnung

$$\frac{g_a(x+h) - g_a(x)}{h} = \frac{1}{h}\left(\int_a^{x+h} f(t)dt - \int_a^x f(t)dt\right) = \frac{1}{h}\int_x^{x+h} f(t)dt = f(y)$$

mit einem y zwischen x und $x+h$. Weil f stetig ist, folgt daraus $g_a'(x) = f(x)$, was zu beweisen war. Obendrein gilt offensichtlich $g_a(a) = 0$. Ist g eine beliebige Stammfunktion zu f, so folgt aus dem vorigen Satz sofort $g(t) = g_a(t) + g(a)$. Das ergibt

$$g(y) - g(x) = g_a(y) - g_a(x) = \int_a^y f(t)dt - \int_a^x f(t)dt = \int_x^y f(t)dt.$$

□

Dieser Satz erlaubt die Berechnung sehr vieler Integrale, weil man die Stammfunktionen kennt.

Beispiele: Monome, trigonometrische Funktionen, Exponentialfunktion, Logarithmus, werden in der Vorlesung dargestellt[1].

13.1.4 Rechenregeln

Die Produktregel für differenzierbare Funktionen f_1, f_2 lautet bekanntlich

$$(f_1(t) \cdot f_2(t))' = f_1'(t) \cdot f_2(t) + f_1(t) \cdot f_2'(t).$$

Sind die Ableitungen noch stetig, können wir integrieren:

$$\begin{aligned}\int_x^y (f_1(t) \cdot f_2(t))'dt &= \int_x^y f_1'(t) \cdot f_2(t)dt + \int_x^y f_1(t) \cdot f_2'(t)dt \\ &= f_1(t) \cdot f_2(t)|_x^y.\end{aligned}$$

[1] `http://de.wikipedia.org/wiki/Tabelle_von_Ableitungs-_und_Stammfunktionen`

Das ist die **partielle Integration** in symmetrischer Formulierung. Die häufigste Form der Anwendung ist unsymmetrisch:

$$\int_a^b u(t)v'(t)dt = u(t)v(t)|_a^b - \int_a^b u'(t)v(t)dt.$$

Die Kettenregel für differenzierbare Funktionen ist

$$(f \circ g)'(t) = f'(g(t)) \cdot g'(t).$$

Sind auch hier die beiden Funktionen stetig differenzierbar, so kann man integrieren und bekommt

$$\begin{aligned} \int_x^y (f \circ g)'(t)dt &= \int_x^y f'(g(t)) \cdot g'(t)dt \\ &= (f \circ g)|_x^y = f(g(y)) - f(g(x)) \\ &= \int_{g(x)}^{g(y)} f'(t)dt. \end{aligned}$$

Hier kann man f' durch eine neue Funktion h ersetzen und bekommt die **Substitutionsregel**

$$\int_x^y h(g(t)) \cdot g'(t)dt = \int_{g(x)}^{g(y)} h(t)dt, \tag{13.8}$$

bei der h nur stetig, die Funktion g aber stetig differenzierbar sein muß. Diese Beziehung wendet man oft in anderer Richtung an:

$$\int_a^b h(t)dt = \int_{g^{-1}(a)}^{g^{-1}(b)} h(g(s)) \cdot g'(s)ds$$

wobei man sich die Eselsbrücke zurechtlegt, dass

$$t = g(s), \frac{dt}{ds} = g'(s), \quad \text{also } \text{“}dt = g'(s)ds\text{”}$$

gilt und dann gern vergißt, die Integrationsgrenzen der rechten Seite richtig einzusetzen.

Der dritte wichtige Trick zum Integrieren ist die **Partialbruchzerlegung**[1]. Sie hat nichts mit Integration zu tun, ist aber dort sehr nützlich. Die Idee ist, eine rationale Funktion

$$f(x) := \frac{P(x)}{Q(x)},$$

[1] `http://de.wikipedia.org/wiki/Partialbruchzerlegung`

d.h. einen Quotienten aus zwei Polynomen P und Q, als Summe einfacherer Funktionen zu schreiben. Falls der Grad von P nicht kleiner ist als der von Q, dividiert man P durch Q mit Rest R und bekommt

$$f(x) := \frac{P(x)}{Q(x)} = \frac{Q(x)W(x) + R(x)}{Q(x)} = W(x) + \frac{R(x)}{Q(x)},$$

wobei nun der Grad von R kleiner ist als der von Q. Deshalb kann man sich auf den Fall beschränken, wo der Grad von P kleiner ist als der von Q. Hat Q den Grad n, so versuche man, Q in seine Faktoren

$$Q(x) = (x - \lambda_1)^{n_1} \cdot (x - \lambda_2)^{n_2} \cdots (x - \lambda_k)^{n_k}$$

zu zerlegen, wobei die schlimmstenfalls komplexen Zahlen $\lambda_1 \ldots, \lambda_k$ die Nullstellen von Q sind, die jeweils mit der Vielfachheit n_1 bis n_k auftreten. Dann kann man $f = P/Q$ umschreiben als

$$f(x) = \sum_{j=1}^{k} \sum_{i=1}^{n_k} \frac{c_{ij}}{(x - \lambda_j)^i}$$

mit geeigneten komplexen Zahlen c_{ij}, die sich durch Anwendung der Bruchrechnung ergeben.

Als Beispiel nehmen wir die Integration von

$$f(x) := \frac{1}{1 - x^2}.$$

Die Partialbruchzerlegung ist

$$f(x) := \frac{1}{2}\left(\frac{1}{1 - x^2} = \frac{1}{1 - x} + \frac{1}{1 + x}\right).$$

und dann ist die Integration über Logarithmen einfach.

Hier kann man sehen, wie MAPLE integriert:

```
> restart;
> f:=x^2;
```

$$f := x^2$$

```
> g:=int(f,x);
```

$$g := 1/3\,x^3$$

```
> h:=int(f,x=-1..1);
```

$$h := 2/3$$

```
> int(exp(-2*x^2),x);
```

$$1/4\,\sqrt{2}\sqrt{\pi}erf\left(\sqrt{2}x\right)$$

```
> int(sqrt(1-x^2),x=-1..1);
```

$$1/2\,\pi$$

```
> g:=int(sqrt(1-x^2),x);
```

$$g := 1/2\,x\sqrt{1-x^2} + 1/2\,\arcsin\left(x\right)$$

```
> simplify(diff(g,x));
```

$$\sqrt{1-x^2}$$

Beispiel:

Mit partieller Integration:

$$\int_a^b \sin^2(x)dx = -\sin(x)\cos(x)|_a^b + \int_a^b \cos^2(x)dx$$

$$\int_a^b \cos^2(x)dx = \int_a^b (1-\sin^2(x))dx$$

$$\int_a^b \sin^2(x)dx = -\sin(x)\cos(x)|_a^b + b - a - \int_a^b \sin^2(x)dx$$

$$\int_a^b \sin^2(x)dx = \frac{b-a}{2} - \frac{1}{2}\sin(x)\cos(x)|_a^b$$

13.2 Anwendungen der Differential– und Integralrechnung

13.2.1 Bogenlänge von Kurven

An dieser Stelle können wir endlich die geometrische Bedeutung von π mit Sinus und Cosinus verbinden. Die Fläche des halben Einheitskreises ist geometrisch gleich $\pi_g/2$, analytisch gleich

$$\int_{-1}^{1} \sqrt{1-x^2}dx.$$

Dabei verstehen wir unter π_g ein "geometrisches" π, das aus der Flächenformel oder der Umfangsformel für Kreise kommt. Dem steht ein "analytisches" π_a gegenüber, das die kleinste positive Nullstelle der durch die Potenzreihe dargestellten Sinusfunktion bezeichnet. Wir wollen zeigen, daß $\pi_a = \pi_g$ gilt.

Eine vernünftige Abbildung, die $[-1, 1]$ mit den trigonometrischen Funktionen verbindet, ist $x = \cos\phi$ wobei wir $\phi \in [0, \pi_a]$ nehmen. Jetzt berechnen wir das Integral mit der obigen Substitution:

$$\begin{aligned}
\int_{-1}^{1} \sqrt{1-x^2}dx &= -\int_{\pi_a}^{0} \sin(\phi) sin(\phi) d\phi \\
&= \int_{0}^{\pi_a} \sin^2(\phi) d\phi \\
&= \frac{\pi_a}{2} - \frac{1}{2}\sin(x)\cos(x)|_0^{\pi_a} \\
&= \frac{\pi_a}{2}.
\end{aligned}$$

Also stimmen π_a und π_g überein, wenn wir π_g aus der Flächenformel für den Kreis nehmen.

Als nächstes wollen wir die Interpretation der Gleichung

$$\exp(i\phi) = \cos\phi + i \cdot \sin\phi$$

am Einheitskreis nachholen, und dazu müssen wir zeigen, daß ϕ die Länge des Kreisbogens ist, der beim Übergang zwischen Polarkoordinaten und cartesischen Koordinaten zu $r\exp(i\phi) = (r\cos\phi, r\sin\phi)^T$ führt. Dazu berechnen wir allgemein die **Bogenlänge**[1] von Kurven. Nehmen wir erst einmal eine planare Kurve $(x, y(x))^T \in I\!R^2$, die nichtparametrisch als Funktion $y(x)$

[1] `http://de.wikipedia.org/wiki/L%C3%A4nge_(Mathematik)`

auf $[a,b]$ geschrieben werden kann. Die beiden Kurvenpunkte $(x,y(x))^T$ und $(x+h,y(x+h))^T$ haben den Abstand

$$\sqrt{h^2+(y(x+h)-y(x))^2}=h\sqrt{1+\left(\frac{y(x+h)-y(x)}{h}\right)^2}.$$

Jetzt verwenden wir eine Zerlegung

$$a=x_0<x_1<\ldots<x_{n+1}=b$$

des Intervalls und summieren alle diese Abstände auf. Es folgt

$$\sum_{j=0}^{n}(x_{j+1}-x_j)\sqrt{1+\left(\frac{y(x_{j+1})-y(x_j)}{x_{j+1}-x_j}\right)^2},$$

und das ist eine Riemannsche Summe, die gegen

$$\int_a^b\sqrt{1+y'(x)^2}dx$$

strebt, weil wir nach dem Mittelwertsatz die Differenzenquotienten durch Ableitungswerte in $[x_j,x_{j+1}]$ ersetzen und dann integrieren können, sofern y' noch stetig ist.

Jetzt rechnen wir die Bogenlänge auf dem Einheitskreis $y(x)=\sqrt{1-x^2}$ aus. Es folgt

$$\begin{aligned}y'(x)&=\frac{-x}{\sqrt{1-x^2}}\\ y'(x)^2&=\frac{x^2}{1-x^2}\\ \sqrt{1+y'(x)^2}&=\frac{1}{\sqrt{1-x^2}}\end{aligned}$$

Wir machen die Substitution $x=\cos\phi$ und berechnen

$$\int_a^b\frac{1}{\sqrt{1-x^2}}dx=\int_{\cos^{-1}(a)}^{\cos^{-1}(b)}(-1)d\phi=\cos^{-1}(a)-\cos^{-1}(b).$$

Das Endstück von $a=\cos\phi$ bis $b=1=\cos(0)$ hat dann die Bogenlänge $\cos^{-1}(a)-\cos^{-1}(b)=\phi-0=\phi$. Damit sehen wir, daß der Kreisbogen wirklich die geometrische Länge ϕ hat. Das verbindet wieder π_a mit π_g, denn der Halbkreis bekommt analytisch die Länge π_a.

Der Vollständigkeit halber berechnen wir auch noch die Bogenlänge von allgemeinen parametrischen Kurven $t \to x(t) \in I\!R^k$. Der Beweisgang ist ähnlich wie oben: wir bilden erst einmal den Abstand zwischen zwei Kurvenpunkten $x(t_{j+1})$ und $x(t_j)$ als

$$\begin{aligned} \|x(t_{j+1}) - x(t_j)\|_2 &= \sqrt{\sum_{m=1}^{k} (x_k(t_{j+1}) - x_k(t_j))^2} \\ &= (t_{j+1} - t_j) \sqrt{\sum_{m=1}^{k} \left(\frac{x_k(t_{j+1}) - x_k(t_j)}{t_{j+1} - t_j} \right)^2}. \\ &= (t_{j+1} - t_j) \sqrt{\sum_{m=1}^{k} x_k'(\tau_{j,k})^2,}. \end{aligned}$$

Jetzt wird mit Hilfe einer Zerlegung aufsummiert und eine Riemannsche Summe gebildet, die dann gegen die Gesamt–Bogenlänge

$$\int_a^b \sqrt{\sum_{m=1}^{k} x_k'(t)^2 dt} = \int_a^b \|x'(t)\|_2 dt$$

konvergiert, sofern $t \mapsto \|x'(t)\|_2$ noch stetig ist.

Theorem 13.9 *Die Bogenlänge einer differenzierbaren parametrischen Kurve ist das Integral über die Längen der Tangentialvektoren.* □

Damit wird das Integral über Stücke des Einheitskreisrandes $t \mapsto x(t) = (\cos(t), \sin(t))^T$ noch einfacher, denn es gilt $\|x'(t)\|_2 = 1$, so dass die Bogenlänge zwischen den Punkten $x(0)$ und $x(t)$ gleich t ist, also gleich dem Winkel t im Bogenmaß.

Ab hier haben wir keine Hemmungen mehr, die zwei Definitionen der trigonometrischen Funktionen (über Reihen und geometrisch) als verträglich anzusehen.

Wenn die Ableitung x' einer Kurve x auf $[a, b]$ noch stetig ist und nirgends verschwindet, können wir die Bogenlänge als Parameter einführen, indem wir

$$\varphi(s) := \int_a^s \|x'(t)\|_2 dt \text{ für alle } s \in [a, b]$$

setzen. Dann bildet φ das Intervall $[a, b]$ streng monoton und bijektiv auf $[0, L]$ ab, wobei L die Gesamtlänge der Kurve ist. Denn nach dem Hauptsatz

der Differential- und Integralrechnung folgt $\varphi'(s) = \|x'(s)\|_2 > 0$. Die Reparametrisierung $y(\varphi(s)) := x(s)$ bzw. $y(t) := x(\varphi^{-1}(t))$ definiert dann y als Kurve auf $[0, L]$ und es gilt nach der Kettenregel

$$x'(s) = y'(\varphi(s))\varphi'(s) = y'(\varphi(s))\|x'(s)\|_2,$$

also $\|y'(t)\|_2 = 1$ für alle $t \in [0, l]$. Damit hat die Kurve y zwischen $y(0)$ und $y(t)$ immer die Bogenlänge t.

Theorem 13.10 *Jede stetig differenzierbare Kurve, deren Ableitung nicht verschwindet, läßt sich so umparametrisieren, daß ihre Bogenlänge als Kurvenparameter auftritt und alle Tangentialvektoren die Länge Eins haben.* □

13.2.2 Spezielle Reihen

Wir haben noch nachzuholen, wie die Potenzreihe

$$\log(1 + x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$$

aus (10.8) auf Seite 260 zustandekommt. Definieren wir $f(x) := \log(1 + x)$, so folgt

$$f'(x) = \frac{1}{1+x} = \frac{1}{1-(-x)} = \sum_{n=0}^{\infty} (-x)^n = \sum_{n=0}^{\infty} (-1)^n x^n \text{ für alle } |x| < 1.$$

Jetzt integrieren wir diese Reihe formell gliedweise und bekommen eine neue Reihe

$$g(x) := \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}.$$

Diese konvergiert absolut für $|x| < 1$, weil sie durch die geometrische Reihe majorisiert werden kann. Deshalb ist sie nach Satz 12.17 auf Seite 317 eine unendlich oft differenzierbare Funktion, deren Ableitung durch das gliedweise Differenzieren der Potenzreihe berechnet werden kann. Also gilt $g'(x) = f'(x)$ für alle $|x| < 1$ und es folgt aus Satz 13.6, daß g bis auf eine Konstante mit f übereinstimmt. Auswertung in $x = 0$ zeigt aber, daß diese Konstante Null ist, denn es gilt $f(x) = g(x) = 0$. Damit haben wir die Korrektheit der obigen Logarithmusreihe bewiesen. Sie konvergiert aber so miserabel, daß sie nur für sehr kleine x brauchbar ist.

Die nächste Altlast betrifft die Leibnizreihe

$$\frac{\pi}{4} = \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1}$$

die wir im Anschluß an Satz 10.3 auf Seite 256 behandelt haben. Damals konnten wir nur die Konvergenz nachweisen, nicht aber den Summenwert ausrechnen. Der Zugang verläuft schrittweise über die Übungsaufgaben

$$\tan'(x) = \frac{1}{\cos^2(x)},\ \arctan'(x) = \frac{1}{1+x^2}$$

und berechnet mit genau denselben Argumenten wie bei der Logarithmusreihe, daß

$$\begin{aligned} \arctan'(x) &= \frac{1}{1+x^2} = \frac{1}{1-(-x^2)} = \sum_{n=0}^{\infty}(-1)^n x^{2n} \\ \arctan(x) &= \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n+1}}{2n+1} \end{aligned}$$

gilt, wobei die Reihen für $|x| < 1$ absolut kovergieren. Wir beweisen hier nicht, daß sich die Reihe in der letzten Gleichung stetig auf $x = 1$ fortsetzen lässt. Aber wenn wir das hinnehmen, folgt

$$\sum_{n=0}^{\infty}(-1)^n \frac{1}{2n+1} = \arctan(1)$$

und weil für 90^o oder $\pi/4$ die Gleichungen $\sin(\pi/4) = \cos(\pi/4)$ und $\tan(\pi/4) = 1$ gelten, folgt $\arctan(1) = \pi/4$.

13.2.3 Uneigentliche Integrale

Definition 13.11 *Ist f auf $(-\infty, b]$ stetig, so definiert man* **uneigentliche Integrale**[1]

$$\int_{-\infty}^{b} f(x)dx := \lim_{a\to-\infty}\int_{a}^{b} f(x)dx$$

wenn dieser Limes existiert. Für stetige Funktionen f auf ganz $\mathbb{R}$ und beliebige Folgen $(a_n)_n \to -\infty$ und $(b_n)_n \to +\infty$ definiert man

$$\int_{-\infty}^{+\infty} f(x)dx := \lim_{n\to\infty}\int_{a_n}^{b_n} f(x)dx$$

sofern der Limes existiert und von den Folgen unabhängig ist.

[1] `http://de.wikipedia.org/wiki/Riemann-Integral%23Uneigentliche_Integrale`

Bei dieser Definition sehe man noch einmal in Definition 8.5 und Abschnitt 8.5 auf den Seiten 220 und 241 nach.

Solche Integrale treten häufig auf, und wir geben ein paar Beispiele an:

$$\begin{aligned}
\int_z^\infty x^\alpha dx &= -\frac{1}{\alpha+1}x^{\alpha+1} \text{ für alle } z>0,\ \alpha<-1\\
\int_{-\infty}^z \exp(x)dx &= \exp(z)\\
\int_{-\infty}^\infty \frac{1}{1+x^2}dx &= \pi \text{ (warum?)}\\
\text{Gammafunktion: } \Gamma(x) &:= \int_0^\infty \exp(-t)t^{x-1}dt, \text{ für alle } x>0\\
\int_0^\infty \exp(-s\cdot t)f(s)ds &=: F(t)\\
&\quad (F = \textbf{Laplace-Transformierte} \text{ von } f)
\end{aligned}$$

Die **Gammafunktion**[1] und die **Laplace-Transformation**[2] treten an verschiedenen Stellen der Mathematik auf, können aber hier nicht genauer untersucht werden.

Ist $f(t) = u(t) + i \cdot v(t)$ eine komplexwertige Funktion, aufgeteilt in Real- und Imaginärteil $u(t)$ und $v(t)$, und sind diese beiden in $[a, b]$ integrierbar, so ist das **komplexe Integral** von f eine komplexe Zahl, nämlich

$$\int_a^b f(t)dt = \int_a^b u(t)dt + i \cdot \int_a^b v(t)dt.$$

Solche Funktionen sind sehr wichtig in der Signalverarbeitung und Elektrotechnik, weil sie Schwingungen und Wechselspannungen beschreiben. Auf ihnen definiert man die **Fourier-Transformation**[3] durch

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty \exp(-ixt)f(t)dt &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty (\cos(t)+i\sin(t))(u(t)+i\cdot v(t))dx\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty (\cos(t)u(t)-\sin(t)v(t))dt\\
&\quad +i\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty (\cos(t)v(t)+\sin(t)u(t))dt\\
&=: \hat{f}(x)\in\mathbb{C}\\
&\quad (\hat{f} = \textbf{Fourier-Transformierte} \text{ von } f)
\end{aligned}$$

[1] http://de.wikipedia.org/wiki/Gammafunktion
[2] http://de.wikipedia.org/wiki/Laplace-Transformation
[3] http://de.wikipedia.org/wiki/Fourier-Transformation

Sowohl bei der Laplace– als auch bei der Fourier–Transformation muss man genau diskutieren, auf welche Funktionen die Transformationen anwendbar sind.

Es ist überraschend, daß die Fourier–Transformation, wenn sie bei einem geeigneten Definitions– und Bildbereich invertierbar ist, eine Inverse hat, die sich fast genauso schreibt:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(+ixt) f(t) dt \quad =: \quad \overset{\vee}{f}(x) \in \mathbb{C}$$

$$(\overset{\vee}{f} = \textbf{inverse Fourier-Transformierte} \text{ von } f)$$

Leider geht es über einen Anfängertext weit hinaus, diese Transformationen genauer zu untersuchen, den sie haben sehr seltsam erscheinende Eigenschaften. Als Beispiel und Übungsaufgabe geben wir an

$$f(t) := \left\{ \begin{array}{ll} 1 & |x| \leq 1 \\ 0 & |x| > 1 \end{array} \right\}, \; \hat{f}(x) = \sqrt{\frac{2}{\pi}} \frac{\sin(x)}{x}.$$

Die Fouriertransformation einer unstetigen stückweise konstanten Funktion ist also eine (nach dem Satz von de l'Hospital und der Potenzreihe der Sinusfunktion) unendlich oft differenzierbare Funktion, die sogenannte **sinc-Funktion**[1]. Es ist aber keineswegs klar, wie man auf der sinc–Funktion die inverse Fourier–Transformation auswerten kann. Die sinc–Funktion und ihre Fouriertransformation spielen eine zentrale Rolle beim Shannon–Whittaker'schen **Sampling–Theorem** der Signalverarbeitung.

13.3 Integrale multivariater Funktionen

Stetige skalare multivariate Funktionen $f : \mathbb{R}^k \to \mathbb{R}$ kann man zunächst längs Kurven integrieren. Das untersuchen wir im nächsten Abschnitt. Man kann aber auch bezüglich einer der Variablen eine Integration ausführen und fragen, welche Eigenschaften die dann entstehende Funktion von $k-1$ Variablen hat, um dann eventuell auch noch bezüglich anderer Variablen zu integrieren. Das ergibt die Mehrfachintegrale des zweiten Abschnitts. Davon sind die danach auftretenden Gebietsintegrale zu unterscheiden.

[1] `http://de.wikipedia.org/wiki/Sinc-Funktion`

13.3.1 Kurvenintegrale

Eine stetige skalare multivariate Funktion $f : \mathbb{R}^k \to \mathbb{R}$ kann man entlang einer stetigen Kurve $x : [a, b] \to \mathbb{R}^k$ integrieren, indem man das Integral

$$\int_a^b f(x(t))dt$$

der zusammengesetzten Funktion $f \circ x$ berechnet. Das ist ein ganz normales Integral über eine reellwertige stetige Funktion. Ein reelles **Kurvenintegral**[1] ist dagegen anders definiert. Man ersetzt die ansonsten übliche Zerlegung

$$a = t_0 < t_1 < \ldots < t_{n+1} = b$$

eines reellen Intervalls $[a, b]$ durch eine Zerlegung des Bildes der Kurve x in die aufeinanderfolgenden Punkte

$$x(a) = x(t_0), x(t_1), \ldots, x(t_{n+1}) = x(b)$$

des $\mathbb{R}^k$, die dann natürlich nicht in einer Ordnungsrelation wie $<$ stehen. Die entsprechende Riemannsche Summe für ein reelles Kurvenintegral ist dann

$$\sum_{j=0}^{n} \|x(t_{j+1}) - x(t_j)\|_2 f(\xi_j)$$

mit Punkten $\xi_j = x(\tau_j),\ \tau_j \in [t_j, t_{j+1}]$ auf den Verbindungsstrecken von $x(t_j)$ und $x(t_{j+1})$. Das kann man in ein normales Integral überführen, indem man die Bogenlänge einführt und

$$\begin{aligned} & \sum_{j=0}^{n} \|x(t_{j+1}) - x(t_j)\|_2 f(\xi_j) \\ = & \sum_{j=0}^{n} (t_{j+1} - t_j) \sqrt{\sum_{m=0}^{k} \left(\frac{x_k(t_{j+1}) - x_k(t_j)}{t_{j+1} - t_j} \right)^2} f(\xi_j) \end{aligned}$$

schreibt. Das ist aber eine Riemannsche Summe für das reelle **Kurvenintegral**

$$\oint f dx := \int_a^b \|x'(t)\|_2 f(x(t))dt,$$

welches offensichtlich wohldefiniert ist, wenn die Kurve x stetig differenzierbar und f stetig ist.

[1] `http://de.wikipedia.org/wiki/Kurvenintegral`

Theorem 13.12 *Kurvenintegrale sind unabhängig von der Parametrisierung.*

Beweis: Reparametrisiert man x über $t = \varphi(\tau)$ mit einer streng monotonen und differenzierbaren Parameterabbildung φ zu einer neuen Kurve $y(\tau) := x(\varphi(\tau))$ mit gleichem Bild, so folgt

$$\begin{aligned}
\oint_C f dx &= \int_a^b \|x'(t)\|_2 f(x(t)) dt \\
&= \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} \|x'(\varphi(\tau))\|_2 f(x(\varphi(\tau)))\varphi'(\tau) d\tau \\
&= \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} \|y'(\tau)\|_2 f(y(\tau)) d\tau \\
&= \oint_C f dy
\end{aligned}$$

weil man nach der Kettenregel $y'(\tau) = x'(\varphi(\tau))\varphi'(\tau)$ hat. □

Das Integral eines **Vektorfeldes** $F : \mathbb{R}^k \to \mathbb{R}^k$ entlang einer stetig differenzierbaren Kurve $x : [a, b] \to \mathbb{R}^k$ ist als

$$\int_a^b F(x(t))^T x'(t) dt$$

gemeint, d.h. man bildet das Skalarprodukt des Vektors $F(t)$ mit $x'(t)$ an jeder Stelle und integriert das skalare Ergebnis. Dann liefert der Hauptsatz der Differential-und Integralrechnung zusammen mit der Kettenregel sofort

Theorem 13.13 *Kurvenintegrale von stetigen Gradientenfeldern $F = \nabla f$ entlang stetig differenzierbaren Kurven x hängen nicht vom Kurvenverlauf, sondern nur vom Anfangs– und Endpunkt der Kurve ab, genauer von der Potentialdifferenz*

$$f(x(b)) - f(x(a)) = \int_a^b (\nabla f)(x(t)) x'(t) dt.$$

□

Komplexe Kurvenintegrale wollen wir hier nicht vertieft behandeln, aber zumindestens gegen die reellen Kurvenintegrale abgrenzen. Sie sind wie Kurvenintegrale zweidimensionaler Vektorfelder entlang von Kurven im $\mathbb{R}^2$ definiert, aber mit komplexer Interpretation der Punkte des $\mathbb{R}^2$. Ist also $f : \mathbb{C} \to \mathbb{C}$

eine komplexwertige Funktion einer komplexen Variablen, so bildet man zu einer Kurve $t \mapsto x(t) \in \mathbb{C}$ die Riemannschen Summen

$$\begin{aligned} & \sum_{j=0}^{n} (x(t_{j+1}) - x(t_j)) f(\xi_j) \\ = & \sum_{j=0}^{n} (t_{j+1} - t_j) \frac{x(t_{j+1}) - x(t_j)}{t_{j+1} - t_j} f(\xi_j) \end{aligned}$$

die gegen das "normal" auswertbare komplexwertige Integral

$$\oint f dx := \int_a^b x'(t) f(x(t)) dt$$

konvergieren, wenn x stetig differenzierbar und f stetig ist. Auch hier bekommt man Invarianz gegen Reparametrisierung.

13.3.2 Mehrfache Integrale

[1] Wir behanden den einfachen Fall einer Funktion $f : [a,b] \times [\alpha, \beta] \to I\!R$. Ist f dort stetig, so ist für festes $x_0 \in [\alpha, \beta]$ die Funktion $f(t, x_0)$ als Funktion von $t \in [a,b]$ stetig. Dazu nehme man ein $t_0 \in [a,b]$ und gebe sich ein $\epsilon > 0$ vor. Es gibt dann ein $\delta > 0$, so dass für alle $(t,x) \in [a,b] \times [\alpha, \beta]$ mit $\|(t,x) - (t_0, x_0)\|_2 < \delta$ auch $|f(t,x) - f(t_0, x_0)| < \epsilon$ folgt. Das liefert aber auch $|f(t, x_0) - f(t_0, x_0)| < \epsilon$ für alle $t \in [a,b]$ mit $|t - t_0| < \delta$, also die verlangte Stetigkeit der eingeschränkten Funktion.

Also ist

$$g(x) := \int_a^b f(t,x) dt \text{ für alle } x \in [\alpha, \beta]$$

eine wohldefinierte Funktion. Ist sie stetig? Dazu bilden wir eine Differenz

$$g(x+h) - g(x) \quad = \quad \int_a^b (f(t, x+h) - f(t,x)) dt$$

und nutzen aus, daß nach Satz 11.35 auf Seite 294 die Funktion f auf ihrem gesamten Definitionsbereich gleichmäßig stetig ist. Zu beliebigem $\epsilon > 0$ gibt es also ein $\delta > 0$, so dass für alle $(t,x)^T, (\tau, \xi)^T \in [a,b] \times [\alpha, \beta]$ mit $\|(t,x)^T - (\tau, \xi)^T\|_\infty < \delta$, d.h. $|t - \tau| < \delta$ und $|x - \xi| < \delta$ auch $|f(t,x) - f(\tau, \xi)| < \epsilon$ gilt. Wenn oben $|h| < \delta$ gilt, folgt

$$|g(x+h) - g(x)| \leq \int_a^b |f(t, x+h) - f(t,x)| dt < \epsilon (b-a)$$

[1] http://de.wikipedia.org/wiki/Integralrechnung%23Integration_.C3.BCber_mehrdimensionale_Ber

und daraus die Stetigkeit von g in x. Diese Stetigkeit ist sogar gleichmäßig, weil die obige Argumentation gar nicht von x abhängt. Also kann man g integrieren und definiert das **Mehrfachintegral** von f durch

$$\int_\alpha^\beta \int_a^b f(t,x)dtdx := \int_\alpha^\beta g(x)dx = \int_\alpha^\beta \left(\int_a^b f(t,x)dt \right) dx.$$

Theorem 13.14 *Jede auf einem endlichen cartesischen Produkt abgeschlossener und beschränkter reeller Intervalle stetige Funktion ist dort durch ein Mehrfachintegral integrierbar, das man duch Hintereinanderausführung der jeweiligen eindimensionalen Integrale ausrechnen kann. Als Satz von* **Fubini**[1] *bezeichnet man die Tatsache, dass das Mehrfachintegral von der Reihenfolge der Einzelintegrationen unabhängig ist.*

Beweis: Wir verzichten auf eine Induktion über die Raumdimension und verzögern den Beweis für den Satz von Fubini auf $[a,b] \times [\alpha,\beta]$ etwas, weil wir dazu noch ein wichtiges Hilfsmittel brauchen.

Theorem 13.15 *Es sei f auf $I := [a,b]\times[\alpha,\beta] \to I\!R$ stetig, und die partielle Ableitung $\frac{\partial f}{\partial x}$ sei ebenfalls dort stetig. Dann gilt*

$$\frac{d}{dx} \int_a^b f(t,x)dt = \int_a^b \frac{\partial f}{\partial x}(t,x)dt,$$

d.h. man kann Integration bezüglich t und Differentiation bezüglich x vertauschen.

Beweis: Man setzt

$$g(x) := \int_a^b f(t,x)dt \text{ für alle } x \in [\alpha,\beta]$$

und sieht sich einen Differenzenquotienten an:

$$\begin{aligned} \frac{g(x+h)-g(x)}{h} &= \frac{1}{h}\left(\int_a^b f(t,x+h)dt - \int_a^b f(t,x)dt \right) \\ &= \int_a^b \frac{f(t,x+h)-f(t,x)}{h} dt \\ &= \int_a^b \frac{\partial f}{\partial x}(t,\xi(t,x,h))dt \end{aligned}$$

[1] `http://de.wikipedia.org/wiki/Satz_von_Fubini`

mit einem $\xi(t,x,h)$ zwischen x und $x+h$. Ferner folgt

$$\frac{g(x+h)-g(x)}{h} - \int_a^b \frac{\partial f}{\partial x}(t,x)dt \;=\; \int_a^b \left(\frac{\partial f}{\partial x}(t,\xi(t,x,h)) - \frac{\partial f}{\partial x}(t,x)\right) dt$$

und man nutzt dann die gleichmäßige Stetigkeit von $\frac{\partial f}{\partial x}$ aus, um das rechts stehende Integral "kleinzukriegen". Diese Details werden unterdrückt. □

Jetzt können wir den Beweis des Satzes von Fubini nachholen, indem wir

$$\begin{aligned} g_1(y) &:= \int_\alpha^y \int_a^b f(t,x)dtdx \\ g_2(y) &:= \int_a^b \int_\alpha^y f(t,x)dxdt \end{aligned}$$

definieren. Nach dem Hauptsatz und nach dem vorigen Satz gilt

$$\begin{aligned} g_1'(y) &= \int_a^b f(t,y)dt \\ g_2'(y) &= \int_a^b \frac{\partial}{\partial y} \int_\alpha^y f(t,x)dxdt \\ &= \int_a^b f(t,y)dt \end{aligned}$$

so dass sich g_1 und g_2 nach Satz 13.6 nur um eine Konstante unterscheiden, die wegen $g_1(a) = g_2(a) = 0$ Null ist. □

Bisher waren unsere Mehrfachintegrale immer gegeben durch nacheinander auszuführende Einzelintegrale mit konstanten Grenzen. geometrisch gesehen sind es also Integrale über Rechtecke, Quader oder im allgemeinen **Parallelepipede** der Form $[a_1,b_1] \times [a_2,b_2] \times \ldots \times [a_k,b_k] \subset I\!R^k$. Die Integrationsgrenzen müssen nicht immer konstant sein, wenn man auf die Vertauschbarkeit der Einzelintegrationen verzichtet, und dann kann das Integrationsgebiet komplizierter aussehen. Man ist dabei aber zunächst an cartesische Koordinaten gebunden, und die äußerste Integration muss über ein Intervall erstreckt werden. Will man z.B. über den Einheitskreis integrieren, schreibt man ihn entweder als

$$\{(x,y) \;:\; -1 \le x \le 1,\; -\sqrt{1-x^2} \le y \le +\sqrt{1-x^2}\}$$

oder als

$$\{(x,y) \;:\; -1 \le y \le 1,\; -\sqrt{1-y^2} \le x \le +\sqrt{1-y^2}\}$$

und kann dann eine dort definierte Funktion f über

$$\int_{-1}^{+1} \int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} f(x,y)dydx$$

oder

$$\int_{-1}^{+1} \int_{-\sqrt{1-y^2}}^{+\sqrt{1-y^2}} f(x,y)dxdy$$

integrieren. Dass man dasselbe Ergebnis herausbekommt, soll hier nicht bewiesen werden. Die obige Technik ist das Standardverfahren zum Ausrechnen konkreter mehrfacher Integrale. Man zerlegt den Integrationsbereich in Teile, die mit dieser Technik behandelbar sind, und summiert dann die Teilintegrale auf. Genaugenommen müßte man an dieser Stelle den Begriff des **Gebietsintegrals** definieren, aber das wird auf später verschoben.

Aber wir wollen noch illustrieren, was der Gaußsche Integralsatz[1] besagt. Dazu sei F ein Vektorfeld auf dem Rechteck $R := [a_1, b_1] \times [a_2, b_2] \subset I\!R^2$ mit Werten im $I\!R^2$. Dann integrieren wie die Divergenz von F auf R und erhalten

$$\begin{aligned}
& \int_{a_1}^{b_1} \int_{a_2}^{b_2} div\ F(x_1,x_2)dx_1dx_2 \\
= & \int_{a_1}^{b_1} \int_{a_2}^{b_2} \left(\frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2}\right)(x_1,x_2)dx_1dx_2 \\
= & \int_{a_1}^{b_1} \int_{a_2}^{b_2} \frac{\partial F_2}{\partial x_2}dx_2dx_1 + \int_{a_2}^{b_2} \int_{a_1}^{b_1} \frac{\partial F_1}{\partial x_1}dx_1dx_2 \\
= & \int_{a_1}^{b_1} (F_2(x_1,b_2) - F_2(x_1,a_2))dx_1 + \int_{a_2}^{b_2} (F_1(b_1,x_2) - F_1(a_1,x_2))dx_2 \\
= & \int_{a_1}^{b_1} F_2(x_1,b_2)dx_1 - \int_{a_1}^{b_1} F_2(x_1,a_2)dx_1 \\
& + \int_{a_2}^{b_2} F_1(b_1,x_2)dx_2 - \int_{a_2}^{b_2} F_1(a_1,x_2)dx_2.
\end{aligned}$$

Die Kanten des Rechtecks sind Geradenstücke, und die Geraden haben jeweils gewisse nach aussen zeigenden Normalenvektoren. Genauer:

$$\begin{array}{lll}
[a_1,b_1] \times b_2 & \text{hat äussere Normale} & e_2 \\
[a_1,b_1] \times a_2 & \text{hat äussere Normale} & -e_2 \\
[a_2,b_2] \times b_1 & \text{hat äussere Normale} & e_1 \\
[a_2,b_2] \times a_1 & \text{hat äussere Normale} & -e_1.
\end{array}$$

[1] `http://de.wikipedia.org/wiki/Gau%C3%9Fscher_Integralsatz`

Damit folgt

$$\begin{aligned}
& \int_{a_1}^{b_1}\int_{a_2}^{b_2} div\ F(x_1,x_2)dx_1dx_2 \\
= & \int_{a_1}^{b_1} F(x_1,b_2)^T n(x_1,b_2)dx_1 + \int_{a_1}^{b_1} F(x_1,a_2)^T n(x_1,a_2)dx_1 \\
& + \int_{a_2}^{b_2} F(b_1,x_2)^T n(b_1,x_2)dx_2 + \int_{a_2}^{b_2} F(a_1,x_2)^T n(a_1,x_2)dx_2 \\
=: & \int_{\partial R} F(z)^T n(z)dz
\end{aligned}$$

wobei ∂R der Rand des Rechtecks R sei und man sich vorstellen sollte, dass $\int_{\partial R}$ als Summe der Integrale über alle Randkanten von $I\!R$ zu verstehen ist. Der obige Spezialfall lässt sich verallgemeinern zum **Divergenzsatz** oder **Gaußschen Integralsatz**[1]

$$\int_G div\ F(x)dx = \int_{\partial G} F(z)^T n(z)dz$$

der das Integral über die Divergenz eines Vektorfelds F auf einem Gebiet G als Integral über den Rand ∂G des Gebiets schreibt, wobei der Integrand das Skalarprodukt von F mit der nach aussen gerichteten Normalen auf dem Gebietsrand ist. Der Beitrag der im Gebiet liegenden Quellen des Vektorfelds ist das Integral über die Divergenz, und dieser Beitrag ist genau gleich dem Überschuss des aus dem Gebiet herausfliessenden über den hineinfliessenden Anteil. Man sieht, dass es nötig ist, Integrale über Gebiete und ihre Ränder umfassender zu definieren. Es sollte aber bis hier schon klar sein, daß der Satz für endliche Vereinigungen von Rechteckgebieten gilt, denn die Randintegrale über die inneren Ränder, an denen die Gebiete aneinanderstoßen, heben sich auf, weil die Normalen entgegengesetztes Vorzeichen haben.

13.3.3 Gebietsintegrale

Integrale über Mengen, die nicht als cartesische Produkte von Intervallen (Parallelepipede) geschrieben werden können, und die sich auch nicht einfach als Mehrfachintegrale mit variablen Grenzen schreiben lassen, kann man mit einer **Substitutionsregel** berechnen, die analog zu (13.8) strukturiert ist. Mit einer Abbildung g, die ein Parallelepiped P auf eine andere Menge $G = g(P)$ abbildet, kann man über die Substitution $x = g(t)$ die Gleichung

$$\int_G f(x)dx = \int_P f(g(t))\det\left(\nabla g(t)\right)dt$$

[1] `http://de.wikipedia.org/wiki/Gau%C3%9Fscher_Integralsatz`

benutzen. Wir sehen diese Beziehung zunächst als Definition eines allgemeinen Gebietsintegrals auf G, obwohl letzteres in der Standardliteratur anders definiert wird, und dann ist die obige Gleichung keine Definition, sondern ein Ergebnis, nämlich der **Transformationssatz**[1] für mehrdimensionale Integrale. Die erforderlichen Voraussetzungen haben wir nicht angegeben, holen das aber jetzt nach. Die Abbildung g muß eine stetig differenzierbare Bijektion zwischen R und G sein, und die Determinante der Jacobimatrix ∇g sollte im Innern von R nirgends verschwinden. Für f reicht Stetigkeit, um die rechte Seite der obigen Gleichung legal zu machen. Dann gilt die Gleichung und kann als Definition der linken Seite durch die rechte dienen.

Die Gleichung gilt aber sinngemäß auch für allgemeine Transformationen zwischen allgemeinen Integrationsgebieten, sofern die beiden Gebiete durch eine Abbildung g wie oben bijektiv auf ein Parallelepiped abbildbar sind. Für kompliziertere Integrationsgebiete muß man sehr viel mehr Arbeit investieren.

Gebietsintegrale schreibt man, obwohl sie sich über höherdimensionale Mengen erstrecken, nur mit **einem** Integralzeichen und gibt unten das Gebiet an. Ist das Gebiet in cartesischen Koordinaten durch ein Parallelepiped mit variablen Grenzen beschreibbar, kann man das Gebietsintegral durch ein Mehrfachintegral ersetzen und ausrechnen.

Als Beispiel nehmen wir die Integration einer in cartesischen Koordinaten $(x, y) \in \mathbb{R}^2$ definierten Funktion f auf einem Gebiet G, das sich besser in Polarkoordinaten (r, φ) des $\mathbb{R}^2$ angeben läßt, z.B. auf einem Sektor S eines Kreisrings. Man hat dann $\varphi \in [\varphi_0, \varphi_1]$ und $r \in [r_0, r_1]$ und nimmt die Transformation

$$\begin{aligned}
(x, y)^T &= g(r, \varphi) = \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} \\
\nabla g &= \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} \\
\det(\nabla g) &= \det \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} = r
\end{aligned}$$

vor. Es folgt

$$\int_S f(z)dz = \int_{\varphi_0}^{\varphi_1} \int_{r_0}^{r_1} f(r\cos\varphi, r\sin\varphi) \cdot r \cdot dr d\varphi.$$

[1] `http://de.wikipedia.org/wiki/Transformationssatz`

In Büchern über Physik oder Ingenieurwissenschaften findet man die lapidare Aussage, daß die Integration in zweidimensionalen Polarkoordinaten über das "Flächenelement" $dz = rdrd\varphi$ erfolgt, und damit ist der obige Sachverhalt gemeint.

Als Beispiel nehmen wir uns das Integral

$$\int_{\mathbb{R}^2} \exp(-\|z\|_2^2)dz = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} exp(-(x^2+y^2))dxdy = \left(\int_{-\infty}^{+\infty} exp(-x^2)dx\right)^2$$

vor und betrachten zuerst das Integral in Polarkoordinaten auf einer Kreisscheibe:

$$\int_0^R \int_{-\pi}^{+\pi} \exp(-r^2)rd\varphi dr = 2\pi \int_0^R \exp(-r^2)rdr.$$

Die Transformation auf Polarkoordinaten beschert uns netterweise ein r im Integranden, so daß man leicht mit der Substitutionsformel und $t := r^2$ weiterkommt zu

$$\begin{aligned} 2\pi \int_0^R \exp(-r^2)rdr &= \pi \int_0^R \exp(-r^2)2rdr \\ &= \pi \int_0^{R^2} \exp(-t)dt \\ &= \pi(1-\exp(-R^2)). \end{aligned}$$

Wenn wir nun den Grenzübergang $R \to \infty$ ausführen, erhalten wir

$$\int_{\mathbb{R}^2} \exp(-\|z\|_2^2)dz = \pi \text{ und } \int_{-\infty}^{+\infty} exp(-x^2)dx = \sqrt{\pi}.$$

Genaugenommen haben wir noch nachzuholen, daß das uneigentliche Integral nicht davon abhängt, dass wir uns auf eine sehr spezielle Weise an Unendlich herangeschlichen haben. Aber das lassen wir weg, denn die Gaußglocke $\exp(-\|x\|_2^2)$ strebt sehr schnell gegen Null, wenn das Argument $\|x\|_2$ gegen Unendlich strebt. Im $\mathbb{R}^n$ folgt aus der obigen Argumentation sofort auch

$$\int_{\mathbb{R}^n} \exp(-\|z\|_2^2)dz = (\pi)^{\frac{n}{2}},$$

und für $n = 1$ ist das Integral mit Schulmethoden nicht zu knacken.

Zur weiteren Veranschaulichung wollen wir jetzt das Volumen von Rotationskörpern berechnen. Dazu zeichne man ein cartesisches dreidimensionales Koordinatensystem hin (Vorlesung) und gebe eine auf $[z_0, z_1]$ definierte positive reelle Funktion f von z vor, deren Rotation um die z–Achse dann einen

Rotationskörper beschreibt. Wie kann man den dadurch definierten Körper K parametrisieren und dann das Volumen als Gebietsintegral der Funktion 1 schreiben?

Wir verwenden **Zylinderkoordinaten**

$$(x, y, z)^T = (r\cos\varphi, r\sin\varphi, z)^T =: g(r, \varphi, z)$$

und beschreiben den Körper (Zeichnung in der Vorlesung) durch

$$\begin{array}{ccccc} z_0 & \leq & z & \leq & z_1 \\ 0 & \leq & \varphi & < & 2\pi \\ 0 & \leq & r & \leq & f(z). \end{array}$$

Die Jacobimatrix ist

$$\begin{pmatrix} \cos\varphi & -r\sin\varphi & 0 \\ \sin\varphi & r\cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

mit Determinante r. Also folgt die schöne Formel

$$\int_K 1ds = \int_{z_0}^{z_1}\int_0^{2\pi}\int_0^{f(z)} rdrd\varphi dz = 2\pi\int_{z_0}^{z_1}\frac{f^2(z)}{2}dz = \pi\int_{z_0}^{z_1} f^2(z)dz.$$

Das Volumen einer Kugel mit Radius R ist damit

$$\pi\int_{-R}^{+R}(R^2 - z^2)dz = \frac{4}{3}\pi R^3.$$

Ebenso kann man mit Zylinderkoordinaten (oder als Rotationskörper) das Volumen eines kühlturmähnlichen Rotationshyperboloids

$$\{(x, y, z)^T \in I\!R^3 \ : \ x^2 + y^2 \leq R^2 + z^2, \ z_0 \leq z \leq z_1\}$$

ausrechnen (Übung). Ferner sollte klar sein, wie man allgemeine Funktionen auf Rotationskörpern integriert, denn in Zylinderkoordinaten ist im Physikerjargon "das Volumenelement gleich $rdrd\varphi dz$".

13.3.4 Flächenintegrale

Wie bei den Kurvenintegralen geht es jetzt darum, skalare Funktionen auf Flächen so zu integrieren, daß das Ergebnis nicht von der Parametrisierung der Fläche abhängt. Zuerst beschreiben wir parametrisierte Flächen durch eine Abbildung

$$g \ : \ (u, v)^T \mapsto g(u, v) := (x(u, v), y(u, v), z(u, v))^T$$

von einem zweidimensionalen in ein dreidimensionales cartesisches Koordinatensystem. Die Jacobimatrix ist

$$\nabla g = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{pmatrix} =: \begin{pmatrix} \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{pmatrix} =: \begin{pmatrix} g_u & g_v \end{pmatrix},$$

aber es ist nicht ohne weiteres klar, wie eine solche Abbildung den Flächeninhalt transformiert.

Wir studieren das an einer 3×2–Matrix A mit den beiden Spalten $a = Ae_1, b = Ae_2$. Wenn diese Vektoren linear unabhängig sind, spannen sie eine Hyperebene auf, deren Normale bis auf einen Faktor durch das **Vektorprodukt**

$$a \times b := (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1)^T$$

gegeben ist, denn es gilt

$$a^T(a \times b) = b^T(a \times b) = 0,$$

wie man leicht nachrechnet oder aus Abschnitt 6.7 noch weiß. Der Flächeninhalt der Bildmenge

$$\{\alpha \cdot a + \beta \cdot b \ : \ \alpha, \beta \in [0,1]\}$$

des Einheitsquadrats auf der Bildebene ist numerisch gleich dem Volumen der durch a, b, c aufgespannten Parallelepipeds, wenn $c = \frac{a \times b}{\|a \times b\|_2}$ gilt, weil die "Dicke" des Parallelepipeds gleich Eins ist. Also ist die Fläche gleich dem Betrag von

$$\det(c, a, b) = \frac{\det(a \times b, a, b)}{\|a \times b\|_2}.$$

Nun folgt aber durch Ausrechnen

$$\det(a \times b, a, b) = \det \begin{pmatrix} a_2b_3 - a_3b_2 & a_3b_1 - a_1b_3 & a_1b_2 - a_2b_1 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} = \|a \times b\|_2^2$$

und damit ist die Fläche des Bildbereichs gleich $\|a \times b\|_2$. Jetzt ist die Formel

$$\int_{u_0}^{u_1} \int_{v_0}^{v_1} \| \left(\frac{\partial g}{\partial u} \times \frac{\partial g}{\partial v} \right)(u, v) \|_2 dv du$$

für den Flächeninhalt eines auf $[u_0, u_1] \times [v_0, v_1]$ über g parametrisierten Flächenstücks plausibel, und sie gilt, sofern die ersten partiellen Ableitungen

von g stetig sind und als Vektoren nirgends verschwinden. Wir werden später mit Hilfe der Substitutionsregel nachweisen, daß diese Formel invariant ist gegenüber Reparametrisierungen der Fläche.

Jetzt wird auch verständlich, wieso man das Integral einer skalaren Funktion f auf einer solchen Fläche $\mathcal{F} := g([u_0, u_1] \times [v_0, v_1])$ durch das **Flächenintegral**

$$\int_{\mathcal{F}} f(x)dx := \int_{u_0}^{u_1} \int_{v_0}^{v_1} f(g(u,v)) \cdot \| \left(\frac{\partial g}{\partial u} \times \frac{\partial g}{\partial v} \right)(u,v) \|_2 dv du$$

definiert und damit Unabhängigkeit von der Parametrisierung bekommt.

Wir halten noch fest, dass die Tangentialebene zur Fläche in $g(u,v)$ durch die Vektoren $\frac{\partial g}{\partial u}$ und $\frac{\partial g}{\partial v}$ aufgespannt wird, und dass deshalb der Vektor $\frac{\partial g}{\partial u} \times \frac{\partial g}{\partial v}$ auf der Fläche senkrecht steht, d.h. die Flächennormale beschreibt. Ob er bei einer Fläche, die einen beschränkten Körper umschließt, nach "außen" oder nach "innen" zeigt, ist nicht klar und erfordert Sonderüberlegungen, die wegen der Existenz des Möbiusbandes und der Boy'schen Fläche alles andere als selbstverständlich sind. Man nennt die stetige und eindeutige Zuweisung einer Flächennormale zu den Punkten einer Fläche auch **Orientierung**, weil man damit ein klares "Außen" und "Innen" definiert. Das geht aber auf den genannten Flächen nicht, sie sind nur lokal und nicht global orientierbar. Der **Normaleneinheitsvektor** wird später auch noch wichtig, er ist bis auf das Vorzeichen gleich

$$n := \frac{\frac{\partial g}{\partial u} \times \frac{\partial g}{\partial v}}{\| \frac{\partial g}{\partial u} \times \frac{\partial g}{\partial v} \|_2}.$$

An dieser Stelle kann man eine kleine Überlegung einschieben, die zeigt, wieso eine Umparametrisierung keine Rolle spielt. Schreibt man die Parameter (u,v) als Bilder eine Umparametrisierung $(u,v) = \varphi(r,s)$ mit zwei neuen Parametern r, s, so hat man die neue Flächenfunktion $h(r,s) := g(\varphi(r,s))$ und es folgt $(\nabla h)(r,s) = (\nabla g)(\varphi(r,s)) \circ (\nabla \varphi)(r,s)$ nach der Kettenregel. Schreibt man

$$\begin{aligned}
g_u &:= \frac{\partial g}{\partial u} \in I\!R^3 \\
g_v &:= \frac{\partial g}{\partial v} \in I\!R^3 \\
h_r &:= \frac{\partial h}{\partial r} \in I\!R^3 \\
h_s &:= \frac{\partial h}{\partial s} \in I\!R^3 \\
\nabla \varphi &=: T \in I\!R^{2\times 2} \\
(h_r, h_s) &= (g_u, g_v) \circ T \text{ (Kettenregel)}
\end{aligned}$$

so zeigt eine elementare Rechnung (Vorlesung...), die man mit allgemeinen Vektoren des $\mathbb{R}^3$ anstellen kann, daß

$$h_r \times h_s = (\det T) \cdot (g_u \times g_v)$$

gilt. Das ist geometrisch klar, denn wir haben schon gesehen, daß $\|a \times b\|_2$ für zwei Vektoren $a, b \in \mathbb{R}^3$ den Flächeninhalt des aufgespannten Parallelogramms angibt. Transformiert man beide Vektoren im Urbildraum mit derselben Transformation T, so ändert sich der Flächeninhalt um den Faktor $\det T$. Damit folgt die Reparametrisierungsinvarianz aus der Substitutionsformel:

$$\begin{aligned} & \int_r \int_s f(h(r,s)) \|h_r \times h_s\|_2 ds dr \\ = & \int_r \int_s f(h(r,s)) \|(g_u \times g_v)(h(r,s))\|_2 \det(\nabla h)(r,s) ds dr \\ = & \int_v \int_u f(u,v) \|(g_u \times g_v)(u,v)\|_2 du dv. \end{aligned}$$

Setzt man $f = 1$, so hat man damit auch die Reparametrisierungsinvarianz des Flächeninhaltsintegrals bewiesen.

Für die Oberfläche einer Kugel im $\mathbb{R}^3$ mit Radius r kann man die Parametrisierung

$$g(\theta, \varphi) = r(\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi)^T,\ \varphi \in [0, \pi],\ \theta \in [0, 2\pi]$$

nehmen. Man berechnet dann (Übung)

$$\|g_\theta \times g_\varphi\|_2 = r^2 \sin\varphi$$

was die Ausartung der Parametrisierung im Süd– und Nordpol demonstriert. Das Integral einer Funktion $f(\theta, \varphi)$ auf Kugelsegmenten ist dann als Mehrfachintegral

$$r^2 \int_{\varphi_0}^{\varphi_1} \int_{\theta_0}^{\theta_1} f(\theta, \varphi) d\theta \sin\varphi d\varphi$$

ausrechenbar. Ingenieure und Physiker würden sagen, das Oberflächenelement einer Kugel mit Radius r sei $r^2 \sin\varphi d\theta d\varphi$ in Kugelkoordinaten.

Jetzt rechnen wir in Zylinderkoordinaten die Oberfläche von Rotationskörpern K aus, wobei wir dieselben Bezeichnungen wie oben verwenden. Die Fläche sei parametrisiert durch $(u, v) = (\varphi, z)$ als

$$(x, y, z)^T = (f(z)\cos\varphi, f(z)\sin\varphi, z) =: g(\varphi, z).$$

Wir bekommen die Vektoren

$$\begin{aligned} g_\varphi &= (-f(z)\sin\varphi, f(z)\cos\varphi, 0)^T \\ g_z &= (f'(z)\cos\varphi, f'(z)\sin\varphi, 1)^T \\ g_\varphi \times g_z &= f(z)\cdot(\cos\varphi, \sin\varphi, -f'(z))^T \\ \|g_\varphi \times g_z\|_2 &= f(z)\sqrt{1+f'(z)^2} \end{aligned}$$

und die Oberfläche

$$\int_{\partial K} 1ds = \int_{z_0}^{z_1}\int_0^{2\pi} f(z)\sqrt{1+f'(z)^2}d\varphi dz = 2\pi\int_{z_0}^{z_1} f(z)\sqrt{1+f'(z)^2}dz.$$

Für die Kugel mit Radius R folgt die Fläche

$$2\pi\int_{-R}^{+R}\sqrt{R^2-z^2}\sqrt{1+\frac{z^2}{R^2-z^2}}dz = 2\pi\int_{-R}^{+R} Rdz = 4\pi R^2.$$

13.4 Anwendungen multivariater Integrale

13.4.1 Integralsätze

Wir wissen nun, was Gebiets– und Flächenintegrale skalarer Funktionen sind und wie man sie über Parametrisierungen ausrechnen kann. Dabei haben wir uns auf Gebiete und Flächen beschränkt, die Bilder von Rechtecken oder Quadern unter stetig differenzierbaren und gutartigen Abbildungen sind. Natürlich lassen sich Gebiets– und Flächenintegrale auch sehr viel allgemeiner definieren, und es gelten auch dann interessante und praktisch wichtige Sätze und Rechenregeln. Die genauen Voraussetzungen werden wir nicht angeben, denn sie können bei Integration über pathologische Definitionsbereiche problematisch sein. Wir setzen von allen auftretenden Ableitungen Existenz und Stetigkeit voraus. Alle Gebiete bzw. Flächen seien über Abbildungen auf cartesischen Produkten von abgeschlossenen und beschränkten Intervallen parametrisiert, deren Jacobimatrizen immer maximalen Rang haben sollen. Unter diesen Umständen geht nichts schief.

Der erste wichtige Resultat ist der **Divergenzsatz** oder **Gaußsche Integralsatz**, den wir auf reinen Parallelepipeden schon kennen:

$$\int_G div\ F(x)dx = \int_{\partial G} F^T(y)n(y)dy.$$

Dabei steht links ein Gebietsintegral über eine skalare Funktion, die Divergenz eines Vektorfelds F ist. Rechts steht ein Flächenintegral über den kompletten Rand ∂G des Gebiets G, und der Integrand ist das Skalarprodukt von

F mit dem nach außen gerichteten Normaleneinheitsvektor n auf der Fläche. Das Gebiet G sollte beschränkt und durch die Fläche oder Kurve ∂G berandet sein. Denn der Satz gilt sinngemäß in allen Raumdimensionen. Im $\mathbb{R}^2$ betrifft er zweidimensionale Gebiete G, die durch eine Kurve ∂G umfahren werden, und die beiden auftretenden Integrale sind ein zweidimensionales Gebiets– und ein Kurvenintegral. Im $\mathbb{R}^3$ hat man dann dreidimensionale Körper G mit einer Randfläche ∂G und je ein dreidimensionales Gebiets– und ein zweidimensionales Flächenintegral. Man sieht an unserem konkreten Beweis für Parallelepipede aus Abschnitt 13.3.2 auf Seite 378, daß der Satz eine mehrdimensionale Variante des Hauptsatzes der Differential– und Integralrechnung ist, der obendrein auch auf "stückweise glatt berandeten" Gebieten wie Rechtecken der Quadern gilt. In der physikalischen Anschauung besagt der Divergenzsatz, daß alles, was die Quellstärke (Divergenz) eines Vektorfeldes im Innern eines Gebietes produziert, den Rand des Gebietes auch verlassen muß.

Jetzt wenden wir den Divergenzsatz auf ein spezielles Vektorfeld der Form $F(x) = f(x)\nabla g(x)$ an, wobei f und g hinreichend oft stetig differenzierbare skalare Funktionen sind. Es folgt

$$\begin{aligned} div\ F &= f \cdot\ div(\nabla g) + (\nabla g) \cdot (\nabla f) \\ &= f \cdot \Delta g + (\nabla g)(\nabla f)^T \\ \int_G (f \cdot \Delta g + (\nabla g)(\nabla f)^T)(x)dx &= \int_{\partial G} f(y)(\nabla g)n(y)dy, \end{aligned}$$

und das ist die erste **Greensche Formel**. Subtrahiert man eine zweite, die durch Vertauschen von f und g entsteht, ergibt sich die zweite Greensche Formel

$$\int_G (f \cdot \Delta g - g \cdot \Delta f)(x)dx = \int_{\partial G} (f(y)(\nabla g) - g(y)(\nabla f))n(y)dy.$$

Die dritte ist ein Spezialfall der ersten für $f = 1$:

$$\int_G (\Delta g)(x)dx = \int_{\partial G} ((\nabla g)n)(y)dy.$$

13.4.2 Partielle Differentialgleichungen

In diesem Abschnitt geben wir einen Ausblick auf einige zentrale Probleme des Wissenschaftlichen Rechnens, wie sie in Göttingen in der entsprechenden Studienrichtung des Bachelor-Master-Studiengangs "Angewandte Informatik" auftreten. Dabei wird deutlich, daß die bis hier entwickelten Begriffe eine zentrale Rolle spielen.

Wir beginnen mit der Modellierung von Potentialen, wie sie etwa in der Elektrostatik oder der Gravitation auftreten. Hat man auf dem Rand eines Gebiets des $\mathbb{R}^3$ eine gewisse Verteilung elektrischer Ladungen, so stellt sich im Innern des Gebiets ein Feld E ein, das sich als Kraftwirkung auf geladene Teilchen wahrnehmen und messen läßt. Es ist ein wirbelfreies Vektorfeld, das sich als Gradient eines skalaren Potentials u schreiben läßt. Diese skalare Potentialfunktion heißt **Spannung**, das Vektorfeld ist die **Feldstärke**. Sind im Innern des Gebietes keine weiteren Ladungen, so ist die Feldstärke auch quellenfrei, d.h. es gilt $div\ E = div\ \nabla u = \Delta u = 0$. Deshalb ist $\Delta u = 0$ die **Potentialgleichung**, und die **Potentialtheorie** ist eine mathematische Disziplin, die sich ausschließlich mit den Lösungen der Potentialgleichung, den **harmonischen Funktionen** befaßt. Eine typische praktische Aufgabe besteht darin, aus der Verteilung der Ladungen auf dem Rand auf das Feld im Innern zu schließen, d.h. eine Lösung der Potentialgleichung zu finden, die auf dem Rand des Definitionsbereichs gewisse vorgegebene Werte annimmt. Sind im Innern Ladungen mit der Dichte ρ vorhanden, so bekommt man die Gleichung $\Delta u = div\ \nabla E = \rho$.

Aber auch für die **Gravitation** gilt die Potentialgleichung, wobei hier der Gradient des Potentials (die Feldstärke) die auf eine im Feld befindliche kleine Masse wirkende Kraft ist. Ist r der Abstand zum Nullpunkt, so kann man zeigen (Übung), daß außerhalb des Nullpunkts die Funktion $f(r) = \frac{1}{r}$ eine Lösung der Potentialgleichung im $\mathbb{R}^3$ ist. Man kennt damit das Potential eines idealisierten Massenpunktes. Plaziert man mehrere Massen im Raum, so trägt jede mit einem solchen abstandsabhängigen Term zum Gesamt-Kraftfeld bei, und zwar durch Bilden einer Linearkombination. Die Gravitationskraft wirkt dann proportional zum Gradienten dieses Kraftfeldes, und eine kleine "Probemasse" bewegt sich dann auf einer Kurve im Raum, deren Tangentialvektor stets proportional zum Gradienten des Potentials ist. Nach Kepler und Newton sind bei zwei Massen stabile periodische Bahnkurven möglich, die man bei geeigneter Wahl des Bezugssystems als Ellipsen schreiben kann. Das sind näherungsweise die Planetenbahnen, und für Kometen gibt es auch noch parabolische Bahnen. Wenn mehr als zwei Massen involviert sind, wird das Studium der möglichen Bahnkurven sehr viel schwieriger. In der Realität bewegen sich aber die Massenpunkte selbst, und dann wird das Gravitationsfeld zeitabhängig. Man kann sich vorstellen, daß es nicht einfach ist, die Bahnkurve zu berechnen, die für eine jahrelange Mission zum Uranus oder zum nahen Vorbeiflug an einem Saturnmond erforderlich ist.

Gehen wir zur Elektrizität zurück, so wissen wir, daß (nach Oersted) ein Stromfluß in einem geraden Leiter ein wirbelförmiges Magnetfeld erzeugt,

und daß Änderungen von Magnetfeldern oder die Bewegung eines Leiters im Magnetfeld (nach Faraday) Ströme erzeugen. Dies wird neben der elektrischen Feldstärke E durch ein zweites Vektorfeld, die **magnetische Induktion** B beschrieben. Die Vektorfelder sind zeitabhängig und bis auf Materialkonstanten durch die **Maxwellschen Gleichungen**

$$\begin{array}{lcll} div\ E & = & \rho & \text{(Coulombsches Gesetz)} \\ div\ B & = & 0 & \text{(Quellenfreiheit des magn. Feldes)} \\ rot\ B & = & I + \dfrac{\partial E}{\partial t} & \text{(Oersted), } I = \text{ Stromdichte} \\ rot\ E & = & -\dfrac{\partial B}{\partial t} & \text{(Faraday)} \end{array}$$

verbunden. Da die Informatik technisch unter Nutzung des Elektromagnetismus realisiert wird, sollten alle Informatik–Studierenden diese Gesetze kennen.

Die Temperatur in einem Körper ist eine skalare Zeit– und ortsabhängige Funktion u, die lokal der **Wärmeleitungsgleichung**

$$\Delta_x u(x,t) = \frac{\partial u}{\partial t}(x.t)$$

genügt, wobei links der **Laplaceoperator** Δ auf die Ortsvariable $x \in I\!R^3$ wirkt. Stationäre, d.h. nicht zeitabhängige Lösungen sind automatisch Lösungen der Potentialgleichung. Der Gradient der Temperatur nach den Ortsvariablen ist ein wirbelfreies Vektorfeld, die **Wärmestromdichte**.

Ene typische Aufgabe besteht darin, aus einer Anfangs–Temperaturverteilung $u(x,t_0)$ und gewissen Randbedingungen, die von der Wärmeübertragung an die Außenwelt abhängen, die gesamte zeitliche Entwicklung der Temperaturverteilung zu berechnen. Die Funktion

$$u(x,t) := \frac{1}{t\sqrt{t}} \exp\left(-\frac{\|x\|_2^2}{4t}\right)$$

ist eine wichtige spezielle Lösung der Wärmeleitungsgleichung im $I\!R^3$ (Übung). Sie zeigt, wie sich eine (im $I\!R^2$ oder $I\!R^1$) glockenförmige Temperaturverteilung mit der Zeit abflacht, ohne ihre Glockenform (**Gaußsche Glockenkurve** $\exp(-x^2)$) zu verlieren. Das Beispiel zeigt aber auch, daß man eine Temperaturverteilung nicht gut "rückwärts" verfolgen kann, weil unerwartete Singularitäten auftreten können. Das ist für **Diffusionsprozesse** typisch und hat etwas mit dem zweiten Hauptsatz der Wärmelehre zu tun.

Eine sich mit der Zeit vorwärtsbewegende Welle ist z.B $u = \sin(t - x)$, weil der Wert zur Zeit $t + \delta$ an der Stelle x gleich dem Wert zur Zeit t an der Stelle $x - \delta$ ist. Die Differentialgleichung so einer Welle ist

$$\frac{\partial^2}{\partial x^2} u(x,t) = \frac{\partial^2}{\partial t^2} u(x,t),$$

die **Wellengleichung**. In mehr als einer Ortsvariablen hat man

$$\Delta_x u(x,t) = \frac{\partial^2}{\partial t^2} u(x,t)$$

zu nehmen. Bei Wellen ist die Umkehrung der Zeitrichtung unproblematisch, denn wenn $u(x,t)$ die Wellengleichung löst, dann auch $u(x,-t)$. Interpretiert man eine Welle zu einer bestimmten Zeit als eine Dichteverteilung von Materie, so bedeutet das Fortschreiten der Welle mit der Zeit einen Materietransport. Deshalb sind **Transportgleichungen** mit der Wellengleichung verwandt.

Ebene Wellen sind räumlich konstant auf allen Ortsvektoren x, die zu einer festen Raumrichtung r mir $\|r\|_2 = 1$ festes Skalarprodukt haben. Man kann sie in allen Raumdimensionen schreiben (Übung) als Funktionen der Form $f(r^T x + t) + g(r^T x - t)$ mit zweimal differenzierbaren skalaren Funktionen f und g. Stammen Wellen von einem weit entfernten Sender, so fallen sie beim Empfänger als ebene Wellen ein. Sie werden von Objekten gestreut, und es ist ein wichtiges Problem des wissenschaftlichen Rechnens, aus dem gestreuten Feld verschiedener einfallender ebener Wellen den streuenden Körper zu ermitteln (**inverses Problem der Streutheorie**). Andere **inverse Probleme** betreffen die Rekonstruktion von Ladungsdichten aus beobachteten Feldstärken oder die Rekonstruktion von Objekten aus Messungen der Absorption von Strahlen, z.B. bei der **Tomographie**.

Besonders interessant, aber auch besonders schwierig ist das Gebiet der Strömungsphänomene bei Flüssigkeiten und Gasen. Klar ist zunächst, daß man eine Strömung durch ein Geschwindigkeits–Vektorfeld V beschreiben sollte, das in jedem Punkt die momentane Bahngeschwindigkeit eines Gas– oder Flüssigkeitsteilchens angibt. Wenn die Strömung sehr "sanft" verläuft, kann dieses Geschwindigkeitsfeld wirbelfrei und damit Gradient eines Potentials sein. Solche Strömungen nennt man **Potentialströmungen**. Ist die strömende Flüssigkeit inkompressibel, so gilt $div\ V = 0$. Denn auf dem Rand jedes Gebiets muß die einströmende Flüssigkeitsmenge gleich der ausströmenden sein, und deshalb ist das Randintegral im Gaußschen Integralsatz immer Null, also auch die Divergenz. Bei kompressiblen Flüssigkeiten ist die Dichte ρ

räumlich und zeitlich veränderlich, und es gilt stattdessen die **Kontinuitätsgleichung**

$$\rho \, div \, V + (\nabla \rho) \cdot V + \frac{\partial}{\partial t} \rho = 0,$$

die bei zeitlich und räumlich konstanter Dichte wieder in $div \, V = 0$ übergeht.

Aber die Betrachtung des Geschwindigkeitsfelds V (und bei kompressiblen Strömungen auch der Dichte) reicht nicht aus, um eine Strömung zu beschreiben. Wenn z.B. eine Strömung durch einen Druck durch eine Verengung gepreßt wird und sich dann in der Verengung die Strömungsgeschwindigkeit erhöht (Düseneffekt), ist klar, daß man den Druck p ebenfalls modellieren muß. Weil er bei idealen Flüsigkeiten und Gasen immer nach allen Seiten gleichartig wirkt, ist er skalar, aber er ist natürlich zeit– und ortsabhängig. Schreibt man das Ganze als Kräftebilanz, so bekommt man für ideale reibungsfreie Flüssigkeiten und Gase die vektoriellen **Navier–Stokes**–Gleichungen

$$\frac{\partial}{\partial t} V + (\nabla V) \cdot V = F - \frac{1}{\rho} \nabla p,$$

wobei F für ein Feld äußerer Kräfte steht. Bei Auftreten von Reibung kommt auf der rechten Seite noch ein Term der Form ΔV hinzu, und bei viskosen Flüssigkeiten ist es oft unstatthaft, den Druck skalar anzusetzen, so daß die Modellierung noch komplizierter wird. Die Nichtlinearität des Terms $(\nabla V) \cdot V$ in V ist schon schlimm genug, und deshalb zählt die praktische Berechnung der Lösungen realistischer Strömungsvorgänge zu den anspruchsvollsten Aufgaben des wissenschaftlichen Rechnens.

14 Fourierreihen und Fouriertransformationen

Hier behandeln wir ein wichtiges Hilfsmittel zur Behandlung von Signalen, bis hin zu deren Kompression. Das Fernziel ist ein extrem wichtiger Algorithmus, die **schnelle Fouriertransformation**[1], die an verschiedenen Stellen der Informatik auftritt, und zwar nicht nur in der Signalverarbeitung. Praktisch noch wichtiger ist die **diskrete Cosinustransformation**[2], die aber hier aus Platzgründen nicht mehr behandelt werden kann. Sie tritt bei der **JPEG-Kompression**[3][4] und ihren Nachfolgern auf. Man kann dazu einiges in einem anderen Skript[5] ab etwa S. 20 nachlesen, leider dort auf Englisch. Die **mp3-Kompression**[6] nutzt sowohl die schnelle Fouriertransformation als auch die Cosinustransformation.

14.1 Fourierreihen

14.1.1 Periodische Funktionen

Wir wiederholen die Definition periodischer Funktionen aus Abschnitt 11.1.1 auf Seite 269.

Definition 14.1 *Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt* periodisch *mit* Periode $h > 0$*, falls $f(x+h) = f(x)$ für alle $x \in \mathbb{R}$ gilt. Der Vektorraum der stetigen und 2π-periodischen Funktionen $f : \mathbb{R} \to \mathbb{R}$ wird mit $C_{2\pi}$ bezeichnet. Ferner bezeichnet $C^k_{2\pi}$ den Raum der k-fach differenzierbaren, 2π-periodischen Funktionen.*

Offensichtlich ist eine h-periodische Funktion eindeutig durch ihre Werte auf $[0, h)$ bestimmt. Hat die Funktion f die Periode $h > 0$, so hat die Funktion $\widetilde{f}(x) := f(hx/(2\pi))$ die Periode 2π. Daher werden wir uns im Folgenden nur noch mit 2π-periodischen Funktionen beschäftigen.

[1] http://de.wikipedia.org/wiki/Schnelle_Fourier-Transformation
[2] http://de.wikipedia.org/wiki/Diskrete_Kosinustransformation
[3] http://de.wikipedia.org/wiki/Jpg
[4] http://www.spemaus.de/studium/visjpeg/applet.html
[5] http://www.num.math.uni-goettingen.de/schaback/teaching/texte/approx/Appverf_I.pdf
[6] http://de.wikipedia.org/wiki/Mp3

14.1.2 Trigonometrische Polynome

Definition 14.2 *Die Elemente der Menge*

$$\mathcal{T}_m^{\mathbb{R}} := \left\{ T(x) = \frac{a_0}{2} + \sum_{j=1}^{m} (a_j \cos jx + b_j \sin jx) : a_j, b_j \in I\!R \right\} \tag{14.2}$$

heißen (reelle) trigonometrische Polynome vom Grad $\leq m$. *Als Erzeugendensystem wählt man*

$$\frac{1}{\sqrt{2}},\ \sin jx,\ \cos jx,\ 1 \leq j \leq m.$$

Offensichtlich ist $\mathcal{T}_m^{\mathbb{R}}$ ein linearer, endlichdimensionaler Vektorraum über $I\!R$ mit Dimension $\leq 2m+1$. Da wir in diesem Abschnitt lineare Räume sowohl über $I\!R$ als auch über $\mathbb{C}$ betrachten, wollen wir in der Bezeichnung etwas formaler sein und $I\!R$ als oberen Index hinzusetzen. Wir werden bald sehen, dass die Dimension in der Tat $2m+1$ ist. Aber zuerst noch eine Begründung für die Bezeichnung "trigonometrisches Polynom".

Theorem 14.3 *Das Produkt zweier trigonometrischer Polynome ist wieder ein trigonometrisches Polynom.*

Dies folgt unmittelbar aus der Definition trigonometrischer Polynome und den folgenden Gleichungen:

$$\begin{aligned} \cos(jx)\cos(kx) &= \frac{1}{2}\left[\cos((j-k)x) + \cos((j+k)x)\right], \\ \sin(jx)\sin(kx) &= \frac{1}{2}\left[\cos((j-k)x) - \cos((j+k)x)\right], \\ \sin(jx)\cos(kx) &= \frac{1}{2}\left[\sin((j+k)x) + \sin((j-k)x)\right], \end{aligned} \tag{14.3}$$

die man leicht verifiziert. □

Die Behandlung reeller trigonometrischer Polynome wird wesentlich erleichtert, indem man ein reelles trigonometrisches Polynom mittels der Eulerschen Formeln

$$e^{ix} = \cos x + i \sin x, \quad \cos x = \frac{1}{2}(e^{ix} + e^{-ix}), \quad \sin x = \frac{-i}{2}(e^{ix} - e^{-ix}), \tag{14.3}$$

in ein komplexes trigonometrisches Polynom überführt:

$$T(x) = \frac{a_0}{2} + \sum_{j=1}^{m} (a_j \cos jx + b_j \sin jx)$$

$$
\begin{aligned}
&= e^{-imx}\left(\frac{a_0}{2}e^{imx} + \sum_{j=1}^{m}\frac{1}{2}(a_j - ib_j)e^{i(m+j)x} + \frac{1}{2}(a_j + ib_j)e^{i(m-j)x}\right) \\
&= : e^{-imx}\sum_{j=0}^{2m} c_j e^{ijx} =: e^{-imx}p(x).
\end{aligned}
$$

Dabei stehen die Koeffizienten von T und p in folgendem Zusammenhang:

$$
\begin{aligned}
c_{m-j} &= \frac{a_j + ib_j}{2}, \qquad 1 \le j \le m, \\
c_{m+j} &= \frac{a_j - ib_j}{2}, \qquad 1 \le j \le m, \\
c_m &= \frac{a_0}{2}.
\end{aligned} \tag{14.3}
$$

Man beachte, dass (14.3) eine bijektive Abbildung zwischen den Koeffizienten von T und p liefert. Startet man allerdings mit den $c_j \in \mathbb{C}$, so ist zunächst nicht garantiert, dass die a_j und b_j in $\mathbb{R}$ liegen, also $T \in \mathcal{T}_m^{\mathbb{R}}$ ist. Dazu benötigt man die zusätzliche Voraussetzung $c_{m-j} = \overline{c_{m+j}}$.

Definition 14.4 *Die Elemente der Menge*

$$
\mathcal{T}_{n-1}^{\mathbb{C}} := \left\{ T : T(x) = \sum_{j=0}^{n-1} c_j e^{ijx} : c_j \in \mathbb{C} \right\} \tag{14.4}
$$

heißen (komplexe) trigonometrische Polynome vom Grad $\le n-1$.

Der Raum $\mathcal{T}_{n-1}^{\mathbb{C}}$ ist ein linearer endlich dimensionaler Raum über $\mathbb{C}$. Die Abbildung $[0, 2\pi) \to \mathbb{C}$, $x \mapsto e^{ix}$ überführt jedes komplexe trigonometrische Polynom in die Einschränkung eines komplexen, algebraischen Polynoms auf den Einheitskreis. Dies motiviert den Begriff Polynom im Namen.

Theorem 14.5 *Der Raum $\mathcal{T}_{n-1}^{\mathbb{C}}$ hat für $n \in \mathbb{N}$ die Dimension n.*

Beweis: Aus $\sum_{j=0}^{n-1} c_j e^{ijx} = 0$ für alle $x \in [0, 2\pi)$ folgt wegen

$$
\int_0^{2\pi} e^{ijx} dx = 2\pi\delta_{j,0}, \qquad j \in \mathbb{Z},
$$

sofort

$$
0 = \int_0^{2\pi} \sum_{j=0}^{n-1} c_j e^{i(j-k)x} dx = 2\pi c_k
$$

für $0 \le k \le n-1$, was die lineare Unabhängigkeit des Erzeugendensystems liefert. □

14.1.3 Fourierreihen

Nach diesem Ausflug ins Komplexe, den wir unten wieder brauchen, gehen wir ins Reelle zurück.

Theorem 14.6 *Mit dem reellwertigen inneren Produkt*

$$(u,v) := \frac{1}{\pi}\int_{-\pi}^{\pi} u(t)v(t)dt$$

auf dem Raum $C_{2\pi}$ sind die Funktionen des Erzeugendensystems aus Definition 14.2 orthonormal, und der Raum $\mathcal{T}_m^{\mathbb{R}}$ hat die Dimension $2m+1$.

Den Beweis lassen wir als Übungsaufgabe zum Integrieren. Dabei kann man die Formeln aus dem Beweis des Satzes 14.3 oder den obigen Umweg übers Komplexe benutzen.

Orthonormalbasen traten schon im Abschnitt 5.4 auf Seite 171 auf. Zu einer gegebenen 2π–periodischen stetigen Funktion f kann man die Projektion auf $\mathcal{T}_m^{\mathbb{R}}$ durch

$$\begin{aligned} P_{\mathcal{T}_m^{\mathbb{R}}}(f)(x) &:= (f(t),\frac{1}{\sqrt{2}})\frac{1}{\sqrt{2}} \\ &\quad +\sum_{j=1}^{m}(f(t),\cos jt)\cos jx \\ &\quad +\sum_{j=1}^{m}(f(t),\sin jt)\sin jx \end{aligned}$$

wobei wir das Argument t bei den inneren Produkten stehengelassen haben. Man nennt $P_{\mathcal{T}_m^{\mathbb{R}}}(f)$ die m–te **Fourier–Partialsumme**. Schreiben wir die Projektion als trigonometrisches Polynom

$$P_{\mathcal{T}_m^{\mathbb{R}}}(f)(x) = \frac{a_0}{2} + \sum_{j=1}^{m}\left(a_j\cos jx + b_j\sin jx\right),$$

so ergeben sich die **Fourier–Koeffizienten** über das Skalarprodukt als

$$\begin{aligned} a_j &= (f(t),\cos jt) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(jt)dt,\ 1\le j\le m \\ b_j &= (f(t),\sin jt) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(jt)dt,\ 1\le j\le m. \end{aligned}$$

Der Koeffizient a_0 ist als Sonderfall anzusehen, weil er nicht mit dem Faktor $\frac{1}{\sqrt{2}}$, sondern mit $\frac{1}{2}$ auftritt. Das hat aber den Vorteil, daß man oben einfach auch $j=0$ zulassen kann und

$$a_0 = (f(t),1) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)dt$$

erhält. Die infiniten Linearkombinationen

$$\frac{a_0}{2} + \sum_{j=1}^{\infty}(a_j \cos(jx) + b_j \sin(jx)) \tag{14.7}$$

nennt man **Fourierreihen**[1] nach Jean Baptiste **Fourier**[2], und man benutzt sie zur Darstellung periodischer Signale. Dazu gibt es ein sehr schönes Java–Applet[3]..

14.1.4 Konvergenz von Fourierreihen

Schön wäre es, wenn zu gegebener Funktion $f \in C_{2\pi}$ die Fourier–Partialsummen $P_{\mathcal{T}_m^{\mathbb{R}}}(f)$ gegen f konvergieren würden, so daß man immer die Gleichung

$$\begin{aligned} f(x) &= \lim_{m\to\infty} P_{\mathcal{T}_m^{\mathbb{R}}}(f)(x) \\ &:= (f(t), \frac{1}{\sqrt{2}})\frac{1}{\sqrt{2}} \\ &\quad + \sum_{j=1}^{\infty}(f(t), \cos jt)\cos jx \\ &\quad + \sum_{j=1}^{\infty}(f(t), \sin jt)\sin jx \end{aligned}$$

benutzen könnte. Das ist aber leider weder einfach noch korrekt. Aber weil es ein periodisches Analogon zum Satz 11.43 von Weierstraß gibt, hat man

Theorem 14.8 *Zu jeder stetigen 2π–periodischen Funktion f konvergieren die Fourier–Partialsummen in der durch das Skalarprodukt erzeugten Norm gegen f. Die Konvergenz ist nicht punktweise, und schon gar nicht gleichmäßig, es gilt nur*

$$0 = \lim_{m\to\infty} \int_{-\pi}^{\pi} \left(f(x) - P_{\mathcal{T}_m^{\mathbb{R}}}(f)(x)\right)^2 dx,$$

und das nennt man auch **Konvergenz im quadratischen Mittel**.

Wir wollen diesen Satz nicht beweisen, sondern nur einige Folgerungen ziehen. Zunächst folgen aus der Orthonormalität des Erzeugendensystems und den

[1] `http://de.wikipedia.org/wiki/Fourierreihe`
[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Fourier.html`
[3] `http://www.falstad.com/fourier/`

in Satz 5.23 bewiesenen Eigenschaften eines Projektors die Aussagen

$$\begin{aligned} f - P_{\mathcal{T}_m^{\mathbb{R}}}(f) &\perp \mathcal{T}_m^{\mathbb{R}} \\ \|f - P_{\mathcal{T}_m^{\mathbb{R}}}(f)\|^2 + \|P_{\mathcal{T}_m^{\mathbb{R}}}(f)\|^2 &= \|f\|^2 \\ \|P_{\mathcal{T}_m^{\mathbb{R}}}(f)\|^2 &= \frac{a_0^2}{2} + \sum_{j=1}^{m}(a_j^2 + b_j^2) \\ \lim_{m\to\infty} \|P_{\mathcal{T}_m^{\mathbb{R}}}(f)\|^2 &= \|f\|^2 \\ &= \frac{a_0^2}{2} + \sum_{j=1}^{\infty}(a_j^2 + b_j^2). \end{aligned}$$

Diese Beziehung heißt **Parsevalsche Gleichung** und ist unter anderem nützlich, um gewisse problematische Reihen auszuwerten, z.B. die Reihe $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$. Wir geben hier nur das Rezept an. Man nehme die 2π–periodische Fortsetzung der Hutfunktion

$$f(x) = \left\{ \begin{array}{cc} x & 0 \le x \le \pi \\ 2\pi - x & \pi \le x \le 2\pi \end{array} \right\} \text{ für alle } x \in [0, 2\pi]$$

und bekommt die Fourierreihe

$$\frac{\pi}{2} - \frac{4}{\pi} \sum_{j=1}^{\infty} \frac{\cos(2j-1)x}{(2j-1)^2}$$

aus der man, wenn man annimmt, dass sie $x = 0$ punktweise gegen f konvergiert, die Gleichung

$$\frac{\pi^2}{8} = \sum_{j=1}^{\infty} \frac{1}{(2j-1)^2}$$

und dann auch die behauptete Reihenformel erhält. Die punktweise Konvergenz in Null kann man mit Hilfe von hier nicht behandelten Techniken erschließen.

Man bekommt ferner eine Fehlerabschätzung aus

$$\|f - P_{\mathcal{T}_m^{\mathbb{R}}}(f)\|^2 = \sum_{j=m+1}^{\infty} (a_j^2 + b_j^2), \tag{14.9}$$

und diese wird noch nützlich sein. Aber man kann aus der Orthonormalität des Erzeugendensystems auch auf eine Kompressionsmethode für periodische Signale schließen. Läßt man in einer Fourierreihe (14.7) alle “kleinen”

Koeffizienten weg, so bekommt man einen Fehler in $\|.\|^2$, der gleich der Quadratsumme der weggelassenen Koeffizienten ist. Dieses Kompressionsprinzip wird bei den hier aus Platzgründen nicht behandelten Entwicklungen nach Wavelet–Basen erfolgreich angewendet.

Wir bekandeln jetzt noch die oft auftretende komplexe Form der Fourierreihen. Die Eulerschen Formeln liefern wie oben

$$P_{\mathcal{T}_m^{\mathbb{R}}}(f)(x) = \frac{a_0}{2} + \sum_{j=1}^{m} \frac{a_j - ib_j}{2} e^{ijx} + \frac{a_j + ib_j}{2} e^{-ijx} = \sum_{j=-m}^{m} c_j e^{ijx}.$$

Dabei rechnet man leicht nach, dass die Koeffizienten $c_j = c_j(f) =: \widehat{f}(j)$ die Form

$$\widehat{f}(j) = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ijx} dx, \qquad j \in \mathbb{Z},$$

annehmen. Die Darstellung (14.7) von f wird damit also zu

$$f(x) = \sum_{j=-\infty}^{\infty} \widehat{f}(j) e^{ijx} \tag{14.9}$$

Neben den schon nach Satz 14.8 gemachten Bemerkungen zur Konvergenz von Fourierreihen tritt hier noch eine eher untypische Definition der biinfiniten Reihe auf, denn wir definieren hier einfach $\sum_{j=-\infty}^{\infty}$ als $\lim_{m\to\infty} \sum_{j=-m}^{m}$. Dies ist allerdings eine weitaus schwächere Definition als die übliche, wo die Summe in zwei einfache infinite Reihen aufgespalten wird, die beide für sich genommen konvergieren müssen.

Definition 14.10 *Die Zahlen $a_j(f)$, $b_j(f)$, $\widehat{f}(j)$ heißen* **Fourier-Koeffizienten** *von f. Die Abbildung $\widehat{f} : C_{2\pi} \to \mathbb{C}$ heißt auch* (semi–diskrete) **Fourier-Transformation** *von f.*

Die Frage nach einer effizienten Berechnung von Fourierkoeffizienten verschieben wir auf später. Stattdessen kümmern wir uns um Fragen der Konvergenzgeschwindigkeit und der punktweisen Konvergenz.

Theorem 14.11 *Sei $f \in C_{2\pi}^k$. Dann gilt*

$$\|f - S_m f\| \le \frac{1}{(m+1)^k} \|f^{(k)} - S_m(f^{(k)})\| = \mathcal{O}(m^{-k}) \qquad \text{für } m \to \infty.$$

Durch partielle Integration und auf Grund der 2π-Periodizität finden wir

$$c_j(f^{(k)}) = \frac{1}{2\pi} \int_0^{2\pi} f^{(k)}(x) e^{-ijx} dx = \frac{ij}{2\pi} \int_0^{2\pi} f^{(k-1)}(x) e^{-ijx} = (ij) c_j(f^{(k-1)}),$$

was per Induktion zu $c_j(f^{(k)}) = (ij)^k c_j(f)$ führt. Aus der komplexen Form von (14.9) erhalten wir damit

$$\begin{aligned}\|f - S_m f\|^2 &= \sum_{|j|\geq m+1} |c_j(f)|^2 \leq \sum_{|j|\geq m+1} |j|^{-2k} |c_j(f^{(k)})|^2 \\ &\leq \frac{1}{(m+1)^{2k}} \|f^{(k)} - S_m f^{(k)}\|^2\end{aligned}$$

und $\|f^{(k)} - S_m f^{(k)}\|$ konvergiert immer noch gegen Null, was den $o(m^{-k})$ Teil rechtfertigt. □

Diese genauere Konvergenzaussage erlaubt uns jetzt auch auf gleichmäßige Konvergenz zu schließen.

Theorem 14.12 *Zu $f \in C^1_{2\pi}$ ist die Fourier-Reihe gleichmäßig konvergent.*

Aus dem Beweis von Satz 14.11 wissen wir bereits $c_j(f') = (ij)c_j(f)$. Daher gilt für die Ableitung von $S_m f$, dass

$$(S_m f)'(x) = \sum_{j=-m}^{m} c_j(f) \left(e^{ijx}\right)' = \sum_{j=-m}^{m} c_j(f)(ij)e^{ijx} = S_m(f')(x).$$

Da $f - S_m f$ senkrecht auf allen trigonometrischen Polynomen vom Grad $\leq m$ steht, folgt insbesondere $0 = (f - S_m f, 1)$, d.h das Integral über $f - S_m f$ verschwindet auf $[0, 2\pi]$. Also hat $f - S_m f$ in $[0, 2\pi]$ eine Nullstelle x^*. Der Hauptsatz der Differential- und Integralrechnung und die Cauchy-Schwarzsche Ungleichung liefern

$$\begin{aligned}|f(x) - S_m f(x)| &= \left| \int_{x^*}^{x} (f - S_m f)'(t) dt \right| \leq \int_{x^*}^{x} |(f' - S_m f')(t)| dt \\ &\leq \sqrt{|x - x^*|}\sqrt{\pi} \|f' - S_m f'\| \leq \sqrt{2\pi}\sqrt{\pi} \|f' - S_m f'\|,\end{aligned}$$

und der letzte Ausdruck strebt gleichmäßig in x gegen Null. □

Theorem 14.13 *Zu $f \in C^k_{2\pi}$ konvergiert die Fourier-Reihe $S_m f$ mindestens wie*

$$\|f - S_m f\|_\infty = o(m^{1-k}) \qquad m \to \infty.$$

14.2 Periodische Interpolation

14.2.1 Interpolation mit trigonometrischen Polynomen

Wir stellen hier einen Zusammenhang zwischen den Koeffizienten und den Funktionswerten trigonometrischer Polynome her.

Theorem 14.14 *Zu $n \in \mathbb{N}$ paarweise verschiedenen* **Stützstellen** $x_0, \ldots, x_{n-1} \in [0, 2\pi)$ *und komplexen Zahlen $f_0, \ldots, f_{n-1}$ gibt es genau ein komplexes trigonometrisches Polynom $p \in \mathcal{T}_{n-1}^{\mathbb{C}}$ mit $p(x_j) = f_j$, $0 \le j \le n-1$. Man sagt, daß p die Werte $f_0, \ldots, f_{n-1}$ in den $x_0, \ldots, x_{n-1} \in [0, 2\pi)$* **interpoliert** *und bezeichnet p als* **Interpolationspolynom**.

Beweis: Die Punkte $z_j := e^{ix_j} \in \mathbb{C}$ sind paarweise verschieden. Nach **Lagrange**[1] betrachtet man das Polynom

$$q(z) := \sum_{j=0}^{n-1} f_j \prod_{\substack{k \neq j \\ 0 \le k < n}} \frac{z - z_k}{z_j - z_k}$$

und stellt fest, daß es ein Polynom maximal $(n-1)$–ten Grades ist, das die Bedingungen $q(z_j) = f_j$, $0 \le j < n$ erfüllt. Dann setzt man

$$p(x) := q(e^{ix})$$

und p ist ein trigonometrisches Polynom, das dem Satz genügt. Gäbe es zwei solche Polynome, so wäre die Differenz ein Polynom, das in n Punkten verschwindet und einen Grad $< n$ hat. Dann muß die Differenz nach dem Fundamentalsatz der Algebra überall Null sein. □

Bei der Anwendung dieses Resultats auf die reelle trigonometrische Interpolation müssen wir nachweisen, dass bei der Rücktransformation in (14.3) tatsächlich auch reelle Koeffizienten a_j, b_j herauskommen. Natürlich müssen wir uns zunächst auf $n = 2m+1$ beschränken.

Theorem 14.15 *Gegeben seien paarweise verschiedene $x_0, \ldots, x_{2m} \in [0, 2\pi)$ und $f_0, \ldots, f_{2m} \in \mathbb{R}$. Dann existiert genau ein reelles trigonometrisches Polynom $T \in \mathcal{T}_m^{\mathbb{R}}$ mit $T(x_j) = f_j$, $0 \le j \le 2m$.*

Beweis: Sei $p(x) = \sum_{j=0}^{2m} c_j e^{ijx}$, das nach Satz 14.14 eindeutig existierende, komplexe trigonometrische Interpolationspolynom mit $p(x_j) = e^{imx_j} f_j$, $0 \le j \le 2m$. Sei $\widetilde{p}$ definiert durch

$$\widetilde{p}(x) := e^{2imx}\overline{p(x)} = \sum_{j=0}^{2m} \overline{c_j} e^{i(2m-j)x} = \sum_{j=0}^{2m} \overline{c_{2m-j}} e^{ijx}, \qquad x \in [0, 2\pi).$$

Dann ist offensichtlich $\widetilde{p} \in \mathcal{T}_{2m}^{\mathbb{C}}$ und $\widetilde{p}(x_j) = f_j e^{imx_j} = p(x_j)$, $0 \le j \le 2m$, da die Funktionswerte f_j reellwertig sind. Aus der Eindeutigkeitsaussage aus

[1] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Lagrange.html`

Satz 14.14 folgt also $\widetilde{p} \equiv p$ und damit nach Satz 14.5 auch $c_j = \overline{c_{2m-j}}$, $0 \leq j \leq 2m$. Aus der Rücktransformation mit (14.3) erhalten wir insbesondere $a_0 = 2c_m \in I\!R$ aber auch $a_j = c_{m-j} + c_{m+j} = 2\Re(c_{m-j}) \in I\!R$ und $b_j = i(c_{m+j} - c_{m-j}) = 2\Im(c_{m-j}) \in I\!R$ jeweils für $1 \leq j \leq m$. □

14.2.2 Äquidistante Stützstellen

Nach dem allgemeinen Interpolationsproblem wollen wir jetzt untersuchen, was man zusätzlich gewinnt, wenn die Stützstellen äquidistant sind, d.h. wenn $x_j = \frac{2\pi j}{n}$, $0 \leq j \leq n-1$, gilt. Es wird sich herausstellen, dass in diesem Fall auch ein reelles trigonometrisches Interpolationspolynom für gerades n existiert und dass sich die Koeffizienten explizit angeben lassen. Eine wichtige Rolle spielen dabei die *n-ten Einheitswurzeln*

$$\zeta_n := e^{\frac{2\pi i}{n}}. \tag{14.15}$$

Sie erfüllen offensichtlich die Beziehungen

$$\zeta_n^n = 1, \quad \zeta_n^j = e^{ix_j}, \quad \zeta_n^{j+k} = \zeta_n^j \zeta_n^k, \quad \zeta_n^{jk} = e^{ijx_k}, \quad \zeta_n^{-j} = \overline{\zeta_n^j}. \tag{14.15}$$

Wesentlich wird noch die folgende Eigenschaft sein.

Theorem 14.16 *Für $n \in I\!N$ und $\ell, k \in I\!N_0$ mit $0 \leq \ell, k \leq n-1$ gilt*

$$\frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{(\ell-k)j} = \delta_{\ell,k}.$$

Beweis: Die Sache ist klar für $\ell = k$. Im Falle $\ell \neq k$ liefert die Einschränkung an ℓ und k, dass $\zeta_n^{\ell-k} \neq 1$, sodass die Behauptung aus

$$\sum_{j=0}^{n-1} \left(\zeta_n^{\ell-k}\right)^j = \frac{\zeta_n^{(\ell-k)n} - 1}{\zeta_n^{\ell-k} - 1} = 0$$

folgt. □

Dieses Lemma erlaubt es uns, den komplexen Fall leicht abzuhandeln.

Theorem 14.17 *Sind für $n \in I\!N$ die Stützstellen $x_j = \frac{2\pi j}{n}$, $0 \leq j \leq n-1$, und die Stützwerte $f_0, \ldots, f_{n-1} \in \mathbb{C}$ gegeben, so hat das eindeutig bestimmte komplexe trigonometrische Interpolationspolynom*

$$p(x) = \sum_{j=0}^{n-1} c_j e^{ijx}, \qquad x \in [0, 2\pi),$$

die Koeffizienten

$$c_j = \frac{1}{n}\sum_{k=0}^{n-1} f_k \zeta_n^{-jk}, \qquad 0 \le j \le n-1. \tag{14.17}$$

Beweis: Da nach Satz 14.14 das Interpolationspolynom eindeutig existiert, reicht es nachzurechnen, dass das hier angegebene Polynom ebenfalls die Daten interpoliert. Dies folgt aber nach Theorem 14.16 aus

$$p(x_\ell) = \sum_{j=0}^{n-1} \frac{1}{n} \sum_{k=0}^{n-1} \zeta_n^{-jk} f_k e^{ijx_\ell} = \sum_{k=0}^{n-1} f_k \frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{(\ell-k)j} = f_\ell$$

für $0 \le \ell \le n-1$. □

Schreibt man zum Vergleich

$$f_k = \sum_{j=0}^{n-1} c_j \zeta_n^{jk} \qquad 0 \le k \le n-1, \tag{14.17}$$

so sieht man, dass die Abbildung $F_n : \mathbb{C}^n \to \mathbb{C}^n$, $\{f_k\} \mapsto \{c_j\}$ und ihre Umkehrabbildung eine sehr ähnliche Struktur haben und deswegen numerisch gleich behandelt werden können.

Definition 14.18 *Die bijektive Abbildung $F_n : \mathbb{C}^n \to \mathbb{C}^n$, $\{f_k\} \mapsto \{c_j\}$, die durch (14.17) definiert ist, heißt die* diskrete Fourier-Analyse *der Daten $\{f_k\}$. Ihre Umkehrabbildung ist gegeben durch (14.17) und heißt* diskrete Fourier-Synthese. *Beide zusammen nennt man* diskrete Fourier-Transformation.

Es folgt die Rücktransformation für die reelle Interpolationsaufgabe.

Theorem 14.19 *Sei $n \in \mathbb{N}$ gegeben als $n = 2m+1$ oder $n = 2m$. Seien $x_j = \frac{2\pi j}{n}$ und $f_j \in \mathbb{R}$ für $0 \le j \le n-1$. Seien*

$$\begin{aligned} a_j &= \frac{2}{n}\sum_{k=0}^{n-1} f_k \cos jx_k, \qquad 0 \le j \le m, \\ b_j &= \frac{2}{n}\sum_{k=0}^{n-1} f_k \sin jx_k, \qquad 1 \le j \le m. \end{aligned}$$

Dann erfüllt das trigonometrische Polynom

$$T(x) := \begin{cases} \dfrac{a_0}{2} + \displaystyle\sum_{j=0}^{m} (a_j \cos jx + b_j \sin jx), & \text{falls } n = 2m+1, \\ \dfrac{a_0}{2} + \displaystyle\sum_{j=0}^{m-1} (a_j \cos jx + b_j \sin jx) + \dfrac{a_m}{2} \cos mx, & \text{falls } n = 2m, \end{cases}$$

die Interpolationsbedingungen $T(x_j) = f_j$, $0 \le j \le n-1$.

Beweis: Wie im Beweis zu Satz 14.15 sei $p \in \mathcal{T}_{n-1}^{\mathbb{C}}$ das trigonometrische Polynom mit $p(x_k) = f_k e^{imx_k}$, $0 \le k \le n-1$. Dann wissen wir, dass die Koeffizienten durch (14.17) gegeben sind als

$$c_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{imx_k} \zeta_n^{-jk} = \frac{1}{n} \sum_{k=0}^{n-1} f_k \zeta_n^{k(m-j)}, \qquad 0 \le j \le n-1.$$

Im Fall $n = 2m$ werden wir p als trigonometrisches Polynom vom Grad $2m$ auffassen, indem wir $c_{2m} = c_n = 0$ explizit setzen. Dann haben wir in beiden Fällen ein komplexes trigonometrisches Polynom, welches vermöge (14.3) in ein reelles trigonometrisches Polynom vom Grad m zurücktransformiert werden kann. Dieses Polynom sei jetzt

$$\widetilde{T}(x) = e^{-imx} p(x) = \frac{\widetilde{a}_0}{2} + \sum_{j=1}^{m} \left(\widetilde{a}_j \cos jx + \widetilde{b}_j \sin jx \right). \tag{14.19}$$

Wir wissen, dass $\widetilde{T}$ die Daten interpoliert, und dass es im Fall $n = 2m+1$ auch reelle Koeffizienten hat. In diesem Fall liefert (14.19) und (14.3) zum einen für $0 \le j \le m$,

$$\widetilde{a}_j = c_{m+j} + c_{m-j} = \frac{1}{n} \sum_{k=0}^{n-1} f_k \left(\zeta_n^{-kj} + \zeta_n^{kj} \right) = \frac{2}{n} \sum_{k=0}^{n-1} f_k \Re(\zeta_n^{jk}) = a_j$$

und zum anderen für $1 \le j \le m$,

$$\widetilde{b}_j = i(c_{m+j} - c_{m-j}) = \frac{1}{n} \sum_{k=0}^{n-1} f_k i \left(\zeta_n^{-kj} - \zeta_n^{kj} \right) = \frac{2}{n} \sum_{k=0}^{n-1} f_k \Im(\zeta_n^{kj}) = b_j,$$

sodass $\widetilde{T} = T$ gilt und damit auch T die Daten interpoliert.

Im Fall $n = 2m$ zeigen obige Rechnungen ebenfalls $\widetilde{a}_j = a_j$ für $0 \le j \le m-1$ und $\widetilde{b}_j = b_j$ für $1 \le j \le m-1$. Ferner gilt

$$\widetilde{a}_m = c_0 = \frac{1}{n}\sum_{k=0}^{n-1} f_k \zeta_n^{km} = \frac{1}{n}\sum_{k=0}^{n-1} f_k \cos mx_k = \frac{a_m}{2},$$

da $c_{2m} = 0$ und $\zeta_n^{km} = \zeta_{2m}^{km} = (-1)^k = \cos(mx_k)$. Also gilt $T(x) = \widetilde{T}(x) - \widetilde{b}_m \sin mx$. Nun wird im Allgemeinen $\widetilde{b}_m = -ic_0$ nicht verschwinden, was uns aber nicht stört, denn da $\sin mx_k = \sin \pi k = 0$, für $0 \le k \le n-1$ gilt, interpoliert mit $\widetilde{T}$ auch T die gegebenen Daten. □

14.3 Die schnelle Fourier-Transformation

Die explizite Formel (14.17) zur Berechnung der Koeffizienten der trigonometrischen Interpolanten erlaubt es, jeden einzelnen Koeffizienten, sofern die Potenzen der Einheitswurzeln vorab bekannt sind, in $\mathcal{O}(n)$ Operationen auszurechnen, sodass man insgesamt $\mathcal{O}(n^2)$ Operationen benötigt, um die Interpolante komplett zu bestimmen. Dies ist im Vergleich zu den üblichen $\mathcal{O}(n^3)$ Operationen, die normalerweise zum Lösen des zugehörigen Gleichungssystem benötigt werden, bereits eine merkliche Verbesserung. Trotzdem lässt sich dieses Resultat noch weiter verbessern.

Bei der Bildung der Summen in (14.17) treten bei geradem $n = 2m$ bei mehreren verschiedenen Funktionswerten f_k numerisch die gleichen (oder nur im Vorzeichen verschiedenen) Faktoren $\zeta_n^{-jk} = e^{-\frac{2\pi i jk}{n}}$ auf. Genauer gilt

$$\zeta_n^{-j(k+m)} = \zeta_n^{-jk}\zeta_n^{-jm} = (-1)^j \zeta_n^{-jk}.$$

Ähnliches gilt natürlich auch für die diskrete Fourier Synthese. Diese Tatsache kann man ausnutzen, um durch geschicktes Zusammenfassen der Terme die Anzahl der Multiplikationen zu reduzieren. Auf dieser Tatsache beruht die *schnelle Fourier-Transformation* (englisch: *Fast Fourier Transform* oder *FFT*)[1].

Bleiben wir bei geradem $n = 2m$, so gilt für die Koeffizienten mit geradem Index $j = 2\ell$ offenbar

$$c_{2\ell} \;=\; \frac{1}{n}\sum_{k=0}^{n-1} f_k \zeta_n^{-2\ell k} = \frac{1}{n}\sum_{k=0}^{m-1} \left(f_k \zeta_n^{-2\ell k} + f_{k+m}\zeta_n^{-2\ell(k+m)}\right)$$

[1] `http://de.wikipedia.org/wiki/Schnelle_Fourier-Transformation`

Tabelle 3: FFT für $n = 8$.

Daten		$m = 4$		$m = 2$		$m = 1$	
f_0	c_0	$f_0^{(1)} = \frac{f_0+f_4}{2}$	c_0	$f_0^{(2)} = \frac{f_0^{(1)}+f_2^{(1)}}{2}$	c_0	$f_0^{(3)} = \frac{f_0^{(2)}+f_1^{(2)}}{2}$	$= c_0$
f_1	c_1	$f_1^{(1)} = \frac{f_1+f_5}{2}$	c_2	$f_1^{(2)} = \frac{f_1^{(1)}+f_3^{(1)}}{2}$	c_4	$f_1^{(3)} = \frac{f_0^{(2)}-f_1^{(2)}}{2}\zeta_2^{-0}$	$= c_4$
f_2	c_2	$f_2^{(1)} = \frac{f_2+f_6}{2}$	c_4	$f_2^{(2)} = \frac{f_0^{(1)}-f_2^{(1)}}{2}\zeta_4^{-0}$	c_2	$f_2^{(3)} = \frac{f_2^{(2)}+f_3^{(2)}}{2}$	$= c_2$
f_3	c_3	$f_3^{(1)} = \frac{f_3+f_7}{2}$	c_6	$f_3^{(2)} = \frac{f_1^{(1)}-f_3^{(1)}}{2}\zeta_4^{-1}$	c_6	$f_3^{(3)} = \frac{f_2^{(2)}-f_3^{(2)}}{2}\zeta_2^{-0}$	$= c_6$
f_4	c_4	$f_4^{(1)} = \frac{f_0-f_4}{2}\zeta_8^{-0}$	c_1	$f_4^{(2)} = \frac{f_4^{(1)}+f_6^{(1)}}{2}$	c_1	$f_4^{(3)} = \frac{f_4^{(2)}+f_5^{(2)}}{2}$	$= c_1$
f_5	c_5	$f_5^{(1)} = \frac{f_1-f_5}{2}\zeta_8^{-1}$	c_3	$f_5^{(2)} = \frac{f_5^{(1)}+f_7^{(1)}}{2}$	c_5	$f_5^{(3)} = \frac{f_4^{(2)}-f_5^{(2)}}{2}\zeta_2^{-0}$	$= c_5$
f_6	c_6	$f_6^{(1)} = \frac{f_2-f_6}{2}\zeta_8^{-2}$	c_5	$f_6^{(2)} = \frac{f_4^{(1)}-f_6^{(1)}}{2}\zeta_4^{-0}$	c_3	$f_6^{(3)} = \frac{f_6^{(2)}+f_7^{(2)}}{2}$	$= c_3$
f_7	c_7	$f_7^{(1)} = \frac{f_3-f_7}{2}\zeta_8^{-3}$	c_7	$f_7^{(2)} = \frac{f_5^{(1)}-f_7^{(1)}}{2}\zeta_4^{-1}$	c_7	$f_6^{(3)} = \frac{f_6^{(2)}-f_7^{(2)}}{2}\zeta_2^{-0}$	$= c_7$

$$= \frac{1}{m}\sum_{k=0}^{m-1}\underbrace{\frac{f_k+f_{k+m}}{2}}_{f_k^{(1)}}\zeta_m^{-\ell k},$$

während für ungeraden Index $j = 2\ell + 1$ anlog

$$c_{2\ell+1} = \frac{1}{m}\sum_{k=0}^{m-1}\frac{f_k - f_{k+m}}{2}\zeta_n^{-(2\ell+1)k} = \frac{1}{m}\sum_{k=0}^{m-1}\underbrace{\frac{f_k - f_{k+m}}{2}\zeta_n^{-k}}_{f_{m+k}^{(1)}}\zeta_m^{-\ell k}$$

folgt. Statt einer Fourier-Transformation der Länge n hat man nun also zwei Fourier-Transformationen der Länge $n/2$, eine für die Koeffizienten mit geradem Index und eine für die Koeffizienten mit ungeradem Index. Ist n nicht nur gerade, sondern eine zweier Potenz $n = 2^p$, lässt sich dieser Prozess iterieren, was in Tabelle 3 für $n = 2^3 = 8$ examplarisch demonstriert wird.

Geht man wieder davon aus, dass die Potenzen der Einheitswurzeln vorliegen, so ergibt sich für die Anzahl der komplexen Multiplikationen und Additionen offenbar $M(n) = n/2 + 2M(n/2)$, bzw. $A(n) = n + 2A(n/2)$, was sich beides zu $\mathcal{O}(n \log n)$ auflösen lässt. Die Anzahl der Multiplikationen ist tatsächlich noch geringer, wenn man berücksichtigt, dass in jedem Schritt $\zeta^{-0} = 1$ vorkommt. Dies ändert aber nicht das asymptotische Verhalten.

Index

Literatur

[1] Brill, M. *Mathematik für Informatiker.* Hanser, 2001.

[2] Drmota, M., Gittenberger, B., Karigl, G., and Panholzer, A. *Mathematik für Informatik.* Berliner Studienreihe zur Mathematik – Band 17, 2007.

[3] Hachenberger, D. *Mathematik für Informatiker.* Pearson, 2007.

[4] Hartmann, P. *Mathematik für Informatiker.* Vieweg, 2003.

[5] Schaback, R., and Wendland, H. *Numerische Mathematik, 5. Auflage.* Springer, 2004.

[6] Teschl, G. und Teschl, S. *Mathematik für Informatiker 1: Diskrete Mathematik und Lineare Algebra.* Springer, 2006.

[7] Teschl, G. und Teschl, S. *Mathematik für Informatiker 2: Analysis und Statistik.* Springer, 2006.

[8] Wolff, M. *Übungsaufgaben zur Mathematik für Informatiker und Bio-Informatiker.* Springer, 2005.

[9] Wolff, M. P., Hauck, P., and Küchlin, W. *Mathematik für Informatik und BioInformatik.* Springer, 2004.